Bayesian Modeling of Cell Neighborhoods from ClumpSeq Data



Atanasiu Stefan Demian

St Peter's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2023

Acknowledgements

First and foremost I would like to acknowledge and thank Edward Morrissey and Marella de Bruijn for their continuous supervision, advice and patience with my project. I feel lucky to have two supervisors with different backgrounds, as I was able to receive advice, knowledge and opinions from both the quantitative and biological points of view. I am especially grateful to Ed for regular meetings and support even after his leaving the WIMM. I would also like to thank Giorgio Anselmi for his help and support with describing the experimental side of the project and in general helping me better understand the biology of EHT.

I am also glad I had the opportunity to spend three months on an internship at BenevolentAI, for which I thank Sam Abujudeh and Meltem Gürel for their support and mentorship.

I enjoyed these last four years in Oxford and I feel lucky to have met so many wonderful and talented people here. Last but not least, I thank my mother for her unlimited support and always believing in me.

Abstract

Cells are often going through many complex dynamical changes, which change their function and transcriptional profile. It is often the case that this process is supported or initiated by neighboring cells in the tissue. The goal of this thesis is to better understand the neighborhood composition of cells and quantitatively describe its effects on the cells' transcriptional activity.

Single-cell RNA sequencing methods have proven valuable in providing high resolution, high throughput and unbiased whole-transcriptome gene expression measurements in individual cells. However, the information about the spatial proximity between cells is lost in single-cell sequencing experiments due to tissue dissociation. We thus focus on a newly developed exploratory tool called ClumpSeq, which involves sequencing small clusters of naturally attached cells via conventional single-cell sequencing protocols.

The focus of this thesis is computational and centered on developing deep learning models for analyzing ClumpSeq data. We validate our model on both simulated and publicly available data to show that it can extract cell type specific information from clump data. Based on these results we show that we can obtain novel biological insights that are otherwise not available from single-cell data.

Furthermore, we apply and adapt the ClumpSeq protocol to study the process of developmental hematopoiesis in the embryo, which is known to be influenced by the surrounding microenvironment. We generate a novel ClumpSeq dataset and show that our computational model provides insights about the composition of the niche surrounding the emerging hemogenic endothelium.

Contents

Contents **i**

1 Introduction **1**

1.1 Cells Interact with the Surrounding Microenvironment 1

1.2 Single-Cell RNA Sequencing Methods 2

1.3 Spatial Sequencing Methods 5

1.4 Sequencing Clumps of Attached Cells via ClumpSeq 7

1.5 Computational Analysis of Single-Cell Data 10

1.6 Thesis Outline and Objectives 12

2 Deep Generative Models for Single-Cell Data **14**

2.1 Background on Generative Models 14

2.2 VAEs for Single-Cell Data 19

2.3 Downstream Analysis 23

3 Cell Type Deconvolution of Clumps **26**

3.1 Single-Cell Generative Model 26

3.2 The Clump Generative Model 29

3.3 Related Work 32

3.4 Validation on Simulated Data 33

3.5 Comparing the Negative Binomial Parameterizations 35

3.6 Ablation Study of the Reference Dataset 37

3.7 Discussion 38

4 Clump Sequencing from the Intestinal Epithelium **41**

4.1 Background . . . 41

4.2 Single Cell Analysis . . . 44

4.3 Clump Model Evaluation . . . 46

4.4 Downstream Analysis . . . 50

4.4.1 Selective Mapping of Retrieved Enterocytes . . . 50

4.4.2 Zonation of Goblet genes . . . 52

4.4.3 Differentially Expressed Genes Within the Enterocyte Population 56

4.5 Co-expression Between Cell Types . . . 58

4.5.1 Background: Identifying Gene Programmes in Single Cells via Hotspot . . . 58

4.5.2 Application of Hotspot to the Single-Cell Gut Data . . . 60

4.5.3 Gene Co-expression in Clumps . . . 60

4.6 Discussion . . . 65

5 Ordering Cells in Pseudotime **68**

5.1 Background: Pseudotime . . . 68

5.2 Single-Cell Pseudotime Generative Model . . . 71

5.3 Application to the Gut Single-Cell Data . . . 75

5.4 Time Deconvolution from Clumps . . . 78

5.5 Identifying Gene Expression Modules Across Pseudotime . . . 80

5.6 Pseudotime, Identifiability and RNA Velocity . . . 87

5.7 Discussion . . . 89

6 Endothelial-to-Hematopoietic Transition **92**

6.1 Background . . . 92

6.2 Microenvironment Plays a Key Role in EHT . . . 95

6.3 Single-Cell and Clump Sequencing via Smart-seq3 . . . 97

6.3.1 Experimental Methods . . . 97

6.3.2 Experimental Details of Clump Generation . . . 100

6.3.3 Single-Cell Analysis . . . 101

6.3.4 Improving Mixing via Adversarial Classifiers 106
6.4 Clump Analysis 108
6.4.1 Composition of E9.5 Clumps 113
6.4.2 Differential Gene Expression in Clumps 114
6.5 Integration with 10x Data 116
6.5.1 Data Integration 117
6.5.2 Analysis 119
6.6 Integrated Model for Deconvolution of E11 Clumps 125
6.7 Discussion 129

7 Conclusions and Future Work **131**

Appendices **134**

Appendix A Data Processing **135**
A.1 Manco et al, 2021 Single-Cell and Clump Datasets 135
A.2 de Bruijn Group EHT Single-Cell and Clump Datasets 135

Appendix B Appendix for Chapter 5 **137**
B.1 β-VAE Rescales the Latent Space 137
B.2 Functional Profiling of Hotspot Modules from Section 5.5 142

Appendix C Appendix for Chapter 6 **144**
C.1 Summary of Clump Generation Experiment 144
C.2 Estimating Re-aggregation Rates 145
C.3 Mixing Metrics 147

References **149**

Chapter 1

Introduction

1.1 Cells Interact with the Surrounding Microenvironment

Cells are the basic units of life, in part because they are discrete and structured building blocks of the organism. All cells carry the same genetic information, yet they develop into a multitude of specialized populations and types, of various phenotypes and performing distinct functions. For instance, immune cells are involved in defense against bacteria, blood cells carry nutrients and oxygen, neurons send signals between each other via neurotransmitters etc. The genetic information inside cells (the DNA) can be considered a comprehensive and complex catalogue of instructions that cells use to conduct their functions. This is achieved via production of proteins; notably, each cell accesses only part of this genetic information in order to produce its characteristic proteins. These instructions are found in genes, and cells process them via a two-step operation, comprising transcription and translation. Briefly, transcription leads to the production of ribonucleic acid (RNA) that carries the information needed to build the protein. Out of the different types of RNA, messenger RNA (mRNA) carries the necessary coding sequences, called transcripts. These are later translated into amino acid chains which fold into functional proteins. The repertoire of different mRNA transcripts within a cell constitutes its transcriptome, and its composition will be defined by the cells' type and function. Importantly, the transcriptome comprises one of the most essential pieces of information about the cell.

The composition of the transcriptome is subjected to various dynamical changes - e.g. cells can differentiate from one type to another, or gain or lose certain functions as they mature etc. It is often the case that cells do not go through these processes by themselves, but are in fact found in frequent interaction with other nearby cells (Singer, 1992; Bloom & Zaman, 2014), which we refer to as the surrounding microenvironment. These cell-cell interactions play a key role in initiating or supporting biological processes. For instance, erythropoiesis is influenced by macrophages in the bone marrow, which were found to contribute to the erythroid proliferation and differentiation (Ramos et al., 2013). Tumours are known to be involved in consistent interaction with diverse nonmalignant cells, which impacts cancer progression and even response to treatment (Quail & Joyce, 2013).

In order to better understand the role of the microenvironment in regulating biological processes, it is necessary to understand its composition. For a given population of cells, we would like to describe what their neighbors are and how their morphological proximity affects their functions. This thesis will focus specifically on the research question of identifying and characterizing the neighborhood structure of a population. The focus of this thesis will be quantitative and adopting a data-driven approach to investigate particular types of data that provide information about the microenvironment composition. In addition to neighborhood information, we are interested in describing these interactions in terms of the cells' transcriptomic identity and how it is influenced by the microenvironment. The first step in tackling this problem is describing the type of data we require. Since we are interested in studying transcriptional profiles, our starting point is exploring RNA sequencing methods.

1.2 Single-Cell RNA Sequencing Methods

Identifying and characterizing cell types is fundamental to understanding any biological process. Even within one cell type population, cells can display a significant level of transcriptional heterogeneity, reflective of e.g. finer sub-types, cell cycle, inherent stochasticity etc. Understanding the properties of cells at an individual

level would provide a wealth of new information in terms of characterizing the cell types and molecular mechanisms contributing to their identity. An instance of such key information are genome-wide mRNA transcripts. Up until recently, mRNA sequencing studies collected bulk RNA-seq data, which consisted of capturing clusters of cells (each with size ranging from dozens to hundreds or thousands of cells) from a tissue. This process was later developed and optimized to sequence individual cells, thus leading to the emergence of single-cell RNA sequencing (scRNA-seq) technologies, which provide insight into the cells' transcriptomic profile (Tang et al., 2009). In this regard, scRNA-seq methods provide high-throughput and high-resolution measurements of individual cells, which have led to new findings on the various biological factors making up their identity (Kolodziejczyk et al., 2015; Wagner et al., 2016). Examples of new insights provided by scRNA-seq measurements include characterizing the heterogeneous composition and multiple sub-types of melanoma tumours (Tirosh et al., 2016) and in the mammalian cerebral cortex (Zeisel et al., 2015), heterogeneous splicing behaviour within one cell type population (Shalek et al., 2013), identifying rare cell types in the intestine (Grün et al., 2015) and many more applications.

In the last decade there has been a tremendous increase in the number of single cells sequenced in one study, ranging up to hundreds of thousands, along with affordable cost and decrease in manual labour (Svensson et al., 2018). This evolution is attributed to improvement of multiple factors involved in the process of sequencing, which consists of several steps. The tissue is first dissociated and cells are isolated into suspensions; a key component in achieving high throughput is optimizing this step in terms of the number of cells. Early methods used manual picking (Tang et al., 2009; Hashimshony et al., 2012) or fluorescence-activated sorting (FACS) (Jaitin et al., 2014) that isolated cells into plates or microfluidic chips (Ramsköld et al., 2012; Islam et al., 2013; Picelli et al., 2014; Hashimshony et al., 2016;). Subsequently, automated and more efficient procedures were developed, resulting in technologies such as droplet-based microfluidics (Mazutis et al., 2013; Macosko et al., 2015; Zheng et al., 2017) capable of processing tens of thousands of samples. Cells contain a very

small amount of mRNA, thus transcribed mRNA is converted into cDNA via reverse transcription and is amplified. First strand of cDNA is usually synthesized using an oligo(dT) primer that anneals to the poly(A) tail at the 3' end of the mRNA molecule. Sequencing methods differ mainly in the generation of the second cDNA strand, construction and amplification of the libraries (Chen et al., 2018b).

Library amplification is usually carried out via PCR. Due to the small mRNA amount, the PCR step is carried out multiple times, resulting in amplification biases. To mitigate this, many protocols now use Unique Molecular Identifiers (UMI) that are added to the cDNA molecules before amplification in order to tag transcripts (Islam et al., 2013). This removes amplification biases as the PCR duplicates will share the same UMI as the original molecule. Efficiency is also enhanced by multiplexing, where cell-specific barcodes are incorporated, usually during the reverse transcription step, and hence multiple cells can be pooled and processed downstream together (Hashimshony et al., 2012).

In practice, a trade-off that is usually encountered in choosing a sequencing technology is between throughput and sequencing depth. Technologies with isolation methods based on droplets or microwells, such as 10x (Zheng et al., 2017) result in higher throughput than FACS or plate-based methods such as Smart-seq2 (Picelli et al., 2014) or Smart-seq3 (Hagemann-Jensen et al., 2020). On the other hand, the latter methods result in many more transcripts captured per cell at the cost of lower throughput. As an example, 10x data contains most often at most a few thousand UMIs per cell, whereas Smart-seq3 can reach even over 100,000 UMIs per cell (Hagemann-Jensen et al., 2020).

There are several measures to assess the quality of scRNA-seq technologies. These include sensitivity (ability to capture low abundance transcripts), accuracy (how close estimated abundances are to the true ones) and sequencing depth. Comprehensive studies of performances of various scRNA-seq technologies are found in (Ziegenhain et al., 2017) and (Svensson et al., 2018). Choosing a sequencing technology for a study always involves a trade-off, and the suitability of any sequencing method is dependent on the study and cell population.

One of the advantages of scRNA-seq experiments is they can capture specific populations enriched via fluorescent expression of cell surface markers. This is important in many biological studies, as it allows us to sort specific cell populations of interest (a specific instance of this will be described in detail in Chapter 6, as well as the trade-off in choosing sequencing methods). However, while scRNA-seq provides key information about the transcriptomic identity of individual cells, their location and information about their proximity to each other is lost due to tissue dissociation. This makes the objective of studying cell interactions challenging or prohibitive. Therefore, we are interested in an additional experimental method that could capture this information. A likely candidate is provided by spatial sequencing methods.

1.3 Spatial Sequencing Methods

Capturing the spatial information of cells can provide key information about phenotype, spatial patterns of gene expression and interactions with other cell types. This requires the ability to conduct transcriptomic studies on intact tissue. In recent years, multiple commercial platforms for spatially resolved measurements of gene expression have become available. These technologies are designed to measure gene transcripts at multiple spatial locations in a tissue. They mainly differ in terms of resolution and technical aspects, such as tissue size, number of genes (ranging from a handful to whole-genome measurements) and size of the spatial units being processed (ranging from a tissue slice or region to an individual cell). As will be discussed below, spatial sequencing methods usually involve a trade-off between the number of genes and spatial resolution.

Broadly speaking, the types of spatially resolved measurements of transcriptomic information can be grouped in two categories: imaging and sequencing based technologies (Moses & Pachter, 2022). The first category consists of *in situ* microscopy imaging of mRNA. These methods use *in situ* hybridization of mRNA via labelled probes that are detected with microscopy. The first quantitative method for imaging these fluorescent probes was developed in smFISH (Femino et al., 1998), where

individual transcripts can be visualized and appear as fluorescent points. Since the first publication, it has been optimized for multiple gene detection (Raj et al., 2008). smFISH maps the spatial location of mRNA transcripts and has nearly 100% sensitivity (Chen et al., 2018b), since it captures almost all endogenous transcripts and requires no amplification. Modern imaging-based methods are improved variants of smFISH, e.g. seqFISH (Lubeck et al., 2014) and MERFISH (Chen et al., 2015), that allow for more genes to be simultaneously measured.

Sequencing-based technologies, making up the second category, consist of capturing mRNA from the tissue, while preserving spatial information, and proceeding with sequencing. There are multiple ways of retaining spatial information. One of the first methods involved Laser Capture Microdissection (LCM; Emmert-Buck et al., 1996) of tissue sections, which were then sequenced with bulk RNA-seq (Zechel et al., 2014) or scRNA-seq (Geo-seq; Chen et al., 2017). While LCM studies can provide whole-transcriptome measurements, the tissue isolating method is quite laborious and prone to mRNA degradation due to the laser application. Later methods retained spatial information by using spatially-barcoded probes positioned on an array. The tissue was mounted on the array and the probes labelled the position of the mRNA (Spatial Transcriptomics; Ståhl et al., 2016). Each unique positional barcode defines a hexagonal spot of cells it captures. This method was later commercialized by 10x Genomics and renamed Visium, and currently has a spatial resolution of $55\mu m$ diameter spots; this can capture between approximately 1 and 30 cells per spot, depending on the cell type (Saiselet et al., 2020).

Detailed reviews of spatial methods can be found in (Rao et al., 2021; Moses & Pachter, 2022; Williams et al., 2022). All types of spatial sequencing methods described here have particular advantages and disadvantages. Imaging based technologies provide spatial information of individual cells and very accurate measurements of mRNA levels; however, the measurements are not genome-wide and the genes have to be selected beforehand. Array based methods such as spatial transcriptomics (Visium) can capture larger tissue sections, provide whole-genome

measurements and retain spatial information, but they have lower sensitivity than FISH methods and do not have single cell resolution.

All three sequencing methods described in this section are summarized in Figure 1.1a. In line with our research objective of studying cell interactions and communication via RNA sequencing, no category fully satisfies the requirements of single cell resolution and preserving cell neighborhood information. Bulk RNA-seq methods do not provide single cell resolution, and scRNA-seq does not preserve neighborhood information. Even if they could be optimized to achieve single cell resolution, array based spatial methods have much lower throughput than scRNA-seq technologies and cannot be enriched for specific cell type populations. For this reason, we focused on a newly developed exploratory method that comprises both advantages of scRNA-seq and information on cell proximity, called ClumpSeq (Manco et al., 2021).

1.4 Sequencing Clumps of Attached Cells via ClumpSeq

As discussed previously, scRNA-seq does not provide information about the proximity of cells relative to each other. This information is crucial in understanding whether and how different cell types interact at gene expression level. Therefore, we studied ClumpSeq, an approach that consists of sequencing small clusters of naturally attached cells, called clumps, via a light dissociation protocol. This results in clumps of 2-10 cells (Manco et al., 2021), but it can also be optimized for capturing cell pairs (Halpern et al., 2018), that are sequenced via conventional scRNA-seq protocols. A schematic of ClumpSeq is shown in Figure 1.1b.

In the original publications from the Itzkovitz lab, the authors generated ClumpSeq datasets of attached cells, as described above, in order to study rare cell types. Many cell types exhibit spatially variable gene expression patterns, a behaviour called zonation. This is difficult to characterize for rare cell types via spatial sequencing, as their low mRNA abundance in multi-cellular samples is diluted.

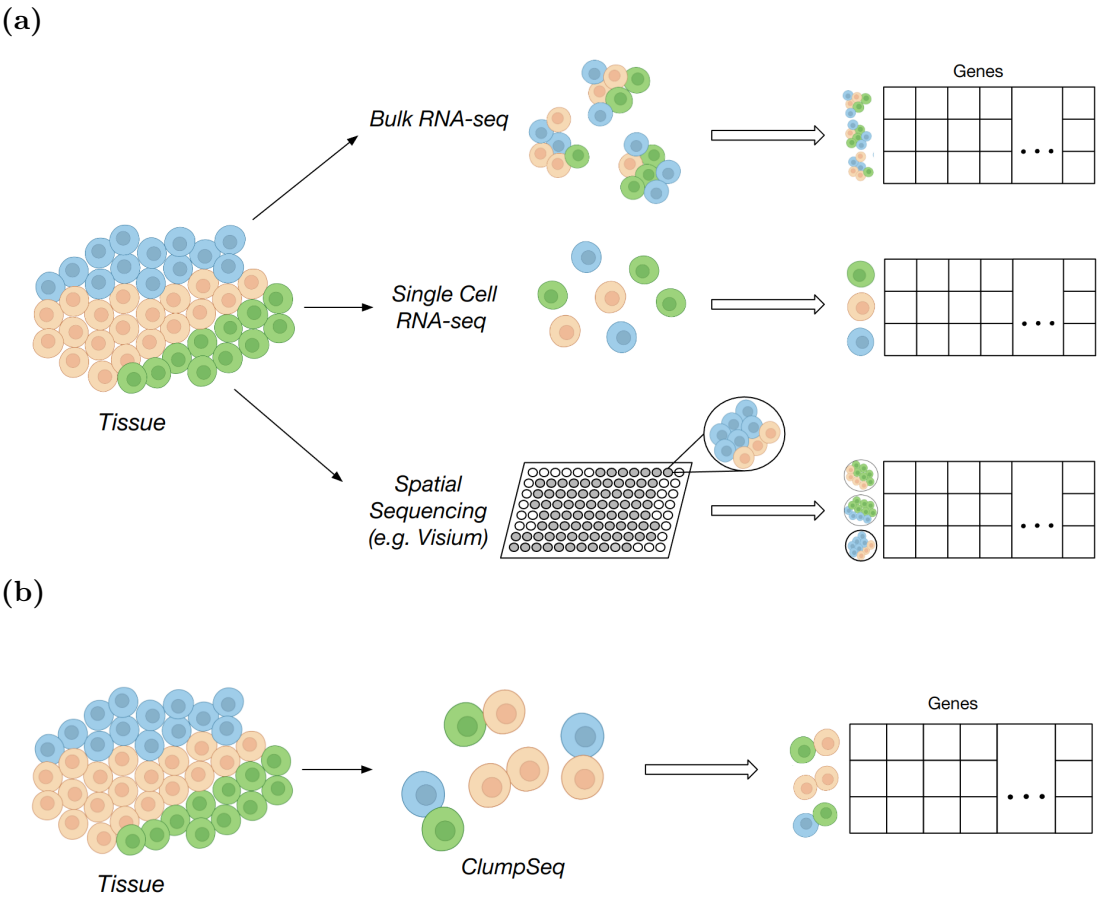


Figure 1.1: Summary and schematic of main RNA sequencing methods. Starting from a tissue (left), each method involves its own procedure for collecting samples, with the output being a gene expression matrix (right) (a) The most widely employed ones are bulk RNA-seq (no single cell resolution and no spatial information), scRNA-seq (single cell resolution but no spatial information) and spatial sequencing methods, Visium shown here as example (preserves spatial information but no single resolution); (b) Newly explored ClumpSeq method, involving light dissociation to capture clumps of cells (we focus on clumps of 2 cells), thus retaining information on cell proximity and providing high resolution. Created with BioRender.com.

They apply this framework to study zonation patterns of rare cell types in the liver (Halpern et al., 2018) and intestinal epithelium (Manco et al., 2021).

We highlight several advantages of the ClumpSeq framework and how it relates to our research objective. First, ClumpSeq contains information about cell neighborhood since the clumps consist of cells that were attached in the tissue. Second, the sequencing protocol leads to high throughput on the same scale as scRNA-seq

methods. Third, we can enrich for cell type populations using fluorescent cell surface markers. This aspect is particularly useful when studying a cell type population that can be enriched by known fluorescent markers and whose proximal neighbors in the tissue are unknown. Fourth, unlike spatial sequencing methods, it does not require isolating physical tissue slices and it can cover a larger range of the tissue.

The focus of this thesis will be on ClumpSeq data, specifically when clumps are made up of two cells. ClumpSeq data can provide important information about how cell types interact with each other, especially when it contains two different cell types. While single-cell data provides information about each cell type population, a key question is whether and how these populations relate and interact with each other at transcriptional level. Another question is whether there exist gene expression patterns conditional on interactions. For instance, for two cell types A and B, we are interested in whether A contains a specific subtype that is attached to cells from B and expresses specific genes only when attached to this cell type. These questions could be extended to developmental processes and cell fate decisions, e.g. whether cells of type A influence cell fate decisions of type B cells whenever they interact. The key point to be made here is that scRNA-seq data does not provide this information. On the other hand, we argue that ClumpSeq data can provide novel insights into these queries.

Having established the experimental method and type of data used for tackling our research question, the next step is developing quantitative methods for studying and modeling clump data. Since a clump is made up of more than once cell, its gene expression measurements will be a sum of signals from each cell. This sum is also influenced by the relative mRNA proportions originating from each cell, which are most likely unequal. For instance, if one cell is physically bigger than the other, it is expected that the sequencing method will capture more mRNA transcripts from the larger cell. Therefore, reconstructing expression of a single cell type from clump data is not trivial.

The next step is therefore establishing the quantitative framework for analysis. A key aspect and basis of our modeling framework is that scRNA-seq provides

the highest resolution representation of cell populations, and analysis of other multicellular modalities (i.e. ClumpSeq) will rely on the quantitative representations obtained from single-cell data. Therefore, we begin with discussing the challenges and computational methods employed for modeling scRNA-seq data, in the upcoming section.

1.5 Computational Analysis of Single-Cell Data

The standard format of single-cell data for analysis is a *cells* $\times$ *genes* expression count matrix. Throughout this thesis we will denote the individual cell expression (matrix row) by $\boldsymbol{x} \in \mathbb{N}^G$, where G is the total number of genes used for analysis. The total sum of gene counts per cell is the library size. Due to the experimental stochasticity in capturing transcripts, each cell has different values of library size, thus gene counts by themselves are not comparable across cells. Therefore, mRNA abundance is best quantified via relative gene proportions, obtained via normalizing the counts by the library sizes, into proportions that sum up to 1 for each cell.

Analysis of single-cell data is difficult, and the challenge is extracting biological signals while accounting for technical variation. This variation includes experimental factors such as fraction of reverse-transcribed molecules, amplification efficiency and differences in sequencing depth. Often single cell studies are conducted across multiple experiments, performed at different times, resulting in multiple batches of data; despite being collected from the same cell type populations, single-cell measurements can show variation between the experiments, known as *batch effects* (Tung et al., 2017). There is also noise due to the inherent stochasticity of gene expression (Raj & Oudenaarden, 2008). Since scRNA-seq provides whole-transcriptome measurements, each cell contains expression values from thousands of genes. Many genes also exhibit sparse values; not all transcripts are captured, therefore the corresponding gene measurements are mostly zeros. For these genes, this reflects the experimental limitation rather than their absence due to biology; this is known as dropout (Stegle et al., 2015). Genes are also not independent: there are groups

of genes whose expression is strongly correlated. All these aspects make analysis challenging and thus developing computational methods for extracting accurate and interpretable biological signals from single-cell data is vital.

In order to account for differences in library size, analysis often involves a normalization step, which conventionally consists of standardizing expression across cells by dividing by library size, and multiplying the resulting proportions by a fixed value, usually the mean or median library size of the dataset. However, this method can lead to loss of information on technical variability due to sequencing depth and might yield spurious or biased results (Vallejos et al., 2017).

Due to the high dimensionality of scRNA-seq data, analysis requires a low-dimensional representation of dimension $D \ll G$. This is meant to summarize the main biological signals in the data while removing technical variation, and is used for downstream analysis, i.e. clustering, visualization, cell type annotation, differential expression, imputation etc. This step is often carried out via PCA, which identifies orthogonal directions of largest variation. A more probabilistic approach, which models samples $\boldsymbol{x}$ as realizations of a random variable with a specified distribution, could be adopted via Factor Analysis or probabilistic PCA. This brings up another important aspect, namely the distributional assumptions for the data. Normalized data is usually modelled via a Normal distribution (e.g. Buettner et al., 2015; Pierson & Yau, 2015; Prabhakaran et al., 2016), since model fitting is easier to carry out with Normal random variables.

The gene expression counts usually exhibit variance larger than the mean, i.e. overdispersion compared to Poisson variation, thus a suitable discrete distribution to model them is the Negative Binomial (Grün et al., 2014; Vallejos et al., 2015; Risso et al., 2018). It is parameterized by $r > 0$ and $p \in [0, 1]$, with probability mass function given by

$$p_{\text{NB}}(x; r, p) = \frac{\Gamma(x + r)}{\Gamma(x + 1)\Gamma(r)}(1 - p)^x \; p^r, \;\; x \in \mathbb{N} \tag{1.1}$$

It can be shown that the Negative Binomial distribution is equivalent to a Gamma-Poisson mixture, from a Gamma random variable with shape r and scale

$\frac{1-p}{p}$. The Gamma step introduces further noise, thus samples drawn from this distribution will have larger variance than Poisson variates. Due to this property, the Negative Binomial distribution has been found to perform well in modeling gene expression counts. For instance, ZINB-WAVE (Risso et al., 2018) considers a distribution that incorporates the effect of gene dropout, called Zero-Inflated Negative Binomial, formulated as a mixture:

$$f_{\text{ZINB}}(x; r, p, \pi) = \pi\delta(x) + (1 - \pi) f_{\text{NB}}(x; r, p) \tag{1.2}$$

where $\delta(\cdot)$ is the Dirac delta function and π is an additional parameter quantifying the probability of observing a 0 due to dropout.

The parameters of ZINB-WAVE, and generally of many models for scRNA-seq, are assumed to be linear functions of the data. Gene expression data can encode complex biological signals, related to e.g. cell subtypes, development, fate choice etc., and therefore the signals could be highly non-linear and not captured accurately by a linear model.

In this thesis, our goal is to work with and develop robust probabilistic models for quantifying important biological signals, as will be detailed below. We focus on an end-to-end framework, where the model output can be used directly for downstream analysis without relying on external methods, which require their own assumptions and might involve further transformations of the data.

1.6 Thesis Outline and Objectives

The aims of this thesis are centered around developing a quantitative model for analyzing ClumpSeq data. From a methodological perspective, this will consist of two main objectives: inferring the cell type composition and cell type specific information from clumps. We adopt a rigorous and probabilistic approach, in order to quantify the variation and noise in gene expression data. For this reason, we are interested in developing generative models that can capture the complex patterns in scRNA-seq data, which will then be extended and tailored for ClumpSeq measurements.

From a biological perspective, our goal is to gain a better understanding of how cells' behaviour is influenced by their neighbors. Since cells consistently exhibit communication patterns between each other and can even be influenced in their fate decisions by the surrounding environment, such findings can provide a better understanding of how these interactions occur. In this thesis we argue that ClumpSeq data can provide such insights, that are otherwise not available from single-cell data. We note that in the context of ClumpSeq, we refer here to a cell type's "microenvironment" or "niche" as relative to that specific cell type. For instance, if we describe the interaction of enterocytes with their niche, here we use the term to refer to cells directly in contact with enterocytes. In addition to development and methodology, we validate our models on experimental ClumpSeq data and investigate the biological conclusions that we can obtain from their application.

The thesis is structured as follows. This chapter provided background on our research question, candidate experimental methods to tackle it and computational methods that have so far been developed for analyzing this data. Chapter 2 provides a quantitative background on generative models, which will be the focus of our model development. We will focus specifically on deep learning models and highlight why they are suitable and efficient in modeling scRNA-seq data. Chapter 3 begins describing our contributions; the focus of this chapter is building up on the deep learning framework to develop a generative model specific to clump data and validate it on simulated data. Chapters 4 and 5 will look at applying our model to the experimental ClumpSeq data from the original publication (Manco et al., 2021). Chapter 6 will introduce a novel ClumpSeq dataset generated in our lab for studying developmental hematopoiesis and will focus on the insights from the clump model. Finally, a summary of conclusions and directions for future work will be presented in Chapter 7.

Chapter 2

Deep Generative Models for Single-Cell Data

2.1 Background on Generative Models

Our approach towards modeling scRNA-seq data is probabilistic - we assume each observed measurement $\boldsymbol{x}$ is a realization of a random variable generated from a sampling distribution of gene counts, with density $p(\boldsymbol{x})$. This represents the true and unknown generative distribution of the data. In a modelling framework, we formulate a sampling distribution with density $p_\theta(\boldsymbol{x})$, indexed by parameters θ, intended to estimate the real data generating process. As mentioned in the previous chapter, scRNA-seq is highly dimensional and noisy, comprising multiple sources of variation, thus making it prohibitive to formulate $p_\theta(\boldsymbol{x})$, let alone fit the data to it. Therefore, we make this task more tractable by focusing on *latent variable models*: we assume $\boldsymbol{x}$ is generated by an unobserved low-dimensional latent variable $\boldsymbol{z} \in \mathbb{R}^D$ via a conditional distribution $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ with parameters θ. In the context of modeling scRNA-seq data, $\boldsymbol{z}$ summarizes the main biological variation in the data and ideally contains no technical variation, which is incorporated in the conditional distribution $\boldsymbol{x}|\boldsymbol{z}$. Computing the low-dimensional representation is a key step in any analysis framework, as this will be used for downstream analysis.

We formulate the above as a generative model: we are interested in modeling the distribution of *both* variables $\boldsymbol{z}$ and $\boldsymbol{x}|\boldsymbol{z}$. We perform inference (i.e. fitting the data to the model) by optimizing the joint likelihood $p_\theta(\boldsymbol{x}, \boldsymbol{z})$, known as the

generative distribution. We note that by integrating over all possible values of $\boldsymbol{z}$ we recover the sampling distribution $p_\theta(\boldsymbol{x})$, referred to as the marginal likelihood. We adopt a Bayesian inference approach, which consists of estimating the *posterior distribution* $p_\theta(\boldsymbol{z}|\boldsymbol{x})$. Assuming a prior distribution $p(\boldsymbol{z})$ for latents, the posterior updates our knowledge of the latent space structure after observing the data $\boldsymbol{x}$, hence leading to an informative representation.

We assume samples are independent and identically distributed, and conditional independence of the data given latents, which allows the joint density to be written as:

$$p_\theta(\boldsymbol{x}, \boldsymbol{z}) = \prod_{n=1}^{N} p(\boldsymbol{z}_n) p_\theta(\boldsymbol{x}_n|\boldsymbol{z}_n) \tag{2.1}$$

Designing generative models involves formulating the (conditional) generative distribution of the data and latent prior. For the former we can choose a distribution suitable for gene counts, such as Negative Binomial, while the latents z are continuous variables and they usually have a $\mathcal{N}(0, I)$ prior to simplify computations.

The next step is fitting the generative model to the data by computing the posterior distribution $p_\theta(\boldsymbol{z}|\boldsymbol{x})$. By Bayes' rule, this is given by

$$p_\theta(\boldsymbol{z}|\boldsymbol{x}) = \frac{p_\theta(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})}{p_\theta(\boldsymbol{x})} \tag{2.2}$$

Apart from a few limited cases of conjugate distributions, the posterior in (2.2) is intractable due to the marginal likelihood term in the denominator, which requires integration over $\boldsymbol{z}$. Therefore, approximate inference methods are required; we focus specifically on using Variational Inference (Jordan et al., 1999; Blei et al., 2017). This consists of formulating a parametric distribution $q_\phi(\boldsymbol{z})$, with parameters ϕ, of a form that is likely to approximate the true posterior, and searching for optimal parameters such that it is as "close" as possible to the true posterior; $q_\phi(\boldsymbol{z})$ is called *variational posterior*. The measure of similarity between the two distributions is quantified via the *Kullback-Leibler (KL) divergence*:

$$\mathrm{KL}(q_\phi(\boldsymbol{z})||p_\theta(\boldsymbol{z}|\boldsymbol{x})) = -\mathbb{E}_q\left[\log \frac{p_\theta(\boldsymbol{z}|\boldsymbol{x})}{q_\phi(\boldsymbol{z})}\right] \tag{2.3}$$

$$= -\int \log\left(\frac{p_\theta(\boldsymbol{z}|\boldsymbol{x})}{q_\phi(\boldsymbol{z})}\right) q_\phi(\boldsymbol{z})d\boldsymbol{z} \tag{2.4}$$

$$= -\int \log\left(\frac{p_\theta(\boldsymbol{z},\boldsymbol{x})}{q_\phi(\boldsymbol{z})p_\theta(\boldsymbol{x})}\right) q_\phi(\boldsymbol{z})d\boldsymbol{z} \tag{2.5}$$

$$= -\int \log\left(\frac{p_\theta(\boldsymbol{z},\boldsymbol{x})}{q_\phi(\boldsymbol{z})}\right) q_\phi(\boldsymbol{z})d\boldsymbol{z} + \int \log p_\theta(\boldsymbol{x}) q_\phi(\boldsymbol{z})d\boldsymbol{z} \tag{2.6}$$

$$= -\mathbb{E}_q\left[\log\left(\frac{p_\theta(\boldsymbol{z},\boldsymbol{x})}{q_\phi(\boldsymbol{z})}\right)\right] + \log p_\theta(\boldsymbol{x}) \tag{2.7}$$

$$= -\mathcal{L}(\boldsymbol{x};\theta,\phi) + \log p_\theta(\boldsymbol{x}) \tag{2.8}$$

KL divergence is always non-negative and achieves minimum of 0 if and only if $p \equiv q$. The above expression can thus be written as a lower bound on the marginal likelihood:

$$\log p_\theta(\boldsymbol{x}) \geq \mathcal{L}(\boldsymbol{x};\theta,\phi) \tag{2.9}$$

$$= \mathbb{E}_q[\log p_\theta(\boldsymbol{x},\boldsymbol{z}) - \log q_\phi(\boldsymbol{z})] \tag{2.10}$$

$$= \mathbb{E}_q[\log p_\theta(\boldsymbol{x},\boldsymbol{z})] + \mathbb{H}[q_\phi(\boldsymbol{z})] \tag{2.11}$$

While the marginal likelihood on the left-hand side of the above expression is intractable to compute, the objective of the variational posterior distribution is to maximize the fit to the data, i.e. be as close as possible to $p_\theta(\boldsymbol{x})$. Based on this, the inequality can be cast as an optimization problem for function $\mathcal{L}(\boldsymbol{x};\theta,\phi)$ of variational parameters ϕ. This term is called ***E**vidence **L**ower **Bo**und (ELBO)*. In practice, the generative parameters θ are often also unknown, in which case the ELBO is a function of both θ and ϕ. The variational inference framework involves an equivalence between the Bayesian inference of posterior $p_\theta(\boldsymbol{z}|\boldsymbol{x})$ and the optimization problem involving $q_\phi(\boldsymbol{z})$. This process consists of choosing a family of distributions for q and optimizing the parameters to maximize the ELBO, which is equivalent to minimizing the KL divergence in (2.3).

We have so far described the structure and inference scheme for the latent variable modeling framework that will be used for single-cell data. Given the highly nonlinear signals in the data, we would like to choose a generative distribution that can accurately capture these signals. As mentioned in the previous chapter,

linear models do not reach the desired levels of expressivity. For this reason, we focused on *Variational Autoencoders* (VAEs; Kingma & Welling, 2013; Rezende et al., 2014), where the variational posterior parameters are specified by neural networks (Goodfellow et al., 2016). VAEs are part of the broader class of deep generative models. For each $\boldsymbol{x}$, we formulate the variational posterior $q_\phi(\boldsymbol{z}|\boldsymbol{x})$, conditioned on $\boldsymbol{x}$, as:

$$\boldsymbol{z}|\boldsymbol{x} \overset{q}{\sim} \mathcal{N}(\mu(\boldsymbol{x}); \text{diag}(\sigma^2(\boldsymbol{x}))) \tag{2.12}$$

where $\mu(\cdot), \sigma(\cdot)$ are usually chosen to be fully-connected multi-layer perceptrons (MLP), mapping from gene expression space onto the latent space. Expression (2.12) is a mean-field assumption, where each latent dimension has its own Normal distribution. We can also incorporate MLPs into the likelihood $p_\theta(\boldsymbol{x}|\boldsymbol{z})$; in this case, θ will be the network weights and also treated as an unknown parameter, jointly optimized with ϕ in the ELBO.

A notable aspect of VAEs is the framework of *amortized* variational inference (Gershman & Goodman, 2014) - it employs functions that map between data points and latents. The mapping from the data space (i.e. gene counts) onto the latent space is called *encoder*; in the above case, the encoder computes the tuple μ, σ^2. We thus employ the specific notation $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ to denote the (conditional) variational distribution for $\boldsymbol{z}$ (also called *inference model*). The function responsible for the reverse mapping is called *decoder*. Thus, the parameters of the generative and variational distributions are each specified by a single function instead of data specific parameters.

VAEs are optimized via the ELBO loss objective (2.9), which can be re-written as:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{x}; \theta, \phi) &= \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}, \boldsymbol{z}) - \log q_\phi(\boldsymbol{z}|\boldsymbol{x})] && (2.13)\\
&= \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z}) + \log p(\boldsymbol{z}) - \log q_\phi(\boldsymbol{z}|\boldsymbol{x})] && (2.14)\\
&= \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] + \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\frac{p(\boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\right] && (2.15)\\
&= \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - \text{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})) && (2.16)
\end{aligned}$$

We highlight how optimization affects each term in (2.16). The first term is called the *reconstruction error*, formulated as the expected conditional likelihood

of data point $\boldsymbol{x}$ over latents $\boldsymbol{z}$ of the corresponding variational distribution. This incentivizes the model to fit the latents and generative distribution parameters to the data. The second term, the KL divergence, is minimized and thus brings together the variational posterior and the generative prior. This component can be seen as a regularization term such that the posterior is similar to the prior. Notably, since both distributions are Normal, the KL term can be computed in analytical form (Kingma & Welling, 2013):

$$\mathrm{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})) = \frac{1}{2}\sum_{i=1}^{d}(\sigma^2(\boldsymbol{x})_i + \mu^2(\boldsymbol{x})_i - 1 - \log(\sigma^2(\boldsymbol{x})_i)) \tag{2.17}$$

The ELBO is optimized jointly with respect to parameters θ and ϕ, via stochastic gradient descent. The ELBO and its gradient are intractable due to the expectation term in (2.16) ; in practice, this is estimated via Monte Carlo sampling. While unbiased estimates w.r.t θ are straightforward to obtain, computing unbiased gradient estimates w.r.t. ϕ is less straightforward, since the expectation is taken w.r.t. the distribution $q_\phi(\boldsymbol{z}|\boldsymbol{x})$, whose parameters are functions of ϕ. This problem can be mitigated via the reparameterization trick (Kingma & Welling, 2013):

$$\begin{aligned}
\nabla_\phi \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p(\boldsymbol{x}|\boldsymbol{z})] &= \nabla_\phi \mathbb{E}_{p(\boldsymbol{\varepsilon})}[\log p(\boldsymbol{x}|\boldsymbol{z})],\ \boldsymbol{z} = \mu(\boldsymbol{x}) + \sigma(\boldsymbol{x}) \odot \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim \mathcal{N}(0, I) \\
&= \mathbb{E}_{p(\boldsymbol{\varepsilon})}[\nabla_\phi \log p(\boldsymbol{x}|\ \mu(\boldsymbol{x}) + \sigma(\boldsymbol{x}) \odot \boldsymbol{\varepsilon})] \\
&\approx \frac{1}{L}\sum_{\ell=1}^{L} \nabla_\phi \log p\left(\boldsymbol{x}|\ \mu(\boldsymbol{x}) + \sigma(\boldsymbol{x}) \odot \varepsilon\right),\ \boldsymbol{\varepsilon}^{(\ell)} \sim \mathcal{N}(0, I)
\end{aligned} \tag{2.18}$$

Therefore, $\boldsymbol{z}$ is written as a deterministic function of random variable $\boldsymbol{\varepsilon}$ [1]. The expectation in (2.16) is then taken w.r.t $\boldsymbol{\varepsilon}$ and thus the Monte Carlo estimate is differentiable w.r.t. ϕ.

We highlight several appealing properties of VAEs and their usefulness for modeling single-cell data. First, incorporating neural networks in the generative and inference distributions offers the model flexibility to capture complex gene expression behaviours and their interactions. Second, due to the amortization scheme, VAEs

[1] Here $\odot$ denotes element-wise product.

easily scale to large data - for a fixed network architecture, the number of parameters is constant regardless of the data size. Furthermore, the training scheme can be conducted in sequential minibatches of data to save memory. Third, the neural network framework is flexible to incorporate more complex models that build upon it - this will be explored in detail in Chapter 3.

Having highlighted the main properties of VAEs, the next step is describing their application to single-cell data in more detail.

2.2 VAEs for Single-Cell Data

Recent work has explored the use of VAEs in modeling single-cell data (Ding et al., 2018; Lopez et al., 2018; Wang & Gu, 2018; Eraslan et al., 2019; Grønbech et al., 2020). We specifically focused on the scvi model (Lopez et al., 2018). We chose this for several reasons: the gene counts are directly modelled via a discrete-valued generative distribution, thus the input requires no normalization step; it performs batch correction by treating batch information as a nuisance variable and incorporating it into the model. We consider a dataset of N cells and G genes $\{(\boldsymbol{x}_n, s_n)\}_{n=1}^N$, with cell specific batch covariates s. The generative model is:

$$\boldsymbol{z}_n \sim \mathcal{N}(0, I) \tag{2.19}$$

$$\ell_n \sim \text{LogNormal}(\ell_\mu, \ell_\sigma^2) \tag{2.20}$$

$$\boldsymbol{\rho}_n = f(\boldsymbol{z}_n | s_n) \tag{2.21}$$

$$w_{ng} | \rho_{ng}, \theta_g \sim \text{Gamma}\left(\theta_g, \frac{\theta_g}{\rho_{ng}}\right) \tag{2.22}$$

$$k_{ng} | w_{ng}, \ell_{ng} \sim \text{Poisson}(\ell_n w_{ng}) \tag{2.23}$$

$$h_{ng} \sim \text{Bernoulli}(f_h(\boldsymbol{z}_n | s_n)_g) \tag{2.24}$$

$$x_{ng} \sim \begin{cases} k_{ng}, & \text{if } h_{ng} = 1 \\ 0, & \text{otherwise} \end{cases} \tag{2.25}$$

The authors define ℓ_n in (2.20) as a nuisance variable for library size, whose prior is given by a LogNormal distribution with mean and scale parameters computed from the empirical library sizes. Function $f(\cdot)$ in (2.21) is the decoder network; we assume

it maps the latents $\boldsymbol{z}_n$ into normalized gene proportions $\boldsymbol{\rho}_n$ via a softmax activation output layer, thus the output lies on the $(G-1)$-dimensional probability simplex. Variables w_{ng} and k_{ng} in (2.22) and (2.23) are part of a compound Poisson sampling scheme (Gamma distribution is parameterized by shape & rate). The underlying reasoning for this formulation is to quantify the overdispersion of gene expression in reference to Poisson variation. Finally, function $f_h(\cdot)$ is a neural network computing the probability of dropout as a parameter of a Bernoulli random variable. All these components lead to the sampling scheme of the gene count x_{ng}.

In our work we used the observed library size N_n instead of variable ℓ_n and did not model dropout (in which case steps (2.20) and (2.25) are excluded), in order to simplify computations in Chapter 3. The variable w_{ng} in the compound Poisson sampling scheme can be marginalized out to obtain a Negative Binomial density (dropping the subscripts):

$$
\begin{aligned}
p(x|\rho,\theta) &= \int p(x|w)p(w|\rho,\theta)\ dw && (2.26)\\
&= \int \frac{(Nw)^x}{\Gamma(x+1)}e^{-Nw}\ \frac{\theta^\theta w^{\theta-1}e^{-\frac{\theta}{\rho}w}}{\rho^\theta\Gamma(\theta)}\ dw && (2.27)\\
&= \frac{N^x\theta^\theta}{\Gamma(x+1)\rho^\theta\Gamma(\theta)}\int w^{x+\theta-1}\exp\left(-\frac{\theta+N\rho}{\rho}w\right)\ dw && (2.28)\\
&= \frac{\Gamma(x+\theta)}{\Gamma(x+1)\Gamma(\theta)}\left(\frac{\theta}{\theta+N\rho}\right)^\theta\left(\frac{N\rho}{\theta+N\rho}\right)^x && (2.29)\\
&= \text{NegativeBinomial}\left(x;\ \theta, \frac{\theta}{\theta+N\rho}\right) && (2.30)
\end{aligned}
$$

This formulation highlights the role of θ_g as gene-specific inverse dispersion parameter, quantifying the level of overdispersion against Poisson variation. We have:

$$
\begin{aligned}
\mathbb{E}[x_{ng}] &= N_n\rho_{ng} && (2.31)\\
\text{Var}(x_{ng}) &= \frac{(N_n\rho_{ng})^2}{\theta_g} + N_n\rho_{ng} && (2.32)
\end{aligned}
$$

highlighting the quadratic relationship between variance and mean. We note that as $\theta_g \to \infty$, the gene count random variable approaches Poisson distribution.

The formulation of the Negative Binomial likelihood in (2.30) is the mean & inverse dispersion parameterization. This provides a more interpretable formulation in the context of gene expression data compared to the conventional (r, p) parameterization. We can thus formulate the generative distribution as

$$x_{ng} \sim \text{NegativeBinomial}\left(N_n f(\boldsymbol{z}_n|s_n),\ \theta_g\right) \tag{2.33}$$

The decoder produces the reconstructed gene proportions $\boldsymbol{\rho}_n$, which are plugged into the parameters in (2.33). The log-likelihood of the data provides the reconstruction error term in (2.16).

The variational inference scheme has the same structure as outlined in the previous section. The parameters of the variational posterior are specified by the encoder: $q_\phi(\boldsymbol{z}_n|\boldsymbol{x}_n, s_n) = \mathcal{N}(\boldsymbol{z}_n; \mu(\boldsymbol{x}_n|s_n), \text{diag}(\sigma^2(\boldsymbol{x}_n|s_n)))$. The mean and variance networks share a hidden layer. A schematic of the VAE framework is shown in Figure 2.1a.

We now focus on the categorical batch covariate s_n. These are cell specific known covariates and are incorporated into the VAE networks. Assuming B experimental batches, scvi incorporates s_n by one-hot encoding the batch vector and appending the B-dimensional standard unit vector to each $\boldsymbol{z}_n$, where the encoder and decoder networks are defined with input dimensions $(G + B)$ and $(D + B)$, respectively. We denote the conditional networks by $f(\boldsymbol{z}_n|s_n), \mu(\boldsymbol{x}_n|s_n), \sigma^2(\boldsymbol{x}_n|s_n)$ to distinguish between variables and the known covariate s_n. This framework defines a conditional VAE (Sohn et al., 2015) and constitutes the basis of how scvi performs batch correction. Incorporating batch labels in the networks leads to a latent space structure agnostic of batch information (i.e. batch-corrected) (Figure 2.1b), because all batch specific variation is captured by the relevant network parameters. In other words, the conditional VAE encourages independence between $\boldsymbol{z}_n$ and s_n. The batch-corrected latent space is then used for downstream analysis.

Throughout our research we trained the single cell VAE models with Adam (Kingma & Ba, 2015). The gene-specific inverse dispersion parameter $\theta_g \in \mathbb{R}_{>0}$ was trained via gradient descent. All networks used dropout to reduce overfitting

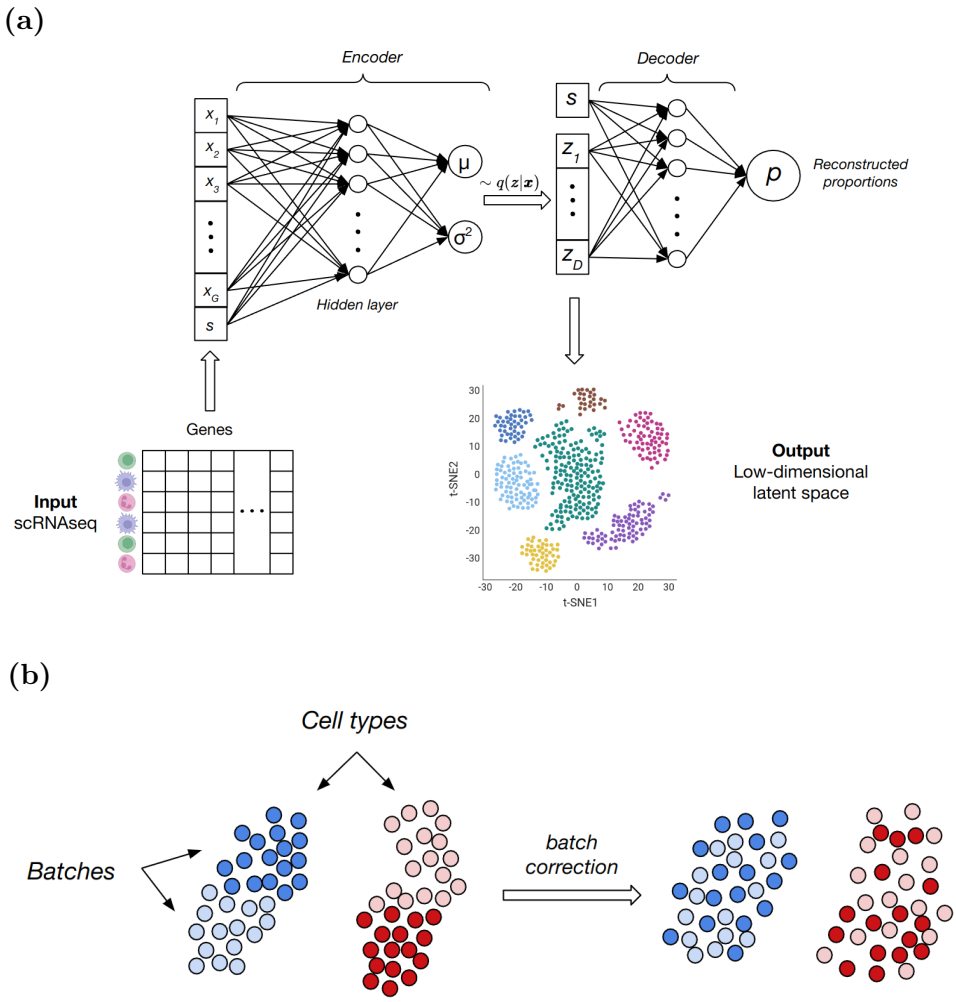


Figure 2.1: Schematic of how VAEs perform dimensional reduction and batch correction. (a) The VAE consists of two fully connected networks (encoder & decoder); the encoder takes as input the single-cell data and computes parameters of the variational posterior; latents are *sampled* and they provide a batch-corrected low-dimensional space. The latents are then mapped back into expression space via the decoder, to obtain normalized proportions. (b) The task of batch correction can be formulated as removing variation related to batch information, while still maintaining biological structure (e.g. cell type clusters). Created with BioRender.com.

(Srivastava et al., 2014)[2], normalization of each input layer to allow for faster convergence and higher learning rates (batch normalization; Ioffe & Szegedy, 2015) and the loss function contained a multiplier of KL divergence that gradually increases

[2]Not to be confused with the term "dropout" we described in single-cell data.

over a pre-defined number of epochs (KL warmup; Bowman et al., 2016). All hidden layers have 128 nodes; for single cell models we generally found that one hidden layer per network provides an expressive representation. Each hidden layer has a ReLU activation function and the decoder output layer has a softmax activation.

VAEs have proven to be great tools in analyzing scRNA-seq, due to their flexibility, scalability, expressivity and accuracy in reconstructing gene expression data. Recent work has shown development of VAE models for integrating multiple scRNA-seq datasets (Lotfollahi et al., 2021; Xu et al., 2021); multi-omic data, for instance single-cell RNA-seq with CITE-seq protein measurements (Gayoso et al., 2021), spatial sequencing methods (Lopez et al., 2019) and single-cell chromatin accessibility data (Ashuach et al., 2022).

Another advantage is their output can be used directly for downstream analysis, without needing to use external methods, thus creating a more robust end-to-end analysis pipeline.

2.3 Downstream Analysis

The VAE formulation allows quantification of uncertainty of latent estimates, via the variational posterior. We use this to form the posterior predictive distribution, which generates replicates $\widetilde{\boldsymbol{x}}$ from the trained model:

$$p_\theta(\widetilde{\boldsymbol{x}}|\boldsymbol{x}, s) = \int p_\theta(\boldsymbol{x}|\boldsymbol{z}, s)p_\theta(\boldsymbol{z}|\boldsymbol{x}, s)d\boldsymbol{z} \tag{2.34}$$

$$\approx \int p_\theta(\boldsymbol{x}|\boldsymbol{z}, s)q_\phi(\boldsymbol{z}|\boldsymbol{x}, s)d\boldsymbol{z} \tag{2.35}$$

where (2.35) comes from estimating the true and intractable posterior $p_\theta(\boldsymbol{x}|\boldsymbol{z}, s)$ with the variational distribution. We note that the marginalization in (2.34) is also intractable. We can characterize the distribution (2.34) for gene proportions (as opposed to gene counts, to eliminate variation from differences in library size) via Monte Carlo estimates (Gayoso et al., 2021), for instance

$$\mathbb{E}[\widetilde{\boldsymbol{x}}|\boldsymbol{x}, s] = \mathbb{E}\Big[\mathbb{E}[\widetilde{\boldsymbol{x}}|\boldsymbol{z}, s] \;\; | \; x, s\Big] \tag{2.36}$$

$$\approx \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x},s)}\left[\mathbb{E}_{p_\theta(\widetilde{\boldsymbol{x}}|\boldsymbol{z},s)}[\widetilde{\boldsymbol{x}}]\right] \tag{2.37}$$

$$= \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x},s)}[\boldsymbol{\rho}_n] \tag{2.38}$$

$$\approx \frac{1}{L} f(\boldsymbol{z}^{(\ell)}|s),\ \ \boldsymbol{z}^{(\ell)} \sim q_\phi(\cdot \mid \boldsymbol{x}, s) \tag{2.39}$$

This approximate distribution can be used for model criticism (how well the model fits the data) via posterior predictive checks (Rubin, 1984), comparing the sampled replicates to the observed data.

The previous section provided details on how the scvi model performs batch correction. The advantage is that by incorporating batch covariates, the resulting latent space is corrected and requires no further adjustments. Well known methods for batch correction used in scRNA-seq analysis include ComBat (Johnson et al., 2006) and mutual nearest neighbors (MNN; Haghverdi et al., 2018). Application of both these methods results in changes to the data in order to obtain a batch-corrected form, which is inconvenient if we are concerned with modeling the data distribution. On the other hand, scvi corrects batch effects via the latent space while leaving the data unchanged; the benchmark examples in (Lopez et al., 2018) show that it aligns batches better than ComBat and MNN.

Another application is differential expression analysis, which consists of identifying genes exhibiting significant expression differences between distinct cell groups. (Boyeau et al., 2023) use the reconstructed gene proportions from scvi and their uncertainty for this task. We consider two non-overlapping groups of cells A and B. For each cell we can sample gene proportions by sampling latents from the variational posterior and mapping them through the decoder. For a gene g and any two cells sampled uniformly at random from A and B each, (Boyeau et al., 2023) formulate the *log-fold change* (LFC) between the two proportions as

$$r^g_{A,B} = \log_2 \frac{\rho^A_g + \epsilon}{\rho^B_g + \epsilon} \tag{2.40}$$

where ϵ is a small offset used for numerical stability.

The log-fold change measures the difference in magnitude between expression in the two groups; high absolute values indicate the gene is differentially expressed

(DE). The offset ϵ plays an important role in filtering out lowly expressed genes. scvi decoders use softmax activation in the output layer in order to produce normalized proportions. If a gene is not expressed or has sparse expression, the softmax can produce arbitrarily small non-zero values (which we always found in practice). Thus, the ratio (2.40) can lead to spurious values. The authors approximate this parameter by querying genes with no expression - in which case the non-zero softmax output for these genes is known to be spurious.

One of the key aspects of the study is that they formulate the LFC as a random variable. Its distribution is estimated by sampling cells from A and B uniformly at random and computing (2.40) for each pairing. They formulate the models for g being up-regulated, down-regulated or not differentially expressed as follows:

$$\mathcal{M}_{+}^{g} : r_{A,B}^{g} > \delta;\ \mathcal{M}_{-}^{g} < -\delta;\ \mathcal{M}_{0}^{g} : |r_{A,B}^{g}| \leq \delta \tag{2.41}$$

Therefore, genes are considered DE if the absolute value of the log-fold change is above some threshold δ. They estimate it by computing the median of the LFC distribution for each gene and fitting a 3-component Gaussian mixture model on the medians; δ is set to the mean of the largest mode in absolute value. The practical task of identifying DE genes via this method will appear several times in this thesis.

While downstream tasks can leverage the sampling properties of the VAE model, an important drawback worth highlighting is that amortised variational methods (and implicitly, VAEs) underestimate the variance of the true posterior and estimates of posterior expectations may be biased (Turner & Sahani, 2011; Lopez et al., 2020). Therefore, care must be taken when trying to interpret the variation in posterior $q_\phi(\cdot|\boldsymbol{x})$ and thereby the confidence intervals resulting from decoding samples into gene expression values. Alternative inference approaches that mitigate this problem include importance-weighted variational inference (Domke & Sheldon, 2018), which optimizes a tighter bound on the marginal log-likelihood via an importance sampling scheme, where the importance weights are:

$$w(x, z) = \frac{p_\theta(x, z)}{q_\phi(z|x)} \tag{2.42}$$

Chapter 3

Cell Type Deconvolution of Clumps

Chapters 1 and 2 described the research question of this thesis and both the experimental and quantitative background for tackling it. As described in Section 1.4, ClumpSeq data is made up of clumps of attached cells; in this thesis we focus on clumps made up of 2 cells with the view that we are interested in how two cell types communicate. The clumps are sequenced via conventional scRNA-seq protocols, hence the gene expression measurements of each clump will be a sum of expression signals from each cell. The computational task is to retrieve these signals in order to characterize the individual cells making up the clump.

Chapter 2 looked at applications of VAEs for modeling single-cell data and highlighted their advantages. This chapter will begin describing the contributions towards addressing the research objectives; here, we will focus on the development of the quantitative deconvolution model for clumps. Our work will build up on the scvi-tools deep learning framework (Gayoso et al., 2022) to show that VAEs can also be applied for this specific type of data.

3.1 Single-Cell Generative Model

In order to characterize each cell in a clump, we must first be able to characterize the cell type population via scRNA-seq data. One of the key aspects of our modeling framework is that we retrieve the cells in clumps by cell type. In every analysis

in this thesis, we assume that each clump dataset is accompanied by a reference single-cell dataset with annotated cell types. We begin with describing how this single-cell data is used for learning cell type representations.

Let C be the number of cell types. For the single-cell VAE model we use scvi with cell type conditional labels:

$$\boldsymbol{z}_n \sim \mathcal{N}(0, I) \tag{3.1}$$

$$x_{ng}|c_n \sim \text{NegativeBinomial}(N_n f(\boldsymbol{z}_n|c_n),\ \theta_g) \tag{3.2}$$

where $c_n \in \{1, \ldots, C\}$ is the cell type of cell n.

The inference model is also similarly defined, with $q_\phi(\boldsymbol{z}|\boldsymbol{x}, c) = \mathcal{N}(\boldsymbol{z};\ \mu(\boldsymbol{x}|c), \text{diag}(\sigma^2(\boldsymbol{x}|c)))$. This model is now conditional on cell type, compared to the previous formulation with batch covariates. Training this will result in a latent space agnostic of cell type. While producing a latent space in which cell types mix is counter-intuitive at first, we provide the reasoning in more detail in the next section and show how it is beneficial for deconvolution.

The objective of the conditional model is to learn the signals that separate the cell types. For now, we emphasize two key aspects of this conditional VAE model. First, the conditional setup forces the decoder to learn the cell-type specific signals. The encoder maps the expression values onto the cell-type agnostic latent space, and the decoder maps the latents back to the corresponding cell type clusters. This latter mapping will reconstruct the cell type specific signals. Suppose we sample $\boldsymbol{z} \sim \mathcal{N}(0, I)$ as per the generative model above or from the variational posterior; reconstructing the expression of an artificial cell of type c is done by plugging the label into the decoder. This way, a fixed $\boldsymbol{z}$ can be mapped to each cell type cluster (different reconstructions) for each label $c = 1, \ldots, C$ plugged into the decoder. We contrast this with a non-conditional model (or where conditional labels are independent of cell type), where the mapping of latents $\boldsymbol{z} \sim q_\phi(\cdot|\boldsymbol{x})$ to each cell type cluster (via the decoder) is based on the *location* in the latent space - i.e. a latent in cluster c_1 will map to reconstructed expression of c_1 cells etc. Second,

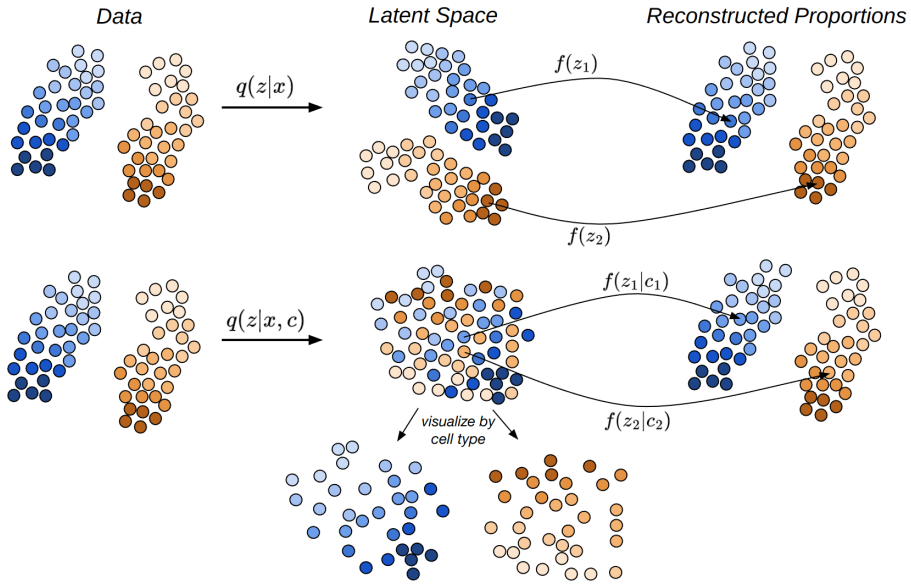


Figure 3.1: Difference in latent space structure resulting from conditioning a VAE model on cell type labels c. Without labels (top), the model constructs a latent space where cell type cluster structure in preserved. On the other hand, the conditional case (bottom) results in latents that mix across cell types but still preserve the cell type specific variation (color gradient).

while cell types are mixing in the latent space, the cell-type specific variation within each population (e.g. subtypes) is preserved.

Figure 3.1 shows a schematic of the expected output when using cell type labels versus when not. We consider two cell type clusters (blue & orange), each comprising specific variation (color gradient). A VAE model with no covariates will output a latent space where the cell type clusters are defined. The reconstructed gene proportion space will also define the two clusters and the decoder takes as input only latents $\boldsymbol{z}$, thus the mapping from a cluster in the latent space to its representation in the expression space is dictated by $\boldsymbol{z}$. On the other hand, a conditional VAE with cell type labels will lead to the two clusters mixing in the latent space - however, the cell type specific variation is preserved. Here, the signals that separate the cell type clusters are learnt by the decoder, thus the mapping into expression space defines the cluster structure again. The mapping is now largely dictated by the conditional label - as shown in the figure, two latents next to each other can map to two different clusters in expression space.

3.2 The Clump Generative Model

We assume we have a clump and a single-cell dataset sequenced via the same technology and pre-processed to share the same set of genes, and the single-cell dataset contains annotated cell types. Our objective is cell type deconvolution; in the context of our work this refers specifically to two tasks: (1) inferring the composition of cell types in clumps - these estimates are a surrogate for the relative proportions of mRNA transcripts of each cell type in a clump; (2) retrieving and constructing gene expression of each cell type found present in a clump for all genes - we argue that the high resolution of ClumpSeq data can provide accurate information to enable retrieving cell type specific signals, in comparison to e.g. bulk RNA-seq data. These reconstructed profiles can be interpreted as artificial cells whose expression, when aggregated as a weighted sum of proportions from (1), approximate the gene expression in clumps. We emphasize the benefits of (2), as reconstructing these whole-transcriptome cell type profiles allows us to compare them to the single cells. ClumpSeq data can provide new information not found in scRNA-seq data, and being able to relate the retrieved profiles with single cells enables us to extend the behaviours seen in clumps to the single cell populations.

We thus quantify the clump generative model for gene counts y_{ng} as follows:

$$\boldsymbol{z}_{n1}, \ldots, \boldsymbol{z}_{nC} \sim \mathcal{N}(0, I) \tag{3.3}$$

$$y_{ng} \sim \text{NegativeBinomial}\left(N_n \sum_{c=1}^{C} \alpha_{nc}\ f(\boldsymbol{z}_{nc}|c), \Theta_g\right) \tag{3.4}$$

$$\sum_{c=1}^{C} \alpha_{nc} = 1,\ \alpha_{nc} \geq 0 \tag{3.5}$$

where N_n is the library size of the clump.

In this model formulation, we assume C latents for each clump n, each latent z_{nc} corresponding to a cell type $c = 1, \ldots, C$. We adopt the mean and inverse dispersion parameterization of the Negative Binomial, as detailed in Section 2.2. Importantly, we use the decoder from the *conditional VAE with cell type labels* (described in the

previous section) in order to construct cell type specific profiles from each latent. The underlying reasoning is that the single-cell conditional VAE optimizes both networks to learn an accurate and expressive representation of the gene expression data, including signatures specific to each cell type. The decoder also produces expression vectors for all genes and incorporating it in the likelihood (3.4) ensures that the clump latents are comparable to the single cell ones; therefore, this setup satisfies condition (2). To facilitate using the optimized decoder network throughout training, we keep the decoder *frozen* - i.e. remove all gradients to avoid training it.

We adopt the mean and inverse dispersion parameterization of the Negative Binomial, detailed in Section 2.2. We take advantage of its interpretability to formulate a sampling model where the mean is a weighted sum of cell type specific profiles. The weights are given by $\boldsymbol{\alpha}_n$, which comprises the proportions of each cell type. Thus, the cell type composition of clumps (condition (1)) is assessed by simply querying $\boldsymbol{\alpha}$. Finally, we assume a gene-specific inverse dispersion parameter Θ_g specific to the clump dataset. Alternative parameterizations of the Negative Binomial distribution are discussed in Section 3.3. Figure 3.2 shows a schematic of how the clump model works.

We highlight several observations about this proposed model. While the ClumpSeq data contains 2 cells (and thus maximum two cell types), this model considers a summation over all cell types. Throughout our research we have found that attempts to "shorten" the summation into two components performed worse than a summation model across all cell types. However, we will also show that despite spanning cell type labels, $\boldsymbol{\alpha}_n$ most often picks up at most two large non-zero values, corresponding to the cell type(s) present in clumps. Additionally, the proposed model (3.4) constructs artificial cells by type - thus if a clump is formed of two attached cells of the same type c', the model output will be only one retrieved profile (with $\alpha_{c'}$=1).

We formulate the inference model as follows:

$$q(\boldsymbol{z}_{n1}, \ldots, \boldsymbol{z}_{nC}|\boldsymbol{y}_n) = \prod_{c=1}^{C} q(\boldsymbol{z}_{nc}|\boldsymbol{y}_n) \tag{3.6}$$

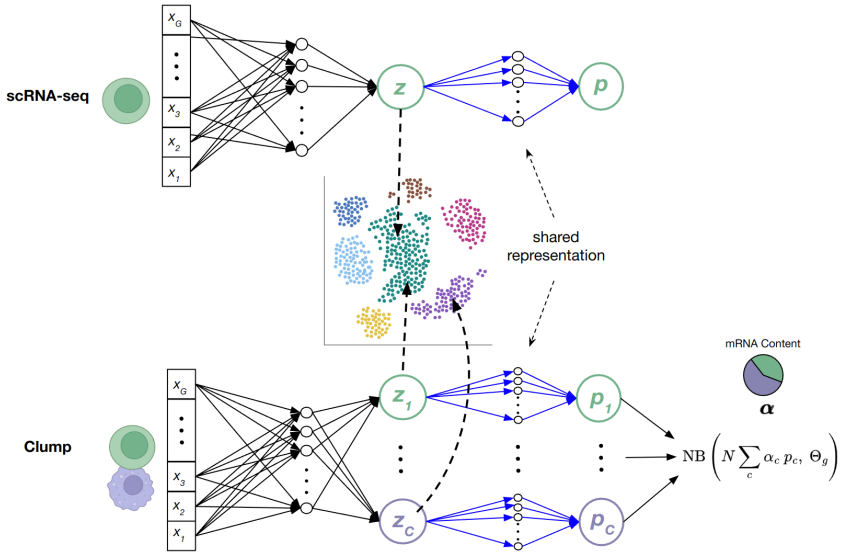


Figure 3.2: The single cell and clump models share the same latent embedding. For each clump, the deconvolution model samples C latents that will each be decoded with a different label. Importantly, using the same decoder as the single cell (without gradients; highlighted by blue) encourages these latents to mix with the single-cell ones. Created with BioRender.com.

$$= \prod_{c=1}^{C} \mathcal{N}(\boldsymbol{z}_{nc};\ \mu(\boldsymbol{y}_n)_{[c]}, \mathrm{diag}(\sigma^2(\boldsymbol{y}_n)_{[c]})) \tag{3.7}$$

$$\boldsymbol{\alpha}_n = f_\alpha(\boldsymbol{y}_n) \tag{3.8}$$

where $\mu, \sigma : \mathbb{R}^G \to \mathbb{R}^{C\times D}$ with $\mu(\boldsymbol{y}_n)_{[c]}, \sigma(\boldsymbol{y}_n)_{[c]} \in \mathbb{R}^D$ denoting cell type specific variational parameters, and the cell type proportions are computed via a fully-connected MLP $f_\alpha : \mathbb{R}^G \to \Delta^{C-1}$. As before, $\Theta_g \in \mathbb{R}_{>0}$ is computed via gradient descent.

The reasoning behind using a conditional VAE on cell types is that the conditional decoder allows identification of each cell type specific summation term. Since the reconstruction of each cell type cluster is mainly done via the label plugged into the decoder (Figure 3.1), each term z_{nc} is assigned a meaning, pertaining to the cell type it constructs. We note how the inference scheme involves sampling C latents for each data point and the generative model (3.4) involves decoding these latents, but each with a different label. Additionally, $\boldsymbol{z}_{nc}$ captures cell type specific variation (e.g. subtypes), given by its location in the latent space.

3.3 Related Work

While deconvolution models published in literature have not been designed specifically for clump data, the deconvolution scheme of ClumpSeq can be seen as a high resolution case for bulk RNA-seq data. However, published models for deconvolution of bulk RNA are usually concerned with cell type composition (condition (1) of our research objective), as samples can be made up of dozens of cells that are not necessarily attached to each other.

Another class of models which employ a similar methodology are the ones designed for spatial sequencing data. Spatial datasets such as Visium consist of samples (spots) made up of multiple cells, potentially from multiple cell types. Similarly, these models are mostly concerned with estimating the cell type composition. The model that is most similar to ours is the recently developed DestVI (Lopez et al., 2022). Their generative model is similar to (3.4), but they use a (r, p) parameterization of the Negative Binomial.

$$x_{ng}|c_n \sim \text{NegativeBinomial}\left(N_n f^g(\boldsymbol{z}_n|c_n)_g,\ p_g\right) \tag{3.9}$$

$$\mathbb{E}[x_{ng}|c_n] = N_n\ f^g(\boldsymbol{z}_n|c_n)_g \frac{1-p_g}{p_g} \tag{3.10}$$

where $p_g \in [0,1]$ and $f^g : \mathbb{R}^D \to \mathbb{R}^G_{>0}$. This decoder no longer computes normalized gene proportions, and the mean proportions in (3.10), multiplied by library size N_n, are a function of both the decoder and parameter p_g, which makes them less interpretable. They formulate a generative model for Visium spots and their reasoning for using the (r, p) parameterization stems from the analytical form of the sum of independent Negative Binomial random variables with identical p: $\sum_c \text{NB}(z_c; r_c, p) = \text{NB}\left(\sum_c z_c;\ \sum_c r_c, p\right)$. Therefore, for a spot s the generative distribution is [1]

$$x_{sg} \sim \text{NegativeBinomial}\left(N_s b_g \sum_{c=1}^{C} \alpha_{sc} f(\boldsymbol{z}_{sc}|c) + \alpha_{s,C+1}\epsilon_g, p_g\right) \tag{3.11}$$

where $\boldsymbol{\alpha}_s \in \mathbb{R}^{C+1}_{\geq 0}$; $b_g > 0$ and $\epsilon_g > 0$ are global parameters quantifying gene specific bias term due to differences in technology and gene specific noise, respectively.

[1] Here the notation is slightly distinct from their paper in order to match the meaning of terms with previous expressions in this chapter.

RCTD (Cable et al., 2021) also performs inference of cell type composition in spatial sequencing data by considering a hierarchical model of counts with a Poisson sampling model. This contains a summation term similar to what we described in the Negative Binomial density (3.4); in their case each term is a constant cell type specific gene expression mean, estimated from a reference scRNA-seq dataset. MuSiC (Wang et al., 2019) also uses scRNA-seq reference datasets to identify cell type compositions in bulk RNA-seq data. In general, many methods for inferring cell type compositon rely on linear models and vector regression, usually preceded by selection of cell type signature genes. More recent approaches have looked at deep learning models, such as scaden (Menden et al., 2020) and TAPE (Chen et al., 2022a).

3.4 Validation on Simulated Data

We validated our proposed model on simulated clump data and compared the results with published models. To simulate clumps, we adopted an approach similar to the one in (Lopez et al., 2022) - instead of sampling cells from the generative model in Section 2.2 and add their counts to obtain clumps, we used a separate sampling model for single cells and clumps and evaluated how well the clump model achieves the 2 research objectives. For the single cell data we chose the PBMC 8k (Zheng et al., 2017), a standard benchmark dataset containing human samples of blood cells, comprising expression counts from 3346 genes.

To this end, we fit PCA on the log-normalized counts [2] of the single cell data. Using the resulting low-dimensional space, we can formulate a sampling model for counts per cell type:

$$\boldsymbol{\omega}_i|c = \exp(\boldsymbol{W}_c\ \boldsymbol{z}_i + \boldsymbol{\mu_c}) - 1,\ \ \boldsymbol{z}_i \sim \mathcal{N}(0, I) \tag{3.12}$$

$$\boldsymbol{\lambda}_i \sim \text{Gamma}(\widehat{\boldsymbol{\theta}},\ \widehat{\boldsymbol{\theta}}/\boldsymbol{\omega}_i) \tag{3.13}$$

$$\boldsymbol{x}_i|c \sim \text{Poisson}(\boldsymbol{\lambda}_i) \tag{3.14}$$

[2] Let $\widehat{N}$ be the mean library size; then we apply PCA to $\log\left(\widehat{N}\frac{x_g}{\sum_g x_g} + 1\right)$

where $\boldsymbol{\mu}_c$ is a cell-type specific (transformed) gene expression mean and $\widehat{\boldsymbol{\theta}}$ is an estimate of inverse dispersion computed from scvi. Steps (3.13) and (3.14) define a Negative Binomial sampling scheme. The procedure is similar for clumps; here we also sample the cell type pairs and relative proportions:

$$c_1 \sim \text{Categorical}(1, \ldots, C) \tag{3.15}$$

$$c_2 \sim \text{Categorical}\left(\{1, \ldots, C\} \setminus \{c_1\}\right) \tag{3.16}$$

$$\alpha_{c_1} \sim \text{Beta}(4, 4) \tag{3.17}$$

$$\boldsymbol{y}_i | c_1, c_2 \sim \text{Poisson}(\alpha_{c_1} \boldsymbol{\lambda}_{i,c_1} + (1 - \alpha_{c_1}) \boldsymbol{\lambda}_{i,c_2})) \tag{3.18}$$

We do not sample doublets here since that would imply two cells of the same type put together in a clump, while the model outputs one latent per cell type. Using the above scheme, we generated 4000 clumps made up of pairings between B cells, CD4 T cells, CD8 T cells, CD14 Monocytes, NK cells and dendritic cells.

We assess the above models by their accuracy in predicting $\boldsymbol{\alpha}$. For each model, we compute the Pearson correlation coefficient between the predicted values versus the ground truth, per cell type; these are shown in Figure 3.3a.

The $\boldsymbol{\alpha}$ values produced by our model match the ground truth ones well in terms of correlation. All models show high values, but all the ones we used for benchmarking showed a larger variance across cell types, usually showing the lowest correlation values for CD8 T cells; meanwhile, our model provided consistently high values. Nevertheless, Pearson correlation does not paint a full picture and thus we also queried the error between estimated $\boldsymbol{\alpha}$ and ground truth values, by cell type, measured by the root mean squared error (RMSE), shown in Figure 3.3b. We also queried the ability of the models to precisely identify the two large α_c values, corresponding to the two cell types in the simulated clump. For each model we sorted the $\boldsymbol{\alpha}$'s and looked at the sum of terms without the 2 largest ones (Figure 3.3c). Since the ground truth is that every clump is made up of 2 distinct cell types, it follows that this quantity should be close to 0; in other words, we expect $\boldsymbol{\alpha}$ to be ideally sparse in all but 2 entries. We find that our model and RCTD provide

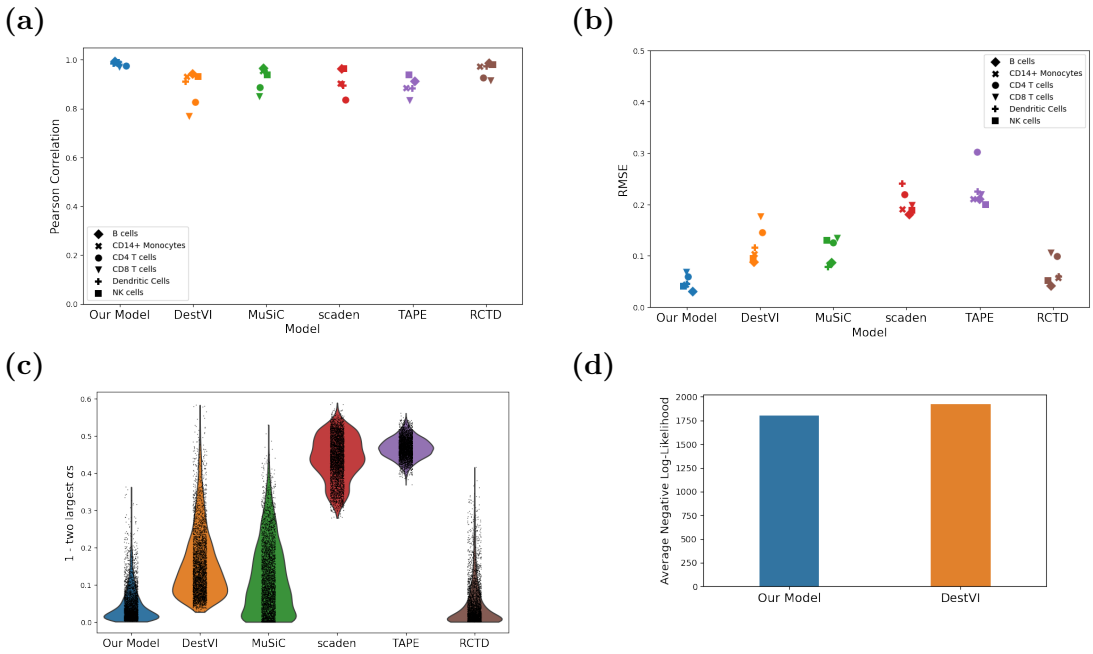


Figure 3.3: Benchmarking different models for cell type deconvolution of simulated clumps. (a) Correlation between predicted and ground truth $\boldsymbol{\alpha}$ per cell type; (b) Root mean squared error between the $\boldsymbol{\alpha}$ estimates and ground truth, per cell type; (c) Sum of terms in each clump specific $\boldsymbol{\alpha}$ without the 2 largest entries; (d) Mean negative log-likelihood for our model and DestVI (lower is better).

similar quality of $\boldsymbol{\alpha}$ reconstruction, based on these metrics. Notably, the entries in these reconstructions are predominantly shifted towards the two cell types.

3.5 Comparing the Negative Binomial Parameterizations

Our model and DestVI have similar formulations, both using a Negative Binomial generative distribution consisting of a sum of cell-type specific gene expression terms (computed via the conditional decoder) and parameters representing the relative contribution of each cell type ($\boldsymbol{\alpha}$'s). The key difference is the Negative Binomial parameterization, which impacts the quality of inference. Our model parameterizes it via the mean and dispersion (μ, θ), while DestVI uses the conventional (r, p) parameters. These can be transformed from one tuple to the other as follows:

$$r,\ p = \theta,\ \frac{\theta}{\theta + \mu} \tag{3.19}$$

Parameterization	(r, p)	(μ, θ)
Formulation for gene count model	Amortized $r(z_i\|c_i)_g$ Gene-specific p_g	Amortized $\mu(z_i\|c_i)_g$ Gene-specific θ_g
$\mathbb{E}[x_{ig}]$	$r_{ig}\frac{1-p_g}{p_g}$	μ_{ig}
Dispersion	r_{ig}	θ_g
$Var(x_{ig})$	$r_{ig}\frac{1-p_g}{p_g^2}$	$\mu_{ig} + \frac{\mu_{ig}^2}{\theta_g}$

Table 3.1: Parameters of the Negative Binomial distribution of gene counts in the 2 parameterizations, for cell i and gene g, showing the mean, dispersion and variance of each as functions of parameters.

$$\mu,\ \theta = \frac{r(1-p)}{p},\ r \tag{3.20}$$

In practice, we found the reconstruction was slightly worse for the (r, p) parameterization; Fig. 3.3d shows the mean negative log-likelihood for both our model and DestVI.

We assessed how the two parameterizations reconstruct the gene expression in clumps. Table 3.1 shows how the parameters are formulated in the single-cell gene count models, in both distributions, including the formulation for mean, dispersion and variance as functions of these parameters. In order to focus on differences due to parameterization, we ran the (r, p) model without the gene-specific bias and other additional parameters implemented in DestVI model (3.11). For our clump model and DestVI, the parameters are defined as follows. In our model, the mean expression of gene g in clump j is a weighted sum of cell-type specific means: $\mu_{jg} = \sum_c \alpha_{jc}\ \mu(z_{jc}|c)_g$, while learning a new dispersion parameter Θ_g. In comparison, in DestVI the r parameter in the density of the spot distribution is $r_{jg} = \sum_c \alpha_{jc}\ r(z_{jc}|c)_g$, i.e. the dispersion is a weighted sum of cell type specific dispersions, while fixing the p_g learned in the single cell model.

We compared the predicted mean and variance of each gene versus the empirical expression. For the data, we normalized the counts of all clumps to the mean library size in order to remove differences due to library size. For the model predictions, we multiplied the predicted gene proportions (the expectations in Table 3.1) by the

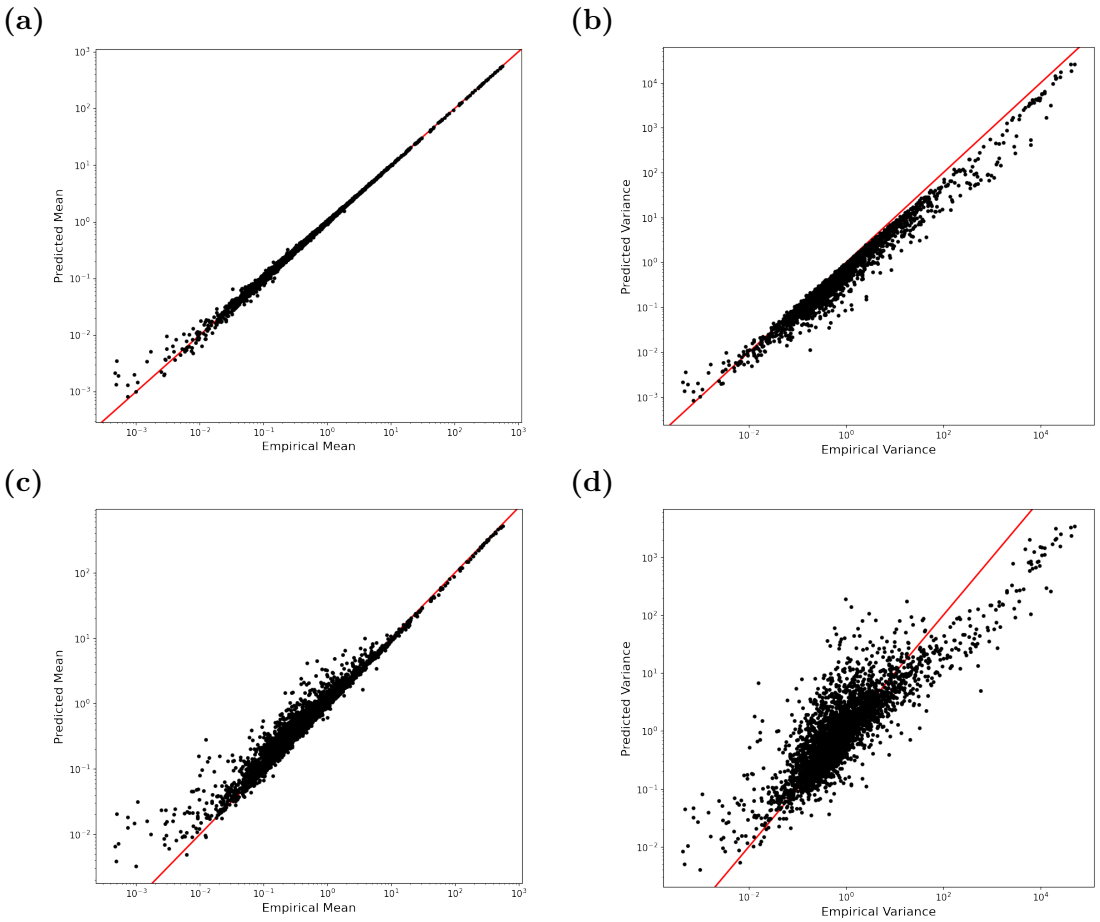


Figure 3.4: Comparison of predicted gene expression mean and variance for simulated clumps versus empirical expression, per gene, for the (a, b) mean-dispersion (RMSE 0.15 for predicted mean, 865 for predicted variance) and (c, d) (r, p) (RMSE 1.6 for predicted mean, 1650 for predicted variance) Negative Binomial parameterizations.

aforementioned mean library size. Fig. 3.4 shows the comparison between the models, showing that the (μ, θ) formulation leads to better reconstructions of gene expression.

3.6 Ablation Study of the Reference Dataset

We assessed the robustness of the deconvolution model with respect to the reference single-cell dataset. To this end, for a given cell type we subsampled the corresponding cell type group to various sizes. We ran the model fitting (both the single cell and clump models) for each value of group size and assessed and quality of deconvolution for the subset of clumps containing that cell type. We ran this analysis for B

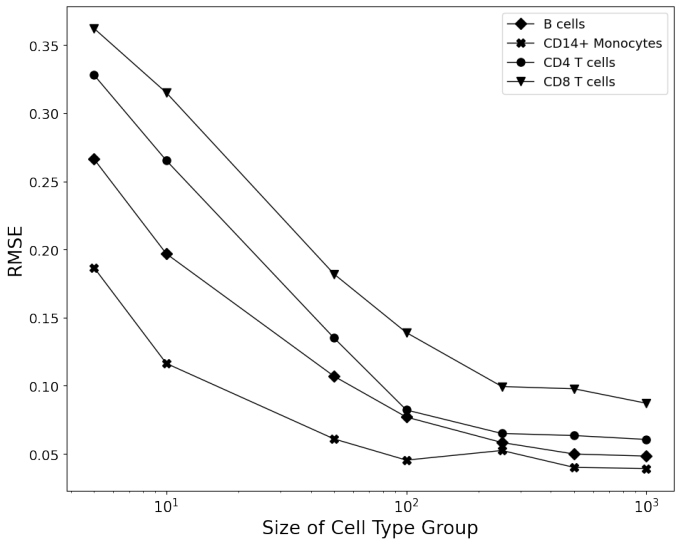


Figure 3.5: For each cell type, the corresponding group in the reference single-cell dataset was subsampled. We assessed quality of deconvolution for clumps containing that cell type, for various sizes of the cell type group ($n = 5, 10, 50, 100, 250, 500, 100$; logarithmic scale) by computing the RMSE between predicted and ground truth cell type proportions.

cells, CD14 Monocytes, CD4 T cells and CD8 T cells as they were the most abundant cell types.

We measured the discrepancy between predicted cell type proportions ($\boldsymbol{\alpha}$'s) and ground truth via the root mean squared error (Fig. 3.5). These results confirm the quality of deconvolution becomes poorer when the reference dataset contains smaller cell type populations, notably if they contain fewer than 100 cells.

3.7 Discussion

In this chapter we developed a model for cell type deconvolution from clump data, able to predict the relative proportions and full gene expression profile of each cell type. We formulated it as a VAE model which maps each clump sample into multiple latents, one for each cell type. The key idea was to plug in the optimized conditional decoder from the single-cell model and not train it, thus ensuring the new latents will map onto the same space and the decoded gene expression will be in the same range as the single cell values. Previous approaches to deconvolution of RNA sequencing data were designed for either bulk RNA-seq or spatial methods.

We generated an artificial clump dataset and compared the performance of our model to other published methods.

It was surprising to see that scaden and TAPE perform poorly at inferring the two cell types; upon inspection we found that the $\boldsymbol{\alpha}$ estimates of these 2 models are much more similar across samples than for the other models. It is worth keeping in mind that these models were developed for bulk RNA-seq data, where samples are composed of many cells and thus sparsity in cell type compositions are less likely. Nevertheless, MuSiC (also developed for bulk RNA-seq) and DestVI show similar levels of performance across the metrics, with MuSiC only marginally more likely to produce "sparser" $\boldsymbol{\alpha}$ values.

It was also surprising to see that DestVI performs poorly at picking out the presence of two cell types, as seen in Figure 3.3c. In this regard, we consistently found that DestVI also performs worse than expected on this dataset - we assumed it would show similar performance to our model since they have similar structure. However, compared to our model, it has more tunable parameters - e.g. the authors provide the option to regularize the gene-specific bias and cell type proportions. Their model also computes maximum a posteriori estimates rather than perform variational inference, in addition to incorporating options to perform either amortization or parameter inference per each data point. For instance, the DestVI version showed here was trained with $\boldsymbol{\alpha}$'s as parameters for each data point; the other configurations performed worse, notably with amortization of $\boldsymbol{\alpha}$ leading to constant output for the entire dataset. In our model, on the other hand, inference of both latent and cell type proportions is amortized, leading to fewer parameters and faster convergence time. We note this disparity in performance is unlikely to be caused by the data simulation process, i.e. the sampling scheme (3.12)-(3.14) does not "favour" a mean based formulation, since a similar procedure was adopted in the original study. We also found that training both the single cell and clump models on the (r, p) Negative Binomial parameterization required many more epochs until convergence, and the reconstruction loss was slightly worse than for the other parameterization (Fig. 3.3d). The negative log-likelihood values are similar despite the disparity in cell

type proportions. This could be attributed to the DestVI likelihood having more flexibility in the sum (3.11) (i.e. b_g, ϵ_g terms); despite this, evaluating the individual summation terms shows that it performs noticeably worse. In our experience, we have not found the (r, p) parameterization to provide a better fit to clump data than what can be obtained via the mean & inverse dispersion formulation (Section 3.5), which is more interpretable.

Given the concentration of α entries on two cell types, seen in our model results (Figure 3.3c), it is expected that the two largest entries per clump correspond to the true pairing. Indeed, we found that our model correctly infers the 2 cell types, with only 35 (out of 4000; $< 1\%$) misclassified pairings. We emphasize the advantage of our model formulation: while RCTD shows similar performance, it only returns cell type composition estimates and does not construct *in silico* cell type profiles, while our model performs both.

One of the drawbacks of this VAE model is the assumption of using the entire frozen decoder network, which could constrain the reconstructed gene expression too much. An instance of this is shown in Chapter 6 where we found a mismatch in expression between the clump and reference datasets. Furthermore, the model assumes we have a fixed number of cell types that are present in both aforementioned datasets. A potential extension could be tackling the situation where a cell type is only contained in clumps, for which there is no single-cell data available. Both these limitations could be approached by relaxing the rigid structure of the decoder, by taking the optimized network, freezing all or part of its parameters and then adding new nodes or layers. This approach has been previously referred to as architectural surgery (scArches; Lotfollahi et al., 2021). The new nodes could allow flexibility in reconstructing expression or introducing parameters corresponding to a new cell type.

Having gained insight into the model performance on simulated data, the next stage is applying it on experimental ClumpSeq data.

Chapter 4

Clump Sequencing from the Intestinal Epithelium

4.1 Background

The gastrointestinal tract is responsible for multiple physiological functions, such as absorption & digestion of nutrients, mucus secretion and protection from external pathogens. The cells inside the gastrointestinal tract (gut lumen) are diverse and involved in a broad range of interactions. The epithelial layer is at the forefront of the defense against pathogens and participates in a complex interaction with the surrounding gut microbiota, which in turn affects the intestinal physiology. Beneath the epithelial layer are tissues and immune cells that form the lamina propria.

The ability of the gastrointestinal tract to perform such a multitude of essential tasks requires a rich composition of cell types. The epithelial tissue is organized in crypt-villus units, where these two regions perform different tasks. The crypt is a cavity-like structure into the lamina propria and its key role is given by the constituent population of stem cells that continuously divide and differentiate into progenitors, which subsequently mature into epithelial cells. These terminally differentiated cells will physically migrate out of the crypt and along the villus, until they reach the villus tip where they undergo apoptosis and are discarded into the lumen. Thus the crypt-villus unit can be seen as an "axis" of migration. This differentiation programme begins in the crypt; once the cells are close to exiting the crypt, the subsequent step is a multi-fate process that leads to formation of distinct

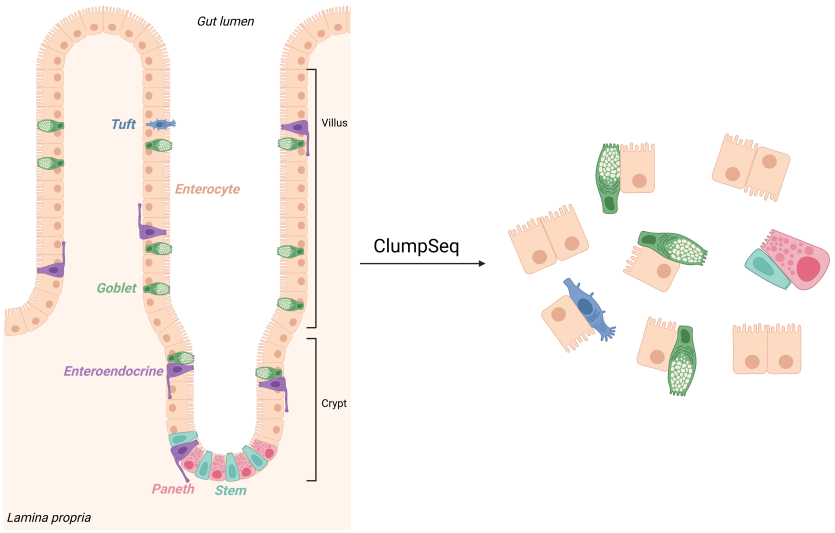


Figure 4.1: Schematic of the crypt-villus unit in the small intestine, with the corresponding cell types. The ClumpSeq experiment of (Manco et al., 2021) consisted of sampling clumps of attached cells. Created with BioRender.com.

cell type lineages. These include absorptive (enteroyctes) and secretory cell types (goblet, paneth, enteroendocrine & tuft cells). Enterocytes have the main function of absorbing nutrients, goblet cells produce mucus that surrounds the epithelial tissue and acts as an additional layer of defense, and enteroendocrine cells are involved in gut hormone production. Paneth cells are an exception to the migration process describe above, as they migrate downward towards the crypt base; their main function is defensive, to protect the stem cell population. (Gehart & Clevers, 2018; Colozza et al., 2022). A schematic of a crypt-villus unit and its constituent cell type populations is shown in Figure 4.1.

Single-cell studies have provided a wealth of novel findings in terms of studying the physiology of the gastrointestinal tissue and identifying cell type and sub-type populations (Grün et al., 2015; Haber et al., 2017; Glass et al., 2017; Gehart et al., 2019). Previous work also looked at spatial organization of cells. For instance, spatial sequencing via Laser Capture Microdissection revealed that enterocytes perform different functions depending on where they are located along the villus (Moor et al., 2018). This spatial heterogeneity of gene expression is called zonation.

In terms of cell type composition of the epithelial tissue, enterocytes are by far the most abundant population. Their abundance is therefore conducive to sampling large and comprehensive sequencing datasets. Nevertheless, secretory cell type populations are sparser and capturing spatial sequencing samples can prove challenging due to the low mRNA abundance. This constitutes the premise of using the ClumpSeq method in (Manco et al., 2021) for identifying zonated genes specific to secretory cell types. The authors sequence clumps from the epithelial tissue in the jejunum (Figure 4.1). Since zonation patterns of enterocytes are well known, the position of a secretory cell can be inferred based on the enterocyte it is attached to.

The purpose of this chapter is twofold. First, we apply and validate the deconvolution model on the clump dataset generated in the original publication (Manco et al., 2021). Second, we argue that ClumpSeq data provides novel information about cell type populations that is not available through single cell measurements. We use the model results for downstream analyses to hypothesize whether we can identify biological signals that are dependent on which cell type a cell is attached to. In addition, clump data can provide information on how different gene programmes are synchronized between cell types. Single cell studies can discover cell type specific gene signatures and migration behaviour along the crypt-villus axis; nevertheless, how the absorptive and secretory populations relate to each other as they migrate remains unclear, as location information is lost. For instance, when a given group of genes becomes upregulated in enterocytes, what happens in the attached goblet cells? Is the gene behaviour between enterocytes and goblet cells mutually independent or can we find correlated genes between the two populations?

This chapter is structured as follows: section 4.2 will look at the analysis of the single-cell data from (Manco et al., 2021) that will be used as reference for deconvolution. Section 4.3 will validate and interpret the clump deconvolution results, showing that our model can predict cell type specific gene expression and infer the composition of clumps. Sections 4.4 and 4.5 will look at downstream analyses for tackling the features described in the previous paragraph.

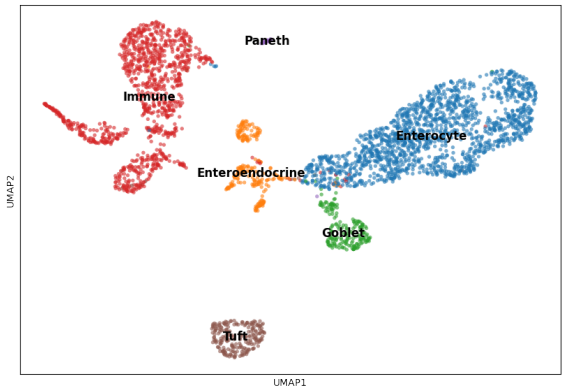
4.2 Single Cell Analysis

The first step of the analysis is to annotate the single-cell reference dataset. For this task I used the set of cell types specific genes from (Manco et al., 2021) for cluster annotation.

A UMAP (McInnes et al., 2018) of all the single cells and annotated cell type clusters is shown in Figure 4.2a. Clusters are made up of enterocytes, secretory cells (goblet cells, tuft cells, enteroendocrine cells & paneth cells) and a population broadly labelled as "Immune", comprising several subclusters with different signatures, as will be detailed below. The identification of all these cell types is straightforward as they all exhibit expression of characteristic marker genes. A handful of these genes are shown in the UMAP in Figure 4.2b, plotting the raw normalized proportions. Expression of *Olfm4* points out to a stem cell population, and the UMAP captures a continuum between this early population and the maturing enterocytes, marked by *Reg3b* and *Apoa1*. The secretory cells are identified via the elevated expression of marker genes specific to each population (Haber et al., 2017): goblet cells (*Muc2*), paneth cells (*Lyz1*), tuft cells (*Dclk1*) and enteroendocrine cells (*Chgb*). Finally, there's quite a large population of cells with expression in genes related to immune functions, which is further divided by expression of different gene groups (Fig. 4.2b). A subset expresses *Gzma* and *Cd8a*, pointing out to a T cell identity. Another subset expresses *H2-Ab1* and other genes related to major histocompatibility complex class II. The third subpopulation expresses genes related to immunoglobulin, for instance through the gene encoding immunoglobulin alpha 1 (*Igha*). The immunoglobulin polymers are linked between each other through a protein called joining chain (*Jchain*) whose corresponding gene shows elevated expression in this subset.

A notable observation in this dataset is that the main variation in the enterocyte population is related to the position along the crypt-villus axis. This can be seen by querying the first principal component of the enterocyte gene expression data and the expression of known zonated genes. For instance, smFISH fluorescence plots show a heterogeneous spatial expression of *Reg1, Slc2a2, Ada* (Figure 4.3a; Moor et al., 2018).

(a)



(b)

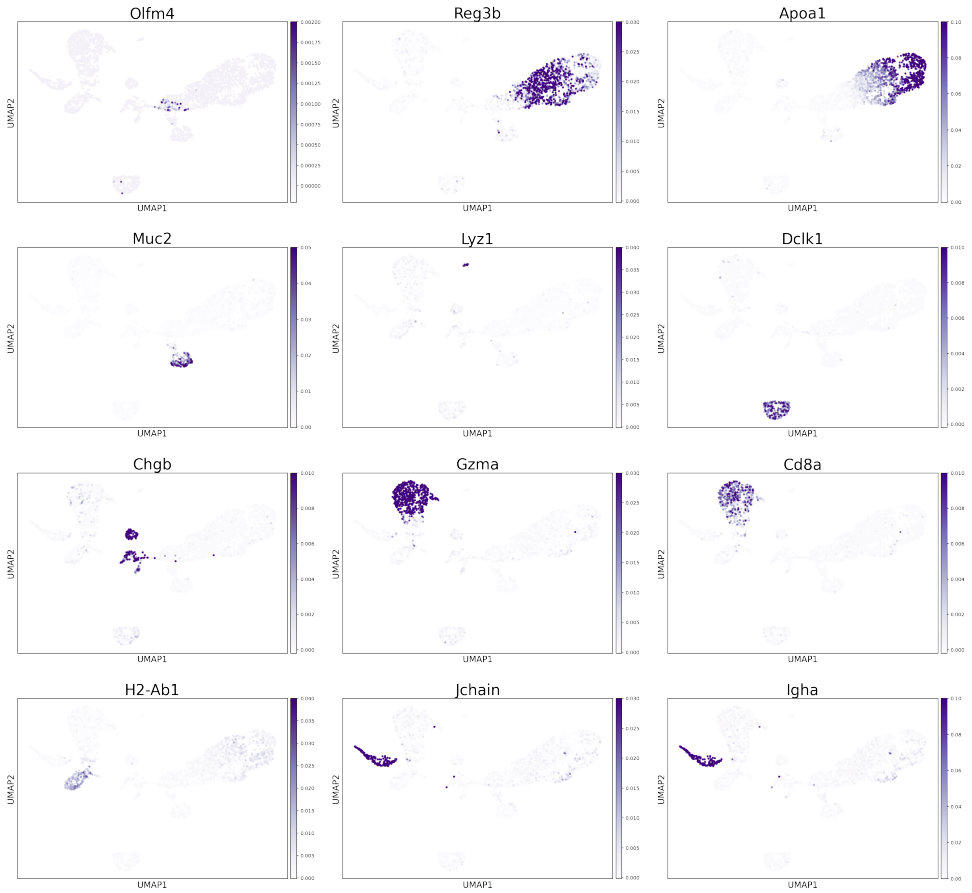


Figure 4.2: Single cell analysis of the intestinal data. UMAP computed on the latents showing (a) annotated cell types; (b) raw normalized gene proportions of key cell type marker genes.

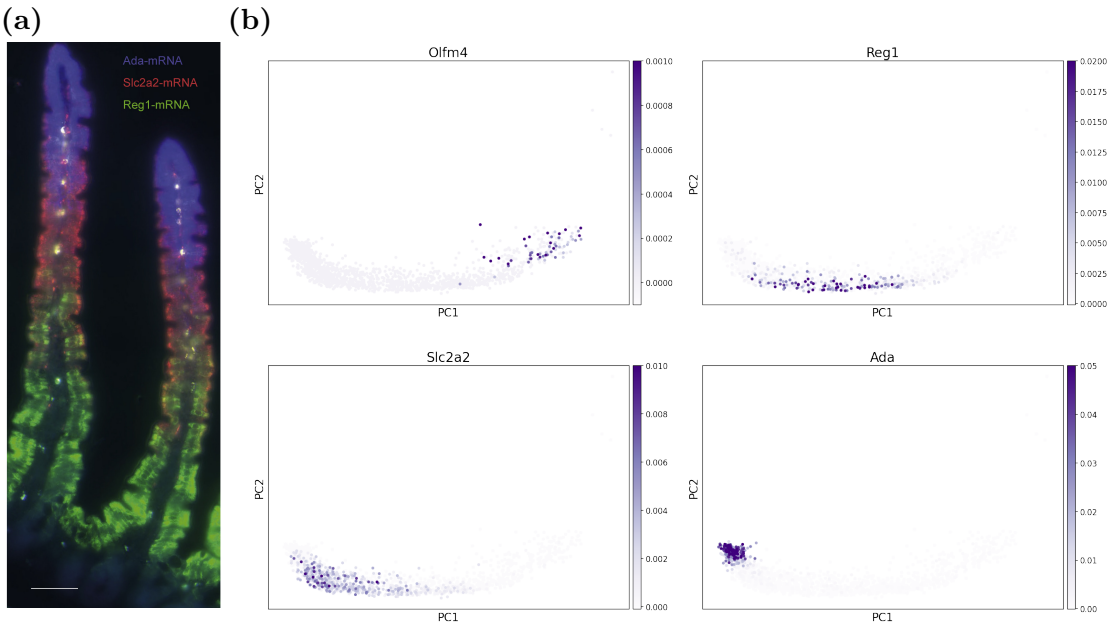


Figure 4.3: (a) smFISH mRNA staining of zonated enterocyte genes *Reg1* (bottom villus), *Slc2a2* (mid-villus), *Ada* (villus tip) from (Moor et al., 2018); (b) Scatter plot of first two principal components of the enterocyte gene expression data colored by expression of stem cell marker *Olfm4* and zonated enterocyte genes from (a). *(a) reprinted by permission from Elsevier (Moor et al., 2018), © 2018, license no. 5705610769563.*

A scatter plot of the first two principal components shows the separation of expression in the aforementioned genes along the first principal component axis (Figure 4.3b).

4.3 Clump Model Evaluation

Quantitative evaluation of the clump model performance boils down to two aspects: (1) the model reconstructs the gene expression in clumps, and (2) each retrieved cell type reconstructs the cell type specific signatures. While (1) is optimized by the loss objective during training, assessing (2) requires more insight.

The advantage of our deconvolution model is that for each clump and each cell type it constructs the full gene profile of an *in silico* single cell, by sampling latents from the variational posterior and querying the frozen decoder. In notation, for $z \sim q(\cdot\,|y)$ each cell type is reconstructed via $f(z|c)$, which returns normalized gene proportions. The conditional decoder assigns each latent to its corresponding cell type (i.e. via the conditional label), thus establishing cell type identifiability among

summation components in the density of the clump distribution. The key aim here is ensuring the decoder reconstructs the cell type specific signatures; for instance, goblet specific genes would only be expressed in the retrieved goblet profile etc.

To evaluate aspect (2), I used a subset of clumps containing an enterocyte and a secretory cell, i.e. multiple cell type signatures. The cell populations studied in this chapter can be identified by expression of corresponding gene markers, as shown in the previous section. These subsets can be selected by querying the inferred cell type composition vector $\boldsymbol{\alpha}$ (see below). To validate whether retrieved gene profiles construct marker expression specific to their corresponding cell type, we compared the predicted expression in clumps with the predicted marker expression values in each individual retrieved cell type. For instance, Figure 4.4a looks at deconvolution of enterocyte-goblet clumps; the scatter plot shows each reconstructed clump as a two-dimensional point (black), where x-axis values are given by the sum of enterocyte markers and the y-axis values are the sum of goblet markers. Since these clumps are all enterocyte-goblet pairs, the points will have non-zero entries in both dimensions. For each clump we plotted two additional points, corresponding to the retrieved enterocyte and goblet cell type profiles (colored), whose dimensions are computed analogously. The retrieved enterocyte profiles (blue) show non-zero expression of enterocyte markers and null values of goblet markers, while the retrieved goblet profiles (orange) exhibit the opposite behaviour. This shows the two retrieved cell type vectors show non-overlapping marker expression and therefore they construct the corresponding cell type specific signatures. This highlights the ability of our model to fit data that expresses both cell type signatures, and retrieve two profiles where each expresses a cell type specific marker set. Figures 4.4b and 4.4c show analogous plots for clumps where enterocytes are paired with tuft and enteroendocrine cells, respectively.

As an exploratory tool, the cell type markers can be used to summarize the composition of each cell type in clumps; this is shown in Figure 4.5a. The majority of clumps show levels of enterocyte signatures; this is expected since enterocytes make up the largest population in the epithelial tissue. Among the secretory cell types, goblet cells show the most frequent marker expression, once again in line with

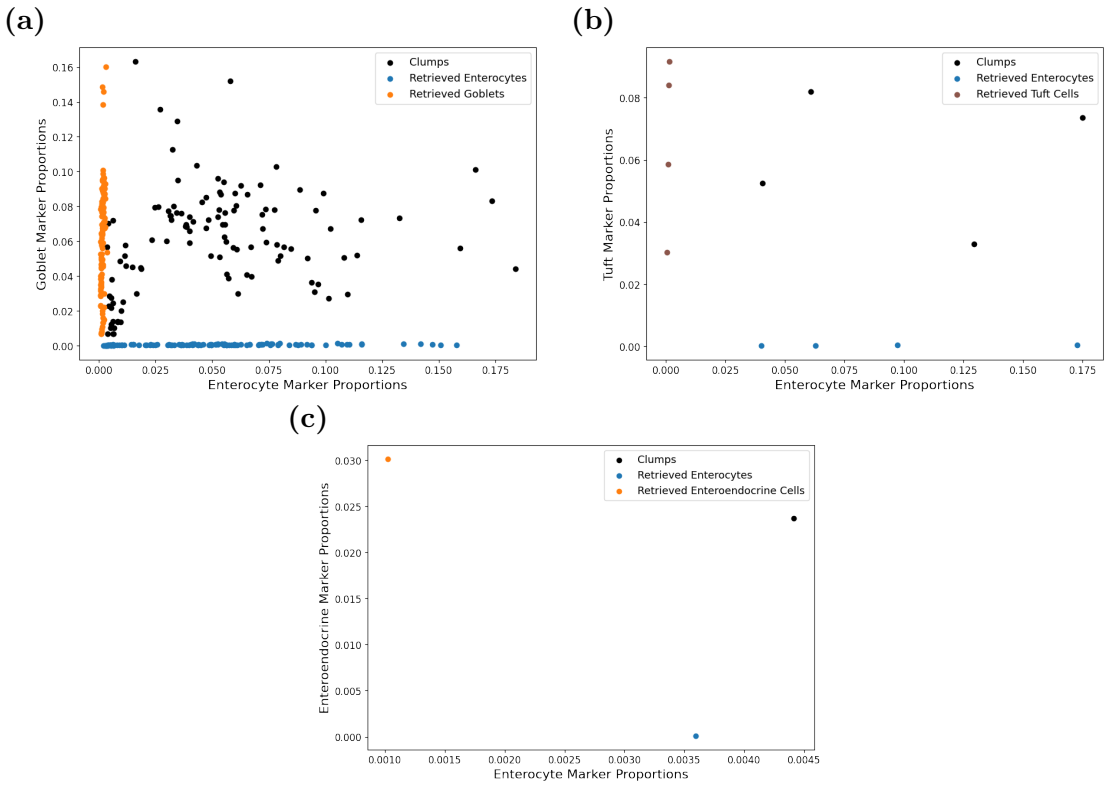


Figure 4.4: Retrieved cell type profiles from clumps capture cell type specific signatures. Scatterplot of cell type specific marker expression queried in the reconstructed clump expression (black) and corresponding retrieved profiles (colored) for pairs of (a) enterocytes & goblet cells; (b) enterocytes & tuft cells; (c) enterocytes & enteroendocrine cells.

the tissue morphology, as they are much more abundant than the other secretory cells. There are clumps containing paneth marker expression, most likely picked up from the crypt. Enteroendocrine and tuft expression is low and most likely only a handful of clumps will contain these cell types. Finally, it is worth noting that clumps show no immune signatures.

In terms of model evaluation, the cell type composition of clumps can be inferred by querying $\boldsymbol{\alpha}$. The amounts of cell type pairs inferred by the model are shown in Figure 4.5b as a dot plot (note the lower triangular matrix form) where dot size is given by the frequency of the corresponding pair. The majority of clumps are enterocyte doublets, followed by enterocyte-goblet clumps and only a handful of other combinations, in line with the empirical observations from Figure 4.5a. Given the abundance of enterocytes in the epithelial tissue, a sampled pair of two attached

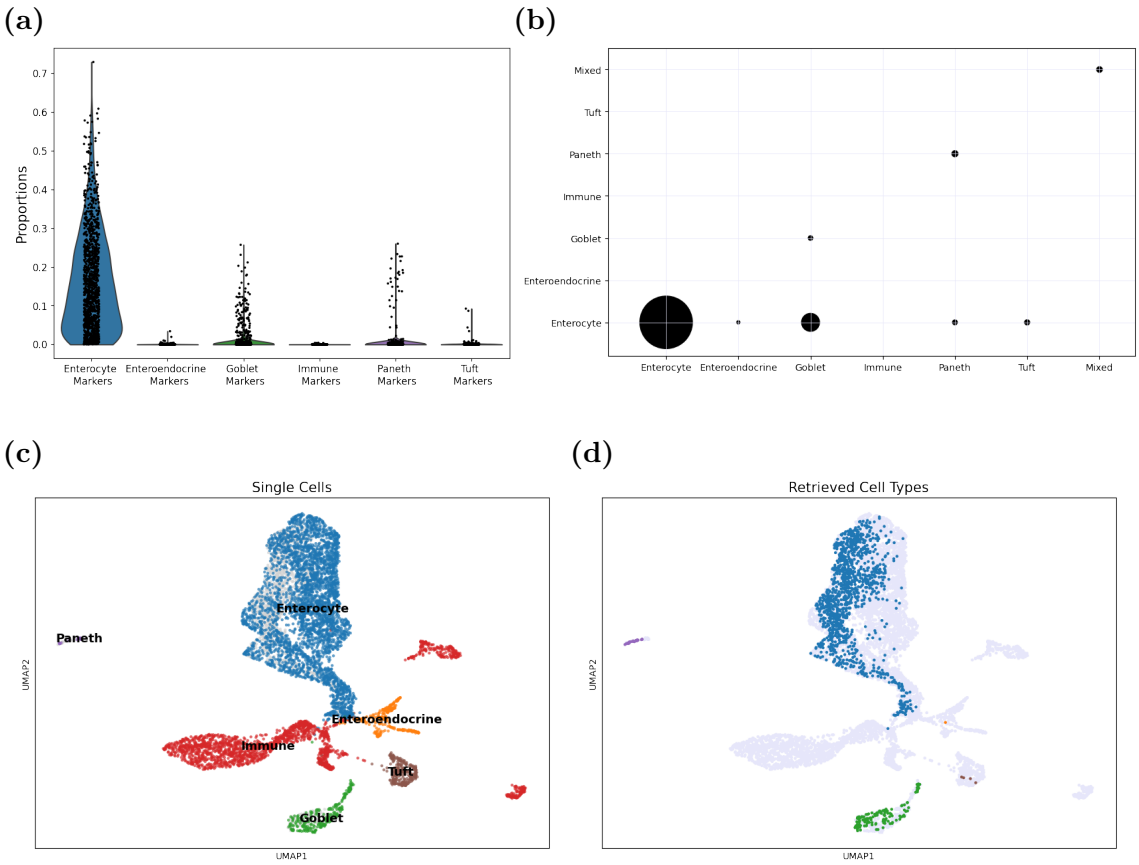


Figure 4.5: (a) Empirical proportion of cell type markers in the clump data; (b) Cell type composition of clumps inferred by the deconvolution model ($\boldsymbol{\alpha}$), where dot size denotes frequency of each pairing (note the lower triangular matrix form). UMAP computed on the joint set of reconstructed proportions of (c) single cells and (d) retrieved profiles from clumps.

cells is very likely to contain two of them.

The retrieved cell type profiles from clumps can be compared to the corresponding single-cell expression vectors, as they are both sampled through the same decoder network. The advantage of this is being able to map the retrieved expression vectors onto the single-cell clusters. Figure 4.5 shows the joint UMAP run in the gene expression space for both sets, highlighting the single-cell clusters (Figure 4.5c) and retrieved cell types (Figure 4.5d). An interesting result is that retrieved enterocytes exhibit a biased mapping onto the enterocyte cluster, while retrieved goblet cells show a much more uniform mapping; this will be explored further in the next section.

This section showed the deconvolution model reconstructs clump gene expression

and retrieves cell type specific expression vectors. One of the main advantages of this model is that the latter vectors predict expression for all genes, for each cell type, and thus opening up applications for downstream analysis, as will be shown in the upcoming sections.

4.4 Downstream Analysis

4.4.1 Selective Mapping of Retrieved Enterocytes

The deconvolution results show an imbalanced mapping of retrieved enterocyte profiles onto the single-cell enterocyte cluster (Figure 4.5d), with most of them mapping onto a subset of this cluster. The entire enterocyte population - including both single cells and retrieved expression profiles from clumps - can be subdivided into multiple subclusters via Louvain clustering (Figure 4.6a). It is straightforward to see that subcluster 1 exhibits the striking imbalance between the two gene profile subsets (Figure 4.6b). We set out to identify potential gene signatures specific to this subcluster. Interestingly, we found it exhibits higher levels of immunoglobulin alpha 1 *Igha* and joining chain coding gene *Jchain* than all the other enterocyte subpopulations (Figures 4.6c and 4.6d). Additionally, subcluster 1 expresses other immunoglobulin related genes, with lower maximal expression, such as *Iglc1, Iglc2, Iglv1* (not shown). As described in Section 4.2, these genes are involved in immunoglobulin secretion and other related immune functions. While this enterocyte subpopulation is found in the single-cell data, the majority of retrieved enterocyte profiles from clumps do not map to this subcluster, implying that these genes constitute signatures not found in the sampled clumps. Indeed, clumps do not show any immune signatures, as shown in the violin plot in Figure 4.5a - we note the set of immune markers in the figure included both *Igha* and *Jchain*.

It was interesting to see that a subset of single-cell enterocytes exhibit such genes related to immune programmes. In fact, the aforementioned immunoglobulin and joining chain genes show the highest expression in 2 out of 4 mice in the single-cell data (Figures 4.6e and 4.6f) - note this does not imply gene expression is exclusive to

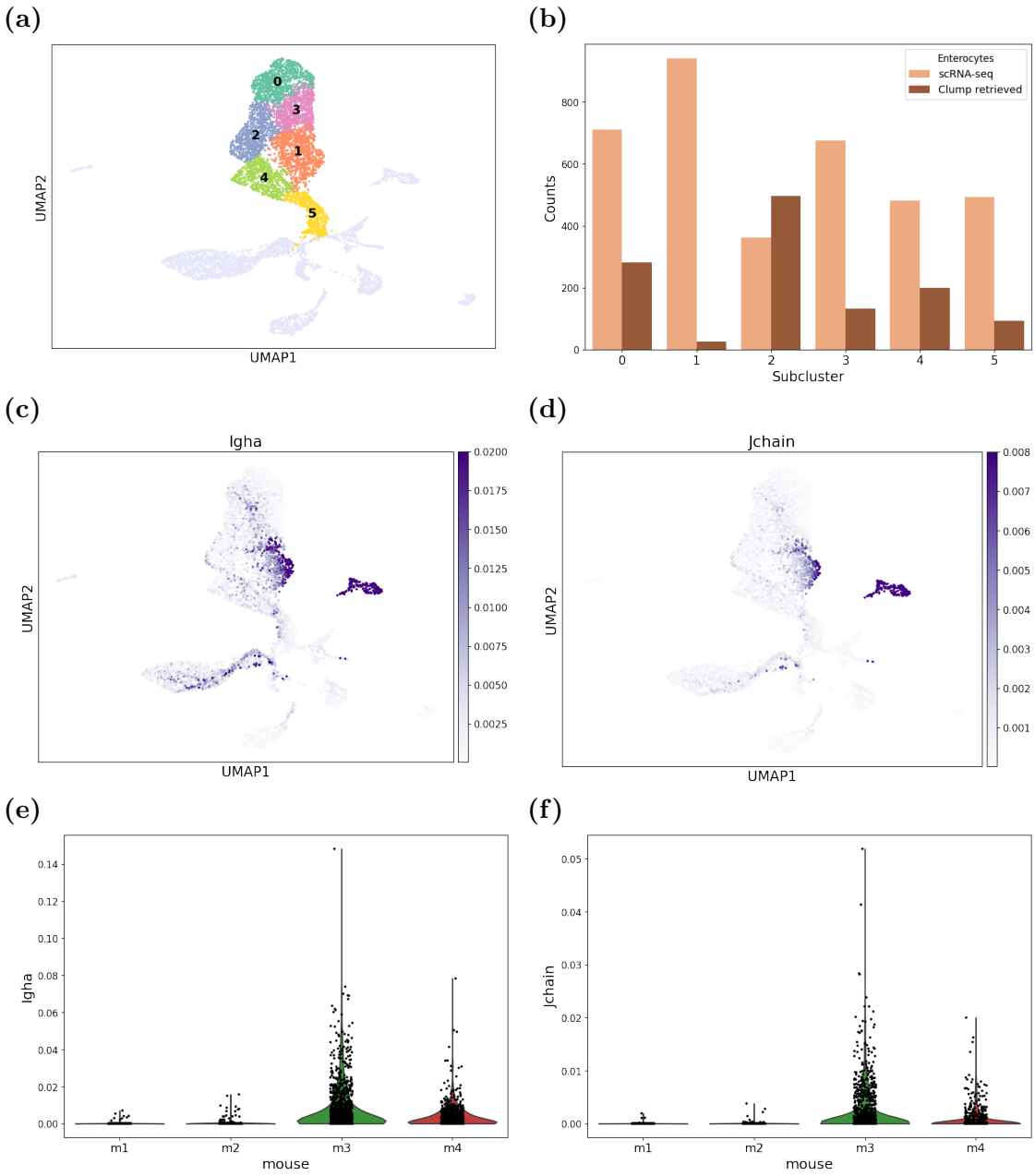


Figure 4.6: Retrieved enterocytes from clumps map to specific areas of the single-cell cluster. (a) UMAP computed on reconstructed proportions of joint set of single cells and retrieved cell type profiles from clumps, highlighting subclusters of the endothelial population; (b) composition of the enterocytes subclusters grouped by modality. Model reconstructed proportions of (c) *Igha* and (d) *Jchain*. Empirical raw proportions of (e) *Igha* and (f) *Jchain* in single-cell enterocytes grouped by mice.

mice 3 and 4, but rather these mice exhibit much higher variation in expression than the other two. This variation in immunoglobulin related gene expression between mice could be the result of technical factors. Nevertheless, it could also reflect mouse specific biological factors, such as amplitude of immune response or interaction with the gut microbiota; for instance, it is known that microbiota and its composition interact with epithelial cells and can induce transcriptional changes in epithelial cells (Peck et al., 2017; Tsang et al., 2022). Expression of immunoglobulin related genes could be a result of an interaction between enterocytes and plasma cells, where the former bind to immunoglobulin receptors (Pabst & Slack, 2020).

We hypothesized why the majority of clumps do not map onto this subcluster. Given the dataset specific variation shown in Figures 4.6f & 4.6e, it is possible the endothelial cells sampled from the mice used in the clump sequencing experiment do not show an enhanced immunoglobulin signature. It is also possible the enterocytes in subcluster 1 are part of a specific subtype, or even a contamination from a nearby tissue, as this subcluster exhibits reduced expression of *Fabp1* and *Reg* genes. For instance, (Haber et al., 2017) describe the proximal (duodenum & jejunum) and distal (ileum) enterocyte populations.

4.4.2 Zonation of Goblet genes

Enterocytes show a striking pattern of zonation, from crypt to villus tip, across many genes (Moor et al., 2018). However, zonation patterns of secretory cells are much less known as they are much rarer than enterocytes and capturing a concentrated population via spatial sequencing can be prohibitive due to their sparsity. The premise of (Manco et al., 2021) is that clumps can be used to infer putative zonation patterns of secretory cells along the crypt-villus axis by querying the zonated genes of the enterocytes they are attached to. This section will therefore look at how the clump deconvolution model can be used for this task.

The single-cell data analyzed in this chapter does not contain information about the location of cells. Therefore, we used a set of known zonated enterocyte genes from previous studies, specifically the enterocyte landmark genes compiled in (Manco

et al., 2021), and we divided these genes into 3 sets, that would approximate 3 discretized zones of the spatial axis: crypt & bottom villus, mid-villus and villus tip. These categories were based on the boundaries used for the LCM experiment in (Moor et al., 2018), shown in Figure 4.7a. Assigning enterocytes to one of the 3 zones was done similarly as in (Manco et al., 2021): for each of the 3 sets of landmark genes, the corresponding normalized gene proportions were summed up per enterocyte and normalized to the unit interval. Each enterocyte had a landmark expression vector in $[0, 1]^3$ space; zones were assigned based on the nearest standard unit vector by Euclidean distance, and are shown in Figure 4.7b. Comparing the estimated zones against the zonated genes outlined in Figure 4.3 confirms the expected continuum of expression (Figure 4.7c).

The deconvolution model allows querying the landmark genes in the retrieved enterocyte expression vector, hence the zone assigment for clump enterocytes was done analogously. Figure 4.7d shows the composition of enterocyte-goblet clumps per zone.

This section will focus on inferring putative zonated genes in goblet cells and therefore will concentrate particularly on the subset of clump data comprising enterocyte-goblet clumps. The clump data provides the ability to learn information about a cell type based on the other cell type it is attached to. For instance, the location of enterocytes can be estimated via the known zonated genes, hence the location of goblet cells retrieved from clumps is given by that of corresponding retrieved enterocytes. Consequently, we can query both gene expression and estimated location of retrieved goblet cells.

Given these resources, we can tackle the problem of identifying goblet zonation patterns across the 3 zones by adopting a differential gene expression framework. For each goblet profile and each zone, we computed denoised samples as in (Gayoso et al., 2021) by sampling latents and subsequently decoding them. These proportions were then plugged into a Gamma distribution with inverse dispersion inferred in the single-cell model:

$$\boldsymbol{z}_{jc} \sim q(\cdot|\boldsymbol{y}_j) \tag{4.1}$$

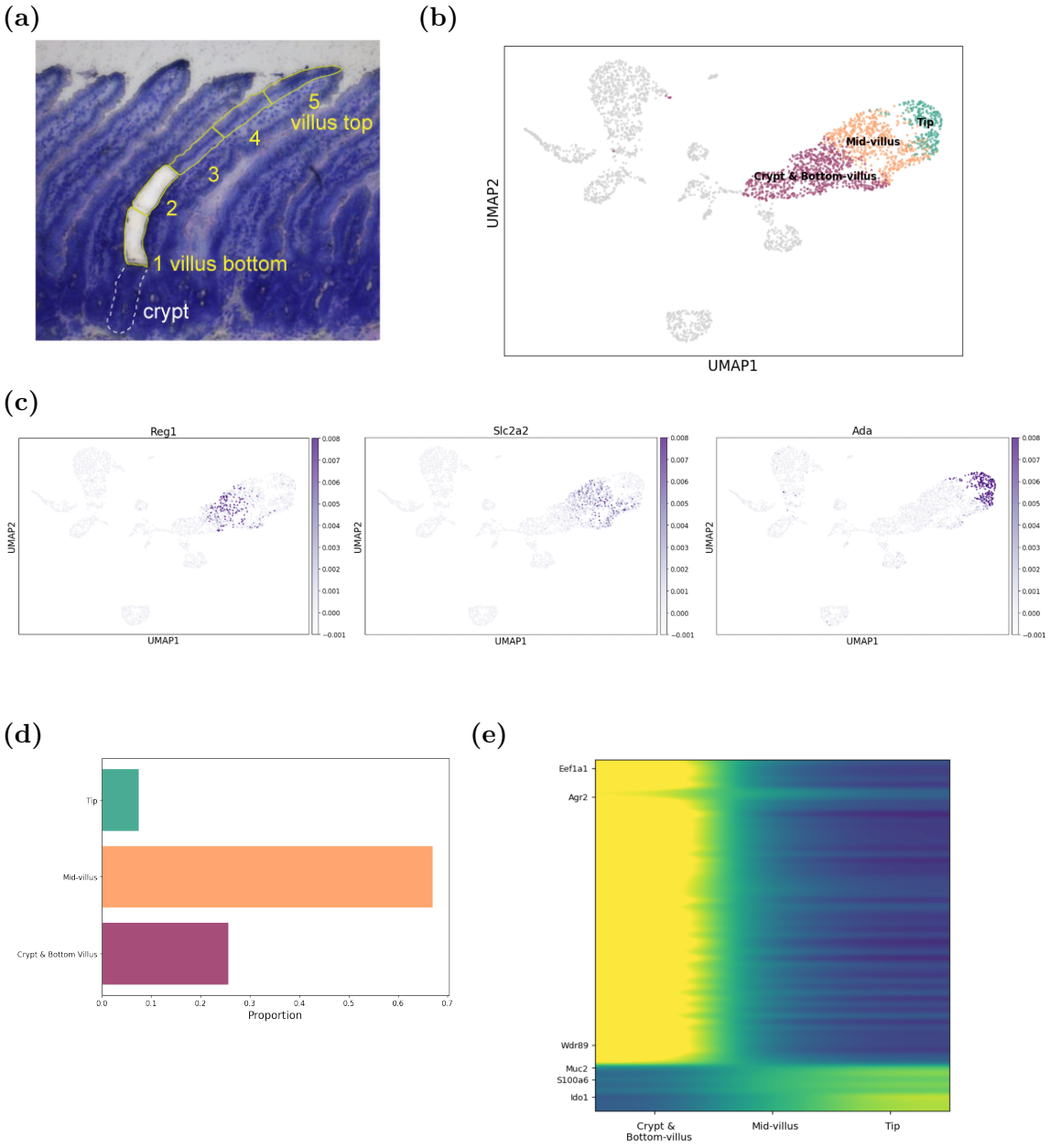


Figure 4.7: (a) Lateral section of mouse jejunum showing the boundaries used in the LCM experiment to cut five equally spaced compartments from the villus bottom to the tip (Moor et al., 2018); (b) estimated zones assigned to single-cell enterocytes based on expression of zonated enterocyte gene expression; (c) raw expression of zonated enterocyte genes from Figure 4.3a; (d) composition of inferred enteroycte-goblet pairs by assigned zone of the retrieved enterocytes; (e) heatmap of constructed proportions of genes found to be zonated in retrieved goblet cells (normalized to unit interval across cells for each gene). *(a) reprinted by permission from Elsevier (Moor et al., 2018), © 2018, license no. 5705610769563.*

$$\boldsymbol{\rho}_{j,c} = f(\boldsymbol{z}_{jc}|c) \tag{4.2}$$

$$\boldsymbol{\mu}_{j,c} = N'\boldsymbol{\rho}_{j,c} \tag{4.3}$$

$$\widehat{x}_{j,c,g} \sim \text{Gamma}\left(\theta_g, \frac{\theta_g}{\mu_{j,c,g}}\right) \tag{4.4}$$

where N' is a fixed scale for all samples, usually the mean library size of the single-cell dataset.

We compare the resulting 3 sets of samples for each gene via a k-sample Anderson-Darling test (Scholz & Stephens, 1987); the null hypothesis is that the 3 samples are generated from the same distribution. The resulting p-values are corrected via FDR (Benjamini & Hochberg, 1995). We restate the differential expression of (Boyeau et al., 2023) from Section 2.3: for any two cell groups A and B the gene specific log fold changes are given by:

$$r^g_{A,B} = \log_2 \frac{\rho^{(A)}_{ig} + \epsilon}{\rho^{(B)}_{jg} + \epsilon} \tag{4.5}$$

for samples indexed by $i \in A$ and $j \in B$, and normalized gene proportions ρ_g. As discussed in Section 2.3, ϵ is an off-set term for lowly expressed genes and in the original publication it is calculated by approximating the mean value produced by the decoder for genes with no expression. In this dataset we found that many genes exhibit sparse counts, with only a handful of non-zero values, thus our analysis required a stricter offset. The mean number of total gene counts per cell is around 1000, thus proportions under 10^{-3} are expected to correspond to counts under 0. We thus heuristically set $\epsilon = 10^{-3}$.

For each zone, we simulated 1000 samples. For each gene, two distributions of log fold changes were computed: crypt & bottom villus versus mid-vilus and mid-villus versus villus tip, and the larger log fold change in absolute value between the two was retained.

Within this framework, a gene is considered zonated if it satisfies two conditions: (1) has a corrected p-value below the significance level and (2) the absolute value of mean log fold change is above a certain threshold: $|r^g_{(A,B)}| \geq \delta$. The latter condition is required in order to eliminate spurious or lowly expressed genes that might be

found significant. In practice we found that most often the tests employed in step (1) lead to a large number of significant p-values, and genes are further filtered in step (2) by inspecting the observed LFC. The ϵ value selected above leads to fewer genes with high LFC in absolute value but whose differential expression is more likely to be biologically meaningful. In this section the threshold δ was computed as described in (Boyeau et al., 2023) - Section 2.3. Expression of genes that met both conditions is shown in Figure 4.7e, grouped by estimated zone. A functional analysis via g:Profiler (Kolberg et al., 2023) showed that this set of genes is mostly enriched for ribosomal functions ribosomal RNA metabolism, specific to crypt cells. This is expected, since most genes in Figure 4.7e show peak expression in the crypt & bottom villus region.

The next step is reconstructing the zonation patterns for single goblet cells - which would require an estimation of their zone - as was done in (Manco et al., 2021); this will be tackled in the next chapter.

4.4.3 Differentially Expressed Genes Within the Enterocyte Population

The clump dataset sequenced from the intestinal epithelium shows a good degree of cell type heterogeneity, with the two most frequent cell type pairs being enterocyte-goblet and homogeneous enterocyte doublets (Figure 4.5b). This poses the general question of whether we can subcluster or partition a cell type population based on what cell they are attached to. For instance, do enterocytes attached to goblet cells express or lack certain gene programmes that are usually found in other enterocytes? For this purpose, this section will focus on identifying differentially expressed genes between enterocytes attached to goblet cells and enterocytes found in homogeneous doublets.

Analogously to the previous section, we can draw *in silico* denoised samples from the two groups. Samples are compared via a two-sample Kolmogorov-Smirnov test, and for each gene we compute the log-fold change distribution (4.5). An important caveat is that enterocytes span the crypt-villus axis and thereby enterocyte specific zonated genes display a broad range of expression. Within the framework of sampling

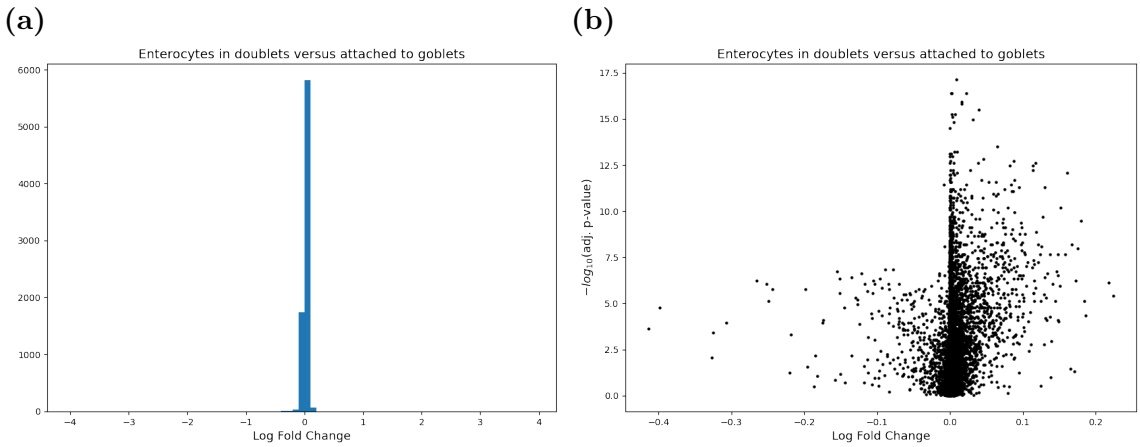


Figure 4.8: Differential gene expression testing between retrieved enterocytes from clump homogeneous doublets and attached to goblet cells. (a) Distribution of mean log fold change, across genes; (b) Volcano plot of adjusted p-values against mean log-fold change. An abundance of genes were found significant in the Kolmogorov-Smirnov test, yet all genes exhibit low absolute values of log-fold change.

gene expression values between the two enterocyte groups, we aim to avoid identifying genes whose differential expression is picked up only due to their zonation properties. To this end, enterocytes are sampled from all zones and we ensure all the cell pairs used in computing the log fold change (4.5) come from the same zone.

Figure 4.8a shows the distribution of mean log fold changes. The methodology of estimating δ outlined in (Boyeau et al., 2023) is based on fitting a 3-component Gaussian mixture model to the distribution of mean log fold changes. This method shows a reduced applicability in this case, as the distribution in Figure 4.8a has heavy tails and no genes show an absolute log fold change value above 0.5. The values in this distribution are predominantly close to 0 despite a large subset genes being found significant in the Kolmogorov-Smirnov test (Figure 4.8b). These findings outline why it is important to consider both the statistical significance and distribution of log fold changes in identifying differentially expressed genes of biological relevance.

On the basis of these findings, we conclude there is no evidence (based on the current clump data) that enterocytes attached to goblet cells display any characteristic genes.

4.5 Co-expression Between Cell Types

The previous sections tackled the hypotheses of whether a cell type population can be divided based on the location of cells along the crypt-villus axis (Section 4.4.2) or which other cell type they are attached to (Section 4.4.3). In other words, clump data can be used to learn variation within a cell type that would otherwise be too laborious or prohibitive to derive from single cell information. Another advantage is being able to query heterogeneous pairs in order to discover groups of genes that are *co-expressed* between two cell types. That is, are there genes g_1 and g_2 specific to cell types 1 and 2, respectively, that show consistent joint expression when the two cell types are attached? This section will focus on enterocyte-goblet clumps, as they are the most abundant heterogeneous pairing.

4.5.1 Background: Identifying Gene Programmes in Single Cells via Hotspot

Quantitatively modeling single-cell data data for the objective of extracting biological signal can be challenging. One of the difficulties stems from the structure of gene expression data; each cell comprises measurements of thousands of genes, many of them with sparse counts due to the limited sensitivity or lowly expressed values with little or no biological relevance. Furthermore, genes are not independent: among the ones encoding relevant biological signals, there are groups made up of strongly correlated genes. One of the challenges is therefore identifying these groups of genes whose variation is biologically driven, referred to as *modules*. This section will focus on the framework of using a quantitative measure of biological similarity between cells in order to infer gene modules. The underlying reasoning is that cells similar to each other will exhibit comparable expression levels of a gene module, whereas cells that are very different are unlikely to share a module. Conversely, a gene can be considered *not* informative under the "null model" that it is randomly expressed across cells, independent of the similarity structure.

An instance of this methodology was developed in Hotspot (DeTomaso & Yosef, 2021). Hotspot uses a weighted graph, computed from the cell-cell similarity

matrix, to identify informative genes that exhibit non-random expression in local neighborhoods. Genes are selected under a testing framework with the null hypothesis that expression is sampled independently from one gene specific distribution. Their test statistic is a measure termed "local autocorrelation", which sums up pairwise gene expression values between a cell and its neighbors in the weighted graph. Specifically, for gene x this is defined as

$$H_x = \sum_i \sum_{j \in \mathcal{G}(i)} w_{ij} x_i x_j \tag{4.6}$$

where i iterates through all the cells, x_i is the expression of gene x in cell i, $\mathcal{G}(i)$ is the set of neighbors of i and w_{ij} denotes the edge weight between i and j.

For the null distribution it is sufficient to compute the first two moments, via either a Bernoulli or Negative Binomial distribution on counts. Counts of x are standardized to null mean and unit variance under the null distributions, followed by standardization of H_x. Thus $H_x \sim \mathcal{N}(0, 1)$ under the null hypothesis.

Hotspot develops this methodology to identify pairs of correlated genes by extending the test statistic (4.6) to capture local correlations between any two genes x and y:

$$H_{xy} = \sum_i \sum_{j \in \mathcal{G}(i)} w_{ij} (x_i y_j + y_i x_j) \tag{4.7}$$

Analogously to H_x, this term is standardized via the moments of the null distribution. Gene pairs for which the null hypothesis is rejected at a fixed significance level are grouped into modules via a hierarchical clustering procedure.

In summary, Hotspot can provide a flexible and useful method for identifying informative genes and grouping them into modules, thus highlighting signatures specific to cell types or sub-types. Its flexiblity comes from being agnostic to the method of computing the cell-cell similarity, as long as the similarity matrix can be compiled into a graph. Therefore, its application to the datasets studied in this chapter can provide biological insights into both single cells and clumps, as will be shown in the upcoming sections.

4.5.2 Application of Hotspot to the Single-Cell Gut Data

The methodology of Hotspot fits well within our objective of developing a robust modeling pipeline, where the results of the generative VAE models can be used in downstream analysis tasks, as was shown in Section 4.4 for identifying differentially expressed genes. The flexible choice of cell-cell similarity metric means the resulting latent space can be used as a measure of proximity between cells. This section will show that Hotspot identifies multiple gene modules within the enterocyte and goblet populations introduced in Section 4.2.

To this end, I ran Hotspot on the two populations individually ($K = 10$ neighbors). The resulting gene modules and local correlation plots are shown in Figure 4.9. It is worth emphasizing that gene modules of a cell type are expressed in different subpopulations. In order to identify these cells, Hotspot computes cell specific module scores. These are shown in Figures 4.9a and 4.9b for the enterocyte and goblet populations, respectively. As it can be seen for both populations, Hotspot modules capture distinct subpopulations, each characterized by module specific gene signatures. We will study the functional enrichment of all modules in the next chapter; the focus of this section is formulating a framework for using the clump data to relate these modules between the cell types.

4.5.3 Gene Co-expression in Clumps

The clump inference model provides retrieved cell type specific gene profiles; each of these *in silico* cells that share a clump can be "linked" as being attached to each other, a unique feature provided by the clump data. Previous work in literature such as Hotspot tackled the problem of identifying groups of genes with correlated expression within a cell type population. This section will use the additional information provided by clumps in order to identify groups of genes that are co-expressed between cell types, thus allowing us to link gene programmes between them. However, we are interested in how these modules are expressed morphologically in the tissue and how they are synchronized between the two populations. For example, are enterocytes

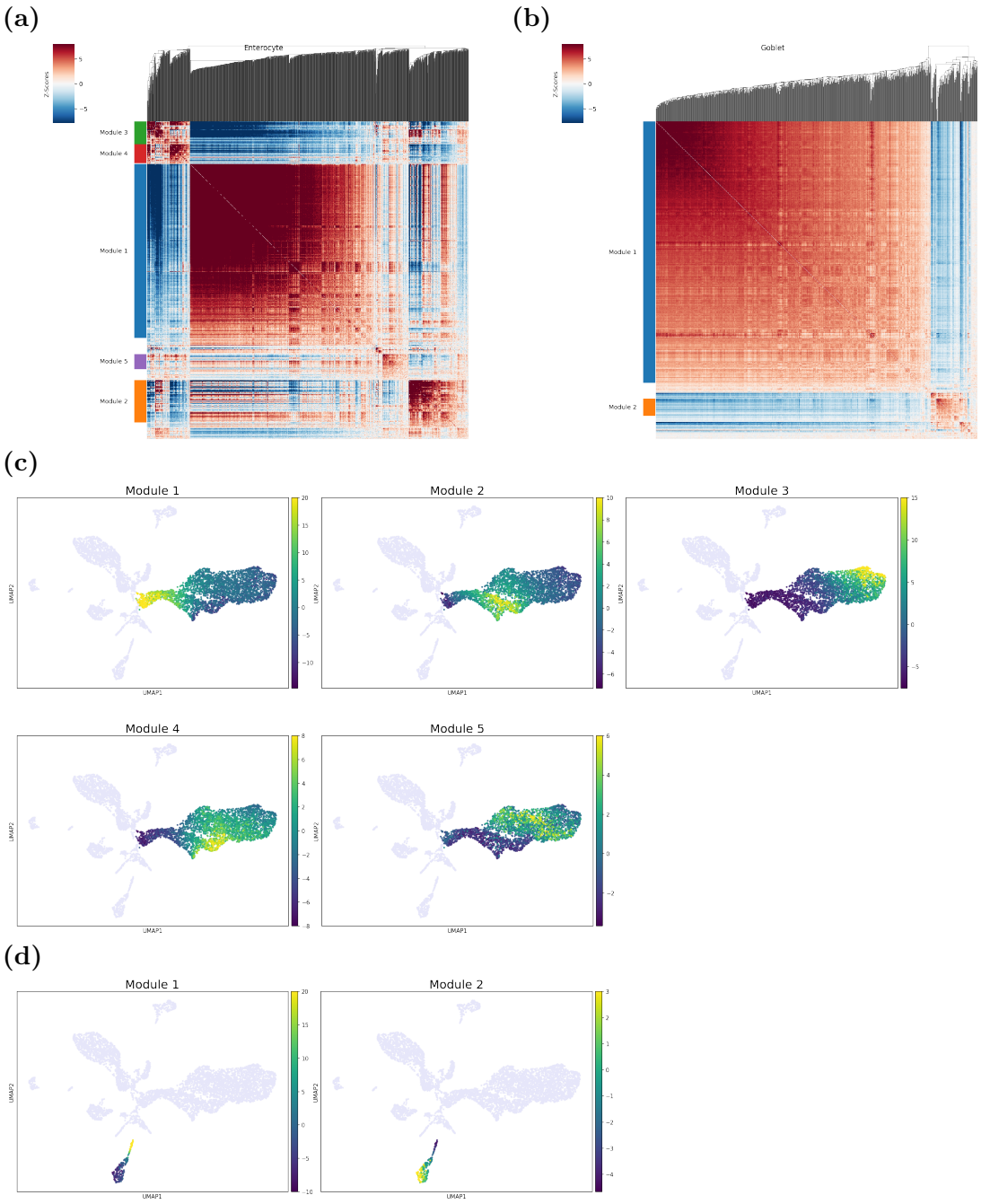


Figure 4.9: Hotspot run on the single-cell enterocyte and goblet populations. Pairwise local correlations in (a) enterocytes and (b) goblet cells. UMAP of single-cell data highlighting module scores for (c) enterocytes and (d) goblet cells.

expressing module 1 attached to goblet cells of module 1 or are these goblet cells attached to enterocytes further along the spatial axis, which express other modules?

The approach of this section consists of adopting a hypothesis testing framework

where we iterate through pairs of informative enterocyte and goblet genes, and assign a pair as correlated based on a null model. In a similar fashion to Hotspot, we summarize the null distribution via its first two moments, as detailed below. We will first outline several important aspects.

First, while computing a cell-cell similarity measure for the single-cell data allows the resulting KNN graph to be used for computing autocorrelation (4.6), there is no such similar measure for clumps. Since they are made up of two cell types, with varying degree of cell type composition, it is difficult to describe a measure of similarity between two clumps.

Second, the null model of Hotspot fits a Bernoulli or Negative Binomial distribution on gene counts, and their method of computing moments takes into account differences in library size. The retrieved cell type profiles given by the clump model are normalized gene proportions, therefore we require an appropriate null model. Additionally, since we work with retrieved cell type profiles for which no individual cell type counts are available, the gene proportions are sampled from the decoder.

Third, lowly expressed genes and their associated proportions can lead to spurious inferences in terms of significance. This problem was mitigated in previous instances of differential expression analysis by introducing the off-set term ϵ that would filter out lowly expressed genes by rescaling both the numerator and denominator in the LFC expression. In this case we do not have a formulation that allows using such term, therefore our approach consisted of pre-selecting genes based on expression heuristics. Specifically for this analysis, I retained genes whose mean proportion was larger than $5 \cdot 10^{-4}$.

With these points in mind, we considered the following correlation test. Let G_1 and G_2 be two gene sets that will be queried in enterocyte and goblet populations, respectively. We are testing whether $x \in G_1$ and $y \in G_2$ show a strong co-expression in clumps, i.e. if x exhibits high expression in enterocytes then so does y in goblet cells. The null hypothesis is that the paired expression of x, y is drawn uniformly at random from a cell type specific distribution. We have the following test statistic:

$$H_{xy} = \sum_j p_{j,1,x}\ p_{j,2,y} \tag{4.8}$$

where j indexes the clumps and $p_{j,c,x}$ is the normalized proportion of gene x in the retrieved cell type profile c.

Computing the test statistic under the null hypothesis requires a null model. This is given by the distribution of single cell gene proportions in each cell type population. To see this, we think of the null model as describing the retrieved cell type gene expression in a clump as a random sample from the single-cell clusters; in other words, expression in the two cell types is independent of whether they are attached to each other. Therefore, the null model requires fitting a distribution on the single-cell gene proportions, which are given by the decoded latents. Given the nonlinearity of the decoder, assuming a Normal distribution for the gene proportions seems inappropriate. In order to sample gene proportions, we make use of the single-cell inference model. We take advantage of the following aspects: (1) sampling latents from the variational posterior and mapping them through the decoder is straightforward and (2) we only need to summarize the null distribution by the first two moments. While the aggregate variational posterior is a mixture of Normal distributions, its samples are mapped through the nonlinear decoder and the resulting distribution cannot be parameterized. However, the moments can be computed via a straightforward Monte Carlo estimate:

$$\ell \sim \text{Uniform}(1, \ldots, N_c) \tag{4.9}$$

$$z_\ell \sim q(\cdot\ |x_\ell, c) \tag{4.10}$$

$$\mathbb{E}[p_{g,c}] = \sum_{\ell=1}^{L} \frac{1}{L} f(z_\ell|c)_g \tag{4.11}$$

$$\mathbb{E}[p_{g,c}^2] = \sum_{\ell=1}^{L} \frac{1}{L} f(z_\ell|c)_g^2 \tag{4.12}$$

where $p_{g,c}$ is the random variable quantifying the expression of g in cell type c, and N_c is the number of single cells of type c.

We use the estimated moments to standardize the gene proportions in (4.8) to zero mean and unit variance:

$$\widehat{H}_{xy} = \sum_j \widehat{p}^{(1)}_{j,1,x} \widehat{p}^{(2)}_{j,2,y} \tag{4.13}$$

Analogously to Hotspot, the null model is conditioned on one of the cell type proportions being fixed. For instance, suppose proportions from the first population are fixed; the mean and variance of the test statistic under the null hypothesis are thus given by:

$$\begin{aligned}
\mathbb{E}\left[\widehat{H}_{xy}\right] &= \sum_{j=1}^{N} \mathbb{E}\left[\widehat{p}_{x,1}\right] \mathbb{E}\left[\widehat{p}_{y,2}\right] && (4.14)\\
&= \sum_{j=1}^{N} \widehat{p}_{x,1}\ \mathbb{E}\left[\widehat{p}_{y,2}\right] && (4.15)\\
&= 0 && (4.16)\\
Var\left(\widehat{H}_{xy}\right) &= \sum_{j=1}^{N} Var\left(\widehat{p}_{x,1}\ \widehat{p}_{y,2}\right) && (4.17)\\
&= \sum_{j=1}^{N} \widehat{p}^{\,2}_{x,1}\ Var\left(\widehat{p}_{y,2}\right) && (4.18)\\
&= \sum_{j=1}^{N} \widehat{p}^{\,2}_{x,1} && (4.19)
\end{aligned}$$

The test statistic (4.13) is standardized using the mean and variance above. It is computed for both situations where either cell type is fixed and the test statistic with lower absolute value is retained.

The application of this correlation test to the retrieved profiles from the inferred enterocyte-goblet pairs is shown in Figure 4.10, displaying the pairwise (standardized) co-expression test statistic between informative genes in enterocytes and goblet cells, respectively. A subset of these genes were assigned to cell type specific modules compiled by Hotspot (Figures 4.9a and 4.9b), and these modules are shown in the figure for both populations across the two axes. A striking pattern is that retrieved profiles from modules 1 of both cell types show a strong degree of co-expression. A functional enrichment query on g:Profiler (Kolberg et al., 2023) revealed that

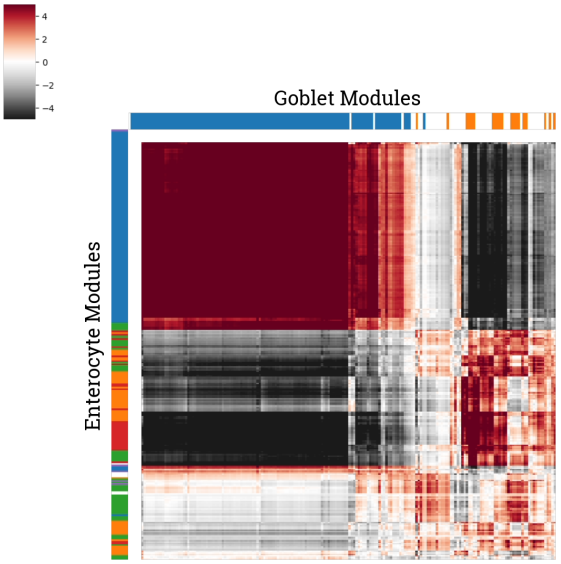


Figure 4.10: Correlation test for gene co-expression in retrieved enterocytes and goblet cells from the correspoding clump pairs, showing the pairwise gene z-scores.

for both cell types, module 1 genes are enriched for ribosome, ribosomal RNA components, ubiquitin ligase inhibitor activity, ribonuclear protein complex; in both cases these genes are expressed in the crypt and bottom villus. Goblet cells with expression of module 2 genes tend to correlate strongly with enterocytes with gene signatures from corresponding modules 2, 3 & 4, which are expressed mainly in the middle and tip of the villus.

Apart from the strong co-expression between crypt modules, the relationship between goblet module 2 and enterocyte modules 2, 3, 4 is less clear - for instance, we are interested in whether these enterocyte modules show a sequential upregulation as the cells migrate and what the expression of attached goblet cells is. We note that the cell representation in this chapter did not account for migration dynamics, which constitute a vital component in describing this dataset. The next chapter will explore this aspect in more detail.

4.6 Discussion

This chapter looked at the application of the deconvolution model to a clump dataset sequenced from the intestinal epithelium (Manco et al., 2021). Using a

single-cell reference dataset from the same source, we annotated the cell types and trained a conditional VAE model. The resulting embedding was used in training the deconvolution model. We found that it was able to retrieve cell type specific gene expression vectors and infer the composition of clumps by cell type.

We highlight several advantages of our modeling framework. First, the deep generative model is flexible in capturing complex non-linearities in gene expression data and encoding them into the latent space via the neural network components. The conditional decoder network is optimized on the single-cell embedding and using it for cell type deconvolution (via the frozen gradient structure) ensures we can construct gene expression profiles comparable the ones of single cells. This brings about the second advantage, namely that we can map the retrieved *in silico* profiles onto the single-cell clusters. Third, the deconvolution inference framework does not require *a priori* knowledge of cell type markers. While the analysis in this chapter made use of known markers for cluster annotation and exploratory plots, we emphasize the clump model makes use of the entire gene profile. The retrieved *in silico* expression vectors capture signatures specific to the cell type they are assigned to without any cell type specific input besides the conditional label for the decoder. For instance, we found that enterocytes in clumps span most of the crypt-villus axis; the mapping of these retrieved enterocytes onto the single-cell cluster was done without any prior input of markers or zonated genes. This aspect highlights an important distinction from the computational methods in (Manco et al., 2021), which rely on identifying markers and landmark genes by directly querying the raw gene expression in clumps.

Clumps provide important information about the original location of a cell population relative to another cell type, which is lost in single-cell experiments due to dissociation. The ability of our model to construct cell type specific profiles allows us to perform downstream tasks in order to identify variation specific to a population based on this new information (Sections 4.4.2 and 4.4.3). We were also able to query multiple cell types and use the retrieved expression of both populations to identify patterns of co-expression (Section 4.5). The advantage of mapping the *in silico* profiles onto the single-cell clusters led to interesting findings that characterize specific

cell type subsets present or absent in clumps - namely a subset of enterocytes with low abundance (Section 4.4.1). This contradicted our prior expectation that enterocytes in clumps would uniformly span the entire single-cell cluster, which was the case e.g. with goblet cells. Whether the low amount of this enterocyte subpopulation in clumps is due to a physiological or technical effect is an open question. Overall, this finding highlights the sensitivity of our deconvolution model to retrieve specific subtypes.

Chapter 5

Ordering Cells in Pseudotime

5.1 Background: Pseudotime

Single-cell data provides whole-transcriptome measurements of gene expression in individual cells, offering the ability to isolate cell populations of multiple types or sub-types. Due to this high resolution we are able to identify gene signatures characteristic to each cell type, which are used to infer important biological functions. While characterizing the gene programmes of various populations is a stepping stone in understanding their biological identity, it is often the case that cell type or sub-type identities are not static. Most biological processes are dynamic, involving many gene regulatory changes and interactions that result in alterations in cells' functions and identity. For instance, differentiation processes involve transitions through multiple cell type stages, which might involve more complex components such as fate choices. The transcriptional identities of the resulting terminal cell types could be vastly different than the one of the original precursors. In these instances we do not characterize the process through one cell type or population, but rather through a *trajectory* comprising sequential states.

Reconstructing the trajectory of transcriptional transitions from single-cell measurements is not straightforward. The biological process under study is often asynchronous and the scRNA-seq samples capture a "snapshot" of a dynamic activity, comprising cells from multiple states pooled together into one dataset, or across multiple experimental batches. The quantitative objective here is to order the cells such that their gene expression reflects the physiological transcriptional changes. This

ordering would be quantified in a variable called *pseudotime* (Trapnell et al., 2014). Conventionally it is defined as a positive real number, with 0 indicating the time of the intial precursor state or "root". Pseudotime is constructed from gene expression data and it estimates developmental time defined by gene expression progression, though it is not necessarily equal to the chronological time of development. A schematic of these points is shown in Figure 5.1a. We also note that pseudotime is also not necessarily related to sampling timepoints; hence estimating gene expression development through a time series experiment (e.g. studying effects of a toxin injected in mouse and collecting cells at various timepoints after application) can lead to inaccurate estimate due to the asynchrony of samples (Figure 5.1b). Therefore, more elaborate computational methods need to be developed.

Categorizing intermediate cell states into discrete groups can prove challenging and phenotypically characterizing them can be prohibitive. Furthermore, changes in gene expression space are smooth and thousands of genes could be going through various changes, thus it is difficult to define boundaries to separate these groups, that would apply to all genes. Pseudotime provides a continuous quantification of this process.

Various computational methods were developed for pseudotime inference, with different approaches. For example, Monocle (Trapnell et al., 2014), Slingshot (Street et al., 2018), scFates (Faure et al., 2022) rely on graph based methods, by fitting a minimum spanning tree in a low dimensional space on cells. Diffusion pseudotime (DPT; Haghverdi et al., 2016) computes a weighted KNN graph on cells and computes pseudotime as distances in random walks of arbitrary length on this graph. Parametric approaches include Gaussian Processes (GPFates; Lönnberg et al., 2017) and Bayesian models of gene expression behaviour (ouija; Campbell & Yau, 2018). Generally, the framework of pseudotime inference includes two steps: (a) dimensionality reduction of expression data and (b) fitting a model on this low dimensional space that reconstructs a trajectory. Most often these two steps are performed separately.

The previous chapter studied the intestinal epithelial tissue, which involved migration of absorptive and secretory cells along the crypt-villus axis. This dynamic process is characterized by changes in gene expression depending on the spatial

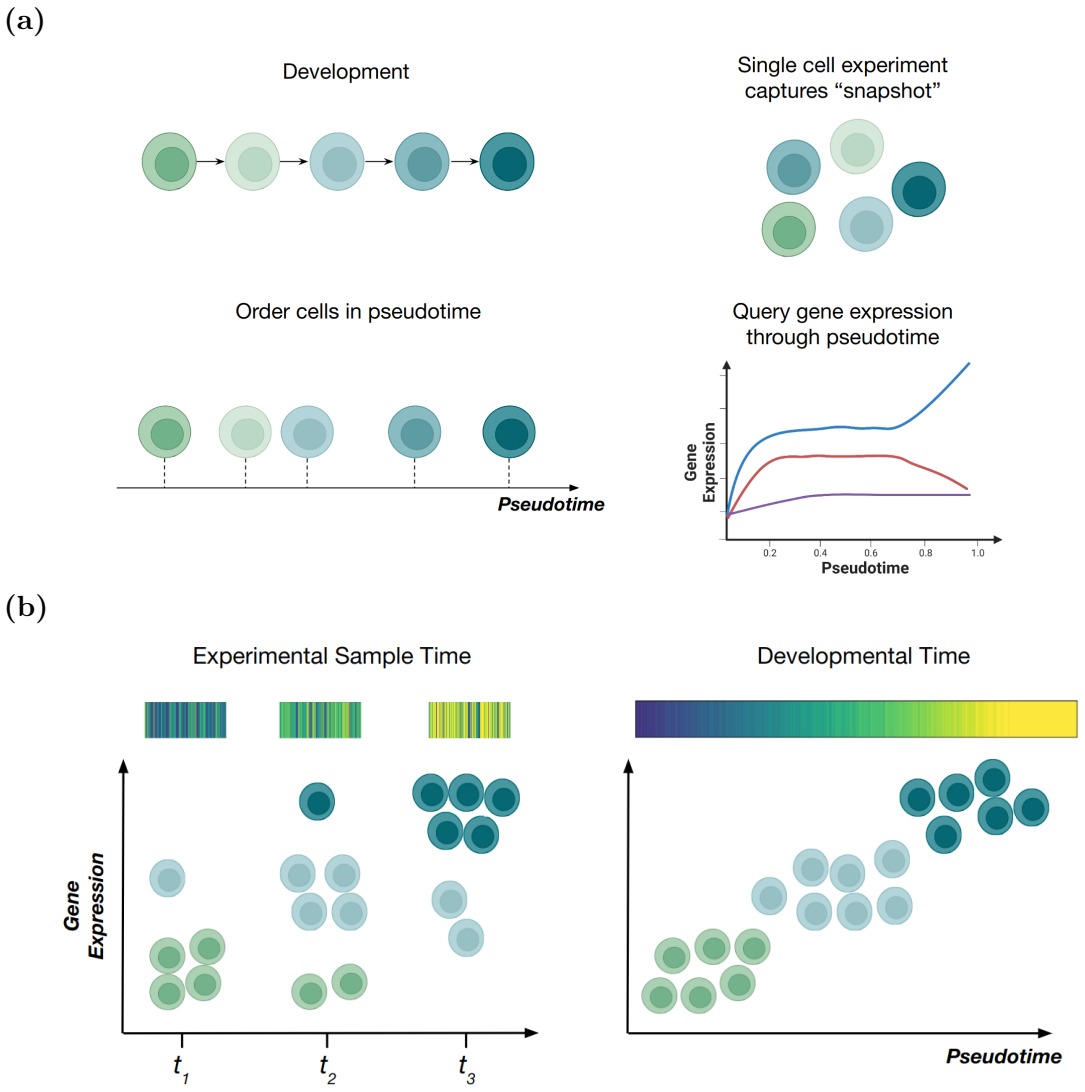


Figure 5.1: (a) A dynamic process is characterized by a sequence of cell states of distinct gene expression levels (top left); experimentally, single-cell data captures only a “snapshot” of this process (top right), made up of multiple asynchronous states; therefore, we reconstruct this trajectory from single-cell data via a pseudotime ordering (bottom left); this allows us to describe gene expression patterns specific to the process (bottom right). (b) Pseudotime does not necessarily coincide with sampling time; dynamic processes are asynchronous and samples contain multiple states, therefore querying gene expression of cells based on sampling timepoints (left) can lead to inaccurate trajectory reconstruction (heatmap, top); on the other hand, pseudotime reconstructs developmental time defined by gene expression changes (right), which leads to a smoother representation. Created with BioRender.com.

position of cells, as was seen with zonation patterns. This chapter will look at a quantitative description of this process within the VAE framework by inferring a trajectory that estimates the migration patterns. For this particular tissue, the developmental time correlates with spatial location, as progenitors in the crypt are differentiating while physically moving out and subsequently along the villus. We will show that even without spatial measurements we can make meaningful inferences about the trajectory of transcriptional changes from the single-cell data. We will then extend this framework to clumps, to infer the position of the cells retrieved from clumps on these trajectories.

5.2 Single-Cell Pseudotime Generative Model

In line with our objective of developing a robust end-to-end modeling framework, the aim of this chapter is to incorporate pseudotime within our VAE model. We highlight the following modeling aspects:

1. The dynamic trajectory is fit in the latent space. As mentioned before, the latent space is a low dimensional representation that summarizes the main biological signals in the data without technical effects. Modeling the trajectory in gene expression space incurs a high computational cost, and it would involve fitting noisy data. We thus define a smooth trajectory through the latent space, describing the developmental path, as a function of the pseudotime variable.

2. Pseudotime and latents are different variables. Previous work on VAEs and pseudotime incorporated the latter by setting the latent space dimension to 1 (e.g. Märtens & Yau, 2020). We argue that a one-dimensional latent space will poorly reconstruct the gene expression due to the constricted bottleneck. Therefore, it is best to retain a multivariate latent space. We also argue that pseudotime should be a separate variable from the latents: the latter encodes biological information from the dataset, e.g. cell types, sub-types, cell cycle etc., while the former specifically describes a trajectory and ordering of cell states.

We formulate the trajectory through the latent space as a function $\mathcal{S} : [0, 1] \to \mathbb{R}^D$, where we assume the pseudotime is normalized to the unit interval. The key component is how this trajectory is fit. Our approach focused on using *neural ODEs* (Chen et al., 2018a). We begin with a general overview below and then describe our extension of the single-cell VAE model for pseudotime inference.

For a function $\mathcal{S}(\cdot)$ and time variable t we consider the following ODE:

$$\frac{d\mathcal{S}}{dt} = h\left(\mathcal{S}(t), t\right) \tag{5.1}$$

with initial value $\mathcal{S}(t_0)$. Here the derivative $h(\cdot)$ is parameterized by a neural network. We chose an ODE formulation in order to describe the continuous dynamics of trajectory $\mathcal{S}(\cdot)$, and choosing a derivative function specified by a neural network allows a fair amount of flexibility. While it is intuitive to parameterize gene specific expression behaviours as interpretable functions of pseudotime, e.g. upregulation & downregulation via sigmoid curves or transient states via Gaussian curves (Campbell & Yau, 2018), this is not the case for dynamics in the latent space.

One of the key contributions in (Chen et al., 2018a) was developing the methodology for solving equation (5.1). The authors develop a black-box solver that scales linearly in time and incurs a low memory cost. Given time points $t_0, t_1, \ldots, t_N$ and initial state $\mathcal{S}(t_0)$, the ODE solver produces $\mathcal{S}(t_1), \ldots, \mathcal{S}(t_N)$.

Given these resources, we formulate the following single-cell generative model:

$$\begin{aligned}
\frac{d\mathcal{S}}{dt} &= h\left(\mathcal{S}(t)\right) \\
\boldsymbol{z}_0 &\sim \mathcal{N}(0, 1) \\
\mathcal{S}_{t_1}, \ldots, \mathcal{S}_{t_N} &\sim \text{ODEsolve}\left(\boldsymbol{z}_0, h, t_1, \ldots, t_N\right) \\
\boldsymbol{z}_n &\sim \mathcal{N}\left(\mathcal{S}_{t_n}, \text{diag}(\boldsymbol{\sigma}^2)\right), \ n = 1, \ldots, N \\
\boldsymbol{x}_{ng} &\sim \text{NB}(N_n f(\boldsymbol{z}_n)_g, \theta_g)
\end{aligned} \tag{5.2}$$

Given initial value $\boldsymbol{z}_0$ at time $t_0 = 0$, the ODE solver returns a deterministic trajectory at given time points $\{t_n\}_{n=1}^N$. The solver implementation is publicly available via open-source code [1]. We note how in this formulation the latents

[1] https://github.com/rtqichen/torchdiffeq

have their generative distribution centered around the trajectory, and the scale σ incorporates the variability of latents at any time point (Figure 5.2a). This model incorporates the 2 aspects discussed above: the multivariate and expressive latent space encodes the main biological signals and the continuous deterministic trajectory $\mathcal{S}(t)$, whose dynamics follow a neural ODE, constructs the gene development. We emphasize the relationship between $\boldsymbol{z}_n$ and $\mathcal{S}_{t_n}$: the latter captures trajectory information and thus using only this term during training (i.e. plugging the decoded $f(\mathcal{S}(t))$ into the Negative Binomial density) would limit the expressivity of reconstructed gene expression; the single-cell data captures other relevant signals besides the ones related to development, and these are captured by the former term.

We formulate the inference model as before, with variational posterior $q(\cdot\,|x_n)$ parameterized in the encoder and an additional function g encoding pseudotime, such that the value for cell n is $t_n = g(x_n)$. The ELBO is given by

$$\mathcal{L} = \mathbb{E}_q\left[\sum_{n=0}^{N} \log \frac{p(x_n|z_n)p(z_n|\mathcal{S}_{t_n})}{q(z_n|x_n)}\right] \tag{5.3}$$

$$= \sum_{n=0}^{N}\Big(\mathbb{E}_q\left[\log p(x_n|z_n)\right] - \mathrm{KL}\left(q(z_n|x_n)||p(z_n|\mathcal{S}_{t_n})\right)\Big) \tag{5.4}$$

including the root latent, with the convention $S_0 = z_0$. A schematic of the generative model and inference framework is shown in Figure 5.2b.

During training, values of z_n for each cell in a minibatch are sampled from the variational posterior, whose parameters are specified by the encoder. The first term in the ELBO, the reconstruction error, optimizes the quality of fit to the data. The second term is the Kullback-Leibler divergence between the variational posterior and generative distribution of the latents; this term provides the regularization between encoder output and pseudotime trajectory solved from the ODE. It is straightforward to compute since it is the KL divergence between two Normal distributions, where the scale σ parameter from (5.2) is inferred via gradient descent. The reconstruction error optimizes the conditional data likelihood given the latents, and the KL term encourages the latents to be close to the determinstic trajectory (Figure 5.2c).

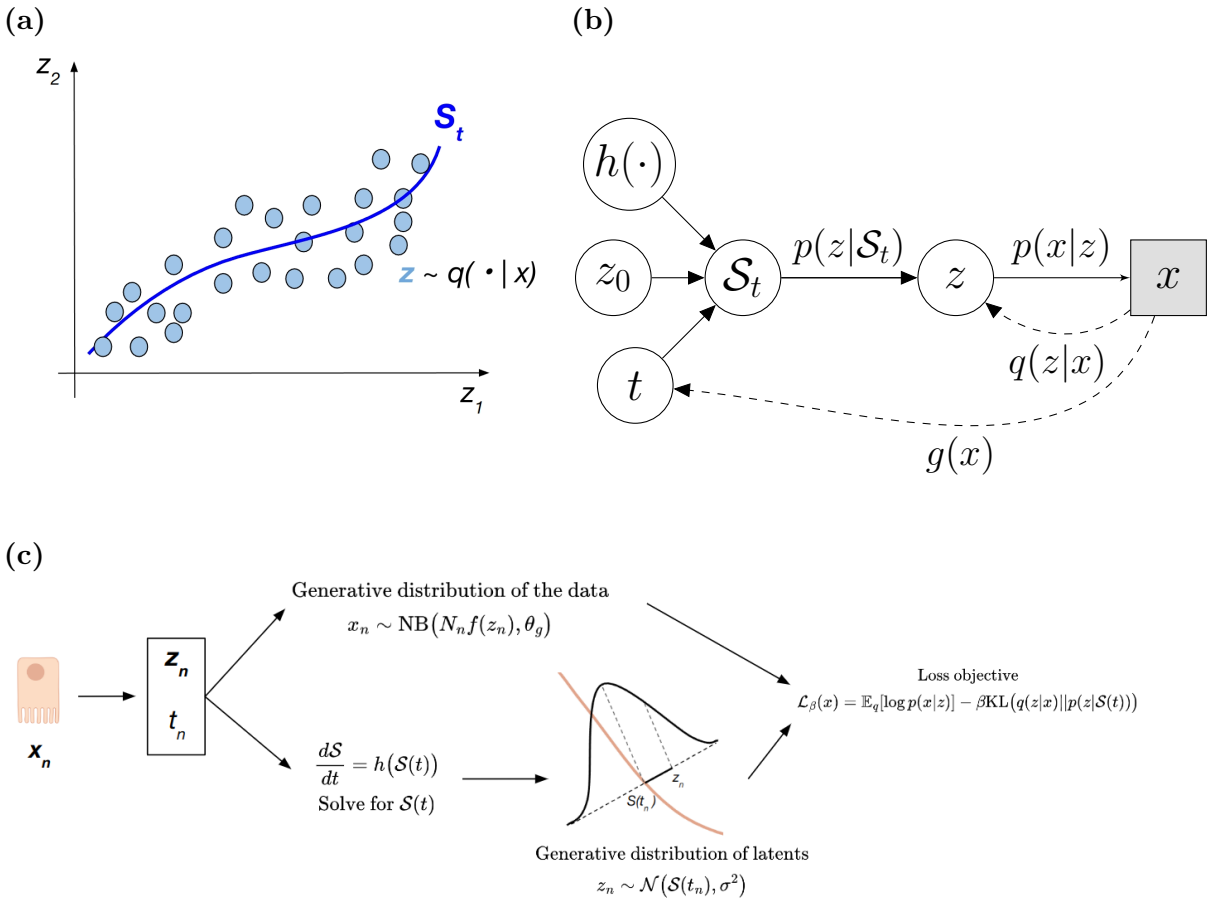


Figure 5.2: (a) Latents z_n sampled from the variational posterior congregate close to the determinstic trajectory $\mathcal{S}(t)$ (example shown for 2-dimensional latent space); (b) Generative model and inference framework for the single-cell pseudotime model; (c) Latents sampled from the variational posterior optimize two quantities in the VAE loss: reconstruction loss of the data and KL divergence between the generative and variational posterior distributions. Created with BioRender.com.

Furthermore, we can strengthen the regularization between the variational posterior and trajectory by formulating a $\beta-$VAE loss (Higgins et al., 2017); hence the ELBO becomes (dropping the subscripts):

$$\mathcal{L}_\beta(x) = \mathbb{E}_q \left[\log p(x|z)\right] - \beta \ \mathrm{KL}\left(q(z|x)||p(z|\mathcal{S}_t)\right) \tag{5.5}$$

The β hyperparameter in loss (5.5) regularizes the strength of the KL term between the variational posterior and generative distribution of the latents. In other words, the latents sampled from the variational posterior congregate closer to the trajectory

$\mathcal{S}(t)$. Supplementary Section B.1 looks at the β-VAE objective in more detail and shows that including the β term in the ELBO results in a rescaled latent space.

5.3 Application to the Gut Single-Cell Data

This chapter will analyze the single-cell and clump datasets sequenced from the intestinal epithelial tissue. The data involves multiple mature cell types and within a cell type population there are transcriptional changes occurring mainly due to the migration along the crypt-villus axis. There are no transitions between cell types and we are interested in modeling the intra-cell type development. For this reason, we extended the generative pseudotime model (5.2) such that trajectories are cell type specific:

$$\begin{aligned} \frac{d\mathcal{S}^{(c)}}{dt} &= h_c\left(\mathcal{S}^{(c)}(t)\right) \\ \boldsymbol{z}_n|c &\sim \mathcal{N}\left(S^{(c)}_{t_n}, \mathrm{diag}(\boldsymbol{\sigma}_c^2)\right) \\ x_{ng}|c &\sim \mathrm{NB}\left(N_n f(\boldsymbol{z}_n|c)_g, \theta_g\right) \end{aligned} \tag{5.6}$$

The main goal of this chapter is to link the pseudotime trajectories of two populations based on information from the clump data. Due to sample sizes and since the most abundant heterogeneous pairing in the current clump dataset is made up of enterocytes and goblet cells, the chapter will focus on these two populations.

Figure 5.3a shows the UMAP computed in the reconstructed gene expression space of the two cell type clusters, colored by the inferred pseudotime. The plot shows the fit for $\beta \in \{1, 5, 10\}$, highlighting how larger values of β in the loss function lead to more compact and constricted representations. Figures 5.3d and 5.3e show values of the negative log-likelihood (negative reconstruction error) and KL divergence, respectively, for these values of β, computed on the validation set. The figures also include the single-cell model from the previous chapter, without pseudotime and with a $\mathcal{N}(0, I)$ prior. The figures show the trade-off resulting from regularizing the KL term: higher values of β lead to lower KL divergence, whilst incurring a cost in the quality of gene expression reconstruction. As large values of

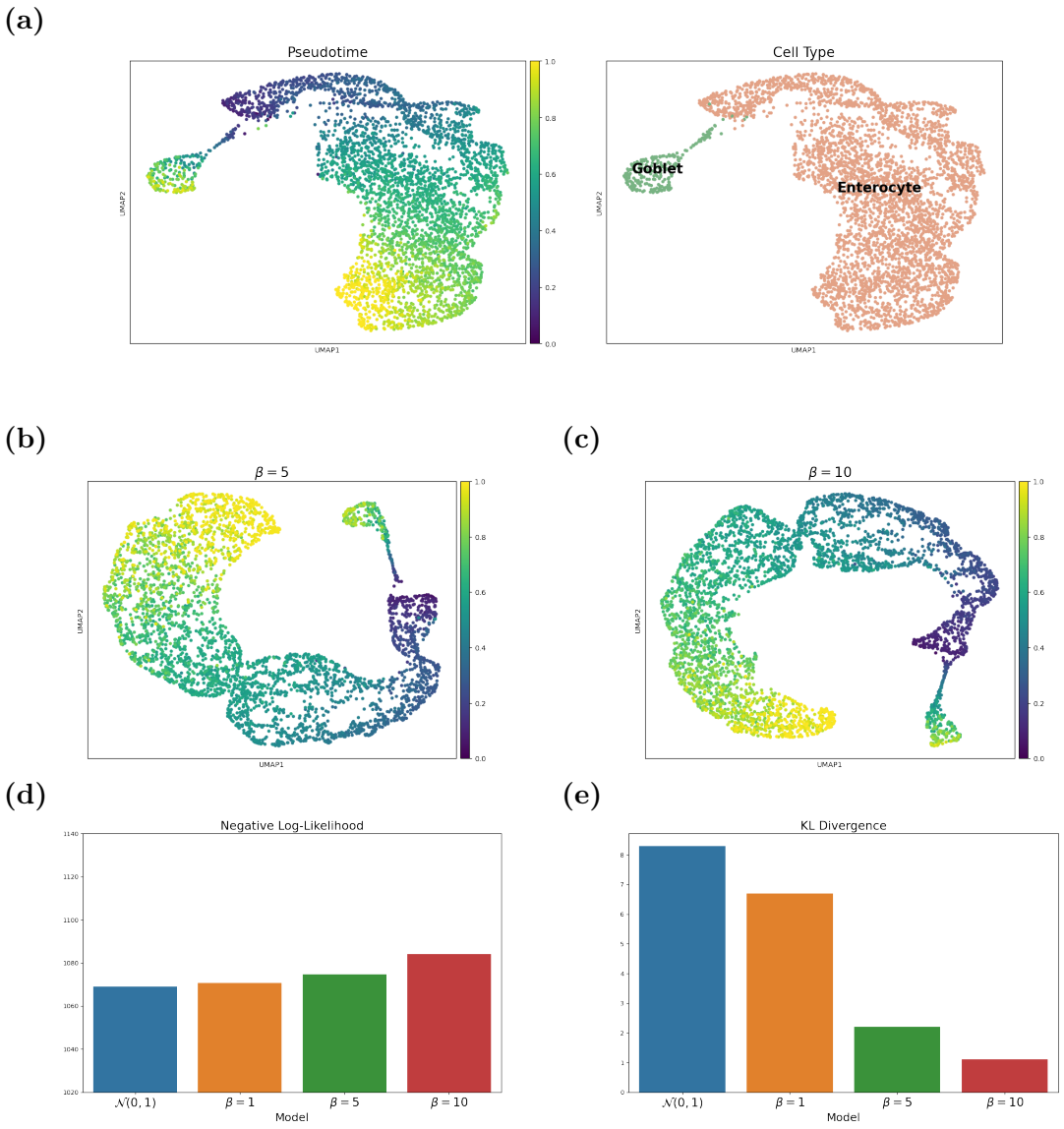


Figure 5.3: UMAP of enterocyte and goblet single-cell clusters computed in gene expression space from the pseudotime model, with (a) $\beta = 1$, (b) $\beta = 5$, (c) $\beta = 10$. Comparison of (d) negative log-likelihood and (e) KL divergence from the VAE loss, evaluated on a validation set, for different values of β and $N(0, I)$ prior (no pseudotime); lower values are better for both.

β force the posterior distribution to remain close to the trajectory, the expressivity of the latent space is reduced.

Using the model results, we can reconstruct gene specific expression patterns through pseudotime. The advantage of this model is that it provides both latents $\{z_n\}_n$ and trajectory values $\mathcal{S}(t)$. Both of these values can be mapped through

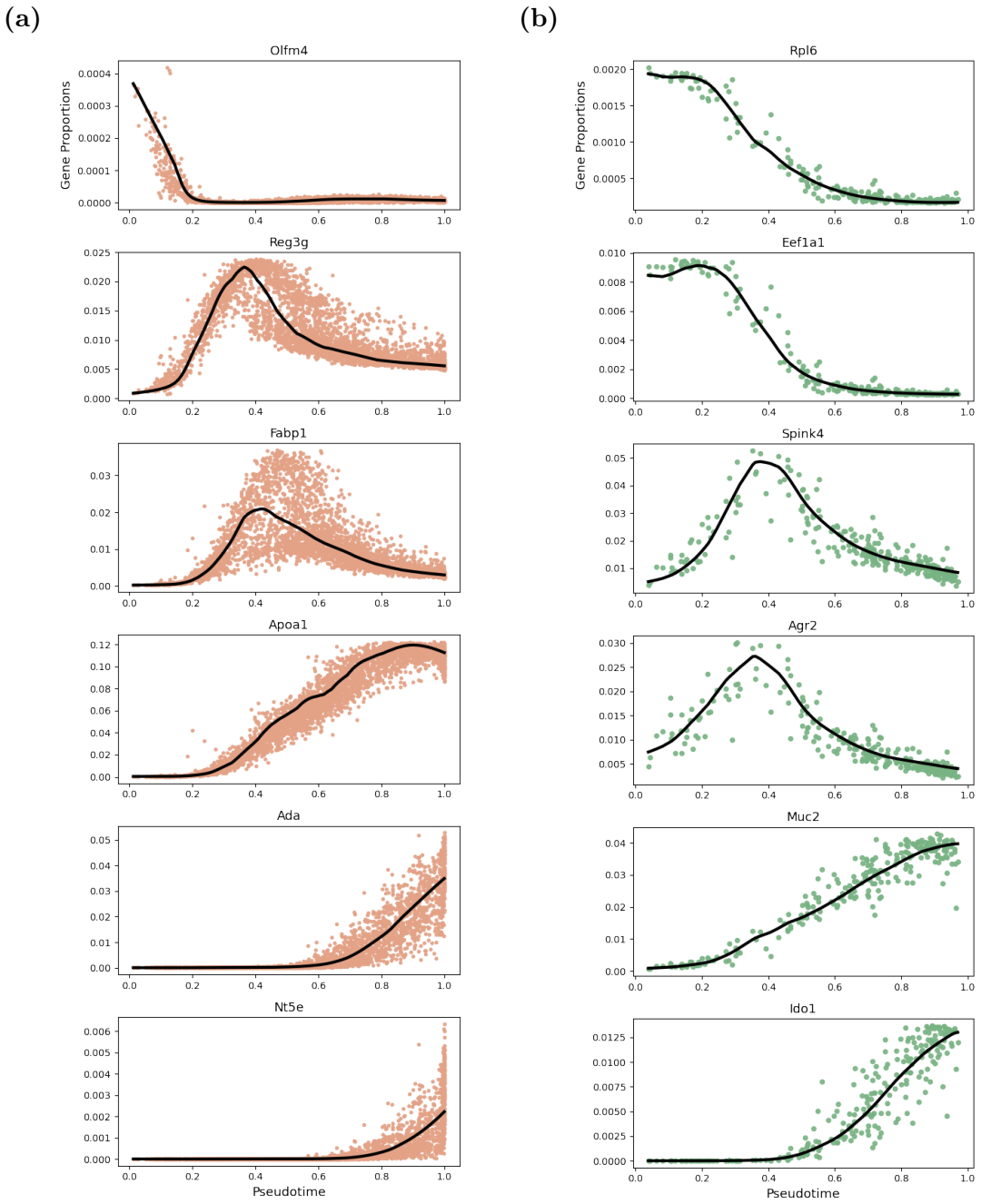


Figure 5.4: Gene expression through pseudotime for (a) enterocytes and (b) goblet cells, for a few key genes, showing reconstructed proportions (colored points) and pseudotime trajectory (black curves); the latter was computed by mapping the latent trajectory through the decoder.

the decoder to obtain reconstructed gene proportions; in particular, the ones obtained from the latter can be interpreted as denoised gene expression specific to the pseudotime trajectory. Figure 5.4 shows a scatter plot of reconstructed gene proportions against pseudotime for several genes specific to enterocytes (Figure 5.4a) and goblet cells (Figure 5.4b), comprising both the decoded expression of latents (colored points) and trajectories (black curve).

This section showed that the single-cell generative model with pseudotime inferred via neural ODEs captures the gene progression trajectory. In the context of the gut data, this trajectory describes the transcriptional changes taking place during the migration of enterocytes and goblet cells along the crypt & villus. These two distinct populations are migrating together, thus it is reasonable to assign each their own pseudotime trajectory. We are able to map this trajectory into gene proportion space to obtain a continuous expression path as function of pseudotime. Neverthelss, as emphasized in the previous chapter, one limitation of single-cell data is that it provides no information on how these two dynamic trajectories relate to each other. To this end, we extended the pseudotime model for clump data.

5.4 Time Deconvolution from Clumps

We assume a dataset of clumps composed of two cell types; either cell will be sampled from some point along its cell type's trajectory. This section focuses on developing a deconvolution model that takes this clump dataset as input and returns, for each clump, two retrieved cell type profiles *and* their corresponding trajectory information.

The approach is similar to the one in the previous chapter: using the frozen neural network components from the single-cell embedding that map latents to gene expression space. In this case, these will also include the pseudotime trajectories. We assume the cells retrieved from the clumps follow the same dynamics as the single-cell populations (Figure 5.5). We note the neural ODE solution does not have a closed form and the solver only returns values at given time points, therefore the deconvolution model uses the frozen derivative function $h(\cdot)$. Querying the derivative

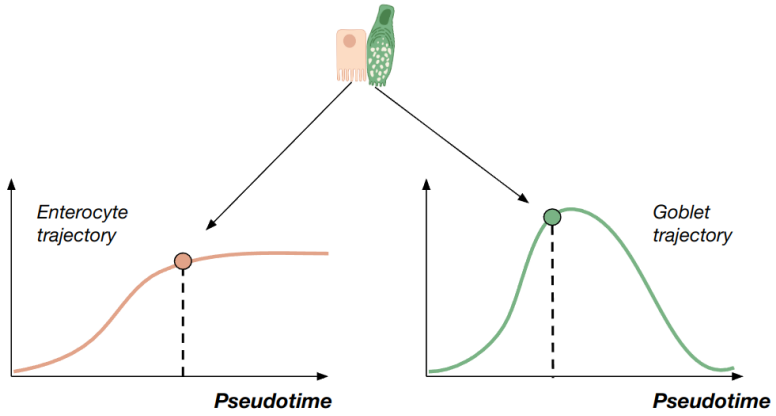


Figure 5.5: Cells in clumps are assumed to follow the same dynamics as single cells. For either cell the clump deconvolution model infers its location on the cell type specific trajectory. Created with BioRender.com.

function is not sufficient for reconstructing the trajectory, thus the model also samples the root latent $\boldsymbol{z}_0^{(c)}$ from the single-cell trajectories, for each cell type population.

We formulate the generative model as follows, with blue indicating frozen components:

$$[t_n^{(1)}, t_n^{(2)}] = \widetilde{g}(\boldsymbol{y}_n) \tag{5.7}$$

$$\boldsymbol{z}_0^{(c)} \sim q(\cdot\,|\boldsymbol{x}_0, c),\ \ c = 1, 2 \tag{5.8}$$

$$\mathcal{S}_{t_1}^{(c)}, \ldots, \mathcal{S}_{t_M}^{(c)} = \text{ODEsolve}\left(z_0^{(c)}, h^{(c)}, t_1^{(c)}, \ldots, t_M^{(c)}\right) \tag{5.9}$$

$$\boldsymbol{z}_n^{(c)} \sim \mathcal{N}\left(\mathcal{S}_{t_n}^{(c)}, \text{diag}(\boldsymbol{\sigma}_c^2)\right) \tag{5.10}$$

where $q(\cdot|\boldsymbol{x})$ is the single-cell model encoder, $\boldsymbol{x}_0$ is the root cell of type c, and $\{\mathcal{S}_{t_n}^{(c)}\}$ are points on the cell type specific trajectory.

Pseudotime of each cell type in the clump is computed from the clump expression data $\boldsymbol{y}_n$ via function $\widetilde{g}$. The inference scheme is similar to the one from the single-cell model, using the same generative distribution for latents. This encourages the latents sampled from the clump model encoder to map close to the single-cell ones and on the pseudotime trajectory, which is computed from the sampled root latent and frozen derivative function. During training, this trajectory is fixed and thus incentivizes the pseudotime function to search for the best value where the latent variable fits along this trajectory.

We applied this time deconvolution model to a subset of the gut clump data from Chapter 4 that expressed both enterocyte and goblet marker signatures (Figure 4.5a). For each clump, the model returns two latents for every cell type: one that incorporates the biological signal and one corresponding to the pseudotime trajectory. The generative distribution encourages these two latents to be close to each other: the former has a level of flexibility that leads to a lower reconstruction error, while the latter follows the dynamics dictated by the neural ODE from the single-cell model. We compared the resulting trajectory latents from clumps with the single-cell ones, for both cell types (Figure 5.6). From the model design, the trajectory latents from clumps follow the same ODE solution as the single-cell ones; however, the flexible elements are the pseudotime values of these latents, which predict their location along the trajectory, and are inferred during training. Figures 5.6a and 5.6b confirm that trajectory latents are mapped at pseudotime points where they coincide with the single-cell trajectory. As a sanity check, we also compared the inferred pseudotime of retrieved cells with the values in neighboring single cells in a KNN graph computed in the latent space (Figure 5.6c, $K = 3$).

5.5 Identifying Gene Expression Modules Across Pseudotime

The aim of this section is to apply the models developed in this chapter to infer gene expression patterns through pseudotime, which would reflect the transcriptional changes occurring during migration along the crypt-villus axis. We will make use of the pairing information from clumps to relate the trajectories of enterocytes and goblet cells.

The previous chapter assumed a partition of the enterocyte dataset based on expression of known zonated genes. Each enterocyte was assigned a zone it most likely came from. Using the retrieved enterocyte profiles from pairings with goblet cells, it was possible to also assign zones to the retrieved goblet cells. The subsequent analysis was based on these discretized subsets. This chapter will focus on a more

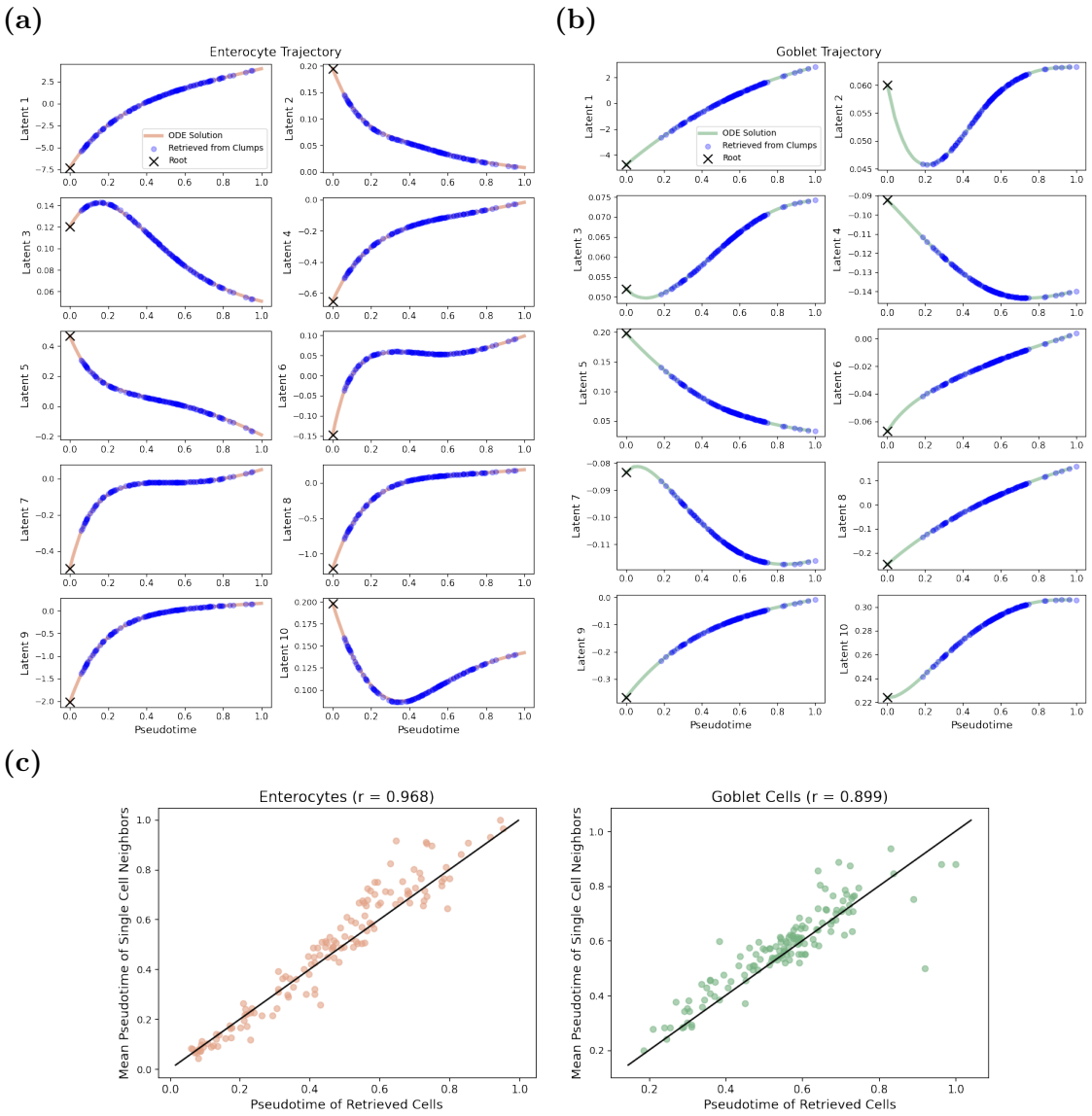


Figure 5.6: Comparing the single-cell latent trajectory (i.e. solution to the neural ODE) against the trajectory latents of retrieved cells from clumps, for (a) enterocytes and (b) goblet cells. (c) Comparing the pseudotime of retrieved cells against values in neighboring single cells in the latent space.

continuous representation of these gene expression changes. While we do not have access to spatial information for this single-cell dataset, we can use the inferred pseudotime as a surrogate for the aforementioned migration process.

The advantage of the pseudotime latent trajectory $\mathcal{S}(t)$ is that it provides a continuous and denoised representation of the developmental process. While the latents $\{z_n\}_n$ sampled from the variational posterior capture a multitude of

biological signals, the trajectory constructs a specific signal. Furthermore, computing pseudotime specific gene expression patterns is straightforward, as it only involves passing $\mathcal{S}(t)$ through the decoder. The resulting gene expression trajectories provide a continuous and smoothed representation, as exemplified in Figures 5.4a and 5.4b.

The latent trajectory can be seen as a cell-cell similarity matrix pertaining strictly to pseudotime ordering, in absence of other noise. For this reason, we ran Hotspot with this representation in order to identify gene modules that vary across pseudotime. The procedure is similar to Section 4.5.2, with the key difference of using the continuous latent trajectory. Using the time deconvolution model for clumps in Section 5.4, we were are able to infer the trajectory positions of both retrieved enterocytes and goblet cells and thereby relate the two trajectories.

The first step is analyzing the gene modules inferred by Hotspot from the enterocyte single-cell data. Figure 5.7a shows the heatmap of local correlations and modules computed by Hotspot. Given specific conditions [2], Hotspot grouped the informative genes into 5 modules. Each module and its associated genes show expression at specific regions in the pseudotime trajectory, as shown in Figure 5.7b, plotting the expression values from the decoded trajectory. We ran a functional profiling analysis via g:Profiler - a few selected entries are shown in Section B.2. We also found that these 5 modules were very similar to the gene clustering results of (Moor et al., 2018). In their case they group genes based on zonation patterns, using known expression references from 7 regions along the crypt-villus axis where they performed spatial sequencing via Laser Capture Microdissection.

We describe the Hotspot enterocyte modules in order of peak expression through pseudotime. Module 2 (Figure 5.7b, orange) is strongly enriched for cell cycle, which is characteristic for enterocyte precursors in the crypt. Module 1 (blue) is enriched for ribosomal activity, peptide & amide metabolic processes. Module 4 (red) is enriched for glutathione binding, oxireductase activity and peroxidase activity. Module 5 (purple) exhibits genes related to carbohydrate digestion & absorption, peptide

[2]Each module to contain a minimum of 100 genes, with adjusted p-values of local correlations below 0.05.

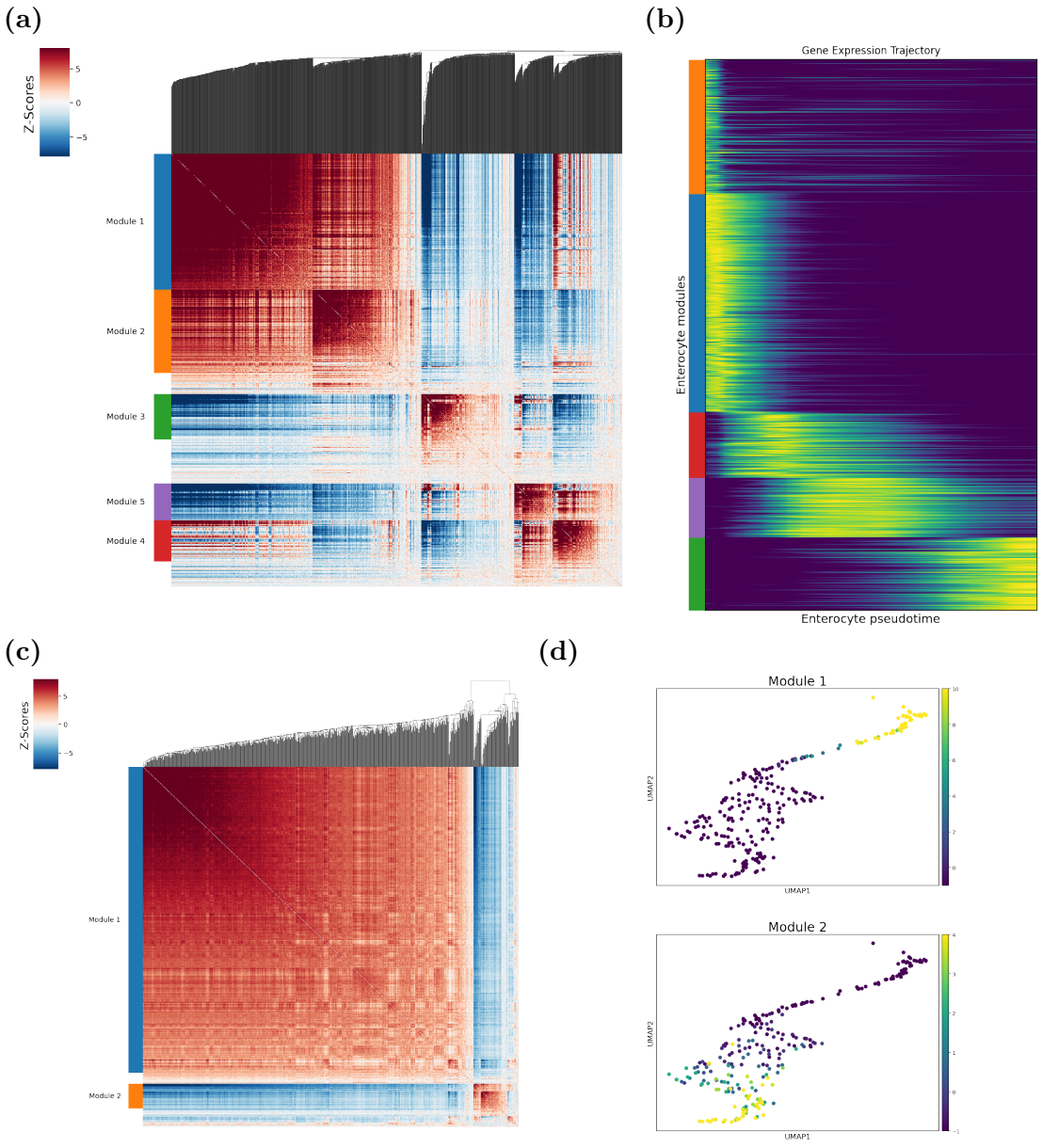


Figure 5.7: Hotspot analysis of enterocyte single-cell data using the latent trajectory as cell similarity graph. (a) Heatmap of pairwise gene local correlations and assigned modules. (b) Heatmap of trajectory expression through pseudotime for genes from each module; in this plot, values across cells are normalized to the unit interval for each gene. (c, d) Hotspot analysis of goblet single-cell data, showing (c) heatmap of gene pairwise local correlations and assigned modules; (d) UMAP of goblet cluster with module scores.

catabolic process, lipid metabolic process; this is related to nutrient transport and these processes were found to be enriched in mid-villus. Finally, module 3 (green) is enriched for cholesterol absorption, interferon-β production & cytokine production. The enrichment of cholesterol aborption in module 3 agrees with findings from (Moor et al., 2018) that these programmes occur further up the villus than nutrient transport. Our findings agree with previous biological results that enterocytes show a division of labour, where they perform different functions depending on their location in the villus (Moor et al., 2018; Adler et al., 2019; Adler et al., 2023).

We next applied Hotspot to the single-cell goblet cluster. This revealed two gene modules, as it did in the previous chapter, with module 1 comprising a much larger amount of genes, as shown in the local correlation plot in Figure 5.7c. The two modules show peak expression in opposite sides of the crypt-villus axis (Figure 5.7d). We ran functional profiling on these 2 modules (Section B.2). Goblet module 1 is predominantly enriched for ribosomal activity; other programmes include peptide metabolic and amide biosynthetic processes. We found much fewer enriched programmes for module 1; these include protein binding and regulation of retinoic acid-inducible gene I (RIG-I) signaling pathway, which is involved in induction of type I interferon (Rehwinkel & Gack, 2020).

The objective of this section is to use information from clumps to relate the transcriptional changes through pseudotime between the two populations. The enterocyte population exhibits multiple gene modules, possibly indicating the zonated distribution of biological functions; the key question is: while enterocytes are migrating and performing different tasks, what transcriptional changes are taking place in the goblet cells attached to them?

We sought out a quantitative approach to relate them in a more precise way. Comparing the two pseudotime values across clumps (Figure 5.8a) shows a potentially non-linear relationship: the goblet pseudotime values are "ahead" of the enterocytes ones, but in later stages they level off. The retrieved cell type profiles from clumps define a correspondence between the trajectories and we sought out to generalize this to single cells. To this end, we fit a B-spline to the pseudotime values and

formulated the goblet pseudotime as a function of the enterocyte one, as shown in Figure 5.8b (alternatively we could also define the reverse relationship). The spline provides a smoothed function that can be queried by enterocyte pseudotime values, and this allows us to extend the ability to compare trajectories to single cells. Therefore, enterocytes following their trajectory can be mapped onto the goblet pseudotime ordering in order to study how gene behaviour is synchronized between the two populations.

We show the importance of relating the trajectories via an example. When looking at the cell type specific pseudotime trajectories, one might wonder whether there is an implied correspondence between them, i.e. if both enterocytes and goblet cells migrate, then the pseudotime values should be roughly aligned anyway. For instance, the genes *Spink4* in single goblet cells and *Reg3b* in enterocytes show a transient behaviour through pseudotime (bell curve shape), and they achieve peak expression around the same values of their own pseudotime (Figure 5.8c - note the pseudotime axis here contains the sets of pseudotime values for both populations), overall showing a highly correlated behaviour. Nevertheless, after applying the spline transformation to the enterocyte pseudotime, the two genes show a different behaviour (Figure 5.8d), namely that *Spink4* becomes upregulated in goblet cells before it does in neighboring enterocytes.

Finally, we use this mapping to relate the enterocyte and goblet modules. For each module we compute the mean expression across pseudotime across all genes in that module, and standardize the expression to the [0, 1] interval (for visualization purposes). We map the enterocyte pseudotime onto the goblet one, as done above, and plot the two sets of modules together. Figure 5.8e shows the enterocyte and goblet modules that show a decreasing behaviour in pseudotime, and Figure 5.8f shows the ones with increasing behaviour. These plots confirm the results from Figure 4.7e. Knowing that the pseudotime trajectories are now scaled, we can draw insights about the joint behaviour of the two populations.

The figures show that goblet cells expressing module 1 genes are likely attached to enterocytes expressing genes of modules 1 & 2. Genes from goblet module

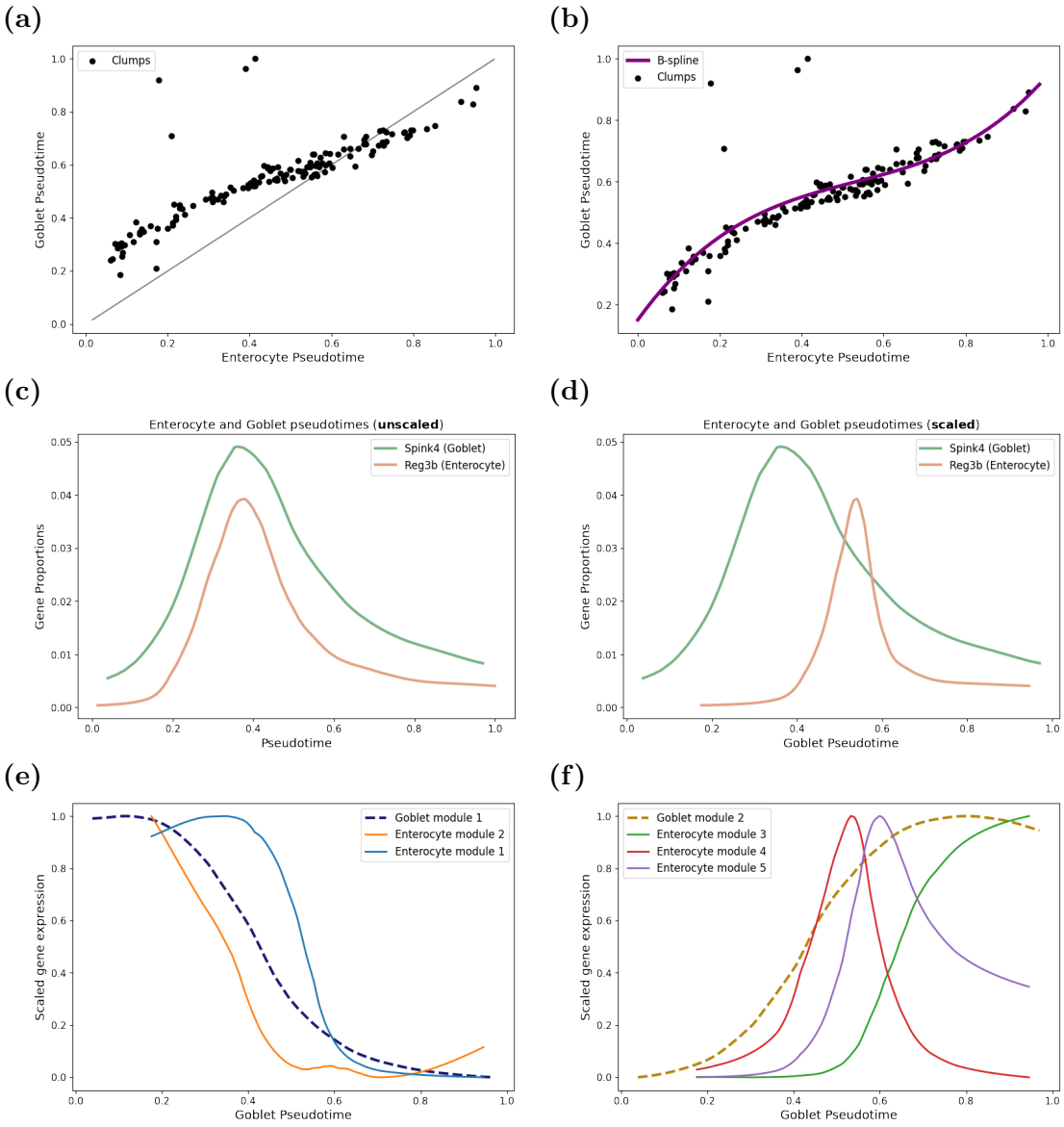


Figure 5.8: Relationship between enterocyte and goblet pseudotime trajectories. (a) Scatterplot between the two sets of pseudotime values in clumps. (b) Fitting a continuous spline to describe the goblet pseudotime as a function of the enterocyte values. (c, d) Expression of *Spink4* in goblet cells and *Reg3b* in enterocytes through pseudotime (gene proportions computed from trajectory latents) in case of (c) overlaying the two cell type specific trajectories and (d) when mapping the enterocyte pseudotime on the goblet one via the spline function. (e, f) Enterocyte and goblet Hotspot modules plotted on the common goblet pseudotime, splitting the modules by (e) decreasing and (f) increasing expression patterns.

2 start becoming upregulated around the same time the attached enterocytes express (their corresponding) module 4. The expression of these goblet specific genes keeps increasing as neighboring enterocytes show expression of enterocyte specific modules 4, 5 & 3.

It is thus interesting to see that modules related to ribosomal functions and cell cycle become downregulated together - this may indicate an almost simultaneous discontinuance of crypt specific programmes in both populations. Goblet module 2 becomes upregulated around the same time as enterocyte module 4, i.e. when enterocytes begin expressing *Reg* genes, *Nlrp6*, *Lypd8* and *Il18*; these genes are involved in microbiota-host interactions (Moor et al., 2018). As previously stated, module 4 is also enriched for glutathione binding, metabolic pathways and oxidoreductase activity. It would be interesting to see if goblets simultaneously also exhibit an immune programme in this region, or whether their change in modules is triggered by an immune response.

We note the results in this section agree with what we found in Section 4.5 (Figure 4.10). We emphasize the advantage of our pseudotime model is that we are able to directly query the two trajectories. Their representations are much less noisier than the space of posterior sampled latents, which Hotspot was applied on in the previous chapter. Furthermore, pseudotime provides a specific ordering for both populations and we are thus able to relate the dynamics of gene expression. It would be interesting to further develop models for co-expression in this framework, to assess how strongly the co-expression of two modules is related to pseudotime (and thereby to location).

5.6 Pseudotime, Identifiability and RNA Velocity

It is worth highlighting the identifiability aspects of the proposed pseudotime model (5.2). It does not require extra information in the data (input is the gene count matrix), thus fitting a trajectory through the latent space relies on using the existing variation and signals in the data. For instance, the data in this chapter exhibits a large signal related to the crypt-villus migration, as that accounts for major

differences in gene expression between cells. In other words, the model introduces more parameters and constraints, without further biological prior knowledge or structure, which leads to identifiability problems.

In recent years, research has been focused on RNA velocity (La Manno et al., 2018), which consists of modeling gene expression data containing counts of unspliced and spliced RNA transcripts (Zeisel et al., 2011). These are outputs of the transcriptional dynamics, consisting of transcription (resulting in unspliced transcripts), splicing and degradation of spliced transcripts. The key concept of RNA velocity is that the ratio of the 2 quantities predicts the progression of gene expression. Namely, a high relative amount of unspliced mRNA indicates gene upregulation (low amount of mature transcripts, succeeded by high rate of splicing), while the opposite indicates downregulation (low rate of splicing, mature transcripts degrade).

Most models quantify the transcriptional dynamics described above via a system of coupled ODEs which have analytical solutions (La Manno et al., 2018; Bergen et al., 2020), where time is the independent variable. If $U(\cdot)$ and $S(\cdot)$ denote the time-dependent amounts of unspliced and spliced transcripts, the dynamics can be represented as follows:

$$\emptyset \xrightarrow{\alpha} U(t) \xrightarrow{\beta} S(t) \xrightarrow{\gamma} \emptyset \tag{5.11}$$

where α, β, γ are the rates of transcription, splicing and degradation. This leads to the following linear ODE:

$$\frac{dU(t)}{dt} = \alpha - \beta U(t) \tag{5.12}$$

$$\frac{dS(t)}{dt} = \beta U(t) - \gamma S(t) \tag{5.13}$$

The RNA velocity is defined as the derivative $\frac{dS(t)}{dt}$.

This framework certainly adds more interpretability towards the concept of trajectory inference. Deep learning models for RNA velocity have also been recently developed, such as VeloVAE (Gu et al., 2022), VeloAE (Qiao & Huang, 2021), VeloVI (Gayoso et al., 2023), DeepVelo (Chen et al., 2022b).

Future work should focus on extending the deep generative model framework for RNA velocity to clump data, opening up research avenues into how neighboring cells may affect the splicing dynamics.

5.7 Discussion

This chapter tackled the problem of inferring an ordering of cells that captures the transcriptional changes they undergo during a dynamic process. In the gut epithelial tissue, studied in this chapter, this dynamic process involves the migration of enterocytes and goblet cells along the crypt-villus axis, characterized by zonation patterns. This reconstructed trajectory is summarized via pseudotime. We extended the VAE framework for single cells to include a dynamic component representing this trajectory, quantified via a neural ODE in the latent space. The pseudotime generative model contains two key latent components, which are optimized for different objectives: (1) the smooth trajectory, representing a solution to the neural ODE; and (2) expressive latents, encoding multiple biological signals and optimizing the reconstruction error of the gene expression data. We formulated the relationship between (1) and (2) via a generative distribution; during inference, it encourages these two components to be close to each other. The advantage of this framework is being able to formulate a multivariate latent space, without sacrificing much of the reconstruction loss *and* quantify the cell ordering via a one-dimensional pseudotime variable. The former encodes other signals besides the ones related to trajectory information, and the latter allows us to explore gene expression patterns through pseudotime.

We extended this single-cell model for deconvolution of cell types from clumps, such that we can retrieve trajectory information for each cell in a clump. Using the ODE solution from the single-cell model, we extended the transfer learning framework to use the frozen trajectory as a prior for the retrieved latents. The latent output of the clump model comprises the same two components described above. While estimating the pseudotime of retrieved cells can be done post-hoc by averaging

values of single cells in a KNN graph, we argue this approach is more noisy and agnostic to the magnitude of distance to the nearest neighbors. This brings about the key utility of the latent trajectory, as detailed below.

The downstream analysis in this chapter focused specifically on variation along pseudotime. We highlight the advantages of using the latent trajectory $\mathcal{S}(t)$ and its decoded version for this purpose. Both components form continuous and smooth representations in their respective spaces. The latent trajectory can be used as a cell similarity matrix, where proximity is evaluated solely based on the pseudotime ordering. Therefore, Hotspot applied on this latent representation identified gene modules expressed in particular regions of the trajectory. We contrast this to the analysis in the previous chapter, where the similarity graph was given by the latent space that encodes multiple biological signals; therefore, assessing the sources of biological similarities between elements in a KNN graph is less straightforward. We note that published pseudotime methods generally rely either on a post-hoc assignment of values to cells, which orders them along a trajectory but does not change their low dimensional representation, or tailoring the latent representation to one dimension, which leads to reduced expressivity.

The results we obtained from this model provided potentially interesting biological insights. Comparing the trajectories of the two populations and being able to query not just the pseudotime of a cell, but also the approximate pseudotime of a potential neighbor in the tissue is certainly a special feature provided by the clump data. Our findings lead to further questions about the cell dynamics in the intestinal epithelium: do enterocyte and goblet cells migrate at different rates? Is the joint and almost simultaneous upregulation of certain programmes between the populations influenced only by their location? We are certain that ClumpSeq can provide many novel insights, especially when paired with expressive and flexible generative VAE models.

In the last 2 chapters we validated our deconvolution model on discrete populations, which are quite different from each other. We next sought out to apply ClumpSeq to study a complex biological process that comprises a continuous differentiation process from one cell type to another, as opposed to sequential

states within discrete populations, and for which the surrounding microenvironment is known to influence cell fate. For this reason, we studied the application of ClumpSeq in the context of the Endothelial-to-Hematopoietic transition, as will be detailed in the next chapter.

Chapter 6

Endothelial-to-Hematopoietic Transition

6.1 Background

Hematopoietic Stem Cells (HSCs) represent the building block of blood cells in vertebrates, having the ability to differentiate into any hematopoietic cell and replenish these cell populations throughout life. They are functionally characterized as capable of long-term repopulating the blood system of irradiated mice after transplantation.

The development of the hematopoietic system consists of multiple asynchronous waves with distinct origins. In the murine embryo, the first wave occurs at embryonic day 7.25 (abbreviated E7.25) and consists of primitive erythroid cells originating in the yolk sac (YS) blood islands. These cells form only during this early stage of development, have a limited lifespan and their main role is to transport oxygen and support embryo growth until definitive erythrocytes are produced (Palis et al., 1999). This primitive erythroid lineage originates directly from mesodermal precursors. Definitive hematopoiesis, defined by the cells' morphological resemblance to adult cells, first takes place at E8.25 and is characterized by emergence of erythromyeloid progenitors (EMPs) within the yolk sac, which give rise to definitive erythrocyes, macrophages, megakaryocytes and other myeloid lineages.

Definitive HSCs with repopulating potential are not formed until E10.5. They first and autonomously emerge in the major arteries of the embryo (Müller et al.,

1994; Medvinsky & Dzierzak, 1996), namely: the dorsal aorta, in the area of the aorta-gonad-mesonephros (AGM) region, the vitelline and umbilical arteries (de Bruijn et al., 2000). HSCs are also found in the yolk sac and fetal liver one day after emergence in the AGM, and are thought to migrate there via the bloodstream. After E11.5 HSCs colonize the fetal liver and later, at E16.5, they start to colonize the bone marrow (Coşkun et al., 2014) where they will reside throughout the adult life, producing a continuous supply of blood cells.

While hematopoietic stem and progenitor cells (HSPCs) are located in the bone marrow during adulthood, their origins are endothelial and they start being produced in the embryo even before the formation of bone marrow. They are generated from a specialized tissue called Hemogenic Endothelium (HE), located in the wall of the main arteries of the embryo, via a process called Endothelial-to-Hematopoietic transition (EHT). This process was first experimentally validated in the chicken embryo (Jaffredo et al., 1998): endothelial cells were labelled with low-density lipoproteins coupled with fluorescent dye DiI; tracing the progeny of the labelled cells revealed expression of the dye in the emerging hematopoietic cells. Later, EHT progression was observed in the zebrafish and mouse embryo via *in vivo* lineage tracing studies (Zovein et al., 2008; Bertrand et al., 2010; Boisset et al., 2010; Kissa & Herbomel, 2010). For example, the study in (Zovein et al., 2008) consisted of labelling VE-Cadherin, a cell-adhesion molecule with expression restricted to endothelial cells. They found the VE-Cadherin labelled endothelium gives rise to emerging functional HSCs, and that AGM hematopoietic cells are exclusively endothelial derived.

The morphology of emerging hematopoietic cells consists of discrete clusters attached to the lining of the aortic wall. These intra-aortic hematopoietic clusters (IAHC) appear during the interval E9.5-E11.5 (Yokomizo & Dzierzak, 2010) and HSCs were found to be located specifically in the ventral side of the Dorsal Aorta (Taoudi & Medvinsky, 2007); see Figure 6.1a for a pictorial representation of the mouse embryo and location of the dorsal aorta. The transcription factor *Runx1* is required for definitive hematopoiesis to proceed in the AGM (Okuda et al., 1996; Wang et al., 1996). It is expressed at E9.5-E10.5 in the IAHC and hemogenic endothelial

cells in the vitelline & umbilical arteries and the dorsal aorta (North et al., 1999). Furthermore, *Runx1* is required for the formation of hematopoietic clusters (North et al., 1999; Cai et al., 2000; Yokomizo et al., 2001), as *Runx1*-deficient mice fail to produce IAHC and hematopoietic cells, and plays a vital role in the formation of IAHC, hematopoietic cells and HSCs specifically during the differentiation of endothelial cells (North et al., 1999; Cai et al., 2000; Chen et al., 2009).

Identification of HE via fluorescent sorting has been difficult due to the large overlap in cell surface marker expression between HE and HSPC. Consequently, many studies used combinations of endothelial and hematopoietic cell surface markers in order to separate the HE population from regular endothelium. An important breakthrough came with identification of the *Runx1* +23 enhancer, active specifically in HE and HSPC populations, which led to designing a transgenic mouse line with +23 enhancer mediated GFP [1] expression, referred to as *23GFP* (Nottingham et al., 2007; Bee et al., 2010; Swiers et al., 2013a). This transgenic mouse model labels all the endothelial cells with hematopoietic potential. Furthermore, transcriptional analysis of cells labelled by 23GFP reveals the continuum of the EHT trajectory, characterized by a gradual activation of the hematopoietic programme (starting as early as E8.5, up to E10.5) with a concomitant decrease in endothelial potential (Swiers et al., 2013a). The authors also tested the functional capacities of 23GFP cells sorted in interval E8.5-E10.5 via clonal cultures, and found that endothelial and hematopoietic potentials are mutually exclusive. Interestingly, *Runx1* expression peaks during E9.5-E10.5 and quickly drops afterwards, pointing out to the transient nature of the process.

Much experimental work has been done to phenotypically characterize the intermediate cell populations throughout EHT. The two broad and successive stages of development that will be referenced in this chapter are pro-HSCs (Rybtsov et al., 2014) and pre-HSCs (Taoudi et al., 2008; Rybtsov et al., 2011); their generation is characterized by upregulation of hematopoietic surface markers CD41 for pro-HSCs and CD43 & CD45 for pre-HSCs. A pictorial representation of the position of

[1] GFP = green fluorescent protein

these cell types within the ventral side of the dorsal aorta, where HSPCs emerge, is shown in Figure 6.1b.

EHT exhibits a complex & coordinated sinergy between various transcription factors throughout the process (Swiers et al., 2013b;, Ottersbach, 2019). While recent research has shed some light on the mechanisms and transcription factors involved in EHT, there is still much to be discovered about how the hemogenic specification of endothelial cells occurs *in vivo*. In addition to the complex molecular mechanisms and interactions, there are also extrinsic factors involved in this regulation, from the neighboring niche.

6.2 Microenvironment Plays a Key Role in EHT

A major role in regulating EHT is played by the AGM niche, which is responsible for producing signals that support the transition. The AGM niche is part of the surrounding microenvironment, whose composition is still poorly understood, especially in terms of how it promotes EHT. The AGM exhibits a dorsal-ventral polarity, with HSCs emerging mainly from the ventral region (Taoudi & Medvinsky, 2007) and its activity is a mixture of signals from the dorsal, ventral and lateral components (Souilhol et al., 2016).

A few studies have focused on the distribution and effect of AGM microenvironment factors *in vivo*. For instance, (Azzoni et al., 2018) found that *Kit* ligand (*Kitl* or Stem Cell Growth Factor), which binds to the tyrosine kinase receptor *Kit*, is required for HSC maturation and survival, specifically beginning at the pre-HSC stage. They also discovered, via a KitlTOM [2] reporter mouse line, that *Kitl* is expressed all ventral mesenchyme, including in particular aortic endothelial cells, perivascular smooth muscle cells & a few hematopoietic cluster cells in E10.5 AGM.

Single-cell transcriptomic analyses of enriched cells with hemogenic potential and undergoing EHT (e.g. identified by 23GFP reporter) have offered a detailed insight into the dynamic & continuous gene expression changes, identification of

[2]TOM = tdTomato fluorescent protein

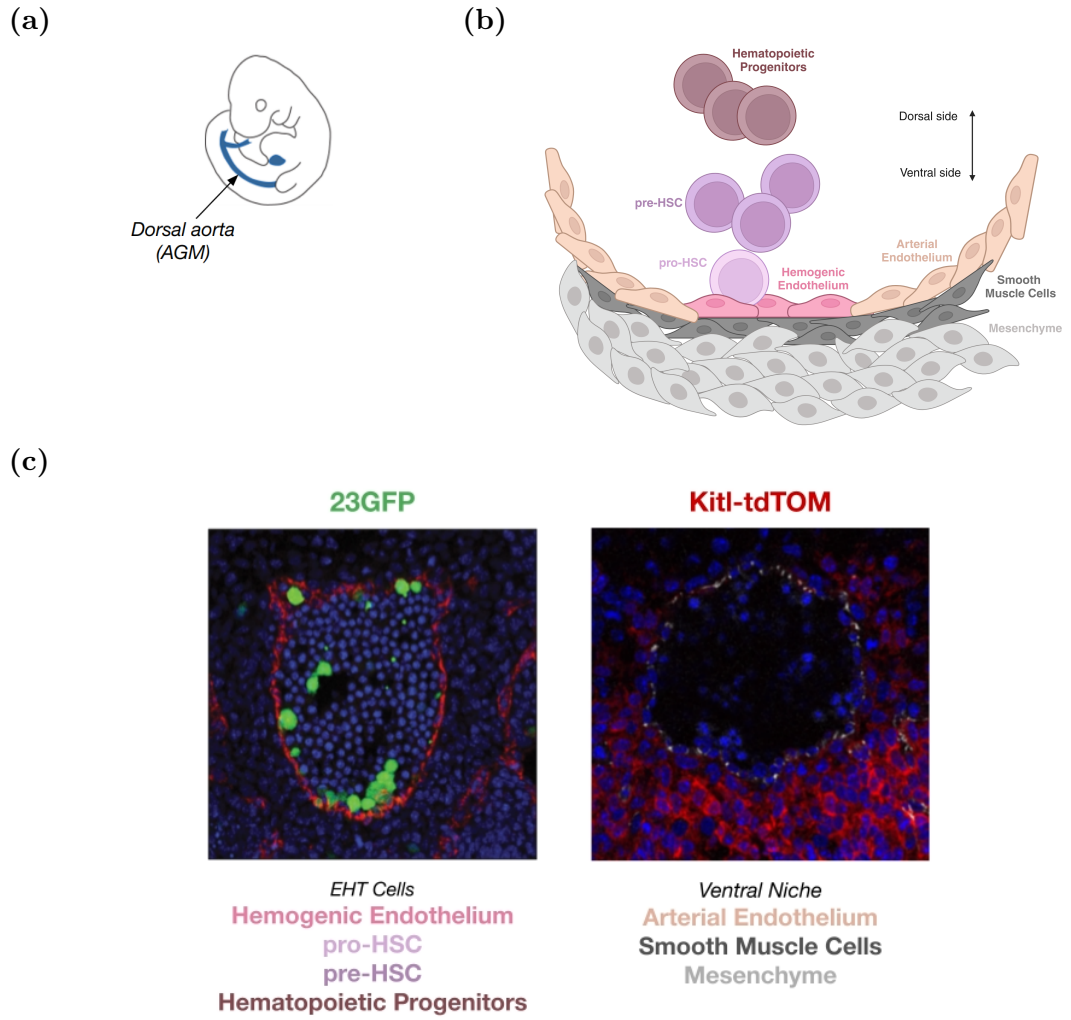


Figure 6.1: (a) Location of the dorsal aorta in the mouse embryo; (b) Schematic of the cross-section of ventral side of the dorsal aorta, where HSPCs predominantly emerge. Created with BioRender.com; (c) Microscopy images of cross-sections of the E10 dorsal aorta showing fluorescent expression of 23GFP and KitlTOM reporter lines, including the cell type populations they label, shown below. *Courtesy of de Bruijn Group, MRC WIMM, University of Oxford.*

key hematopoietic transcription factors and composition of IAHCs (e.g. Swiers et al., 2013a; Baron et al., 2018; Zhu et al., 2020; Hou et al., 2020). Both the EHT trajectory and individual cell types within the surrounding niche can be studied and characterized in detail from single-cell data, however the essential missing information is how they are physiologically positioned and interacting with each other. To this end, we will generate a novel dataset to tackle this problem.

6.3 Single-Cell and Clump Sequencing via Smart-seq3

As detailed in the previous section, the niche surrounding the arterial endothelium plays a key role in regulating EHT. We motivate that Clump-Seq could provide essential information about the interaction of HE with the surrounding cells in the aortic wall, which would not be available otherwise from single-cell data. While the composition and dynamic changes occurring in the niche directly attached to the HE are poorly understood, the EHT trajectory has been studied in detail.

Here, we set out to generate a novel Clump-Seq dataset in order to study the interactions between HE and the surrounding microenvironment. The objective is to obtain clumps made up of one HE cell attached to a neighboring cell. Using the clump deconvolution model developed in the previous chapters, our goal was to reconstruct the EHT trajectory and simultaneously describe the behavior of neighbors throughout development. The advantage provided by ClumpSeq data is being able to specifically characterize niche cells that are in direct contact with HE, as these are the primary cells driving the interactions. We emphasize this data generation framework provides key insights into the profile of neighbor cells that would otherwise not be available from single-cell data, as the latter involves tissue dissociation and therefore loss of information about where the microenvironment cells were located in reference to HE.

6.3.1 Experimental Methods

The first step of this study was generating a reference single-cell dataset comprising both (1) cells in different stages of EHT and (2) cell populations from the surrounding niche. The study focused on the E9.5 embryo, as at this stage EHT is just starting and we were interested in capturing signals triggering EHT initiation.

A 23GFP mouse reporter line was used for identifying the cells spanning the EHT trajectory, as the 23GFP labels all cells with hemogenic and hematopoietic potential, including HE, thus capturing population (1). To further enrich for more

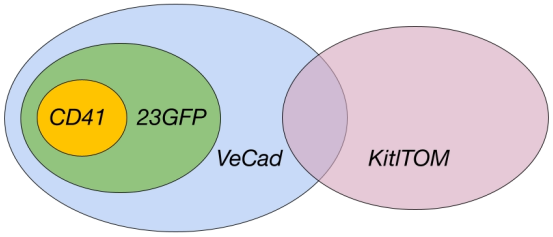


Figure 6.2: Venn diagram of cell type populations labelled by markers used in our single-cell study.

Cell Type	Phenotype
Endothelium	VeCad$^+$ or KitlTOM$^+$
HE	VeCad$^+$23GFP$^+$
pro-HSC, pre-HSC, HPC	VeCad$^+$23GFP$^+$CD41$^+$
Ventral Mesenchyme	Kitl$^+$

Table 6.1: Phenotype of each cell type population based on markers used for gating cells in the Smart-seq3 study.

specialized populations we also used the CD41 marker; at this developmental stage, this labels pro-HSCs and hematopoietic progenitor cells (HPC).

The niche surrounding the hemogenic endothelium, i.e. population (2), is broadly made up of the ventral mesenchyme and other arterial endothelial cells that might not further differentiate or have no hemogenic potential (Fig. 6.1b). In order to capture the latter population (which would not be identified by 23GFP) we sorted cells based on the VeCad endothelial marker. The ventral mesenchyme can be picked up via expression of *Kit* ligand (*Kitl*), which labels the mesenchymal niche predominantly in the ventral side of the dorsal aorta. To this end, we used a KitlTOM mouse reporter line.

To show how the transgenic reporter lines label the relevant cells *in vivo*, Figure 6.1c exhibits fluorescent confocal images from cross-sections of the dorsal aorta in E10 *Runx1* 23GFP and *Kitl*TOM transgenic embryos, along with a list of cell populations labelled by their respective fluorescent proteins.

A diagram of all the markers used in the single-cell study and the populations they label is shown in Figure 6.2 and Table 6.1. In particular, for the endothelial cells and HE we can use a sequential combination of markers to increase the level of enrichment. VeCad captures the broadest population (since it is a conventional endothelial marker); a subset of this population will be labelled by 23GFP, enriching for cells with hemogenic potential. A futher subset of this VeCad$^+$23GFP$^+$ population is labelled by CD41, which consists of cells that are transcriptionally and phenotypically further along the hemogenic development process, starting from pro-HSCs.

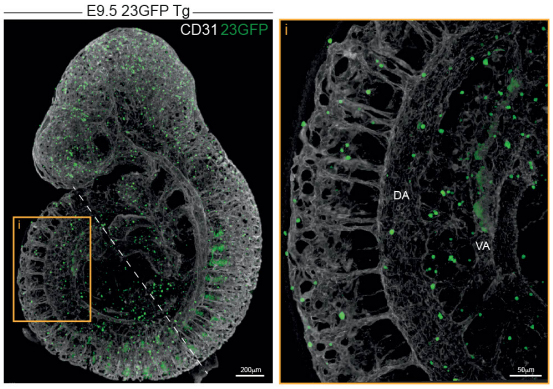


Figure 6.3: Whole mount staining of a E9.5 embryo (left), showing an approximate boundary where the embryo is cut to isolate the caudal parts (below the dotted line) and a zoomed in section (right) showing a lateral view of the dorsal aorta (DA) and vitelline arteries (VA). *Courtesy of de Bruijn Group, MRC WIMM, University of Oxford.*

The caudal parts - portion of the embryo proper cut underneath the liver primordium and heart - of E9.5 embryos were dissected [3]. Figure 6.3 shows a whole-mount microscopy image of a E9.5 embryo, providing a lateral view of the dorsal aorta; caudal parts dissected in the experiment are located below the dotted line. After establishing the gating strategy, the next step was sorting and preparing cell libraries for sequencing. In order to obtain a high resolution and comprehensive scRNA-seq dataset, we set out to perform sequencing with Smart-seq3 (Hagemann-Jensen et al., 2020). Since Smart-seq3 is a plate-based method, cells were index sorted into 384-well plates across multiple experiments. A total of 5 plates were obtained: 4 plates of sorted hemato-endothelial populations, denoted A-D, and one plate of sorted niche cells, denoted M. Each of the A-D plates contained cells sorted via a different gating strategy, and plate M cells were sorted from a KitlTOM mouse. All 5 single-cell plates and their associated gates are shown in Table 6.2.

The purpose of the scRNA-seq experiment was to compile a single-cell reference dataset comprising both cell populations that are involved in EHT and are part of the surrounding niche, respectively. We now arrive at the culmination of this chapter: generating a novel ClumpSeq dataset from the embryo. Our overarching objective is

[3] "caudal parts" are dissected in E9.5 embryos as the AGM is not fully developed at this stage yet.

Plate	*Gate*	*Size*
A	**23GFP$^+$**	56
B	**CD41$^+$**	131
C	**VeCad$^+$**	209
D	**23GFP$^+$ CD41$^+$**	139
M	**KitlTOM$^+$**	345

Table 6.2: Composition of the 5 plates prepared for the Smart-seq3 experiment, showing for each plate the gate and number of single cells retained for analysis after quality control.

to study the cell type composition and transcriptional profile of the microenvironment population directly attached to cells committed to a hemogenic fate. Therefore, the aim is to capture 2-cell clumps containing one cell from either population. In practice this can be achieved by enriching for clumps with 23GFP expression.

All experimental work was designed and performed by Dr. Giorgio Anselmi, de Bruijn Group, MRC WIMM. Smart-seq3 library preparation was performed by Elitza Deltcheva and John Brown, Enver Group, UCL Cancer Institute. The clump isolation procedure was optimized for mouse embryonic tissues starting from the Clump-Seq protocol described in (Manco et al., 2021), and the consistent isolation of 2-cell clumps was further validated by imaging flow cytometry. Section 6.3.2 describes the clump experiments in more detail.

All clumps were sorted on VeCad$^+$23GFP$^+$. Analogously to the single-cell experiment, sorted clumps were prepared into 384-well plates for sequencing with Smart-seq3. Two clump plates were sorted from a different embryonic day each: E9.5 and E11, denoted by C_1 and C_2.

6.3.2 Experimental Details of Clump Generation

This section will present an outline of the experimental steps and challenges involved in the clump data generation. These can be divided into the following categories: establishing a gating strategy to enrich for 2-cell clumps; sorting for clumps with heterogeneous composition of cell types; and minimizing re-aggregation events. A summary of these steps is shown in Figure C.1.

First step was setting up a gating strategy to separate clumps from singlets. In a calibration experiment, tissue from the caudal parts of E9.5 embryos was dissociated using 0.06% Collagenase, a lower concentration than traditionally used for generating single-cell suspensions, in order to obtain lighter dissociation. Flow cytometer CytPix (Attune) was used for capturing images of the individual clumps and confirming they are made up of two attached cells. Dead cells were removed and live cells were stained with DRAQ5 in order to separate clumps from the singlets. Fluorescence Activated Cell Sorting (FACS) plots of area against height based on DRAQ5 staining exhibited a clear separation of clumps from the singlets.

The next step was ensuring clumps capture cell types from the niche. In the context of studying potential interactions between HE and the surrounding niche, clumps were sorted from a E9.5 *Runx1* 23GFP transgenic mouse embryo. The objective was for each clump to contain at least one HE cell, hence the clumps were gated on $23GFP^{+}CD31^{+}$ expression. ImageStream (Luminex) analysis of fluorescent marker expression showed a heterogeneous composition of clumps.

Finally, the third step was concerned with re-aggregation events; after the dissociation protocol, some singlets in suspension might become attached to other singlets and thus appear as 2 clumped cells in contact. In order to quantify the extent of this spurious coupling, clumps were sorted from both ubiquitous *GFP* and *TOM* reporter mice (i.e. constitutively expressing the fluorescent reporter gene in all cells). The sorted clumps of homogeneous paired expression *GFP-GFP* and *TOM-TOM* will be a mixture of real clumps of two attached cells and spurious re-aggregates, while clumps expressing both *GFP* and *TOM* will certainly be spurious as they come from two different embryos. By measuring the quantity of these latter clumps we estimated the proportion of re-aggregates to be around 11%. Section C.2 explains our estimation procedure in more detail.

6.3.3 Single-Cell Analysis

Figures 6.4a and 6.4b show the distribution of the library sizes (number of UMIs) and number of expressed genes per cell after quality control filtering, for all 5 plates.

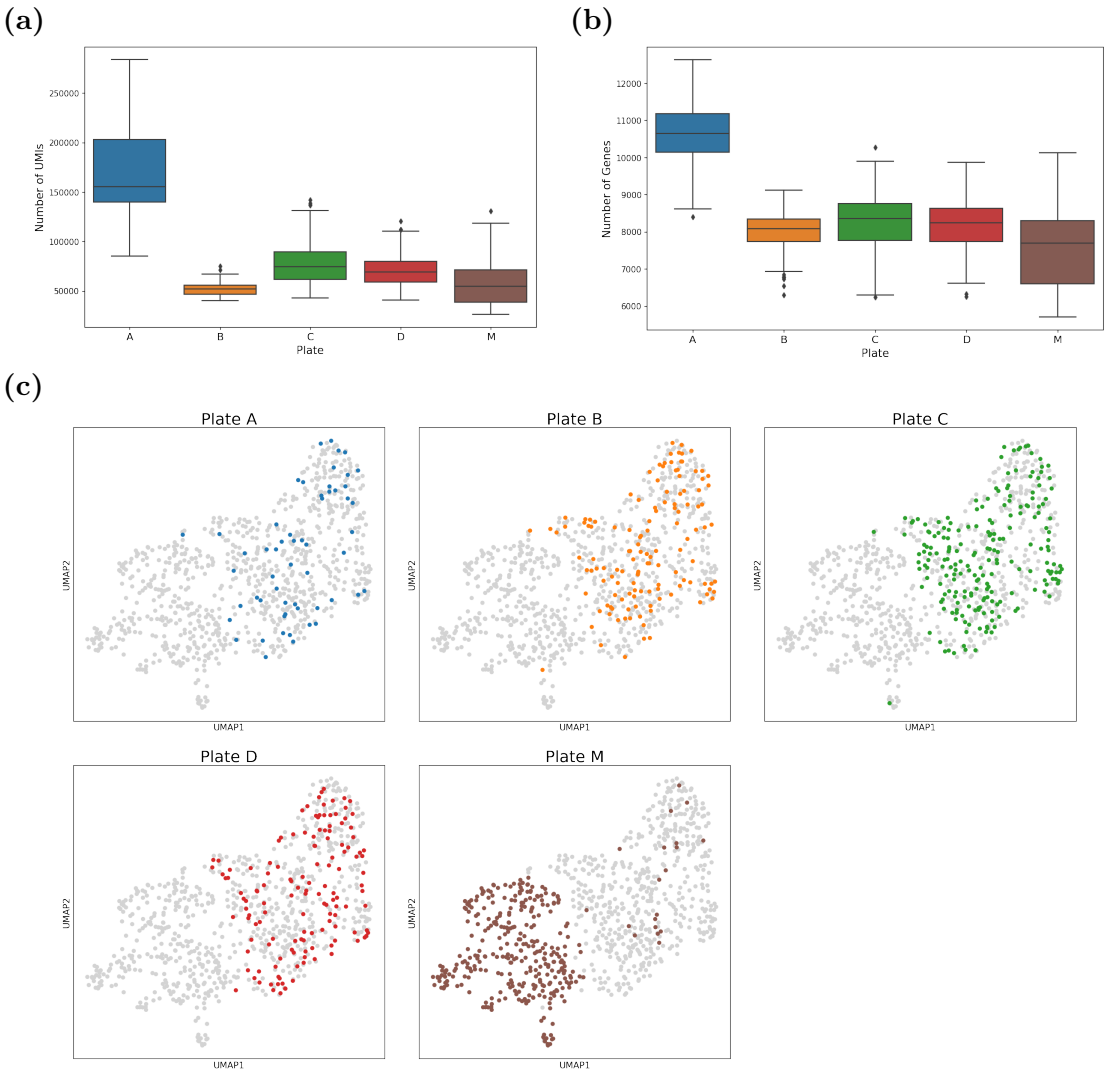


Figure 6.4: Analysis of Smart-seq3 single-cell data by Plate. (a) Distribution of the number of UMIs; (b) Distribution of the number of genes; (c) UMAP computed on the latent space highlighting each Plate.

Plate A in particular shows larger number of UMIs and genes per cell than the other plates. In addition, the plates were sequenced across distinct experiments and runs, thus part of the technical variation in the data could stem from batch effects by plate.

To counteract potential batch effects, we trained a conditional VAE model with a categorical covariate given by plate. Cells from plates A-D are expected to overlap since they are all made up of endothelial and EHT cells, while cells from plate M were sorted from the niche and therefore should only (at most) partially mix with the others. Fortunately, the latent space exhibits this expected behaviour, thus proving

a good level of batch correction without over-correcting. Fig. 6.4c shows the UMAP computed on the resulting latent space, highlighting each plate.

Figure 6.5a shows the UMAP of Leiden clusters. The split between endothelial and mesenchymal populations is evidenced by the disjoint expression of cell type specific markers VE-Cadherin (*Cdh5*) for the former and Pdgfr-α receptor (*Pdgfra*) for the latter. Clusters 5,6,7 are characterized by expression of venous markers *Aplnr, Nr2f2, Nrp2*. These cells will be labelled as Vascular Endothelial Cells - we emphasize that while the aforementioned genes are known to be venous markers, we refrain from labelling these cells as venous as the default programme of endothelial development is more "venous-like" and therefore expression of the aforementioned genes is likely to indicate early endothelial progenitors rather than venous endothelial cells (Trimm & Red-Horse, 2022). Clusters 8 & 9 express arterial genes *Gja4, Unc5b, Dll4*, with cluster 9 cells exhibiting varying levels of transcription factors *Mecom, Meis1* and *Cd44*, all known to be expressed in HE. Furthermore, only a handful of cells from cluster 9 express *Runx1*, pointing out to the beginning phase of a hematopoietic transition. Therefore, we further subclustered this population into Arterial Endothelium (high expression of *Gja5* but low levels of *Meis1* and *Mllt3)* and HE (high expression of *Meis1* and *Mllt3*, and a subset of cells expressing *Runx1*). Figure 6.6 shows the UMAP colored by raw proportions of key genes and transcription factors.

Our findings on clusters and differentially expressed genes share similarities with a single-cell sequencing study of vascular development (Hou et al., 2022). Authors sorted populations of vascular endothelial cells (VEC) from multiple parts of mouse embryos (yolk sac, limb, head, body, AGM) at time points ranging from E8 to E11. Their single cell analysis of embryo proper VEC (comprising all sites except yolk sac) revealed the transcriptional landscape of vasculogenesis, comprising a continuum of differentiation from venous VEC to arterial vasculature. In particular, they identified two populations of arterial endothelial cells, defined by their location: major artery (including the dorsal aorta) and arterial plexus. In particular, cluster 8 from our analysis shows similar differential gene expression to what are identified in (Hou et al.,

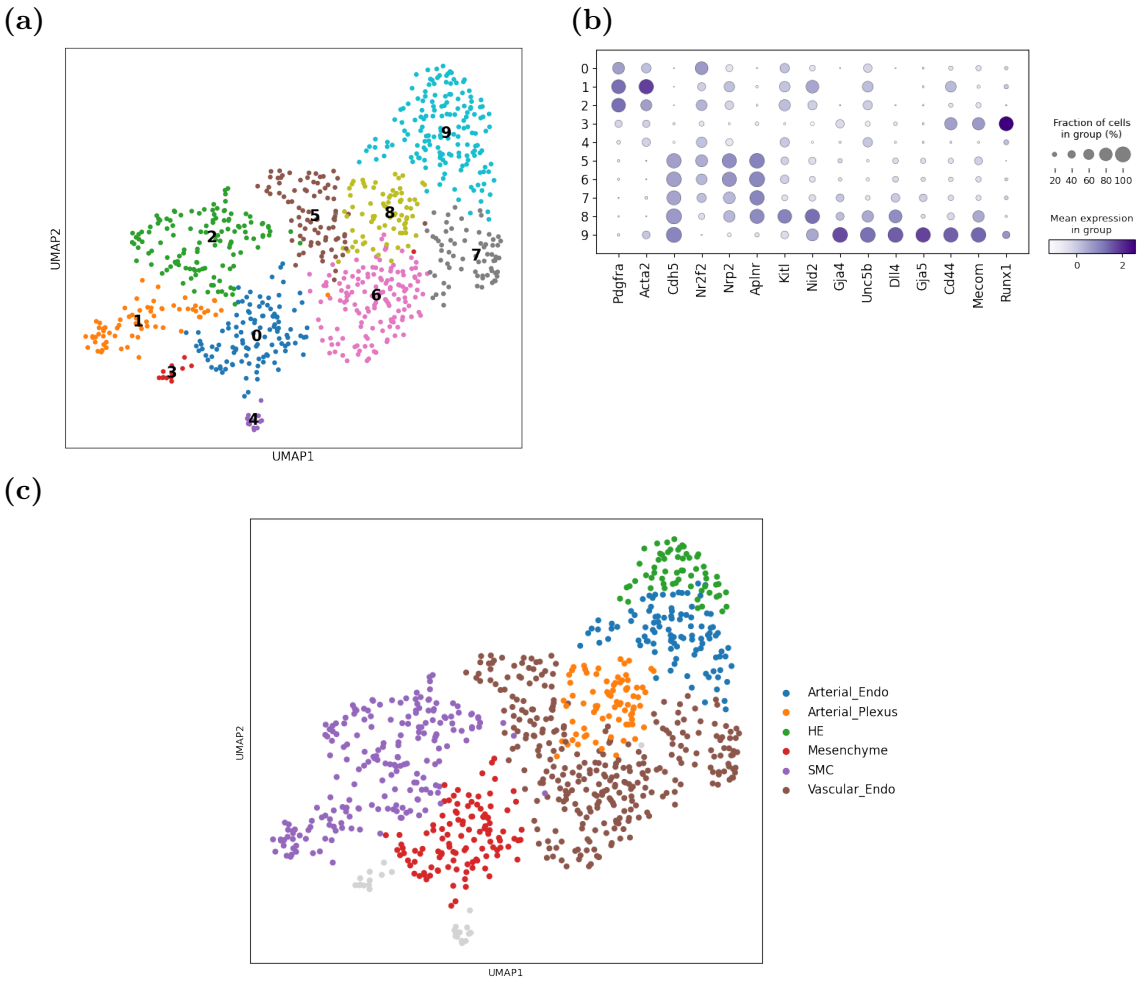


Figure 6.5: Clustering of Smart-seq3 single-cell data. (a) Cluster labels; (b) Dot plot showing expression of key markers across clusters; (c) Cell type annotations. Greyed out cells are contaminants.

2022) as arterial plexus cells. They specifically define the arterial plexus as "small vessels widely distributed throughout the whole embryo, including the perineural vascular plexus, vascular sprouts within the neural tube, intersegmental vessels at the body trunk, and capillaries in the limb buds". The authors characterized the arterial plexus as exhibiting venous expression but also expression of arterial markers such as *Gja4, Dll4* and *Unc5b*, however no major artery genes such as *Gja5* and *Cd44*. Other genes with elevated expression in these cells include *Kitl, Vwa1* and *Nid2*. The authors described this population as migrating cells moving from the venous plexus to the major arteries, thus supporting arterial growth during development. We find a similar pattern of expression in cluster 8, showing levels of arterial markers



Figure 6.6: UMAP of latent space of the single-cell VAE model showing normalized raw proportions of various genes and transcription factors.

but lacking expression of *Gja5* and *Cd44*, which are only expressed in cluster 9.

As expected, most cells sorted in plate M are mesenchymal, as evidenced by the levels of *Pdgfra* in clusters 0, 1, 2. Clusters 1 & 2 show expression of α-smooth muscle actin (*Acta2*), specific to smooth muscle cells. Therefore, these 2 clusters will be annotated with this cell type and 0 remains labelled as mesenchymal. Finally, there are two small clusters 3 & 4, most likely unwanted contaminants, in the sense of not being labelled by KitlTOM. Cluster 3 expresses *Runx1* and the hematopoietic marker *Hlf*, thus it is most likely made up of hematopoietic progenitors.

Now that the dataset is annotated, the next step is running a conditional VAE model with cell type labels, whose embedding will be used in the clump deconvolution model. The 2 contaminant clusters from the previous paragraph were not included in the final annotated dataset as they consist of only a handful of samples.

6.3.4 Improving Mixing via Adversarial Classifiers

As mentioned in previous chapters, the motivation behind conditioning the single-cell VAE model on cell type labels is for the decoder to learn the cell type specific variation. Given samples from the latent space, the decoder is able to reconstruct cell type specific gene profiles based on the additional label input. This framework allows formulating the parameters of the clump generative distribution. As a result, the cell type specific signals are removed from the latent space, which will now capture signals of continuous variation within each type.

A potential model misbehaviour is when latents still have a tendency to separate by label. In this case the latent space still contains signals that separate cell types and this can be problematic in the decoding stage. To see this, we can consider a simple two-label latent space, c_1 and c_2, with mapping $\Phi(z) \in \{c_1, c_2\}$ denoting the cell type of z. If the sets $\mathcal{I}_{c_i} = \{z' : \Phi(z') = c_i\}$ are separating and $z \in \mathcal{I}_{c_1}$, then $f(z|c_1)$ will yield a gene profile specific to c_1, yet $f(z|c_2)$ is likely to yield a profile specific to c_1 in spite of the label information passed to the decoder. This happens because the similarity of the decoded gene vector to either cell type is largely determined by the location of z rather than the label.

There have been methods developed for encouraging conditional VAEs to learn representations that overlap across labels, e.g. trVAE (Lotfollahi et al., 2020) and CoMP (Foster et al., 2022), which rely on adding a penalty to the loss function. trVAE uses a Maximum Mean Discrepancy kernel, which can be cumbersome to tune; CoMP formulates a penalty based on the variational posterior, however it requires fixing the variational scale. In line with our objective of retaining the structure of our VAE model and training efficiently, we opted for an adversarial scheme to encourage mixing, which has been used in single-cell studies (Cao & Gao, 2022; Lotfollahi et al., 2023). We add a classifier to the VAE model, consisting of an MLP with a softmax output layer, that takes as input latents and predicts their cell type label. This classifier can be trained in an adversarial manner via the following two steps: (1) Training the classifier to predict the labels (the reward is the classification loss, e.g. cross-entropy loss) and (2) training the encoder to "fool" the classifier by generating latents with stronger mixing (the reward is proportional to how often the classifier predicts the wrong labels). The strength of classifier training can be varied by regularizing the classifier loss in step (2); this regularization term will be denoted by γ.

This subsection will look at the application of the adversarial classifier to the conditional VAE model trained on the Smart-seq3 data, using the annotated cell type labels from the previous subsection. Mixing can be assessed via the silhouette width (Rousseeuw, 1987); for each label c we have:

$$s_c = \frac{1}{|\mathcal{I}_c|} \sum_{i \in \mathcal{I}_c} \frac{b_i - a_i}{\max\{a_i, b_i\}} \tag{6.1}$$

where a_i and b_i are the mean Euclidean distances between latents $\boldsymbol{z}_i$ and the sets of all other latents of the same label and of different labels, respectively, computed in a k-nearest neighbor graph. The silhouette width measures the level of clustering between labels, by measuring distance-wise similarity between latents of the same label as cell i and latents of all other labels. In a well-mixed latent space the values of s_c should be close to 0.

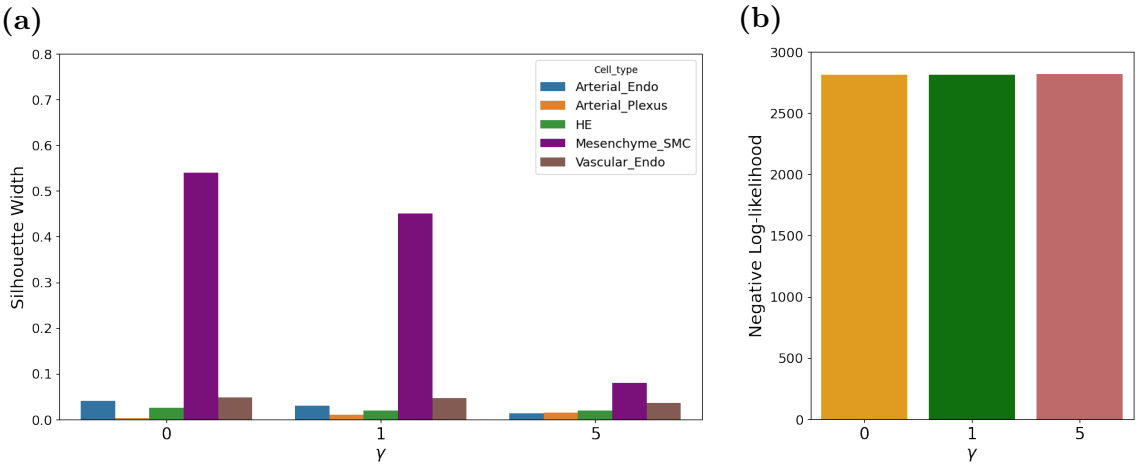


Figure 6.7: Training with an adversarial classifier to increase mixing of latents, with different values of classifier training strength γ. (a) Absolute values of silhouette width per cell type group; values closer to 0 are better. (b) Negative log-likelihood values for cells in a validation set.

When running the conditional VAE on the Smart-seq3 data, cells from plate M (mostly mesenchymal and SMC) do not usually mix with the ones from the other plates. Fig. 6.7a shows the silhouette width for values of γ, computed for cells from each cell type group. To highlight how plate M does not mix with the others, we grouped Mesenchyme and SMC cells into one category (`Mesenchyme_SMC`; since both cell types are found in plate M). Incorporating the adversarial classifier brings down the silhouette width significantly with a negligible change in reconstruction error (Fig. 6.7b).

Given the conditional VAE model where latents mix across the cell type labels, the upcoming section will look at how the clump deconvolution performs on the novel clump dataset. There are two clump plates, C_1 and C_2, collected at E9.5 and E11. The following section will examine the deconvolution of clumps in plate C_1 using the reference Smart-seq3 dataset collected at the same embryonic day.

6.4 Clump Analysis

During evaluation of the clump model performance, we noticed some genes were poorly reconstructed when comparing the predicted model clump proportions with

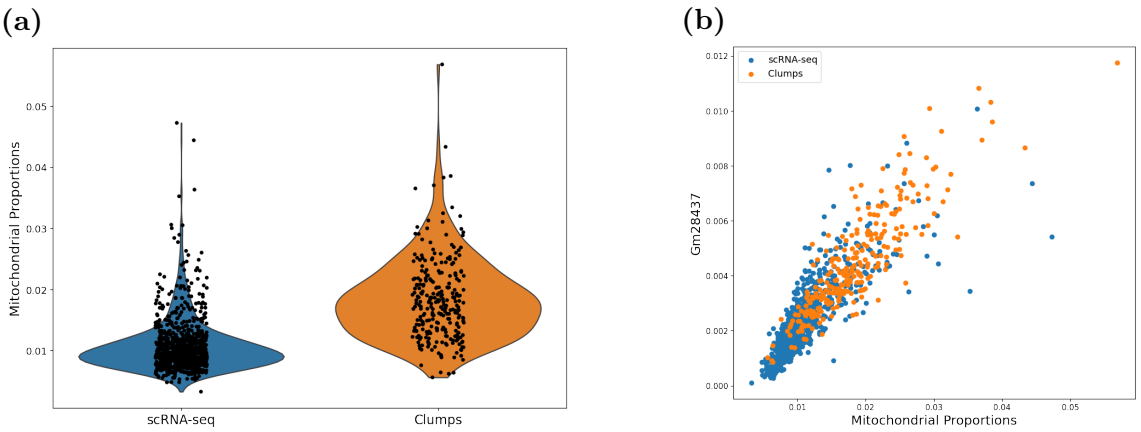


Figure 6.8: Mitochondrial proportions, defined as the sum of proportions of all mitochondrial genes, for Smart-seq3 single-cell and clump data. (a) Boxplot; (b) Scatter plot with proportions of *Gm28437*.

the raw data. Specifically, these genes were always underestimated by the model. We subsequently looked at the distribution of raw proportions of these genes in the clump and single cells and found that some of them exhibit higher raw proportions in clumps than in single cells. For instance, mitochondrial content is shown in Figure 6.8a as a violin plot of total mitochondrial expression, calculated as the sum of proportions of all mitochondrial genes. We note that all the clumps and single cells shown in this figure have passed quality control filtering on mitochondrial content based on conventional thresholds (Luecken & Theis, 2019). Therefore, it is unlikely these samples have low quality. Nevertheless, the discrepancy between mitochondrial quantities in clumps and single cells poses problems for the clump model, since it is formulated on the assumption that for each gene the mean clump gene expression is a sum of cell type specific gene proportions. If all the cell types exhibit a lower range of mitochondrial proportions than the clumps, then the model will struggle to reconstruct these genes, explaining the observed behaviour.

A possible reason for the mismatch in gene expression values in Fig. 6.8a are technnical variation differences between samples obtained in the single-cell and clump experiments. Since these datasets were collected and sequenced in different experiments, with different mice, it is likely this mismatch is part of a batch effect. It is also possible there are genes whose expression is much higher or lower than what

would be predicted by the summation model due to potential cell interactions or effects that would otherwise not be seen in dissociated samples, i.e. single-cell data. For example, (van den Brink et al., 2017) found a subpopulation of muscle stem cells that express particular genes in the scRNA-seq dataset but not *in vivo*; they concluded that the dissociating procedure induces transcriptome changes within this subpopulation . Other genes appear to be strongly correlated with mitochondrial content, for example non-coding gene *Gm28437* (Figure 6.8b); hence the model will struggle with reconstructing this gene as well. In conclusion, it is difficult or prohibitive to identify all these genes *a priori*. While metrics such as clump model reconstruction error per gene could be used to select them, retaining them hinders the deconvolution ability of the model and removing them might throw out potential biologically relevant signals due to cells interacting or communicating with each other.

We adjusted the clump deconvolution model to incorporate signals not captured in the reference single-cell data. For a clump of library size N, let $\xi \in [0, 1]$ be the proportion of transcripts that can be reconstructed by a sum of cell type profiles decoded from the reference data embedding. The remaining $N(1 - \xi)$ transcripts represent additional variation that will be learned during model training. Writing the mean parameter of the Negative Binomial distribution in terms of gene proportions leads to the following formulation:

$$y_g \sim \text{NegativeBinomial}\left(N\left(\xi \sum_{c=1}^{\mathcal{C}} \alpha_c \; f(z_c|c)_g + (1-\xi)\varepsilon_g\right), \theta_g\right) \tag{6.2}$$

where $\boldsymbol{\varepsilon}$ is the gene proportion vector of clump-specific signal, summing up to 1. Initial implementations had this term formulated as a gene specific parameter, however we found the clump reconstruction was still poor. To allow more flexibility, this term is amortized and put through an information bottleneck by introducing an auxiliary low dimensional variable $\boldsymbol{u}$ encoded from $\boldsymbol{y}$ and decoded into $\boldsymbol{\varepsilon}$. The mappings $\boldsymbol{y} \to \boldsymbol{u}$ and $\boldsymbol{u} \to \boldsymbol{\varepsilon}$ are one MLP each, with an output softmax activation for the latter one.

Introducing an amortization scheme in the clump generative distribution is harmful for the inference of cell type reference profiles, as ξ will be shifted towards 0 and the model trivially learns an embedding of the data via $\boldsymbol{u}$. To counteract this effect, we set out to strengthen the information bottleneck between $\boldsymbol{y}$ and $\boldsymbol{\varepsilon}$ by setting the dimension of $\boldsymbol{u}$ to 1.

This section will look at the application of the adjusted parameters in (6.2) to clump plate C_1 and show that it provides more flexibility in fitting the aforementioned genes without trading off important signals. We first characterize the poor reconstruction of the initial model more robustly by querying the posterior predictive distribution (Section 2.3). We sample from the variational posterior multiple times and compute the expected gene proportions via Monte Carlo estimates - see (2.39). Figure 6.9a shows a scatterplot between mean raw gene proportions in clumps and estimated posterior predictive means, with a few annotated genes (with highest mean raw expression). The genes with highest mean expression are also the ones showing the poorest reconstruction.

Figure 6.9b shows the distribution of ξ for the C_1 clump dataset (265 samples). Repeating the estimation of the posterior predictive mean for the adjusted model results in a better fit (Figure 6.9c), especially for genes with high mean expression that previously showed poor reconstruction. To see how much weight vector $\boldsymbol{\varepsilon}$ places on each gene, we sorted its mean gene expression across clumps (Fig. 6.9d). Incidentally, the genes with highest mean values of ε_g are the ones with biggest reconstruction improvement in the adjusted model. We can also confirm the better fit by querying the negative log-likelihood between the two models (Figure 6.9e). Furthermore, looking at this difference in log-likelihood values between the two models at gene level (Figure 6.9f) shows that most genes exhibit a better fit. We note the largest changes in log-likelihood occur in only a handful of genes, and these genes correspond to the ones with highest expression in $\boldsymbol{\varepsilon}$ in Figure 6.9d. These findings confirm that $\boldsymbol{\varepsilon}$ picks out the poorly reconstructed genes and fixes the expression in the likelihood without significantly changing the fit of other genes; therefore, it works as intended.

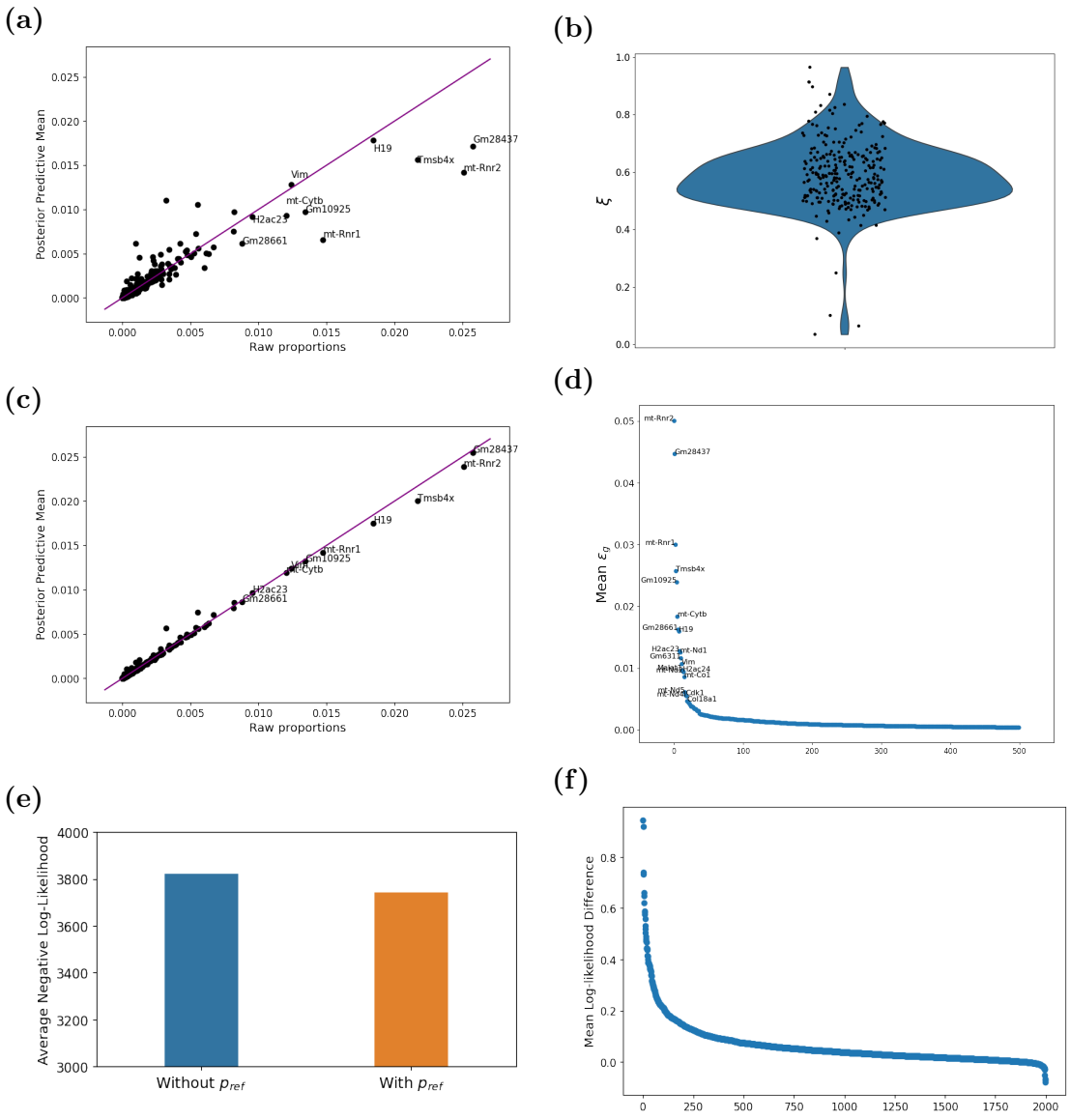


Figure 6.9: Adjusting the parameters of the clump generative distribution to incorporate extra variation. (a) Estimated posterior predictive means versus mean raw expression for the clump model without adjustment, showing selected annotated genes; (b) Values of ξ (see (6.2)) for clump plate C_1; (c) Estimated posterior predictive means versus mean raw expression for the clump model *with* adjustment; (d) Mean gene-specific values of $\boldsymbol{\epsilon}$ sorted in decreasing order (truncated to 500 genes); (e) Mean negative log-likelihood between the two models, with and without adjustment, evaluated on a held-out set (lower is better); (f) Mean difference in log-likelihood between the models (not adjusted versus adjusted) at gene level, sorted by magnitude in mean difference.

Now that the model is better calibrated to reconstruct genes, the following subsection will look at its application on the E9.5 clump dataset.

6.4.1 Composition of E9.5 Clumps

After quality control filtering, the clump dataset from plate C_1 is made up of 265 samples. The results of the clump deconvolution model applied on this dataset reveals pairs of multiple cell types, including of doublets of the same cell type (α_c close to 1), summarized in Figure 6.10. As shown in the figure, most heterogeneous pairs are made up of Arterial Plexus + Vascular Endothelium and HE + Arterial Endothelium.

While most clumps had values of $\boldsymbol{\alpha}$ that allowed a straightforward inference of a 2-cell type composition, there were clumps where the composition was more heterogeneous (shown as "Mixed" in Figure 6.10). Besides the possibility of those clumps being made up of more than 2 cell types, a technical reason could be an overlap between the reconstructed cell type profiles due to similarities in gene signatures. For instance, a clump might contain 60% transcripts from an HE cell, but the model might output $\boldsymbol{\alpha}$ as 40% HE and 20% Arterial Endothelium, given the many gene expression similarities between the two cell types.

The inferred cell type composition of E9.5 clumps agrees with the tissue physiology. The abundance of Arterial Plexus + Vascular Endothelium clumps is expected since the former population most likely represents cells migrating from the venous plexus towards the major arteries, thus they are likely to be in contact with vascular endothelium. The abundance of HE + Arterial Endothelium pairs is also expected, since these two cell types are embedded in the aortic wall.

One of the advantages of the generative model framework is being able to reconstruct cells by sampling the latent variables. In particular, for each clump the model can reconstruct an "artificial" cell of type c by sampling z_c from the encoder and query the decoder $f(\cdot|c)$ to obtain a gene expression profile. Thus, for instance, for a clump of HE + Arterial Endothelium we can reconstruct two single-cell profiles, each corresponding to a cell type. To see this, the gene expression profiles retrieved from clumps were put together with the decoded proportions from the single-cell

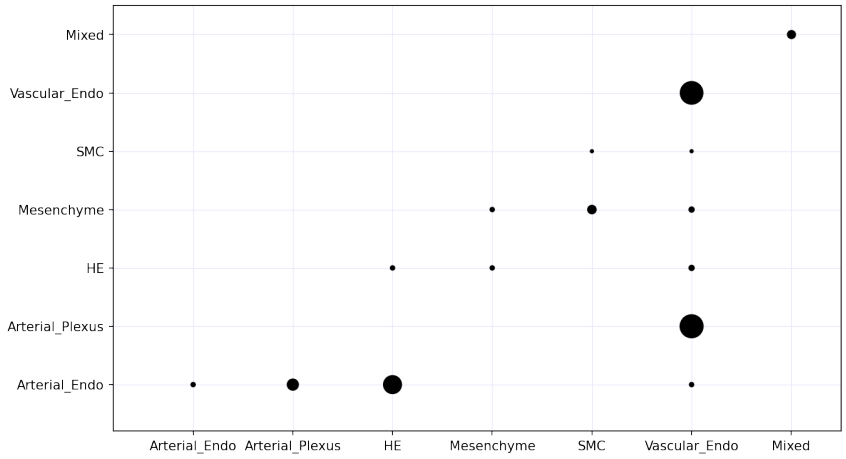


Figure 6.10: Distribution of cell type pairs inferred in E9.5 clumps (note the lower triangular matrix form). "Mixed" refers to clumps where it was not clear or possible to distinguish exactly 2 cell types in the composition.

model (as both models use the same decoder); Figure 6.11 shows a UMAP run on the gene expression space. As seen in the figure, each cell type retrieved from clumps mixes with its single-cell counterpart. Figure 6.11b shows the all the relevant profiles identified in each clump, as per Figure 6.10, without pairing information. For the 2 most abundant heterogeneous pairings (Arterial Plexus + Vascular Endothelium and Arterial Endothelium + HE), Figures 6.11c and 6.11d highlight the relevant *in silico* cells in the UMAP and the mapping between them.

For any cell type, it is therefore possible to query the gene profiles of both the single cells and *in silico* cells retrieved from the relevant clumps where that cell type is present. Clump data provides us with the ability to subset these cells based on which cell types they are attached to. This leads to the question of whether we can find gene signatures that strongly correlate with the neighborhood composition.

6.4.2 Differential Gene Expression in Clumps

For a given cell population, we are interested in identifying genes whose expression depends on the cell type those cells are attached to. For instance, are there gene programmes specifically upregulated in HE cells attached to Arterial Endothelium?

Given the generative nature of the model, we can formulate a differential gene expression analysis framework comparing the retrieved cell type profiles with the

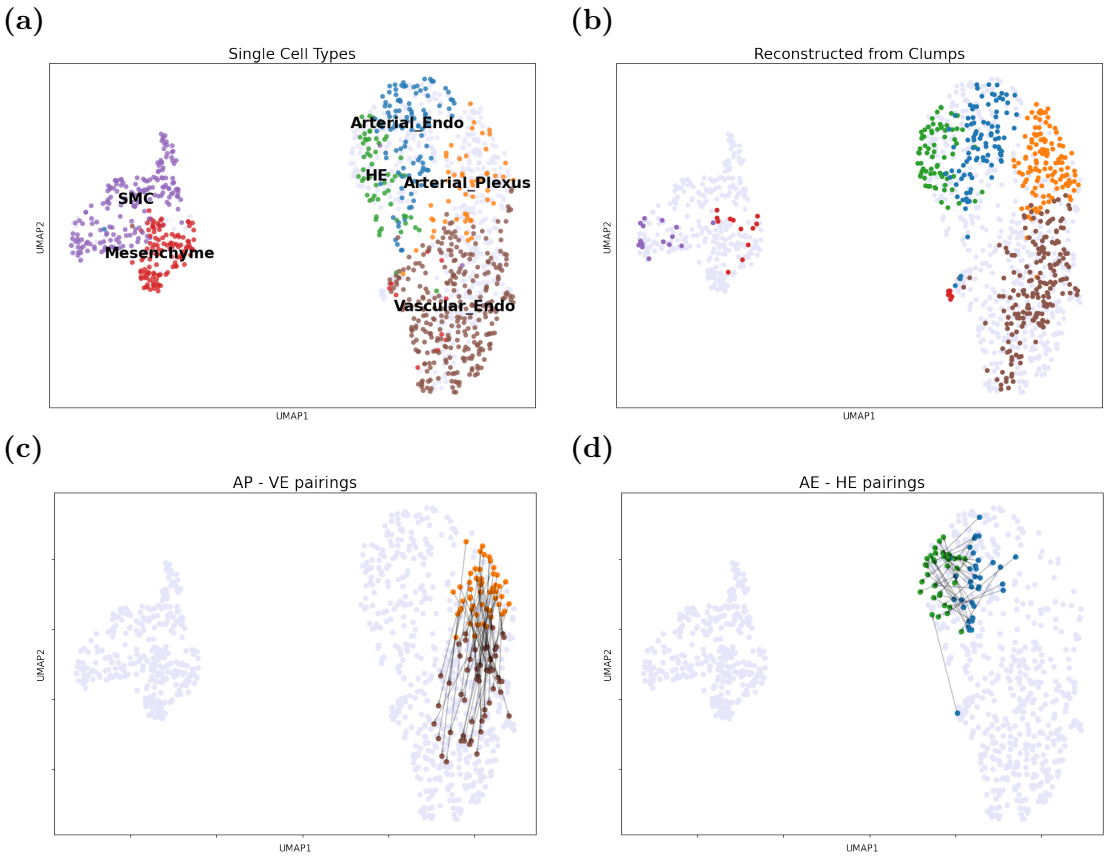


Figure 6.11: UMAP computed on a joint set of decoded proportions of both single cells and cell types retrieved from clumps, highlighting each (a, b). Subsets of the latter population originate from pairs of (c) Arterial Plexus + Vascular Endothelium and (d) Arterial Endothelium + HE.

single-cell ones, within a cell type. Both the clump and single-cell models use the same decoder, which returns the normalized gene proportions. Following the approach in Section 4.4.2, we can simulate multiple *in silico* samples for both sets of cells by sampling the latents $z_c^{(1)}$ & $z_c^{(2)}$, and subsequently decoding them to obtain denoised samples from the Gamma distributions. LFC is computed similarly as in (2.40).

Samples are compared via a two-sample Kolmogorov-Smirnov test, and p-values are corrected using the false-discovery rate. Genes are considered differentially expressed when the corrected p-values are smaller than 0.05 and the LFC is above a certain threshold; both this threshold and ϵ in the LFC were computed as in (Boyeau et al., 2023).

For instance, we can consider the HE - Arterial Endothelium clumps (Fig. 6.11d)

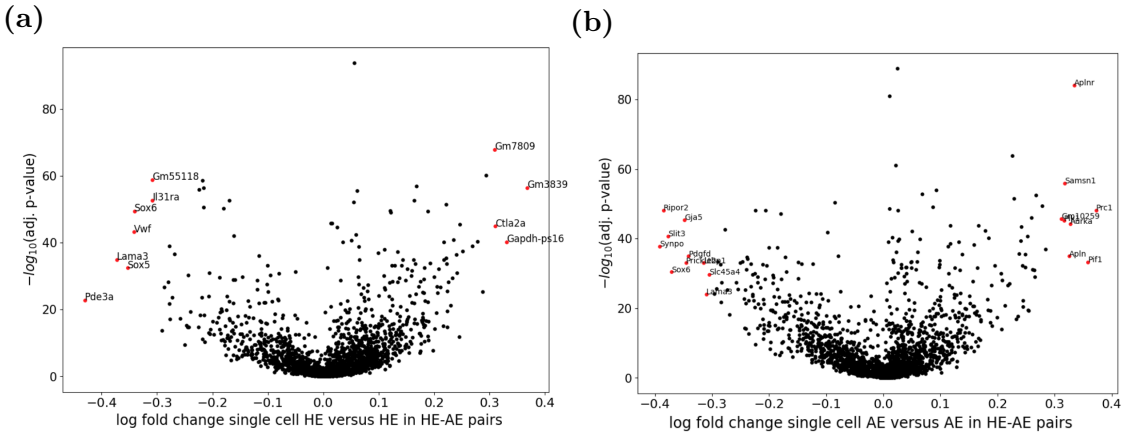


Figure 6.12: Volcano plots showing the adjusted p-values and log-fold change for (a) HE and (b) Arterial Endothelium *in silico* cells retrieved from HE-Arterial Endothelium clump pairs, compared to their single-cell counterparts. Genes are considered significant (red) if the adjusted p-value is lower than 0.05 and the absolute mean log-fold change is above a threshold (see main text).

and run tests for HE and Arterial Endothelium separately, comparing the retrieved profiles with the single-cell ones. Volcano plots are shown in Figure 6.12.

6.5 Integration with 10x Data

EHT is characterized by a transitory nature, with HE undergoing significant transcriptomic changes within a short window of time. Consequently, capturing this continuum of gene expression at a high level of granularity can be difficult. While Smart-seq3 provides single-cell measurements with high resolution and sensitivity, it is quite limited in throughput. For instance, the sequencing experiment described in the previous section resulted in around 820 cells after quality control filtering. We would like to obtain a broader and more granular representation of the gene expression landscape and since the EHT has been studied in previous scRNA-seq experiments, it would be beneficial for our analysis to query other datasets. Nevertheless, these datasets were collected across multiple experiments, in different labs, often with different sequencing technologies, and most likely consist of distinct cell type compositions. Consequently, comparing biological signals between them can prove challenging due to these technical effects. This leads to the general problem of *data*

integration: combining multiple scRNA-seq datasets from the same tissue into one joint and consistent embedding to be used for downstream analysis.

6.5.1 Data Integration

The task of data integration can become challenging, as it involves removing the technical variation between datasets. Some approaches have relied on using generalized linear models, e.g. ZINB-WAVE (Risso et al., 2018); Harmony (Korsunsky et al., 2019), where the data specific effects are quantified via linear regression coefficients. However, linear models show limited potential, especially when the datasets have distinct compositions of cell types or only some datasets contain a certain cell type. Subsequent integration methods consisted of factorization of data matrices into a low-dimensional representation, usually followed by alignment of these representations to encourage mixing between datasets, often measured by heuristic metrics. For example, LIGER (Welch et al., 2019) uses non-negative matrix factorization followed by quantile normalization; Seurat anchors (Stuart et al., 2019) performs dimensional reduction via canonical correlation analysis and aligns the resulting vectors using mutual nearest neighbors. VAE models for data integration have been explored in trVAE (Lotfollahi et al., 2020), CoMP (Foster et al., 2022), SCANVI (Xu et al., 2021).

Quantitative evaluation of integration requires 2 measures. First, the latents encoded from different datasets should mix well; in other words, the k-nearest neighbor graph of any cell should consist of a balanced composition across datasets. Second, cells from each dataset should retain their original embedding; that is, for each dataset the KNN graph of its cells should be similar in the integrated latent space as it is when encoding it in isolation. As noted in (Xu et al., 2021), these two measures are somewhat antithetic and neither is sufficient without the other. The first metric can be maximized by randomizing the data and the second can be optimized by trivially concatenating the latent vectors. Therefore, an integration model performs well when both metrics are maximized.

The first metric can be measured via the entropy of batch mixing (Haghverdi et al., 2018), which is the average entropy of the composition of batch proportions in the KNN graph of each cell in the integrated space. Higher values indicate good mixing, as the KNN graph contains cells from multiple datasets. There are situations, as will be shown in the next section, where one dataset contains much fewer samples than the other. In this case the entropy of batch mixing can be misleading, since most cells in the KNN graph are expected to come from the larger dataset even when the mixing is good. Therefore, the metric of batch mixing, measuring composition of local graphs, should incorporate the global composition of datasets. To this end, we used a modified version from (Gayoso et al., 2021): for fixed k, let $p_i^{(n)}$ be the frequency of batch i in the k-nearest neighbor graph of cell n, and q_i the global frequency of i. The metric is the relative entropy between local frequencies in the KNN graph and global frequencies:

$$\mathrm{KL}(p^{(n)}||q) = \sum_{n=1}^{N} p_i^{(n)} \log \frac{p_i^{(n)}}{q_i} \tag{6.3}$$

The second metric, conservation of the latent structure for each dataset, can be measured via the average Jaccard index. For each dataset we consider two KNN graphs: one computed in the integrated latent space and the other in the space obtained from running the model on just this dataset. For cell n and the two local KNN neighborhoods $\mathcal{N}_1^{(n)}$ and $\mathcal{N}_2^{(n)}$, the Jaccard index measures the consistency between them and is given by:

$$\mathcal{J}_n = \frac{|\mathcal{N}_1^{(n)} \cap \mathcal{N}_2^{(n)}|}{|\mathcal{N}_1^{(n)} \cup \mathcal{N}_2^{(n)}|} \tag{6.4}$$

In summary, higher values of both (6.3) and (6.4) indicate good mixing in the latent space.

6.5.2 Analysis

In this section we set out to obtain a more comprehensive representation of the EHT trajectory and cell type composition of the niche by integrating with another single-cell dataset. The objective is to integrate the Smart-seq3 dataset analyzed in Section 6.3.3 and use the joint latent representation to annotate the cell type clusters. Since only data labels are required for integration, we used a conditional scvi model.

We chose a 10x dataset generated by Dr. Giorgio Anselmi, de Bruijn Group (unpublished). This dataset is part of a large sequencing experiment of mouse embryos from days E9, E10 and E11, using multiple reporter lines. Cells were sorted from caudal parts of E9 embryos, and AGM & VU arteries in E10 and E11 embryos. The reporter mouse lines consisted of 23GFP, 204GFP and KitlTOM. The 23GFP line picks up cells with hemogenic potential, including (pre-)HE, as well as all hematopoietic cells. The 204 enhancer is expressed in a small fraction of arterial endothelial cells, including HE, and niche cells in the subaortic tissue; a subset of these labelled endothelial cells will also express 23 enhancer activity (de Bruijn group; unpublished). Finally, Kit ligand (*Kitl*) labels a broader range of niche cells predominantly in the ventral side of the dorsal aorta. Figure 6.13a reiterates the fluorescent confocal images from cross-sections of the dorsal aorta for different reporter mice, including 204GFP. In summary, these 3 reporter lines can be used to compile a dataset constituting both the subset of cells with potential undergo EHT (captured via 23GFP) and the ones composing mainly the niche (204GFP, KitlTOM).

The objective is therefore to integrate 10x and Smart-seq3 datasets into one joint latent embedding, which will be used for clustering cells into one common set of cell types between the two datasets. A schematic of the reporter composition and integration is shown in Fig. 6.13b. The two datasets exhibit distinct levels of technical variation, with 10x consisting of shallower reads and high throughput while Smart-seq3 has deeper reads and less throughput. For integration we implemented a Conditional VAE model where the networks are conditioned on data modality. In order to mitigate difficulties in inference due to the stark sensitivity and depth

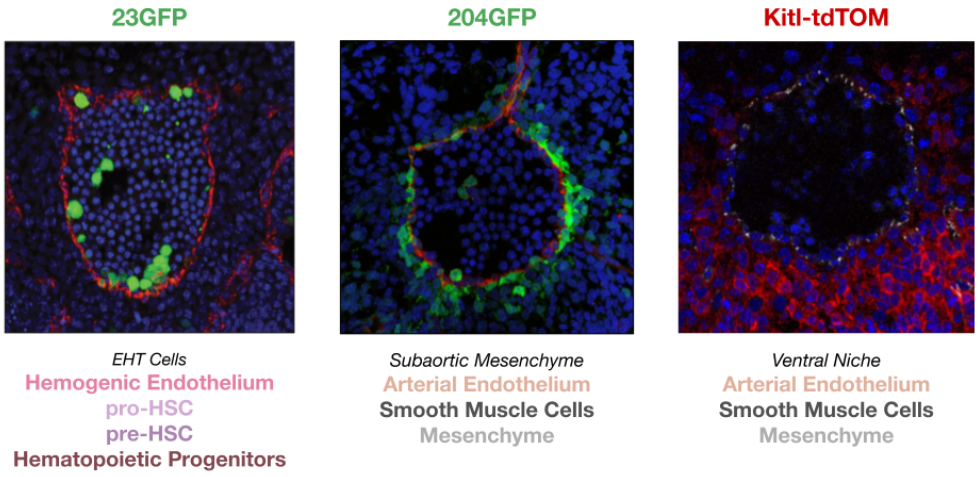
differences between the two sequencing technologies, we found two minor modifications that helped with improving integration. The first was using $x \to \log(x+1)$ as input for the encoder, to reduce the gap in input values between the two modalities (Smart-seq3 will have much higher count values than 10x). Secondly, in order to incorporate differences in depth and sensitivity, the gene dispersion parameter θ_g in the Negative Binomial likelihood is data-specific.

Figure 6.14a shows the UMAP computed on the latent space of the joint model. The plot shows good mixing between 10x and Smart-seq3 samples, where the latter mix with a subset of the Endothelial and Mesenchyme 10x population, as expected. Figure 6.14b shows the UMAP colored by embryonic day of 10x samples. Comparing this with Figure 6.14a confirms that all Smart-seq3 samples mix with the 10x E9 samples only, as expected. The distribution of the 3 values of embryonic day across endothelial and hematopoietic clusters agree with the timeline of development. There is a cluster of E9 endothelial cells (EC), with a subset that are primed for a hemogenic fate. The latter cells, constituting the hemogenic endothelium, undergo EHT at E10; subsequent hematopoietic development of the progenitors occurs at days E10 and E11. There is also a separation by day in the mesenchymal cluster. Mixing metrics for measuring the quality of integration are shown in Figures 6.14c and 6.14d. These indicate a good quality of mixing; as a reference, Section C.3 shows an example of these metrics when samples do not mix.

We first analyzed the composition of the 10x data as it appears within this joint embedding, showing that it provides a good representation of the cell types present in the dorsal aorta. The next step describes how Smart-seq3 samples map within this embedding via gene expression, followed by jointly clustering the cells from both data modalities.

Figure 6.15a highlights the 10x cells, colored by the pre-annotated cell type clusters (shorthand annotations for brevity). The endothelial and hematopoietic cell clusters are adjacent, outlining the continuous developmental path of the hematopoietic cells from the endothelial precursors. The endothelial cells with EHT potential differentiate into hematopoietic progenitor cells (HPC) via the

(a)



(b)

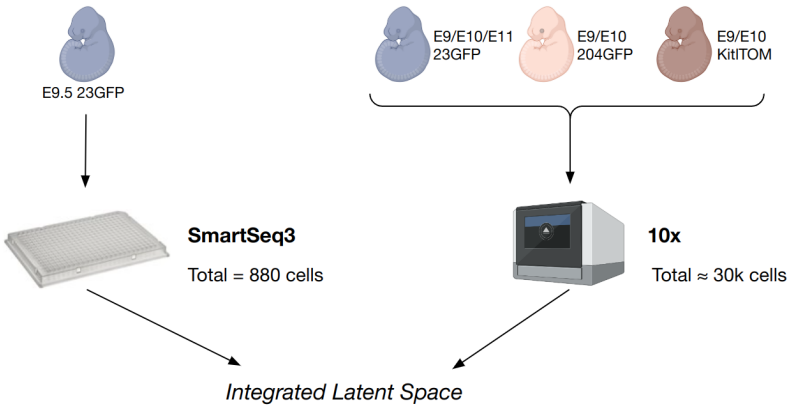


Figure 6.13: (a) Microscopy images of cross-sections of the E10 dorsal aorta showing fluorescent expression of 23GFP, 204GFP and KitlTOM reporter lines, including the cell type populations they label, shown below. *Courtesy of de Bruijn Group, MRC WIMM, University of Oxford.* (b) Schematic of the composition of Smart-seq3 and 10x datasets by embryo day and reporter line. The two datasets were generated with different sequencing technologies, with distinct levels of sensitivity and throughput; an integration model removes the technical variation between the datasets and summarizes the main biological signals in one joint low dimensional representation. Created with BioRender.com.

pre-HSC intermediate stage. The progenitors will further develop into a multitude of hematopoietic cell types: definitive erythrocytes (EryD), megakaryocytes (Mk), neutrophiles (Neu), macrophages (Mac1, Mac2). The other cell types, constituting the ventrally-distributed non-hematoendothelial populations in the AGM, are clustered

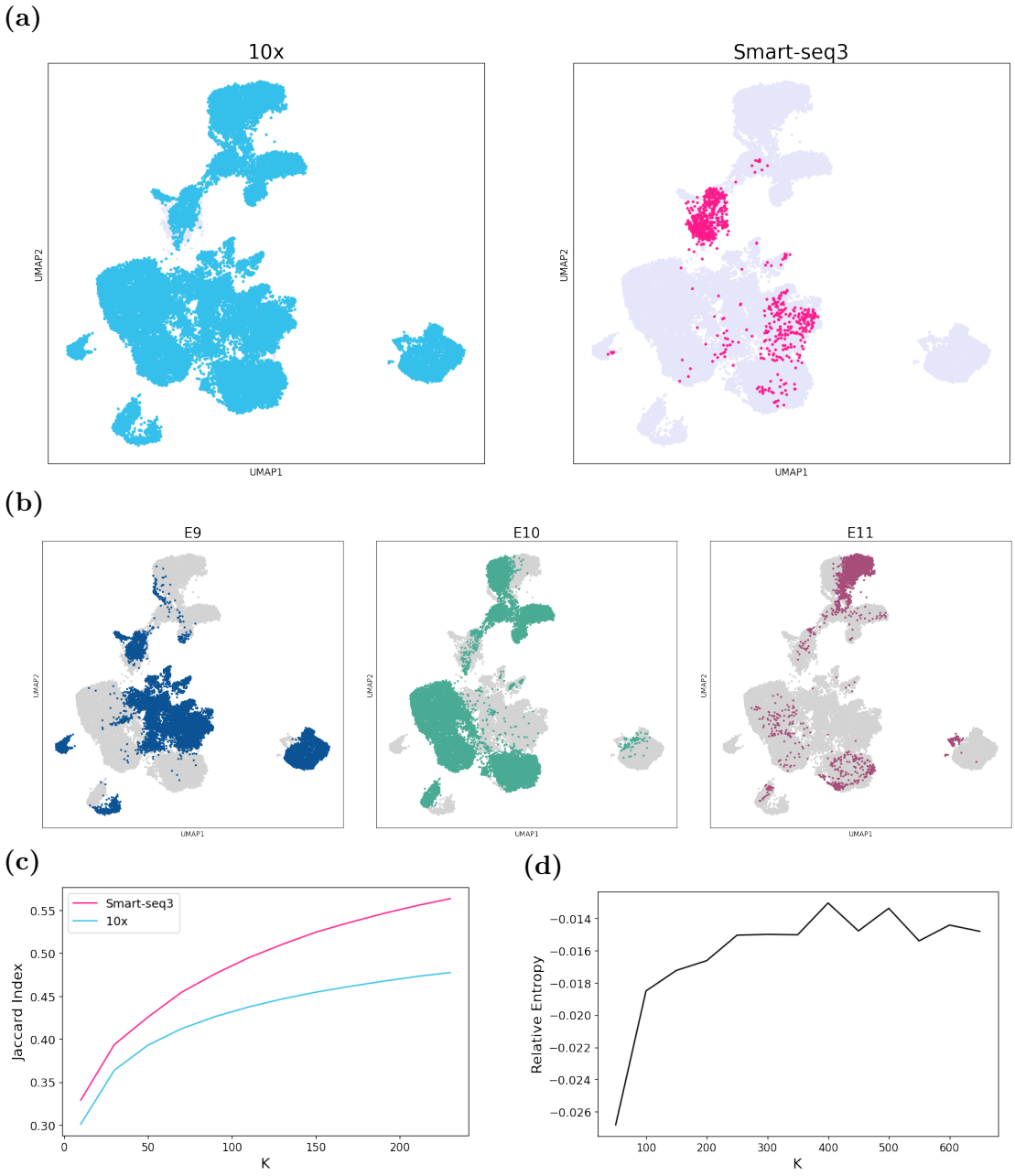
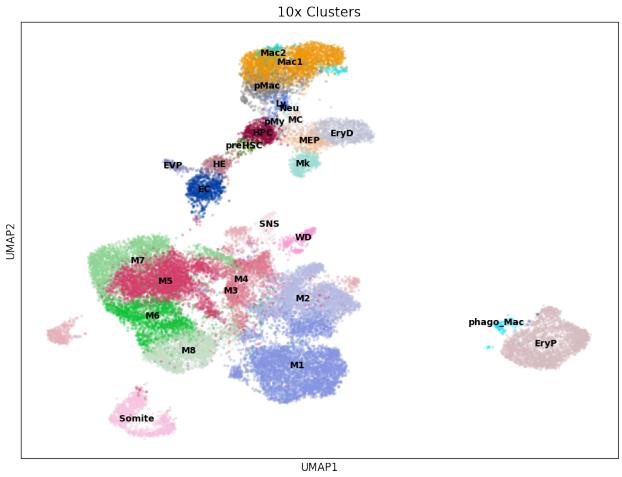


Figure 6.14: UMAP of the latent space of the integration model, highlighting (a) Smart-seq3 and 10x datasets, and (b) embryonic day. (c,d) Mixing metrics for the integration model: (c) Jaccard index and (d) relative entropy of batch mixing as function of nearest neighbor graph size K; higher values are better for each.

(a)



(b)

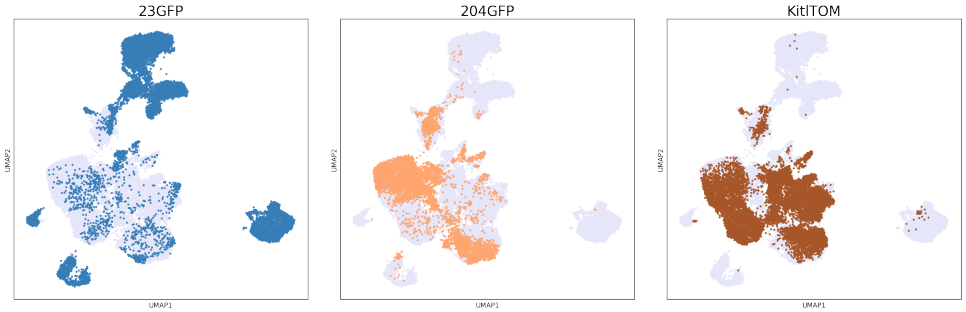


Figure 6.15: Analysis of 10x data. UMAP of the latent space of the integration model showing sample covariates for 10x data: (a) pre-annotated cell type clusters and (b) reporter line.

separately. The biggest population is represented by the mesenchymal cells, further subclustered into 8 populations. There is a population of primitive erythroids (EryP) and phagocytic macrophages (phago_Mac), both developing early at E7 in the blood islands of the yolk sac (Section 6.1) and subsequently migrating towards the dorsal aorta. The remaining cells types are Wolffian Duct (WD), Sympathetic Nervous System (SNS) and Somites.

Plotting the cells by reporter origin (Fig. 6.15b) shows the E9 endothelial population is made up of samples from all 3 reporter lines; while this population does not express *Runx1*, the cells with EHT potential are already picked up by the 23GFP expression. The UMAP clearly outlines the separation between endothelial cells and hemogenic endothelium undergoing EHT at E10 onwards, where the latter population and its progenitors are labelled predominantly by 23GFP only. As expected, the

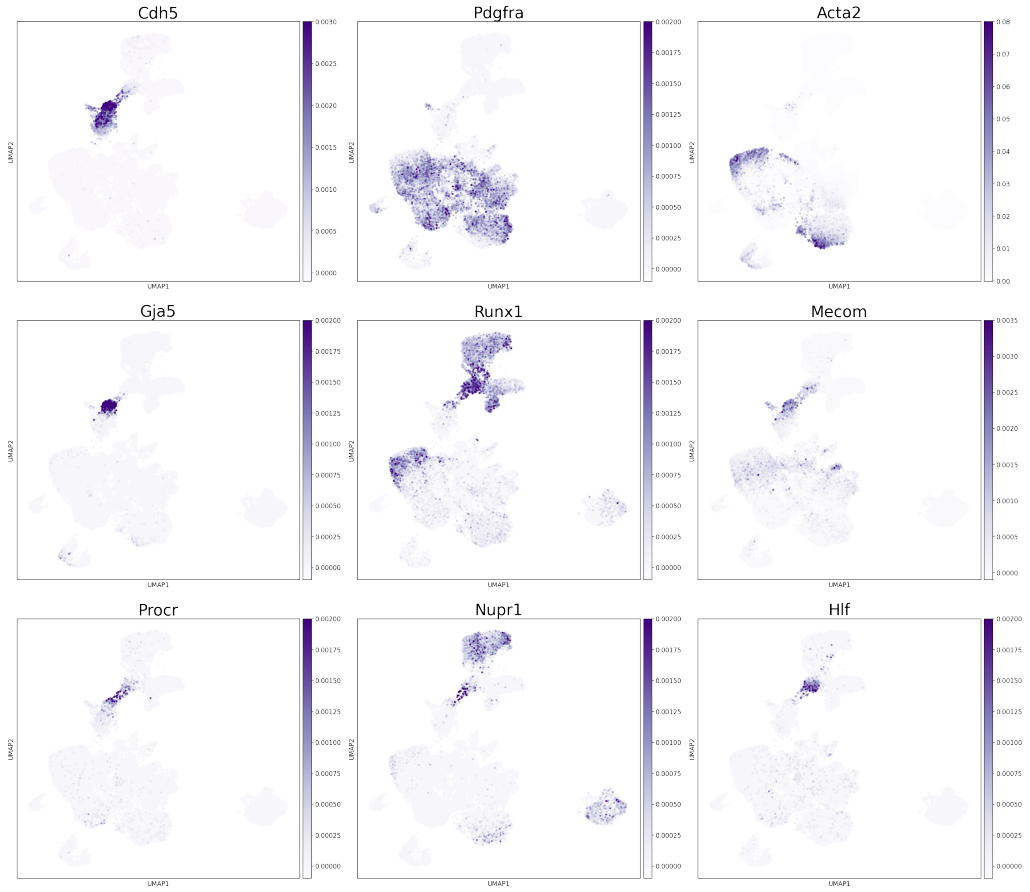


Figure 6.16: UMAP of the latent space of the integration model, colored by normalized gene expression.

majority of mesenchymal cells are captured by the 204GFP and KitlTOM lines.

The next step is analyzing where Smart-seq3 cells map and which 10x populations they mix with (Fig. 6.14a). Comparing the regions of mixing with the embryonic day of 10x samples confirms that Smart-seq3 endothelial cells mix mostly with corresponding E9 10x cells, agreeing with the timeline of sampling. Furthermore, Smart-seq3 cells from Plate M (sorted with KitlTOM) mix with the 10x mesenchymal clusters M1, M4, M5. The two small contaminant clusters found in the analysis in Section 6.3.3 map to Hematopoietic Progenitor Cells (HPC) and Wolffian Duct (WD) populations. Agreement between the mappings of the two data modalities is confirmed by gene expression, as shown in Figure 6.16. Marker expression confirms the presence of endothelial cells (*Cdh5*), hemogenic endothelium (*Mecom*, *Procr*), pre-HSCs (*Nupr1*) and hematopoietic progenitors (*Hlf*).

To obtain a more granular representation of the EHT trajectory, we further analyzed these aforementioned populations in more detail and used gene expression markers to annotate the cell type clusters. Figure 6.17 shows a UMAP computed on latents from a model run on just those cells, from both data modalities, colored by dataset (Fig. 6.17a) and reporter (Fig. 6.17b), while keeping in mind that Smart-seq3 cells were sampled from 23GFP and KitlTOM reporters. Cells were clustered on this joint representation (Fig. 6.17c). Raw proportions of relevant genes and transcription factors offer insight into the transcriptomic identities. Clusters 3, 4, 5 express venous markers *Aplnr* and *Nrp2*, thus they were labelled Vascular Endothelium. Expression of *Gja4* and *Gja5* points out to the arterial identity of cluster 2. Receptor *Cd44* and Transcription Factor *Meis1* in cluster 1 point out the hemogenic endothelium population. Therefore clusters 1 and 2 were named HE and Arterial, respectively. Nevertheless, expression of late HE & pre-HSC markers *Mllt3* and *Procr* shows a divide in cluster 1, that can also be seen by the distribution between E9 and E10 samples (Fig. 6.14b) and reporter lines (Fig. 6.17b). Therefore, with the same rationale as in Section 6.3.3, cluster 1 was further divided into pre-HE and HE, corresponding to E9 and E10, respectively. Cluster 6 exhibits levels of *Nupr1*, which begins expression at the pre-HSC stage, hence this cluster was annotated accordingly. Finally, cluster 6 is made up of hematopoietic progenitors, as confirmed by expression of *Hlf* and *Cd27*. A summary of gene expression across clusters and composition by data modality are shown in Figures 6.17e and 6.17f.

6.6 Integrated Model for Deconvolution of E11 Clumps

The clump experiment resulted in two plates, sorted from E9.5 and E11 embryos, respectively, and sequenced with Smart-seq3. Section 6.4 looked at the application of the deconvolution model to the E9.5 plate, using a reference conditional VAE model run on the E9.5 Smart-seq single-cell dataset. Retrieving cell type profiles from clumps requires corresponding single cell samples in the reference dataset for

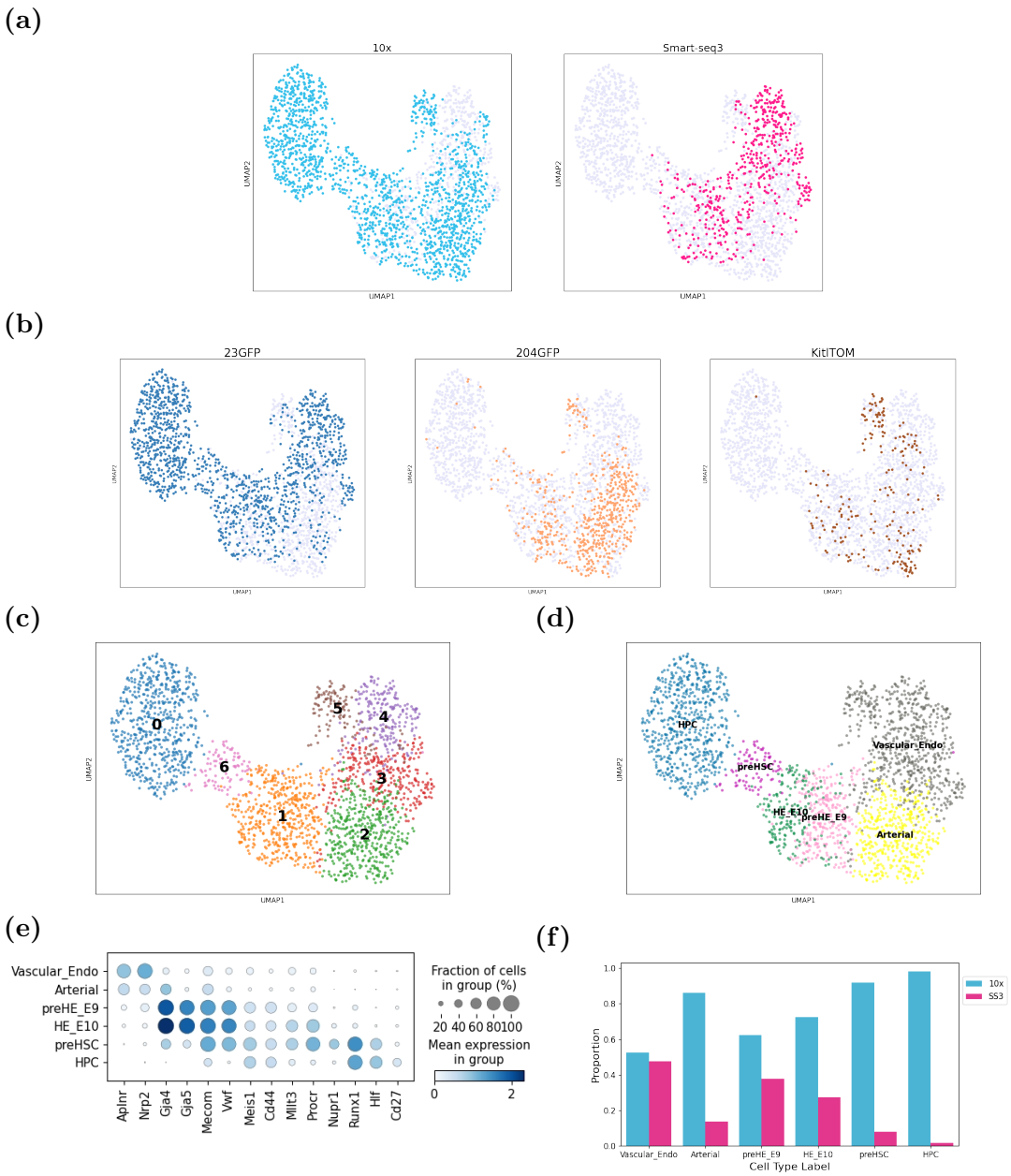


Figure 6.17: Integration of a selected population from Smart-seq3 and 10x datasets, consisting of endothelial cells, hemogenic endothelium, pre-HSCs and hematopoietic progenitors. UMAP of latent space highlighting (a) data modalities; (b) reporter line (from both datasets); (c) cluster labels; (d) cell type annotations. (e, f) Gene expression and distribution of data modalities, respectively, across the newly annotated cell types.

each cell type. Given the dynamic nature of EHT, the cell type composition of the dorsal aorta at E11 is expected to be different than at E9.5, including newly differentiated cell types. These populations will not be present in the E9.5 reference dataset. Nevertheless, to mitigate this deficiency in the reference dataset, we are able to integrate multiple single-cell datasets, as shown in Section 6.5. While the Smart-seq3 single-cell dataset generated in this study captures the beginning stages of EHT, the 10x data captures a broader landscape where EHT is well under way. Therefore, we expect the 10x dataset to capture the cell types present in E11 clumps.

We consider the two datasets, with two sample covariates: cell types c (annotated in the previous section) and data modalities $s \in \{1, 2\}$. We extend the conditional VAE structure such that the networks are conditioned on both covariates; thus the generative and inference models are $p(x|z, s, c)$ and $q(z|x, s, c)$. The resulting latent space will show mixing across both cell types and modalities.

We next used this single-cell embedding for the deconvolution model, applied to both clump plates. We modified the parameters of the generative distribution as follows:

$$y_{ng} \sim \text{NB}\left(N_n\Big(\sum_c \alpha_{nc}\, b_g \widetilde{f}(z_{nc}|c)_g\Big), \theta'_g\right) \tag{6.5}$$

$$\widetilde{f}(z_{nc}|c)_g = \frac{1}{2}(f(z_{nc}\,|c, s = 1)_g + f(z_{nc}\,|c, s = 2)_g) \tag{6.6}$$

All the clump data is Smart-seq3. Since the distribution in (6.5) contains expression decoded with 10x labels, we introduced a gene-specific bias term b_g to mitigate differences in expression between the two modalities. We eliminate the data modality conditioning by marginalizing them in (6.6).

Figure 6.18 shows the UMAP computed on a joint set of single-cell and retrieved cell type proportions. We note that these gene proportions were computed via the marginalization in (6.6), including for single cells. Figure 6.18a shows the distribution of the retrieved profiles on the UMAP, split by E9.5 and E11 clumps. We note how almost all E9.5 clumps contain endothelial cells; on the other hand, E11 clumps show a richer composition of cell types, including macrophages, hematopoietic

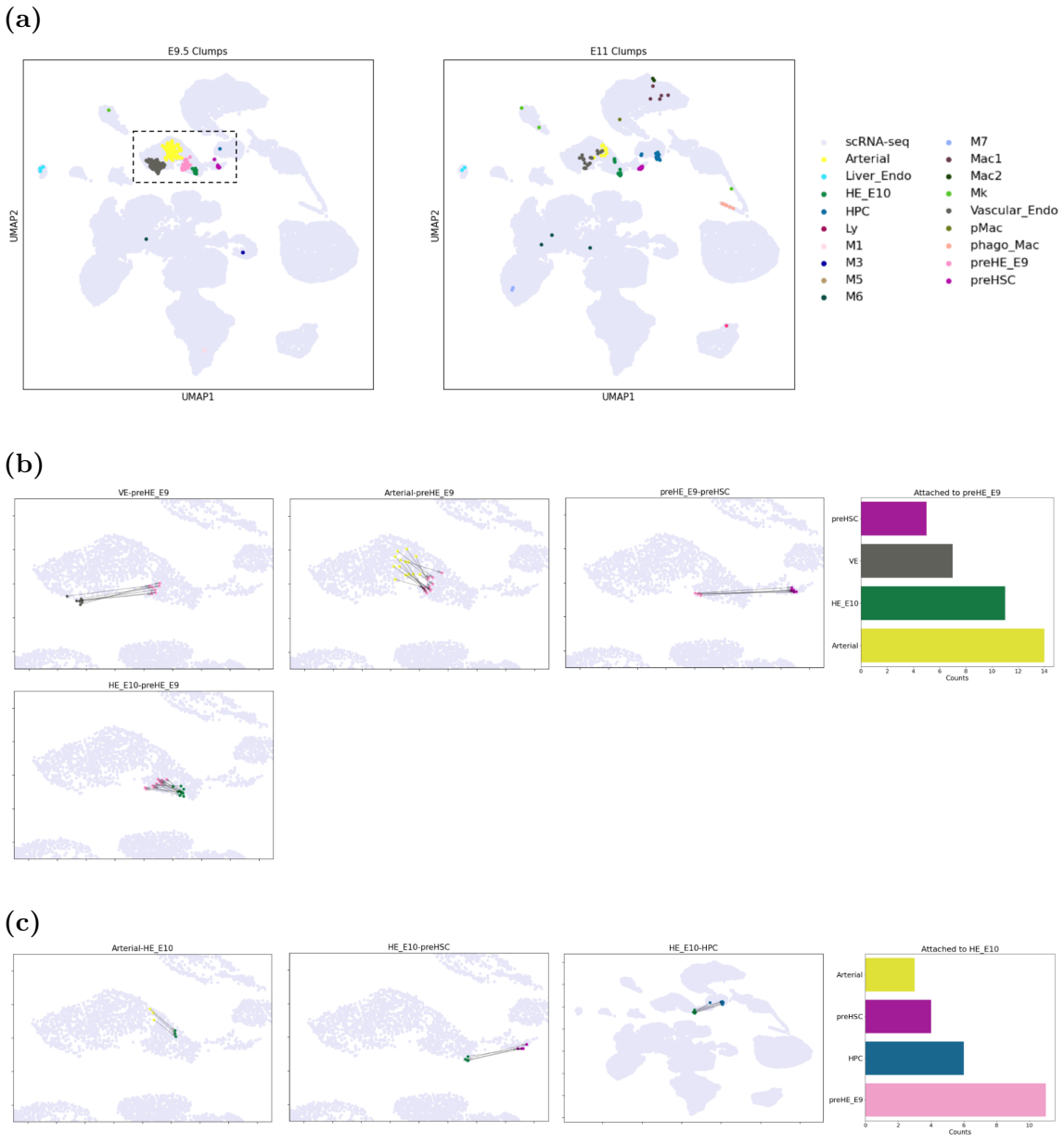


Figure 6.18: Cell type deconvolution of Smart-seq3 clumps from E9.5 and E11 plates. (a) Retrieved cell type profiles mapped onto the integrated Smart-seq3 & 10x single clusters, plotted by embryonic day of clump plates. The rectangle highlights the niche of endothelial cells, HE and preHSCs; (b, c) Mapping of pairings of (b) preHE (E9) and (c) HE (E10) to neighboring cell types that were identified from the deconvolution model.

progenitors, megakaryocytes and preHSCs. This finding is in line with the expected tissue composition, since the dorsal aorta of E11 embryos will contain differentiated hematopoietic cells.

The main goal of our study was to analyze how the AGM niche influences EHT, by identifying the profiles of cell types attached to HE. These populations might interact with HE and provide signals to initiate (at E9) or support (at E10) the process. Therefore, we focused on the cell types attached to preHE (E9) and HE (E10) (Figures 6.18b and 6.18c). We found that populations from both cell types are attached to arterial endothelial cells and preHSCs. It also appears that HE is attached to hematopoietic progenitors. Interestingly, we found a population of vascular endothelial cells (transcriptionally venous-like) exclusively attached to preHE cells. This leads to the question of whether these cells are part of the aortic wall or are endothelial progenitors migrating towards the aorta. The findings in this chapter lay the basis for future work that will study whether these populations in the AGM niche play an important role in interacting with EHT.

6.7 Discussion

This chapter tackled the complex question of how the microenvironment surrounding endothelial cells with hemogenic potential influences their fate choice and regulates EHT. Identifying and sorting the endothelial cells involved or which are about to be involved in EHT is straightforward, and allows us to generate single-cell datasets in order to study and characterize the transcriptional changes occuring throughout this differentiation process. While sorting cells from the niche is also possible (as seen e.g. with the KitlTOM reporter), the key information on how these niche cells interact with the hemogenic endothelium, or even the composition of cells attached to it, is not available in single-cell data. For this reason we turned our attention to ClumpSeq as a way to learn this information. By adapting the ClumpSeq protocol to the embryo tissue, we were able to generate a novel dataset consisting of samples

of hemogenic endothelium attached to its neighbors. To our knowledge, there has not been any other clump dataset generated for studying embryonic hematopoiesis.

The novel insights that could be provided by this data motivated us to develop a quantitative model that reconstructs individual cell type profiles from clumps. Deep generative models have been shown to work well on fitting single-cell data, thus it was natural to build up the clump deconvolution model on top of this framework. First, the single-cell VAE model trained on the Smart-seq3 data revealed the main cell types found present in the dorsal aorta at E9.5. Secondly, the deconvolution model developed in previous chapters proved to successfully retrieve the cell type profiles from E9.5 clumps, as shown by the quality of mixing between the retrieved and single-cell profiles. One of the key advantages of this deep generative modeling framework was that throughout this analysis we were able to use a single end-to-end robust pipeline for both model training and downstream analysis. As a result, the latent space and neural network components of the VAE were used for clustering (Section 6.3.3) and differential gene expression (Section 6.4.2).

Furthermore, the flexibility of conditional VAEs allowed us to integrate the newly generated Smart-seq3 data with a high throughput 10x dataset. Despite the technical variation differences between the two datasets, the integration model was able to learn a biologically meaningful latent representation. The data generated in this study included a plate of E11 clumps; despite the Smart-seq3 single-cell dataset not containing E11 populations, we were able to use the embedding of the integrated model for deconvolution.

The main limitation of our analysis is the number of clumps sequenced in the experiment. The most abundant pairings found in the analysis were endothelial doublets and hemato-endothelial; these are expected to be abundant, especially the former since endothelial cells are the most frequent and have stronger connections. However, while not found in our dataset, interactions with the mesenchyme cannot be excluded, as it is likely these interactions are rarer. Having gained experience in the experimental generation and quantitative analysis of clumps, we can expand our study of microenvironment interactions with EHT by conducting further experiments.

Chapter 7

Conclusions and Future Work

This thesis tackled the objective of investigating and better understanding the composition of cell neighborhoods. We described the comprehensive insights that can be obtained from single-cell RNA sequencing data, yet highlighting an important drawback, namely the lack of spatial information on the proximity between sequenced cells. We thus turned our attention to a newly developed exploratory method called ClumpSeq, which provides both the high resolution of single-cell sequencing and neighborhood information.

The focus of this thesis was quantitative, consisting of developing a computational model for analyzing ClumpSeq data in order to extract relevant biological information that could contribute towards understanding how cells are influenced by their neighbors. Our data driven approach consisted of formulating a probabilistic model, in order to incorporate the generative properties of the data and quantify uncertainty. This allowed us to perform an end-to-end analysis starting from modeling the raw gene counts to performing downstream tasks. We focused on the Variational Autoencoder framework, which has been shown to work well in modeling single-cell data, and extended it into a clump deconvolution model. Comparison with other published methods, designed for deconvolution of spatial or bulk RNA-seq data, showed that our model achieved good performance in extracting cell type specific signals.

We first validated our model on a publicly available dataset, describing the epithelial tissue in the gut. Our deconvolution model successfully retrieved profiles of absorptive and secretory cells. The high resolution of clump data (since its expression

only comes from 2 cells) makes it flexible in integrating it with single-cell clusters. Interestingly, it also proved to be sensitive enough to map retrieved enterocyte profiles to specific areas of the enterocyte single-cell cluster, avoiding regions of single cells that express genes not found in clumps. We were also able to develop a pseudotime deconvolution model in order to relate the trajectories of enterocyte and goblet populations. It was interesting to see that there is not necessarily a 1-to-1 mapping between these trajectories. The key point to emphasize is the ability to link the trajectories was solely provided by the clump data, which highlights its advantages.

We then applied the ClumpSeq experimental framework to generate a novel dataset to study the composition of the microenvironment surrounding the emerging HE during embryogenesis. We also encountered a scenario where the clump data contained signals not found in the single cells (a somewhat "reverse" case of the above), and the flexibility of the deconvolution model made it straightforward to adjust it. Finally, we were able to successfully integrate the new Smart-seq3 dataset with 10x data, two technologies that exhibit very different levels of variation, and also perform deconvolution. The model results aligned with our expectations of what cell types might reasonably be found near HE. It was also interesting to see that we are able to retrieve profiles of hematopoietic cells from E11 Smart-seq3 clumps, despite there not being a Smart-seq3 single-cell subset of these types. We find these results motivating in terms of continuing to analyze co-expression patterns between cell types and conducting more sequencing experiments.

While VAEs provide a multitude of advantages, it is important to note their drawbacks. The flexibility of neural networks comes at the cost of interpretability, hence why VAEs are considered 'black box' solvers. Furthermore, in differential expression analyses we consistently found that VAEs have a tendency to output very small non-zero gene proportions that do not correspond to any counts. Such behaviour can results in biased results if these spurious values are not accounted for. This motivated us to be careful in choosing the offset terms ϵ for computing the log-fold change.

In terms of future work, we envision several extensions: one would be to model in more detail the potential interactions found in clumps but not single cells, which might lead to novel discoveries about expression behaviour initiated or disrupted by single-cell dissociation protocols. This could also be extended to use clumps to infer new cell types that are difficult to sequence in single-cell format; for instance, (Halpern et al., 2018) sequenced pairs of liver endothelial cells and hepatocytes, but their analysis did not contain hepatocyte single-cell data, as it is difficult to obtain due to the inherent low mRNA abundance. It would also be interesting to apply ClumpSeq to the tumour microenvironment, which is known to comprise various cell interactions, to investigate e.g. if tumour progression is affected by its neighbors.

From a methodological point of view, it would beneficial to incorporate a measure or regularization of boundaries between cell type clusters. Currently cell type labels are assigned based on clustering or marker expression; the latter is especially relevant for continuous trajectories. We noticed that our model might find it difficult to perform deconvolution if cell types are mostly similar - as it is the case with pre-HE and HE, for example. It would be interesting to see whether this partition of cells is instead done according to a measure of "dissimilarity", i.e. perform a cell label assignment that maximizes the model's ability to separate them in a deconvolution setting.

Overall, we find that clump data can provide valuable biological insights and we hope that it can become a useful exploratory tool in many future studies.

Appendices

Appendix A

Data Processing

All preprocessing was done with the *scanpy* Python package (Wolf et al., 2018), version 1.9.2.

A.1 Manco et al, 2021 Single-Cell and Clump Datasets

The data was downloaded from Gene Expression Omnibus, accession code GSE154714.

We selected the set of genes shared between the two datasets. We removed genes that were expressed in fewer than 3 cells or clumps. We then selected the top 5000 highly variable genes in each dataset separately, using the count based method in (Stuart et al., 2019). The final set of genes was the union of the 2 gene sets, totalling 7668 genes. Finally, we retained cells and clumps with library size greater than 300 counts.

A.2 de Bruijn Group EHT Single-Cell and Clump Datasets

Raw sequencing files were converted to .FASTQ files via the bcl2fastq package (Illumina, version v2.20.0.422). Reads were aligned to the mouse reference genome (GRCm38) via STAR (Dobin et al., 2012, version 2.7.3a) and subsequently processed into gene expression counts with zUMI (Parekh et al., 2018, version 2.9.7d).

Analogously to the previous section, we filtered genes that were shared between the two datasets. We removed genes that were expressed in fewer than 3 cells or clumps. We retained cells and clumps with library size greater than 25000 UMI counts. We selected the top 2000 highly variable genes for each dataset, similarly as in the previous section, totalling 3044 genes.

Appendix B

Appendix for Chapter 5

B.1 β-VAE Rescales the Latent Space

We begin with Corollary 1 from (Mathieu et al., 2019):

Proposition. (Mathieu et al., 2019; Corollary 1) *If* $p(\boldsymbol{z}) = \mathcal{N}(z; 0, \Sigma)$ *and* $q_\phi(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}; \mu_\phi(\boldsymbol{x}), S_\phi(\boldsymbol{x}))$, *then*

$$\mathcal{L}_\beta(\boldsymbol{x}; \theta, \phi) = \mathcal{L}(\boldsymbol{x}; \theta', \phi') + \frac{(\beta - 1)}{2} \log|S_{\phi'}(\boldsymbol{x})| + c \tag{B.1}$$

where θ' *and* ϕ' *represent rescaled networks such that*

$$\begin{aligned} p_{\theta'}(\boldsymbol{x}|\boldsymbol{z}) &= p_\theta\left(\boldsymbol{x}|\boldsymbol{z}/\sqrt{\beta}\right) \\ q_{\phi'}(\boldsymbol{z}|\boldsymbol{x}) &= \mathcal{N}(\boldsymbol{z}; \mu_{\phi'}(\boldsymbol{x}), S_{\phi'}(\boldsymbol{x})) \\ \mu_{\phi'}(\boldsymbol{x}) &= \sqrt{\beta}\mu_\phi(\boldsymbol{x}); \quad S_{\phi'}(\boldsymbol{x}) = \beta S_\phi(\boldsymbol{x}), \end{aligned}$$

This proposition shows the ELBO of a β-VAE can be written in terms of the loss objective of a standard VAE in a latent space rescaled by a factor of $\sqrt{\beta}$. This section will rephrase the results of (Mathieu et al., 2019) to emphasize how the β-VAE loss objective rescales the latent space when the prior and variational posterior are Normal distributed. We note that the above proposition assumes a generative distribution centered at 0. This section will consider the pseudotime generative model (5.2), with sample specific prior $p(z_n) = \mathcal{N}(\mathcal{S}_{t_n}, \sigma^2)$, and will show how the $\beta-$VAE rescales the parameters of the variational posterior for a prior distribution with non-zero mean.

The first step is interpreting the above proposition. Let $p_\theta(\boldsymbol{x}; \boldsymbol{z})$ be a generative model with prior latent distribution $p(\boldsymbol{z})$. We restate Theorem 1 from (Mathieu

et al., 2019), stating that the β-VAE loss can be interpreted in terms of the ELBO of a standard VAE with annealed prior $p(\boldsymbol{z})^\beta / F_\beta$, where F_β is a normalization constant. Let $q_{\phi'}(\boldsymbol{z}|\boldsymbol{x})$ be the variational distribution of the latter VAE model. Then:

$$\mathcal{L}_\beta(\boldsymbol{x}) = \mathcal{L}(\boldsymbol{x}; \theta, \phi') + (\beta - 1)\mathbb{H}[q_{\phi'}(\boldsymbol{z}|\boldsymbol{x})] + \log F_\beta \tag{B.2}$$

The right-hand side of (B.2) is the loss objective of a standard VAE with a maximum entropy regularization (while ignoring constant F_β since it does not contribute to the optimization), composed of the ELBO and entropy of $q_{\phi_1}(\boldsymbol{z}|\boldsymbol{x})$.

The special case when both the prior and variational distribution are Normal distributed provides more interpretable results. We assume a generative model with prior distribution $\mathcal{N}(\boldsymbol{z}; 0, \Sigma)$. The aim is to characterize the resulting latent space from a β-VAE model trained with this prior. The corresponding annealed prior is $\mathcal{N}(\boldsymbol{z}; 0, \Sigma/\beta)$. We note that in the Normal case the annealed prior only rescales the latent space (and hence the variational posterior parameters) without changing the ELBO. To see this, we consider the following models:

1) $p_{\theta_1}(\boldsymbol{z}) = \mathcal{N}(0, \Sigma)$ and $q_{\phi_2}(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}; \mu(\boldsymbol{x}), S(\boldsymbol{x}))$, with generative model $p_{\theta_1}(x|z)$.

2) $p_{\theta_2}(\boldsymbol{z}) = \mathcal{N}(0, \Sigma/\sqrt{\beta})$ and $q_{\phi_2}(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{z}; \frac{\mu(\boldsymbol{x})}{\sqrt{\beta}}, \frac{S(\boldsymbol{x})}{\beta}\right)$, with generative model $p_{\theta_2}(x|z)$. This corresponds to the VAE model with annealed prior in (B.2).

These 2 models have equivalent ELBO objectives. We consider the following ELBO formulation:

$$\mathcal{L}(\boldsymbol{x}; \theta_1, \phi_1) = \mathbb{E}_{q_{\phi_1}}\left[\log p_{\theta_1}(\boldsymbol{x}|\boldsymbol{z}) + \log p_{\theta_1}(\boldsymbol{z})\right] + \mathbb{H}[q_{\phi_1}(\boldsymbol{z}|\boldsymbol{x})] \tag{B.3}$$

Each term can be written as follows:

$$\mathbb{E}_{q_{\phi_1}}\left[\log p_{\theta_1}(\boldsymbol{x}|\boldsymbol{z})\right] = \mathbb{E}_{q_{\phi_2}}\left[\log p_{\theta_1}\left(\boldsymbol{x}|\boldsymbol{z}\sqrt{\beta}\right)\right] \tag{B.4}$$

$$= \mathbb{E}_{q_{\phi_2}}\left[\log p_{\theta_2}(\boldsymbol{x}|\boldsymbol{z})\right] \tag{B.5}$$

$$\tag{B.6}$$

$$\mathbb{E}_{q_{\phi_1}}[\log p_{\theta_1}(\boldsymbol{z})] = \mathbb{E}_{q_{\phi_1}}\left[-\frac{1}{2}\boldsymbol{z}^\top \Sigma^{-1}\boldsymbol{z} - \frac{1}{2}\log|\Sigma| - \frac{D}{2}\log(2\pi)\right] \tag{B.7}$$

$$= \mathbb{E}_{q_{\phi_1}}\Bigg[-\frac{1}{2}\left(\boldsymbol{z}\sqrt{\beta}\right)^{\top}(\beta\Sigma)^{-1}\left(\boldsymbol{z}\sqrt{\beta}\right) - \frac{1}{2}(\log|\Sigma| + D\log\beta) - \tag{B.8}$$

$$- \frac{D}{2}\log(2\pi)\Bigg] + \frac{D}{2}\log\beta \tag{B.9}$$

$$= \mathbb{E}_{q_{\phi_2}}\left[\log p_{\theta_1}\left(\boldsymbol{z}\sqrt{\beta}\right)\right] + \frac{D}{2}\log\beta \tag{B.10}$$

$$= \mathbb{E}_{q_{\phi_2}}[\log p_{\theta_2}(\boldsymbol{z})] + \frac{D}{2}\log\beta \tag{B.11}$$

$$\tag{B.12}$$

$$\mathbb{H}[q_{\phi_1}(\boldsymbol{z}|\boldsymbol{x})] = \frac{D}{2}(1+\log 2\pi) + \frac{1}{2}\log\left|\frac{S(\boldsymbol{x})}{\beta}\right| \tag{B.13}$$

$$= \frac{D}{2}(1+\log 2\pi) + \frac{1}{2}(\log|S(\boldsymbol{x})| - D\log\beta) \tag{B.14}$$

$$= \frac{D}{2}(1+\log 2\pi) + \frac{1}{2}\log|S(\boldsymbol{x})| - \frac{D}{2}\log\beta \tag{B.15}$$

$$= \mathbb{H}[q_{\phi_2}(\boldsymbol{z}|\boldsymbol{x})] - \frac{D}{2}\log\beta \tag{B.16}$$

Putting all terms together leads to:

$$\mathcal{L}(\boldsymbol{x};\theta_1,\phi_1) = \mathbb{E}_{q_{\phi_2}}\left[\log p_{\theta_2}(\boldsymbol{x}|\boldsymbol{z})\right] + \mathbb{E}_{q_{\phi_2}}[\log p_{\theta_2}(\boldsymbol{z})] + \frac{D}{2}\log\beta + \mathbb{H}[q_{\phi_2}(\boldsymbol{z}|\boldsymbol{x})] - \frac{D}{2}\log\beta \tag{B.17}$$

$$= \mathcal{L}(\boldsymbol{x};\theta_2,\phi_2) \tag{B.18}$$

The entropy regularized loss objective (right-hand side of (B.2)) can be re-written in terms of the variational distribution parameters, by plugging in the entropy expression of the Normal distribution:

$$\mathcal{L}_{H,\beta}(\boldsymbol{x};\theta_1,\phi_1) = \mathcal{L}(\boldsymbol{x};\theta_1,\phi_1) + (\beta-1)\mathbb{H}[q_{\phi_1}(\boldsymbol{z}|\boldsymbol{x})] \tag{B.19}$$

$$= \mathcal{L}(\boldsymbol{x};\theta_1,\phi_1) + (\beta-1)\left(\frac{D}{2}(1+\log 2\pi) + \frac{1}{2}\log|S(\boldsymbol{x})|\right) \tag{B.20}$$

$$= \mathcal{L}(\boldsymbol{x};\theta_1,\phi_1) + \frac{\beta-1}{2}\log|S(\boldsymbol{x})| + c \tag{B.21}$$

where c is a generic notation for a constant, as it is not relevant for the optimization.

We also note the entropy regularized loss objectives for the 2 models are equivalent up to a constant:

$$\mathcal{L}_{H,\beta}(\boldsymbol{x};\theta_2,\phi_2) = \mathcal{L}(\boldsymbol{x};\theta_2,\phi_2) + \frac{(\beta-1)}{2}\log\left|\frac{S(\boldsymbol{x})}{\beta}\right| + c \tag{B.22}$$

$$= \mathcal{L}(\boldsymbol{x}; \theta_1, \phi_1) + \frac{(\beta - 1)}{2} (\log|S(\boldsymbol{x})| - D \log \beta) + c \tag{B.23}$$

$$= \mathcal{L}(\boldsymbol{x}; \theta_1, \phi_1) + \frac{(\beta - 1)}{2} \log|S(\boldsymbol{x})| + c \tag{B.24}$$

$$= \mathcal{L}_{H,\beta}(\boldsymbol{x}; \theta_1, \phi_1) + c \tag{B.25}$$

We now consider training a β-VAE on a generative model $p_\theta(\boldsymbol{x}, \boldsymbol{z})$ with prior $\mathcal{N}(\boldsymbol{z}; 0, \Sigma)$. Let q' be the corresponding variational posterior, parameterized by ϕ. The β-VAE loss objective is equivalent to the entropy regularized loss objective of the above VAE model (2) with annealed prior $\mathcal{N}(\boldsymbol{z}; 0, \Sigma/\beta)$, up to a constant:

$$\mathcal{L}_\beta(\boldsymbol{x}; \theta, \phi) = \mathbb{E}_{q'_\phi}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - \beta \mathrm{KL}\left(q'_\phi(\boldsymbol{z}|\boldsymbol{x}) || p_\theta(\boldsymbol{z})\right) \tag{B.26}$$

$$= \mathcal{L}_{H,\beta}(\boldsymbol{x}; \theta, \phi) \tag{B.27}$$

$$= \mathcal{L}_{H,\beta}(\boldsymbol{x}; \theta', \phi') + c \tag{B.28}$$

Parameters θ', ϕ' correspond to VAE model (1) and θ, ϕ correspond to model (2), where the networks are rescaled. For the Normal variational distributions, this means $\mu_\phi = \mu_{\phi'}/\sqrt{\beta}$ and $S_\phi(\boldsymbol{x}) = S_{\phi'}(\boldsymbol{x})/\beta$. In the same spirit as Corollary 2 of (Mathieu et al., 2019), it follows that objectives (B.27) and (B.28) share the same stationary points. Therefore, the networks resulting from optimization of (B.26), including $q'_\phi(\boldsymbol{z}|\boldsymbol{x})$, are equivalent to the rescaled optimal networks from (B.28).

We now focus on the pseudotime model, with generative distribution $\mathcal{N}(\boldsymbol{z}_n;\ \mathcal{S}_{t_n}, \mathrm{diag}(\boldsymbol{\sigma}^2))$ and posterior $q_\phi(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}; \mu_\phi(\boldsymbol{x}), S_\phi(\boldsymbol{x}))$, with $S_\phi(\boldsymbol{x}) = \mathrm{diag}(\sigma^2_\phi(\boldsymbol{x}))$. We rephrase it in terms of a latent variable $\boldsymbol{z}'$ centered at 0, with inference model $\widetilde{q}_\phi(\cdot\ |\boldsymbol{x})$:

$$\begin{aligned} \widetilde{p}(\boldsymbol{z}') &= \mathcal{N}(\boldsymbol{z}'; 0, \mathrm{diag}(\boldsymbol{\sigma}^2)) \\ \boldsymbol{x_n} &\sim p_\theta(\cdot\ |\boldsymbol{z}'_n + \mathcal{S}_{t_n}),\ n = 1, \ldots, N \end{aligned} \tag{B.29}$$

For each cell $n = 1, \ldots, N$ we can write the original latents as $\boldsymbol{z}_n = \boldsymbol{z}'_n + \mathcal{S}_{t_n}$. Therefore,

$$\widetilde{q}_\phi(\boldsymbol{z}'_n|\boldsymbol{x}_n) = q_\phi(\boldsymbol{z}_n + \mathcal{S}_{t_n}|\boldsymbol{x}_n) \tag{B.30}$$

$$= \mathcal{N}\left(\boldsymbol{z}'_n; \mu_\phi(\boldsymbol{x}_n) - \mathcal{S}_{t_n}, S_\phi(\boldsymbol{x}_n)\right) \tag{B.31}$$

$$= \mathcal{N}\left(\boldsymbol{z}'_n; \widetilde{\mu}_\phi(\boldsymbol{x}_n), S_\phi(\boldsymbol{x}_n)\right) \tag{B.32}$$

Parameters of the generative distribution are cell specific (due to S_{t_n}) and neural network parameters are shared across all data points due to the amortization scheme. The ELBO taken across all data points is

$$\mathcal{L}_\beta(\theta, \phi) = \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}_\beta(\boldsymbol{x}_n; \theta, \phi) \tag{B.33}$$

Training a β-VAE leads to optimal values identical to the ones in the rescaled inference distribution, given by

$$\widetilde{q}_{\phi'}(\boldsymbol{z}'_n|\boldsymbol{x}_n) = \mathcal{N}\left(\boldsymbol{z}'_n; \frac{\widetilde{\mu}_\phi(\boldsymbol{x}_n)}{\sqrt{\beta}}, \frac{S_\phi(\boldsymbol{x}_n)}{\beta}\right) \tag{B.34}$$

$$= \mathcal{N}\left(\boldsymbol{z}'_n + \mathcal{S}_{t_n}; \frac{\widetilde{\mu}_\phi(\boldsymbol{x}_n)}{\sqrt{\beta}} + \mathcal{S}_{t_n}, \frac{S_\phi(\boldsymbol{x}_n)}{\beta}\right) \tag{B.35}$$

$$= \mathcal{N}\left(\boldsymbol{z}_n; \frac{\mu_\phi(\boldsymbol{x}_n) - \mathcal{S}_{t_n}}{\sqrt{\beta}} + \mathcal{S}_{t_n}, \frac{S_\phi(\boldsymbol{x}_n)}{\beta}\right) \tag{B.36}$$

$$= q_{\phi'}(\boldsymbol{z}_n|\boldsymbol{x}_n) \tag{B.37}$$

Therefore, regularizing the KL term by a factor of β in loss (5.5) will rescale the parameters of the variational posterior distribution for each cell. The rescaled mean is $\mu'_n = \frac{\mu_n - \mathcal{S}_{t_n}}{\sqrt{\beta}} + \mathcal{S}_{t_n}$; this scales down the distance $|\mu_n - \mathcal{S}_{t_n}|$ between the variational and generative means by a factor of $\sqrt{\beta}$.

B.2 Functional Profiling of Hotspot Modules from Section 5.5

Module	*Functional profiling*
1	rRNA binding ubiquitin ligase inhibitor activity ribonucleoprotein complex binding peptide biosynthetic process peptide metabolic process amide metabolic process cellular metabolic process metabolism of RNA
2	mitotic cell cycle DNA metabolic process cell cycle DNA replication Mitotic cell cycle G1 to S cell cycle control
3	intestinal cholesterol absorption interferon-beta production immune response type I interferon production cytokine production lipoprotein particle
4	oxidoreductase activity antioxidant activity biological oxidations glutathione binding peroxidase activity metabolic pathways oxidative stress and redox pathway small molecule metabolic process
5	carbohydrate digestion and absorption peptide catabolic process fat digestion and absorption retinol metabolism lipid metabolic process

Table B.1: Enterocyte Hotspot modules and corresponding selected functional profiling queries from g:Profiler

Module	*Functional profiling*
1	rRNA binding structural constituent of ribosome ribonucleoprotein complex binding peptide metabolic process amide biosynthetic process large ribosomal subunit polysomal ribosome
2	protein binding regulation of RIG-I signaling pathway positive regulation of defense response

Table B.2: Goblet Hotspot modules and corresponding selected functional profiling queries from g:Profiler

Appendix C

Appendix for Chapter 6

C.1 Summary of Clump Generation Experiment

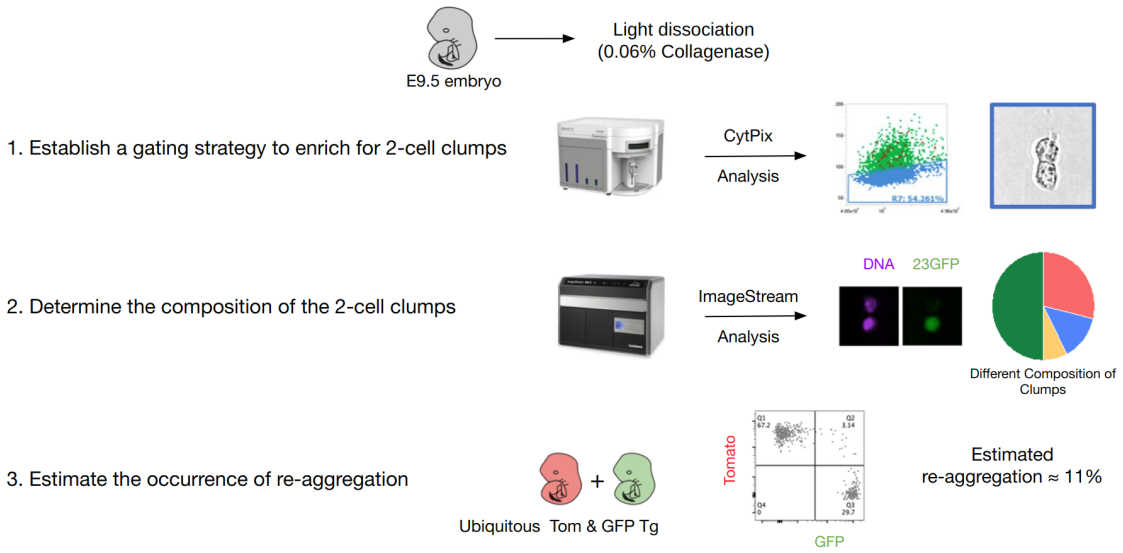


Figure C.1: Pictorial representation of the steps involved in the clump generation experiment: (1) FACS plots of area versus height of DRAQ5 stained samples show a separation between clumps and singlets. CytPix images confirm the clumps were composed of two attached cells; (2) Clumps were gated on $23GFP^{+}CD31^{+}$ expression. ImageStream analysis of fluorescent marker expression confirms a heterogeneous composition of cell type pairs, e.g. one cell is $CD31^{+}23GFP^{-}$ and the other is $CD31^{-}23GFP^{+}$ etc.; (3) We estimated the rates of re-aggregation by sorting doublets from ubiquitous GFP and TOM reporter mice. Created with BioRender.com.

C.2 Estimating Re-aggregation Rates

In the clump generation experiment we expect that clumps of attached cells are captured and sequenced. There are cases where some single cells appear after dissociation, bind to another single cell in suspension and are sequenced together as a pair. We refer to these samples as re-aggregates, as they are not made up of attached cells but may pass quality control. In our experimental study, we sought to estimate the rates of re-aggregation. To this end, we ran the clump gating experiment with ubiquitous GFP and TOM mice (where all cells in the mouse constitutively express the green and red fluorescent proteins, respectively) and measured the fluorescent protein levels in captured cell pairs. We refer to the pairings of attached cells as clumps. The samples expressing only GFP or only TOM will be mixtures of clumps or re-aggregates, but pairings expressing both are certainly re-aggregates. We ran several experiments with varying ratios between GFP and TOM cells (given by ratio of corresponding mice); each experiment resulted in singlets and pairings, that were separated. We used the distribution of fluorescent expression in singlets to estimate the ratio between clumps and re-aggregates in the pairing set.

We made the following assumptions: let p_r be the proportion of TOM *singlets* relative to the GFP ones; in complete absence of re-aggregation we would expect the rates of TOM-only and GFP-only clumps to also be p_r and $1 - p_r$. Hence we assume the rates of TOM-only and GFP-only clumps are p_r and $1 - p_r$ out of the (unknown) amount of clumps in the experiment.

Let N be the total number of samples (clumps + re-aggregates) resulting in the pair gating experiment, and let τ be the proportion of re-aggregates out of N. The samples in the pairing set were divided into 3 categories, whose numbers were normalized to proportions: Q_1 (pairings expressing TOM only), Q_2 (TOM & GFP) and Q_3 (GFP only). As mentioned above, Q_1 and Q_3 are mixtures of clumps and re-aggregates. For the re-aggregation process we assume a Binomial sampling model. Thus, out of τN total re-aggregates we expect the following proportions:

(TOM-TOM, TOM-GFP, GFP-GFP) = $(p_r^2,\ 2p_r(1-p_r),\ (1-p_r)^2)$. We can now write the composition of Q_1, Q_2, Q_3 explicitly:

$$Q_1 = \frac{(1-\tau)Np_r + \tau Np_r^2}{N} = (1-\tau)p_r + \tau p_r^2 \tag{C.1}$$

$$Q_2 = \frac{2N\tau p_r(1-p_r)}{N} = 2\tau p_r(1-p_r) \tag{C.2}$$

$$Q_3 = \frac{(1-\tau)N(1-p_r) + \tau N(1-p_r)^2}{N} = (1-\tau)(1-p_r) + \tau(1-p_r)^2 \tag{C.3}$$

$$\implies \widehat{\tau} = \frac{Q_2}{2p_r(1-p_r)} \tag{C.4}$$

We are also interested in the proportion of true clumps within each fluorescent population:

$$P_r = \frac{(1-\tau)p_r}{(1-\tau)p_r + \tau p_r^2} \tag{C.5}$$

$$P_g = \frac{(1-\tau)(1-p_r)}{(1-\tau)(1-p_r) + \tau(1-p_r)^2} \tag{C.6}$$

Figure C.2 shows the FACS plots for 3 experiments, comprising the distribution of singlets and pairs, shown separately, for distinct ratios between GFP and TOM mice used in the experiment. For singlets we estimated p_r based on values from Q_1 and Q_3 in the FACS plots. The experiment was conducted with different ratios of GFP:TOM mice in order to obtain a broader range of p_r values. Table C.1 shows the re-aggregation estimates from 3 experiments, including the ratio between the GFP and TOM mice, estimated proportion of TOM singlets (p_r), and estimated proportion of TOM and GFP clumps out of each population (P_r and P_g), respectively. Based on these experiments, we estimated the re-aggregation rate τ by averaging the results, yielding 10.6%.

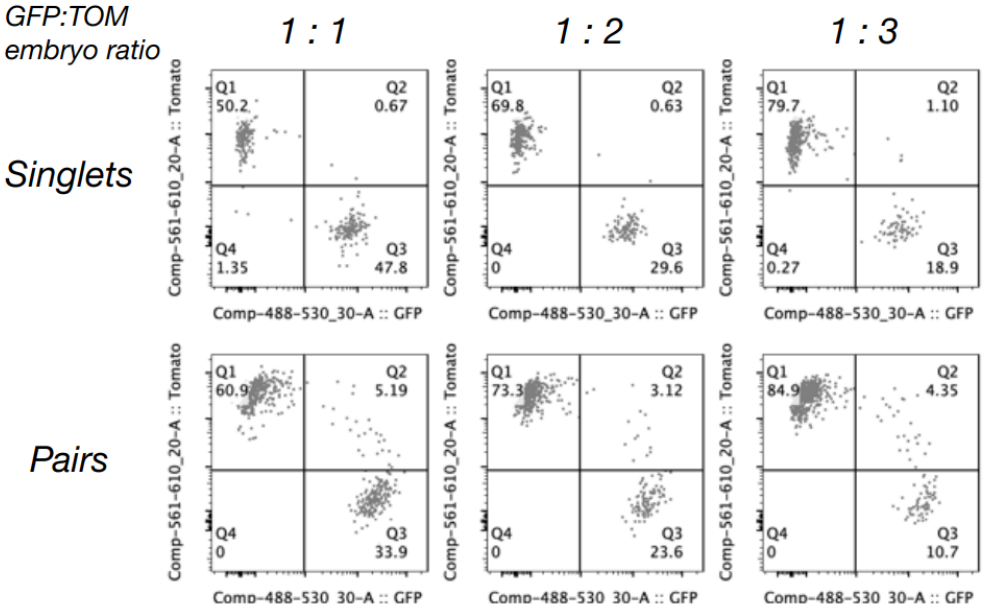


Figure C.2: Cells were sorted from mice with ubiquitous Tomato & GFP expression in different ratios, to evaluate the rate of re-aggregation (step 3 in Figure C.1). FACS plots of Tomato versus GFP fluorescent expression in singlets (top) and pairs (bottom) of cells, showing the proportion of samples in each quadrant ($Q1 - Q4$). Data was generated by Dr. Giorgio Anselmi, de Bruijn Group, MRC WIMM, University of Oxford, unpublished (2023).

GFP:TOM embryo ratio	Singlet p_r	τ	P_r	P_g
1:1	51.2%	10.4%	94.4%	94.6%
1:2	70.2%	7.4%	94.6%	97.7%
1:3	80.8%	14.0%	88.4%	97.0%

Table C.1: Estimates of re-aggregation rates (τ) for the FACS plots in Figure C.2, including the estimated proportion of true Tomato (P_r) and GFP (P_g) clumps.

C.3 Mixing Metrics

To show an example of what the mixing metrics in Section 6.5.1 and Figures 6.14c & 6.14d look like when there is no mixing, we simulate two random sets of data $\boldsymbol{w}_i \sim \mathcal{N}(\boldsymbol{u}_i, I) \in \mathbb{R}^2$, where $\boldsymbol{u}_1 = (0, 0)$ and $\boldsymbol{u}_2 = (1, 1)$, with labels $\{1, 2\}$ and 500 samples each.

A scatter plot of the concatenated dataset is shown in Figure C.3a. We first compute the relative entropy for this dataset, where the size K of the KNN graph is a

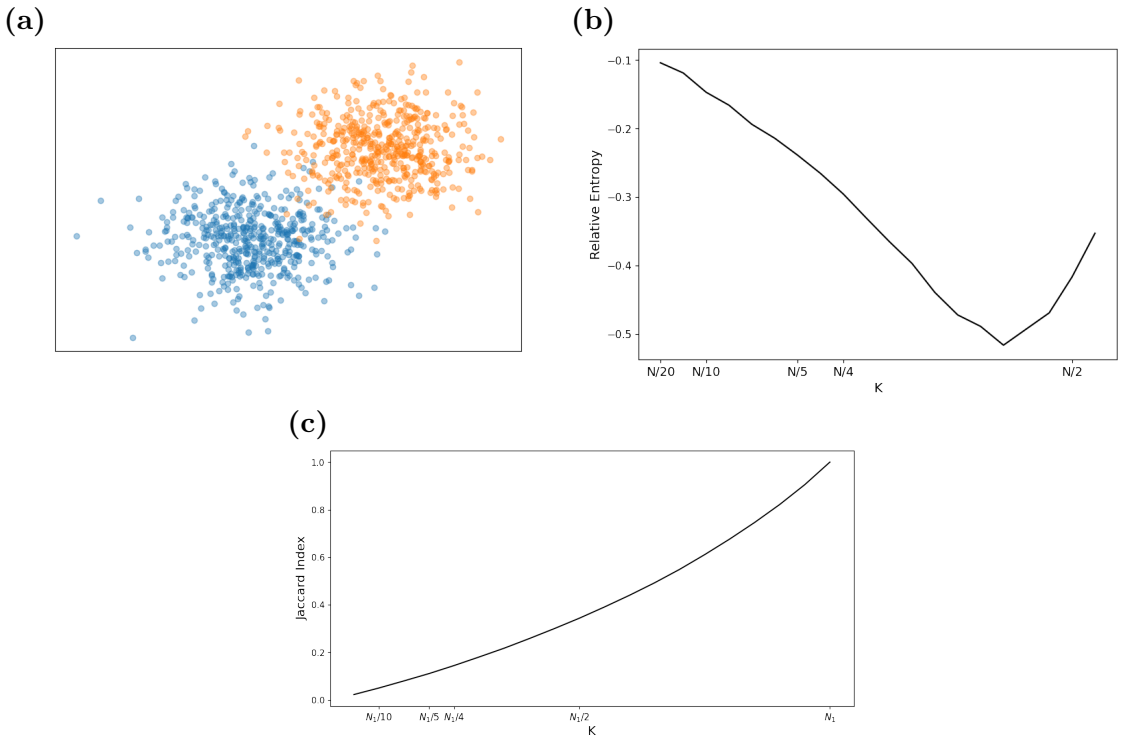


Figure C.3: Mixing metrics for a random dataset with no mixing. (a) example of two clusters in $\mathbb{R}^2$; Mixing metrics plotted as function of data sizes for (b) Relative entropy of batch mixing between the two clusters; (c) Jaccard index between the blue cluster and random permutation of it.

function of data size (here $N = 1000$), shown in Figure C.3b. We note the magnitude of the relative entropy and that it is decreasing as K becomes larger. At round $N/2$ the values are rising; this is because each label has $N/2$ samples, thus for $K > N/2$ the KNN graph contains the entire cluster of the label of the indexed sample.

For the Jaccard index, we need two representations of the same data. To this end, we take the samples of label 1 and permute them at random. The resulting Jaccard index is plotted in Figure C.3c with K as a function of $N_1 = 500$. We note the increasing behaviour is simply due to KNN graph size and the magnitude of values is low $K < N_1/2$.

We compare these values with the ones in Figures 6.14c & 6.14d and we note the values in those Figures are higher, while K is a small proportion of the dataset size.

References

Adler M., Korem Kohanim Y., Tendler A., Mayo A. & Alon U. (2019). "Continuum of Gene-Expression Profiles Provides Spatial Division of Labor within a Differentiated Cell Type". *Cell Systems* 8 (1), 43–52.e5. ISSN: 2405-4712. DOI: 10.1016/j.cels.2018.12.008.

Adler M., Moriel N., Goeva A., Avraham-Davidi I., Mages S., Adams T. S., Kaminski N., Macosko E. Z., Regev A., Medzhitov R. & Nitzan M. (2023). "Emergence of division of labor in tissues through cell interactions and spatial cues". *Cell Reports* 42 (5), p. 112412. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2023.112412.

Ashuach T., Reidenbach D. A., Gayoso A. & Yosef N. (2022). "PeakVI: A deep generative model for single-cell chromatin accessibility analysis". *Cell Reports Methods* 2 (3), p. 100182. ISSN: 2667-2375. DOI: https://doi.org/10.1016/j.crmeth.2022.100182.

Azzoni E., Frontera V., McGrath K. E., Harman J., Carrelha J., Nerlov C., Palis J., Jacobsen S. E. W. & de Bruijn M. F. (2018). "Kit ligand has a critical role in mouse yolk sac and aorta–gonad–mesonephros hematopoiesis". *EMBO reports* 19 (10). DOI: 10.15252/embr.201745477.

Baron C. S., Kester L., Klaus A., Boisset J.-C., Thambyrajah R., Yvernogeau L., Kouskoff V., Lacaud G., van Oudenaarden A. & Robin C. (2018). "Single-cell transcriptomics reveal the dynamic of haematopoietic stem cell production in the aorta". *Nature Communications* 9 (1). DOI: 10.1038/s41467-018-04893-3.

Bee T., Swiers G., Muroi S., Pozner A., Nottingham W., Santos A. C., Li P.-S., Taniuchi I. & Bruijn M. F. T. R. de (2010). "Nonredundant roles for Runx1 alternative promoters reflect their activity at discrete stages of developmental hematopoiesis". *Blood* 115 (15), pp. 3042–3050. ISSN: 1528-0020. DOI: 10.1182/blood-2009-08-238626.

Benjamini Y. & Hochberg Y. (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". *Journal of the Royal Statistical Society: Series B (Methodological)* 57 (1), pp. 289–300. ISSN: 2517-6161. DOI: 10.1111/j.2517-6161.1995.tb02031.x.

Bergen V., Lange M., Peidli S., Wolf F. A. & Theis F. J. (2020). "Generalizing RNA velocity to transient cell states through dynamical modeling". *Nature Biotechnology*. DOI: 10.1038/s41587-020-0591-3.

Bertrand J. Y., Chi N. C., Santoso B., Teng S., Stainier D. Y. R. & Traver D. (2010). "Haematopoietic stem cells derive directly from aortic endothelium during development". *Nature* 464 (7285), pp. 108–111. ISSN: 1476-4687. DOI: 10.1038/nature08738.

Blei D. M., Kucukelbir A. & McAuliffe J. D. (2017). "Variational Inference: A Review for Statisticians". *Journal of the American Statistical Association* 112 (518), pp. 859–877. DOI: 10.1080/01621459.2017.1285773.

Bloom A. B. & Zaman M. H. (2014). "Influence of the microenvironment on cell fate determination and migration". *Physiological Genomics* 46 (9), pp. 309–314. ISSN: 1531-2267. DOI: 10.1152/physiolgenomics.00170.2013.

Boisset J.-C., Cappellen W. van, Andrieu-Soler C., Galjart N., Dzierzak E. & Robin C. (2010). "In vivo imaging of haematopoietic cells emerging from the mouse aortic endothelium". *Nature* 464 (7285), pp. 116–120. ISSN: 1476-4687. DOI: 10.1038/nature08764.

Bowman S., Vilnis L., Vinyals O., Dai A., Jozefowicz R. & Bengio S. (2016). "Generating sentences from a continuous space". English (US). CoNLL 2016 - 20th SIGNLL Conference on Computational Natural Language Learning, Proceedings. 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016 ; Conference date: 11-08-2016 Through 12-08-2016. Association for Computational Linguistics (ACL), pp. 10–21.

Boyeau P., Regier J., Gayoso A., Jordan M. I., Lopez R. & Yosef N. (2023). "An empirical Bayes method for differential expression analysis of single cells with deep generative models". *Proceedings of the National Academy of Sciences* 120 (21). ISSN: 1091-6490. DOI: 10.1073/pnas.2209124120.

Buettner F., Natarajan K. N., Casale F. P., Proserpio V., Scialdone A., Theis F. J., Teichmann S. A., Marioni J. C. & Stegle O. (2015). "Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells". *Nature Biotechnology* 33 (2), pp. 155–160. ISSN: 1546-1696. DOI: 10.1038/nbt.3102.

Cable D. M., Murray E., Zou L. S., Goeva A., Macosko E. Z., Chen F. & Irizarry R. A. (2021). "Robust decomposition of cell type mixtures in spatial transcriptomics". *Nature Biotechnology* 40 (4), pp. 517–526. ISSN: 1546-1696. DOI: 10.1038/s41587-021-00830-w.

Cai Z., Bruijn M. de, Ma X., Dortland B., Luteijn T., Downing J. R. & Dzierzak E. (2000). "Haploinsufficiency of AML1 Affects the Temporal and Spatial Generation of Hematopoietic Stem Cells in the Mouse Embryo". *Immunity* 13 (4), pp. 423–431. ISSN: 1074-7613. DOI: 10.1016/s1074-7613(00)00042-x.

Campbell K. R. & Yau C. (2018). "A descriptive marker gene approach to single-cell pseudotime inference". *Bioinformatics* 35 (1). Ed. by I. Birol, pp. 28–35. DOI: 10.1093/bioinformatics/bty498.

Cao Z.-J. & Gao G. (2022). "Multi-omics single-cell data integration and regulatory inference with graph-linked embedding". *Nature Biotechnology* 40 (10), pp. 1458–1466. ISSN: 1546-1696. DOI: 10.1038/s41587-022-01284-4.

Chen J., Suo S., Tam P. P., Han J.-D. J., Peng G. & Jing N. (2017). "Spatial transcriptomic analysis of cryosectioned tissue samples with Geo-seq". *Nature Protocols* 12 (3), pp. 566–580. DOI: 10.1038/nprot.2017.003.

Chen K. H., Boettiger A. N., Moffitt J. R., Wang S. & Zhuang X. (2015). "Spatially resolved, highly multiplexed RNA profiling in single cells". *Science* 348 (6233). ISSN: 1095-9203. DOI: 10.1126/science.aaa6090.

Chen M. J., Yokomizo T., Zeigler B. M., Dzierzak E. & Speck N. A. (2009). "Runx1 is required for the endothelial to haematopoietic cell transition but not thereafter". *Nature* 457 (7231), pp. 887–891. DOI: 10.1038/nature07619.

Chen R. T. Q., Rubanova Y., Bettencourt J. & Duvenaud D. (2018a). "Neural Ordinary Differential Equations". *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. NIPS'18. Montréal, Canada: Curran Associates Inc., pp. 6572–6583.

Chen X., Teichmann S. A. & Meyer K. B. (2018b). "From Tissues to Cell Types and Back: Single-Cell Gene Expression Analysis of Tissue Architecture". *Annual Review of Biomedical Data Science* 1 (1), pp. 29–51. DOI: 10.1146/annurev-biodatasci-080917-013452.

Chen Y., Wang Y., Chen Y., Cheng Y., Wei Y., Li Y., Wang J., Wei Y., Chan T.-F. & Li Y. (2022a). "Deep autoencoder for interpretable tissue-adaptive deconvolution and cell-type-specific gene analysis". *Nature Communications* 13 (1). ISSN: 2041-1723. DOI: 10.1038/s41467-022-34550-9.

Chen Z., King W. C., Hwang A., Gerstein M. & Zhang J. (2022b). "DeepVelo: Single-cell transcriptomic deep velocity field learning with neural ordinary differential equations". *Science Advances* 8 (48). ISSN: 2375-2548. DOI: 10.1126/sciadv.abq3745.

Colozza G., Park S.-Y. & Koo B.-K. (2022). "Clone wars: From molecules to cell competition in intestinal stem cell homeostasis and disease". *Experimental and Molecular Medicine* 54 (9), pp. 1367–1378. ISSN: 2092-6413. DOI: 10.1038/s12276-022-00854-5.

Coşkun S., Chao H., Vasavada H., Heydari K., Gonzales N., Zhou X., Crombrugghe B. de & Hirschi K. K. (2014). "Development of the Fetal Bone Marrow Niche and Regulation of HSC Quiescence and Homing Ability by Emerging Osteolineage Cells". *Cell Reports* 9 (2), pp. 581–590. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2014.09.013.

de Bruijn M. F., Speck N. A., Peeters M. C. & Dzierzak E. (2000). "Definitive hematopoietic stem cells first develop within the major arterial regions of the mouse embryo". *The EMBO Journal* 19 (11), pp. 2465–2474. DOI: 10.1093/emboj/19.11.2465.

DeTomaso D. & Yosef N. (2021). "Hotspot identifies informative gene modules across modalities of single-cell genomics". *Cell Systems* 12 (5), 446–456.e9. ISSN: 2405-4712. DOI: 10.1016/j.cels.2021.04.005.

Ding J., Condon A. & Shah S. P. (2018). "Interpretable dimensionality reduction of single cell transcriptome data with deep generative models". *Nature Communications* 9 (1). DOI: 10.1038/s41467-018-04368-5.

Dobin A., Davis C. A., Schlesinger F., Drenkow J., Zaleski C., Jha S., Batut P., Chaisson M. & Gingeras T. R. (2012). "STAR: ultrafast universal RNA-seq aligner". *Bioinformatics* 29 (1), pp. 15–21. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bts635.

Domke J. & Sheldon D. R. (2018). "Importance Weighting and Variational Inference". *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi & R. Garnett. Vol. 31. Curran

Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2018/file/25db67c5657914454081c6a18e93d6dd-Paper.pdf.

Emmert-Buck M. R., Bonner R. F., Smith P. D., Chuaqui R. F., Zhuang Z., Goldstein S. R., Weiss R. A. & Liotta L. A. (1996). "Laser Capture Microdissection". *Science* 274 (5289), pp. 998–1001. ISSN: 1095-9203. DOI: 10.1126/science.274.5289.998.

Eraslan G., Simon L. M., Mircea M., Mueller N. S. & Theis F. J. (2019). "Single-cell RNA-seq denoising using a deep count autoencoder". *Nature Communications* 10 (1). DOI: 10.1038/s41467-018-07931-2.

Faure L., Soldatov R., Kharchenko P. V. & Adameyko I. (2022). "scFates: a scalable python package for advanced pseudotime and bifurcation analysis from single-cell data". *Bioinformatics* 39 (1). Ed. by O. Vitek. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btac746.

Femino A. M., Fay F. S., Fogarty K. & Singer R. H. (1998). "Visualization of Single RNA Transcripts in Situ". *Science* 280 (5363), pp. 585–590. ISSN: 1095-9203. DOI: 10.1126/science.280.5363.585.

Foster A., Vezer A., Glastonbury C. A., Creed P., Abujudeh S. & Sim A. (2022). "Contrastive Mixture of Posteriors for Counterfactual Inference, Data Integration and Fairness". *Proceedings of the 39th International Conference on Machine Learning*. Ed. by K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu & S. Sabato. Vol. 162. Proceedings of Machine Learning Research. PMLR, pp. 6578–6621. URL: https://proceedings.mlr.press/v162/foster22a.html.

Gayoso A., Lopez R., Xing G., Boyeau P., Valiollah Pour Amiri V., Hong J., Wu K., Jayasuriya M., Mehlman E., Langevin M., Liu Y., Samaran J., Misrachi G., Nazaret A., Clivio O., Xu C., Ashuach T., Gabitto M., Lotfollahi M., Svensson V., Veiga Beltrame E. da, Kleshchevnikov V., Talavera-López C., Pachter L., Theis F. J., Streets A., Jordan M. I., Regier J. & Yosef N. (2022). "A Python library for probabilistic analysis of single-cell omics data". *Nature Biotechnology* 40 (2), pp. 163–166. ISSN: 1546-1696. DOI: 10.1038/s41587-021-01206-w.

Gayoso A., Steier Z., Lopez R., Regier J., Nazor K. L., Streets A. & Yosef N. (2021). "Joint probabilistic modeling of single-cell multi-omic data with totalVI". *Nature Methods* 18 (3), pp. 272–282. ISSN: 1548-7105. DOI: 10.1038/s41592-020-01050-x.

Gayoso A., Weiler P., Lotfollahi M., Klein D., Hong J., Streets A., Theis F. J. & Yosef N. (2023). "Deep generative modeling of transcriptional dynamics for RNA velocity analysis in single cells". *Nature Methods* 21 (1), pp. 50–59. ISSN: 1548-7105. DOI: 10.1038/s41592-023-01994-w.

Gehart H. & Clevers H. (2018). "Tales from the crypt: new insights into intestinal stem cells". *Nature Reviews Gastroenterology and Hepatology* 16 (1), pp. 19–34. ISSN: 1759-5053. DOI: 10.1038/s41575-018-0081-y.

Gehart H., Es J. H. van, Hamer K., Beumer J., Kretzschmar K., Dekkers J. F., Rios A. & Clevers H. (2019). "Identification of Enteroendocrine Regulators by Real-Time Single-Cell Differentiation Mapping". *Cell* 176 (5), 1158–1173.e16. ISSN: 0092-8674. DOI: 10.1016/j.cell.2018.12.029.

Gershman S. J. & Goodman N. D. (2014). "Amortized Inference in Probabilistic Reasoning". *Cognitive Science* 36. URL: https://api.semanticscholar.org/CorpusID:924780.

Glass L. L., Calero-Nieto F. J., Jawaid W., Larraufie P., Kay R. G., Göttgens B., Reimann F. & Gribble F. M. (2017). "Single-cell RNA-sequencing reveals a distinct population of proglucagon-expressing cells specific to the mouse upper small intestine". *Molecular Metabolism* 6 (10), pp. 1296–1303. ISSN: 2212-8778. DOI: 10.1016/j.molmet.2017.07.014.

Goodfellow I., Bengio Y. & Courville A. (2016). *Deep Learning*. http://www.deeplearningbook.org. MIT Press.

Grønbech C. H., Vording M. F., Timshel P., Sønderby C. K., Pers T. H. & Winther O. (2020). "scVAE: Variational auto-encoders for single-cell gene expression data". *Bioinformatics*. DOI: 10.1093/bioinformatics/btaa293.

Grün D., Kester L. & Oudenaarden A. van (2014). "Validation of noise models for single-cell transcriptomics". *Nature Methods* 11 (6), pp. 637–640. ISSN: 1548-7105. DOI: 10.1038/nmeth.2930.

Grün D., Lyubimova A., Kester L., Wiebrands K., Basak O., Sasaki N., Clevers H. & Oudenaarden A. van (2015). "Single-cell messenger RNA sequencing reveals rare intestinal cell types". *Nature* 525 (7568), pp. 251–255. ISSN: 1476-4687. DOI: 10.1038/nature14966.

Gu Y., Blaauw D. & Welch J. D. (2022). "Bayesian Inference of RNA Velocity from Multi-Lineage Single-Cell Data". DOI: 10.1101/2022.07.08.499381.

Haber A. L., Biton M., Rogel N., Herbst R. H., Shekhar K., Smillie C., Burgin G., Delorey T. M., Howitt M. R., Katz Y., Tirosh I., Beyaz S., Dionne D., Zhang M., Raychowdhury R., Garrett W. S., Rozenblatt-Rosen O., Shi H. N., Yilmaz O., Xavier R. J. & Regev A. (2017). "A single-cell survey of the small intestinal epithelium". *Nature* 551 (7680), pp. 333–339. DOI: 10.1038/nature24489.

Hagemann-Jensen M., Ziegenhain C., Chen P., Ramsköld D., Hendriks G.-J., Larsson A. J. M., Faridani O. R. & Sandberg R. (2020). "Single-cell RNA counting at allele and isoform resolution using Smart-seq3". *Nature Biotechnology* 38 (6), pp. 708–714. ISSN: 1546-1696. DOI: 10.1038/s41587-020-0497-0.

Haghverdi L., Büttner M., Wolf F. A., Buettner F. & Theis F. J. (2016). "Diffusion pseudotime robustly reconstructs lineage branching". *Nature Methods* 13 (10), pp. 845–848. ISSN: 1548-7105. DOI: 10.1038/nmeth.3971.

Haghverdi L., Lun A. T. L., Morgan M. D. & Marioni J. C. (2018). "Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors". *Nature Biotechnology* 36 (5), pp. 421–427. DOI: 10.1038/nbt.4091.

Halpern K. B., Shenhav R., Massalha H., Toth B., Egozi A., Massasa E. E., Medgalia C., David E., Giladi A., Moor A. E., Porat Z., Amit I. & Itzkovitz S. (2018). "Paired-cell sequencing enables spatial gene expression mapping of liver endothelial cells". *Nature Biotechnology* 36 (10), pp. 962–970. DOI: 10.1038/nbt.4231.

Hashimshony T., Senderovich N., Avital G., Klochendler A., Leeuw Y. de, Anavy L., Gennert D., Li S., Livak K. J., Rozenblatt-Rosen O., Dor Y., Regev A. & Yanai I. (2016). "CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq". *Genome Biology* 17 (1). ISSN: 1474-760X. DOI: 10.1186/s13059-016-0938-8.

Hashimshony T., Wagner F., Sher N. & Yanai I. (2012). "CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification". *Cell Reports* 2 (3), pp. 666–673. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2012.08.003.

Higgins I., Matthey L., Pal A., Burgess C., Glorot X., Botvinick M., Mohamed S. & Lerchner A. (2017). "beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework". *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=Sy2fzU9gl.

Hou S., Li Z., Dong J., Gao Y., Chang Z., Ding X., Li S., Li Y., Zeng Y., Xin Q., Wang B., Ni Y., Ning X., Hu Y., Fan X., Hou Y., Li X., Wen L., Zhou B., Liu B., Tang F. & Lan Y. (2022). "Heterogeneity in endothelial cells and widespread venous arterialization during early vascular development in mammals". *Cell Research* 32 (4), pp. 333–348. ISSN: 1748-7838. DOI: 10.1038/s41422-022-00615-z.

Hou S., Li Z., Zheng X., Gao Y., Dong J., Ni Y., Wang X., Li Y., Ding X., Chang Z., Li S., Hu Y., Fan X., Hou Y., Wen L., Liu B., Tang F. & Lan Y. (2020). "Embryonic endothelial evolution towards first hematopoietic stem cells revealed by single-cell transcriptomic and functional analyses". *Cell Research* 30 (5), pp. 376–392. DOI: 10.1038/s41422-020-0300-2.

Ioffe S. & Szegedy C. (2015). "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*. ICML'15. Lille, France: JMLR.org, pp. 448–456.

Islam S., Zeisel A., Joost S., La Manno G., Zajac P., Kasper M., Lönnerberg P. & Linnarsson S. (2013). "Quantitative single-cell RNA-seq with unique molecular identifiers". *Nature Methods* 11 (2), pp. 163–166. DOI: 10.1038/nmeth.2772.

Jaffredo T., Gautier R., Eichmann A. & Dieterlen-Lièvre F. (1998). "Intraaortic hemopoietic cells are derived from endothelial cells during ontogeny". *Development* 125 (22), pp. 4575–4583. ISSN: 1477-9129. DOI: 10.1242/dev.125.22.4575.

Jaitin D. A., Kenigsberg E., Keren-Shaul H., Elefant N., Paul F., Zaretsky I., Mildner A., Cohen N., Jung S., Tanay A. & Amit I. (2014). "Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types". *Science* 343 (6172), pp. 776–779. ISSN: 1095-9203. DOI: 10.1126/science.1247651.

Johnson W. E., Li C. & Rabinovic A. (2006). "Adjusting batch effects in microarray expression data using empirical Bayes methods". *Biostatistics* 8 (1), pp. 118–127. ISSN: 1465-4644. DOI: 10.1093/biostatistics/kxj037.

Jordan M. I., Ghahramani Z., Jaakkola T. S. & Saul L. K. (1999). "An Introduction to Variational Methods for Graphical Models". *Machine Learning* 37 (2), pp. 183–233. ISSN: 0885-6125. DOI: 10.1023/A:1007665907178.

Kingma D. P. & Welling M. (2013). "Auto-Encoding Variational Bayes". *arXiv*. URL: https://arxiv.org/abs/1312.6114v10.

Kingma D. P. & Ba J. (2015). "Adam: A Method for Stochastic Optimization". *3rd International Conference on Learning Representations*. URL: https://arxiv.org/abs/1412.6980.

Kissa K. & Herbomel P. (2010). "Blood stem cells emerge from aortic endothelium by a novel type of cell transition". *Nature* 464 (7285), pp. 112–115. ISSN: 1476-4687. DOI: 10.1038/nature08761.

Kolberg L., Raudvere U., Kuzmin I., Adler P., Vilo J. & Peterson H. (2023). "g:Profiler—interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update)". *Nucleic Acids Research* 51 (W1), W207–W212. ISSN: 1362-4962. DOI: 10.1093/nar/gkad347.

Kolodziejczyk A. A., Kim J. K., Svensson V., Marioni J. C. & Teichmann S. A. (2015). "The Technology and Biology of Single-Cell RNA Sequencing". *Molecular Cell* 58 (4), pp. 610–620. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2015.04.005.

Korsunsky I., Millard N., Fan J., Slowikowski K., Zhang F., Wei K., Baglaenko Y., Brenner M., Loh P.-r. & Raychaudhuri S. (2019). "Fast, sensitive and accurate integration of single-cell data with Harmony". *Nature Methods* 16 (12), pp. 1289–1296. ISSN: 1548-7105. DOI: 10.1038/s41592-019-0619-0.

La Manno G., Soldatov R., Zeisel A., Braun E., Hochgerner H., Petukhov V., Lidschreiber K., Kastriti M. E., Lönnerberg P., Furlan A., Fan J., Borm L. E., Liu Z., Bruggen D. van, Guo J., He X., Barker R., Sundström E., Castelo-Branco G., Cramer P., Adameyko I., Linnarsson S. & Kharchenko P. V. (2018). "RNA velocity of single cells". *Nature* 560 (7719), pp. 494–498. DOI: 10.1038/s41586-018-0414-6.

Lönnberg T., Svensson V., James K. R., Fernandez-Ruiz D., Sebina I., Montandon R., Soon M. S. F., Fogg L. G., Nair A. S., Liligeto U. N., Stubbington M. J. T., Ly L.-H., Bagger F. O., Zwiessele M., Lawrence N. D., Souza-Fonseca-Guimaraes F., Bunn P. T., Engwerda C. R., Heath W. R., Billker O., Stegle O., Haque A. & Teichmann S. A. (2017). "Single-cell RNA-seq and computational analysis using temporal mixture modeling resolves TH1/TFHfate bifurcation in malaria". *Science Immunology* 2 (9), eaal2192. DOI: 10.1126/sciimmunol.aal2192.

Lopez R., Boyeau P., Yosef N., Jordan M. I. & Regier J. (2020). "Decision-making with auto-encoding variational Bayes". *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS '20. , Vancouver, BC, Canada, Curran Associates Inc. ISBN: 9781713829546.

Lopez R., Li B., Keren-Shaul H., Boyeau P., Kedmi M., Pilzer D., Jelinski A., Yofe I., David E., Wagner A., Ergen C., Addadi Y., Golani O., Ronchese F., Jordan M. I., Amit I. & Yosef N. (2022). "DestVI identifies continuums of cell types in spatial transcriptomics data". *Nature Biotechnology* 40 (9), pp. 1360–1369. DOI: 10.1038/s41587-022-01272-8.

Lopez R., Nazaret A., Langevin M., Samaran J., Regier J., Jordan M. I. & Yosef N. (2019). "A joint model of unpaired data from scRNA-seq and spatial transcriptomics for imputing missing gene expression measurements". *ICML Workshop in Computational Biology*.

Lopez R., Regier J., Cole M. B., Jordan M. I. & Yosef N. (2018). "Deep generative modeling for single-cell transcriptomics". *Nature Methods* 15 (12), pp. 1053–1058. DOI: 10.1038/s41592-018-0229-2.

Lotfollahi M., Klimovskaia Susmelj A., De Donno C., Hetzel L., Ji Y., Ibarra I. L., Srivatsan S. R., Naghipourfar M., Daza R. M., Martin B., Shendure J.,

McFaline-Figueroa J. L., Boyeau P., Wolf F. A., Yakubova N., Günnemann S., Trapnell C., Lopez-Paz D. & Theis F. J. (2023). "Predicting cellular responses to complex perturbations in high-throughput screens". *Molecular Systems Biology* 19 (6). ISSN: 1744-4292. DOI: 10.15252/msb.202211517.

Lotfollahi M., Naghipourfar M., Luecken M. D., Khajavi M., Büttner M., Wagenstetter M., Avsec Ž., Gayoso A., Yosef N., Interlandi M., Rybakov S., Misharin A. V. & Theis F. J. (2021). "Mapping single-cell data to reference atlases by transfer learning". *Nature Biotechnology* 40 (1), pp. 121–130. ISSN: 1546-1696. DOI: 10.1038/s41587-021-01001-7.

Lotfollahi M., Naghipourfar M., Theis F. J. & Wolf F. A. (2020). "Conditional out-of-distribution generation for unpaired data using transfer VAE". *Bioinformatics* 36 (Supplement˙2), pp. i610–i617. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btaa800.

Lubeck E., Coskun A. F., Zhiyentayev T., Ahmad M. & Cai L. (2014). "Single-cell in situ RNA profiling by sequential hybridization". *Nature Methods* 11 (4), pp. 360–361. ISSN: 1548-7105. DOI: 10.1038/nmeth.2892.

Luecken M. D. & Theis F. J. (2019). "Current best practices in single-cell RNA-seq analysis: a tutorial". *Molecular Systems Biology* 15 (6). ISSN: 1744-4292. DOI: 10.15252/msb.20188746.

Macosko E. Z., Basu A., Satija R., Nemesh J., Shekhar K., Goldman M., Tirosh I., Bialas A. R., Kamitaki N., Martersteck E. M., Trombetta J. J., Weitz D. A., Sanes J. R., Shalek A. K., Regev A. & McCarroll S. A. (2015). "Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets". *Cell* 161 (5), pp. 1202–1214. ISSN: 0092-8674. DOI: 10.1016/j.cell.2015.05.002.

Manco R., Averbukh I., Porat Z., Bahar Halpern K., Amit I. & Itzkovitz S. (2021). "Clump sequencing exposes the spatial expression programs of intestinal secretory cells". *Nature Communications* 12 (1). ISSN: 2041-1723. DOI: 10.1038/s41467-021-23245-2.

Märtens K. & Yau C. (2020). "BasisVAE: Translation-invariant feature-level clustering with Variational Autoencoders". *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. Ed. by S. Chiappa & R. Calandra. Vol. 108. Proceedings of Machine Learning Research. PMLR, pp. 2928–2937. URL: https://proceedings.mlr.press/v108/martens20b.html.

Mathieu E., Rainforth T., Siddharth N. & Teh Y. W. (2019). "Disentangling Disentanglement in Variational Autoencoders". *Proceedings of the 36th International Conference on Machine Learning*. Ed. by K. Chaudhuri & R. Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 4402–4412. URL: https://proceedings.mlr.press/v97/mathieu19a.html.

Mazutis L., Gilbert J., Ung W. L., Weitz D. A., Griffiths A. D. & Heyman J. A. (2013). "Single-cell analysis and sorting using droplet-based microfluidics". *Nature Protocols* 8 (5), pp. 870–891. ISSN: 1750-2799. DOI: 10.1038/nprot.2013.046.

McInnes L., Healy J., Saul N. & Großberger L. (2018). "UMAP: Uniform Manifold Approximation and Projection". *Journal of Open Source Software* 3 (29), p. 861. DOI: 10.21105/joss.00861.

Medvinsky A. & Dzierzak E. (1996). "Definitive Hematopoiesis Is Autonomously Initiated by the AGM Region". *Cell* 86 (6), pp. 897–906. DOI: 10.1016/S0092-8674(00)80165-8.

Menden K., Marouf M., Oller S., Dalmia A., Magruder D. S., Kloiber K., Heutink P. & Bonn S. (2020). "Deep learning–based cell composition analysis from tissue expression profiles". *Science Advances* 6 (30). ISSN: 2375-2548. DOI: 10.1126/sciadv.aba2619.

Moor A. E., Harnik Y., Ben-Moshe S., Massasa E. E., Rozenberg M., Eilam R., Halpern K. B. & Itzkovitz S. (2018). "Spatial Reconstruction of Single Enterocytes Uncovers Broad Zonation along the Intestinal Villus Axis". *Cell* 175 (4), 1156–1167.e15. DOI: 10.1016/j.cell.2018.08.063.

Moses L. & Pachter L. (2022). "Museum of spatial transcriptomics". *Nature Methods* 19 (5), pp. 534–546. ISSN: 1548-7105. DOI: 10.1038/s41592-022-01409-2.

Müller A. M., Medvinsky A., Strouboulis J., Grosveld F. & Dzierzakt E. (1994). "Development of hematopoietic stem cell activity in the mouse embryo". *Immunity* 1 (4), pp. 291–301. ISSN: 1074-7613. DOI: 10.1016/1074-7613(94)90081-7.

North T., Gu T.-L., Stacy T., Wang Q., Howard L., Binder M., Marín-Padilla M. & Speck N. A. (1999). "Cbfa2 is required for the formation of intra-aortic hematopoietic clusters". *Development* 126 (11), pp. 2563–2575. ISSN: 1477-9129. DOI: 10.1242/dev.126.11.2563.

Nottingham W. T., Jarratt A., Burgess M., Speck C. L., Cheng J.-F., Prabhakar S., Rubin E. M., Li P.-S., Sloane-Stanley J., Kong-a-San J. & Bruijn M. F. T. R. de (2007). "Runx1-mediated hematopoietic stem-cell emergence is controlled by a Gata/Ets/SCL-regulated enhancer". *Blood* 110 (13), pp. 4188–4197. ISSN: 1528-0020. DOI: 10.1182/blood-2007-07-100883.

Okuda T., Deursen J. van, Hiebert S. W., Grosveld G. & Downing J. R. (1996). "AML1, the Target of Multiple Chromosomal Translocations in Human Leukemia, Is Essential for Normal Fetal Liver Hematopoiesis". *Cell* 84 (2), pp. 321–330. ISSN: 0092-8674. DOI: 10.1016/s0092-8674(00)80986-1.

Ottersbach K. (2019). "Endothelial-to-haematopoietic transition: an update on the process of making blood". *Biochemical Society Transactions* 47 (2), pp. 591–601. DOI: 10.1042/BST20180320.

Pabst O. & Slack E. (2020). "IgA and the intestinal microbiota: the importance of being specific". *Mucosal Immunology* 13 (1), pp. 12–21. ISSN: 1933-0219. DOI: 10.1038/s41385-019-0227-4.

Palis J., Robertson S., Kennedy M., Wall C. & Keller G. (1999). "Development of erythroid and myeloid progenitors in the yolk sac and embryo proper of the mouse". *Development* 126 (22), pp. 5073–5084. ISSN: 1477-9129. DOI: 10.1242/dev.126.22.5073.

Parekh S., Ziegenhain C., Vieth B., Enard W. & Hellmann I. (2018). "zUMIs - A fast and flexible pipeline to process RNA sequencing data with UMIs". *GigaScience* 7 (6). ISSN: 2047-217X. DOI: 10.1093/gigascience/giy059.

Peck B. C. E., Shanahan M. T., Singh A. P. & Sethupathy P. (2017). "Gut Microbial Influences on the Mammalian Intestinal Stem Cell Niche". *Stem Cells International* 2017, pp. 1–17. ISSN: 1687-9678. DOI: 10.1155/2017/5604727.

Picelli S., Faridani O. R., Björklund Å. K., Winberg G., Sagasser S. & Sandberg R. (2014). "Full-length RNA-seq from single cells using Smart-seq2". *Nature Protocols* 9 (1), pp. 171–181. ISSN: 1750-2799. DOI: 10.1038/nprot.2014.006.

Pierson E. & Yau C. (2015). "ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis". *Genome Biology* 16 (1). DOI: 10.1186/s13059-015-0805-z.

Prabhakaran S., Azizi E., Carr A. & Pe'er D. (2016). "Dirichlet Process Mixture Model for Correcting Technical Variation in Single-Cell Gene Expression Data". *Proc. 33nd Int. Conf. Mach. Learn., ICML 2016* 48, pp. 1070–1079.

Qiao C. & Huang Y. (2021). "Representation learning of RNA velocity reveals robust cell transitions". *Proceedings of the National Academy of Sciences* 118 (49). ISSN: 1091-6490. DOI: 10.1073/pnas.2105859118.

Quail D. F. & Joyce J. A. (2013). "Microenvironmental regulation of tumor progression and metastasis". *Nature Medicine* 19 (11), pp. 1423–1437. ISSN: 1546-170X. DOI: 10.1038/nm.3394.

Raj A. & Oudenaarden A. van (2008). "Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences". *Cell* 135 (2), pp. 216–226. ISSN: 0092-8674. DOI: 10.1016/j.cell.2008.09.050.

Raj A., van den Bogaard P., Rifkin S. A., van Oudenaarden A. & Tyagi S. (2008). "Imaging individual mRNA molecules using multiple singly labeled probes". *Nature Methods* 5 (10), pp. 877–879. DOI: 10.1038/nmeth.1253.

Ramos P., Casu C., Gardenghi S., Breda L., Crielaard B. J., Guy E., Marongiu M. F., Gupta R., Levine R. L., Abdel-Wahab O., Ebert B. L., Van Rooijen N., Ghaffari S., Grady R. W., Giardina P. J. & Rivella S. (2013). "Macrophages support pathological erythropoiesis in polycythemia vera and β-thalassemia". *Nature Medicine* 19 (4), pp. 437–445. ISSN: 1546-170X. DOI: 10.1038/nm.3126.

Ramsköld D., Luo S., Wang Y.-C., Li R., Deng Q., Faridani O. R., Daniels G. A., Khrebtukova I., Loring J. F., Laurent L. C., Schroth G. P. & Sandberg R. (2012). "Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells". *Nature Biotechnology* 30 (8), pp. 777–782. ISSN: 1546-1696. DOI: 10.1038/nbt.2282.

Rao A., Barkley D., França G. S. & Yanai I. (2021). "Exploring tissue architecture using spatial transcriptomics". *Nature* 596 (7871), pp. 211–220. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03634-9.

Rehwinkel J. & Gack M. U. (2020). "RIG-I-like receptors: their regulation and roles in RNA sensing". *Nature Reviews Immunology* 20 (9), pp. 537–551. ISSN: 1474-1741. DOI: 10.1038/s41577-020-0288-3.

Rezende D. J., Mohamed S. & Wierstra D. (2014). "Stochastic Backpropagation and Approximate Inference in Deep Generative Models". *Proceedings of the 31st International Conference on Machine Learning*. Ed. by E. P. Xing & T. Jebara. Vol. 32. Proceedings of Machine Learning Research 2. Bejing, China: PMLR, pp. 1278–1286. URL: https://proceedings.mlr.press/v32/rezende14.html.

Risso D., Perraudeau F., Gribkova S., Dudoit S. & Vert J.-P. (2018). "A general and flexible method for signal extraction from single-cell RNA-seq data". *Nature Communications* 9 (1). DOI: 10.1038/s41467-017-02554-5.

Rousseeuw P. J. (1987). "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis". *Journal of Computational and Applied Mathematics* 20, pp. 53–65. ISSN: 0377-0427. DOI: 10.1016/0377-0427(87)90125-7.

Rubin D. B. (1984). "Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician". *The Annals of Statistics* 12 (4). ISSN: 0090-5364. DOI: 10.1214/aos/1176346785.

Rybtsov S., Batsivari A., Bilotkach K., Paruzina D., Senserrich J., Nerushev O. & Medvinsky A. (2014). "Tracing the Origin of the HSC Hierarchy Reveals an SCF-Dependent, IL-3-Independent CD43- Embryonic Precursor". *Stem Cell Reports* 3 (3), pp. 489–501. DOI: 10.1016/j.stemcr.2014.07.009.

Rybtsov S., Sobiesiak M., Taoudi S., Souilhol C., Senserrich J., Liakhovitskaia A., Ivanovs A., Frampton J., Zhao S. & Medvinsky A. (2011). "Hierarchical organization and early hematopoietic specification of the developing HSC lineage in the AGM region". *Journal of Experimental Medicine* 208 (6), pp. 1305–1315. DOI: 10.1084/jem.20102419.

Saiselet M., Rodrigues-Vitória J., Tourneur A., Craciun L., Spinette A., Larsimont D., Andry G., Lundeberg J., Maenhaut C. & Detours V. (2020). "Transcriptional output, cell-type densities, and normalization in spatial transcriptomics". *Journal of Molecular Cell Biology* 12 (11). Ed. by Z. Wang, pp. 906–908. ISSN: 1759-4685. DOI: 10.1093/jmcb/mjaa028.

Scholz F. W. & Stephens M. A. (1987). "K-Sample Anderson–Darling Tests". *Journal of the American Statistical Association* 82 (399), pp. 918–924. ISSN: 1537-274X. DOI: 10.1080/01621459.1987.10478517.

Shalek A. K., Satija R., Adiconis X., Gertner R. S., Gaublomme J. T., Raychowdhury R., Schwartz S., Yosef N., Malboeuf C., Lu D., Trombetta J. J., Gennert D., Gnirke A., Goren A., Hacohen N., Levin J. Z., Park H. & Regev A. (2013). "Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells". *Nature* 498 (7453), pp. 236–240. ISSN: 1476-4687. DOI: 10.1038/nature12172.

Singer S. J. (1992). "Intercellular Communication and Cell-Cell Adhesion". *Science* 255 (5052), pp. 1671–1677. ISSN: 1095-9203. DOI: 10.1126/science.1313187.

Sohn K., Lee H. & Yan X. (2015). "Learning Structured Output Representation using Deep Conditional Generative Models". *Advances in Neural Information Processing Systems*. Ed. by C. Cortes, N. Lawrence, D. Lee, M. Sugiyama & R. Garnett. Vol. 28. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2015/file/8d55a249e6baa5c06772297520da2051-Paper.pdf.

Souilhol C., Gonneau C., Lendinez J. G., Batsivari A., Rybtsov S., Wilson H., Morgado-Palacin L., Hills D., Taoudi S., Antonchuk J., Zhao S. & Medvinsky A. (2016). "Inductive interactions mediated by interplay of asymmetric signalling underlie development of adult haematopoietic stem cells". *Nature Communications* 7 (1). DOI: 10.1038/ncomms10784.

Srivastava N., Hinton G., Krizhevsky A., Sutskever I. & Salakhutdinov R. (2014). "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". *Journal of Machine Learning Research* 15 (56), pp. 1929–1958. URL: http://jmlr.org/papers/v15/srivastava14a.html.

Ståhl P. L., Salmén F., Vickovic S., Lundmark A., Navarro J. F., Magnusson J., Giacomello S., Asp M., Westholm J. O., Huss M., Mollbrink A., Linnarsson S., Codeluppi S., Borg Å., Pontén F., Costea P. I., Sahlén P., Mulder J., Bergmann O., Lundeberg J. & Frisén J. (2016). "Visualization and analysis of gene expression in tissue sections by spatial transcriptomics". *Science* 353 (6294), pp. 78–82. DOI: 10.1126/science.aaf2403.

Stegle O., Teichmann S. A. & Marioni J. C. (2015). "Computational and analytical challenges in single-cell transcriptomics". *Nature Reviews Genetics* 16 (3), pp. 133–145. ISSN: 1471-0064. DOI: 10.1038/nrg3833.

Street K., Risso D., Fletcher R. B., Das D., Ngai J., Yosef N., Purdom E. & Dudoit S. (2018). "Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics". *BMC Genomics* 19 (1). ISSN: 1471-2164. DOI: 10.1186/s12864-018-4772-0.

Stuart T., Butler A., Hoffman P., Hafemeister C., Papalexi E., Mauck W. M., Hao Y., Stoeckius M., Smibert P. & Satija R. (2019). "Comprehensive Integration of Single-Cell Data". *Cell* 177 (7), 1888–1902.e21. ISSN: 0092-8674. DOI: 10.1016/j.cell.2019.05.031.

Svensson V., Vento-Tormo R. & Teichmann S. A. (2018). "Exponential scaling of single-cell RNA-seq in the past decade". *Nature Protocols* 13 (4), pp. 599–604. DOI: 10.1038/nprot.2017.149.

Swiers G., Baumann C., O'Rourke J., Giannoulatou E., Taylor S., Joshi A., Moignard V., Pina C., Bee T., Kokkaliaris K. D., Yoshimoto M., Yoder M. C., Frampton J., Schroeder T., Enver T., Göttgens B. & de Bruijn M. F. T. R. (2013a). "Early dynamic fate changes in haemogenic endothelium characterized at the single-cell level". *Nature Communications* 4 (1). DOI: 10.1038/ncomms3924.

Swiers G., Rode C., Azzoni E. & Bruijn M. F. de (2013b). "A short history of hemogenic endothelium". *Blood Cells, Molecules, and Diseases* 51 (4), pp. 206–212. ISSN: 1079-9796. DOI: 10.1016/j.bcmd.2013.09.005.

Tang F., Barbacioru C., Wang Y., Nordman E., Lee C., Xu N., Wang X., Bodeau J., Tuch B. B., Siddiqui A., Lao K. & Surani M. A. (2009). "mRNA-Seq whole-transcriptome analysis of a single cell". *Nature Methods* 6 (5), pp. 377–382. DOI: 10.1038/nmeth.1315.

Taoudi S. & Medvinsky A. (2007). "Functional identification of the hematopoietic stem cell niche in the ventral domain of the embryonic dorsal aorta". *Proceedings of the National Academy of Sciences* 104 (22), pp. 9399–9403. DOI: 10.1073/pnas.0700984104.

Taoudi S., Gonneau C., Moore K., Sheridan J. M., Blackburn C. C., Taylor E. & Medvinsky A. (2008). "Extensive Hematopoietic Stem Cell Generation in the AGM Region via Maturation of VE-Cadherin+CD45+ Pre-Definitive HSCs". *Cell Stem Cell* 3 (1), pp. 99–108. ISSN: 1934-5909. DOI: 10.1016/j.stem.2008.06.004.

Tirosh I., Izar B., Prakadan S. M., Wadsworth M. H., Treacy D., Trombetta J. J., Rotem A., Rodman C., Lian C., Murphy G., Fallahi-Sichani M., Dutton-Regester K., Lin J.-R., Cohen O., Shah P., Lu D., Genshaft A. S., Hughes T. K., Ziegler C. G. K., Kazer S. W., Gaillard A., Kolb K. E., Villani A.-C., Johannessen C. M., Andreev A. Y., Van Allen E. M., Bertagnolli M., Sorger P. K., Sullivan R. J., Flaherty K. T., Frederick D. T., Jané-Valbuena J., Yoon C. H., Rozenblatt-Rosen O., Shalek A. K., Regev A. & Garraway L. A. (2016). "Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq". *Science* 352 (6282), pp. 189–196. ISSN: 1095-9203. DOI: 10.1126/science.aad0501.

Trapnell C., Cacchiarelli D., Grimsby J., Pokharel P., Li S., Morse M., Lennon N. J., Livak K. J., Mikkelsen T. S. & Rinn J. L. (2014). "The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells". *Nature Biotechnology* 32 (4), pp. 381–386. ISSN: 1546-1696. DOI: 10.1038/nbt.2859.

Trimm E. & Red-Horse K. (2022). "Vascular endothelial cell development and diversity". *Nature Reviews Cardiology* 20 (3), pp. 197–210. ISSN: 1759-5010. DOI: 10.1038/s41569-022-00770-1.

Tsang D. K., Wang R. J., De Sa O., Ayyaz A., Foerster E. G., Bayer G., Goyal S., Trcka D., Ghoshal B., Wrana J. L., Girardin S. E. & Philpott D. J. (2022). "A single cell survey of the microbial impacts on the mouse small intestinal epithelium". *Gut Microbes* 14 (1). ISSN: 1949-0984. DOI: 10.1080/19490976.2022.2108281.

Tung P.-Y., Blischak J. D., Hsiao C. J., Knowles D. A., Burnett J. E., Pritchard J. K. & Gilad Y. (2017). "Batch effects and the effective design of single-cell gene expression studies". *Scientific Reports* 7 (1). ISSN: 2045-2322. DOI: 10.1038/srep39921.

Turner R. E. & Sahani M. (2011). "Two problems with variational expectation maximisation for time-series models". URL: https://api.semanticscholar.org/CorpusID:18708421.

Vallejos C. A., Risso D., Scialdone A., Dudoit S. & Marioni J. C. (2017). "Normalizing single-cell RNA sequencing data: challenges and opportunities". *Nature Methods* 14 (6), pp. 565–571. DOI: 10.1038/nmeth.4292.

Vallejos C. A., Marioni J. C. & Richardson S. (2015). "BASiCS: Bayesian Analysis of Single-Cell Sequencing Data". *PLOS Computational Biology* 11 (6), e1004333. DOI: 10.1371/journal.pcbi.1004333.

van den Brink S. C., Sage F., Vértesy Á., Spanjaard B., Peterson-Maduro J., Baron C. S., Robin C. & Oudenaarden A. van (2017). "Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations". *Nature Methods* 14 (10), pp. 935–936. ISSN: 1548-7105. DOI: 10.1038/nmeth.4437.

Wagner A., Regev A. & Yosef N. (2016). "Revealing the vectors of cellular identity with single-cell genomics". *Nature Biotechnology* 34 (11), pp. 1145–1160. ISSN: 1546-1696. DOI: 10.1038/nbt.3711.

Wang D. & Gu J. (2018). "VASC: Dimension Reduction and Visualization of Single-cell RNA-seq Data by Deep Variational Autoencoder". *Genomics, Proteomics & Bioinformatics* 16 (5), pp. 320–331. DOI: 10.1016/j.gpb.2018.08.003.

Wang Q., Stacy T., Binder M., Marin-Padilla M., Sharpe A. H. & Speck N. A. (1996). "Disruption of the Cbfa2 gene causes necrosis and hemorrhaging in the central nervous system and blocks definitive hematopoiesis." *Proceedings of the National Academy of Sciences* 93 (8), pp. 3444–3449. ISSN: 1091-6490. DOI: 10.1073/pnas.93.8.3444.

Wang X., Park J., Susztak K., Zhang N. R. & Li M. (2019). "Bulk tissue cell type deconvolution with multi-subject single-cell expression reference". *Nature Communications* 10 (1). ISSN: 2041-1723. DOI: 10.1038/s41467-018-08023-x.

Welch J. D., Kozareva V., Ferreira A., Vanderburg C., Martin C. & Macosko E. Z. (2019). "Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity". *Cell* 177 (7), 1873–1887.e17. DOI: 10.1016/j.cell.2019.05.006.

Williams C. G., Lee H. J., Asatsuma T., Vento-Tormo R. & Haque A. (2022). "An introduction to spatial transcriptomics for biomedical research". *Genome Medicine* 14 (1). ISSN: 1756-994X. DOI: 10.1186/s13073-022-01075-1.

Wolf F. A., Angerer P. & Theis F. J. (2018). "SCANPY: large-scale single-cell gene expression data analysis". *Genome Biology* 19 (1). ISSN: 1474-760X. DOI: 10.1186/s13059-017-1382-0.

Xu C., Lopez R., Mehlman E., Regier J., Jordan M. I. & Yosef N. (2021). "Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models". *Molecular Systems Biology* 17 (1). ISSN: 1744-4292. DOI: 10.15252/msb.20209620.

Yokomizo T. & Dzierzak E. (2010). "Three-dimensional cartography of hematopoietic clusters in the vasculature of whole mouse embryos". *Development* 137 (21), pp. 3651–3661. DOI: 10.1242/dev.051094.

Yokomizo T., Ogawa M., Osato M., Kanno T., Yoshida H., Fujimoto T., Fraser S., Nishikawa S., Okada H., Satake M., Noda T., Nishikawa S.-I. & Ito Y. (2001). "Requirement of Runx1/AML1/PEBP2αB for the generation of haematopoietic cells from endothelial cells". *Genes to Cells* 6 (1), pp. 13–23. ISSN: 1365-2443. DOI: 10.1046/j.1365-2443.2001.00393.x.

Zechel S., Zajac P., Lönnerberg P., Ibáñez C. F. & Linnarsson S. (2014). "Topographical transcriptome mapping of the mouse medial ganglionic eminence by spatially resolved RNA-seq". *Genome Biology* 15 (10). DOI: 10.1186/s13059-014-0486-z.

Zeisel A., Köstler W. J., Molotski N., Tsai J. M., Krauthgamer R., Jacob-Hirsch J., Rechavi G., Soen Y., Jung S., Yarden Y. & Domany E. (2011). "Coupled pre-mRNA and mRNA dynamics unveil operational strategies underlying transcriptional responses to stimuli". *Molecular Systems Biology* 7 (1), p. 529. DOI: 10.1038/msb.2011.62.

Zeisel A., Muñoz-Manchado A. B., Codeluppi S., Lönnerberg P., La Manno G., Juréus A., Marques S., Munguba H., He L., Betsholtz C., Rolny C., Castelo-Branco G., Hjerling-Leffler J. & Linnarsson S. (2015). "Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq". *Science* 347 (6226), pp. 1138–1142. ISSN: 1095-9203. DOI: 10.1126/science.aaa1934.

Zheng G. X. Y., Terry J. M., Belgrader P., Ryvkin P., Bent Z. W., Wilson R., Ziraldo S. B., Wheeler T. D., McDermott G. P., Zhu J., Gregory M. T., Shuga J., Montesclaros L., Underwood J. G., Masquelier D. A., Nishimura S. Y., Schnall-Levin M., Wyatt P. W., Hindson C. M., Bharadwaj R., Wong A., Ness K. D., Beppu L. W., Deeg H. J., McFarland C., Loeb K. R., Valente W. J., Ericson N. G., Stevens E. A., Radich J. P., Mikkelsen T. S., Hindson B. J. & Bielas J. H. (2017). "Massively parallel digital transcriptional profiling of single cells". *Nature Communications* 8 (1). ISSN: 2041-1723. DOI: 10.1038/ncomms14049.

Zhu Q., Gao P., Tober J., Bennett L., Chen C., Uzun Y., Li Y., Howell E. D., Mumau M., Yu W., He B., Speck N. A. & Tan K. (2020). "Developmental trajectory of prehematopoietic stem cell formation from endothelium". *Blood* 136 (7), pp. 845–856. DOI: 10.1182/blood.2020004801.

Ziegenhain C., Vieth B., Parekh S., Reinius B., Guillaumet-Adkins A., Smets M., Leonhardt H., Heyn H., Hellmann I. & Enard W. (2017). "Comparative Analysis of Single-Cell RNA Sequencing Methods". *Molecular Cell* 65 (4), 631–643.e4. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2017.01.023.

Zovein A. C., Hofmann J. J., Lynch M., French W. J., Turlo K. A., Yang Y., Becker M. S., Zanetta L., Dejana E., Gasson J. C., Tallquist M. D. & Iruela-Arispe M. L. (2008). "Fate Tracing Reveals the Endothelial Origin of Hematopoietic Stem Cells". *Cell Stem Cell* 3 (6), pp. 625–636. DOI: 10.1016/j.stem.2008.09.018.

-