UNIVERSITY OF SOUTHAMPTON

Faculty of Social Sciences
School of Mathematical Sciences

An Information-Theoretic Definition of Cell Type

by

Michael John Casey
BA, MSci
ORCiD: 0000-0002-0322-4224

A thesis for the degree of
Doctor of Philosophy

October 2021

University of Southampton

Abstract

Faculty of Social Sciences
School of Mathematical Sciences

Doctor of Philosophy

An Information-Theoretic Definition of Cell Type

by Michael John Casey

Individual cells are often classified into cell 'types' based on the expression of so-called marker genes. Such marker-based classification assumes that cells of a given type are (at least approximately) interchangeable with respect to the expression of their associated markers. This traditional approach to cellular classification has been disrupted by single-cell RNA-sequencing technologies, which are able to measure genome-wide gene expression across thousands of individual cells. While potentially providing a wealth of data for cellular classification, these technologies have revealed that cells ostensibly of the same type are often highly heterogeneous (i.e. not interchangeable) with respect to the expression of established marker genes.

A myriad of single-cell clustering methods has recently been developed to overcome the issue of heterogeneity with respect to marker gene expression and identify cell types directly from single-cell expression data. These methods typically proceed via: (1) unsupervised identification of clusters from single-cell expression data sets; (2) mapping of identified clusters to known cell types based on the expression of previously established marker genes. However, this two-step cluster-based approach to cellular classification is less biologically intuitive than the traditional marker-based approach, involving substantial mathematical and biological assumptions regarding the nature of cell type.

In this thesis, I formalise the traditional marker gene approach to cellular classification using notions from information theory, and show how this formalism can be applied to identifying cell types from single-cell RNA-sequencing data. Specifically, I develop a novel clustering method based on the assumption that cells of the same type should be minimally heterogeneous – i.e. approximately interchangeable – with respect to the measured expression of a set of genes. Thus, this work offers an intuitive, formal definition of cell type that unites the traditional and current approaches to cellular classification through the mathematics of information theory.

Contents

List of Figures **vii**

List of Tables **ix**

Declaration of Authorship **xi**

Acknowledgements **xiii**

1 Introduction **1**

1.1 Phenotypic Classification of Cell Types . . . 3

1.1.1 Gene Expression . . . 5

1.1.2 Single-cell RNA-sequencing . . . 6

1.2 Unsupervised Clustering of Cell Types . . . 8

1.2.1 Gene Regulatory Networks . . . 10

1.2.2 Dynamical Systems Theory . . . 14

1.2.3 Unsupervised Clustering . . . 19

1.2.4 Pre-processing . . . 25

1.3 Classifying Clusters . . . 26

2 Information-Theoretic Clustering of Cell Types **31**

2.1 Information Theory . . . 33

2.2 Quantifying Heterogeneity . . . 37

2.3 Heterogeneity by Cluster . . . 41

2.3.1 Inter-cluster Heterogeneity . . . 42

2.3.2 Intra-cluster Heterogeneity . . . 44

2.3.3 Additive Decomposition . . . 45

2.3.4 Mathematical Derivation . . . 48

2.4 Heterogeneity of Many Genes . . . 50

2.5 Implementation & Validation . . . 53

2.5.1 Normalisation . . . 55

2.5.2 Population Heterogeneity . . . 57

2.5.3 Differential Gene Expression . . . 61

2.6 Discussion . . . 62

3 Numerical Optimisation of Information-Theoretic Clustering **67**

3.1 Optimising Inter-cluster Heterogeneity . . . 70

3.1.1 Fuzzy Clustering . . . 72

3.1.2 Differentiation of Inter-cluster Heterogeneity . . . 74

3.1.3 Implementation of the *L-BFGS-B* Algorithm 77
3.1.4 Validation 79
3.2 Semi-supervised Classification 87
3.3 Discussion 90

4 Estimating Cluster Number 95
4.1 Introduction to Graph Theory 97
4.2 Hypergraph Construction 100
4.2.1 Single-cell Hypergraphs 101
4.3 Quantifying Heterogeneity 103
4.4 Estimating Cluster Number 104
4.4.1 Intra-type Heterogeneity 106
4.5 Discussion 108

5 Discussion 111

Appendix A Single-cell RNA-sequencing 125

Appendix B Further Methods for Single-cell Analysis 129

Appendix C Supplementary Code 141

References 151

List of Figures

1.1 Phenotypic classification by marker gene expression 3
1.2 Schematic of the gene regulatory network of λ phage 12
1.3 Schematic of the gene regulatory network of sonic hedgehog 14
1.4 Waddington's epigenetic landscape 16
1.5 Example of gene expression space 17
1.6 Transitions in cellular gene expression 22
1.7 Visualisation of differential gene expression 28
1.8 tSNE of the Tabula Muris 29

2.1 Example of branching 35
2.2 Entropy of distributions 36
2.3 Distribution of transcripts per cell 38
2.4 Uniform distribution of transcripts on the set of cells 39
2.5 Gene expression heterogeneity 40
2.6 Quantifying heterogeneity 41
2.7 Inter-cluster heterogeneity 42
2.8 Intra-cluster heterogeneity 44
2.9 Branching in transcript assignment 46
2.10 Additive decomposition 48
2.11 Branching assignment of transcripts to genes and cells 52
2.12 Effect of sparsity on population heterogeneity 56
2.13 James-Stein-type $I(g)$ against mean gene expression 57
2.14 Population heterogeneity of sequencing data sets 59
2.15 UMAP of mouse bone marrow cells 60
2.16 Significance of $H_S(g)$ with respect to randomisations 63
2.17 Normalised inter-cluster heterogeneity of Zheng et al. (2017) data 64

3.1 Landscape of an objective function 69
3.2 Effect of fuzziness on inter-cluster heterogeneity 78
3.3 Batch effect and the James-Stein-type estimator 83
3.4 Visualisation of scEC clustering of mouse bone marrow data set 85

4.1 Intra-type R 107

5.1 Example of a microscopy experiment 112
5.2 The cell as a dynamical system 114
5.3 Gene regulatory networks 115
5.4 Inter and intra-cluster heterogeneity during cellular differentiation . . . 122

List of Tables

2.1 Data sets for use in validation 54
2.2 H_S of established and randomised cellular clusterings 62

3.1 Similarity of clustering results to established classifications 81
3.2 Information-theoretic clustering of immune cells 82
3.3 Information-theoretic clustering of mouse bone marrow data set 84
3.4 Cancer cell line data sets 90
3.5 Semi-supervised classification of cancer cell lines 90

4.1 Toy examples of R 104
4.2 R of validation data sets 105

Declaration of Authorship

I declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:

   (a) M. J. Casey, R. J. Sanchez-Garcia, and B. D. MacArthur. Measuring the information obtained from a single-cell sequencing experiment. *bioRxiv*, 2020a

   (b) M. J. Casey, P. S. Stumpf, and B. D. MacArthur. Theory of cell fate. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 12(2):e1471, 2020b

   (c) R. Mulas and M. J. Casey. Estimating cellular redundancy in networks of genetic expression. *Mathematical Biosciences*, page 108713, 2021. ISSN 0025-5564

Signed:... Date:..................

Acknowledgements

First and foremost, I would like to thank my supervisors, Ben D. MacArthur and Rubén J. Sánchez-García, for providing me with guidance and support during my candidature. Thank you both for helping a biologist begin to navigate the world of mathematics.

I would like to thank my colleagues, Patrick Stumpf, Joe Egan, Raffaella Mulas and David Méndez Martínez, for lending me their expertise in biology and mathematics. I would like to thank Patrick in particular for so readily sharing his insights, knowledge and experience in the field of single-cell biology.

I would like to thank Jörg Fliege for sharing his expertise on numerical optimisation.

Finally, thanks to my friends, especially my flatmate, Ben Kitching-Morely, and to the Lunch Beacon, Ben Batchelor, Katie Roe, Rachel Bolton and Dan Noel for many lunches, coffee breaks and sessions of dungeons and dragons.

To My Family

My Parents, Valerie and Sean
And My Brother, Ruairí

"These ambiguities, redundancies, and deficiencies recall those attributed by Dr. Franz Kuhn to a certain Chinese encyclopedia called the Heavenly Emporium of Benevolent Knowledge. In its distant pages it is written that animals are divided into (a) those that belong to the emperor; (b) embalmed ones; (c) those that are trained; (d) suckling pigs; (e) mermaids; (f) fabled ones; (g) stray dogs; (h) those that are included in this classification; (i) those that tremble as if they were mad; (j) innumerable ones; (k) those drawn with a very fine camel's-hair brush; (l) et cetera; (m) those that have just broken the flower vase; (n) those that at a distance resemble flies."

From *The Analytical Language of John Wilkins*, Borges (1952)

Chapter 1

Introduction

Robert Hooke first coined the term *cell* in the 1665 Micrographia (Hooke, 1665). Nearly 200 years later, Schleiden and Schwann concluded that all living matter was composed of cells (Schleiden, 1838; Schwann, 1839). Specifically, Schwann formulated *cell theory*: that "the elementary parts of all tissues are formed of cells" (Schwann, 1839). Cell theory has since become the "most general structural paradigm in biology" (Mazzarello, 1999), providing a universal basis for the study of life.

Much like the multicellular organisms they constitute, cells come in many varieties. Cells display a range of morphologies and functions, occupying distinct phenotypic niches known as *cell types* (Regev et al., 2017). These cell types form the elements of a classification system of cells that stretches across organs and species – as cell theory provides a general structural paradigm, cell types provide a general organisational paradigm for the study of cell biology.

However, there is no universally acceptable criteria by which to classify a cell type (Regev et al., 2017; Clevers et al., 2017; Xia and Yanai, 2019). An early notion was that each distinct cell type carried only part of the genome. The observed differences in cellular phenotype arose because each cell type had access to a different set of "instructions" from the genome. Thus, cell types could be precisely identified by genome sequence. However, the pioneering experiments of Laskey and Gurdon disproved this idea, showing that all cells of a given organism share the same DNA (aside from specific lymphocytes, some neurons and anuclear cells)(Laskey and Gurdon, 1970). Distinct cell types arise not due to differences in the genome but due to differences in expression of the same genome.

Therefore, instead of classifying cells by genotype, cell types are identified by phenotype. Cells are classified into distinct types based on aspects or features of cellular phenotype, including morphology, biological function, and gene expression (Regev et al., 2017; Tanay and Regev, 2017; Alberts, 2017; Mescher, 2018). These identifying features are shared by cells of the same type and distinguish cells of different types. For example, neurons and Spermatozoa can be identified by their distinct morphology's, B cells and T cells by function, and pluripotent stem cells by gene expression (Xia and Yanai, 2019). This feature-based identification of cell types is termed phenotypic classification (Guillemin and Stumpf, 2020).

The phenotypic classification of cells assumes that the cells of each type are interchangeable with respect to features of interest. For example, all Spermatozoan are assumed to have the same morphology, and all erythrocytes to have the same function. Conversely, phenotypic classification assumes cells of different types are distinguishable with respect to features of interest. A feature can only be used to classify a cell type if it is fulfils both criteria, with cells of the same type being interchangeable and cells of different types being distinguishable with respect to the feature.

However, the phenotypic classification of cell types has been disrupted with the advent of high-throughput single-cell methods, primarily single-cell RNA-sequencing (Trapnell, 2015; Regev et al., 2017; Casey et al., 2020b). Single-cell sequencing technologies quantitatively profile the gene expression of individual cells (Svensson et al., 2017). This profiling has revealed that cells ostensibly of the same type are often heterogeneous, i.e. distinguishable, with respect to established molecular characteristics, as illustrated in **Fig 1.1** (Trapnell, 2015; Regev et al., 2017; Luecken and Theis, 2019).

This thesis aims to quantify the heterogeneity of cells with respect to gene expression in the context of single-cell RNA-sequencing. Through this quantification, I aim to formalise the phenotypic classification of cells with respect to gene expression heterogeneity and extend the principles of phenotypic classification to single-cell expression data. To that end, in this introductory chapter, I will discuss the phenotypic classification of cell types and the application of phenotypic classification to single-cell RNA-sequencing data. I will introduce the dominant approach to cellular classification with respect to single-cell sequencing data, unsupervised clustering. I will detail the theory justifying the cluster-based classification of cells and discuss some of the unsupervised clustering methods applied to single-cell RNA-sequencing. Finally, I will conclude this chapter by introducing how clusters are identified with established cell types via differential gene expression analysis.

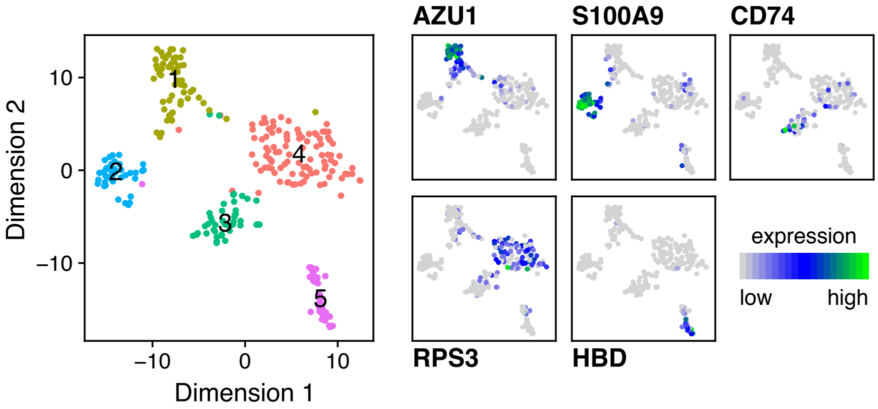


FIGURE 1.1: **Phenotypic classification by marker gene expression**. Each point corresponds to the gene expression vector of an individual cell sampled from human bone marrow projected into two dimensions using the t-SNE nonlinear dimension reduction method (Hinton and Roweis, 2003). (*Left*) The set of cells are classified into five different cell types, highlighted in distinct colours: (1) Myeloblasts, (2) Monoblasts, (3) Lymphoid Cells (4) Stem and Progenitors (5) Erythroblasts. (*Right*) Each cell type is identified based on the expression of established marker genes. Despite each cell type having been previously defined with respect to marker gene expression, the cells of each type are heterogeneous with respect to the expression of their associated marker gene. Figure reproduced from Casey et al. (2020b).

1.1 Phenotypic Classification of Cell Types

Phenotypic classification provides the *de facto* working definition of cell type. Cells are identified as specific cell types based on shared aspects or features of cellular phenotype (Tanay and Regev, 2017). Under this classification, cells are identified with specific cell types when they match key *marker* characteristics of the type, e.g. morphology, biological function or gene expression.

In phenotypic classification, cells of each type are assumed to be (approximately) interchangeable with respect to their associated marker features. Practically, this assumption is necessary as there are too many cells in a given organism to characterise individually (there are at least 10^{13} cells in an adult human, (Bianconi et al., 2013)). Moreover, the majority of cells cannot be one-to-one mapped between individuals in complex multicellular organisms (unlike with some simpler organisms, e.g. *Caenorhabditis elegans*) (Sender et al., 2016; Sulston et al., 1983). Instead, phenotypic features are identified with individual cell types.

Thus, where cell theory provides a general structural paradigm for cell biology, cell types provide a general organisational paradigm. By assuming that cells of each type are interchangeable (same morphology, same function, same gene expression, etc.), the properties of one cell of a type can be generalised to all cells of that type. A complete description of cellular phenotypes requires characterising only the relatively small set of cell types, as opposed to individually characterising each cell (Trapnell, 2015).

In contrast, while cells of the same type must be interchangeable with respect to a marker feature, cells of different types must be distinguishable with respect to that feature. A given cell type is unlikely to be uniquely distinguishable based on a single feature (morphology, function, gene expression, etc.), but each marker feature must distinguish a cell type or set of cell types from at least some other types of cell (Fischer and Gillis, 2021). For example, the various types of T-cells can be distinguished based on the expression of a set of surface proteins (Zheng et al., 2017).

Importantly, different features can disagree on the identity of a cell; for example, cells can be of different developmental lineages but have similar functions, e.g., trunk skeletal muscle cells and cranial skeletal muscle cells have the same function but are from different lineages with different gene expression patterns (Sambasivan et al., 2009). Thus, the identity of a given cell may only be resolved based an accumulation of many different marker features (Fischer and Gillis, 2021).

Sets of cells that do not belong to an established cell type and are distinguishable with respect to a given feature may represent novel cell types (Trapnell, 2015). However, establishing a set of cells as constituting a novel cell type requires identifying many such features: the biological function of a given cell type depends on the coordination of various aspects of cellular phenotype, including morphology, gene expression and location. Therefore, cells of the same type should be interchangeable with respect to many different features, i.e. true cell types are distinguished from random groupings of cells based on accumulating many distinct marker features (Melé et al., 2015).

Given that cells are only classifiable with respect to observable or measurable aspects of cellular phenotype, the identification of novel cell types is method dependent, i.e. not all cell types will be identifiable based on existing experimental methods. For example, many of the earliest identified cell types (e.g. Erythrocytes, Schwann cells, Purkinje cells, Sertoli cells, Astrocytes and Spermatozoon) have distinctive whole-cell morphology's (Mescher, 2018; Xia and Yanai, 2019). Such large-scale morphologies are observable under a traditional light microscope, the only method available for use in classifying cells for some time.

As such, many cell types have only been made discoverable with the advent of new experimental methodologies and technologies (Tanay and Regev, 2017). A leading example of such methodological dependence lies in the tools of molecular biology, which enable the direct measurement of the molecular products of gene expression. Based on the tools of molecular biology, the molecular classification of cells has become a leading approach in identifying known and novel cell types (Regev et al., 2017).

1.1.1 Gene Expression

Genes are expressed through a process known as the 'central dogma' (Krebs et al., 2017; Alberts, 2017). Molecules of messenger RNA (mRNA) are transcribed from the coding sequence of a given gene; the mRNA is then exported from the nucleus and molecules of protein translated from each molecule of mRNA. The level of expression of a given gene can be changed by varying the rate of transcription or translation (Gygi et al., 1999). Note that as the complete set of DNA molecules in a cell forms the genome, the complete set of mRNA transcripts form the transcriptome, and the complete set of protein molecules form the proteome.

Within molecular biology, tools have been developed to measure the mRNA and protein content of individual cells (Krebs et al., 2017; Alberts, 2017; Mescher, 2018). Instead of relying on potentially subjective descriptions of morphology or function, cell types can be identified through the expression of so-called *marker genes*. The expression of characteristic genes, such as those encoding cell surface proteins or transcription factors provides a reliable way to identify cell types (Thomson et al., 1998; Akashi et al., 2000; Lv et al., 2014; Tapscott et al., 1988; Mitsui et al., 2003).

Classifying cells into types based on the expression of marker genes provides a common language for the identification of cell types, with each cell type being identifiable based on the expression of a unique combination of $\sim$ 10-200 genes (Arendt, 2008; Arendt et al., 2016; Xia and Yanai, 2019; Fischer and Gillis, 2021). Nevertheless, the marker gene approach to cell type discovery and classification is limited by throughput. Traditionally, molecular biology methods profile the expression of only one or a handful of genes at a time (Tanay and Regev, 2017). However, with the advent of single-cell sequencing technologies, it is now possible to measure the expression of all genes simultaneously. By measuring the whole-transcriptome, single-cell sequencing enables the systematic and comprehensive discovery of marker genes for existing cell types and, importantly, provides sufficient data to identify all cells types, known and novel, distinguishable at the level of the transcriptome (Regev et al., 2017).

1.1.2 Single-cell RNA-sequencing

Single-cell RNA-sequencing profiles the genome-wide gene expression of individual cells, counting the number of mRNA molecules (which I will refer to as 'transcripts' from here on) transcribed from each gene within individual cells (Svensson et al., 2017; Ziegenhain et al., 2017). Note that the process of sequencing requires lysing (destroying) the measured cells. The final output of the experiment is a count matrix detailing the number of transcripts of each gene measured in each cell. The count matrix represents a static snapshot of the transcriptome of a cellular population. See the box *"Single-cell RNA-sequencing"* for a brief outline of the stages of single-cell RNA-sequencing and see **Appendix A** for more detail (Luecken and Theis, 2019).

Single-cell sequencing measures the whole transcriptome, profiling the expression of every gene, and so capturing all possible differences in gene expression between cell types (Tanay and Regev, 2017). Thus, instead of conducting many successive experiments, single-cell sequencing provides sufficient data to classify the complete set of cell types present in a population from a single experiment (Trapnell, 2015; Regev et al., 2017). The marker genes expressed by each cell can be identified, allowing the rapid and automatic classification of each cell.

However, single-cell sequencing data has revealed substantial inconsistencies in the expression of established marker genes (Brennecke et al., 2013; Dillies et al., 2013; Grün et al., 2014; Kim et al., 2015; Vallejos et al., 2017). In sequenced populations, each cell of a given type tends to express only a subset of relevant marker genes or express those genes at varying levels (see again the example in **Fig 1.1**). This inconsistency in marker gene expression prohibits the reliable classification of cells based on the expression of any single marker gene.

These revealed inconsistencies in expression are termed heterogeneity, or more formally, intra-type heterogeneity. Throughout this thesis, I define a set of cells as *heterogeneous* with respect to the expression of a given gene when the cells are distinguishable, i.e. not interchangeable, based on the observed expression of the gene. When cells of the same type are heterogeneous with respect to the expression of a given gene, any given cell cannot be reliably classified based on that gene.

Intra-type heterogeneity with respect to maker gene expression is ubiquitous (Regev et al., 2017). The revealed gene expression heterogeneity can be broadly explained by two characteristics of single-cell RNA-sequencing: single-cell technologies measure gene expression genome-wide, and single-cell

Single-cell RNA-sequencing

Isolation The input for single-cell RNA-sequencing is a population of cells, typically sampled from a biological tissue. The cellular population is dissociated into a suspension of single cells, with each cell isolated separately (Luecken and Theis, 2019). Throughout this thesis, I will assume cells have been isolated by high-throughput droplet-based methods (e.g. Drop-seq or inDrop) that capture cells in microfluidic droplets (Macosko et al., 2015; Klein et al., 2015; Ziegenhain et al., 2017; Papalexi and Satija, 2018).

Barcoding Post-isolation, the mRNA content of each cell (or nucleus) is released. Each transcript is associated with two nucleic acid barcodes, one identifying the cell of origin (the cellular barcode) and the other identifying the individual transcript molecule (the unique molecular identifier or UMI) (Papalexi and Satija, 2018; Kivioja et al., 2012; Islam et al., 2014; Svensson et al., 2017). (Note that the use of UMIs is not universal, but for simplicity, I will assume the use of UMIs throughout this thesis).

Sequencing Transcripts from each cell are pooled and amplified, producing many DNA molecules complementing each transcript. The amplified DNA molecules are then sequenced, converting the set of nucleic acid bases of each physical transcript into a digital sequencing read (Goodwin et al., 2016).

Mapping The sequencing reads are then mapped back 1) onto the genome to identify the expressed gene; 2) to the cell of origin by cellular barcode; and, 3) to the individual transcript molecule via UMI. The result of this mapping is the count matrix, detailing the number of individual transcripts expressed by each gene in each cell.

technologies quantify gene expression on the level of individual transcript molecules (Svensson et al., 2017).

By measuring the whole transcriptome, single-cell sequencing allows for and enforces a higher stringency in classifying cell types. With lower-throughput measurements, where cells are classified based on the expression of a single gene, gene expression heterogeneity is hidden by the misclassification of individual cells. Whereas, with whole-genome sequencing, each cell type is identifiable based on all possible marker genes simultaneously. Therefore, heterogeneity with respect to the expression of any one gene is revealed by the classification of cells with respect to the remaining genes (Tanay and Regev, 2017).

Single-cell RNA-sequencing is a high-resolution measurement, measuring single molecules of mRNA in individual cells. Single-cell sequencing, therefore, captures previously unobservable differences between cells (Buettner et al., 2015; Björklund et al., 2016; Stuart et al., 2019). However, single-cell sequencing is also a highly inefficient process, measuring only 3-10% of transcripts per cell; this inefficiency results in cells of the same type being distinguishable solely due to technical error (Papalexi and Satija, 2018). These alternative sources of gene expression heterogeneity can be sorted into three main categories: biological function, biological noise and technical error. See the box *"Alternative Sources of Heterogeneity"* for a detailed explanation of each source of alternative heterogeneity.

In response to the revealed intra-type heterogeneity of cells with respect to established marker genes, a novel classification approach has been developed for identifying cell types from single-cell RNA-sequencing data. Instead of classifying individual cells, cells are first grouped into clusters based on the 'similarity' of cellular gene expression across all genes measured (Trapnell, 2015; Regev et al., 2017; Kiselev et al., 2019). Each cluster is then classified based on the relative expression of marker genes within each cluster. The clustering of cells is undertaken computationally via a process termed unsupervised clustering; in the next section, I will outline the theoretical motivation for the unsupervised clustering approach to classifying cell types, before moving to discuss the practical application of unsupervised clustering to single-cell data.

1.2 Unsupervised Clustering of Cell Types

Phenotypic classification identifies cells based on the univariate expression patterns of individual marker genes. However, intra-type heterogeneity with respect to marker gene expression prohibits the phenotypic classification of individual cells based on single-cell RNA-sequencing data. Instead, cells are first grouped into clusters of 'similar' cells (Kiselev et al., 2019). Clusters are then identified with established cell types by determining those marker genes that are on average up or down-regulated in each cluster, with each cell inheriting the identity of its assigned cluster (Love et al., 2014; Luecken and Theis, 2019).

Alternative Sources of Heterogeneity

Biological function Gene expression heterogeneity arising from biological processes orthogonal to the emergence and maintenance of cell type. The most prominent example of these functions is the cell cycle, where gene expression undergoes a series of transitions through the different stages of DNA replication and mitosis (Stuart et al., 2019).

Biological noise Gene expression heterogeneity arising from stochasticity in transcription. Gene expression involves the interaction of small numbers of molecules. The motion of individual molecules is inherently random, introducing substantial stochasticity in gene expression (Raj et al., 2006). Moreover, transcription occurs in 'bursts', not continuously, increasing the range in transcripts numbers present in the cell at any given time (Raj et al., 2006).

Technical error Gene expression heterogeneity arising from the measurement process (see **Appendix A** for a detailed description of the measurement process). Single-cell sequencing is highly inefficient, capturing only 3-10% of transcripts in a given cell (Papalexi and Satija, 2018). Moreover, the total number of transcripts measured per cell can vary by orders of magnitude (Dillies et al., 2013). Thus, even if the relative proportion of transcripts mapping to a given gene in each cell is constant, the absolute number of transcripts measured per cell can vary dramatically. Technical error is a substantial source of intra-type heterogeneity (Stuart et al., 2019; Townes et al., 2019; Breda et al., 2021; Lause et al., 2020; Ahlmann-Eltze and Huber, 2020).

Cells are grouped into clusters via a set of computational techniques known as unsupervised clustering (Freytag et al., 2018; Duò et al., 2018; Kiselev et al., 2019). Unsupervised clustering methods cluster cells based on overall similarity in gene expression, leveraging the expression of many genes to overcome intra-type heterogeneity with respect to the expression of any single gene. Specifically, involvement in cell types induces a high level of dependency or coordination between the expression of various genes. Unsupervised clustering methods leverage this multivariate dependency in gene expression to identify cells that are similar with respect to the whole-transcriptome.

The classification of cells on the level of clusters assumes that all cells assigned to a given cluster are of the same type. Therefore, unsupervised clustering methods must be able to identify cell types from single-cell RNA-sequencing data without knowledge of which genes have been established as markers for each cell type. Unsupervised clustering methods must be able to recover the the inherent structure of a cellular population via solely computational and mathematical means.

In this section, I will outline the justification for why unsupervised clustering methods are able to group cells into types without human supervision. The motivation for unsupervised clustering is substantially more involved than for phenotypic classification: as phenotypic classification represents an empirically-driven approach to cell type identification, unsupervised clustering represents a theory-driven approach. Therefore, I will begin this section by outlining the theoretical justification for the unsupervised clustering identification of cell types. I will introduce how genes act collectively to give rise to distinct cellular identities, demonstrating this collective action first through small scale examples and then discussing how this occurs genome-wide. I will discuss how this collective action makes it possible to identify cell types without reference to any particular marker gene and how cells types are identified by unsupervised clustering methods. Finally, I will outline how unsupervised clustering is implemented in practice. Note that parts of this section, **Section 1.2**, have been published in my review article Casey et al. (2020b).

1.2.1 Gene Regulatory Networks

Cell types emerge through the coordinated action of many genes. This coordination takes the form of regulatory interactions between genes, and in particular between genes coding for transcription factors (transcription factors being proteins that control the rate of transcription). The interactions are fixed by the DNA of a given organism, encoded through regulatory sequences of DNA (sequences of DNA to which transcription factors can bind to regulate the expression of neighbouring genes), and through chemical interactions between the RNA and protein outputs of gene expression (Krebs et al., 2017).

The coordinated action of set of genes is typically modelled through a gene regulatory network (Britten and Davidson, 1969). A gene regulatory network is the collection of genes and the regulatory interactions between each pair of genes for a given cell. For a small set of genes, gene regulatory networks are typically illustrated as a circuit diagram, with 'wires' representing interactions between genes, typically inhibition or activation (Peter and Davidson, 2015). These networks process the information provided by external signals, leading to a specific set of transcription factors being expressed in the cell, which then activate downstream 'batteries' or 'schedules' of genes, resulting in the acquisition of a distinct cell type.

Gene regulatory networks strongly constrain the expression patterns of the individual genes: the expression of each gene in the network, including those genes in downstream 'batteries', depends on the expression of the remaining genes in the network. In the following section, **Section 1.2.2**, I will discuss how this dependency, or coordination, in gene expression results in cells of the same type being similar with respect to genome-wide gene expression. It is this within-type similarity that enables unsupervised clustering methods to identify cell types *de novo*. However, in this section, **Section 1.2.1**, I will first illustrate the concept of the gene regulatory network through two smaller-scale examples, involving only two/three genes: lambda phage and the ventral neural tube (Ptashne, 2004; Balaskas et al., 2012).

The lambda phage was one of the first characterised examples of genetic decision making (Jacob and Monod, 1961; Monod and Jacob, 1961). Phages are acellular complexes of nucleic acid and protein that invade and replicate within bacteria. They have a limited number of genes and posses minimal gene regulatory networks. In lambda phage the gene regulatory network decides between two possible "lifestyles" for the phage: while phages are not cells, these "lifestyles" offer a simplified model of how gene regulatory networks activate different cell types.

After introducing gene regulatory networks in phages, I will discuss a more complex example in eukaryotes: the patterning of the ventral neural tube (Balaskas et al., 2012). The ventral neural tube is an example of a morphogen gradient. Morphogens are signalling molecules that trigger cellular differentiation during development. But a single morphogen does not trigger a single cell type; rather, a single morphogen can induce multiple distinct cell types, depending on the exact conditions of the signalling (e.g. concentration of the morphogen molecule). The ventral neural tube serves as an example of how interactions between a small set of genes can give rise to multiple stable cell types.

Lambda Phage

The lambda phage is a bacteriophage that has two distinct lifestyles: lysis and lysogeny (Ptashne, 2004). Lysis involves mass replication of the phage through hijacking of the bacteria's DNA replication machinery, resulting in bacterial cell death (lysis) and the release of many phage particles. In lysogeny, the

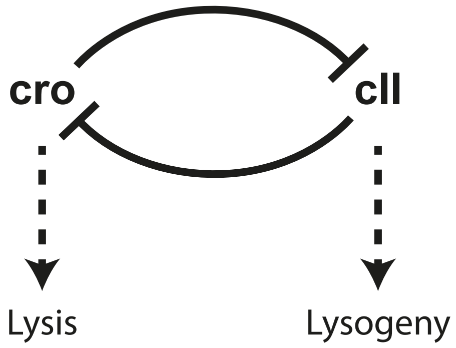


FIGURE 1.2: **Schematic of the gene regulatory network of λ phage**. The genes *cro* and *cII* repress each others expression, represented by the block-head arrows, and activates the phage lifestyles of lysis and lysogeny respectively, as represented by the dashed arrows. The mutual repression forms a 'genetic-switch' where only one of *cro* or *cII* is expressed at a given time. The purpose of the network is to determine the level of nutrients in the environment via the level of protease expressed in the cell. In a high-nutrition environment with high protease expression, the protein *cII* is broken down, leading to expression of *cro* and the energetically expensive lysis lifestyle.

phage integrates itself into the host's genome, replicating with the host bacteria's DNA.

Lysis is energetically expensive, so is preferred only when nutrients are abundant; lysogeny is the preferred lifestyle for survival in a low-nutrient environment. The phage makes this decision through a network of two key proteins: *cro* and *cII*, shown in **Fig 1.2**.

As detailed in Ptashne (2004), each gene inhibits the transcriptional activation of the other, forming a genetic switch: only one gene can be 'on' at a time. The protein *cII* is broken down in the presence of bacterial proteases, and bacterial proteases are produced in high-nutrient conditions: the circuit turns lysogeny 'off' and lysis 'on' in the presence of nutrients.

The genetic switch forces the phage into one of two mutually exclusive, stable states. When *cro* is expressed, it inhibits the expression of *cII*, relieving any inhibition on its own expression, leading to greater *cro* expression; the same is true for *cII*. Intermediary states are unstable, as any imbalance in the levels of *cro* or *cII* will self-amplify, shifting the phage into lysis or lysogeny (Assaf et al., 2011).

More generally, such positive feedback circuits, where two competing genes (or sets of genes) are mutually inhibitory, and so self-reinforcing, are ubiquitous in gene regulatory networks (Milo et al., 2002; Soulé, 2006). They provide an immediate resolution to how multiple cell types can stably emerge from a single gene regulatory network: each cell type will correspond to a stable state of the network. The choice of cell type depends on the imbalance in the circuit induced by the external cues, e.g. protease concentration in the case of lambda phage.

The network as shown in **Fig 1.2** is a minimal representation of the lambda phage gene regulatory network: there are many proteins involved in the mediation and fine-tuning of the shown interaction, see Casjens and Hendrix (2015). In particular, the remaining genes of the network 1) aid in interpreting the protease concentration, i.e. the external signal, and 2) make the lysis/lysogeny decision reversible. Unlike with the majority of eukaryotic cell types, the decision between lysis and lysogeny will at some point be reversed, when the nutrient conditions change.

Nevertheless, the minimal representation demonstrates how interactions between sets of genes give rise to distinct cell types. Distinct cell types, or in the case of the lambda phage, distinct lifestyles, rely on the coordinated expression of multiple genes: a given cell type is 'defined' by both the individual genes expressed and the interactions between those genes.

Morphogen Gradient

The sonic hedgehog gene regulatory network represents a substantially more complex example of a gene regulatory network than the lambda phage. As described in Balaskas et al. (2012), the sonic hedgehog network illustrates the action of gene regulatory networks during the development of a complex multicellular organism.

Sonic hedgehog is a morphogen, a signalling molecule involved in triggering and regulating cellular differentiation during development. Sonic hedgehog triggers the cells of the ventral neural tube to differentiate into one of three distinct cell types: motor neurons, V2 neurons and V3 interneurons. The three-way decision is made through a single gene regulatory network composed of three transcription factors: *Olig2*, *Pax6* and *Nkx2.2*.

These transcription factors are involved in an asymmetric and nested set of positive and negative feedback loops, leading to the emergence of three stable states of the network, which induce the three possible cell types. The network in **Fig 1.3** is complex, see the box *"Sonic Hedgehog Gene Regulatory Network"* for a detailed description of the activity of the network.

As with *cro* and *cII* of the lambda phage, the expression of the three transcription factors of the sonic hedgehog network are highly coordinated. Each of the three cell types encoded by the network are defined with respect to

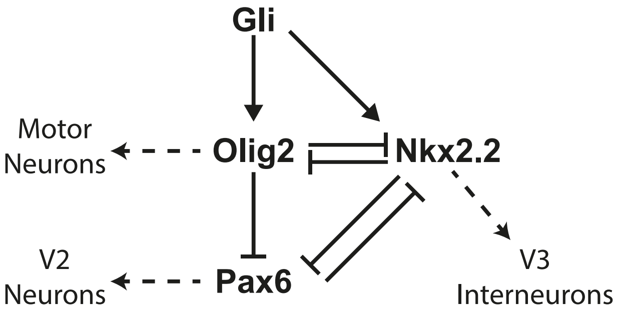


FIGURE 1.3: **Schematic of the gene regulatory network of sonic hedgehog**. The extracellular morphogen sonic hedgehog activates the intra-cellular intermediate Gli. Gli activates the expression of the transcription factors *Olig2* and *Nkx2.2*, as represented by the pointed arrows. Both *Olig2* and *Nkx2.2* repress the expression of *Pax6*, as represented by the block-head arrows. Expression of each of *Pax6*, *Olig2* and *Nkx2.2* leads to a cell differentiating into a different cell type. The final state of the network depends on the strength and length of sonic hedgehog signalling, with the dominant transcription factor swapping from *Pax6* to *Olig2* to *Nkx2.2* over the course of prolonged signalling.

the expression of all three transcription factors. However, despite the network's complexity, the sonic hedgehog network remains only a (very) small scale example of genetic regulation. Single-cell RNA-sequencing measures all genes simultaneously; therefore, there is a need to consider how gene regulatory networks, and the coordination in gene expression imposed by genetic regulation, extends to the whole genome. Specifically, the mathematics of dynamical systems theory and attractors is required.

1.2.2 Dynamical Systems Theory

Imagine there was sufficient data to determine an organism's complete gene regulatory network in fine detail: all the regulatory interactions are known, with each interaction fully characterised with respect to both strength and nature (inhibitory or activatory), enabling all the regulatory networks can be knitted together into a cohesive whole. From this complexity, do distinct cell types emerge as with the smaller scale gene regulatory networks?

This question was first addressed by Waddington in 1939 in the context of development (Waddington et al., 1939; Waddington, 2014). During an organism's development, cells differentiate from embryonic and intermediate cell types into adult cell types. Waddington imagined that the regulatory interactions between genes formed a landscape along which the cell travelled during development. The position of the cell on this landscape represents the current developmental state of the cell. The movement of the cell down the landscape represents the development of the cell, as constrained by the regulatory interactions between genes.

Sonic Hedgehog Gene Regulatory Network

In the absence of sonic hedgehog, *Pax6* is active, leading to V2 neuron cell type. Sonic hedgehog, through the intermediate *Gli*, activates *Olig2*, which in turn inhibits *Pax6*, leading to a switch in fated identity from V2 neuron to motor neuron. If sonic signalling is then lost, *Olig2* remains expressed, inhibiting the return of *Pax6* expression.

Gli also activates *Nkx2.2*, but while *Pax6* remains, expression of *Nkx2.2* is inhibited; as *Olig2* inhibits *Pax6*, this relieves the inhibition of *Nkx2.2*, allowing it to increase in expression. *Nkx2.2*, in turn, increases the inhibition of *Pax6*, accelerating its own activation, and inhibits *Olig2*, leading to another switch in eventual identity from motor neuron to V3 interneuron.

As can been seen in **Fig 1.3**, *Olig2* and *Nkx2.2* are involved in a genetic switch, as are *Nkx2.2* and *Pax6*. These genetic switches induce a tristability in the network, with sonic hedgehog signalling pushing the network from one stable state to the next. If sonic hedgehog signalling is withdrawn at an intermediate step, i.e. when *Olig2* is expressed, but before substantial build-up of Nkx2.2, then the *Olig2*-*Nkx2.2* genetic switch maintains motor-neuron cell type. Under sustained sonic hedgehog signalling, this switch flips as described, leading to stable V3 interneuron cell type.

While moving down the landscape, the cell will pass through various forks, representing the decisions in cell type, like those encoded by the small-scale, modular gene regulatory networks discussed above (see **Fig 1.4** for illustration). Starting from the peak of the landscape, representing the undifferentiated unicellular zygote, Waddington imagined that cells would move down through various forks before eventually coming to rest at a specific valley representing the adult cell type. These valleys represent stable configurations of the whole-genome regulatory network.

While proposed as a metaphor, Waddington's approach has, with some adjustment, substantial mathematical backing (Kauffman, 1969; Huang et al., 2005; Huang, 2012; Weinreb et al., 2018; Strogatz, 2018; Casey et al., 2020b; Newman, 2020; Greulich et al., 2020). Assume that the *state* of a cell can be described by the vector $x(t)$, where $x_i(t)$ is the level of expression of the i^{th} gene at time t (expression level can correspond to various measures, for example, number of mRNA transcripts or number of protein molecules). The change in cell state over time can be described by a set of coupled ordinary

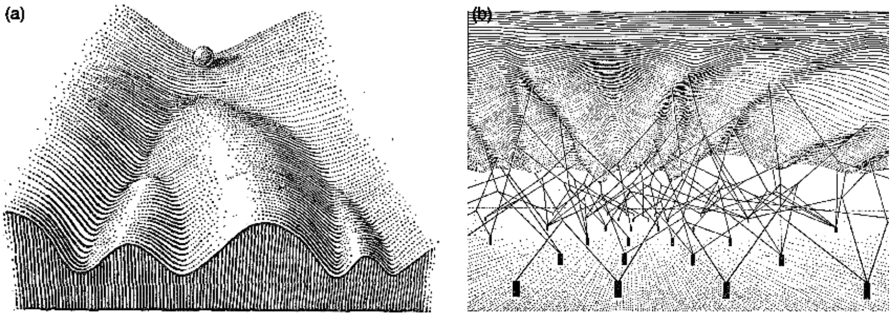


FIGURE 1.4: **Waddington's epigenetic landscape**. **a)** Waddington's visualisation of a cell (the ball) undergoing development by travelling through some landscape shaped by **b)** the gene regulatory network. The position of the cell on the landscape represents the cells developmental state. In **b)** the influence of genes, represented as black pins, on the landscape is represented by the 'ropes'. The collective action of genes shape the landscape. Figure reproduced from Waddington (2014).

differential equations,

$$\frac{dx}{dt} = F(x), \tag{1.1}$$

where $F(x)$ is a set of functions encoding the dependence of changes in the expression each gene, x_i, on the expression of all genes (including itself), x. Thus, $F(x)$ encodes the regulatory interactions in gene regulatory networks. Formally, **Eqn 1.1** is a dynamical system describing how the gene expression of a cell changes with time (Strogatz, 2018). More specifically, as cells exchange energy/mass with the environment, **Eqn 1.1** is a dissipative dynamical system (Strogatz, 2018; Greulich et al., 2020).

$F(x)$ encodes an organism's complete gene regulatory network and formalises Waddington's landscape, detailing the constraints on how a cells state can change over time. Moreover, $F(x)$ encodes the possible cell types of a given organism, as an emergent property of the dynamics of gene expression.

Imagine a random assortment of cells of varying initial states. Following the evolution of these cells' states over time, the cells converge towards isolated subsets of states, with cells of similar initial states tending to converge towards the same subset of states (Kauffman, 1969; Huang et al., 2005; Huang, 2012; Weinreb et al., 2018; Strogatz, 2018; Casey et al., 2020b; Newman, 2020; Greulich et al., 2020). These isolated subsets of states are termed the *attractors* of the dissipative dynamical system.

Broadly, an attractor of a dissipative dynamical system is an isolated subset of states toward which the system (cell) will evolve for a subset of initial states (gene expression profiles). Note that an attractor can consist of multiple,

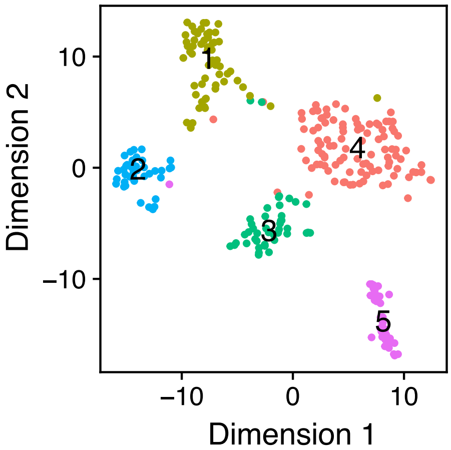


FIGURE 1.5: **Example of gene expression space**. Each point corresponds to the gene expression vector of an individual cell projected onto two dimensions using t-SNE (Hinton and Roweis, 2003). Cell type represented by colour of cell. Figure reproduced from Casey et al. (2020b).

contiguous states, i.e. gene expression profiles. In the case where an attractor does consist of numerous states, the system will undergo periodic changes in states, oscillating through the states making up the attractor (I will discuss such oscillations in more detail in **Chapter 5**). The subset of initial states that converge on a given attractor is termed that attractor's basin of attraction (Strogatz, 2018). All possible states belong to some basin of attraction, with complex dynamical systems admitting multiple possible attractors, so partitioning up states into different basins of attraction. Note that, if the system's initial state is in an attractor, the system will not leave the attractor.

Each attractor corresponds to a valley in Waddington's landscape, the stable sets of states towards which cells will evolve based on the constraints imposed by the gene regulatory network. Thus, the attractors of an organism's gene regulatory network correspond to cell types.

In the context of single-cell RNA-sequencing, dynamical system theory predicts that cell types will emerge as attractors in an abstract, high-dimensional gene expression space. In this space, the measured set of transcriptional counts is represented as a position vector (Casey et al., 2020b). Each dimension of the space corresponds to a gene and a cell's position along each dimension corresponds to the level of expression of the gene. Thus, each cell corresponds to a single position vector in this space. Cell types emerge as attractors in this space, subsets of gene expression space towards which cells will evolve with time. **Fig 1.5** illustrates gene expression space projected onto two dimensions.

Attractors provide a general explanatory theory for the emergence of cell types. While characterising an organism's complete gene regulatory network is impossible, dynamical systems theory predicts various testable phenomena. For example, Kauffman (1969) simulated randomly assembled Boolean networks (genes are restricted to being 'on' or 'off') of size and complexity similar to that found in nature. Kaufman's networks converged onto sets of oscillatory attractors, confirming the emergence of cellular attractors under reasonable conditions.

Furthermore, Huang et al. (2005) experimentally tested the phenomenological prediction from attractor theory that a given attractor will be approachable from different states. Huang et al. (2005) triggered the differentiation of one cell type into another using two distinct signals and followed the state of each population of cells using microarrays – a precursor to single-cell RNA-sequencing, measuring the gene expression of populations of cells as opposed to single-cells – to measure the expression of thousands of genes simultaneously. Initially distinct, the trajectory of the two populations converged before arriving at the same cell type, empirical evidence for the presence of a multi-dimensional attractor.

However, the theory of attractors relies on changes in gene expression being deterministic. As discussed in **Section 1.1.2**, gene expression is noisy, with changes in cell state having a stochastic component. Incorporating stochastic noise into **Eqn 1.1** gives rise to the Fokker-Planck equation, where cellular gene expression evolves by the interaction of deterministic and stochastic effects (Greulich et al., 2020). The Fokker-Planck equation does not encode attractors; instead, the equation predicts that certain subsets of states will be visited with high probability (Greulich et al., 2020). These high probability sets of states are analogous to noisy attractors.

Attractors, deterministic or noisy, provide the (often implicit) justification for the application of unsupervised clustering to single-cell RNA-sequencing data. As I will outline in detail in the following section, unsupervised clustering methods identify dense groupings of cells in gene expression space. These dense groupings correspond to the attractors of the whole-genome regulatory network. Thus, dynamical systems theory explains the ability of unsupervised clustering methods to classify cells into types *de novo*, without explicit reference to established marker genes.

1.2.3 Unsupervised Clustering

Unsupervised clustering methods identify groups of cells that are 'similar' in high-dimensional gene expression space, where each unsupervised clustering method's measure of cellular similarity is encoded mathematically via the so-called *objective function* (Jain, 2010; Trapnell, 2015; Kiselev et al., 2019). As discussed above, the regulatory interactions between genes result in cell types presenting as high-dimensional attractors in gene expression space. Therefore, by identifying cells that are similar in gene expression, unsupervised clustering methods can group cells into types.

In the rest of this section, I outline the specifics of how unsupervised clustering methods identify the dense groupings of cells in gene expression space. I focus on two clustering methods, *k*-means and the Louvain method (Lloyd, 1982; Newman and Girvan, 2004; Blondel et al., 2008). The Louvain method method is the best performing algorithm for clustering single-cell RNA-sequencing data; however, the method is mathematically complex, utilising a graphical representation of the data (Freytag et al., 2018; Duò et al., 2018; Luecken and Theis, 2019; Stuart et al., 2019; Blondel et al., 2008). Therefore, I will first discuss unsupervised clustering through the application of the more intuitive *k*-means algorithm (Lloyd, 1982; Jain, 2010).

k-means

The *k*-means method measures the similarity of cells using the Euclidean distance. The Euclidean distance is the 'shortest-path distance' between each pair of cells in gene expression space, i.e. between each pair of gene expression vectors. The distance is calculated across all genes, providing an overall measure of similarity between cells. The generic equation for the Euclidean distance between two vectors, p and q, each of length n is,

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}. \tag{1.2}$$

The Euclidean distance measures how similar or coordinated the gene expression profiles of a pair of cells are: the more genes that are similar in expression, the less the Euclidean distance between a pair of cells.

The *k*-means method seeks to maximise the overall similarity of cells within each cluster. The algorithm does so by quantifying the total dissimilarity of cells within each cluster as the sum of the Euclidean distances of each cell, x, to the centre of the cluster. The position of the centre of the cluster in gene space is given by the mean expression, μ_i, of all cells in the cluster i. For a chosen number of clusters, k (a hyperparameter that gives k-means its name), k-means identifies the set of k clusters that minimises the total distance of each cell from the centre of its cluster (Jain, 2010). This process is encoded by the k-means objective function,

$$\underset{S}{arg\,min} \sum_{i=1}^{k} \sum_{x \in S_i} ||x - \mu_i||^2, \tag{1.3}$$

where $S = S_1, \ldots, S_k$ is the set of k clusters and $arg\,min$ corresponds to identifying the clustering S that minimises the distance of each cell to the centre of its assigned cluster. Importantly, the number of clusters has to be chosen as a hyperparamter – minimising **Eqn 1.3** only identifies the optimal clustering with k clusters.

Optimising **Eqn 1.3** identifies a clustering of cells wherein each cell is maximally similar to the average expression of its assigned cluster. In doing so, the k-means method identifies dense groupings of cells, putatively identifying cell type-defining attractors in gene expression space. The total similarity of each cluster is defined with respect to the mean gene expression profile of each cluster, resulting in the k-means method finding clusters that are approximately hyperspherical in gene expression space (Kiselev et al., 2017). The explicit dependence on the average expression of the cluster is advantageous with respect to the cluster-wise classification of cells, as the mean expression of each cluster is broadly representative of the constituent cells of the cluster. However, as I will now discuss, the k-means method is a relatively inflexible approach to clustering, specifically with respect to quantifying within-cluster similarity.

Note briefly that the identification of the optimal clustering is computationally hard (formally, NP-hard) (Friedman et al., 2001). Accordingly, various heuristic algorithms have been developed that do not attempt to find the single best partition of cells; instead, they try to rapidly identify a good partition of cells, without any guarantee that this is the best partition (Jain, 2010). Unsupervised clustering is a hard problem generally, with many methods employing heuristics for more rapid clustering.

Difficulty with Distance

The *k*-means method utilises the Euclidean distance of each cell to the centre of its assigned cluster. However, not all distances are equally reliable. Estimations of cellular distances can be severely affected by both noise (technical and biological) and importantly, by gene regulatory interactions.

Consider the notion of cellular distance with respect to the gene regulatory network presented in **Fig 1.3**. Recall that the three transcription factors change expression level on exposure to the sonic hedgehog morphogen in a highly coordinated way: initially, *Pax6* is highly expressed, then *Olig2* and finally *Nkx2.2*. Discretising, this corresponds to three possible position vectors in gene expression space:

$$\begin{pmatrix} Pax6 \\ Olig2 \\ Nkx2.2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

The Euclidean distance between any pair of these is the same, $\sqrt{2}$, i.e. the Euclidean distance assumes that cells transition from one gene expression profile to another along the shortest possible path in gene expression space. However, in the above example, the regulatory interactions between genes constrain cells to pass through the three states in order: gene regulatory networks place highly non-linear constraints on changes in cellular gene expression (see **Fig 1.6** for illustration). These constraints make the Euclidean distance an unreliable measure of cellular similarity with respect to gene expression (Kim et al., 2019; Schiebinger et al., 2019).

However, the Euclidean distance is not uniformly unreliable. Transcription is a stochastic process, yet cellular function remains consistent, robust to small fluctuations in transcript numbers. For cellular function to be robust to small changes in gene expression, cells that are close in gene space must be more likely to be biologically similar (Raj et al., 2006; Casey et al., 2020b). Thus, the Euclidean distance as a measure of cellular similarity must be reliable over short distances, becoming increasingly unreliable over longer distances, i.e. changes in gene expression can be assumed to occur linearly (in straight lines) in gene expression space over short distances.

The assumption of linearity over short distances is not unique to single-cell biology. Instead, it is a common refrain in mathematics to approximate complex functions as locally linear; for example, Taylor's theorem relates to

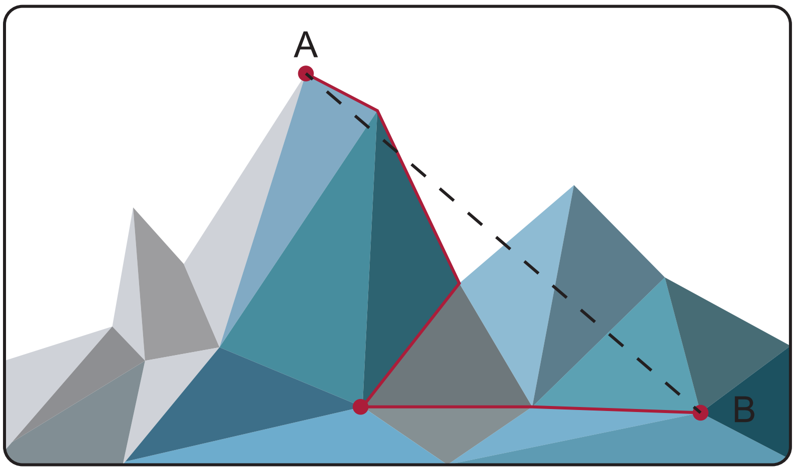


FIGURE 1.6: **Transitions in cellular gene expression**. Diagrammatic representation of the constraints imposed on changes in cell state by the non-linear interactions of the gene regulatory network. The red line represents a cell changing in expression from gene expression profile A to profile B via the path of steepest descent. This path differs from the path assumed by the Euclidean distance represented by the dashed black line. Figure reproduced with modification from Casey et al. (2020b).

the approximation of complex differentiable functions using *k*-order polynomials, where the first-order Taylor polynomial is the linear approximation of the function (Voit, 2017; Strogatz, 2018). Linear and polynomial approximations enable the analysis of otherwise too complex functions, such as those required to encode the non-linear dependencies between different gene's expressions (see **Eqn 1.1**).

The Louvain method accommodates the variable reliability of distance when quantifying within-cluster similarity, utilising only the distances between 'neighbouring' cells (in gene expression space). This accommodation provides a theoretical explanation for the relative success of the Louvain method in clustering single-cell RNA-sequencing data (Casey et al., 2020b). I will now discuss how the Louvain method quantifies within-cluster similarity, and how this leads to a robust identification of cell types in gene expression space.

Louvain Method

Instead of utilising the distances between all cells in a cluster, the Louvain method utilises only the distances between each cell and its nearest neighbours (Newman and Girvan, 2004; Reichardt and Bornholdt, 2006). Total within-cluster similarity is not defined with respect to a single point in gene expression space, as with *k*-means, but with respect to the distance between each cell and the nearest cells assigned to the same cluster. The Louvain method ignores the longer, less reliable distances between cells.

The Louvain method utilises a graphical representation of the data (Newman and Girvan, 2004; Reichardt and Bornholdt, 2006). An undirected graph, $G = \{V, E\}$ consists of a set of N vertices, $V = 1, \ldots, N$, and a set of G edges, $e_{ij} \in E$, where each edge is defined with respect to a pair of vertices i and j. The edges can be weighted or unweighted, with weighted edges assigned some coefficient, w_{ij}. In a weighted graph $w_{ij} = 0$ indicates that the vertices i and j are not connected by an edge.

Graphs offer a flexible representation of single-cell RNA-sequencing data. For a k-nearest-neighbours graph, each cell is encoded as a vertex in a graph, $v_i \in V$. Each cell has an edge to its k nearest cells by Euclidean distance (Von Luxburg, 2007). The (inverse of the) distance between cells is encoded as the weight of the edge. Thus, a kNN graph only encodes the subset of shorter, more reliable distances between cells (Stuart et al., 2019). Changes in cell state are explicitly treated as locally linearly, without any assumption of linearity over longer scales.

Note that the kNN graph is only one possible graphical representation of the data. Another, closely related form, is the shared-nearest-neighbours graph (sNN) (Von Luxburg, 2007). In the sNN graph, an edge is only included if both cells are k-nearest-neighbours of each other, avoiding biologically erroneous edges to outlier cells (Stuart et al., 2019).

The Louvain method quantifies the total within-cluster similarity of a given clustering, S, using a graph-theoretic measure termed modularity, Q_S. High modularity indicates that the cells of each cluster are densely interconnected, compared to the interconnection density expected at random (see **Eqn 1.4**). As cells are only connected if they have similar gene expression, high modularity indicates a clustering with high within-cluster similarity.

Modularity is calculated based on the adjacency matrix. The adjacency matrix of a graph, A, is a matrix representation of the graph. Each element of the matrix, A_{ij}, encodes the weight of the edge e_{ij}, provided an edge is present between cells i and j in the graph – $A_{ij} = 0$ implies no edge is present. The degree of a vertex (cell), v_i, is the sum of the weights of each edge connected to the vertex, $v_i = \sum_j A_{ij}$.

Based on the adjacency matrix, the modularity of a clustering of cells is found by (Newman and Girvan, 2004),

$$Q_S = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \gamma \frac{v_i v_j}{2m} \right] \delta(S_i, S_j), \tag{1.4}$$

where S is the clustering of the cells and $m = \frac{1}{2} \sum_{ij} A_{ij}$ is the total degree or weight of the graph. Note that $\delta(S_i, S_j)$ is 1 if cells i and j are in the same cluster and 0 otherwise, and that γ is the 'resolution' hyperparameter.

The Louvain method seeks to maximise the modularity of the graphical clustering (Blondel et al., 2008),

$$\underset{S}{arg\,max} \quad Q_S. \tag{1.5}$$

Modularity measures the density of connections within each cluster against the density expected at random. The resolution hyperparameter sets the density of connections within graphical clusters expected at random. Thus, the greater the resolution, the greater density expected, and so the greater number of (denser) clusters returned.

As with the k-means clustering method, the Louvain method identifies dense groupings of similar cells, putatively identifying cell type-defining attractors in gene expression space. Unlike with the k-means method, the resulting clusters can take any contiguous 'shape' in gene expression space, as modularity is maximised by ensuring cells are close to at least some other cells in the same cluster. This flexibility in shape enables the Louvain method to accommodate substantial intra-type heterogeneity (Blondel et al., 2008; Kiselev et al., 2019). However, the trade-off when compared to k-means is that the mean expression of each cluster does not necessarily provide a reasonable representation of the constituent cells of the cluster. Thus, the Louvain method can accommodate substantially more intra-type heterogeneity than k-means, albeit at the expense of an increase in conceptual complexity with respect to the cluster-based classification of cells into types.

By identifying sets of cells that are similar with respect to gene expression, unsupervised clustering identifies the attractors of gene expression space and the putative mix of cell types in a cellular population. Unsupervised clustering does not classify cells; instead, each cluster must be classified based on the relative expression of known marker genes. I will discuss how clusters are classified in the **Section 1.3**. First, however, it is important to note that while the correspondence between dense clusters of similar cells and cell types is theoretically justified (through dynamical systems theory), it is not trivial. The process of clustering single-cell RNA-sequencing data involves a series of pre-processing stages.

1.2.4 Pre-processing

Pre-processing of single-cell expression data is an essential part of single-cell clustering analysis (Vieth et al., 2019; Luecken and Theis, 2019). Single-cell data must be pre-processed as the differences in gene expression between cell types are not guaranteed to be the primary source of heterogeneity with respect to gene expression. As discussed, there are various alternative sources of heterogeneity, including biological noise and technical error. Unsupervised clustering methods can cluster cells based on these alternative sources of heterogeneity. For example, if the cell cycle has a larger influence on gene expression heterogeneity than cell type, unsupervised clustering would group cells by cell cycle stage (Xue et al., 2020). Alternatively, the total number of transcripts measured per cell could be the dominant source of gene expression heterogeneity, resulting in cells being clustered by cellular count depth (Dillies et al., 2013).

Single-cell RNA-sequencing pre-processing constitutes three main steps: normalisation, feature selection and linear dimension reduction (Vallejos et al., 2017; Luecken and Theis, 2019; Yip et al., 2019; Vieth et al., 2019). These steps seek to minimise the effects of alternative sources of heterogeneity. There are many competing methods and algorithms available for each pre-processing step, with a diversity of biological and mathematical assumptions about gene expression and cell type (Dillies et al., 2013; Brennecke et al., 2013; Grün et al., 2014; Love et al., 2014; Townes et al., 2019; Hafemeister and Satija, 2019; Breda et al., 2021; Lause et al., 2020; Ahlmann-Eltze and Huber, 2020; Andrews and Hemberg, 2019; Sparta et al., 2021; Jiang et al., 2016; Liu et al., 2020; Ranjan et al., 2021; Ascensión et al., 2021; Hotelling, 1933). However, the details of these methods are largely technical. Therefore, I will briefly outline each step in the box *"Single-cell Pre-processing"*, and survey the range of methods available for each step in more detail in **Appendix B**.

Together with unsupervised clustering, pre-processing methods form a computational pipeline for the automatic grouping of cells into cell types. However, as discussed, these clusters must still be classified. I will now discuss how clusters are identified based on the differential expression of marker genes between clusters.

Single-cell Pre-processing

Normalisation The total number of transcripts (count depth) measured per cell can vary by orders of magnitude within a single data set (Dillies et al., 2013). This variation is largely technical, resulting from inefficient measurement of individual transcripts (Papalexi and Satija, 2018). Normalisation methods estimate the relative level of gene expression in each cell, with the principal aim of minimising the effects of count depth variation (Vallejos et al., 2017).

Feature selection Cellular similarity is calculated based on the expression of all genes. Genes not variably expressed between cell types introduce noise in calculating cellular similarity with respect to the goal of clustering cells by type. Feature selection methods select only those genes likely to be informative in distinguishing between cell types for inclusion in clustering (Yip et al., 2019).

Linear dimension reduction Euclidean distance is calculated based on the sum of gene-wise differences in expression between cells. Over a large number of genes, small differences in expression accumulate, inflating the final measure of cell-to-cell distance into meaninglessness (Beyer et al., 1999). Linear dimension reduction techniques identify and collapse sets of linearly dependent genes into 'meta-genes', reducing the overall dimensionality of the data (Hotelling, 1933).

1.3 Classifying Clusters

In phenotypic classification, the type of each cell is identified via the presence or absence of a phenotypic feature. With respect to gene expression, cells are identified based on the expression of established marker genes. However, as measured by single-cell RNA-sequencing, cells of each type often only inconsistently express their associated marker genes, displaying substantial intra-type heterogeneity. Accordingly, cells are not classified individually, but on the level of clusters. Assuming that each cluster represents a single cell type, an assumption justified by dynamical systems theory, each cell can be classified by identifying its assigned cluster.

The marker genes expressed by each cluster are not identified absolutely but relative to the remaining clusters in the population. Specifically, marker genes are identified by differential gene expression analysis. For each gene and for each cluster, a hypothesis test is conducted against the remaining clusters, determining if there is any significant difference in expression (Robinson et al., 2010; Love et al., 2014; Soneson and Robinson, 2018; Wang et al., 2019; Luecken and Theis, 2019). Those genes that do significantly differ in expression are termed differentially expressed (or differentially expressing) genes.

Differential expression analysis overcomes intra-type heterogeneity by pooling the expression of individual cells. Droplet-based single-cell sequencing platforms typically measure thousands to tens-of-thousands of cells (Macosko et al., 2015; Klein et al., 2015; Papalexi and Satija, 2018; Svensson et al., 2020). Therefore, with a reasonable number of clusters, each cluster will likely contain at least hundreds of cells. Leveraging each cell as one sample of a cluster, hypothesis testing can robustly identify the marker genes differentially expressed with respect to each cluster.

Qualitatively, differential gene expression between clusters can be assessed via non-linear dimension reduction and visualisation. Explained in more detail in **Appendix B**, non-linear dimension reduction techniques, such as UMAP (Uniform Manifold Approximation and Projection) or tSNE (*t*-distributed Stochastic Neighbour Embedding) project the position of cells in gene expression space onto a 2-dimensional plot (Hinton and Roweis, 2003; McInnes et al., 2018). These reduction techniques preserve the distances between neighbouring cells at the expense of distances between less similar cells, so are able to project the cluster structure of the data onto two dimensions; see **Fig 1.8** for illustration. The expression of various marker genes can then be overlaid on the projection, allowing for a visual assessment of the differential expression of marker genes and providing "proof by visualisation" for the correspondence of unsupervised clusters with the phenotypic classification of cells; see **Fig 1.7** for an example of visual localisation of marker gene expression (Fox Keller, 2002; Luecken and Theis, 2019).

If none of the differentially expressed genes in a given cluster correspond to known marker genes, then the cluster potentially represents a novel cell type. The combination of single-cell RNA-sequencing, unsupervised clustering and differential expression analysis has led to an explosion in the number of known cell types, particularly through the whole organism cell atlas projects, the Tabulas' Sapiens, Mouse and Fly (The Tabula Sapiens Consortium and Quake, 2021; Tabula Muris Consortium et al., 2018; Li et al., 2021a). Each of these projects have identified hundreds of cell types. For example, the Tabula Sapiens project identified 400 distinct cell types, expanding on the 200-300 cell types that have traditionally been characterised in humans (Junqueira et al., 1992; The Tabula Sapiens Consortium and Quake, 2021). The ongoing consortium project, the Human Cell Atlas, aims to identify all cell types in

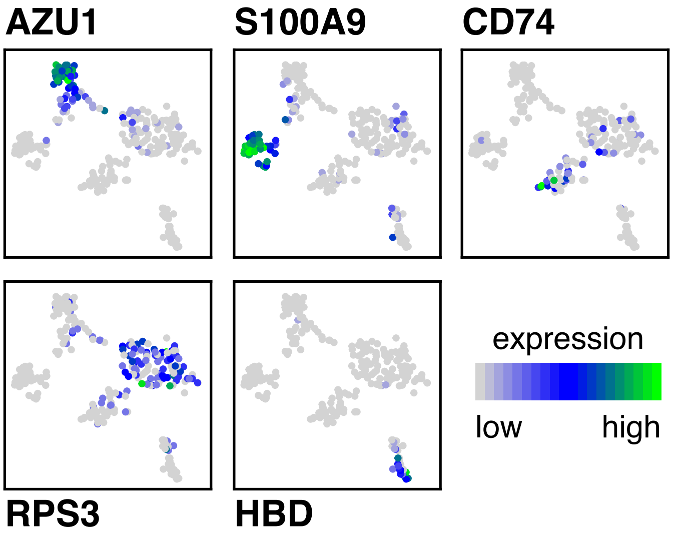


FIGURE 1.7: **Visualisation of differential gene expression.** Expression of known marker genes of five cell types overlaid on tSNE visualisation of human bone marrow population (Hinton and Roweis, 2003). Figure reproduced from Casey et al. (2020b).

humans, sequencing and clustering each organ in the adult human (Rozenblatt-Rosen et al., 2017).

Importantly, in the context of single-cell RNA-sequencing, differential expression analysis provides the *de facto* working definition of cell types: cells of different types form differentially expressed clusters in a population of cells. Where previously, cell types would be defined by the absolute expression of marker genes, with respect to single-cell sequencing data, known and novel cell types are identified via differential expression.

This break from the traditional approach to cellular classification – from classifying cells based on marker gene expression to classifying clusters based on differential marker gene expression – is in response to the gene expression heterogeneity revealed by single-cell RNA-sequencing. Therefore, it is important to ask how well does this change accommodate gene expression heterogeneity. Moreover, how does the differential expression conception of cell type relate to the phenotypic classification of cell type. This thesis aims to answer these questions by first quantifying the proportion of gene expression heterogeneity attributable to differential expression between clusters and then formalising the connection between differential gene expression analysis and the phenotypic classification of cells.

In **Chapter 2**, I will develop a novel information-theoretic framework for quantifying heterogeneity with respect to the expression of individual genes. I will define a novel measure of heterogeneity with respect to the measured expression of one gene or many. I will show that this measure is additively

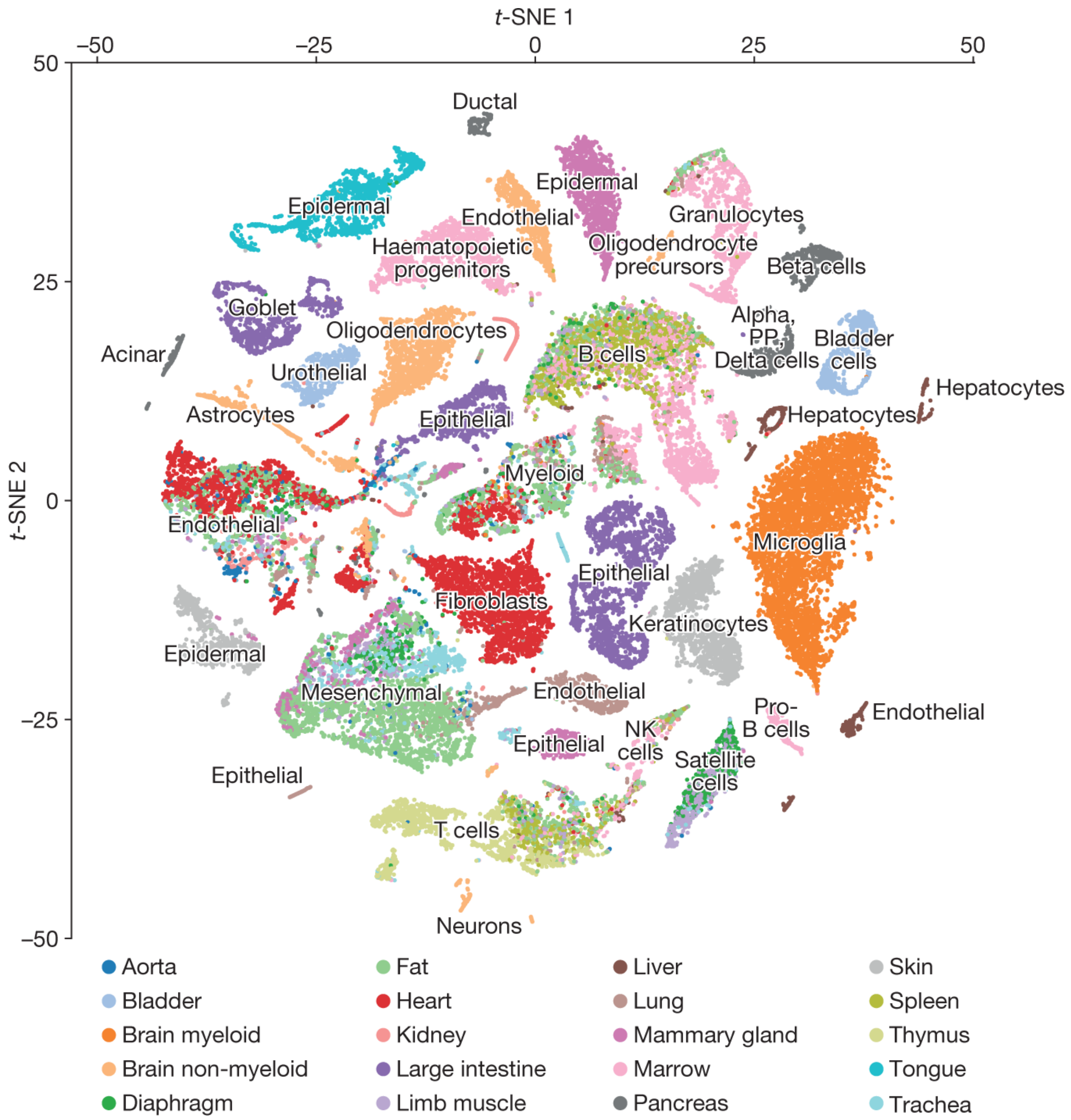


FIGURE 1.8: **tSNE of the Tabula Muris**. Two-dimensional tSNE embedding of Tabula Muris data, coloured by organ. Figure reproduced from Tabula Muris Consortium et al. (2018)

decomposable with respect to a given clustering of cells into that heterogeneity attributable to differential gene expression and that resulting from differences in expression within each cluster. Through this decomposition, I will quantify: (1) the proportion of heterogeneity attributable to differential gene expression; and (2) the divergence of a set of clusters from the assumption that cells of the same type are interchangeable with respect to measured gene expression.

Then in **Chapter 3**, I will utilise the mathematics developed in **Chapter 2** to develop a novel unsupervised clustering method. The unsupervised clustering method will directly identify the set of clusters that are maximally differentially expressed and are minimally divergent from the assumption that the cells of each cluster should be interchangeable. Importantly, the developed method is justified not by dynamical systems theory and attractors but by the principles of phenotypic classification.

Finally, in **Chapter 4**, I will return to the multivariate view of gene expression utilised by traditional unsupervised clustering methods. Unsupervised clustering methods, including the method developed in **Chapter 3**, require the number of clusters to be specified via an external hyperparameter: unsupervised clustering methods generally only identify the optimal clustering for a set number of clusters. Therefore, I will develop a tool for estimating the number of cell types in a population based on quantifying the total heterogeneity of a cellular population with respect to the joint distribution of gene expression. In doing so, I will introduce the mathematics of hypergraph theory to single-cell analysis.

Chapter 2

Information-Theoretic Clustering of Cell Types

Introduction

In phenotypic classification, cell types are identified based on the presence or absence of specific cellular phenotypes. Specifically, when observing cellular phenotype at the level of gene expression, cell types are identified based on the expression of specific genes. Genes whose expression is informative in classifying cell types are termed marker genes (Tanay and Regev, 2017; Luecken and Theis, 2019).

Single-cell RNA-sequencing measures gene expression genome-wide, counting the number of transcripts (mRNA molecules) expressed by each gene (Svensson, 2020). Single-cell RNA-sequencing experiments therefore provide sufficient data to classify all cells types distinguishable at the level of the transcriptome (Regev et al., 2017). However, single-cell sequencing has revealed that cells thought to be of the same type are often heterogeneous, i.e. distinguishable, with respect to the expression of established marker genes. Such intra-type heterogeneity prohibits reliable identification of cells based on the expression of any single gene.

Instead, cells are clustered into groups based on overall similarity in gene expression (Kiselev et al., 2019). Unsupervised clustering approaches leverage the expression of all genes (or a large subset of genes) to identify groups of cells with similar gene expression profiles. Clusters are then classified into different cell types based on which marker genes are differentially expressed with respect to each cluster, i.e. which marker genes are relatively up or down-regulated in each cluster (Love et al., 2014; Luecken and Theis, 2019). Thus, differential expression analysis provides a classification scheme for cells, defining cell types as differentially expressing non-overlapping sets of cells (Trapnell, 2015).

Differential expression is a major source of gene expression heterogeneity, resulting in cells of different types being distinguishable with respect to gene expression (Robinson et al., 2010; Love et al., 2014). I will call the heterogeneity resulting from differential gene expression inter-type heterogeneity, where the greater the differential expression of a given gene between different cell types, the greater the inter-type heterogeneity.

By phrasing differential expression in terms of heterogeneity, I aim to determine how successfully unsupervised clustering accommodates gene expression heterogeneity. Specifically, I aim to quantify the proportion of gene expression heterogeneity attributable to differential expression between a set of clusters. Given that cell types are defined by differential gene expression, I expect a large proportion of gene expression heterogeneity to be attributable to differential gene expression when clusters correspond to the set of cell types in a population.

This chapter quantifies the proportion of gene expression heterogeneity attributable to differential expression between clusters across all genes. To that end, in the first half of this chapter, I will formally develop a framework for quantifying heterogeneity with respect to the observed expression of each gene, measuring: the heterogeneity of a population of cells with respect to the expression of each gene; the proportion of gene expression heterogeneity resulting from differential gene expression between clusters; and, the proportion of gene expression heterogeneity resulting from differences in gene expression within each cluster. I will then extend the framework to quantifying heterogeneity with respect to the expression of many genes, deriving a single measure of the proportion of heterogeneity attributable to differential expression genome-wide.

In the second half of this chapter I will apply the developed framework to publicly available single-cell RNA-sequencing data sets. I will demonstrate the biological relevance of the proposed measure of gene expression heterogeneity and confirm that the established cellular classification of each data set explains a substantial proportion of gene expression heterogeneity (note that by established classification, I refer to the classification of cells provided with the original publications; see **Table 2.1** later on for details). Principally, I will establish between-cluster heterogeneity as a robust measure of differential expression, significantly associated with the *true* clustering of cells into cell types (where I assume that the established classification represents the true cellular classification for sequenced population).

The developed framework is based on the mathematics of information theory. Therefore, I will begin this chapter by briefly introducing information theory and specifically two key information-theoretic quantities, entropy and relative entropy (Shannon, 1948; Kullback and Leibler, 1951).

2.1 Information Theory

In this and the following chapter, I will quantify gene expression heterogeneity using tools from information theory. Information theory deals with the quantification of information. Originally developed by Claude Shannon to quantify the information content of messages, information theory has found broad application, particularly in its intersection with statistics (Shannon, 1948; Kullback and Leibler, 1951). I will here introduce the notions of information theory used in the rest of the chapter.

Entropy

Shannon's key insight was that the information content of a message depends on context, i.e. on how probable the message is (Shannon, 1948; Cover and Thomas, 2012; Smith and MacArthur, 2017). To illustrate this, consider the game of hangman, where a player tries to identify a word by sequentially guessing letters: what letters have the greatest potential information gain with respect to identifying the word? Words containing the letter Z are substantially rarer than those containing the letter E (letters capitalised without loss of generality to avoid clash with mathematical nomenclature). Thus, knowing a word contains the letter Z provides substantially more information about the identity of the word than knowing the word contains the letter E.

Shannon demonstrated that this potential gain in information could be quantified by the negative logarithm of the probability, $-\log x_i$, where x_i is the probability of the i^{th} possible outcome (Shannon, 1948). If the letter Z occurs in 0.44% of words, and the letter E in 11%, then the respective potential information gains are $-\log_2 0.0044 = 7.83$ bits and $-\log_2 0.11 = 3.18$ bits (Lewand, 2000). The typical units of information are termed bits, where n bits of information are sufficient to discriminate between 2^n possible choices of equal probability, e.g. gaining 1 bit of information in a game of hangman halves the number of words possible. Alternatively, information can be measured in nats, where the log is taken to base e, and where n nats of information are sufficient to discriminate between e^n possible choices of equal probability.

0However, in choosing a letter, there is a trade-off, as the less frequent a letter, the less chance there is to realise the potential information gain. Instead, information would be gained from knowing a letter is not present. For example, in choosing Z, there is a chance, with probability equal to 0.0044 (0.44%), to gain 7.83 bits of information if the letter is present; conversely, there is a non-overlapping chance, with probability equal to 0.9956 (99.56%), to gain $(-\log_2 0.9956 =)$ 0.0064 bits of information if the letter is not present. The average information gain from choosing the letter Z is given by $(0.0044)(-\log 0.0044) + (0.9956)(-\log 0.9956) = 0.041$ bits. More generally, the average information gain from realising a discrete random variable is,

$$H(X) = -\sum_{i=1}^{N} x_i \log x_i, \tag{2.1}$$

where $H(X)$ is the information entropy (also known as Shannon's entropy), X is a discrete random variable on the set of integers $\{1, ..., N\}$, with probabilities $p(X = i) = x_i$, and N is the total number of choices (in the above example, $N = 2$, where the letter is either present or not). By convention, $0 \cdot \log 0 = 0$, that is, there is zero information to be gained from impossible events.

Information entropy quantifies the expected information gain from realising a discrete random variable, taking into account the probability of an event occurring, and the information gain that event carries. In a game of hangman, to continue our example, the letter E should be chosen as it has a greater entropy, $-(0.11 \cdot \log_2 0.11 + 0.89 \cdot \log_2 0.89) = 0.50$ bits, than the letter Z, $-(0.0044 \cdot \log_2 0.0044 + 0.9956 \cdot \log_2 0.9956) = 0.041$ bits.

Shannon's entropy is non-negative: you cannot lose information from observing a random variable. Moreover, Shannon proved that $\log x$ is the only function (up to a multiplicative constant) to satisfy three basic constraints expected of information: monotonicity, independence, and branching (Shannon, 1948). Monotonicity ensures that all else being equal, the entropy of a random variable will increase with the number of possible outcomes, i.e. more information will be gained from realising a random variable with more possible outcomes, and independence ensures that the information

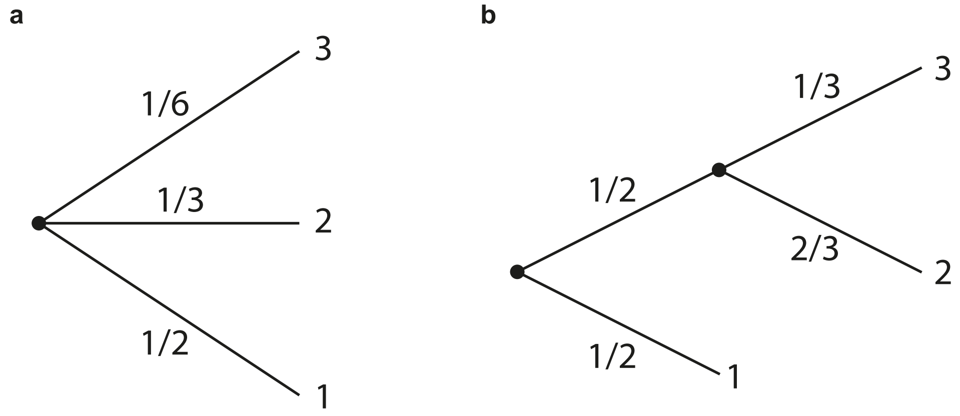


FIGURE 2.1: **Example of branching**. The discrete random variable X can be resolved in **a)** one or **b)** two stages. There is an assigned probability for each possible choice at each branch point. The entropy of both structures is the same, being the weighted sum of the entropy of each branch point.

gained from realising two independent random variables is the sum of the information gained from realising each of them separately. The constraint of branching concerns the different possible ways to realise a discrete random variable: "if a choice be broken down into two successive choices, the original [entropy] should be the weighted sum of the individual values of [entropy]", Shannon (1948).

I will illustrate branching by example. Consider the discrete random variable X on the set of integers $\{1,2,3\}$ with probabilities $\{\frac{1}{2},\frac{1}{3},\frac{1}{6}\}$. In realising X, I can either make the choice directly or through a series of hierarchical choices. For example I can first make the choice between the sets $\{1\}$ and $\{2,3\}$, and, if choosing the second one, make a further choice between $\{2\}$ and $\{3\}$. This results in two non-trivial branch-points for X, choosing between $\{1\}$ and $\{2,3\}$, and choosing between $\{2\}$ and $\{3\}$ (see **Fig 2.1**). Under the constraint of branching, the total entropy of X is the sum of the entropy at each branch point, weighted by the probability of arriving at that branch point,

$$H\left(\frac{1}{2},\frac{1}{3},\frac{1}{6}\right) = H\left(\frac{1}{2},\frac{1}{2}\right) + \frac{1}{2}H\left(\frac{2}{3},\frac{1}{3}\right),$$

where the first branch point with entropy $H\left(\frac{1}{2},\frac{1}{2}\right)$ is arrived at with probability 1. The constraint of branching leads to the property of additive decomposition, which I will discuss and utilise later in this chapter.

Discrete random variables with associated probabilities correspond to discrete probability distributions: it is clear (see **Eqn 2.1**) that entropy is a property of discrete probability distributions. The minimum value of entropy, 0, is

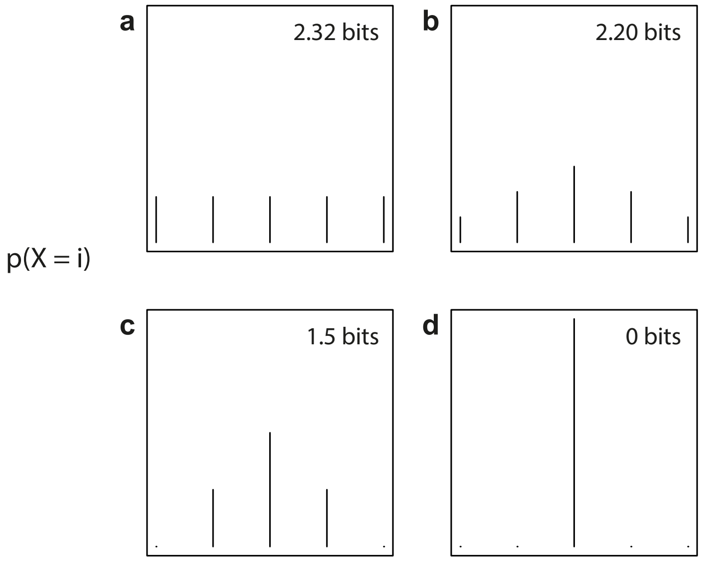


FIGURE 2.2: **Entropy of distributions**. Entropy of discrete distributions decreases from a maximum, $\log 5 = 2.32$ bits, with the uniform distribution (**a**) towards 0 bits with the point distribution (**d**).

achieved only when $p(X = i) = 1$ and $p(X = j) = 0$ for all $j \neq i$, i.e. no information can be gained by sampling from a probability distribution when there is only one possible outcome. Conversely, the value of entropy is at a maximum, $\log N$, for a uniform distribution over N possible outcomes, $p(X = i) = 1/N$ for all i.

Entropy is a measure of distribution uncertainty: the more uncertain a statistical process, the broader its probability distribution, the greater the entropy (see **Fig 2.2** for illustration). The more uncertain a statistical process, the more information is required to encode the possible outcomes. For example, consider the distribution of a given letter in the the game of hangman. If 50% of words contain a letter, the entropy is at a maximum, $\log 2$, and there is maximal uncertainty over whether the unknown word will contain the letter. Conversely, as more and more, or fewer and fewer, words contain the letter, the lower the entropy and the less uncertainty over whether the word will contain the letter. When all or no words contain the letter, there is zero entropy and zero uncertainty in whether the word contains the letter.

Relative Entropy

Consider quantifying the average information gain from a discrete random variable based on incorrect or approximate probabilities, for example, if an English game of hangman were to be played based on the frequency of letters in French. What is the reduction in average information gain, i.e. entropy, based on using the alternative probabilities?

Relative entropy, also known as the Kullback-Leibler Divergence, measures the information lost when realising one random variable, X, based on the probabilities of another, Y (Kullback and Leibler, 1951). Let X be a discrete random variable on the set of integers $\{1, ..., N\}$ with probabilities $p(X = i) = x_i$, and let Y be a discrete random variable on the set of integers $\{1, ..., N\}$ with probabilities $p(Y = i) = y_i$. The relative entropy of X with respect to Y is defined as,

$$D(X||Y) = \sum_{i=1}^{N} x_i \log\left(\frac{x_i}{y_i}\right), \tag{2.2}$$

with the provision that $q_i = 0$ implies $p_i = 0$, and that $0 \cdot \log(\frac{0}{0}) = 0$. For each possible choice i, $\log\,(x_i/y_i)$ quantifies the loss in average information gain due to the difference in the probabilities of X and Y.

Importantly, relative entropy provides a general measure of the difference between two discrete distributions (Cover and Thomas, 2012). In the following sections, I will develop a measure of gene expression heterogeneity based on relative entropy, quantifying heterogeneity as the divergence of the observed distribution of gene expression from the hypothetical case where all cells are interchangeable with respect to gene expression.

2.2 Quantifying Heterogeneity

Recall from **Section 1.1.2** that a set of cells are *heterogeneous* with respect to the expression of a gene g when the cells are distinguishable, i.e. not interchangeable, based on the observed expression of g. In this section, I develop a measure of heterogeneity with respect to the expression of each gene based on the relative entropy of the observed distribution of transcripts from the hypothetical case where all cells are interchangeable with respect to g.

To quantify heterogeneity using relative entropy, I must first choose a distribution on which to view the expression of each gene. Typically, each cell's measured set of transcript counts is represented as a position vector in a high-dimensional gene expression space (see **Section 1.2.2** for discussion on gene expression space). Concerning the expression of a single gene, cells are distributed with respect to the number of transcripts per cell (see **Fig 2.3** for

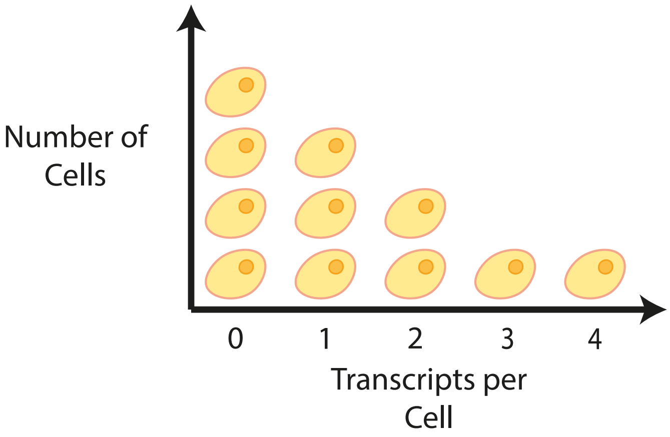


FIGURE 2.3: **Distribution of transcripts per cell**. Conventionally, with respect to a single gene, cells are distributed with respect to the number of transcripts per cell.

illustration). This distribution, of transcripts per cell, is the conventional view of a single gene's expression (Robinson et al., 2010; Brennecke et al., 2013; Grün et al., 2014; Love et al., 2014; Hafemeister and Satija, 2019; Townes et al., 2019; Liu et al., 2020).

However, I instead introduce and utilise an alternative distribution (novel in the context of single-cell analysis), inspired by experimental cell biology. Imagine a population of cells as viewed under a microscope, where the expression of some gene of interest, g, has been marked, e.g. by fluorescent tagging or antibody staining. Under the microscope, a population of cells appears as a field of individual cells, each with some read-off of the expression level of g. If the set of cells are (approximately) interchangeable with respect to the expression of g, then the expression level will be (approximately) uniform across the field of cells. Conversely, the more heterogeneous the set of cells are with respect to the expression of g, the greater expression levels will deviate from uniformity

I construct the mathematical analogue of the view under the microscope for single-cell count data. Consider the expression of a gene g across a population of N cells. Let m_i^g be the number of transcripts of gene g measured in cell i and let $\sum_{i=1}^N m_i^g = M^g$ be the total number of transcripts of gene g measured. Note that the measured number of transcripts may differ from the true number due to error in the measurement process.

Now consider the stochastic process of assigning the M^g observed transcripts to the N cells profiled. Let X^g be the discrete probability distribution on the set of cells $\{1, 2, \ldots, N\}$, where $p(X^g = i) = x_i^g$ is the probability of assigning a

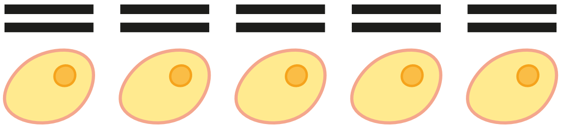


FIGURE 2.4: **Uniform distribution of transcripts on the set of cells**. Transcripts (black bars) are distributed uniformly on a set of five cells. The set of cells are interchangeable with respect to the expression this gene.

transcript of gene g to cell i. I estimate x_i^g as $p_i^g = m_i^g/M$, the proportion of the total number of transcripts of g measured in cell i. Note that p_i^g is the maximum likelihood estimate of x_i^g, see Townes et al. (2019).

This process of stochastic assignment constitutes a generative model of gene expression. If the cells of the population are statistically interchangeable with respect to the expression of g, then the M^g transcripts of g will be assigned to the N cells independently and with equal probability, i.e. each transcript has a probability of $1/N$ of being assigned to a given cell. This assignment corresponds to the (discrete) uniform distribution U, where U is the discrete probability distribution on the set of cells $\{1, \ldots, N\}$, with probability $1/N$ for $i = 1, \ldots, N$. See **Fig 2.4** for an illustration of U.

If the population of cells is heterogeneous with respect to the expression of g, then transcripts will not be assigned uniformly. If the heterogeneity is due to differential expression between a mix of cell types, transcripts will instead be assigned preferentially to distinct subsets of cells, see **Fig 2.5**. Heterogeneity in observed gene expression can therefore be quantified in terms of deviation from the uniform distribution of transcripts, U.

I quantify the deviation from the uniform distribution via relative entropy. From **Eqn 2.2**, I can derive the relative entropy of the observed transcript distribution from U:

$$I(g) = \sum_{i=1}^{N} x_i^g \log\left(\frac{x_i^g}{1/N}\right) = \sum_{i=1}^{N} x_i^g \log\left(N x_i^g\right). \tag{2.3}$$

I call $I(g)$ the *population heterogeneity* of g, as it measures the heterogeneity of the population of cells with respect to the expression of g. More precisely, $I(g)$ measures the information lost by approximating the observed expression distribution with the uniform distribution, or conversely, the amount of information required to encode the observed heterogeneity with respect to the expression of g.

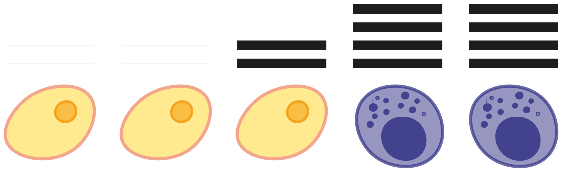


FIGURE 2.5: **Gene expression heterogeneity**. Transcripts (black bars) are distributed heterogeneously on a set of five cells of two differing types, where I define heterogeneity as deviation from uniformity. Aside from the the pair of purple cells, the cells are not interchangeable with respect to the transcript distribution of this gene.

Intuitively, when a cellular population is unstructured with respect to the expression of g, i.e. in the absence of a mixture of differentially expressing cell types, then the assumption of uniformity is correct and $I(g) = 0$, its minimum value. Conversely, the maximum value of $I(g)$, $\log N$, is reached when a gene is only expressed in a single cell ($x_i^g = 0$ for all $i \neq j$, $x_j^g = 1$). Note that $I(g)$ does not require any *a priori* assumptions about the particular expression pattern of g in the population, i.e. $I(g)$ is a minimally assuming measure of gene expression heterogeneity (with respect to the maximum entropy principle of model construction; see Jaynes (1957) for discussion of maximum entropy principle).

Broadly, $I(g)$ increases as fewer cells express a given gene; see **Fig 2.6** for illustration. Accordingly, $I(g)$ is particularly sensitive to differential gene expression between discrete subsets of cells (see **Fig 2.6**, top right). Such differential expression distinguishes the respective subsets of cells with respect to the expression of g.

$I(g)$ directly relates heterogeneity to Shannon's entropy and uncertainty in cellular classification. Expanding **Eqn 2.3** (using $\log(xy) = \log(x)\log(y)$), I get

$$I(g) = \sum_{i=1}^{N} x_i^g \log(N) + \sum_{i=1}^{N} x_i^g \log\left(x_i^g\right) = \log(N) - H(X^g) \tag{2.4}$$

where $H(X^g)$ is the entropy of the observed distribution of transcripts of g on the set of cells.

$H(X^g)$ measures the uncertainty in determining which cell expresses a given transcript of gene g. $H(X^g) = \log N$ indicates maximum possible uncertainty, occurring when cells uniformly express a given gene. Thus, $I(g)$ measures the reduction in uncertainty with respect to the distribution of transcripts. Increasing gene expression heterogeneity reduces the uncertainty over which cell expresses a given transcript of g, as illustrated in **Fig 2.6**.

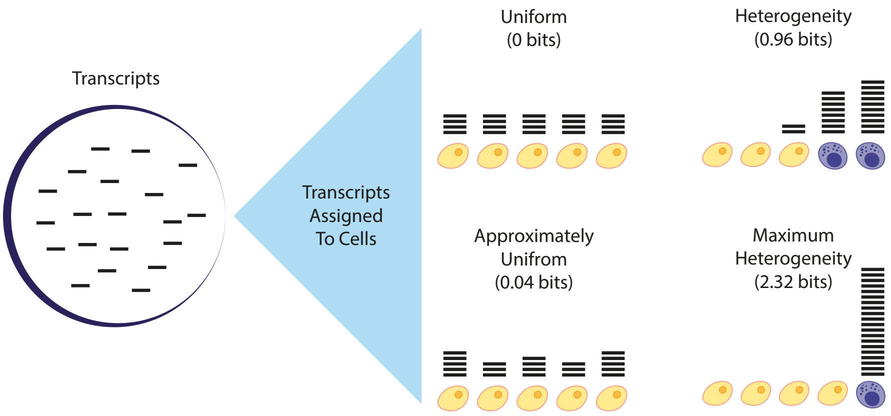

Transcripts
Transcripts Assigned To Cells
Uniform (0 bits)
Heterogeneity (0.96 bits)
Approximately Unifrom (0.04 bits)
Maximum Heterogeneity (2.32 bits)

FIGURE 2.6: **Quantifying heterogeneity**. Transcripts (black bars) are distributed in a population of cells. If the gene is expressed uniformly, then the heterogeneity is zero (top left). As the population of cells increasingly deviates from uniformly expressing the gene, the measured $I(g)$ increases, reaching a maximum of $\log N$, where N is the number of cells, when only one cell expresses the gene (bottom right).

In the next section, I will repeat the above stochastic process but assign transcripts first to clusters then to cells. By doing so, I can define two further quantities related to $I(g)$, measuring first the gene expression heterogeneity between clusters and then the expression heterogeneity between cells assigned to each cluster.

2.3 Heterogeneity by Cluster

In this section, I develop two further measures of gene expression heterogeneity with respect to the decomposition of a cellular population into non-overlapping clusters, where each cluster represents a (putative) cell type. The first measure quantifies the expression heterogeneity attributable to differential gene expression between clusters, and the second quantifies the expression heterogeneity attributable to differences in gene expression between cells within each cluster. Following the introduction of these measures, I show that $I(g)$ is the sum of these two measures, i.e. that population heterogeneity is a sum of the heterogeneity between clusters and the heterogeneity within each cluster. Together, the three measures of gene expression heterogeneity provide a quantitative framework for assessing the proportion of expression heterogeneity attributable to differential expression with respect to a given clustering.

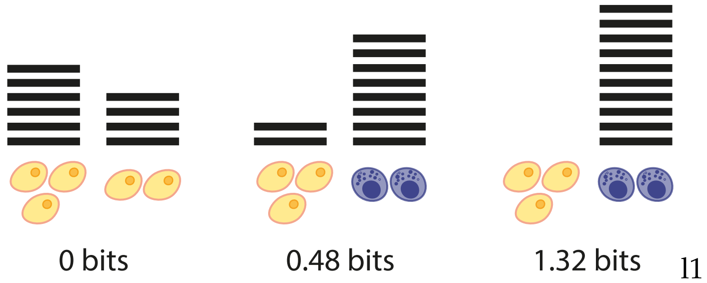


FIGURE 2.7: **Inter-cluster heterogeneity**. Transcripts (black bars) are distributed (*left*) uniformly, (*middle*) differentially between clusters, and (*right*) exactly by cluster. The value of $H_S(g)$ increases as the uncertainty in classifying clusters with respect to gene expression decreases. Importantly, $H_S(g)$ is concerned only the total number of transcripts assigned to each cluster – a gene can be expressed uniformly with respect to a set of cluster, while deviating from uniformity within each cluster. For example, a single cell in a given cluster could express all copies of the transcripts assigned to that cluster.

2.3.1 Inter-cluster Heterogeneity

Consider again the stochastic process of assigning M^g transcripts to N cells. Now consider grouping cells into C clusters $S_1, \ldots, S_C$ of sizes $N_1, \ldots, N_C$, where $\sum_{k=1}^{C} N_k = N$, so that transcripts are assigned to clusters as opposed to cells. Each cell is unambiguously assigned to one of the C non-intersecting clusters. Let Y^g be the discrete probability distribution on the set of clusters $S = \{S_1, \ldots, S_C\}$, where $p(Y^g = S_k) = y_i^g$ is the probability of assigning a transcript of gene g to cluster S_k. Recalling that x_i^g is the probability of assigning a transcript of gene g to cell i, the total probability of assigning a transcript of g to S_k is $y_k^g = \sum_{i \in S_k} x_i^g$, i.e. the total proportion of transcripts of gene g measured in the cells assigned to cluster S_k.

Recall that if the cells of the population are statistically interchangeable with respect to the expression of g, then $x_i^g = 1/N$ for all i in $1 \leq i \leq N$. Thus, when all cells in a population are interchangeable, the total probability of a transcript of g being assigned to the cluster S_k is proportional to the the number of cells assigned to the cluster, N_k. This corresponds to the (discrete) uniform distribution U_{group}, where U_{group} is the discrete probability distribution on the set $S = \{S_1, \ldots, S_C\}$ with probabilities N_k/N for $k = 1, \ldots, C$. Carrying forward the example of U in **Fig 2.4**, I illustrate an example of U_{group} in **Fig 2.7** (*left*).

Importantly, a gene can be expressed uniformly with respect to a set of clusters while deviating from uniformity within each cluster. For example, in **Fig 2.7** (*left*), the transcripts expressed by each cluster could be expressed in any combination by the cells of each cluster. U_{group} is only defined with respect to the total number of transcripts assigned to each cluster.

I can again quantify the deviation from the uniform distribution via relative entropy. The relative entropy of the observed cluster-wise distribution of transcripts, Y^g, relative to U_{group} is,

$$H_S(g) = \sum_{k=1}^{C} y_k^g \log\left(\frac{y_k^g}{N_k/N}\right), \tag{2.5}$$

where $H_S(g)$ measures the heterogeneity of the set of clusters with respect to the expression of g. I call $H_S(g)$ *inter-cluster heterogeneity*, as it quantifies how distinguishable, i.e. heterogeneous, clusters are with respect to the expression of g. The clustering S is not guaranteed to coincide with the true mix of cell types in a population: when the clustering S does coincide with the set of cell types, $H_S(g)$ measures inter-type heterogeneity with respect to the expression of g.

$H_S(g)$ quantifies the differential expression between clusters: the greater the difference in mean expression between clusters, the greater the inter-cluster heterogeneity. Each additive term making up $H_S(g)$ quantifies the divergence of an individual cluster from the assumption that the set of clusters are interchangeable with respect to g. As one cluster (or several) preferentially expresses g, the other clusters must comparatively down-regulate g. Thus, $H_S(g)$ behaves similarly to $I(g)$, where the more restricted expression of g is between clusters, the greater the measured heterogeneity; see **Fig 2.7** for illustration.

As with $I(g)$, $H_S(g)$ has an information-theoretic interpretation as the amount of information lost by assuming a set of clusters are interchangeable with respect to the measured expression of g. Moreover, $H_S(g)$ can be interpreted as quantifying the uncertainty in assigning transcripts to clusters. The less uncertain the assignment (i.e. the greater certainty over which cluster expresses a given transcript of g) the greater the inter-cluster heterogeneity.

$H_S(g)$ provides an absolute measure of the heterogeneity attributable to differential gene expression. However, to contextualise $H_S(g)$ and determine the proportion of gene expression heterogeneity attributable to differential gene expression, a further measure, quantifying the heterogeneity resulting from differences in gene expression within each cluster is required. In the following section, I will develop this further measure of gene expression heterogeneity.

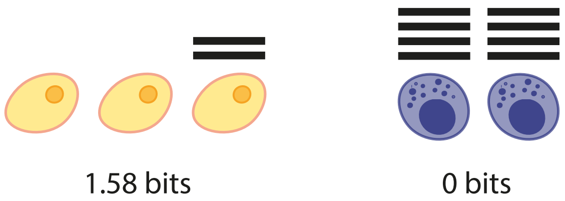


FIGURE 2.8: **Intra-cluster heterogeneity**. Diagram representing two clusters, where the first cluster (*left*) is an example of within-cluster heterogeneity with respect gene expression, $D(Z_k^g||U_k) = 1.58$ bits and the second cluster (*right*) is an example of within-cluster uniformity, $D(Z_k^g||U_k) = 0$ bits.

2.3.2 Intra-cluster Heterogeneity

Consider again the stochastic process of assigning transcripts to C clusters, focusing on the assignment of transcripts to the N_k cells within each cluster, S_k. Let Z_k^g be the discrete probability distribution on the set of cells $i \in S_k$ (i.e., the cells assigned to cluster S_k), with probabilities $z_i^g = x_i^g / y_k^g$, where z_i^g is the conditional probability of a transcript being assigned to cell i given that the transcript has been assigned to cluster S_k.

If the cells of the cluster are statistically interchangeable with respect to the expression of g, then the transcripts of g assigned to cluster S_k will be assigned uniformly to the N_k cells of cluster S_k. This corresponds to the uniform distribution U_k, where U_k is the discrete probability distribution defined on the set $i \in S_k$ with probabilities $1/N_k$ for $k = 1, \dots, C$.

U_k differs from U and U_{group} in applying to only a subset of the cellular population, with C distinct localised uniform distributions, U_1 to U_C, as illustrated in **Fig 2.5**. Accordingly, I separately quantify the heterogeneity of each cluster S_k, measuring the relative entropy of the observed transcript distribution Z_k^g from the uniform distribution U_k,

$$D(Z_k^g||U_k) = \sum_{i \in S_k} \frac{x_i^g}{y_k^g} \log \left(\frac{x_i^g / y_k^g}{1/N_k} \right), \tag{2.6}$$

$$= \sum_{i \in S_k} z_i^g \log \left(N_k \, z_i^g \right). \tag{2.7}$$

$D(Z_k^g||U_k)$ measures the divergence of the observed transcript distribution from the assumption of consistent, uniform expression within a given cluster. When all cells are assigned to a single cluster, $D(Z_k^g||U_k) = I(g)$.

The expression heterogeneity of a given gene within each cluster can be summed, with the contribution of each cluster to the overall expression heterogeneity weighted by the proportion of transcripts assigned to the cluster (this weighting derives from the branching property of entropy, and will be further discussed in **Section 2.3.3**). Accordingly, the total *intra-cluster heterogeneity* is given by,

$$h_S(g) = \sum_{k}^{C} y_k^g \, D(Z_k^g || U_k), \tag{2.8}$$

where $h_S(g)$ measures the heterogeneity resulting from differences in expression of g within each cluster. Importantly, $h_S(g)$ quantifies the total expression heterogeneity within each cluster with respect to a given gene. The measure is weighted by the proportion of transcripts assigned to each cluster; therefore, clusters with greater expression of a given gene can contribute more to the total intra-cluster heterogeneity.

In terms of information, $h_S(g)$ quantifies the amount of information lost by assuming that the cells within each cluster are interchangeable with respect to the measured expression of g. The further cells within each cluster are from being interchangeable, the greater the intra-cluster heterogeneity. Conversely, when $h_S(g) = 0$, the cells of each cluster are exactly interchangeable, with g expressed consistently and uniformly within each cluster.

Based on $h_S(g)$ and the other measures of gene expression heterogeneity developed above, I will construct a framework for quantifying the proportion of expression heterogeneity attributable to differential gene expression. Specifically, population heterogeneity, $I(g)$, is the sum of inter-cluster heterogeneity, $H_S(g)$, and intra-cluster heterogeneity, $h_S(g)$. Thus, $H_S(g)$ quantifies the proportion of $I(g)$ attributable to differential expression, with $h_S(S)$ quantifying the remaining, unattributed gene expression heterogeneity. In the following sections, I will derive this relation, showing that it follows from the constraint of branching.

2.3.3 Additive Decomposition

I now have three measures of heterogeneity, measuring 1) the total population heterogeneity, 2) the heterogeneity due to differential expression between clusters, and 3) the heterogeneity remaining within each cluster. The measures

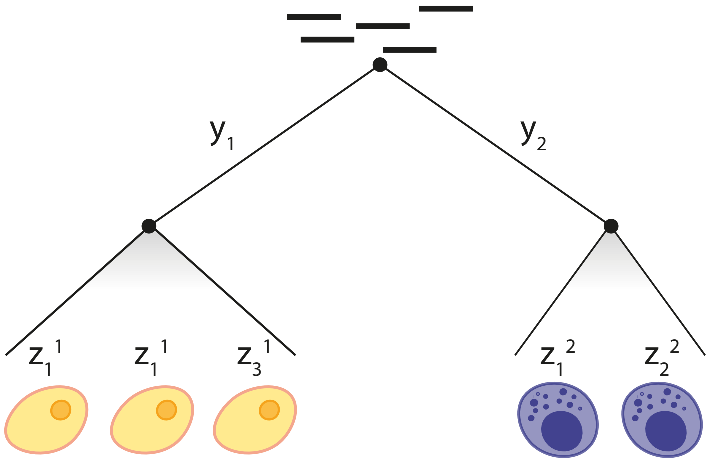


FIGURE 2.9: **Branching in transcript assignment**. Branching underlies the additive decomposition of inter- and intra-cluster heterogeneity. Transcripts (black bars) are first assigned to one of two clusters, according to the distribution Y. Transcripts are then assigned to a cell within each cluster k, according to the distribution Z_k. y_k is the proportion of transcripts assigned to cluster k and z_i^k the proportion of transcripts assigned to cell i in cluster k. The entropy of the population, $H(X)$ (where $x_i \in X$ is the proportion of transcripts assigned to cell i) is equal to the sum $H(Y) + y_1 H(Z^1) + y_2 H(Z^2)$.

are related through the constraint of branching, discussed earlier in **Section 2.1**.

To apply the constraint of branching to the context of clustering, consider again the stochastic assignment of a transcript to one of N cells, a process described by the distribution X. Based on a clustering, S, I separate the assignment process into two stages: first assigning transcripts to clusters, a process described by the distribution Y, then to the cells within each cluster, a non-overlapping set of processes described by the C distributions Z_k (see **Fig 2.9** for illustration).

By the constraint of branching, the entropy of X is equal to the sum of the entropies of Y and Z_k, for k = 1,. . . , C, weighted by the fraction of transcripts assigned to each distribution. The weighting of the entropies is

$$H(X) = H(Y) + \sum_{k=1}^{C} y_k^g H(Z_k). \tag{2.9}$$

This relation (which holds with respect to any clustering, or grouping of cells, S) is called *additive decomposition*. In other words, the entropy of the distribution X decomposes into additive components of the entropies of Y and each Z_k.

Relative entropy is similarly additively decomposable (Theil, 1967; Shorrocks, 1980). Accordingly, population heterogeneity is equal to the weighted sum of inter-cluster and intra-cluster heterogeneities,

$$I(g) = H_S(g) + h_S(g), \tag{2.10}$$

where the heterogeneity with respect to the expression of g must arise from either differential expression between clusters, or from differences in expression within each cluster (see **Section 2.3.4** for a full derivation). In the absence of an assigned clustering, all cells can be considered as belonging to a single cluster so that $h_S(g) = I(g)$ and $H_S(g) = 0$ (equally reasonably, in the absence of an assigned clustering, each of the N cells in a population could be considered as uniquely belonging to one of N clusters, so that $h_S(g) = 0$ and $H_S(g) = I(g)$).

Eqn 2.10 provides a framework for quantifying the proportion of heterogeneity attributable to differential expression, with respect to the expression of a single gene. Additive decomposition guarantees that all heterogeneity with respect to a given gene is attributable to either the differential expression between clusters or differences in expression within clusters. Therefore, $H_S(g)$ measures the heterogeneity attributable to differential expression as part of the total gene expression heterogeneity, $I(g)$. In doing so, $H_S(g)$ provides a basis for assessing the success of a given clustering in accommodating expression heterogeneity with respect to a single gene.

For example, consider the population illustrated in **Fig 2.10**. The population of cells is heterogeneous with respect to gene expression, with $I(g) = 0.8$ bits. This heterogeneity must result from either differential expression between clusters, or differences in expression within clusters. For the proposed clustering of cells in **Fig 2.10**, the majority of gene expression heterogeneity results from differential expression, $H_S(g) = 0.48$ bits, with the remaining heterogeneity resulting from differences in gene expression within one of the clusters, with $h_S(g) = 0.32$ bits.

In the next section, I will mathematically derive the property of additive decomposition for relative entropy. Following that derivation, I will extend each measure of heterogeneity to the case of many genes, developing whole-genome measures of population, inter-cluster and intra-cluster heterogeneity. Through this extension, I can measure the proportion of gene expression heterogeneity attributable to differential expression across all genes, with respect to a given clustering.

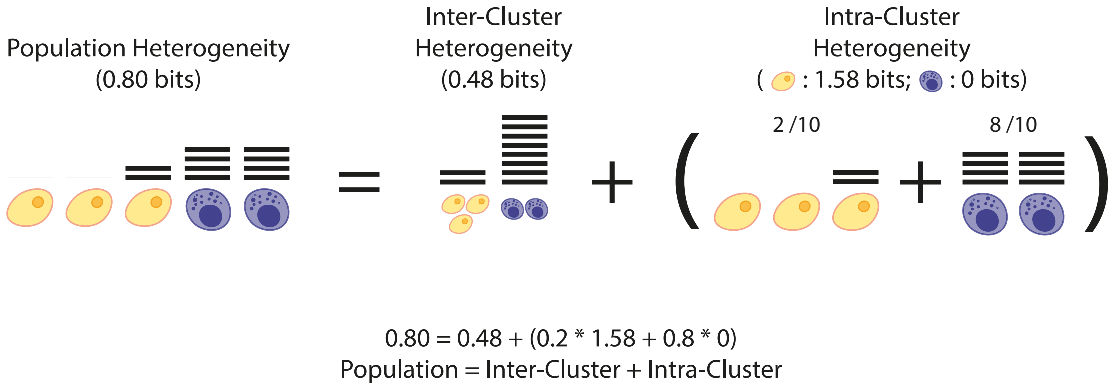


FIGURE 2.10: **Additive decomposition**. Diagrammatic representation of the additive decomposition of population heterogeneity into inter and intra-cluster heterogeneity. Population heterogeneity is the sum of inter-cluster heterogeneity and intra-cluster heterogeneities, where the contribution to intra-cluster heterogeneity from each cluster is weighted by the proportion of transcripts assigned to that cluster.

2.3.4 Mathematical Derivation

I will now go through a self-contained derivation of additive decomposition for relative entropy (cf. Theil (1967)), returning **Eqn 2.10**.

Recall that $x_i \in X$ is the proportion of transcripts of assigned to cell i, that $y_k \in Y$ is the proportion of transcripts of assigned to the cells of cluster S_k, and that $z_i^k \in Z_k$ is the proportion of transcripts of assigned to the cell i in cluster S_k. Note that

$$\sum_{i \in S_k} z_i = \sum_{i \in S_k} \frac{x_i}{y_k} = \frac{1}{y_k} \sum_{i \in S_k} x_i = 1, \tag{2.11}$$

so z_i, for $i \in S_k$ form a (discrete) probability distribution on S_k, for each $k = 1, \ldots, C$.

$I(g)$ may be rewritten in terms of Y and Z_k, as follows:

$$
\begin{aligned}
I(g) &= \sum_{i=1}^{N} x_i \log\left(N x_i\right), && (2.12)\\
&= \log\left(N\right) - \sum_{i=1}^{N} x_i \log\left(\frac{1}{x_i}\right), && (2.13)\\
&= \log\left(N\right) - \sum_{k=1}^{C} \sum_{i \in S_k} x_i \log\left(\frac{1}{x_i}\right), && (2.14)\\
&= \log\left(N\right) - \sum_{k=1}^{C} y_k \sum_{i \in S_k} \frac{x_i}{y_k} \left(\log\left(\frac{1}{x_i/y_k}\right) + \log\left(\frac{1}{y_k}\right)\right), && (2.15)\\
&= \log\left(N\right) - \sum_{k=1}^{C} y_k \underbrace{\sum_{i \in S_k} \frac{x_i}{y_k} \log\left(\frac{1}{x_i/y_k}\right)}_{H(Z_k)} - \sum_{k=1}^{C} y_k \sum_{i \in S_k} \frac{x_i}{y_k} \log\left(\frac{1}{y_k}\right), && (2.16)\\
&= \log\left(N\right) - \sum_{k=1}^{C} y_k\, H(Z_k) - \underbrace{\sum_{k=1}^{C} \log\left(\frac{1}{y_k}\right) \overbrace{\sum_{i \in S_k} x_i}^{y_k},}_{H(Y)} && (2.17)\\
&= \log\left(N\right) - \sum_{k=1}^{C} y_k\, H(Z_k) - H(Y), && (2.18)\\
&= \underbrace{\log\left(N\right) - H(Y) - \sum_{k=1}^{C} y_k \log\left(N_k\right)}_{A} + \underbrace{\sum_{k=1}^{C} y_k \log\left(N_k\right) - \sum_{k=1}^{M} y_k\, H(Z_k)\,.}_{B} && (2.19)
\end{aligned}
$$

Expression A may be rewritten as:

$$
\begin{aligned}
A &= \log\left(N\right) - H(Y) - \sum_{k=1}^{C} y_k \log\left(N_k\right), && (2.20)\\
&= \sum_{k=1}^{C} y_k \log\left(N\right) - \sum_{k=1}^{C} y_k \log\left(\frac{1}{y_k}\right) - \sum_{k=1}^{C} y_k \log\left(N_k\right), && (2.21)\\
&= \sum_{k=1}^{C} y_k \log\left(\frac{y_k}{N_k/N}\right), && (2.22)\\
&= H_S(g). && (2.23)
\end{aligned}
$$

Expression B may be rewritten as:

$$\begin{aligned}
B &= \sum_{k=1}^{C} y_k \log\left(N_k\right) - \sum_{k=1}^{C} y_k\, H(Z_k), && (2.24)\\
&= \sum_{k=1}^{C} y_k \left(\log\left(N_k\right) - H(Z_k)\right), && (2.25)\\
&= \sum_{k=1}^{C} y_k \sum_{i \in S_k} \frac{x_i}{y_k} \log\left(N_k \frac{x_i}{y_k}\right), && (2.26)\\
&= \sum_{k=1}^{C} y_k \sum_{i \in S_k} \frac{x_i}{y_k} \log\left(\frac{x_i/y_k}{1/N_k}\right), && (2.27)\\
&= \sum_{k=1}^{C} y_k\, D(Z_k||U_k) && (2.28)\\
&= h_S(g). && (2.29)
\end{aligned}$$

Therefore, $I(g)$ may be expressed as:

$$I(g) = H_S(g) + h_S(g). \tag{2.30}$$

2.4 Heterogeneity of Many Genes

I have introduced how the heterogeneity attributable to differential expression can be quantified with respect to a single gene. I will now extend this quantification to many genes, developing a measure of the total amount of heterogeneity attributable to differential expression across all genes. Given that cell types are differentially expressing subsets of cells, I expect that the better a given clustering represents the true set of cell types in a population, the more heterogeneity should be attributable to differential expression genome-wide.

The intuitive solution to measuring heterogeneity genome-wide is to simply sum across the gene-wise measures. Indeed, the intuitive approach is correct, up to a multiplicative factor: each measure of heterogeneity, $I(g)$, $H_S(g)$ and $h_S(g)$ is additive across genes, so that the total heterogeneity based on a set of G genes is simply $I = \sum_g^G I(g)$, $H_S = \sum_g^G H_S(g)$ or $h_S = \sum_g^G h_S(g)$.

This gene-wise additivity stems from the property of branching. I will spend the rest of this section, **Section 2.4**, proving the property of additivity by considering the assignment of transcripts to both cells and genes. The assignment of transcripts to cells is formalised as before, see **Sections 2.2** & **2.3**;

I will therefore begin by introducing the process of assigning transcripts to genes.

First, recall that M^g is the total number of transcripts of gene g. Let $L = \sum_{g=1}^{G} M^g$ be the total number of transcripts measured in a single experiment, where G is the number of genes. Consider the stochastic process of assigning L transcripts to each gene in each cell. Let V be the discrete probability distribution on the set of gene-cell pairs $\{1, \ldots, G \cdot N\}$, where $v_{i,g}$ is the probability of assigning a transcript to gene g in cell i.

Now consider the stochastic process of assigning L transcripts to the G genes profiled. Let W be the discrete probability distribution on the set of genes $\{1, \ldots, G\}$, where $p(W = g) = w_g$ is the probability of assigning a transcript to gene g. Let $w_g = M^g / L$, the measured proportion of transcripts assigned to gene g.

As illustrated in **Fig 2.11**, the assignment of transcripts can be arranged as a branching process, first to genes and then to cells. Therefore, the entropy of the overall assignment process, V is additively decomposable into the sum of the entropy of the assignment of transcripts to genes, W, and the gene-wise distribution of transcripts to cells, X^g:

$$H(V) = H(W) + \sum_{g=1}^{G} w_g \, H(X^g). \tag{2.31}$$

Note that $H(V)$ is independent of the branching structure – the assignment processes can be swapped, first assigning transcripts to cells and then to genes without affecting the total entropy of the assignment process (Shannon, 1948).

To restate **Eqn 2.31** in terms of $I(g)$ instead of $H(X^g)$, I must define distributions against which the relative entropy of V and W can be measured. For this, I will assume that transcripts are distributed with equal probability to each gene in each cell. With respect to V, this assumption corresponds to the discrete uniform distribution, $U_{\{genes, cells\}}$ defined on the set $\{1, \ldots, G \cdot N\}$ with equal probabilities $1/(G \cdot N)$. With respect to W, the assumption instead corresponds to the discrete uniform distribution U_{genes}, defined on the set $\{1, \ldots, G\}$ with probabilities $1/G$.

Restating **Eqn 2.31** with respect to the newly defined uniform distributions,

$$D(V||U_{\{genes, cells\}}) = D(W||U_{genes}) + \sum_{g=1}^{G} w^g \, I(g), \tag{2.32}$$

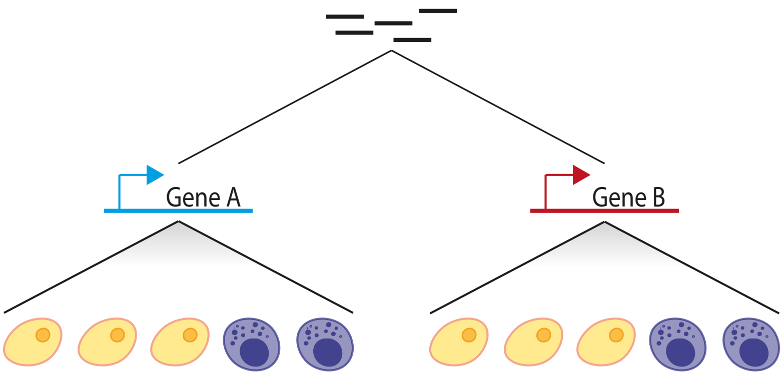


FIGURE 2.11: **Branching assignment of transcripts to genes and cells**. Diagrammatic representation of the process of assigning transcripts (black bars) first to specific genes, then to individual cells. The assignment process follows a branching pattern analogous to **Fig 2.10**. The relative entropy of this assignment process is additively decomposable into the relative entropy of the assignment of transcripts to genes, and the relative entropy of the assignment of transcripts to cells.

where $I(g)$ can itself be additively decomposed,

$$D(V||U_{\{genes,cells\}}) = D(W||U_{genes}) + \sum_{g=1}^{G} w^g\, H_S(g) + \sum_{g=1}^{G} w^g\, h_S(g). \tag{2.33}$$

Therefore, the many-gene measure of each type of heterogeneity is the sum of each gene's heterogeneity, weighted by the proportion of transcripts assigned to each gene. This means that higher expressing genes contribute more heavily to the total heterogeneity. Given that the heterogeneity of cells with respect to lowly expressed genes, especially transcription factors, can be critical to biological function (e.g. *Nanog*, see Smith et al. (2017)), I will instead assume transcripts to be equally likely to be assigned to each gene.

Based on this assumption that $w^g = \frac{1}{G}$ for $g = 1, \ldots, G$,

$$D(V||U) = \overbrace{D(W||U_{genes})}^{=0} + \sum_{g=1}^{G} w^g\, I(g) \tag{2.34}$$

$$= \frac{1}{G} \sum_{g=1}^{G} I(g), \tag{2.35}$$

$$= \frac{1}{G} \left[\sum_{g=1}^{G} H_S(g) + \sum_{g=1}^{G} h_S(g) \right]. \tag{2.36}$$

Moving forward, I drop the $\frac{1}{G}$ constant so that each many-gene measure of heterogeneity is simply the sum of each gene's heterogeneity. Accordingly, I define $I = \sum_g^G I(g)$, $H_S = \sum_g^G H_S(g)$ and $h_S = \sum_g^G h_S(g)$ as the many-gene analogues of each of $I(g)$, $H_S(g)$ and $h_S(g)$.

Importantly, the many-gene analogues retain the property of additive decomposition, i.e.

$$I = H_S + h_S. \tag{2.37}$$

Eqn 2.37 provides a framework for assessing the proportion of heterogeneity attributable to differential expression between a set of clusters S, with respect to all genes. The total amount of heterogeneity attributable to genome-wide differential expression is simply the sum of gene-wise inter-cluster heterogeneities. Between the single-gene and many-gene measures, differential expression can be intuitively quantified in terms of gene expression heterogeneity.

In the rest of this chapter, I will apply both the single-gene and many-gene measures to publicly available single-cell RNA-sequencing data sets. I will introduce a method for robustly estimating expression heterogeneity from single-cell RNA-sequencing data. I will validate that $I(g)$ is a biologically relevant measure of gene expression heterogeneity. I will demonstrate $H_S(g)$ as a practical measure of differential gene expression and establish the association between H_S and the true clustering of cells into cell types (where I assume that the established classification represents the true classification of cells into types for each data set).

2.5 Implementation & Validation

In this section, I apply the framework developed above to a range of single-cell RNA-sequencing data sets, detailed in **Table 2.1**. These data sets cover a wide range with respect to the number of cell types present in each population, and the cells of each data set have been classified by a diverse set of methods, both computational and experimental. In particular, the data sets from Svensson et al. (2017) and Tabula Muris Consortium et al. (2018) represent extremes with respect to number of classified types, being a technical control data set and a mouse cell atlas respectively. Note that from here on, I will refer to the Tabula Muris Consortium et al. (2018) data set as the Tabula Muris).

The remaining data sets are chosen as they have well-established cellular classifications derived from a variety of methods, i.e. the published

Data Set	Cells (N)	Genes (G)	Cell Types (C)	Classification
Svensson et al. (2017)	4000	4483	1	None
Tian et al. (2019)	902	14718	3	Genotypic
Zheng et al. (2017)	85423	11811	4	Phenotypic
Stumpf et al. (2020)	5504	8768	14	Clustering
Tabula Muris Consortium et al. (2018)	55656	16062	56	Clustering

TABLE 2.1: **Data sets for use in validation**. Table details the number of cells and number of genes included post normalisation. Clustering refers to computational unsupervised clustering with differential gene expression analysis.

classifications are likely to represent the true classification of cells into types. For example, the Tian et al. (2019) data set consists of a mixture of three cancerous cell lines, with each cell line acting as a proxy for a cell type. The cells of the Tian et al. (2019) population were classified genotypically, with the cells of each cell line having distinct genotypes.

The Zheng et al. (2017) data set represents an example of traditional phenotypic classification. The data set concatenates multiple sequencing runs, where each sequencing run consists of cells of a single cell type, with cells having been sorted into different types prior to sequencing. Each cell type was sorted based on surface protein expression, with all cells belonging to a distinct peripheral blood mononuclear cell (PBMC) type, namely one of B-cells, T-cells, monocytes (CD14+) and natural killer cells (CD56+). (Note that the T-cells were sequenced in six different runs, each isolating a distinct T-cell sub-type; I treat these cells as belonging to a single T-cell identity as in Zou et al. (2021)).

The Stumpf et al. (2020) data set consists of cells sampled from mouse bone marrow. The cells are largely derived from the hematopoietic stem cell lineage: cells of this developmental lineage are actively transitioning in type from a stem cell identity to one of several possible mature cell types. Thus, the cells of the Stumpf et al. (2020) data set do not strictly belong to discrete cell types; instead, several of the published clusters group cells that are actively transitioning from one type to another. As such, the Stumpf et al. (2020) data set involves both discrete differences in expression between cell types and substantial continuous variation in expression within each cluster (i.e. there is substantial intra-type heterogeneity with respect to gene expression). Stumpf et al. (2020) clustered the cells of the data set using the Louvain method and classified each cluster based on differential expression of known marker genes (Blondel et al., 2008; Stuart et al., 2019).

I will use these data sets to assess the relationship between $I(g)$ and cell type diversity, expecting $I(g)$ to positively correlate with the number of cell types in the data set. I will then confirm that a significant proportion of gene expression heterogeneity can be attributed to differential gene expression between the established clusters of each of the Tian et al. (2019), Zheng et al. (2017) and Stumpf et al. (2020) data sets.

Throughout this section, I am interested in the expression heterogeneity arising from differential gene expression between cell types. However, single-cell RNA-sequencing is a noisy process, with substantive technical error, causing biological uniform cells to be heterogeneous with respect to gene expression. Therefore, I first need to introduce a method for denoising the distribution of transcripts, X^g, minimising the effect of technical error on the measured gene expression heterogeneity.

2.5.1 Normalisation

Single-cell RNA-sequencing counts the number of individual mRNA molecules in single cells. However, the measurement of single molecules is both stochastic and inefficient: recall that single-cell sequencing experiments only capture 3-10% of the total mRNA molecules in a given cell (Papalexi and Satija, 2018). This low capture rate results in the sparse detection of lowly expressed genes (Risso et al., 2018; Lopez et al., 2018; Eraslan et al., 2019; Svensson, 2020; Lause et al., 2020; Sarkar and Stephens, 2021).

The effect of this sparsity can be observed in the Svensson et al. (2017) technical control data set. In Svensson et al. (2017), cell-equivalents have been generated from a mixed solution of endogenous human brain RNA and External RNA Control Consortium spike-ins. Without any biological function, the heterogeneity of the pseudo-population should be minimal, arising only from technical error.

Nevertheless, I find that $I(g)$ increases with respect to decreasing mean gene expression, see **Fig 2.12**. Specifically, below an average of 1 transcripts per cell (0 log mean expression), I find a log-linear relationship between mean gene expression and $I(g)$. With fewer transcripts than cells, the distribution of transcripts cannot be uniform, as clearly, when $M^g < N$, there is no way to distribute M^g transcripts among N cells uniformly. Instead, for $M^g < N$, the minimum value of $I(g) = \log(\frac{N}{M^g})$, where M^g cells each express a single transcript.

When calculating $I(g)$, it is natural to assume $x_i^g = p_i^g$, the measured proportion of transcripts of g assigned to cell i. Indeed, this represents the maximum likelihood estimate for X^g (Townes et al., 2019). However,

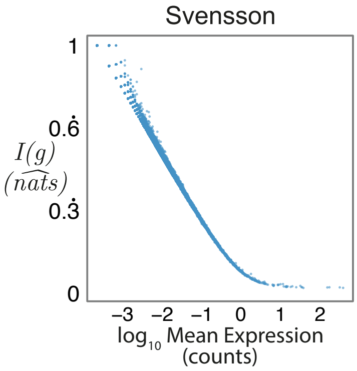


FIGURE 2.12: **Effect of sparsity on population heterogeneity**. Plot of population heterogeneity, $I(g)$, against $\log_{10}$ mean gene expression for each gene in the Svensson et al. (2017) data set. Values of $I(g)$ are normalised by data set specific theoretical maximum of $I(g)$, $\log N$, where N is the number of cells in the data set. Below $\sim 0 \log_{10}$ mean gene expression, i.e. a mean expression of 1 transcript per cell, $I(g)$ increases linearly with decreasing (log) mean gene expression.

maximum likelihood estimators performs poorly whenever there are fewer observations, M^g, than variables being estimated, N (James and Stein, 1992). Therefore, to account for the effect of sparsity, I adopt an alternative estimator for the transcript distribution X^g: the James-Stein-type shrinkage estimator (James and Stein, 1992).

Previously applied to microarray gene expression data in Hausser and Strimmer (2009), the James-Stein-type estimator shrinks the maximum likelihood estimate of the expression distribution of each gene towards the uniform distribution, U, thus reducing $I(g)$ (recall that $I(g)$ measures the divergence of the observed gene expression distribution from U). The strength of the shrinkage determined by a scale factor. Specifically, the James-Stein-type estimator strongly shrinks the distribution of those genes with fewer measured transcripts, and those genes with greater variance in p_i (where the observed transcripts are closer to being uniformly distributed).

The shrinkage strength for each gene, λ^g, is,

$$\lambda^g = \frac{\sum_{i=1}^{N} \widehat{Var}(p_i^g)}{\sum_{i=1}^{N}(\frac{1}{N} - p_i^g)^2} = \frac{1 - \sum_{i=1}^{N}(p_i^g)^2}{(M^g - 1)(\sum_{i=1}^{N}(\frac{1}{N} - p_i^g)^2)}, \tag{2.38}$$

where p_i^g is the maximum likelihood estimation of the frequency of gene g in cell i and $\frac{1}{N} \in U$ is the shrinkage probability. Note that $\lambda^g \in [0, 1]$.

The James-Stein-type distribution is a compromise between the maximum likelihood estimate and the uniform distribution:

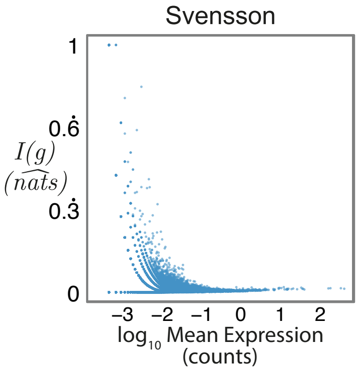


FIGURE 2.13: **James-Stein-type $I(g)$ against mean gene expression**. Plot of population heterogeneity, $I(g)$, against $\log_{10}$ mean gene expression for each gene in the Svensson et al. (2017) data set. $I(g)$ is calculated based on James-Stein-type shrinkage estimations of X^g. Using the James-Stein-type estimator abolishes the strict log-linear relationship between mean gene expression and $I(g)$, as was observed with respect to values of $I(g)$ calculated based on the maximum likelihood estimation of X^g.

$$x_i^g = \lambda^g \frac{1}{N} + (1 - \lambda^g) p_i^g. \tag{2.39}$$

At $\lambda^g = 0$, the estimator returns the maximum likelihood distribution, and at $\lambda^g = 1$, the estimator returns the uniform distribution.

Returning to the Svensson et al. (2017) data set, it is clear from **Fig 2.13** that calculating $I(g)$ based on the James-Stein-type estimation of X^g minimises the effect of sparsity without abolishing the potential for heterogeneity with respect to lowly-expressed genes. However, for very lowly-expressed genes, the James-Stein-type estimator remains insufficient with respect to minimising the effect of sparsity. Therefore, in the following sections, I will remove the most severe cases of sparsity in each data set analysed, namely, those genes with less than 100 transcripts in total across all cells.

2.5.2 Population Heterogeneity

There are various sources of heterogeneity in a cellular population: differential expression between cell types, biological functions orthogonal (unrelated) to the differential gene expression between cell types, biological noise and technical error (see **Section 1.1.2** for greater discussion of each source of gene expression heterogeneity). Through use of the James-Stein type estimator, I aim to limit the effects of the mainly stochastic biological noise and technical error. I therefore expect the remaining population heterogeneity to positively correlate with the number of cell types and the extent of any other biological function. $I(g)$ should capture both sources of heterogeneity through inter-cluster and intra-cluster heterogeneity, respectively.

To confirm $I(g)$ as a biologically relevant measure of gene expression heterogeneity, I measure the gene-wise $I(g)$ of data sets listed in **Table 2.1**. The the number of cell types, increases across the collection of data sets. Therefore, I expect the number of genes with substantial $I(g)$ to increase across the data sets as ordered in **Table 2.1**.

As shown in **Fig 2.14**, I find the expected correlation between gene-wise $I(g)$ and number of cell types. The relationship between number of cell types and $I(g)$ is made visually obvious by the differences between the Svensson et al. (2017) technical control data set, $C = 1$, and the Tabula Muris cell atlas data set, $C = 56$. These data sets demonstrate clear extremes in $I(g)$, reflecting their differing number of cell types (recall the value of C is the number of cell type clusters reported with each data set). The differences between the remaining data sets are more muted, but still reflective of a correlation between gene-wise $I(g)$ and number of cell types: the Zheng et al. (2017) data set ($C = 4$) has a higher average value of $I(g)$ with a mean of 0.24 nats compared to the mean of 0.20 nats of Tian et al. (2019) ($C = 3$).

Note that while the gene-wise measure of heterogeneity, $I(g)$, strongly correlates with the number of cell types, the association between the total gene expression heterogeneity, I, and number of cell types is weaker. This is due to the so-called 'curse of dimensionality' (Beyer et al., 1999). Even with adjustment by the James-Stein-type estimator, most genes will likely be associated with some non-zero level of $I(g)$ due to biological noise and technical error. Over thousands of genes, these small amounts of gene expression heterogeneity accumulate, resulting in the genome-wide measure of population heterogeneity, I, being dominated by biological and technical effects other than the differential expression between cell types. This results, for instance, in the total heterogeneity of Svensson et al. (2017) exceeding that of Tian et al. (2019), 4250 nats to 2930 nats. Though the total heterogeneity still depends in part on the number of cell types: the Stumpf et al. (2020) and Tabula Muris data sets have values of I of 7040 nats and 16500 nats, respectively.

To confirm that $I(g)$ additionally captures heterogeneity arising from biological functions unrelated to differential expression, I examine the Stumpf et al. (2020) data set in more detail. The cell types of the population sequenced in Stumpf et al. (2020) are actively undergoing differentiation and maturation.

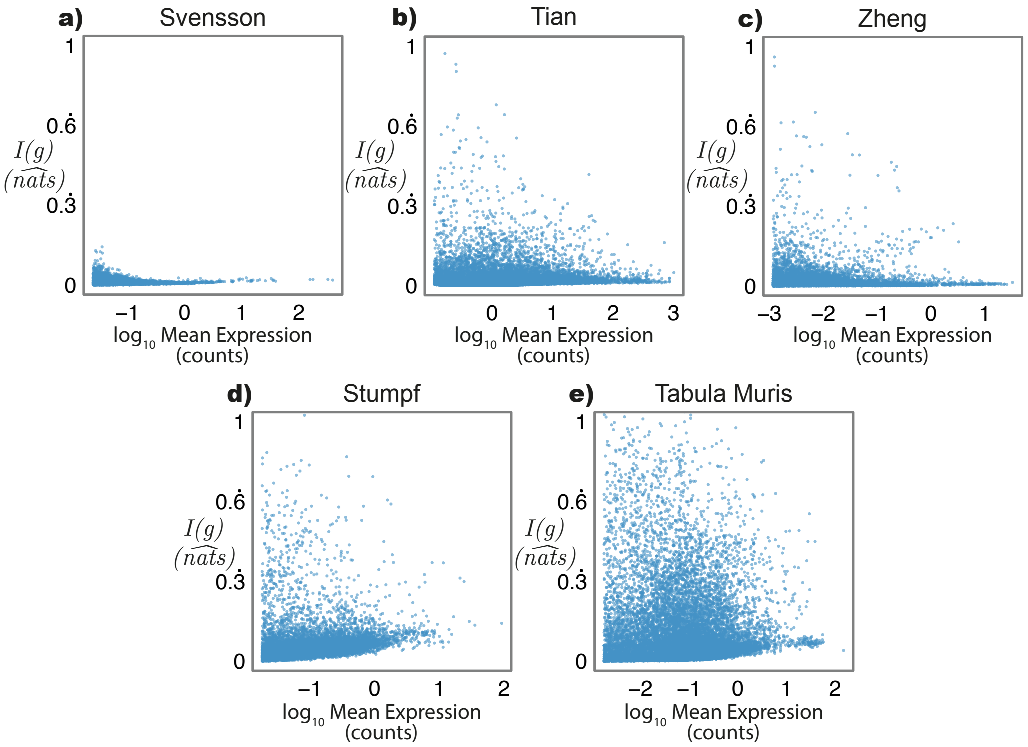


FIGURE 2.14: **Population heterogeneity of sequencing data sets**. Plot of population heterogeneity normalised by theoretical maximum, $I(g)/\log N$, against $\log_{10}$ mean gene expression for each gene in the **a)** Svensson et al. (2017) (number of cell types, $C = 1$), **b)** Tian et al. (2019) ($C = 3$), **c)** Zheng et al. (2017) ($C = 4$), **d)** Stumpf et al. (2020) ($C = 14$), and **e)** the Tabula Muris ($C = 56$) data sets (Tabula Muris Consortium et al., 2018). The number of genes cells are heterogeneous with respect to, as measured by $I(g)$, broadly increases with increasing number of cell types, C, in the population. See **Table 2.1** for detailed information on each data set.

Therefore, there is substantial gene expression heterogeneity arising from biological processes other than differential expression between cell types.

I confirm $I(g)$ captures this heterogeneity by visualisation, projecting the gene expression profiles of each cell in the Stumpf et al. (2020) data set down onto two dimensions by non-linear dimension reduction (McInnes et al., 2018). I restrict the dimension reduction to include only the top 500 genes by $I(g)$. As shown in **Fig 2.15**, these 500 genes are sufficient to capture the complex topology of the (Stumpf et al., 2020) data set. Thus, genes with high values of $I(g)$ are associated with gene expression heterogeneity arising from both the differential expression between cell types and the biological processes of differentiation and maturation

The value of population heterogeneity, $I(g)$, broadly corresponds to the number of cell types in a population. In the next section, I will test the relationship between information-theoretic heterogeneity and cell type

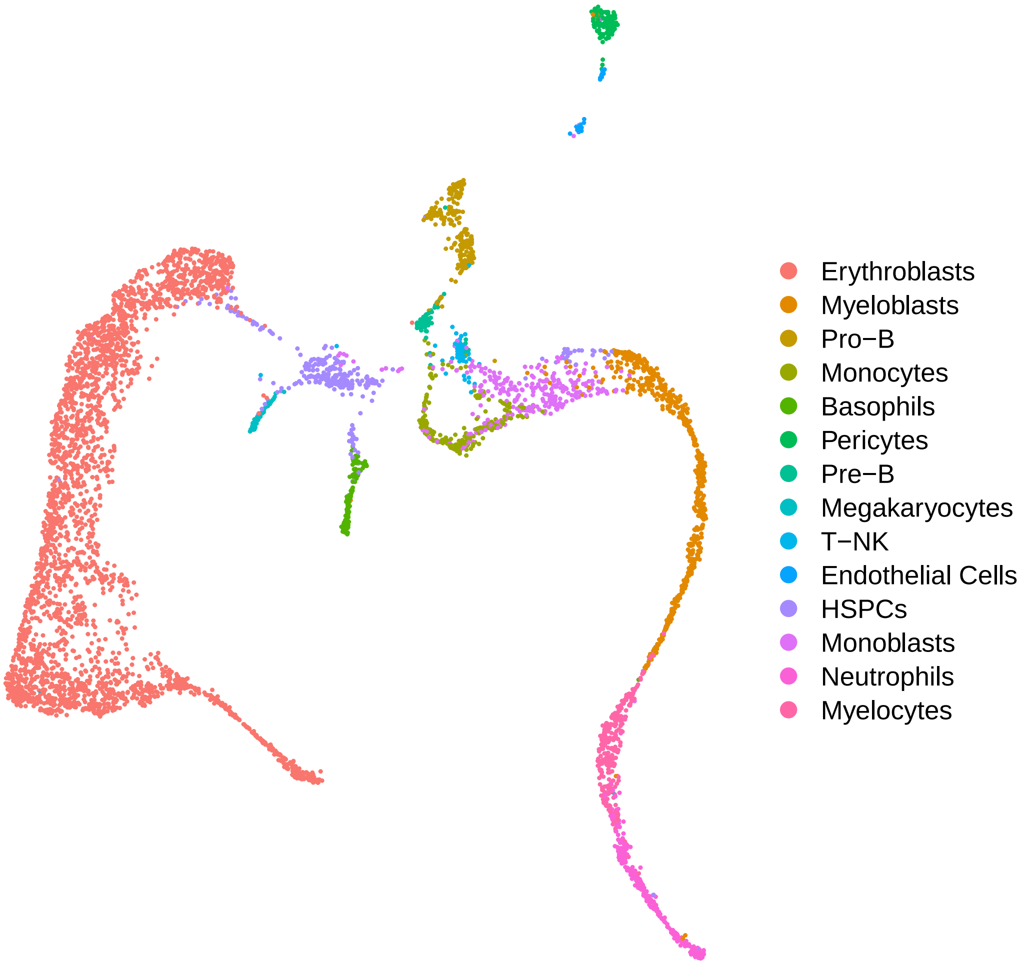


FIGURE 2.15: **UMAP of mouse bone marrow cells.** Non-linear dimension reduction of the expression of the top 500 genes by $I(g)$, with cells coloured by their classification in Stumpf et al. (2020). Top 500 genes by $I(g)$ capture both the differential expression between cell types, leading to clear separation of each identity, and the continuous variation associated with cellular differentiation and maturation.

explicitly, calculating inter-cluster heterogeneity with respect to the established classifications for the Tian et al. (2019), Zheng et al. (2017) and Stumpf et al. (2020) data sets. Each of these classifications has been robustly established, through genotype, surface protein expression or differential gene expression analysis: these classifications represent a suitable approximation of the true mixture of cell types in a population (up to the resolution afforded by each classification method, see discussion of Zheng et al. (2017) data set in **Section 3.1.4**).

2.5.3 Differential Gene Expression

Inter-cluster heterogeneity quantifies the gene expression heterogeneity attributable to differential gene expression between clusters. In the context of single-cell RNA-sequencing data, cell types are realised as differentially expressing discrete sets of cells. Therefore, I expect the correct clustering of cells into types to be associated with substantial inter-cluster heterogeneity, both in terms of number of genes identified as differentially expressed by $H_S(g)$ and in terms of overall H_S.

I confirm the association between inter-cluster heterogeneity and the true clustering of cells by comparing the inter-cluster heterogeneity of established cellular classifications to randomly constructed clusterings of cells. I construct these random clusterings through permuting the cell-wise classifications of each data set, keeping the number and size of clusters the same between the established and randomised clusterings.

For each data set where the classification has been robustly established, i.e. the Tian et al. (2019), Zheng et al. (2017) and Stumpf et al. (2020) data sets, I compare the associated values of $H_S(g)$ and H_S between the established and randomised clusterings. These comparisons are formalised as exact, one-sided significance tests, generating an exact p-value for each gene and genome-wide, based on 10^4 randomisations of each data set (Fisher, 2017).

For each gene, I test whether the amount of heterogeneity attributable to differential expression between cell types exceeds that expected from a random clustering of the same structure, i.e. a clustering with the same number of clusters, C, and with the same proportion of cells assigned to each cluster. Note that as many statistical tests are being carried out simultaneously, the false discovery rate of the gene-wise comparisons must be controlled for with respect to each data set (Benjamini and Hochberg, 1995). To maintain the power of the analysis (and to avoid the computational cost of randomising all $> 10^4$ genes of the large Zheng et al. (2017) data set), I restrict testing to the top 500 genes by $I(g)$ for each data set. By doing so, I include only those genes where substantial values of $H_S(g)$ are possible (by additive decomposition, the value of $H_S(g)$ cannot exceed the value of $I(g)$).

For each data set, the majority of genes are significantly differentially expressed as measured by $H_S(g)$. Of the 500 genes tested in each data set, 485 in Tian et al. (2019), 472 in Zheng et al. (2017) and 498 in Stumpf et al. (2020) significantly exceed the values of $H_S(g)$ calculated with respect to the the randomised clusterings (one-sided exact test, $\alpha = 0.05$, false discovery rate

Data Set	Known	Mean	SD	Max
Tian et al. (2019)	199	6.04	0.61	10.1
Zheng et al. (2017)	161	0.796	0.16	1.77
Stumpf et al. (2020)	608	29.3	3.4	44.6

TABLE 2.2: H_S **of established and randomised cellular clusterings**. H_S of the established classification of cells and the mean, standard deviation and maximum H_S of 10^4 randomisations. All numeric values are in units of nats.

correction for 500 trials; see **Fig 2.16**). Moreover, for the majority of genes tested in each data set, the value of $H_S(g)$ with respect the established classification exceeds that of all 10^4 randomisations, resulting in an exact p-value of 0 (p-value $= 0$ for 459, 415 and 487 out of 500 genes tested in the Tian et al. (2019), Zheng et al. (2017), and Stumpf et al. (2020) data sets, respectively). See **Fig 2.17** for illustration of Zheng et al. (2017) data set.

Furthermore, the overall inter-cluster heterogeneity, H_S, with respect to each established classification significantly and substantially exceeds that of all randomised clusterings, see **Table 2.2**. Thus, the true clustering of cells into types is associated with significantly greater H_S than expected for a given cluster structure. This association holds across the range of classification methods used: genotypic, surface protein expression and unsupervised clustering. The true clustering of cells into types is robustly associated with substantial H_S, and by additive decomposition, low h_S.

2.6 Discussion

This chapter developed a formal framework for the measurement of heterogeneity with respect to gene expression, explicitly quantifying the contribution of differential gene expression to observed heterogeneity. The framework is based on the language of information theory, quantifying gene expression heterogeneity as the amount of information required to encode the observed distribution of gene expression on the set of cells/clusters. Overall, this framework represents a novel approach to the quantification of heterogeneity in the context of single-cell RNA-sequencing data (Brennecke et al., 2013; Grün et al., 2014; Townes et al., 2019; Hafemeister and Satija, 2019; Breda et al., 2021; Lause et al., 2020).

Throughout this chapter, I have focused on cellular classification via differential gene expression analysis, assuming that cells of different types form differentially expressing clusters. I have assessed the success of a given clustering in attributing gene expression heterogeneity to differential gene

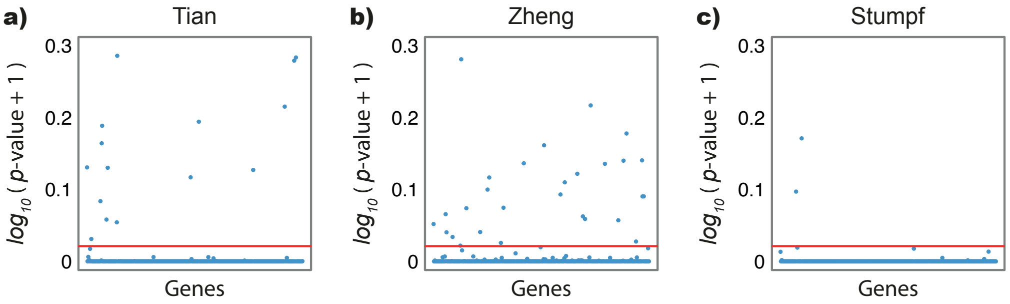


FIGURE 2.16: **Significance of $H_S(g)$ with respect to randomisations.** Plot of $log_{10}(p\text{-values} + 1)$, where the p-values are exact and controlled for false discovery rate, for each gene in the **a)** Tian et al. (2019), **b)** Zheng et al. (2017), and **c)** Stumpf et al. (2020) data sets. A pseudo-count is added to exact p-values as the majority of p-values $= 0$. The horizontal red line at $log_{10}(\alpha + 1)$ represents the chosen significance threshold of $\alpha = 0.05$; genes below the threshold are significantly differentially expressed as measured by $H_S(g)$.

expression. However, the developed framework is also applicable to the more traditional approach to cellular classification, phenotypic classification (discussed in **Section 1.1**). Specifically, through the quantity $h_S(g)$, the principles of phenotypic classification are generalised to the case of single-cell RNA-sequencing data.

In phenotypic classification, cells are classified based on the expression of marker genes. In classifying cells based on marker genes, cells of different types are assumed to be (at least approximately) interchangeable with respect to marker gene expression. Intra-cluster heterogeneity measures how well a given clustering holds to this assumption, quantifying the divergence from the assumption that cells of the same type should be interchangeable with respect to measured gene expression.

Importantly, $h_S(g)$ generalises the fundamental assumption of phenotypic classification: $h_S(g)$ quantifies the divergence from intra-cluster interchangeability with respect to all clusters, as opposed to only a single cluster (note that each additive component of $h_S(g)$, see **Eqn 2.8**, quantifies the divergence from interchangeability within each individual cluster). The assumption is generalised further through use of h_S, which quantifies the divergence from intra-cluster interchangeability with respect to all genes, as opposed to a single marker gene.

H_S, as established in **Sections 2.3.1** & **2.4**, represents a analogous generalisation of differential gene expression, measuring differential expression with respect to all clusters and all genes. Thus, the developed framework formalises the correspondence between phenotypic classification and differential gene expression analysis through the property of additive

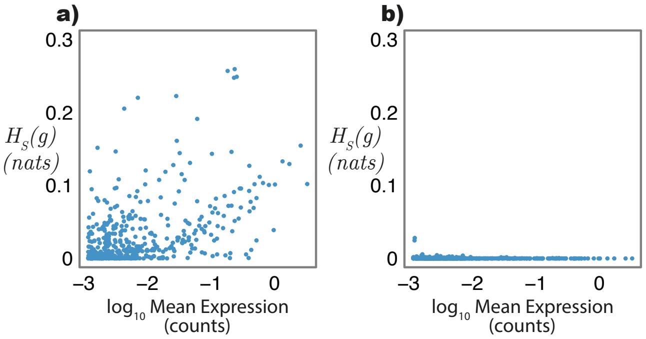


FIGURE 2.17: **Normalised inter-cluster heterogeneity of Zheng et al. (2017) data**. Plot of $H_S(g)$ for each gene in the Zheng et al. (2017) data set based on **a)** the established surface protein classification or **b)** the best performing random clustering with respect to each gene.

decomposition: as $I = H_S + h_S$, the greater the differential expression between a set of clusters, the less divergent the clusters from the assumption that cells of the same type are interchangeable with respect to measured gene expression. Thus, in establishing an empirical association between H_S and the classification of cells into cell types, I have demonstrated that the fundamental assumption of phenotypic classification – that cells of the same type are approximately interchangeable with respect to measured gene expression – extends to single-cell expression data.

In the next chapter, I will build on the empirical association established in **Section 2.5.3**. I will assume that for a given data set, the clustering most likely to represent the true clustering of cells into types maximises H_S and, by additive decomposition, minimises h_S. This assumption stems from both the phenotypic and differential expression approaches to classification: the clustering that maximises H_S will be both minimally divergent from the assumption that the cells of each cluster should be interchangeable (with respect to gene expression) and maximally differentially expressed (as measured by H_S). Formally, in the next chapter, I will introduce a method for the numerical optimisation of the following statement,

$$\underset{S}{arg\,max}\ H_S, \tag{2.40}$$

where $\underset{S}{arg\,max}$ refers to the clustering S that maximises H_S.

The numerical optimisation of **Eqn 2.40** represents a novel unsupervised clustering method. Notably, the novel clustering algorithm departs from the traditional approach to unsupervised clustering: rather than implicitly assuming notions from dynamical systems theory (as discussed in **Section 1.2.2**), the novel method represents a mathematical formalisation of the empirical principles of phenotypic classification.

Chapter 3

Numerical Optimisation of Information-Theoretic Clustering

Introduction

Chapter 2 formalised an information-theoretic framework for quantifying the proportion of heterogeneity attributable to genome-wide differential expression. I demonstrated that established cellular classifications, where the clusters likely represent the true set of cell types in the population, attribute significantly more gene expression heterogeneity to differential gene expression between clusters (i.e. greater inter-cluster heterogeneity, H_S) and significantly less heterogeneity to differences in gene expression within clusters (i.e. less intra-type heterogeneity, h_S).

Building on this association, in this chapter, I will assume that the *true* clustering of cells into types maximises H_S. This assumption builds on both the phenotypic and differential expression approaches to classification: the clustering that maximises H_S is both minimally divergent from intra-cluster interchangeability (with respect to gene expression) and maximally differentially expressing.

The assumption can be formalised as the partial optimisation statement,

$$\underset{S}{arg\,max}\ H_S, \tag{3.1}$$

where the true clustering of cells into types, S, maximises H_S.

Note that **Eqn 3.1** is only a *partial* optimisation statement – optimising **Eqn 3.1** will only identify the optimal clustering with respect to a specified number of clusters C. There will therefore be a 'true' clustering for each possible number of clusters. Determining the correct number of clusters is not a trivial task, as it requires that the number of cell types in a population is known. I will present one approach to inferring the correct choice of C in **Chapter 4**. For the majority of this chapter, specifically **Section 3.1**, I will assume that the true number of clusters is known.

This chapter will develop a computational approach to the optimisation of inter-cluster heterogeneity, identifying the clustering of cells that maximises H_S. This computational optimisation corresponds to a novel supervised clustering method for single-cell expression data, clustering cells based on the univariate expression distributions of individual genes. I will, therefore, first discuss the general problem of unsupervised clustering before introducing the specifics of the approach taken in this chapter.

Recall from **Chapter 1** that unsupervised clustering methods group cells into clusters without reference to any external knowledge or information; supervised methods, by contrast, group cells based on a training set of pre-classified cells. Unsupervised clustering methods consist of two parts: an *objective function* to be optimised and an algorithm for the optimisation of the objective function (Jain, 2010).

Objective functions of unsupervised clustering methods are often difficult to optimise (Jain, 2010; Von Luxburg et al., 2012; Kiselev et al., 2019). Many clustering objective functions are non-linear, so they are intrinsically more difficult to solve for than linear functions (Bradley et al., 1977). Moreover, for real-world large data sets (e.g. single-cell data sets containing thousands to millions of cells), calculating the objective functions may be highly demanding with respect to computational time and memory (Kiselev et al., 2019; Svensson et al., 2020).

Furthermore, along with the single globally optimal solution, many objective functions have multiple additional locally optimal solutions (Bradley et al., 1977). In the context of single-cell clustering, a local optimum occurs when a given clustering scores better by the objective function than any neighbouring clustering, where neighbouring clusterings differ by the cluster assignment of a single cell (this definition of neighbouring assumes that the clustering is discrete; see **Section 3.1.1** for discussion of non-discrete clusterings). These local optima score less well with respect to the objective function than the

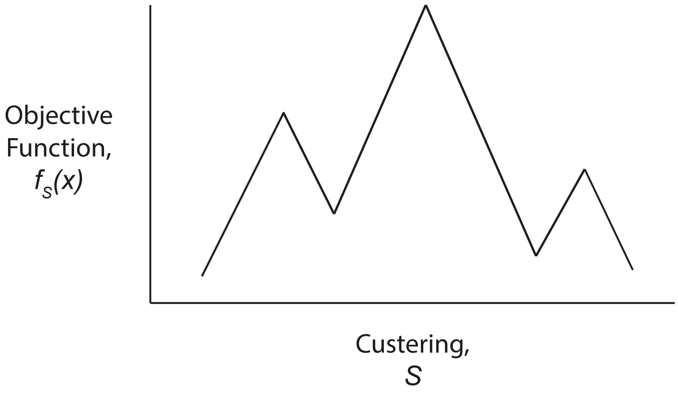


FIGURE 3.1: **Landscape of an objective function**. Each potential clustering S of the data x is associated with some value of the non-linear objective function $f_S(x)$. The objective function can be visualised as a landscape over the range of possible clusterings. The objective function may have only a single peak in the landscape, representing the sole optimal solution. Alternatively, there may be multiple optima with respect to S, with only one global optimum (centre peak) and some number of local optima (two side peaks). Local optimisation algorithms typically 'move' along the landscape, so they can get 'trapped' at local optima (Bradley et al., 1977).

global optimum but score better than any neighbouring clustering (see **Fig 3.1** for illustration).

Given the difficulties involved with many clustering optimisation problems, different optimisation algorithms can be developed for a single objective function; for example, the Louvain and Leiden methods are both algorithms for modularity optimisation used in the context of single-cell clustering (Blondel et al., 2008; Traag et al., 2019). Different optimisation algorithms have different advantages and trade-offs. For instance, a major division in how algorithms operate is whether they search locally or globally (Bradley et al., 1977). Global optimisation algorithms attempt to find the single best solution to the objective function. In contrast, local optimisation algorithms only attempt to find a local optimum, usually in exchange for a substantial reduction in the search time required.

Concerning inter-cluster heterogeneity, several algorithms have already been developed for optimising (relative) entropy as an objective function for unsupervised clustering, see Roberts et al. (2000, 2001), and Li et al. (2004). However, the objective functions used by these methods do not exactly correspond to inter-cluster heterogeneity, and moreover, these algorithms were developed for substantially smaller data sets; the large size of single-cell sequencing data sets requires a more scalable approach, able to cluster thousands of cells in a reasonable length of time.

To that end, I will begin this chapter by developing a novel approach to optimising relative entropy in the context of unsupervised clustering. I will adopt a fuzzy notion of clustering, allowing the adaption of an existing, efficient local optimisation algorithm, the *L-BFGS-B* (Limited-memory Broyden–Fletcher–Goldfarb–Shanno Bound-constrained) algorithm (Zhu et al., 1997). The *L-BFGS-B* algorithm is a very well established non-linear numerical optimisation algorithm, developed for dense, large non-linear optimisation problems, such as maximising H_S (Zhu et al., 1997).

Later in the chapter, in **Section 3.1.4**, I will validate the novel unsupervised clustering method, which I call scEC (single-cell Entropic Clustering), on a range of publicly available single-cell RNA-sequencing data sets with known cellular classifications (see **Table 2.1**). I will compare the scEC clustering of each data set to the respective established cellular classification and to a clustering produced by a state-of-the-art single-cell clustering method. Note that the code for scEC can be found in **Appendix C** and is available online at `https://github.com/mjcasy/scEC`.

I will conclude the chapter by extending the developed unsupervised clustering method to the semi-supervised setting. Semi-supervised methods leverage some external information to enhance clustering. In the context of single-cell clustering, semi-supervised methods allow cells to be clustered directly into established types, directed by the known cellular classification of a reference data set.

3.1 Optimising Inter-cluster Heterogeneity

To realise **Eqn 3.1** as a practical clustering method, an efficient algorithm needs to be adopted for the purpose of maximising H_S. Brute force optimisation (iterating all possible clusterings and computing the objective function for each) is not feasible as there is an exponential number of ways (k^n) to partition a set of n objects into k subsets.

Instead, I will adopt the limited-memory, box constrained BFGS (*L-BFGS-B*) optimisation algorithm from the `Python3` (v3.8.2) package `SciPy` (v1.5.3) (Byrd et al., 1995; Zhu et al., 1997; Van Rossum and Drake, 2009; Virtanen et al., 2020). The *L-BFGS-B* algorithm is a non-linear local optimisation method developed for solving large, dense problems, such as that of clustering single-cell RNA-sequencing data (Zhu et al., 1997). Concerning the optimisation of a given vector of variables, w_{ik}, with respect to some objective function, the algorithm starts from a specified initial vector and searches the space of possible values of w_{ik}. The search is directed by the gradient of the objective function and the inverse of a limited-memory approximation to the Hessian of the objective function (Byrd et al., 1995). Note that the gradient encodes the partial derivative of the objective function with respect to each variable, w_{ik}, and the Hessian encodes the second partial derivative of the objective function with respect to each pair of variables, w_{ik} & w_{rq}.

L-BFGS-B is ideally suited to large optimisation problems due to the algorithm's efficient approximation of the Hessian. The Hessian encodes additional information on the shape of the optimisation landscape beyond the 'slope' along each dimension encoded by the gradient. However, the Hessian is large – for an objective function of n variables, the gradient has n elements, and the Hessian has n^2 elements – so the matrix can be slow to compute for large, dense data sets. The *L-BFGS-B* algorithm does not compute the Hessian directly; instead, it derives an efficient, limited-memory representation of the Hessian, based on the gradient, that scales linearly (instead of quadratically) in memory with the number of variables n (Byrd et al., 1995). Thus, the *L-BFGS-B* algorithm gains much of the advantage of the Hessian without incurring the computational cost.

In the case of clustering, the variables being optimised are the cluster identities of each cell. The *L-BFGS-B* algorithm utilises the derivative H_S with respect to S. Carrying out such differentiation requires that H_S be continuous with respect to S; however, the clustering S is discrete, with each cell assigned solely to a single cluster. Therefore, I will adopt a fuzzy notion of clustering, allowing each cell to be admitted to multiple clusters with differing degrees of membership (Peters et al., 2013; Tasic et al., 2016). By adopting a fuzzy notion of clustering, the objective function, H_S, is made continuous with respect to the cluster membership of each cell, enabling the differentiation of H_S with respect to S (or more strictly, with respect to the fuzzy cluster memberships of each cell as encoded in S).

In the following sections (**Sections 3.1.1** & **3.1.2**), I will establish the necessary mathematical machinery for implementing the *L-BFGS-B* optimisation algorithm. I will begin by formally introducing a notion of fuzzy clustering, with the corresponding fuzzy versions of inter and intra-cluster heterogeneity. I will then demonstrate that the property of additive decomposition applies in the fuzzy setting, confirming that maximising H_S minimises h_S with respect to the fuzzy clustering S. I will then derive the gradient of H_S with respect to S by differentiating H_S with respect to the cluster memberships of each cell.

Finally, based on the developed mathematics, I will adapt the *L-BFGS-B* algorithm for the maximisation of H_S.

3.1.1 Fuzzy Clustering

I adopt a fuzzy conception of clustering in which cells are assigned to C (possibly) overlapping, fuzzy clusters, $S_1, \ldots, S_C$. Each cell i has C corresponding membership functions $\mu_k(i) \colon \{1, \ldots, N\} \to [0,1]$ for $k = 1, 2, \ldots, C$, where $\mu_k(i) = \mu_{ik}$ is the membership of cell i to cluster S_k for $1 \leq i \leq N$ and $1 \leq k \leq C$. I assume that the membership functions are normalised,

$$\sum_{k=1}^{C} \mu_{ik} = 1 \text{ with } \mu_{ik} \geq 0, \tag{3.2}$$

for each $i = 1, 2, \ldots, N$, guaranteeing that every cell belongs to at least one cluster and possibly partially to several clusters. Based on **Eqn 3.2**, the normalised value μ_{ik} can be interpreted as the probability of the cell i being assigned to the (fuzzy) cluster S_k.

The information-theoretic framework developed in **Chapter 2** must be adapted to the fuzzy setting. Population heterogeneity, $I(g)$, is independent of the chosen clustering, so it remains unaffected by the adoption of fuzzy clusters, i.e. the total gene expression heterogeneity is unaffected by the choice of discrete or fuzzy cluster memberships. For the calculations of inter-cluster and intra-cluster heterogeneities, I extend the discrete random variables Y^g and Z_k^g to the fuzzy setting, where Y^g now measures the expression distribution of the gene g across the C fuzzy clusters, and Z_k^g measures the expression distribution of the gene g within fuzzy cluster S_k, as follows.

I define the discrete random variable Y^g on the set of clusters $S_k \in \{S_1, \ldots, S_C\}$ with probabilities y_k^g given by,

$$y_k^g = \sum_{i=1}^{N} \mu_{ik}\, x_i^g. \tag{3.3}$$

I also define, for each $S_k \in S$, a discrete random variable Z_k^g on the set of cells, $i = 1, \ldots, N$, with probabilities z_{ik}^g given by,

$$z_{ik}^{g} = \mu_{ik} \frac{x_i^g}{y_k^g}. \tag{3.4}$$

Note that because μ_{ik} defines the membership of every cell with respect to the cluster S_k (with $\mu_{ik} = 0$ implying no membership), I define the fuzzy version of Z_k^g on the set of all cells; whereas, the non-fuzzy Z_k^g was defined only on the subset of cells included in each cluster.

Finally, I also extend N_k, the number of cells in the kth cluster, to the fuzzy setting, as

$$N_k = \sum_{i=1}^{N} \mu_{ik}. \tag{3.5}$$

Based on these extensions, $H_S(g)$ and $h_S(g)$ can be redefined for fuzzy clusterings as

$$H_S(g) = \sum_{k=1}^{C} y_k^g \log\left(\frac{y_k^g}{N_k/N}\right), \tag{3.6}$$

and

$$h_S(g) = \sum_{k=1}^{C} y_k^g \sum_{i=1}^{N} \mu_{ik} \frac{x_i^g}{y_k^g} \log\left(\frac{x_i^g/y_k^g}{1/N_k}\right). \tag{3.7}$$

Note that the form of $H_S(g)$ is unaffected by the extension to fuzzy clustering (see **Eqn 2.5** for the non-fuzzy version). However, the fuzzy version of $h_S(g)$ involves an additional μ_{ik} term (see **Eqn 2.8** for the non-fuzzy version). This term is included as the contribution of each cell i to $h_S(g)$ must be weighted by the probability, μ_{ik}, that the cell is assigned to the cluster S_k. This weighting means our measure of relative entropy, $h_S(g)$, corresponds not to Shannon's information entropy but rather to the generalisation of entropy to fuzzy sets introduced by Zadeh (1968). See the box *"Fuzzy Entropy"* for a brief explanation of fuzzy entropy.

With these extensions, the proof for additive decomposition then follows as before, see **Eqn 2.12** through **Eqn 2.30**, so that the additive decomposition of heterogeneity on a fuzzy clustering is

Fuzzy Entropy

Zadeh (1968) extended Shannon's entropy to fuzzy subsets. For example, let X be a discrete random variable on the set $\{1, \ldots, N\}$, with probabilities $p(X = i) = x_i$. The fuzzy subset k of the set $\{1, \ldots, N\}$ is defined by weighting the inclusion of each member i of the set by a membership μ_{ik}. The entropy of X with respect to the set $\{1, \ldots, N\}$ is defined in **Eqn 2.1**. The entropy of X with respect to the fuzzy subset k is,

$$H_k = -\sum_{i=1}^{N} \mu_{ik} x_i \log x_i,$$

where each element of the summation is weighted by its membership of the subset. Note that when $\sum_{i=k}^{C} = \mu_{ik}$, the memberships of the fuzzy subset k can be interpreted as the probability that each member i is included in the fuzzy set.

$$I(g) = \sum_{k=1}^{C} y_k^g \log\left(\frac{y_k^g}{N_k/N}\right) + \sum_{k=1}^{C} y_k^g \sum_{i=1}^{N} \mu_{ik} \frac{x_i^g}{y_k^g} \log\left(\frac{x_i^g / y_k^g}{1/N_k}\right) \tag{3.8}$$

$$= H_S(g) + h_S(g). \tag{3.9}$$

As information-theoretic heterogeneity remains additively decomposable with respect to the fuzzy clustering S, only one of H_S or h_S needs to be optimised and so only one of H_S or h_S needs to differentiated. In the following section, I will differentiate H_S with respect to the cluster memberships of each cell, μ_{ik}.

3.1.2 Differentiation of Inter-cluster Heterogeneity

In this section, I find the gradient of H_S with respect to μ_{ik} for use in the *L-BFGS-B* algorithm. Recall that the gradient of a function encodes the partial derivative of the function with respect to each element of a vector (or matrix) of variables. Concerning H_S, I aim to find the partial derivative of H_S with respect to each element μ_{ik}, the membership of the cell i with respect to fuzzy cluster S_k.

(Note that I will not be interpreting the gradient of H_S in terms of the biology of gene expression or cell type. While the derivative may have some biological interpretation, within the scope of this thesis, the gradient of H_S has a purely technical role in the optimisation of H_S with respect to S.)

I begin by repeating the definition of H_S in full,

$$H_S = \sum_{g=1}^{G} \sum_{k=1}^{C} y_k^g \log \left(\frac{y_k^g}{N_k/N} \right). \tag{3.10}$$

(Recall that y_k^g and N_k depend on μ_{ik}, as per **Eqn 3.3 & 3.5**.)

I now differentiate H_S with respect to the $N \cdot C$ membership functions μ_{rq}, where $1 \leq r \leq N, 1 \leq q \leq C$.

From **Eqn 3.3** and **Eqn 3.5**,

$$\frac{\partial y_k^g}{\partial \mu_{rq}} = \begin{cases} x_r^g & k = q \\ 0 & k \neq q \end{cases} \tag{3.11}$$

$$\frac{\partial N_k}{\partial \mu_{rq}} = \begin{cases} 1 & k = q \\ 0 & k \neq q \end{cases} \tag{3.12}$$

Using the product, chain and quotient rules, $H_S(g)$ can be differentiated with respect to the cluster memberships of each cell μ_{rq}:

$$\begin{aligned}
\frac{\partial H_S(g)}{\partial \mu_{rq}} &= \frac{\partial}{\partial \mu_{rq}} \left(\sum_{k=1}^{C} y_k^g \log \left(\frac{y_k^g}{N_k/N} \right) \right) && (3.13)\\
&= \sum_{k=1}^{C} \frac{\partial}{\partial \mu_{rq}} \left(y_k^g\right) \log \left(\frac{y_k^g}{N_k/N} \right) + \sum_{k=1}^{C} y_k^g \frac{\partial}{\partial \mu_{rq}} \left(\log \left(\frac{y_k^g}{N_k/N} \right) \right) && (3.14)\\
&= x_r^g \log \left(\frac{y_q^g}{N_q/N} \right) + \sum_{k=1}^{C} y_k^g \frac{N_k/N}{y_k^g} \frac{\partial}{\partial \mu_{rq}} \left(\frac{y_k^g}{N_k/N} \right) && (3.15)\\
&= x_r^g \log \left(\frac{y_q^g}{N_q/N} \right) + \sum_{k=1}^{C} \frac{N_k}{N} N \frac{\partial}{\partial \mu_{rq}} \left(\frac{y_k^g}{N_k} \right) && (3.16)\\
&= x_r^g \log \left(\frac{y_q^g}{N_q/N} \right) + \sum_{k=1}^{C} N_k \frac{1}{N_k^2} \left(\frac{\partial}{\partial \mu_{rq}} \left(y_k^g\right) N_k - y_k^g \frac{\partial}{\partial \mu_{rq}} \left(N_k\right) \right) && (3.17)\\
&= x_r^g \log \left(\frac{y_q^g}{N_q/N} \right) + \frac{1}{N_q} \left(x_r^g N_q - y_q^g \right) && (3.18)\\
&= x_r^g \left(\log \left(\frac{y_q^g}{N_q} \right) + \log(N) + 1 \right) - \frac{y_q^g}{N_q}. && (3.19)
\end{aligned}$$

Eqn 3.19 is the gradient of $H_S(g)$ with respect to cluster memberships, where each element of $\frac{\partial H_S(g)}{\partial \mu_{rq}}$ encodes the partial derivative of $H_S(g)$ with respect to the membership of the cell r to fuzzy cluster S_q.

The gradient of the objective function, H_S, is then,

$$\frac{\partial H_S}{\partial \mu_{rq}} = \sum_{g=1}^{G} \frac{\partial H_S(g)}{\partial \mu_{rq}}. \tag{3.20}$$

Recall that the memberships of each cluster are constrained, so that $\mu_{rq} \geq 0$ and that $\sum_{q=1}^{C} \mu_{rq} = 1$. A convenient way to incorporate these constraints is to introduce the unconstrained variables w_{ik} related to the cluster memberships μ_{ik} through the softmax function,

$$\mu_{ik} = \frac{e^{w_{ik}}}{\sum_{l=1}^{C} e^{w_{il}}}. \tag{3.21}$$

Regardless of the values of the variables w_{ik}, the softmax function guarantees that $\mu_{ik} \geq 0$ and that $\sum_{k=1}^{C} \mu_{ik} = 1$.

Making the objective function H_S a function of w_{ik}, the gradient of H_S can be found with respect to w_{ik} by the chain rule,

$$\frac{\partial H_S}{\partial w_{ik}} = \sum_{r=1}^{n} \sum_{q=1}^{C} \frac{\partial H_S}{\partial \mu_{rq}} \frac{\partial \mu_{rq}}{\partial w_{ik}}. \tag{3.22}$$

To determine $\frac{\partial \mu_{rq}}{\partial w_{ik}}$, I write $M_i = \sum_{l=1}^{C} e^{w_{il}}$ so **Eqn 3.21** becomes $\mu_{ik} = e^{w_{ik}}/M_i$. Then, by the quotient rule,

$$\frac{\partial \mu_{rq}}{\partial w_{ik}} = \frac{\partial}{\partial w_{ik}} \left(\frac{e^{w_{rq}}}{M_r} \right) = \begin{cases} \frac{e^{w_{ik}} M_i - e^{2w_{ik}}}{M_i^2} & \text{if } r = i \text{ and } q = k, \\ \frac{-e^{w_{iq}} e^{w_{ik}}}{M_i^2} & \text{if } r = i \text{ and } q \neq k, \\ 0 & \text{if } r \neq i. \end{cases} \tag{3.23}$$

Having derived suitably constrained objective and gradient functions, I will now implement the *L-BFGS-B* algorithm in optimising H_S.

3.1.3 Implementation of the *L-BFGS-B* Algorithm

Starting from an initial vector, w_{ik}, the *L-BFGS-B* algorithm carries out a search for a local optimum in H_S (Byrd et al., 1995; Zhu et al., 1997). The search is carried out via step-wise updates of w_{ik}. At each step of the search, the algorithm determines a new direction for the search, in part, by computing the gradient of H_S with respect to w_{ik} (given by **Eqn 3.22**). The search is bounded, with the algorithm only assessing values of w_{ik} between some range, $lb \leq w_{ik} \leq ub$, where the specific choice of upper (ub) and lower bounds (lb) is arbitrary. Importantly, this optimisation is for a fixed number of clusters, where the number of clusters is set via the size of the w_{ik} matrix.

The result of the *L-BFGS-B* algorithm algorithm is a vector encoding the variables w_{ik} for $i = 1, \ldots, N$ and $k = 1, \ldots, C$, which are converted to cluster memberships μ_{ik} through the softmax function, see **Eqn 3.21**. These memberships encode a fuzzy clustering corresponding to a (local or global) optimum in H_S, within the bounds of the search.

Concerning the initial vector, the cluster-wise memberships must be varied for each cell. If all the membership values for a given cell are the same, i.e. $w_{ik} = w_{ij}$ for all $j \neq k$, then the initial gradient with respect to each membership value w_{ik} will be identical, resulting in no net direction for the search. In other words, at each step, the relative cluster memberships μ_{ik} will remain unchanged. I break this symmetry through random initialisation, choosing the initial vector w_{ik} through random sampling of a uniform distribution centred on zero.

The specific clustering found by the algorithm is determined by the choice of initial vector. The *L-BFGS-B* algorithm is only guaranteed to identify local optima, so different initial vectors can result in finding different optima (Byrd et al., 1995; Zhu et al., 1997). To make the implementation robust to the exact choice of initial vector, the optimisation can be repeated multiple times with different initial vectors, choosing the clustering S with the greatest inter-cluster heterogeneity H_S.

The returned clustering is fuzzy. Nevertheless, biologically, cell types are typically assumed to be discrete. Moreover, H_S is generally maximised when there is no fuzziness in cluster membership; see **Fig 3.2** for a diagrammatic explanation. For example in the following section, I will cluster the Tian et al. (2019) and Zheng et al. (2017) data sets. Concerning both data sets, the greatest fuzzy membership weight for each cell μ_i, where $\mu_i = max(\mu_{ik})$ for $k = 1, \ldots, C$, is approximately equal to 1 for nearly all cells. Specifically,

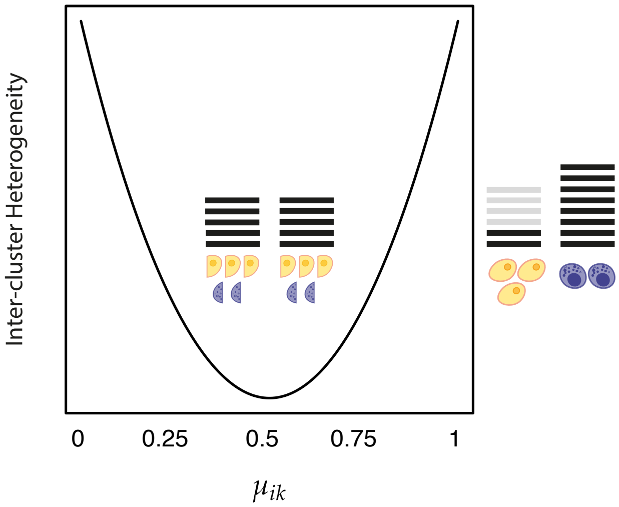


FIGURE 3.2: **Effect of fuzziness on inter-cluster heterogeneity**. At maximum fuzziness, where every cell belongs equally to each cluster, with $y_k = N_k/N = 1/C$, there is zero inter-cluster heterogeneity. Here the number of clusters is two, so maximum fuzziness corresponds $\mu_{ik} = 0.5$. As fuzziness reduces and cells are increasingly preferentially assigned to one cluster or another, inter-cluster heterogeneity can emerge, generally reaching a maximum when cells are discretely assigned to each cluster, represented in the diagram on the right-hand side. Importantly, not all discrete clusterings will exceed all fuzzy clusterings with respect to H_S. However, when the clustering is correct, the less fuzzy the clustering, the greater inter-cluster heterogeneity can be captured. Importantly, this only applies when the true clustering is genuinely discrete; see the results of clustering the Stumpf et al. (2020) data set in **Section 3.1.4**.

concerning the returned clustering, $\mu_i > 0.99$ for all cells in the Tian et al. (2019) data set and for 99.7% of cells in the Zheng et al. (2017) data set.

Note that the *L-BFGS-B* algorithm cannot return discrete memberships as 1) the algorithm is bounded and 2) discrete cluster memberships are asymptotic with respect to w_ik, with $\mu_{ik} = 1$ (and $\mu_{ij} = 0$ for $j \neq k$) only when $w_{ik} = \infty$. Therefore, I discretise the returned cluster memberships, assigning each cell solely to the cluster with the highest membership. The overall unsupervised clustering method, scEC, is therefore discrete (though I will discuss an exception to this discrete view regarding the Stumpf et al. (2020) data set in **Section 3.1.4**).

In the next section, I will validate scEC against various data sets. Specifically, I will confirm that scEC is able to (largely) recover the established classifications for the Tian et al. (2019), Zheng et al. (2017) and Stumpf et al. (2020) data sets where the cellular classification has been derived via experiment or differential gene expression analysis.

3.1.4 Validation

In **Chapter 2**, I introduced three data sets where the classification of cells has been established, either through experimental evidence (separate from gene expression as measured by single-cell RNA-sequencing) or through differential expression analysis of a set of clusters derived by unsupervised clustering. Specifically, the cells of the Tian et al. (2019) data belong to three different cancerous cell lines so are classified by cellular genotype; the cells of the Zheng et al. (2017) data set were separated based on surface protein expression prior to sequencing; and the cells of the Stumpf et al. (2020) data set were grouped into cell types via unsupervised clustering, with each cluster identified via differential gene expression analysis.

In this section, I cluster each of these three data sets using the scEC (single-cell Entropic Clustering) clustering method developed above. (Note that for each data set, I will cluster based on the top 500 genes by $I(g)$; thus, each clustering depends on only those genes likely to be differentially expressed). I will compare the scEC clustering of each data set with the established classification through the Adjusted Rand Index (ARI), where the ARI is a measure of similarity between two classifications, adjusted for similarity that may emerge from chance (Rand, 1971; Hubert and Arabie, 1985). An ARI of 1 indicates perfect alignment, and an ARI of 0 indicates no greater similarity than expected from chance. See the box *"Adjusted Rand Index"* for details of the adjusted rand index.

The ARIs achieved by scEC are benchmarked against those achieved by an alternative unsupervised clustering algorithm, the Louvain method (as implemented in the R package `Seurat`, v3.2.3) (Blondel et al., 2008; Stuart et al., 2019). I choose the Louvain method as it has been repeatedly recognised as the best performing unsupervised clustering algorithm for single-cell RNA-sequencing data, so it provides a benchmark for state-of-the-art clustering performance (Freytag et al., 2018; Duò et al., 2018; Luecken and Theis, 2019). For both scEC and the Louvain method, I specify the number of clusters, choosing the same number of cell types as identified in the established classification in each case, where $C = 3$ for Tian et al. (2019), $C = 4$ for Zheng et al. (2017) and $C = 14$ for Stumpf et al. (2020). The results of these analyses are detailed in **Table 3.1**.

The Tian et al. (2019) data set is relatively simple, with clear, discrete clusters of cell types. Both methods successfully recover the genotyped classification of the Tian et al. (2019) population, with both methods achieving an ARI of 0.99.

Adjusted Rand Index

The Adjusted Rand Index (ARI), as described in Hubert and Arabie (1985), is a measure of the overlap between two sets of cluster annotations, adjusted for the amount of overlap expected by chance. The ARI is 1 when two annotations, X and Y, overlap entirely, and 0 when they have a level of overlap that could be expected to arise solely by chance. Negative ARIs can be achieved if the overlap is worse than that expected by chance, though Hubert and Arabie (1985) did not derive a specific lower bound on ARI.

The calculation of the ARI is based on a contingency table of the form presented below, recording the number of cells, n_{ij}, belonging to cluster i in clustering X and cluster j in clustering Y. The total number of clusters in each data set is R for X and C for Y. The total number of cells in each data set, n, is the same.

Class	Y_1	Y_2	$\cdots$	Y_C	Sums
X_1	n_{11}	n_{12}	$\cdots$	n_{1C}	$n_{1\cdot}$
X_2	n_{21}	n_{22}	$\cdots$	n_{2C}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_R	n_{R1}	n_{R2}	$\cdots$	n_{RC}	$n_{R\cdot}$
Sums	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot C}$	n

The ARI of the above contingency table is found as,

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{n_{i\cdot}}{2} \sum_j \binom{n_{\cdot j}}{2}\right] \Big/ \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{n_{i\cdot}}{2} + \sum_j \binom{n_{\cdot j}}{2}\right] - \left[\sum_i \binom{n_{i\cdot}}{2} \sum_j \binom{n_{\cdot j}}{2}\right] \Big/ \binom{n}{2}},$$

where $\binom{n}{k}$ is the binomial coefficient, detailing the number of ways to choose an (unordered) subset of k elements from a fixed set of n elements.

Both methods perform well on the Zheng et al. (2017) data set; however, the Louvain method achieves a higher ARI (0.99) than scEC (0.87). Interestingly, the surface protein classification of the Zheng et al. (2017) cells captures less inter-cluster heterogeneity, 177 nats, than the scEC clustering, 199 nats. Thus, the difference in the cellular classification reflects a genuine disagreement in classification rather than poor optimisation of the objective function.

Importantly, both the established classification and the scEC clustering of the Zheng et al. (2017) data set represent forms of phenotypic classification (as

Data Set	scEC	Louvain
Tian et al. (2019)	0.99	0.99
Zheng et al. (2017)	0.87	0.99
Stumpf et al. (2020)	0.69	0.35

TABLE 3.1: **Similarity of clustering results to established classifications**. The similarity of the unsupervised clustering produced by either scEC or the Louvain method to the established classification for three data sets. Similarity is measured by the Adjusted Rand Index, where a value of 1 indicates perfect alignment and a value of 0 indicates no greater similarity than expected from chance.

discussed in **Section 2.6**). Indeed, where the established classification identifies cells with respect to the measured expression of a handful of marker genes, scEC clusters cells with respect to the measured expression of a large subset of all genes (recall that the scEC clustering is based top 500 genes by $I(g)$). The scEC clustering therefore represents a more exhaustive application of the principles of phenotypic classification.

Moreover, the Zheng et al. (2017) data set is a concatenation of multiple sequencing runs, where the cells of each type are sequenced separately. The apparent success of the Louvain method could be interpreted as the successful recovery of the separate sequencing runs via the associated batch effects (small differences in expression across many genes resulting from cells being in different batches) as opposed to the recovery of genuine biology. The scEC method is robust to such batch effects, as the James-Stein-type estimator minimises the influences of small changes in gene expression (see **Section 2.5.1** for discussion of James-Stein-type estimator). The absence of heterogeneity attributable to batch effects can be seen with respect to the Svensson et al. (2017) data set, which is a concatenation of two separate sequencing runs.

The $H_S(g)$ values associated with the novel scEC clustering and the clustering in the original Zheng et al. (2017) publication are strongly positively correlated (Pearson's correlation coefficient of 0.94; including all genes with at least 100 transcripts expressed across all cells) with a mean gene-wise difference, $\Delta H_S(g) = H_{novel}(g) – H_{original}(g)$, of 0.0019. Those genes used to separate the cell types on the level of surface protein expression – *CD14*, *CD4*, *CD8* and *NCAM1* – are each associated with a significant value of $H_S(g)$ with respect to the novel clustering (Zheng et al., 2017). Performing a gene ontology enrichment analysis of those genes where the value of $H_S(g)$ with respect to the novel clustering outperforms that of the originally published clustering by at least 0.1 nat, i.e. $\Delta H_S(g) > 0.1$, reveals a significant enrichment of cell cycle genes (p-value $= 1.33 \cdot 10^{-8}$, correction for false discovery rate) (Ashburner et al., 2000; Gene Ontology Consortium, 2021; Benjamini and Hochberg, 1995). Conversely, those genes where the original clustering substantially

	1	2	3	4
T-Cells	62929	1015	285	112
NK-Cells	983	7379	13	10
B-Cells	754	17	9311	3
Monocytes	35	6	25	2546

TABLE 3.2: **Information-theoretic clustering of immune cells**. Contingency table between surface protein-derived classification and the scEC clustering of the Zheng et al. (2017) data set. The columns represent the four scEC clusters and the rows represent the classification based on surface protein expression. Each element of the table encodes the number of cells in both an scEC cluster and a surface protein-based cell type.

outperforms the novel clustering, $\Delta H_S(g) < -0.1$, are enriched with respect to the immune response ontology term (p-value $= 1.49 \cdot 10^{-3}$, correction for false discovery rate). Therefore, the divergence in clustering is driven by the greater influence of the cell cycle on scEC. This influence results from the large number of genes involved in the cell cycle; in maximising the additive sum of $H_S(g)$ across individual genes, scEC is biased towards biological processes involving larger tranches of genes.

The greatest difference in performance between the two unsupervised clustering methods is with respect to the Stumpf et al. (2020) data set, with scEC achieving an ARI of 0.69 and the Louvain method an ARI of 0.35 (see **Table 3.3** for contingency table of the scEC clustering against the established classification). The Stumpf et al. (2020) data set is the most biologically complex, with cells sampled from the haemopoietic stem cell lineage. The haemopoietic stem cell lineage is a dynamic biological system in which cells are actively transitioning from one cell type identity to another. Cells transition along one of several different continuous developmental trajectories, each associated with a distinct lineage of cell types. The data set is therefore more difficult to classify, as the cells of the population do not truly identify with a single discrete type. Experimental classifications of the type in Tian et al. (2019) and Zheng et al. (2017) cannot be obtained for such biologically dynamic systems; instead, approximate (discrete) classifications have to be derived from computational clustering and validated based on differential expression of established marker genes (as discussed in **Section 1.3**).

The substantially better performance of scEC compared to the Louvain method on the Stumpf et al. (2020) data set is somewhat surprising, as scEC assumes each cell type to consist of a set of approximately interchangeable cells. In contrast, the Louvain method requires only that cells be similar to at least some other cells of the same type. Given that the scEC objective function H_S largely measures discrete differences in expression, it is interesting that

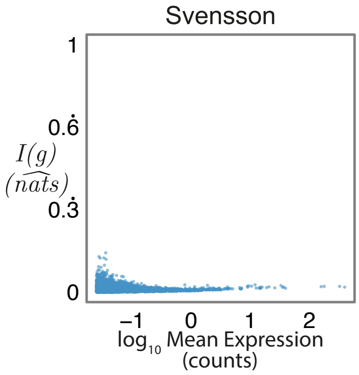


FIGURE 3.3: **Batch effect and the James-Stein-type estimator**. The Svensson et al. (2017) data set concatenates two separate sequencing runs, each of an equal number of cells. There is no visible effect of these separate batches on $I(g)$ as calculated based on James-Stein-type estimator of the distribution of the expression of each gene. Through additive decomposition, population heterogeneity represents the maximum possible inter-cluster heterogeneity, $H_S < I$, so batch effect likely has only a minimal influence over the clusterings produced by scEC.

information-theoretic clustering can recover the quasi-continuous cellular identities of the Stumpf et al. (2020) data set.

The primary difference between the three classifications (scEC, Louvain and established) is in the identities of the Erythrocyte clade (see **Fig 2.15** and **Fig 3.4**). In the established classification of the Stumpf et al. (2020) data set, the Erythrocyte cell type consists of cells drawn from along a continuous maturation process. scEC appears to identify three stages of maturation: stem-like, maturing and terminal Erythrocytes (see **Fig 3.4a** and **Table 3.3**). The Louvain method, by contrast, breaks the Erythrocytes clade into five distinct stages, substantially over-clustering the lineage relative to the established classification (see **Fig 3.4b**). The number of cell types is fixed at $C = 14$, so the over-clustering of Erythrocytes is balanced by the merging of other cell types originally identified as distinct in Stumpf et al. (2020). The merging of established cell types substantially reduces the ARI achieved.

The uncertainty in Erythrocyte identity is reflected in the original (i.e. pre-discretisation) fuzzy membership values, μ_{ik}, of the Stumpf et al. (2020) scEC clustering. Only 49.3% of cells are assigned to a single cluster with a weighting of at least $\mu_i > 0.95$. Of the cells with less certain assignments, $\mu_i < 0.95$, 87.3% are preferentially assigned to the Erythrocyte identity, i.e. cluster number 7 in **Table 3.3**. These Erythrocyte cells have equal probability, $\mu_i k$ of being assigned to both cluster number 7 and cluster number 14. I discretise the identity of the cells by arbitrarily choosing one of the two clusters, assigning all cells to cluster 7 (this results in one fewer discrete cluster being returned than originally specified). Importantly, this uncertainty is

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Erythroblasts	0	0	233	0	0	0	2436	1	0	171	1	0	3	0
Myeloblasts	313	0	0	0	245	0	39	0	0	4	0	0	2	0
Pro-B	0	0	0	187	0	0	38	0	0	0	0	0	0	0
Monocytes	0	2	0	23	23	165	0	0	0	0	4	0	0	0
Basophils	0	0	2	0	1	0	0	130	0	0	0	0	0	0
Pericytes	0	2	0	0	0	0	0	0	0	0	0	110	0	0
Pre-B	0	0	0	58	2	0	0	0	0	0	2	0	0	0
Megakaryocytes	0	0	21	0	0	0	0	0	40	0	0	0	0	0
T-NK	0	0	0	0	1	0	3	0	0	0	56	0	0	0
Endothelial Cells	0	41	4	0	0	1	0	0	0	0	0	4	0	0
HSPCs	3	0	182	5	50	1	49	7	0	4	2	0	0	0
Monoblasts	16	0	3	7	200	66	7	0	0	0	0	1	0	0
Neutrophils	0	0	0	1	8	1	0	1	0	0	1	0	255	0
Myelocytes	4	0	0	3	93	0	2	0	0	0	0	0	164	0

TABLE 3.3: **Information-theoretic clustering of mouse bone marrow data set**. Contingency table between differential gene expression analysis validated classification and the scEC clustering of the Stumpf et al. (2020) data set. The columns represent the scEC clusters, the rows the established classification. Each element of the table encodes the number of cells in both an scEC cluster and an established cell type. Note that while the number of clusters was set to 14 for scEC, only 13 clusters were realised when discretised: in the developed implementation of scEC, only the maximum number of clusters can be specified.

genuine. The discretised memberships outperform the fuzzy memberships, with $H_S = 608$ nats and $H_S = 574$ nats, respectively; however, the discrete clustering is itself slightly outperformed by the case where only those cells with near-certain assignments, $\mu_{il} > 0.95$, are discretised, with $H_S = 610$ nats. Thus, scEC can, to an extent, identify cells with uncertain cellular identities.

The $H_S(g)$ values of the novel scEC clustering and the clustering in the original Stumpf et al. (2020) publication are strongly positively correlated (Pearson's correlation coefficient of 0.99), with a mean $\Delta H_S(g)$ of 0.024. All of the primary lineage marker genes (stem and progenitor cells, *Cd34*; niche cells, *Kitl*; myeloid lineage, *Spi1*; erythroid lineage, *Gata1*; lymphoid lineage, *Pax5*) and all of the cell type specific markers (Erythrocytes: *Car2*, *Hemgn*, *Hba-a2*, and *Aldh1a1*; Granulocytes: *Elane*, *Ltf*, *Retnlg*, and *Mcpt8*; Monocytes: *Irf8*, *Klf4*, *Ccr2*, and *Cx3cr1*; Stem cells: *Kit*, *Angpt1*, *Kitl*, and *Tek*; Endothelial cells: *Cdh5*, *Cxcl12*, *Kdr*, and *Lepr*; Lymphocytes: *Flt3*, *Il7r*, *Cd19*, and *Ms4a1*) identified in Stumpf et al. (2020) are associated with significant values of $H_S(g)$ with respect to the novel (and original) clustering. With respect to gene ontology terms, those genes with $\Delta H_S(g) > 0.1$ are significantly enriched with respect to Erythrocyte differentiation and homeostasis (p-value $= 1.80 \cdot 10^{-2}$ and p-values $= 6.17 \cdot 10^{-4}$, correction for false discovery rate); conversely, those genes with $\Delta H_S(g) < -0.1$ are significantly enriched with respect to the

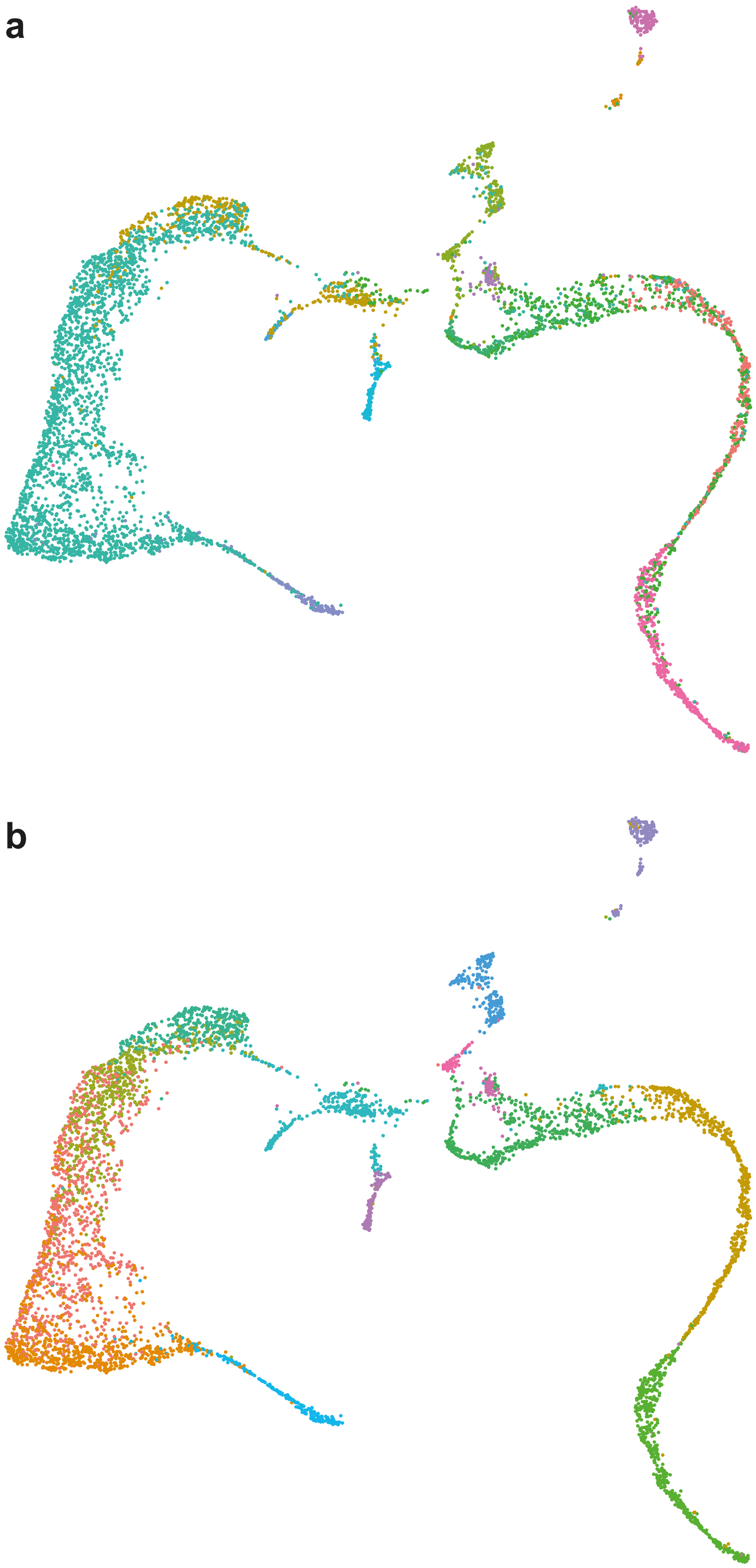


FIGURE 3.4: **Visualisation of scEC clustering of mouse bone marrow data set**. Nonlinear dimension reduction (UMAP) of cellular gene expression (McInnes et al., 2018). Cells are coloured either by **a)** scEC clustering or **b)** Louvain clustering results. The Erythrocyte clade (left-hand side) is split into either **a)** three or **b)** five clusters. See **Fig 2.15** for visualisation with established cluster identities. Processing and visualisation of genes performed using the `Seurat` package (Stuart et al., 2019).

immune response (p-value $= 2.57 \cdot 10^{-18}$, correction for false discovery rate)(Ashburner et al., 2000; Gene Ontology Consortium, 2021; Benjamini and Hochberg, 1995). This difference in genes corresponds in the shift in the novel clustering from separating the cell types of the Neutrophil lineage to those of the Erythrocyte clade, suggesting that the various stages of the Erythrocyte clade are separable enough to be considered distinct cell types, in contrast to the considerations of Stumpf et al. (2020).

The reduced ARIs of both clustering methods on the Stumpf et al. (2020) data set relates to a broader point – that the performance of an unsupervised clustering method depends strongly on the number of clusters specified. Both scEC and the Louvain method optimise their respective objective functions for a given number of clusters, with the number of clusters set via hyperparameter and not learned from the data. More generally, deciding on the appropriate number of clusters is an active and important choice in unsupervised clustering, with most methods requiring that the number of clusters be specified by hyperparameter (Von Luxburg et al., 2012; Kiselev et al., 2019).

There are generic tools available for choosing the number of clusters, e.g. the GAP statistic, see Tibshirani et al. (2001). However, in specifying the number of clusters, a strong statement is being made about the biology of a given cellular population, with such tools attempting to infer the number of cell types present in a population. In **Chapter 4**, I will develop one such tool for estimating the true number of clusters in a single-cell expression data set, based on a novel quantification of heterogeneity with respect to the joint distribution of gene expression.

An alternative approach to learning the true cluster number is to leverage a reference data set. For example, it is increasingly common to have access to a previously classified single-cell RNA-sequencing data set of the biological system of interest (Lotfollahi et al., 2021; Regev et al., 2017; Tabula Muris Consortium et al., 2018; Li et al., 2021a; The Tabula Sapiens Consortium and Quake, 2021). These data sets can act as references for the semi-supervised or supervised classification of novel data sets, with the number of cell types in the reference providing the maximum number of clusters in the novel test data set. In the next section, I will develop a semi-supervised classification method for single-cell expression data, extending scEC to the semi-supervised setting.

3.2 Semi-supervised Classification

The success of single-cell RNA-sequencing and unsupervised clustering has led to the ambition to sequence every cell type in humans and model organisms. These large atlas projects can take the form of a single large experiment, as typified by the Tabula projects, or a distributed organ-wise approach, as seen with the massive Human Cell Atlas project (Tabula Muris Consortium et al., 2018; Li et al., 2021a; The Tabula Sapiens Consortium and Quake, 2021; Han et al., 2020; Regev et al., 2017).

These atlases can serve as *reference transcriptomes*, in analogy to reference genomes (Nurk et al., 2021). Thus, for a given experimental data set: instead of undertaking the full process of pre-processing, unsupervised clustering and annotation by marker genes, cells can instead be mapped to the established cell types of a relevant cell atlas (Kiselev et al., 2018; Luecken and Theis, 2019). Each cell can then be identified not by unsupervised clustering but by supervised or semi-supervised classification based on known cell types. Much as sampling reads are not *de novo* characterised each time but mapped to the reference genome, the repeated *de novo* classification of cells can be avoided with cells instead mapped to a reference transcriptome.

However, this process is not as simple as with the genome, in which sequences of DNA can be mapped exactly to their place along the 1-dimensional genome. With the transcriptome, the problem is both high-dimensional and inexact, with the intra-type heterogeneity arising from biological noise and technical error prohibiting exact matching of cells to types. Nevertheless, methods for mapping data sets to reference transcriptomes, such as Kiselev et al. (2018), Pliner et al. (2019), Stuart et al. (2019), & Lotfollahi et al. (2021), are a fast-growing area of single-cell analysis (Zappia and Theis, 2021).

In this section, I will extend scEC to the problem of semi-supervised classification. That is, I will cluster an unclassified test data set based on a mixed data set of both classified and unclassified cells, where the classified cells are sourced from an established reference transcriptome (Regev et al., 2017; Tanay and Regev, 2017; Stuart et al., 2019; Luecken and Theis, 2019). To cluster the mixed data set, I again maximise inter-cluster heterogeneity, with the distinction that a subset of cells, i.e. those from the reference data set, have pre-assigned discrete cluster memberships. Importantly, this clustering directly classifies the cells of the test data set, as each cell inherits the classification of the reference cluster it is assigned to. The resulting method is semi-supervised because the classification of each cell depends both on the reference data set and on the test data set. (In contrast, in a supervised method, the classification of each cell would depend solely on the reference data set).

More formally, let us consider a pair of count matrices, $T = \{T_{ref}, T_{test}\}$, where one count matrix, which I will term the reference, T_{ref}, has a known cluster structure (and therefore each cell has a unique, discrete cell type classification). The goal of reference mapping is to classify cells in the test matrix, T_{test}, based on cellular classifications of the reference matrix, which I accomplish by semi-supervised clustering of the mixed data set.

I first normalise each count matrix separately, using the James-Stein-type estimator so that $\sum_{i \in T_k} x_i^g = 1$ (see **Section 2.5.1** for discussion on James-Stein-type estimator) (James and Stein, 1992). To combine the normalised count matrices, I assume that T_{ref} and T_{test} are interchangeable with respect to gene expression. Recall that when clusters of cells are interchangeable with respect to gene expression, the proportion of transcripts expressed in each cluster is N_k/N, where N_k is the number of cells in that cluster (see **Section 2.3.1**). Therefore, treating both normalised count matrices as interchangeable with respect to gene expression, the normalised, combined count matrix, X_{mix}, is given by the weighting:

$$X_{mix} = \left[\frac{N_{ref}}{N} X_{ref}, \frac{N_{test}}{N} X_{test} \right], \tag{3.24}$$

where N_{ref} is the number of cells in the reference data set, N_{test} is the number of cells in the unclassified data set, and N is the total number of cells across both data sets.

As with scEC, I cluster the combined data set, X_{mix}, by maximising the inter-cluster heterogeneity of the combined data set with respect to the fuzzy clustering S. However, unlike with scEC, a subset of cells have already been identified. Let $R = R_1, \ldots, R_C$ be the discrete clustering of the reference data set, with cluster sizes $n_1, \ldots, n_C$, where $\sum_{k=1}^{C} n_k = N_{ref}$. I define each cluster in the reference data set as a subset of a cluster in the mixed data set, $R_k \in S_k$, where $S = S_1, \ldots, S_C$ is the clustering of the mixed data set. Thus, the number of clusters in the mixed data sets equals the number of clusters in the reference.

Based on the mixture of known and unknown cellular identities, I define y_k^g and N_k of the combined data set X_{mix},

$$y_k^g = \sum_{i=1}^{N_{test}} \mu_{ik}\, x_i^g + \sum_{i \in R_k} x_i^g, \text{ and} \tag{3.25}$$

$$N_k = \sum_{i=1}^{N_{test}} \mu_{ik} + n_k, \tag{3.26}$$

where each variable is a sum of both fuzzy and discrete terms.

Using the above formulations of y_k^g and N_k, the definition of H_S follows as before, see **Eqn 2.5**, and so the objective function remains the same, **Eqn 3.10**. The derivative of the objective function follows as before, see **Eqns 3.13** through **3.19**, except that the membership function, μ_{rq} is only defined with respect to cells of the test data set, $1 \leq r \leq N_{test}$. Therefore, the gradient function is the softmax formulation, **Eqn 3.22**, defined for $1 \leq i \leq N_{test}$.

I adopt the same implementation of the *L-BFGS-B* algorithm as for unsupervised clustering, with the exception that each element of the initial vector of cluster memberships, w_{ik} is set to 0. The known cellular identities of the reference data set introduce the necessary asymmetry for numerical optimisation; this eliminates the stochasticity of the initial choice of w_{ik} and the corresponding need for randomised initialisation. Moreover, because the reference transcriptome provides the number of clusters, the optimisation statement is now complete, and so has no tunable parameters.

For computational efficiency and the minimise the effects of technical and biological noise, I include only those genes associated with substantial $H_S(g)$ with respect to the reference data set. Implicitly this restriction assumes that the reference transcriptome provides a complete description of the biological system: if the reference data captures the complete set of cell types in the biological system, only those genes with substantial $H_S(g)$ with respect to the reference data set should be relevant to the cellular classification. This reference-based feature selection increases the degree to which the classification of the test data set is 'supervised' by the reference transcriptome.

Robustly validating semi-supervised or supervised clustering methods is difficult as two distinct data sets with independently established cellular classification are required, where the set of cell types in the test data set is a subset of the cell types in the reference. Tian et al. (2019) generated such a pair of data sets, namely the *Tian3* and the *Tian5* data sets. Up to now, I have utilised only the *Tian3* data set, sequenced from a mixture of three cancerous

Data Set	Cells (N)	Cell Lines
Tian3	902	H1975, H2228, HCC827
Tian5	3918	H1975, H2228, HCC827, H838, A549

TABLE 3.4: **Cancer cell line data sets**. Tian et al. (2019) generated two data sets, one sequenced from a mixture of three cancerous cell lines, *Tian3*, and one sequenced from a mixture of five cancerous cell lines, *Tian5*.

	H1975	H2228	HCC827
H1975	276	2	0
H2228	0	308	1
HCC827	0	0	258
A549	29	5	15
H838	8	0	0

TABLE 3.5: **Semi-supervised classification of cancer cell lines.** Semi-supervised classification of *Tian3*, with *Tian5* serving as the reference data set (Tian et al., 2019). The columns represent the established genotypic classification of the cells of the three cell lines data set, *Tian3*, and the rows represent the semi-supervised classification of cells based on the *Tian5* data set. The semi-supervised classification achieves an adjusted rand index of 0.90.

cell lines (recall cell lines are proxies for cell types). The *Tian5* data set is larger, containing both more cells and an additional two cell lines, see **Table 3.4**.

I cluster both data sets together, with *Tian3* serving as the test data set and *Tian5* as the reference data set. The resulting clustering substantially overlaps with the established genotypic classification of the *Tian3* data set, achieving an ARI of 0.90, see **Table 3.5**. Thus, the semi-supervised classification method successfully recovers the established cell types of the *Tian3* data set.

3.3 Discussion

In this chapter, I have developed a pair of novel single-cell clustering methods based on the assumption that cells of the same type should be (at least approximately) interchangeable with respect to the measured expression of a set of genes. I have built on the information-theoretic framework developed in **Chapter 2**, extending that framework to quantify the proportion of heterogeneity attributable to differential gene expression between a set of overlapping, fuzzy clusters. The extension to the fuzzy setting enables the efficient optimisation of H_S with respect to S, identifying the set of clusters for each data set that is maximally differentially expressed and minimally divergent from the assumption that cells of the same type are interchangeable with respect to measured gene expression (by the additive decomposition of I; see discussion in **Section 2.6**).

The scEC method applies the process of phenotypic classification to single-cell RNA-sequencing data. The fundamental assumption of phenotypic classification is that the cells of each type are interchangeable with respect to marker gene expression. As outlined in **Section 2.6**, the measure h_S extends this assumption genome-wide, quantifying the deviation of the measured patterns of gene expression from the hypothetical case where the cells of each cluster are interchangeable with respect to the expression of many genes. Thus, given that maximising H_S minimises h_S, scEC can be interpreted as automating the phenotypic classification of cells with respect to single-cell expression data.

The scEC method directly quantifies the importance of each gene to the final clustering through $H_S = \sum_g H_S(g)$, with those genes associated with high values of $H_S(g)$ representing putative marker genes. The final clustering, and so the list of genes identified as putative markers, does not depend on the differential expression of any one gene: when clustering based on hundreds or thousands of genes, maximising the total H_S depends on maximising the value of $H_S(g)$ across many individual genes. Thus, the optimal clustering depends on the number of differentially expressed genes with respect to each clustering and the strength of each gene's differential expression (measured through $H_S(g)$).

Therefore, scEC uncovers and leverages the coordination (statistical dependency) between the expression patterns of different genes. Given that the value of $H_S(g)$ associated with each gene is found with respect to the same clustering, the expression of those genes that contribute most to the optimal H_S will tend to be correlated with other highly-contributing, i.e. high $H_S(g)$, genes. Thus, the scEC method implicitly assumes that cell types emerge from the coordinated expression of many genes, an assumption similarly made by the more traditional single-cell clustering methods (as discussed in **Section 1.2** in reference to gene regulatory networks and dynamical systems theory).

Traditional clustering methods uncover the coordination in gene expression in a two-step process: first, the Euclidean distance quantifies the similarity in gene expression between individual pairs of cells; then, cells are clustered based on the Euclidean distance, indirectly identifying any coordination in gene expression patterns (as discussed in **Section 1.2.3**). The scEC method avoids the intermediate step, directly identifying any coordination between the expression patterns of individual genes.

This difference in identifying coordination with respect to the expression patterns of individual genes reflects a more fundamental difference between scEC and traditional single-cell clustering methods: traditional unsupervised clustering methods assume cell types emerge as regions of high probability density in gene expression space. Philosophically (although often implicitly) this assumption stems from a dynamical systems theory view of cell fate (as discussed in **Section 1.2.2**) (Greulich et al., 2020). In contrast, scEC assumes that cells of the same type are approximately interchangeable with respect to measured gene expression, an assumption stemming from the empirical principles of phenotypic classification.

With these considerations in mind it is notable that both scEC and the Louvain method performed comparably well with respect to retrieving established cellular classifications, suggesting that both conceptions of cell type are reasonable despite their apparent philosophical differences. While the full reasons for this concordance are not clear (see **Chapter 5** for discussion on the philosophical connections between the two approaches), it does suggest that cells of the same type are at least approximately interchangeable and the traditional working definition of cell type still has relevance in the world of high-throughput single-cell RNA-sequencing data (Casey et al., 2020b).

Moreover, although the Louvain method is a more generally powerful unsupervised clustering method than scEC (not least because it has been validated across diverse data sets beyond single-cell RNA-sequencing), scEC has a more intuitive biological interpretation that will be familiar to many experimentalists. While the assumption that cell types emerge as dynamic attractors in gene space is mathematically appealing, it is empirically non-intuitive and can confound, rather than help, the distillation of biological knowledge from complex single-cell data sets (Newman, 2020). Thus, the choice of clustering method depends on whether power or interpretability is preferred, the specifics of the data and the practitioner's background.

The supervised and semi-supervised classification of cells based on reference transcriptomes is a burgeoning and rapidly growing field of single-cell analysis (Zappia and Theis, 2021). Where scEC extends the principles of phenotypic classification genome-wide, (semi-)supervised classification directly extends the process of phenotypic classification genome-wide. In supervised classification, each cell or cluster of a test data set is compared to the clusters of the reference data set. With respect to the reference data, each cluster represents a cell type: thus, instead of classifying cells/clusters based on the expression of a handful of marker genes, cells/clusters are classified with respect to the expression of all genes (or at least a large subset of all genes). By classifying each cell type with respect to genome-wide gene expression, semi-supervised and supervised classification approaches offer a more robust approach to classification than the traditional marker gene-based approach to classifying clusters (Luecken and Theis, 2019).

Existing computational classification approaches tend to be fully supervised, directly mapping cells or clusters of cells in the unclassified data set onto the reference (Kiselev et al., 2018; Pliner et al., 2019; Stuart et al., 2019; Lotfollahi et al., 2021). With supervised approaches, classification depends solely on the gene expression patterns of the reference data set, and with unsupervised clustering, cellular classification depends solely on the gene expression patterns of the test data set. The approach I have developed in **Section 3.2** is distinguished by being semi-supervised, leveraging information from both test and reference data sets in classifying cells.

Semi-supervised classification provides a more robust approach to classifying cells than wholly supervised methods. The differential expression patterns of specific genes in the reference may result from technical error or from some biological process specific to the reference population. Cellular classification based solely on the reference data set can therefore be overfit (i.e. be overly specific) to the reference data. The semi-supervised approach developed here minimises the risk of overfitting by utilising the expression patterns of the test data set and the reference.

The need to minimise overfitting is particularly important with respect to cross-species classification. Many cell types are present in multiple species – for example, the cell types of the hematopoietic stem cell lineage are largely conserved between mouse and human, see Stumpf et al. (2020). Importantly, reference data sets may be available for a more experimentally tractable species (e.g. mouse) than for other, less easily studied species (e.g. humans). However, in classifying cells cross-species, the risk of overfitting to one species is substantial: semi-supervised classification offers a potentially more robust approach to computational classification.

Having validated the semi-supervised classification algorithm on the dual Tian et al. (2019) data sets, in future, I would aim to apply the semi-supervised classification method to larger-scale reference transcriptomes, for instance, the Human Cell Atlas (Regev et al., 2017). I would also aim to test the classification of cells cross-species, a task in which semi-supervised classification should have a substantial advantage.

In **Chapters 2** & **3**, I have formalised and automated the clustering and classification of cells. However, as discussed, scEC only maximises H_S over a subset of all possible clusterings. Specifically, scEC maximises H_S over the set of clusterings, S_C, containing a fixed number of clusters, C. Indeed, the value of $H_S(g)$ will, all else being equal, increase monotonically with the number of clusters (see description of the property of monotonicity in **Section 2.1**) (Shannon, 1948; Kullback and Leibler, 1951). Thus, the number of clusters must be set as a hyperparameter for scEC. In the next chapter, I will develop a method for estimating the true number of clusters (and thereby number of cell types) in a heterogeneous population of cells. For this, I will return to the more traditional multivariate view of gene expression, i.e. the joint distribution of gene expression. However, instead of representing cellular gene expression profiles as position vectors in a high-dimensional space (as is usual), I will develop and utilise a novel hypergraph representation of gene expression data.

Chapter 4

Estimating Cluster Number

Introduction

In **Chapter 3**, I developed a novel unsupervised clustering algorithm based on the optimisation of inter-cluster heterogeneity. However, this optimisation problem, and the optimisation problems underlying many unsupervised clustering methods, are incomplete, requiring the number of clusters to be specified via a hyperparameter (Von Luxburg et al., 2012). There are generic tools available for inferring the number of clusters, e.g. the GAP statistic, but choosing the *true* number of clusters is a domain-dependent problem; for instance, choosing the number of clusters specifies the number of cell types in a cellular population (Tibshirani et al., 2001; Von Luxburg et al., 2012). Therefore, this chapter will develop a novel tool for inferring the true number of clusters in a single-cell RNA-sequencing data set.

In **Chapters 2** & **3**, I adopted a univariate view of gene expression, viewing the expression distribution of each gene individually. However, as discussed in **Section 1.2**, cell types emerge from the coordinated action of many genes, with this coordination inducing statistical dependency between the expression of different genes. For example, recall the Sonic Hedgehog gene regulatory network from **Section 1.2.1**: the expression of any one of *Olig2*, *Pax6* or *Nkx2.2* depends on the expression of at least one of the others (Balaskas et al., 2012). To infer the number of clusters, independent of any specific clustering, a multivariate view of gene expression is required, viewing the joint distribution of the expression of a set of genes.

The joint distribution of gene expression encodes the full expression profile of each cell, capturing any coordination between the expression of different genes. The joint distribution is typically represented as a high-dimensional gene expression space, where each cell's gene expression profile positions the cell in the space (Kiselev et al., 2019). Under the dynamical systems framework, cell types emerge as regions of high probability in gene expression space, analogous to 'noisy' attractors (see **Section 1.2.2**)(Greulich et al., 2020). Thus, if this space could be visualised without distortion, i.e. without dimension reduction, the number of cell types would be evident, with cells densely grouped in distinct regions of gene expression space.

Nevertheless, the full, high-dimensional joint distribution of gene expression cannot be visualised. Instead, the number of cell types must be inferred indirectly through the mathematical properties of the joint distribution of gene expression. Specifically, I propose that the number of cell types (and clusters) in a cellular population can be inferred by quantifying heterogeneity with respect to the joint distribution of gene expression.

Recall that a set of cells are heterogeneous with respect to the expression of a single gene g when the cells are distinguishable, i.e. are not interchangeable, based on the observed expression of g. I define a set of cells as heterogeneous with respect to the expression of the set of genes $g = 1, \ldots, G$ when the cells are distinguishable based on the joint distribution of gene expression.

I have shown that gene expression heterogeneity broadly increases with more cell types present: recall that the number of genes associated with substantial $I(g)$ increased with an increasing number of cell types. However, the number of cell types cannot be inferred from the sum of each gene's associated value of heterogeneity, $I = \sum_g I(g)$, due to the 'curse of dimensionality' (see **Section 2.5.2**)(Beyer et al., 1999). Instead, an inherently multivariate measure is required, quantifying the heterogeneity with respect to the joint distribution of gene expression.

In this chapter, I will assume (and later prove) that the heterogeneity with respect to the joint distribution of gene expression increases monotonically with the number of cell types. Based on this assumption, I will infer the number of clusters in a given test data set based on the expression heterogeneity of previously classified data sets. Importantly, this inference does not require that the novel and classified data sets be related, i.e. the number of cell types can be determined without access to a specific reference transcriptome (see **Section 3.2** for discussion of reference transcriptomes).

To that end, this chapter adopts an alternative view of the joint distribution of gene expression to the conventional high-dimensional gene expression space. Specifically, I will introduce a novel mathematical form for representing gene expression data, the hypergraph. Hypergraphs are a generalisation of graphs, which themselves have a range of powerful applications in single-cell analysis, for example, in unsupervised clustering and non-linear dimension reduction (Blondel et al., 2008; McInnes et al., 2018). Moreover, recent advances in hypergraph theory make hypergraphs an increasingly attractive option for representing gene expression data (Klamt et al., 2009; Jost and Mulas, 2021).

I will begin this chapter by discussing graph theory and the generalisation of graph theory to the setting of hypergraphs. I will demonstrate that hypergraphs offer a natural representation of the joint distribution of gene expression and show that the mathematical properties of hypergraphs enable the development of a novel measure of gene expression heterogeneity. I will then apply this measure to a range of data sets, see **Tables 2.1** & **3.4**, illustrating the use of this measure in inferring the number of cell types in a cellular population.

Note that the work of this chapter has been published in Mulas and Casey (2021), with the work of that publication carried out in collaboration between the present author and Raffaella Mulas (RM). Specifically, the present author and RM contributed equally to the conception of the idea to apply hypergraphs to single-cell RNA-sequencing data as outlined in **Section 4.2**, RM derived the mathematical properties presented in **Section 4.3** with biological interpretation provided by the present author, and the present author undertook the computational work and interpretation of the results in **Sections 4.4** & **4.5**. The initial introduction to graph theory in **Section 4.1** is not included in Mulas and Casey (2021).

4.1 Introduction to Graph Theory

Recall from **Section 1.2.3** that an undirected graph, $G = \{V, E\}$, consists of a set of N vertices, $V = 1, \dots, N$, and a set of G edges, $e_{ij} \in E$, where each edge is defined with respect to a pair of vertices i and j. The edges can be weighted or unweighted, with weighted edges assigned some coefficient, w_{ij}. In a weighted graph, $w_{ij} = 0$ indicates that the vertices i and j are not connected by an edge (Von Luxburg, 2007).

Graphs are useful in representing the connectivity or shape of a given data set. For example, as discussed in **Section 1.2.3**, the Louvain method encodes single-cell sequencing data sets in a graph structure, representing each of N cells as one of N vertices (Blondel et al., 2008). Specifically, the Louvain method encodes the distance between each cell i and the k most similar cells via weighted edges, forming a k-nearest-neighbour graph. This graph captures the core shape of the data, allowing the identification of clusters via modularity optimisation (Newman and Girvan, 2004).

Modularity, the objective function of the Louvain method, is not directly a property of the graph G, but of the associated adjacency matrix A. Recall that the weighted adjacency matrix A_{ij} encodes the weight of the edge e_{ij}, provided an edge is present between cells i and j in the graph and that the degree of each vertex (cell), v_i, is the sum of the weights of each edge connected to it, $v_i = \sum_j A_{ij}$. Recall that the modularity of a weighted graph with respect to a given clustering of vertex S is then

$$Q_S = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \gamma \frac{v_i v_j}{2m} \right] \delta(S_i, S_j), \tag{4.1}$$

where $m = \frac{1}{2} \sum_{ij} A_{ij}$ is the total number of edges in the graph, and $\delta(S_i, S_j)$ is 1 if cells i and j are in the same cluster and 0 otherwise (Newman and Girvan, 2004; Blondel et al., 2008).

In general, matrix representations of graphs are useful as they are more computationally tractable. Specifically, matrix representations allow us to apply the tools of linear algebra to compute informative quantitative summaries of graphs, such as Q_S.

An alternative set of matrix representations are the graph Laplacians (Chung and Graham, 1997). The graph Laplacian is inspired by a continuous operator from physics of the same name that quantifies diffusion through a continuous space; analogously, the graph Laplacian quantifies the 'diffusion' of information through a graph structure (Evans, 2010). There are various ways to compute a graph Laplacian, L, with each Laplacian capturing different structural properties of a graph (Chung and Graham, 1997; Von Luxburg, 2007). One such Laplacian, known as the normalised Laplacian, is defined as,

$$L = I - D^{-1} A, \tag{4.2}$$

where I is the identity matrix and D is the degree matrix of a graph (D is a diagonal matrix where the i^{th} diagonal element encodes the degree, v_i, of the vertex i and the off-diagonal elements are 0). Importantly, in the next section, **Section 4.2**, I will introduce a hypergraph analogue of the normalised Laplacian.

As a matrix, the various tools of linear algebra can be applied to the normalised Laplacian. In particular, the set of N eigenvalues of L can be computed, where the eigenvalues of the Laplacian are known as the *spectrum* of the graph (Chung and Graham, 1997). The spectrum of a graph captures information on the connectivity of the graph; for example, the multiplicity of zero eigenvalues encodes the number of connected components in a graph (Von Luxburg, 2007). Moreover, when the vertices of a graph represent a set of data points or objects, the eigenvalues and eigenvectors of the Laplacian can be used in spectral clustering, an alternative graphical clustering method to modularity optimisation with demonstrated application to single-cell expression data (Li et al., 2021b).

Graph theory yields various powerful tools for use in data analysis. However, graphs are limited with respect to representing higher-order relationships between objects or data points. For example, in a protein complex, many proteins can be bound together, with the collective interaction of these proteins giving rise to the function of the complex (Klamt et al., 2009). Such a complex can be represented using a graph, encoding the complex as a series of connections between each pair of proteins. However, the functionality of the complex does not emerge from the pair-wise connections but the collective interaction of all the proteins in the complex (Alberts, 2017).

There are two generalisations of graphs that can represent such higher-order relationships: simplicial complexes and hypergraphs. Simplicial complexes encode higher-order relationships hierarchically; for example, three-way connections are built from three pair-wise interactions, four-way connections are built from four three-way connections, etc. (Zomorodian, 2005). Hypergraphs are a further generalisation, encoding higher-order connections without the requirement for lower-order connections between vertices (Jost and Mulas, 2021).

In the next section, I will exploit the substantial generality and flexibility of hypergraphs to develop a natural representation of single-cell sequencing count data. I will begin by introducing hypergraphs and the computation of a specific hypergraph Laplacian. I will then demonstrate how hypergraphs encode a natural representation of the joint distribution of gene expression, with correspondence to the microscopy-view of gene expression adopted in **Chapters 2** & **3**. Finally, I will demonstrate how the spectrum of the single-cell hypergraph Laplacian provides a measure of heterogeneity with respect to the joint distribution of gene expression.

4.2 Hypergraph Construction

Hypergraphs can be constructed analogously to graphs. A hypergraph Γ, consists of a set of N vertices, $V = \{1, \dots, N\}$ and G hyperedges, $E = \{e_1, \dots, e_G\}$, where each hyperedge is an arbitrary non-empty set of vertices, that is, $e_g \subseteq V$ for each $g = 1, \dots, G$. The cardinality of a hyperedge e_g is the number of vertices, $|e_g|$, that are associated with the edge. The notion of edge weights generalises to the hypergraph setting: in any hypergraph, a real coefficient, $c(i, g)$, can be associated with each vertex i for each hyperedge e_g.

Thus, the key distinction between a graph and a hypergraph is that a single edge (now hyperedge) can encode a relationship between any number of vertices. In graphs, the cardinality of each edge must be two, whereas, with hypergraphs, any number of vertices can be associated with a given hyperedge. Hypergraphs are, as such, a more general mathematical formulation than graphs.

However, with this generalisation, it is not immediately clear what properties of graphs carry through to hypergraphs. Jost and Mulas (2021) derive a generalisation of the normalised Laplacian and the associated spectrum (introduced above; see **Section 4.1**) to hypergraphs with normalised real coefficients.

First, note that the coefficients of a hypergraph are *normalised* when the degree of each vertex is 1, i.e.,

$$\sum_g c(i, g) = 1 \quad \text{for each vertex } i. \tag{4.3}$$

The calculation of the normalised Laplacian then follows the same form as in **Eqn 4.2**, with adjustment of each term for the hypergraph setting. The degree of each vertex i is now given by,

$$v_i = \sum_g c(i, g)^2, \tag{4.4}$$

with the degree matrix, D of Γ again being a diagonal matrix where the i^{th} element of the diagonal encodes the degree, v_i, of vertex i, with off-diagonal values of 0.

The adjacency matrix is given by,

$$A_{ij} = -\sum_{g} c(i,g) \cdot c(j,g) \quad \text{for all } i \neq j, \tag{4.5}$$

where the diagonal elements of the adjacency matrix are 0, i.e. no vertex is connected to itself. The strength of the relationship between each pair of vertices i and j, encoded in the element A_{ij}, depends on the magnitude of the coefficients of each vertex with respect to each hyperedge.

Finally, an NxG matrix specific to hypergraphs, the incidence matrix Id, is required,

$$Id_{ig} = c(i,g), \tag{4.6}$$

encoding the coefficient of each vertex with respect to each edge.

The normalised hypergraph Laplacian (from here on referred to as the Laplacian) is then,

$$L = Id - D^{-1}A. \tag{4.7}$$

As in the graph setting, the N eigenvalues of the hypergraph Laplacian can be computed, with the eigenvalues of L forming the spectrum of a hypergraph. As with the graph spectrum, the hypergraph spectrum encodes various useful summaries of data (Jost and Mulas, 2021). Specifically, it is from the hypergraph spectrum that I will derive a measure of heterogeneity with respect to the joint distribution of gene expression. However, I will first introduce how hypergraphs can be used to represent single-cell RNA-sequencing data.

4.2.1 Single-cell Hypergraphs

Consider the expression of a set of G genes in a population of N cells. Let m_i^g be the number of transcripts of gene g measured in cell i and let $\sum_g m_i^g = M_i$ be the total number of transcripts measured in cell i. Note that the measured number of transcripts may differ from the true number due to error in the measurement process.

As in the graph setting, each of the N cells can be represented as one of N vertices in the hypergraph Γ. Each of the G genes can be represented as a hyperedge, $e_g \in E$, and the expression of each gene in each cell as the coefficient $c(i,g)$. (Note that I utilise the same indexing for both cells and vertices and for both genes and hyperedges). Thus, the count matrix of a single-cell RNA-sequencing experiment can be directly encoded as a hypergraph.

As discussed above, the graph spectrum has been generalised to hypergraphs with normalised coefficients. To satisfy the constraint **Eqn 4.3**, the expression level of gene g in cell i is taken to be the proportion of transcripts measured in cell i that are assigned to gene g, that is:

$$c(i,g) = m_i^g / M_i. \tag{4.8}$$

The coefficient $c(i,g)$ is an estimate of the relative level of gene expression in each cell. As discussed in **Section 1.2.4**, such cell-wise normalisation is required in many single-cell analyses, as the number of counts per cell can vary by orders of magnitude solely due to technical error (Dillies et al., 2013; Townes et al., 2019; Lause et al., 2020).

Note that each hyperedge, $e_g \in E$, encodes an analogous view of gene expression as used in **Chapters 2 & 3**, i.e. the view of a field of cells through a microscope. Each hyperedge encodes the expression of a gene over the set of cells, with the expression distribution of a given gene defined via the coefficients $c(i,g)$ for $i = 1, \dots, N$. The hypergraph view differs from that used in the developed information-theoretic framework with respect to normalisation: in **Chapter 2**, normalisation is carried out gene-wise, based on the total number of transcripts assigned to each gene; here, normalisation is carried out cell-wise, based on the total number of transcripts assigned to each cell.

Thus, hypergraphs provide a natural representation of the joint distribution of gene expression over the set of cells. Based on this hypergraph representation of gene expression, the hypergraph spectrum can be computed as described above. In Mulas and Casey (2021), RM derives a set of properties of the largest Laplacian eigenvalue. These properties make the largest eigenvalue an appropriate quantity for measuring gene expression heterogeneity. In the next section, **Section 4.3**, I will discuss these properties, omitting the mathematical proofs to make clear my own scientific contribution to this work. I will demonstrate the use of the largest Laplacian eigenvalue as a measure of gene expression heterogeneity and illustrate its utility on a range of single-cell RNA-sequencing data sets.

4.3 Quantifying Heterogeneity

By encoding single-cell expression data as a hypergraph, the spectrum associated with the data can be computed. The spectrum of a normalised hypergraph is the set of N real, non-negative eigenvalues of the Laplacian matrix of the hypergraph (Jost and Mulas, 2021; Mulas and Casey, 2021). These eigenvalues encode different structural properties of the hypergraph, and therefore, of the data (Oellermann and Schwenk, 1991; Jost and Mulas, 2021).

Specifically, Mulas and Casey (2021) show that the largest Laplacian eigenvalue λ_N is such that $1 \leq \lambda_N \leq N$ and, moreover,

- $\lambda_N = 1$ if and only if each edge has cardinality 1;
- $\lambda_N = N$ if and only if all edges have cardinality N and, for each edge k and for all vertices $i \neq j$,
$$c(i,k) = c(j,k).$$

Recall that the cardinality of a hyperedge g is the number of vertices with non-zero coefficients with respect to g. In the context of single-cell RNA-sequencing data, the cardinality of a gene is the number of cells that express that gene. Therefore, $\lambda_N = 1$ if and only if each gene is expressed in a single cell. In such a case, each cell in the population is expressing a different combination of genes, and all cells are absolutely distinguishable with respect to the joint distribution of gene expression. Thus, $\lambda_N = 1$ when the population is maximally heterogeneous with respect to the joint distribution of gene expression.

In contrast, $\lambda_N = N$ if and only if each gene is expressed in all cells to the same level, $c(i,k) = c(j,k)$ for $i \neq j$. In such a case, every cell is identical with respect to the joint distribution of gene expression, i.e. every cell is exactly interchangeable with respect to the gene expression. Thus, $\lambda_N = N$ in the absence of any heterogeneity with respect to the joint distribution of gene expression.

A normalised measure of gene expression heterogeneity, R, can be derived by dividing through by the number of vertices, N, so that

$$R = \frac{\lambda_N}{N}, \tag{4.9}$$

	Gene 1	Gene 2
Cell 1	0.3	0.7
Cell 2	0.3	0.7

	Gene 1	Gene 2
Cell 1	0.1	0.9
Cell 2	1	0

TABLE 4.1: **Toy examples of** R. Two examples of normalised single-cell sequencing data. In the first, the cells are interchangeable with respect to the normalised gene expression values, so $R = 1$. In the second, both genes are differentially expressed. Therefore, the cells are distinguishable with respect to gene expression, and $R = 0.55$ (note that the minimum value of R is 0.5 in this case as $N = 2$)

and where R is constrained to the range

$$\frac{1}{N} \leq R \leq 1. \tag{4.10}$$

Thus, with respect to the joint distribution of gene expression, as represented by the hypergraph Γ, $R = 1$ if and only if cells are interchangeable and $R = \frac{1}{N}$ if and only if cells are absolutely distinguishable: R quantifies heterogeneity with respect to the joint distribution of gene expression.

To illustrate, consider the toy examples in **Table 4.1**. In the first example, both cells are exactly interchangeable with respect to normalised gene expression, and $R = 1$. In the second, the cells are heterogeneous with respect to the expression of both genes, leading to a reduction in the value of R to 0.55.

The toy examples in **Table 4.1** illustrate the relationship of R and gene expression heterogeneity on a small scale. As discussed, the level of heterogeneity with respect to the joint distribution should be associated with the number of cell types in a population; in the next section, I will demonstrate this association empirically, showing that R can serve as a tool in inferring the number of cell types in a cellular population.

4.4 Estimating Cluster Number

I have previously introduced a range of data sets with various numbers of cell types present, see **Tables 2.1 & 3.4**. The cells of these data sets have been classified using a diversity of approaches, specifically by genotype, surface protein expression and by unsupervised clustering (Svensson et al., 2017; Tian et al., 2019; Zheng et al., 2017; Stumpf et al., 2020; The Tabula Sapiens Consortium and Quake, 2021). Importantly, the number of clusters, C, in each data set has been determined independently. Moreover, these data sets were generated by various labs, each using different sequencing protocols. As such,

Data Set	R	C
Svensson et al. (2017)	0.953	1
Tian et al. (2019) (a)	0.825	3
Zheng et al. (2017)	0.793	4
Tian et al. (2019) (b)	0.771	5
Stumpf et al. (2020)	0.398	14
Tabula Muris Consortium et al. (2018)	0.493	56

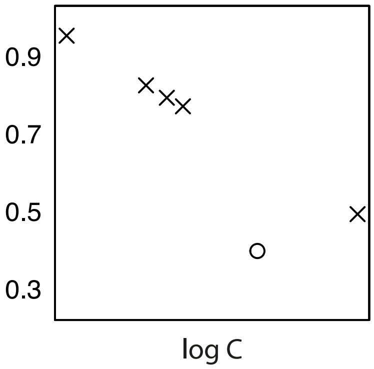


TABLE 4.2: *R* **of validation data sets**. Heterogeneity with respect to joint distribution of gene expression, *R*, and number of cell types, *C*, for range of data sets (*left*) detailed and (*right*) plotted, demonstrating a log-linear relationship between *C* and *R*. The outlying Stumpf et al. (2020) data set is highlighted in the plot, represented by a circle.

these data sets represent a diverse sampling of single-cell RNA-sequencing data sets for testing the relationship between R and C.

Unfortunately, the computation of the hypergraph spectrum is computationally demanding, with the matrix multiplication $D^{-1}A$ (see computation of the hypergraph Laplacian, **Eqn 4.7**) scaling quadratically in computational complexity with respect to the number of cells, N. This computational complexity prohibits the direct computation of R for the larger data sets from Zheng et al. (2017) and Tabula Muris Consortium et al. (2018) which each consist of tens of thousands of cells (see **Table 2.1**). Instead, I use a stratified sampling of 5,000 cells for both data sets, maintaining the relative number of cells assigned to each cell type.

Across the six data sets, R strongly negatively correlates with the number of cell types; see **Table 4.2**. Specifically, there is an apparent log-linear relationship between number of cell types and R, with a Pearson's correlation coefficient between $\log C$ and R of -0.886 (p-value $= 0.0186$, two-sided, 95% confidence interval of -0.988 to -0.267).

The Stumpf et al. (2020) data set is an outlier to this log-linear relationship, with substantially lower R than expected for the number of cell types present ($C = 14$), especially when compared to the Tabula Muris ($C = 56$). As discussed in **Section 2.5.2**, the Stumpf et al. (2020) data sets consists of cells actively transitioning in cellular identity from one type to another. Thus, there is substantial intra-type heterogeneity arising from biological functions other than the differential expression between cell types.

Restricting our analysis to those data sets where heterogeneity is predominantly attributable to differential gene expression between cell types, the strength of the correlation between R and $\log C$ increases to -0.999

(p-value $= 4.04 \cdot 10^{-7}$, two-sided, 95% confidence interval of -0.999 to -0.999). Thus, for biological systems where cellular identity is approximately static, R correlates near-exactly with the number of cell types.

R provides a measure of heterogeneity with respect to the joint distribution of gene expression, robust to the choice of cellular classification method (see **Table 2.1** for list of classification methods). In the absence of substantial alternative sources of gene expression heterogeneity, the relationship between R and cluster number is effectively exact. As such, the value of R measured in an unclassified test data set relative to classified data sets provides a powerful method for estimating the true number of clusters in a test population. In the next section, I will explore further the case where there are substantial alternative sources of gene expression heterogeneity, calculating R for each cell type in the Stumpf et al. (2020) population.

4.4.1 Intra-type Heterogeneity

Subsetting the hypergraph of the full population, I measure the value of R for each cell type in the Stumpf et al. (2020) data set. Note that unlike with the previously presented information-theoretic framework, there are no theoretical guarantees on the relation between the gene expression heterogeneity within each type and the total population heterogeneity, i.e. unlike $I(g)$, R is not additively decomposable. Thus, there is no prior theoretical expectation on the value of R for each cell type.

As seen in **Fig 4.1**, the value of R with respect to each cell type is remarkably consistent. However, there are two distinct sets of outliers, one group with lower R, the Pro-B, Pre-B and T-NK (Natural Killer) cell types, and another group with greater R, the Neutrophil and Myelocyte cell types.

Interestingly, these outliers form two coherent developmental clades: the Lymphocyte (Pro-B, Pre-B and T-NK cell types) and Neutrophil (Neutrophil and Myelocyte cell types) lineages, respectively; see **Fig 2.15** (Stumpf et al., 2020). This coherence suggests that the value of R is informative of some underlying biology of these lineages. For example, the cells of the Lymphocyte lineage are functional in the immune response (Alberts, 2017). Given the diversity of pathogens that the cells of the immune system respond to, immune cells possess substantial heterogeneity with respect to cellular function: this increased functional heterogeneity should be reflected in an increased level of gene expression heterogeneity (Satija and Shalek, 2014).

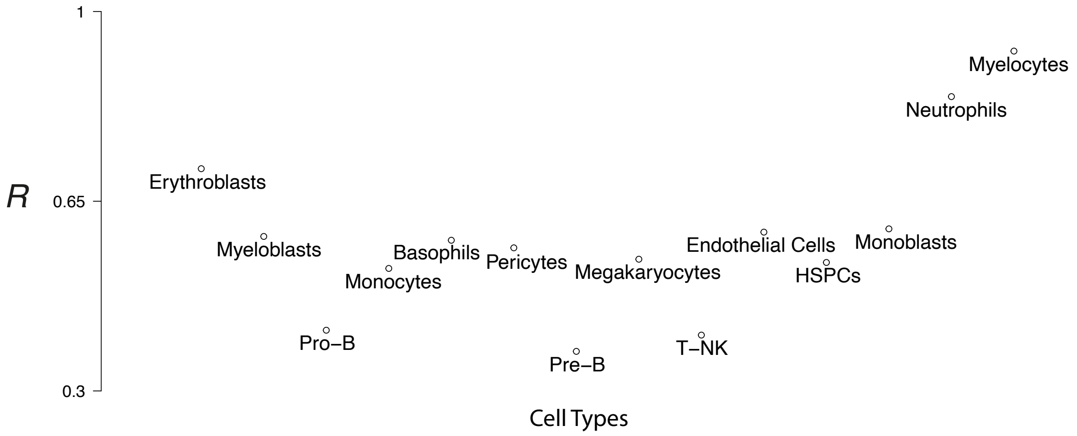


FIGURE 4.1: **Intra-type** R. R of each cell type in the Stumpf et al. (2020) data set. R largely consistent across types indicating similar levels of gene expression heterogeneity. The exceptions are Pre-B, Pro-B and T-NK cell types, which form a coherent developmental lineage, and Neutrophils and Myelocytes, which form another coherent developmental lineage.

However, given the diversity of biological functions of the remaining cell types, each with relatively consistent value of R, it is unlikely that this difference in R is solely biological. Instead, the differences in R could be an artefact of clustering – R may detect over-/under-clustering of cells.

As discussed, the cell types of the Stumpf et al. (2020) are not truly discrete; rather, they each capture a continuum of cellular differentiation and maturation. Thus, the demarcation of each cell type along a given lineage is partially arbitrary. For example, both the scEC and Louvain clusterings of the Stumpf et al. (2020) data set presented in **Section 3.1.4** merge the majority (78.6% and 98.1%, respectively) of cells classified as Neutrophils or Myelocytes in the established classification into a single cluster. This merging of cell types with respect to both unsupervised clustering methods suggests that these cells are over-clustered in the original classification. Such over-clustering would result in each cell type being less heterogeneous with respect to gene expression, as detected by the increase in R for the Neutrophil lineage cell types.

Conversely, low values of R, such as with the Lymphocyte lineage, may indicate under-clustering. However, unlike the cell types of the Neutrophil lineage, the individual cell types of the Lymphocyte lineage are recovered by both scEC and the Louvain method; see **Section 3.1.4**. Therefore, the cells may be appropriately classified, with the decreased values of R reflecting the increase in gene expression heterogeneity expected of immune cells.

4.5 Discussion

This chapter aimed to complete the optimisation problem posed in **Chapter 3**, developing a measure for inferring the number of cell types in a cellular population. I have developed such a measure through hypergraph theory, quantifying the heterogeneity of a cellular population with respect to the joint distribution of gene expression via the largest eigenvalue of the normalised hypergraph Laplacian. I have shown that the proposed measure of gene expression heterogeneity, R, strongly correlates with the number of discrete cell types present in a population, providing a simple way to infer the number of clusters in a novel data set based on a random sampling of previously classified data sets. Thus, comparing the relative value R offers one approach to completing the optimisation problem underlying scEC (and other unsupervised clustering methods).

R correlates strongly with the number of cell types in non-dynamic biological systems, where each cell can be reasonably assumed to be uniquely assigned to a single cell type, i.e. biological systems where the cell types are discrete. For example, when excluding the biological outlier of the Stumpf et al. (2020) data set, R and the (log) number of cell types correlate near-exactly, with a Pearson's correlation coefficient of -0.999. This correlation strength is particularly notable given the diversity of approaches used in classifying the cells of each data set, including both experimental and computational approaches.

However, it should be noted that this correlation is based on relatively few data sets. Moreover, the utilised data sets represent a poor sampling with respect to the number of cell types in each cellular population: there is only one non-outlier data set, the Tabula Muris, with > 5 cell types. Furthermore, the strength of the association between R and $\log C$ is suggestive of confounding technical effects. While I have made best efforts to ensure that there are no obvious technical confounders between the data sets, notably, all utilised data sets were generated via droplet-based sequencing methods (see **Appendix A** for an explanation of droplet-based sequencing). Therefore, moving forward with this work, I would expand the number of data sets involved in testing the association between R and the number of cell types, specifically including more data sets with a large number of cell types and data sets generated via plate-based sequencing methods (the primary alternative to droplet-based sequencing methods) (Papalexi and Satija, 2018).

Methods for inferring the number of cell types will be increasingly valuable as more and more single-cell data sets become available (Svensson et al., 2020).

While (semi-)supervised classification methods can be used if there is an available reference transcriptome, estimating the number of cell types in a population *de novo* is a time-consuming and often arbitrary task (Luecken and Theis, 2019). In the absence of a relevant reference transcriptome, R can be used to estimate the true number of clusters based on a sampled variety (with respect to number of cell types) of single-cell RNA-sequencing data sets.

Throughout this thesis, I have dealt with the classification of cells with respect to single-cell RNA-sequencing data. I have formalised the notion of gene expression heterogeneity in two mathematical languages: first in terms of information theory; then in terms of hypergraph theory. I have utilised these dual frameworks to determine the optimal clustering of cells and the optimal number of clusters. In the next chapter, I will discuss how these distinct frameworks can contribute to the concept and definition of cell type. Specifically, I will contrast the notion of cell type used throughout this thesis – i.e. that the cells of each type are interchangeable with respect to the measured expression of a set of genes – with the conception of cell type stemming from dynamical systems theory (as introduced in **Section 1.2.2**).

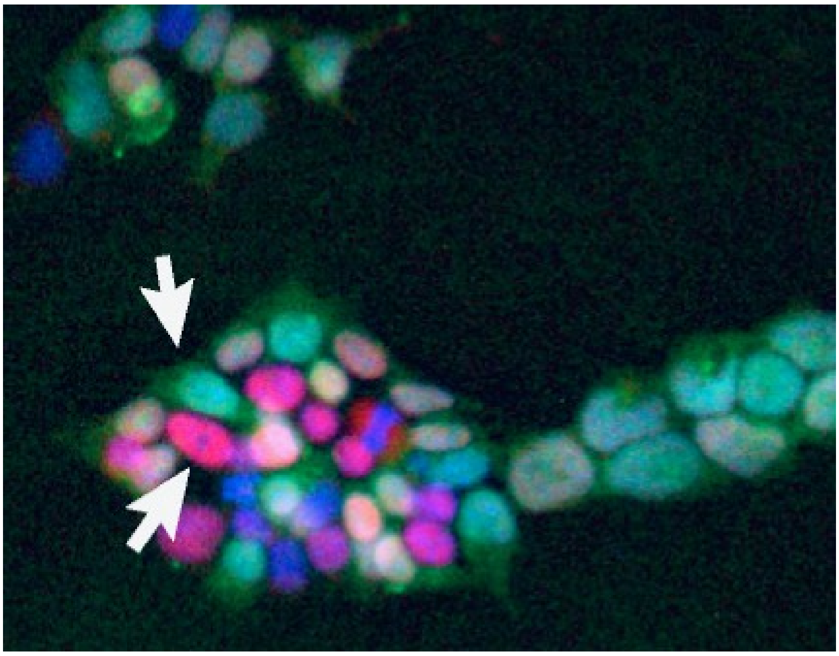
Chapter 5

Discussion

The technology of single-cell RNA-sequencing has revolutionised the classification of cells, but this revolution has precipitated an increase in conceptual complexity with respect to cellular classification and the notion of cell type (Moris et al., 2016; Clevers et al., 2017; Weinreb et al., 2018; Kiselev et al., 2019; Greulich et al., 2020). The traditional approach to classifying cells, phenotypic classification, is conceptually straightforward, identifying cells based on the presence or absence of observable features, such as gene expression (Mescher, 2018). However, single-cell sequencing technologies have revealed that cells of the same type are heterogeneous, inconsistently expressing established marker genes (Trapnell, 2015). Such intra-type heterogeneity prohibits the application of phenotypic classification to single-cell expression data at the level of individual cells.

Accommodating this heterogeneity, conceptually and practically, has become a central theme of single-cell analysis (Soneson and Robinson, 2018; Yip et al., 2019; Wang et al., 2019; Becht et al., 2019; Luecken and Theis, 2019). Instead of classifying cells individually, cells are first clustered into types, with each cluster identified based on the differential expression of marker genes relative to the remaining clusters in the population. The process of cluster-wise classification assumes that unsupervised clustering methods are able to group cells into types without specific reference to established marker genes of the phenotypic classification scheme, an assumption justified by dynamical systems theory (Greulich et al., 2020).

Dynamical systems theory is the field of mathematics concerned with systems that evolve in time (Strogatz, 2018). Of relevance to the classification of cells, dynamical systems theory predicts that cellular gene expression profiles will evolve towards certain subsets of gene expression space with time (see **Section**



et

FIGURE 5.1: **Example of a microscopy experiment**. Image from microscopy experiment observing gene expression in a population of cells. Cells are immediately classifiable based on the expression of *Nanog*, a key marker gene for pluripotent stem cells, with the two white arrows highlighting examples of cells with high/low *Nanog* expression. Image reproduced from Smith et al. (2017).

1.2.2). These subsets of the space of possible gene expression profiles are termed attractors (Strogatz, 2018). Cells of a given type will evolve towards the same attractor in time, resulting in cells of the same type being similar with respect to gene expression. Unsupervised clustering methods detect these groupings of similar cells and cluster accordingly.

This thesis develops a distinct motivation for the application of unsupervised clustering based on the principles of phenotypic classification. Recall that the process of phenotypic classification resembles a microscopy experiment. Imagine looking down a microscope, observing the expression of a given gene with respect to each cell. The cells can be sorted accordingly, classified into different types based on the observed expression of different genes (see **Fig 5.1** for an example of a microscopy image). Importantly, this view corresponds to an empirically intuitive definition of cell type: cells of the same type are (at least approximately) interchangeable with respect to gene expression.

The empirical conception of cell type does not accommodate intra-type heterogeneity with respect to gene expression: when genes are expressed inconsistently, the cells identified with each type cannot be treated as interchangeable. **Chapter 2** reconciled the gene expression heterogeneity observed in single-cell sequencing data with the empirical conception of cell type by developing a formal, information-theoretic framework for quantifying gene expression heterogeneity. One measure in this framework – h_S, intra-cluster heterogeneity – quantifies the divergence of a given classification of cells from the assumption that cells of the same type are interchangeable with respect to genome-wide gene expression. Thus, h_S quantifies how well a given classification holds to the traditional, empirical conception of cell type.

Through h_S, the cellular classification or clustering of a given population that best approximates the empirical conception of cell type can be identified. **Chapter 3** developed an algorithm for the minimisation of h_S through maximising another element of the information-theoretic framework, H_S. H_S quantifies the heterogeneity arising from differential expression between cell types, and by additive decomposition, the clustering of cells that maximises H_S minimises h_S. Thus, the developed algorithm, scEC, identifies the clustering that is minimally divergent from the fundamental assumption of phenotypic classification, with clusters that are maximally differentially expressed.

The scEC algorithm clusters cells into types based on the expression of each gene in the genome. Thus, scEC represents a high-throughput, automatic implementation of the principles of phenotypic classification, extended to classify cells with respect to the expression of every gene. Such an extension is only possible due to the quantitative nature of the developed information-theoretic framework, weighing up the potentially conflicting clusterings of cells preferred by each gene in the genome.

Importantly, the information-theoretic approach represents a distinct conception of cell type to that of the attractor. The information-theoretic framework encodes an empirically intuitive conception of cell type, drawing from the principles of phenotypic classification. In contrast, the attractor conception of cell type is inherently theoretical, drawing from the theory of dynamical systems. However, the differing conceptions of cell type are not unrelated. As I will outline below, the information-theoretic conception of cell type is a static approximation of the dynamic cell type represented by attractors.

As discussed, attractors emerge from the complex regulatory interactions of different genes, confining the space of possible gene expression profiles for each cell. Importantly, however, these attractors do not necessarily consist of a single point in gene expression space. Instead, attractors can consist of a contiguous series of gene expression profiles, forming non-trivial shapes in gene expression space; see **Fig 5.2** for example illustrations (Weinreb et al., 2018; Greulich et al., 2020). For these attractors to be stable over time, they have to be periodic structures, i.e. over time, cells repeatedly oscillate in

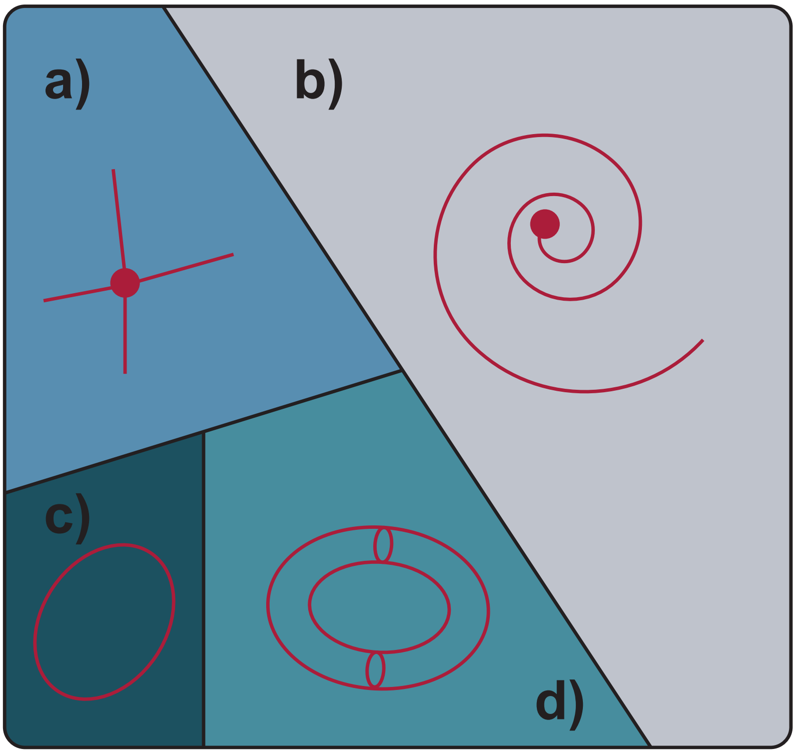


FIGURE 5.2: **The cell as a dynamical system**. The expression of each gene is regulated by the expression of other genes, collectively giving rise to a complex gene regulatory network. This network encodes a complex dynamical system that may admit numerous attractors, each corresponding to a distinct cell type. These attractors partition gene expression space into discrete regions, termed basins of attraction. The gene regulatory network of a cell may admit many different kinds of attractor including various different kinds of fixed-point such as **a)** stable nodes and **b)** stable spirals, as well as **c)** limit cycles, and **d** more topologically complex structures such as limit tori. Figure reproduced with modification from Casey et al. (2020b).

expression through the constituent profiles of the attractors. Thus, these complex attractors define each cell type not with respect to a singular gene expression profile but with respect to a collection of different gene expression profiles (Casey et al., 2020b).

One example of such oscillatory behaviour is the cell cycle (Kruse and Jülicher, 2005). A cell of a given type can progress through the cell cycle while maintaining a single type identity: the cell type is defined with respect to the gene expression profiles at each cell cycle stage (Weinreb et al., 2018). More generally, an attractor can involve multiple overlapping oscillations, resulting in cell types being defined with respect to remarkably complex multivariate gene expression patterns, such as the limit tori in **Fig 5.2d**.

Indeed, such complex attractors are expected to be the norm with respect to cellular transcriptomes: single-state, so-called fixed-point attractors only

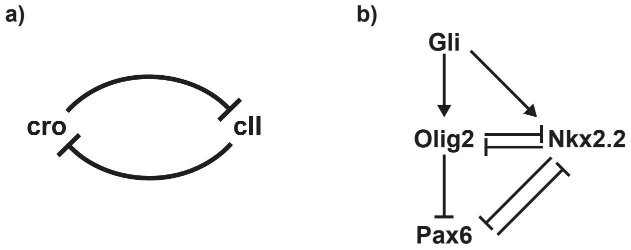


FIGURE 5.3: **Gene regulatory networks**. Diagrammatic representations of an **a)** symmetric network and **b)** and asymmetric network (Ptashne, 2004; Balaskas et al., 2012).

emerge when the defining gene regulatory network is symmetric (that is all regulatory interactions are bidirectional), such as in the lambda phage network, see **Fig 5.3a** (Ptashne, 2004). However, symmetry is a strong constraint to place on a network, and not one that will be generally met for more complex gene regulatory networks, e.g. the Sonic hedgehog network, see **Fig 5.3b** (Balaskas et al., 2012). As such, oscillatory cell types defined by complex attractors are predicted to be the norm (Weinreb et al., 2018).

Accordingly, at a given moment in time, cells may be distributed at different points along an attractor. Thus, at any specific time-point, the set of cells associated with a given attractor, i.e. cells of the same type, will be heterogeneous with respect to gene expression. However, these cells are not truly biologically heterogeneous – over time, the behaviour of these cells is interchangeable. Thus, the population is only heterogeneous when observed at a single time point.

Therefore, instead of representing opposing conceptions of cell type, the developed information-theoretic framework and dynamical systems theory both realise the same conception of cell type, considered over different time scales. The information-theoretic realisation is strict, expecting the cells of each type to be interchangeable at any given point in time; the attractor realisation expects cells of the same type to be interchangeable over more extended periods of time. Thus, the information-theoretic conception of cell type is a static approximation of the dynamic cell type formalised by attractors.

This simplifying approximation represents a trade-off between power and interpretability. Single-cell RNA-sequencing captures a static snapshot of the transcriptomes of a cellular population (Trapnell, 2015). In this static snapshot, the oscillatory dynamics underpinning each cell type are frozen, resulting in substantial intra-type heterogeneity with respect to gene expression (Weinreb et al., 2018; Casey et al., 2020b). Dynamical systems theory predicts that this gene expression heterogeneity is structured, corresponding to the complex shapes formed by attractors in gene expression shape (see again **Fig 5.2**).

The Louvain method – the best performing single-cell unsupervised clustering method for single-cell RNA-sequencing data – identifies these complex cluster structures, accommodating the associated intra-type heterogeneity with respect to gene expression (Blondel et al., 2008; Luecken and Theis, 2019). The Louvain method utilises a k-nearest-neighbours representation of the data, so cells need not be similar to all other cells in their assigned cluster, but only to some (Newman and Girvan, 2004; Blondel et al., 2008; Kiselev et al., 2019). Indeed, the clustering method is able to identify clusters of arbitrary shape, with the shape of each cluster limited only to being contiguous (unlike the k-means clustering method, wherein clusters tend to be hyperspherical) (Lloyd, 1982; Kiselev et al., 2019). Therefore, dynamical systems theory justifies both the success of traditional unsupervised clustering generally and the Louvain method in particular with respect to identifying cell types.

By contrast, scEC assumes that all intra-type heterogeneity with respect to gene expression needs to be minimised. The scEC method is, therefore, less accommodating and so less powerful than the Louvain method in identifying complex cell types. In theory, the scEC method will perform less well when cell types are defined with respect to complex attractors, consisting of numerous, biologically important cell states or subtypes.

However, dynamical systems theory does not necessarily hold in practice. As discussed in **Chapter 1**, the concept of attractors assumes that the system is deterministic. When a system is stochastic, the correspondence of cell types to attractors becomes more complex: cell types instead correspond to regions of high probability density in the joint distribution of gene expression (Greulich et al., 2020).

Unlike deterministic attractors, these regions of high probability density do not necessarily imply that cells are interchangeable over time. To make that inference, i.e. to identify periodic oscillations in gene expression, the expression of a given set of cells would have to be followed through time. Without such temporal information, whether the clusters identified by the Louvain method are indeed interchangeable over time cannot be confirmed. In contrast, the scEC framework does not require temporal information: scEC assumes that cells are interchangeable within each type at the measured time-point. Thus, scEC clusters are substantially simpler to interpret concerning the static snapshot of the transcriptome produced by single-cell RNA-sequencing.

In future, the temporal information required to test the assumption that cells are equivalent over time may become accessible; for example, a new single-cell RNA-sequencing technology, Live-seq, is reportedly able to sequence cells at multiple time points (the technology has not been published at the time of writing, see Chen et al. (2021) for preprinted manuscript). Alternatively, it may be possible to identify the specific oscillatory structures predicted by dynamical systems theory at a single time point: a burgeoning set of tools, collectively termed Topological Data Analysis, analyse the shape of data. These tools, in particular that of persistent homology which quantifies the number of 'holes' or 'cycles' in the shape of the data, could allow the gene expression patterns associated with complex attractors to be identified from the snapshot of gene expression space provided by single-cell RNA-sequencing (Zomorodian, 2005; Carlsson, 2009; Rizvi et al., 2017; Rabadan and Blumberg, 2019).

Nevertheless, the information-theoretic framework developed in **Chapters 2** & **3** provides a formalised approach to clustering based in the empirically intuitive conception of cell type utilised in phenotypic classification. The comparable performance of the developed clustering method scEC to the state-of-the-art Louvain method suggests that the relatively simple conception of cell type, namely that cells of the same type are approximately interchangeable, is sufficient to accommodate the gene expression heterogeneity observed in single-cell sequencing data.

Information Theory for Single-cell Analysis

The central contribution of this work is conceptual, formalising the phenotypic classification of cells in the language of information theory. However, the practical undertaking of clustering single-cell RNA-sequencing data is not trivial, requiring substantial pre-processing and *post-hoc* analysis of clusters (Luecken and Theis, 2019). Alongside the unsupervised clustering algorithm, scEC, the individual measures of the information-theoretic framework offer practical alternatives and supplements to existing elements of single-cell clustering analysis. Specifically, $I(g)$ and $H_S(g)$ readily provide a basis for feature selection and differential expression analysis, respectively (see **Sections 1.2.4** & **1.3** for discussions on feature selection and differential expression analysis, respectively).

The goal of feature selection is to identify, prior to clustering, those genes that are likely to be differentially expressed between cell types (Yip et al., 2019).

Differential gene expression increases the heterogeneity of cells of different types with respect to gene expression. $H_S(g)$ quantifies the heterogeneity attributable to differential gene expression as an additive component of $I(g)$. Thus, $I(g)$ represents the maximum possible differential expression: high values of $I(g)$ directly identify those genes more likely to be differentially expressed between cell types.

Self-evidently, differential gene expression analysis aims to identify those genes that are differentially expressed between clusters (Soneson and Robinson, 2018; Wang et al., 2019). $H_S(g)$ quantifies the heterogeneity attributable to differential expression between all clusters, so significant values of $H_S(g)$ (identified through exact testing, see **Section 2.5.3**) identify those genes that are differentially expressed between clusters in a population.

As discussed in **Section 1.3**, clusters are classified based on which genes are identified as differentially expressed with respect to each individual cluster. Typical differential expression analyses therefore identify genes that are significantly up or down-regulated in the cells of a specific cluster relative to the rest of the cellular population (though there are exceptions; for example, Rackham et al. (2016) calculates a single score for the differential expression of a given gene across multiple cell types). However, when used in conjunction with unsupervised clustering, differential gene expression has a highly-inflated false positive rate, falsely identifying many genes as differentially expressed (Luecken and Theis, 2019; Gao et al., 2020). This false identification can lead to the misclassification of clusters and the generation of flawed hypotheses with respect to the biological function of specific genes (Squair et al., 2021).

In future, $H_S(g)$ could be used in conjunction with traditional differential gene expression testing, first identifying which genes are differentially expressed and then identifying the clusters up or down-regulated with respect to the expression of each gene. Such a composite two-step process to differential expression would improve the statistical robustness of differential expression analysis, limiting the false identification of genes as being differentially expressed.

The information-theoretic framework developed in **Chapter 2** offers a cohesive mathematical framework for carrying out single-cell analysis in the language of gene expression heterogeneity. However, as an analytic framework, it is incomplete: the framework does not provide a visualisation of the distribution of transcripts on the set of cells for large numbers of genes; the framework provides no estimation of the true number of clusters in a data set; and, the framework (as implemented) is discrete, without accommodation for quasi-continuous cellular identities, such as those found during cellular differentiation (recall that the returned fuzzy membership values are discretised in the scEC method). I will address the third point later (see *"An Information-Theoretic Approach to Cellular Differentiation"*) as a matter for future work, but first, I will discuss the domain-specific implications of a lack of visualisation.

By visualisation, I refer to the various dimension reduction projections, both linear and non-linear, that project the high-dimensional gene expression profiles of individual cells onto two or three dimensions (Hotelling, 1933; Hinton and Roweis, 2003; McInnes et al., 2018; Becht et al., 2019). These reductions view cells as position vectors in some abstract gene expression space; in contrast, in the information-theoretic framework, an alternative view of gene expression is adopted, distributing transcripts onto the set of cells. This view lacks any notion of distance between cells. Thus, the results of scEC have to be projected and visualised using existing visualisation tools, such as UMAP (McInnes et al., 2018).

The lack of visualisation is problematic as biology is an inherently visual field: biologists require 'proof by visualisation' (Fox Keller, 2002). For example, while a given clustering may be formally validated by differential expression analysis, different clusterings are informally analysed via visualisation (Luecken and Theis, 2019; Becht et al., 2019; Chari et al., 2021). While such visualisation often distorts the data, visualising results is an inherent part of the culture of biological research (see Chari et al. (2021) for a discussion of the distorting effects of non-linear dimension reduction and specifically UMAP on single-cell expression data). Indeed, the motivation for scEC, phenotypic classification, is a fundamentally visual approach to classification, as illustrated in **Fig 5.1**. The lack of an inherent tool for visualisation limits the biological interpretability of the developed information-theoretic framework.

The problem of estimating the true number of clusters in a data set is not limited to scEC, being a general problem for unsupervised clustering methods. This thesis has developed two methods for inferring the number of cell types. Firstly, in **Section 3.2**, I developed a semi-supervised version of scEC to classify cells based on an established reference transcriptome: in semi-supervised classification, the number of cell types is derived from the reference. Secondly, in the absence of a direct reference, in **Chapter 4**, I developed a hypergraph theory method for inferring the true number of clusters in a data set based on the relative level of gene expression heterogeneity in a variety of independent single-cell sequencing data sets.

I have validated both methods for cluster number estimation; however, both methods would benefit from testing on additional data sets. For the semi-supervised classification algorithm, there are a limited number of test-reference data set pairs where the classification of both data sets are known, so validation is more difficult than for unsupervised clustering methods. In contrast, the hypergraph theory approach depends on collections of biologically unrelated data sets, so is more easily tested than the semi-supervised method. Recall that the relationship between cluster number and the developed measure, R, was extremely strong for the trialled data sets, with the notable biological outlier of the Stumpf et al. (2020) data set. In future, by adding further test data sets, I aim to validate the strength of this relationship, and determine the type of cellular populations for which R is an informative measure of cluster number (e.g. dynamic versus static cell types).

An Information-Theoretic Approach to Cellular Differentiation

The clustering methods I developed in **Chapter 3** rely on a generalisation of the information-theoretic framework introduced in **Chapter 2** to the case of fuzzy cellular identities. I introduced fuzzy cellular identities primarily to enable the efficient clustering of thousands of cells. However, as shown with respect to the scEC clustering of the Stumpf et al. (2020) population, fuzzy cluster identities may be of biological relevance (see **Section 3.1.4**). I have assumed throughout this thesis that all cells will occupy a discrete cellular identity, i.e. that each cell should be assigned to only a single cell type. Moreover, even with fuzzy identities, discrete cellular identities are optimal with respect to maximising H_S for most cell types trialled. However, cells do not always occupy a discrete identity – during cellular differentiation, cells change between discrete cell types. Fuzziness offers a mechanism for formally encoding differentiation through fuzzy mixtures of intermediate cellular identities as cells transition from one discrete cell type to another.

Viewing differentiating cells as identifying with a fuzzy set of different cell types, the dynamics of cellular differentiation can be examined with respect to the developed information-theoretic quantities. Specifically, H_S can be used as a measure of the favourability for a given (fuzzy) set of cellular identities. Assuming that the partition of cells into types maximises inter-cluster heterogeneity (as I assumed for the scEC unsupervised clustering method), the favourability of a cell differentiating from one type to another can be assessed.

For example, consider a population of five cells with respect to the expression of a single gene, g. Let four cells have discrete identities, with two belonging to a high-expression cell type, which I term type A, and two to a low-expression cell type, type B; let the final cell have a variable fuzzy membership with respect to each type.

As shown in **Fig 5.4**, if the cell has a high expression of g, matching the cells of type A, differentiation of the cell from type A to type B is unfavourable with respect to H_S. Conversely, if the cell has a low expression of g, matching the cells of type B, differentiation from A to B becomes favourable. Thus, the favourability of differentiation is determined by the similarity of the gene expression profile of the differentiating cell to the gene expression profiles of the cells of the starting and terminal cell types.

When the cell has an intermediate expression of g, the favourability of differentiation with respect to H_S becomes more complex. In the example shown in **Fig 5.4**, the difference in H_S between $\mu = 0$ and $\mu = 1$ means that the differentiation from type A to B is favourable. However, the process of differentiation itself is unfavourable, where the intermediate identity of the cell, a fuzzy mixture of both type A and type B identities is associated with a lower value of H_S than either end state. Thus, when a cell has a gene expression profile intermediate between two cell types, either identity can be stably adopted – the identity of the cell is *bistable* with respect to H_S (Ferrell Jr, 2012).

Expanding on this toy example, I suggest an informal model of cellular differentiation. Consider a cell of some initial type that is similar in gene expression to other cells of the type. The cell becomes primed for differentiation into some terminal type through random fluctuations in gene expression or some active process, gaining an expression pattern intermediate between the two types. Priming makes available a route for differentiation, albeit only through an unfavourable intermediate. At the unfavourable intermediate, transitioning to either cellular identity, the initial or the terminal, is approximately equally favourable with respect to H_S – the cell could return to identifying as the initial type. Post transition, it is favourable, with respect to H_S, for the cell newly of the terminal type to increasingly resemble the other cells of the terminal type, stabilising the new identity of the cell.

The prediction of an unstable intermediate qualitatively resembles the transition state model proposed in Moris et al. (2016), itself inspired by the transition states of chemical reaction dynamics. In this model, each cell type is defined by a collection of gene expression profiles. A differentiating cell

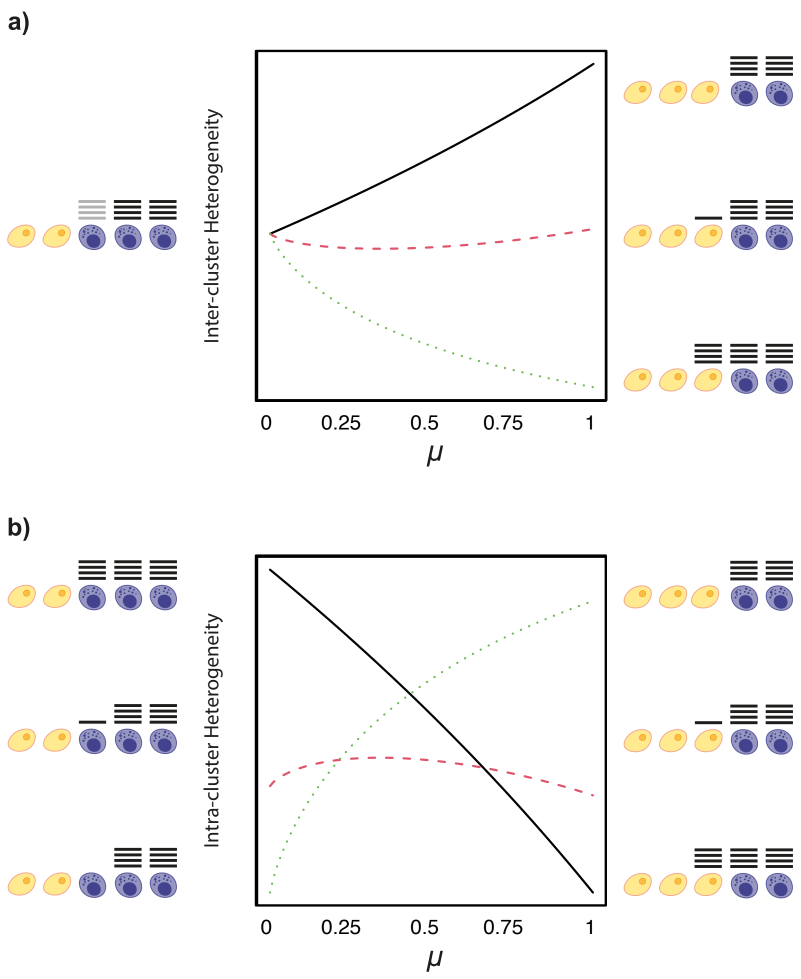


FIGURE 5.4: **Inter and intra-cluster heterogeneity during cellular differentiation**. Population consisting of five cells, where cells are assigned to one (or both) of two cell types, A and B. Two cells of type A (purple cells) highly express the gene g, and two cells of type B (yellow cells) do not express g. The expression and cell type membership of the final (middle) cell are variable. The variable cell can express g at a high level (solid black line), at an intermediate level (red dashed line), or have no expression of g (green dotted line). The cell's identity varies, being fully of type A at $\mu = 0$ and fully of type B at $\mu = 1$. Intermediate values of μ indicate a fuzzy mixture of identities, e.g. $\mu = 0.6$ means that the cell has a membership strength of type A of 0.4 and a membership strength of type B of 0.6. The **a)** inter-cluster and **b)** intra-cluster heterogeneity of the population changes with both cellular identity and expression (note that the population heterogeneity, $I(g)$, differs between the different expression levels of the variable cell). The favourability of the variable cell's identity depends on its expression, with $\mu = 0$ being favoured when g is highly expressed and $\mu = 1$ when g is lowly expressed. Either identity is stable when the cell has an intermediate level of expression, with the transition between identities involving decreasing inter-cluster heterogeneity and increasing intra-cluster heterogeneity.

adopts a range of possible ‘transition state’ gene expression profiles, intermediate between the expression profiles of two different cell types. The available transition states are unstable, eventually collapsing to give rise, with some probability, to either the new or original cell type.

Notably, the toy model in **Fig 5.4** that gives rise to these qualitative dynamics is defined with respect to the expression of only a single gene. Concerning H_S, the favourability of a given gene expression profile is not determined by the gene regulatory network but by the relative expression level of a given cell with respect to the other cells in the population. Of course, differentiation is driven by the dynamics of gene regulatory networks; nevertheless, H_S provides a simple, heuristic approach to considering the favourability of cell type transitions (Waddington et al., 1939; Britten and Davidson, 1969; Alberts, 2017; Greulich et al., 2020).

Such a simplification is useful, as gene regulatory networks are challenging to impute quantitatively. Moreover, the dynamical systems theory measure of favourability, potential, cannot be realistically computed for high-dimensional systems (Wang et al., 2010). Thus, in future, H_S could provide a practical, heuristic measure of favourability with respect to cell type transitions, building on the use of H_S as an objective function for unsupervised and semi-supervised clustering.

To provide a formal theory of cellular differentiation and not simply a heuristic measure of favourability, H_S would have to be motivated from physical principles. Importantly, information theory has deep links with statistical mechanics and thermodynamic entropy, and statistical mechanics has been repeatedly proposed as a physical basis for understanding cellular differentiation (Jaynes, 1957; Guillemin and Stumpf, 2020; Teschendorff and Feinberg, 2021). In the future, it would be interesting to investigate the relation between inter-cluster heterogeneity and thermodynamic entropy to develop a formal, physical theory of cellular differentiation based on H_S. In doing so, the information-theoretic definition of cell type developed here would emerge as a property of the underlying thermodynamics of gene expression.

Appendix A

Single-cell RNA-sequencing

Single-cell RNA-sequencing

The experimental process of single-cell RNA-sequencing consists of four principle steps: dissociation, isolation, library construction and sequencing (Luecken and Theis, 2019). In this appendix, we discuss each of these steps in turn, briefly surveying some of the choices available in a single-cell RNA-sequencing experiment.

Dissociation

The input for single-cell RNA-sequencing is typically a sample of biological tissue, but for single-cell sequencing, we require a suspension of dissociated single cells. Therefore single-cell RNA-sequencing begins with the digestion of the tissue sample (Luecken and Theis, 2019).

Isolation

Following dissociation, each cell must be separately isolated. How cells are isolated forms the major split in single-cell technologies, with two main types: plate-based and droplet-based (Papalexi and Satija, 2018). Plate-based methods (e.g. SMART-seq v1-v3) separate cells into wells by micro-pipetting, microfluidics or FACS sorting (Svensson et al., 2017; Ramsköld et al., 2012; Picelli et al., 2013; Hagemann-Jensen et al., 2020). Droplet-based methods (e.g. Drop-seq or inDrop) capture cells directly in microfluidic droplets (Macosko et al., 2015; Klein et al., 2015).

Droplet-based approaches offer massive parallelization, able to isolate ten to a thousand times more cells than plate-based approaches; in contrast, plate-approaches have much greater sequencing depth, with more transcripts measured (10–20% of transcripts measured versus 3–10% for droplet-based methods), and fewer restrictions on the size and type of cells that can be isolated (Ziegenhain et al., 2017; Papalexi and Satija, 2018). The preferred approach depends on what is required: deep sequencing of a few cells or an unbiased screen of as many cells as possible. Note that with both types of platforms, multiple cells can be captured in the same well or droplet, an error known as doublets or multiplets (McGinnis et al., 2019).

Library Construction

Within each well or droplet, cellular membranes are broken down, and intracellular mRNA is released. The released mRNA is reversed transcribed into cDNA (complementary DNA) using a poly-T primer (the majority of single-cell RNA-sequencing technologies only measure polyadenylated mRNA). The primer contains up to three additional sequences: a cellular barcode, a Unique Molecular Identifier (UMI) and a priming sequencing for amplification (Ziegenhain et al., 2017). The cellular barcode is unique to each plate or droplet, mapping each RNA to its cell of origin (except in the case of doublets or multiplets, where multiple cells will share a single barcode).

The UMI is unique to every mRNA molecule: the UMI preserves the identity of the mRNA molecule through the cDNA amplification required for sequencing; UMIs provide a direct measure of the number of transcripts in a cell and eliminates most bias introduced by preferential amplification (Kivioja et al., 2012; Islam et al., 2014; Svensson et al., 2017).

Not all technologies use UMIs: the use of UMIs extends the length of nucleic acid to be sequenced, restricting sequencing to the $3'$ end of the mRNA molecule, preventing full-length sequencing as in SMART-seq v1-v2 (Ziegenhain et al., 2017). A single gene can produce multiple types of mRNA through alternative splicing of exons: full-length sequencing is required to distinguish between different splice isoforms of a given gene (Wilkinson et al., 2020). The third iteration of SMART-seq enables both UMIs and sequencing of internal segments of the mRNA for mapping of splice isoforms (Hagemann-Jensen et al., 2020).

Sequencing

High-throughput short-read sequencing requires many copies of cDNA, requiring amplification of the sequence mediated through the priming sequence (Goodwin et al., 2016). The sequencing libraries from each well/droplet are pooled together for a single sequencing run, a technique known as multiplexing. Each cell is demultiplexed post-sequencing based on the cellular barcode, with sequencing reads mapped to the genome of the relevant species. If UMIs were used, mapped transcript reads are collapsed down into digital transcript counts. The final result of the sequencing process is a cell by gene count (or read) matrix where each element encodes the number of transcripts (or reads) of each gene measured in each cell.

In the main text of this thesis we solely consider data sets produced by droplet-based sequencing with UMIs. This combination has emerged as the most popular approach to sequencing: UMIs minimise a source of noise – variation in amplification efficacy – while losing information on splicing that is only of interest in specific studies. Moreover, UMIs directly represent the number of transcript molecules in a cell, rather than a first-order approximation: models of transcript counts represent direct, physical models of gene expression.

Droplet-based approaches allow for orders of magnitude more cells to be sequenced in a single experiment, increasing the chances of sampling rarer cell types, which are more likely to have been missed in low-throughput taxonomic identification. Many plate-based experiments would have to be run in parallel to achieve similar amounts, introducing the potential for batch effects between different plates. Droplet-based UMI data provides the greatest number of cells with the least noise, albeit at relatively shallow sequencing depths per cell.

Appendix B

Further Methods for Single-cell Analysis

In this appendix, we detail the methods involved in clustering analysis that are peripheral to unsupervised clustering itself. These methods are vital to the process of cellular classification, but are largely technical in detail, so are included here.

In the first section of this appendix, we discuss the methods used to prepare single-cell sequencing count data for unsupervised clustering. In the second section, we discuss how count data, and the results of clustering, are visualised through non-linear dimension reduction.

Clustering Pre-processing

The goal of unsupervised clustering with respect to dingle-cell RNA-sequencing data is the grouping of cells into type. However, this task is not trivial, and is particularly prone to error when there are substantive sources of heterogeneity other than differential gene expression between clusters.

The goal of pre-processing is to minimise the effect of these alternative sourced of heterogeneity with respect to gene expression. In this section, we discuss the various methods involved in pre-processing. We begin by discussing normalisation, which minimises the effect of technical noise. We then discuss feature selection, which selects *a priori* those genes most likely to be differentially expressed. Finally, we discuss linear dimension reduction, which transforms the data to emphasise the coordinated aspects of heterogeneity.

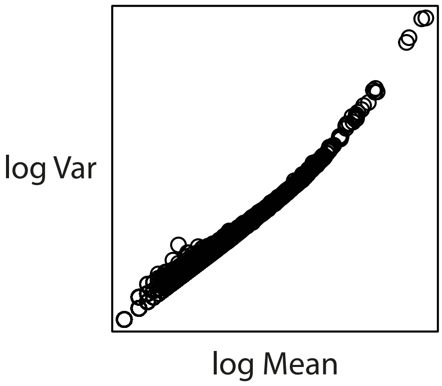


FIGURE B.1: **Mean-Variance Relationship of Count Data.** The log mean and log variance of gene transcript counts from data in (Svensson et al., 2017). Variance increases quadratically with increasing mean expression. To use variance as a measure of heterogeneity, it must be normalised by mean expression, otherwise only highly expressed genes will be selected.

Normalisation

Normalisation is the first step once the final gene expression count matrix has been determined. The purpose of normalisation is twofold: 1) to minimise any noise introduced by the sequencing experiment itself, thus isolating the biological component of gene expression heterogeneity, and 2) to transform the data so that it is more amenable to subsequent data analytic processes, specifically by stabilising the variance of the gene expression distribution of each gene (Vallejos et al., 2017; Hafemeister and Satija, 2019; Ahlmann-Eltze and Huber, 2020).

Single-cell sequencing is a count process, counting the number of transcripts in each cell. For a transcript to be counted, it must first bind to a barcode, then be successfully amplified (see **Appendix A**). Failure at either stage leads to the transcript being 'missed'. The probability of a given transcript being measured is low, with the total fraction of transcripts successively measured estimated at 3 to 10% (Papalexi and Satija, 2018).

This low probability of success has numerous knock-on effects. Firstly, the count matrix is sparse. Many genes will have few transcripts present in any given cell, whereas others will have orders of magnitude more transcript molecules. When the number of transcripts measured is substantially lower than the number of transcripts present in the cell, highly expressed genes can crowd out lowly expressed genes. This leads to many zero values in the count matrix, a phenomenon known as sparsity. Secondly, the total number of transcripts sequenced in a given cell (the count depth) can vary greatly. Variation in count depth can be biological, as cells will vary in how many transcripts they have, e.g. if they actively going through the cell cycle.

However, the effect is predominantly technical, with a variable number of missed transcripts per cell. The result is substantial variation in the count depth of measured cells (Dillies et al., 2013).

Various methods have been suggested to deal with both sparsity and count depth variation. Initially, normalisation methods were imported from analysis of bulk RNA-sequencing, an older technology where cell were not isolated for sequencing. In bulk-like normalisation, the expression vector of each cell is normalised by the count depth of the cell (or some other estimate of the 'size' of the cell) (Dillies et al., 2013). The normalised count values are then log-transformed to remove the observed dependence of the variance on the mean, see **Fig B.1** (Vallejos et al., 2017; Lause et al., 2020; Ahlmann-Eltze and Huber, 2020).

However, the extreme sparsity of single-cell sequencing data undermines these traditional approaches to RNA-sequencing noise: the size effect factors cannot be stably estimated on severely sparse data. Moreover, zero values cannot be log-transformed, so a pseudo-count is added to every gene-cell element. This $\log(x+1)$ transformation of the data fails to stabilise the variance of lowly expressed genes, as well as introducing the false appearance of zero-inflation from the pseudo-count(Lun et al., 2016; Lun, 2018; Townes et al., 2019; Sparta et al., 2021). These difficulties have required the development of normalisation method specific to single-cell data, not imported from the analysis of bulk data.

Statistical modelling approaches have been developed to deal with the problem of sparsity which seek to leverage distributional models of the data to isolate that part of heterogeneity that is due to differential expression. The goal of these modelling approaches is to fit some error model to the distribution of transcript counts. The error model estimates the amount of variance we expect to see from purely technical effects. This excess variance can then be regressed, leaving only the heterogeneity due to biological effects.

The most common model used is the negative binomial distribution. The negative binomial is an extension of the Poisson distribution, the archetypal distribution for count processes (Haight, 1967). The Poisson distribution models the number of counts expected from a process in a set time/space, e.g. the cell, and is parameterised solely by the distribution mean, λ. The Poisson distribution is statistically elegant, with the variance of the distribution equalling the mean, mathematically $\sigma^2 = \mu$.

However, while statistically simple and archetypal for count processes, the Poisson distribution is an inadequate model of single-cell RNA-sequencing data (Brennecke et al., 2013; Grün et al., 2014; Love et al., 2014; Townes et al., 2019; Cameron and Trivedi, 2013). Single-cell sequencing counts are overdispersed, having an excess of variance compared to that predicted by the Poisson, where variance is fixed by the mean. This excess variance is thought to arise from a range of sources, including from biological variability between samples/cells and from both biological and technical noise (Love et al., 2014; Grün et al., 2014; Brennecke et al., 2013; Raj et al., 2006).

The negative binomial is an overdispersed Poisson distribution, where mean and variance are related as, $\sigma^2 = \mu + \phi \cdot \mu^2$, where σ^2 is the variance, μ is the mean and ϕ is the overdispersion coefficient (Svensson, 2020). The overdispersion parameter can be adjusted to cope with a range of excess variance; however, this introduces a free parameter into the model - the negative binomial distribution must be fit separately for each data-set, and in some procedures, fit separately for each gene.

Assuming one takes the maximum likelihood estimate of the mean (which all discussed methods do with the exception of Breda et al. (2021), which instead adopts a Bayesian approach to parameter estimation), overdispersion is the only parameter to 'fit' for negative binomial gene-expression models. Accordingly, there is much debate about how to model overdispersion and over what the biological or technical sources of the overdispersion are.

The popular `Seurat` software package takes a maximally flexible approach to overdispersion, fitting the overdispersion of each gene separately (Hafemeister and Satija, 2019). This data-driven approach makes minimal assumptions about the source and nature of overdispersion. However, to avoid over-fitting, Hafemeister and Satija (2019) regularises the gene-wise estimates of overdispersion using a kernel smoothing routine. The regression does admit some commonality in the source of overdispersion between genes, despite fitting an individual overdispersion term for each gene.

Townes et al. (2019), in contrast, builds a null model of gene expression from statistical first principles. They consider single-cell RNA-sequencing as a multinomial process, where there are N trials (total number of transcripts measured in a given cell) and where each gene has some probability of being measured (ideally equal to the proportion of transcripts belonging to that gene). Overdispersion is introduced through the use of a Dirichlet-multinomial distribution, which following the same assumptions as above, can be approximated as a series of negative binomial distributions when probabilities are sufficiently low and the number of trials sufficiently large (both valid assumptions in single-cell sequencing, where the large number of genes ensures that no single gene has a high probability of being sequenced). Townes et al. (2019) explicitly states that this overdispersion is biological in origin but does not provide specific guidance on fitting the overdispersion parameters, gene-wise or otherwise.

Lause et al. (2020) provides specific guidance for fitting overdispersion in the Townes et al. (2019) framework. Lause et al. (2020), shows that the `Seurat` approach is a rank-1 approximation of the Townes et al. (2019) framework and that `Seurat` overfits overdispersion. Instead, they suggest a single overdispersion parameter for all genes (a similar suggestion was made in Svensson (2020)). Moreover, they find that if count depth variation is regressed as an offset term of the model, that the estimated overdispersion parameter for technical control data is consistently small. Outside of count depth variation, which we can explicitly model, Lause et al. (2020) suggests overdispersion is primarily biological in origin. However, the authors of `Seurat` rebut this claim in Choudhary and Satija (2021).

Feature Selection

While the goal of normalisation is to minimise the technical noise component of gene expression heterogeneity, the goal of feature selection is to maximise the heterogeneity attributable to differential expression. Feature selection achieves this by selecting only those genes likely to be differentially expressed for inclusion in clustering (Yip et al., 2019; Luecken and Theis, 2019).

Furthermore, feature selection reduces the dimensionality of the data and improves the algorithmic run-time of downstream analyses (Saeys et al., 2007). Feature selection methods substantially reduce the dimensionality of the data, with typically 500-5000 genes selected. Note that while the number of genes chosen is arbitrary, downstream analysis is largely robust to the exact number of genes chosen (Klein et al., 2015; Luecken and Theis, 2019). This reduction in dimensionality is essential when downstream clustering depends on the Euclidean distance: when measured over a large number of genes, the small-scale noise of each gene adds up, inflating the total distance between cells. Over a sufficient number of genes, the Euclidean distance become effectively ‘meaningless’ in discriminating how close cells are (Beyer et al., 1999), a general feature of working in high dimensions known as the ‘curse of dimensionality’ (Beyer et al., 1999).

Feature selection also improves the statistical power of downstream analyses (Saeys et al., 2007). Post-clustering, the genes that are differentially expressed between clusters are identified through statistical testing. Due to the large number of genes involved, multiple-test corrections must be employed, lowering the power of any given test and increasing the type II error rate (Benjamini and Hochberg, 1995; Ascensión et al., 2021). By using only a reduced number of genes, the stringency of the multiple-test correction can be reduced, decreasing the type II error rate and increasing the number of differentially expressed genes identified correctly.

The disadvantage with feature selection is that it can be a self-fulfilling prophecy. Unsupervised clustering will cluster in the absence of any genuine structure, leveraging random, non-functional noise in gene expression to separate cells into clusters. Therefore, unsupervised clustering will tend to return the inputted genes as differentially expressed, even if they are not (Gao et al., 2020). Accordingly, a large number of genes are typically selected to ensure those genes genuinely differentially expressing will be included.

We will now introduce a selection of single-cell feature selection methods, grouping the methods into two main classes : model-based and model-free. The model-based approaches measure how well the observed count distribution holds to some statistical null model of gene expression – specifically those used in normalisation – and consist themselves of two further sub-classes: variance-based and drop-out based. Model-free approaches utilise some other measure of sequencing data without reference to the statistical null models used in normalisation.

Model-based

In the normalisation section, we discussed the use of statistical modelling in quantifying the fraction of heterogeneity due to noise. For example, we discussed how the negative binomial (the typical distribution for measuring count processes with greater than expected variance when compared with Poisson distribution) provides an appropriate error model for single-cell sequencing (Brennecke et al., 2013; Grün et al., 2014). Model-based feature selection methods utilise these error models as null models of gene expression, testing whether the observed gene expression matches that predicted from the null. If a gene matches the null, then that gene's expression distribution can be explained as resulting from noise; if not, it suggests that the gene is involved in some biological process. Model-based approaches select those genes with the least explainable expression distributions.

Model-based approaches vary in both the assumed null model and how deviation from the null is measured. There are various ways to quantify deviation, but the most popular in single-cell analysis are variance and drop-out (percent of zero counts). Both measures are mathematically simple and easy to compute, so they are useful for sorting through the deviation of thousands of genes quickly.

Note that model-based methods do not employ formal hypothesis testing against the error model. Feature selection is inherently limited, as differential expression cannot be measured pre-clustering; instead, only a gross estimate can be made. This limitation means that more exact approaches to estimating deviation from the null, e.g. hypothesis testing, are unnecessary, and faster simpler alternatives can be used.

Variance. The statistical models presented in normalisation, and specifically the negative binomial, predict genes to have a certain amount of variance based on their mean (Svensson, 2020). Variance-based approaches compare the observed and expected variance (while controlling for mean expression) to identify those genes that deviate from the null. The genes identified by variance approaches are referred to as HVGs – highly variable genes (Brennecke et al., 2013; Yip et al., 2019).

The simplest approach to variance selection is to compare the observed variance to the expected, both normalised by mean expression. This is the approach taken in Brennecke et al. (2013), which assumes expression data to follow a negative binomial distribution. However, this simple approach has been found to over-inflate the heterogeneity of lowly expressed genes and only marginally improve downstream analyses compared to a random selection of genes (Andrews and Hemberg, 2019; Townes et al., 2019; Kiselev et al., 2018).

Both Hafemeister and Satija (2019) & Townes et al. (2019) take a more sophisticated approach, measuring the variance of the normalised expression data (the Pearson residuals from the model). The approaches use slightly different null models but take the same approach to variance; in each, the top X genes by variance of normalised counts are selected as highly variable. Both approaches benefit from the enhanced variance stabilisation offered by normalisation; the highly variable genes approach employed by Brennecke et al. (2013) only normalised the variance, not the individual count values.

Drop-out. The second type of model-based feature selection method exploits the sparsity of single-cell data. Where variance-based methods measure deviation from the error model as the increase in variance, drop-out methods measure deviation as the increase in zero values compared with expected. The number of zeroes is quicker to estimate in sparse data than variance, and drop-out more directly assess the potential for differential gene expression -

broadly, differential expression will result in genes that are effectively 'on' in one cluster and 'off' in the remaining (Sparta et al., 2021). The 'off' clusters, even if they have mostly low but non-zero expression, will cause, due to sparsity, a large increase in zero values compared to that expected from the mean expression of the gene, which is set from the average of 'on' and 'off' clusters.

Drop-out methods differ in the assumed statistical null model of the data, e.g. Andrews and Hemberg (2019) presents two drop-out models, one based on Michaelis-Menten kinetics and the other on the negative binomial. The drop-out rate is easier to compute than the expected variance and does not require a full specification of the null model, only the expected number of zero values, so drop-out methods have access to a greater variety of practicable null models. For instance, Townes et al. (2019) developed a drop-out approach based on the less computationally tractable but more theoretically robust multinomial distribution (multinomial is more theoretical robust, as it models the competition of genes with each other to be sequenced, see above).

Sparta et al. (2021) argues against the use of the multinomial in drop-out based feature selection, asserting that the multinomial model assumes saturation in sequencing. Instead, Sparta et al. (2021) assume every transcript has an equal probability of being sequenced, p_c. The number of transcripts of a given gene in each cell is then the mean count value normalised by p_c; this establishes the null hypothesis that every cell expresses each gene equally. From this Sparta et al. (2021) develops a binomial distribution to obtain the probability of X number of cells with zero expression for a given gene. Sparta et al. (2021) has the advantage of producing a p-value, so standard significance testing can be used; the other methods produce an ordered set of genes, from which the top X is selected. However, given the blunt nature of feature selection, the need for this additional specificity is questionable.

Model-free

Model-free feature selection methods do not utilise error models of gene expression. Instead of testing whether a gene's expression distribution can be explained by noise, each method attempts to directly infer a gene's involvement in differential expression between cell types. Each method assumes a gene's expression distribution being not solely attributable to noise is insufficient criteria for assessing a gene's likeliness of being differentially expressed.

Ranjan et al. (2021) introduces DUBStepR, which identifies sets of highly correlated genes. DUBStepR assumes that differentially expressed genes will correlate strongly with each but not with other genes, whereas more ubiquitously expressed genes will have a uniform level of correlation.

Jiang et al. (2016) builds off the fact that differential expression restricts gene expression – for a given number of measured transcripts, differential expression involves distributing transcript counts to fewer and fewer cells. This restriction is analogous to the restriction in wealth associated with income equality: Jiang et al. (2016) proposes using a measure of income inequality, the Gini coefficient, to measure restriction in expression. Liu et al. (2020) proposes using entropy, another measure of wealth inequality, in a similar fashion.

`Triku`, introduced in Ascensión et al. (2021), measures how similar the expression of each gene is between each cell and its k-nearest-neighbours. `Triku` then compares the measured similarity to groupings of each cell with k random cells. The approach is equivalent to forming many small clusters and asking how coherently they express each gene; `Triku` identifies the genes that are likely to be expressed at similar levels within each cluster, abiding by the intra-cluster homogeneity sought by both clustering methods discussed above.

Linear Dimensions Reduction

Biology is intrinsically lower-dimensional than the total number of genes – the expression of each gene is dependent on the expression of the rest (Huang et al., 2005). In particular, involvement in biological processes, e.g. cell type, induces substantial correlation between genes, and the corresponding expression heterogeneity, reducing the effective dimensionality of the data. Dimension reduction techniques exploit these dependencies between genes, projecting the gene expression vectors of cells onto a lower number of dimensions.

The most popular linear dimension reduction technique is principal component analysis (PCA) (Luecken and Theis, 2019; Hotelling, 1933). Linear dimension reduction techniques project the data onto a linear subspace of the original space – each dimension of the reduced space is a linear combination of dimensions in the non-reduced space.

Principal component analysis changes the basis of the data so that the first principal component (dimension) captures the greatest variation possible, the second principal component the second most, and so on, under the constraint of orthogonality (Hotelling, 1933). Only the top d principal components are kept, with the value of d chosen either by default (commonly 50, or heuristically by an elbow analysis).

Each principal component is then a linear combination of genes, where each gene has some weight or loading, making principal component analysis highly interpretable. PCA removes much of the uncoordinated variability from a data set and identifies tranches of linearly co-dependent genes. As differential expression is one of the major sources of coordination in gene expression, these linearly co-dependent genes likely represent sets of deferentially expressing genes. Therefore, PCA represents a form of fuzzy feature selection, where genes are included in analysis with differing weights.

By capturing the linear dependencies between genes, PCA is able to reduce the dimension of the data substantially, typically down to 5 to 50 principal components (Stuart et al., 2019). Distances between cells calculated in this low-dimensional space are then used in clustering, either in totality for traditional unsupervised clustering or only a subset of the smallest distances in graphical clustering. Note that this reduction in dimensionality also serves to avoid the 'curse of dimensionality', discussed in reference to feature selection.

Visualisation

Biology is an inherently visual field; biologists require 'proof by visualisation' (Fox Keller, 2002). However, to visualise data, we need to map it onto a 2-dimensional plane (or 3-dimensional) and given the non-linear nature of multivariate gene expression (see **Section 1.2.3**), non-linear dimension reduction methods are required.

Non-linear dimension reduction methods exploit the non-linear dependencies between genes to generate a low-dimensional embedding of the data for visualisation. The most popular reduction techniques are tSNE (t-distributed Stochastic Neighbourhood Embedding) and UMAP (Uniform Manifold Approximation and Projection) (McInnes et al., 2018; Hinton and Roweis, 2003; van der Maaten and Hinton, 2008). Both UMAP and tSNE operates on the same principle and foundation as the Louvain method for graphical clustering – that we want to prioritise smaller distances between cells, and that we can achieve this through a k-nearest-neighbours graphical representation of the data.

UMAP, in particular, has become ubiquitous in single-cell analysis (Luecken and Theis, 2019). In brief, UMAP projects a transformed version of the kNN graph of the data onto two dimensions. This transformed version has additional topological guarantees on how well it captures the 'shape' of the data, compared to the untransformed version (see Nerve Theorem, (Zomorodian, 2005)), at the cost of the introduction of further hyperparameters.

UMAP maps the kNN onto low dimensions by constructing another graph, constrained to fit on a 2-dimensional Euclidean plane. UMAP then adjusts the low dimensional graph to minimise the divergence with the high dimensional kNN. The resulting low dimensional graph is visualised as a point cloud of cells, where cells that are close together on the UMAP projection are close together in the full-dimensional gene space. Accordingly, the low-dimensional embedding to can be used assess a proposed clustering – cells of the same type should have similar expression, so colouring cells by type on the visualisation should reveal contiguous groupings of cells.

Dimension reduction by tSNE operates along similar principles to UMAP (Becht et al., 2019). However, instead of maximising the similarity between kNN graphs constructed in high and low-dimensions, tSNE maximises the similarity between probability distributions in high and low-dimensions (Hinton and Roweis, 2003). The distributions encode the probability that cells are 'similar' or are 'neighbours' in high and low-dimensions respectively, i.e. are close by Euclidean distance. The low-dimension distributions have a longer tail of probability, prioritising the preservation of smaller distances through the projection onto lower dimensions. Informally, tSNE is a continuous analogue of the discrete UMAP.

UMAP and tSNE have proven remarkably successful at visualising single-cell data (for examples, see (Becht et al., 2019)). Both methods operate on the same principle as graphical clustering – small distances are more reliable than large ones. This preserves the neighbourhood structure of the data, with cells that are close in high dimensions being close in low-dimensions, but means that larger distances cannot be directly interpreted on a UMAP or tSNE visualisation (unlike PCA, see **Section 1.2.4**). Thus, the results of clusterings can be qualitatively assessed – cells assigned to the same cluster should be close together on the projection. Note however, that UMAP and tSNE preserve only small scales distances; the relative size of larger distances between cells on the low-dimensional embedding cannot be interpreted.

Appendix C

Supplementary Code

This appendix includes the code developed for **Chapters 2** & **3**. The code is available as an R package on GitHub https://github.com/mjcasy/scEC.

Normalisation

The below code is for the calculation of the James-Stein-type shrinkage estimate of the frequency of transcripts expressed in each cell. The first function, GetFreqShrink, estimates the frequency for a single gene. The second function, GetFreq, applies GetFreqShrink to all genes, returning a gene by cell matrix of James-Stein estimator frequencies. Note that neither function is exported for use; both are used within other functions.

```
#' James-Stein Frequency Estimator
#'
#' @param transposeCounts Transposed sparse count matrix
#' @param ind Integer indicating chosen gene (row number in count matrix)
#' @param N N Number of cells
#' @param Total Total Integer of total counts per cell
#'
#' @return Numeric vector of shrinkage cell frequencies
#'
GetFreqShrink <- function(transposeCounts, ind, N, Total){

  tk <- rep(1/N, N)
  tkadj <- tk

  count <- transposeCounts@x[(transposeCounts@p[ind]+1) : transposeCounts@p[ind+1]]
  elements <- transposeCounts@i[(transposeCounts@p[ind]+1) : transposeCounts@p[ind+1]]+1

  freq <- count / Total[ind]

  num <- 1 - sum(freq^2)
  tkadj[elements] <- tkadj[elements] - freq
  den <- (sum(count) - 1)*sum(tkadj^2)
  lambda <- num/den
  lambda[lambda > 1] <- 1

  freqshrink <- lambda*tk
  freqshrink[elements]  <- freqshrink[elements] + (1 - lambda)*freq

  freqshrink
}


#' Normalise Frequencies
#'
#' @param CountsMatrix Feature x cell sparse counts matrix of class dgCMatrix
#'
#' @return Feature x cell dense matrix of frequencies
#'
GetFreq <- function(CountsMatrix){

  Total <- Matrix::rowSums(CountsMatrix)
  transposeCounts <- Matrix::t(CountsMatrix)
  Indices <- length(transposeCounts@p)-1
  N <- dim(CountsMatrix)[2]

  freq <- matrix(NA, nrow = nrow(CountsMatrix), ncol = N)
  for (ind in 1:Indices) {
    freq[ind,] <- GetFreqShrink(transposeCounts, ind, N, Total)
  }

  freq
}
```

Feature Selection

Functions for the calculation of population heterogeneity and subsequent feature selection based on population heterogeneity. The function `Population` calculates gene-wise population heterogeneities. The intended use of `Population` is for plotting $I(g)$ against mean expression, of the type seen in **Fig 2.13**. The function `FeatureSelection` returns the names of the top `nGenes` by population heterogeneity that have at least `minCounts` total transcripts.

```
#' Population Heterogeneity
#'
#' @param CountsMatrix Feature x cell sparse counts matrix of class dgCMatrix
#'
#' @return Numeric vector of gene-wise population heterogeneities
#' @export
#'
#' @examples
Population <- function(CountsMatrix) {

  Total <- Matrix::rowSums(CountsMatrix)
  N <- ncol(CountsMatrix)

  transposeCounts <- Matrix::t(CountsMatrix)
  Indices <- length(transposeCounts@p)-1
  Pop <- vector("numeric", length(Indices))

  for (ind in 1:Indices) {
    freqshrink <- GetFreqShrink(transposeCounts, ind, N, Total)
    LogNfreq <- log(N*freqshrink)
    LogNfreq[LogNfreq == -Inf] <- 0
    Pop[ind] <- t(freqshrink) %*% LogNfreq
  }


  Pop[is.infinite(Pop)] <- 0

  names(Pop) <- colnames(transposeCounts)

  Pop
}

#' Feature Selection by Population Heterogeneity
#'
#' @param CountsMatrix Feature x cell sparse counts matrix of class dgCMatrix
#' @param minCounts Minimum number of transcripts per gene
#' @param nGenes Number of genes selected
#'
#' @return Vector of top genes by population heterogeneity
#' @export
#'
#' @examples
FeatureSelection <- function(CountsMatrix, minCounts = 100, nGenes = 500){

  Exp <- rownames(CountsMatrix)[Matrix::rowSums(CountsMatrix) >= minCounts]

  Div <- Population(CountsMatrix[Exp,])
  GOI <- names(sort(Div, decreasing = T)[1:nGenes])

  GOI
}
```

Unsupervised Clustering

Function for unsupervised clustering. For efficiency, numerical optimisation is carried out in `Python3` instead of `R`, with the `Python3` code detailed below. The `Cluster` function does not carry out feature selection, allowing alternative feature selection methods to that introduced above to be used.

The `Cluster` function carries out clustering for and up to `numClus` clusters. The function allows for mutlistart, and has an option to set a seed. Note the seed must be set internally in the function, as opposed to using the standard `R` function `set.seed`, as the standard function will fail to set the seed for the `Python3` code.

```
#' Single-Cell Entropic Clustering
```

```
#'
#' @param CountsMatrix Feature x cell sparse counts matrix of class dgCMatrix
#' @param numClus Number of clusters (or maximum number of clusters for greedy)
#' @param Multistart Number of restarts at each step
#' @param Greedy Boolean. Should greedy algorithm be used.
#' @param Seed Seed
#'
#' @return When Greedy = F, vector of integer identities. When Greedy = T,
#' matrix where each column is a vector of integer identities. The Nth column
#' encodes N clusters.
#' @export
#'
#' @examples
Cluster <- function(CountsMatrix, numClus, Multistart = 5, Seed){

  if(!missing(Seed)){
    set.seed(Seed)
    reticulate::py_set_seed(seed = Seed)
  }

  G <- dim(CountsMatrix)[1]
  N <- dim(CountsMatrix)[2]

  FullFreq <- GetFreq(CountsMatrix)

  mu <- PyFunc$multiStartClusterCells(freq = FullFreq, numClusters = numClus,
  multistart = Multistart)
  Ident <- apply(mu, 1, which.max) - 1

  Ident
}
```

Below is detailed the `Python3` code for use in clustering. Note this code is not exported, serving solely as background for the `Cluster` function.

```
import numpy as np
from scipy import optimize

def ident2mu(ident):

  N = ident.size
  C = np.unique(ident).size
  mu = np.zeros((N,C), 'int')

  for i in range(C):
    mu[ident == i, i] = 1

  return mu


def pop(freq):

  freq = np.array(freq)

  N = freq.shape[1]

  popG = np.sum(freq * np.log(N * freq), 1)
  return popG


def intercluster(freq, ident):

  ident = np.array(ident).astype(int)
  freq = np.array(freq)

  mu = ident2mu(ident)

  y = freq @ mu
  N = ident.size

  Nk = np.sum(mu, 0)
  logNk = np.log(Nk)

  interclusterG = np.sum(y * (np.log(N * y) - logNk), 1)

  return interclusterG


def funcCost(wVec, freq, tfreq):
  N = freq.shape[1]
  C = int((wVec.size / N))

  wVecar = np.array(wVec)
  W = wVecar.reshape(N, C)
  M = np.sum(np.exp(W), 1)
  mu = np.exp(W) / M[:,None]

  y = freq @ mu

  Nj = np.sum(mu, 0)
  logNj = np.log(Nj)

  Is = -1*np.sum(y * (np.log(N * y) - logNj))

  return Is


def gradCost(wVec, freq, tfreq):
  N = freq.shape[1]
  C = int(wVec.size / N)

  wVecar = np.array(wVec)
  W = wVecar.reshape(N, C)
  M = np.sum(np.exp(W), 1)
  mu = np.exp(W) / M[:,None]

  y = freq @ mu

  Nj = np.sum(mu, 0)
  yNj = y / Nj

  term1 = np.log(yNj) + np.log(N) + 1
  dIsdu = tfreq @ term1
  term2 = -1*np.sum(yNj, 0)
  dIsdu = dIsdu + term2
  dIsdw = dIsdu * mu * (1 - mu)
  iterC = np.arange(C)

  for j in iterC:
    ind = iterC[iterC!=j]
    term2 = np.transpose(-1 * np.transpose(mu[:,ind]) * mu[:,j])
    dIsdw[:,j] = dIsdw[:,j] + np.sum(dIsdu[:,ind] * term2, 1)

  dIsdwVec = -1 * dIsdw.reshape((dIsdw.size,))
  return dIsdwVec


def clusterCells(freq, numClusters):
  freq = np.array(freq)
  numClusters = int(numClusters)

  N = freq.shape[1]
  tfreq = freq.T
  wVec = np.random.uniform(low = -0.5, high = 0.5, size = (numClusters*N,))

  bounding = 3
  Bounds=optimize.Bounds(lb=-bounding, ub=bounding)

  Out = optimize.minimize(funcCost,
                          x0 = wVec,
                          args = (freq, tfreq),
                          method = 'L-BFGS-B',
                          jac=gradCost,
                          bounds=Bounds)

  W = Out.x.reshape(N, numClusters)
  M = np.sum(np.exp(W), 1)
  mu = np.exp(W) / M[:,None]

  return mu


def multiStartClusterCells(freq, numClusters, multistart):

  freq = np.array(freq)
  numClusters = int(numClusters)
  multistart = int(multistart)

  maxScore = 0

  for i in range(multistart):
    tempMu = clusterCells(freq, numClusters)
    newIdent = tempMu.argmax(1)
    Score = intercluster(freq, newIdent).sum()

    if Score > maxScore:
      maxmu = tempMu
      maxScore = Score

  return maxmu
```

Semi-supervised Classification

Functions for the semi-supervised classification of a data set based on another. As with clustering, underlying optimisation carried out in `Python3`. `MapFeatureSelection` selects genes with high $H_S(g)$ in reference data set; note that it is not exported. `Map` carries out the semi-supervised classification.

```
#' Feature Selection by Inter-cluster Heterogeneity
#'
#' @param ReferenceCountsMatrix Reference count matrix; feature x cell sparse counts matrix of c
#' @param ReferenceID Factor of reference cell identities
#' @param minCounts Minimum number of transcripts per gene
#' @param nGenes Number of genes selected
#'
#' @return Vector of top genes by inter-cluster heterogeneity
#'
#' @examples
MapFeatureSelection <- function(ReferenceCountsMatrix, ReferenceID, minCounts, nGenes){

  Exp <- rownames(ReferenceCountsMatrix)[Matrix::rowSums(ReferenceCountsMatrix) >= minCounts]

  Div <- DifferentialExpression(ReferenceCountsMatrix[Exp,], ReferenceID)
  GOI <- names(sort(Div, decreasing = T)[1:nGenes])

  GOI
}

#' Mapping of Cells Onto Reference Cell Identities
#'
#' @param MapCountsMatrix Count matrix of cells to be mapped
#' @param ReferenceCountsMatrix Count matrix of cells with known identities
#' @param ReferenceID Factor of reference cell identities
#' @param minCounts Minimum number of transcripts per gene
#' @param nGenes Number of genes selected
#' @param Seed Seed set for both R and Python components
#'
#' @return Mapped cellular identities
#' @export
#'
#' @examples
Map <- function(MapCountsMatrix, ReferenceCountsMatrix, ReferenceID, minCounts = 100, nGenes = 1

  if(!missing(Seed)){
    set.seed(Seed)
    reticulate::py_set_seed(seed = Seed)
  }

  RefN <- ncol(ReferenceCountsMatrix)
  MapN <- ncol(MapCountsMatrix)
  N <- RefN + MapN

  GOI <- MapFeatureSelection(ReferenceCountsMatrix, ReferenceID, minCounts = minCounts, nGenes =
  ReferenceCountsMatrix <- ReferenceCountsMatrix[GOI,]
  MapCountsMatrix <- MapCountsMatrix[GOI,]

  RefFreq <- GetFreq(ReferenceCountsMatrix)
  MapFreq <- GetFreq(MapCountsMatrix)
  FullFreq <- cbind((RefN/N) * RefFreq, (MapN/N) * MapFreq)
  RefID <- as.numeric(ReferenceID)-1

  mu <- PyFunc$meld(freq = FullFreq, refID = RefID)
  Ident <- apply(mu, 1, which.max) - 1

  levels(ReferenceID)[Ident+1]

}
```

Python3 background functions semi-supervised classification.

```
def funcCostMeld(wVec, freq, tfreq, refW):
  N = freq.shape[1]
  mapN = N - refW.shape[0]
  C = int((wVec.size / mapN))

  wVecar = np.array(wVec)
  mapW = wVecar.reshape(mapN, C)

  W = np.concatenate((refW, mapW))

  M = np.sum(np.exp(W), 1)
  mu = np.exp(W) / M[:,None]

  y = freq @ mu

  Nj = np.sum(mu, 0)
  logNj = np.log(Nj)

  Is = -1*np.sum(y * (np.log(N * y) - logNj))

  return Is


def gradCostMeld(wVec, freq, tfreq, refW):
  N = freq.shape[1]
  mapN = N - refW.shape[0]
  C = int((wVec.size / mapN))

  wVecar = np.array(wVec)
  mapW = wVecar.reshape(mapN, C)

  W = np.concatenate((refW, mapW))

  M = np.sum(np.exp(W), 1)
  mu = np.exp(W) / M[:,None]

  y = freq @ mu

  Nj = np.sum(mu, 0)
  yNj = y / Nj

  term1 = np.log(yNj) + np.log(N) + 1
  dIsdu = tfreq @ term1
  term2 = -1*np.sum(yNj, 0)
  dIsdu = dIsdu + term2
  M = np.sum(np.exp(mapW), 1)
  mu = np.exp(mapW) / M[:,None]

  dIsdw = dIsdu * mu * (1 - mu)

  iterC = np.arange(C)

  for j in iterC:
    ind = iterC[iterC!=j]
    term2 = np.transpose(-1 * np.transpose(mu[:,ind]) * mu[:,j])
    dIsdw[:,j] = dIsdw[:,j] + np.sum(dIsdu[:,ind] * term2, 1)

  dIsdwVec = -1 * dIsdw.reshape((dIsdw.size,))
  return dIsdwVec

def meld(freq, refID):
  freq = np.array(freq)
  N = freq.shape[1]

  refID = np.array(refID).astype(int)
  refN = refID.size

  C = np.unique(refID).size

  refW = np.ones(shape = (refN,C))
  refW = -10*refW
  refW[np.arange(refN),refID] = 10

  mapN = N - refN
  wVec = np.random.uniform(low = -0.5, high = 0.5, size = (C*mapN,))
  wVec = np.zeros(shape = (C*mapN,))

  rangeN = np.arange(start = refN, stop = N)
  tfreq = freq[:,rangeN].T

  bounding = 3
  Bounds=optimize.Bounds(lb=-bounding, ub=bounding)

  Out = optimize.minimize(funcCostMeld,
                          x0 = wVec,
                          args = (freq, tfreq, refW),
                          method = ’L-BFGS-B’,
                          jac=gradCostMeld,
                          bounds=Bounds)

  W = Out.x.reshape(mapN, C)
  M = np.sum(np.exp(W), 1)
  mu = np.exp(W) / M[:,None]

  return mu
```

Differential Gene Expression

The function `DifferentialExpression` returns the inter-type heterogeneity of each gene based on the discrete clustering `Identity`. Differentially expressed genes are the chosen manually on the basis of the returned values or through randomisation and exact testing as in **Chapter 2**.

```
#' Differential Expression by Inter-Type Heterogeneity
#'
#' @param CountsMatrix Feature x cell sparse counts matrix of class dgCMatrix
#' @param Identity Factor of cell identities
#'
#' @return Numeric vector of gene-wise inter-type heterogeneities
#' @export
#'
#' @examples
DifferentialExpression <- function(CountsMatrix, Identity) {

  if(length(Identity) != ncol(CountsMatrix)){
    stop("Inconsistent number of cells between objects:\n\tlength(Identity) !=
    ncol(CountsMatrix)")
  }

  Total <- Matrix::rowSums(CountsMatrix)
  N <- ncol(CountsMatrix)

  Ng <- as.vector(table(Identity))

  transposeCounts <- Matrix::t(CountsMatrix)

  Indices <- length(transposeCounts@p)-1
  InterType <- vector("numeric", length(Indices))

  for (ind in 1:Indices) {
    freqshrink <- GetFreqShrink(transposeCounts, ind, N, Total)
    groupedfreqshrink <- tapply(freqshrink, Identity, sum)

    NonZero <- which(groupedfreqshrink != 0)

    InterType[ind] <- t(groupedfreqshrink[NonZero]) %*% log(N*groupedfreqshrink[NonZero] /
    Ng[NonZero])
  }

  InterType[is.infinite(InterType)] <- 0

  names(InterType) <- rownames(CountsMatrix)

  InterType
}
```

References

C. Ahlmann-Eltze and W. Huber. glmgampoi: fitting gamma-poisson generalized linear models on single cell count data. *Bioinformatics*, 36(24): 5701–5702, 2020.

K. Akashi, D. Traver, T. Miyamoto, and I. L. Weissman. A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. *Nature*, 404(6774): 193, 2000.

B. Alberts. *Molecular Biology of the Cell*. W.W. Norton, 2017. ISBN 9781317563754.

T. S. Andrews and M. Hemberg. M3drop: dropout-based feature selection for scrnaseq. *Bioinformatics*, 35(16):2865–2867, 2019.

D. Arendt. The evolution of cell types in animals: emerging principles from molecular studies. *Nature Reviews Genetics*, 9(11):868–882, 2008.

D. Arendt, J. M. Musser, C. V. Baker, A. Bergman, C. Cepko, D. H. Erwin, M. Pavlicev, G. Schlosser, S. Widder, M. D. Laubichler, et al. The origin and evolution of cell types. *Nature Reviews Genetics*, 17(12):744–757, 2016.

A. M. Ascensión, O. Ibañez-Solé, I. Inza, A. Izeta, and M. J. Araúzo-Bravo. Triku: a feature selection method based on nearest neighbors for single-cell data. *bioRxiv*, 2021.

M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

M. Assaf, E. Roberts, and Z. Luthey-Schulten. Determining the stability of genetic switches: explicitly accounting for mrna noise. *Physical review letters*, 106(24):248102, 2011.

N. Balaskas, A. Ribeiro, J. Panovska, E. Dessaud, N. Sasai, K. M. Page, J. Briscoe, and V. Ribes. Gene regulatory logic for reading the sonic hedgehog signaling gradient in the vertebrate neural tube. *Cell*, 148(1-2):273–284, 2012.

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

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

K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is "nearest neighbor" meaningful? In *International conference on database theory*, pages 217–235. Springer, 1999.

E. Bianconi, A. Piovesan, F. Facchin, A. Beraudi, R. Casadei, F. Frabetti, L. Vitale, M. C. Pelleri, S. Tassani, F. Piva, et al. An estimation of the number of cells in the human body. *Annals of human biology*, 40(6):463–471, 2013.

Å. K. Björklund, M. Forkel, S. Picelli, V. Konya, J. Theorell, D. Friberg, R. Sandberg, and J. Mjösberg. The heterogeneity of human cd127+ innate lymphoid cells revealed by single-cell rna sequencing. *Nature immunology*, 17 (4):451, 2016.

V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

J. L. Borges. El idioma analítico de john wilkins. *Otras inquisiciones*, 2, 1952.

S. Bradley, A. Hax, A. Hax, and T. Magnanti. *Applied Mathematical Programming*. Addison-Wesley Publishing Company, 1977. ISBN 9780201004649.

J. Breda, M. Zavolan, and E. van Nimwegen. Bayesian inference of gene expression states from single-cell rna-seq data. *Nature Biotechnology*, pages 1–9, 2021.

P. Brennecke, S. Anders, J. K. Kim, A. A. Kołodziejczyk, X. Zhang, V. Proserpio, B. Baying, V. Benes, S. A. Teichmann, J. C. Marioni, et al. Accounting for technical noise in single-cell rna-seq experiments. *Nature methods*, 10(11): 1093–1095, 2013.

R. J. Britten and E. H. Davidson. Gene regulation for higher cells: a theory. *Science*, 165(3891):349–357, 1969.

F. Buettner, K. N. Natarajan, F. P. Casale, V. Proserpio, A. Scialdone, F. J. Theis, S. A. Teichmann, J. C. Marioni, and O. Stegle. Computational analysis of cell-to-cell heterogeneity in single-cell rna-sequencing data reveals hidden subpopulations of cells. *Nature biotechnology*, 33(2):155, 2015.

R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on scientific computing*, 16(5): 1190–1208, 1995.

A. C. Cameron and P. K. Trivedi. *Regression analysis of count data*, volume 53. Cambridge university press, 2013.

G. Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

M. J. Casey, R. J. Sanchez-Garcia, and B. D. MacArthur. Measuring the information obtained from a single-cell sequencing experiment. *bioRxiv*, 2020a.

M. J. Casey, P. S. Stumpf, and B. D. MacArthur. Theory of cell fate. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 12(2):e1471, 2020b.

S. R. Casjens and R. W. Hendrix. Bacteriophage lambda: early pioneer and still relevant. *Virology*, 479:310–330, 2015.

T. Chari, J. Banerjee, and L. Pachter. The specious art of single-cell genomics. *bioRxiv*, 2021.

W. Chen, O. Guillaume-Gentil, R. Dainese, P. Y. Rainer, M. Zachara, C. G. Gabelein, J. A. Vorholt, and B. Deplancke. Genome-wide molecular recording using live-seq. *bioRxiv*, 2021.

S. Choudhary and R. Satija. Comparison and evaluation of statistical error models for scrna-seq. *bioRxiv*, 2021. .

F. R. Chung and F. C. Graham. *Spectral graph theory*. Number 92. American Mathematical Soc., 1997.

H. Clevers, S. Rafelski, M. Elowitz, A. Klein, J. Shendure, C. Trapnell, E. Lein, E. Lundberg, M. Uhlen, A. Martinez-Arias, et al. What is your conceptual definition of "cell type" in the context of a mature organism? *Cell Systems*, 4 (3):255–259, 2017.

T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

M.-A. Dillies, A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, G. Marot, D. Castel, J. Estelle, et al. A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. *Briefings in bioinformatics*, 14(6):671–683, 2013.

A. Duò, M. D. Robinson, and C. Soneson. A systematic performance evaluation of clustering methods for single-cell rna-seq data. *F1000Research*, 7, 2018.

G. Eraslan, L. M. Simon, M. Mircea, N. S. Mueller, and F. J. Theis. Single-cell rna-seq denoising using a deep count autoencoder. *Nature communications*, 10(1):1–14, 2019.

L. Evans. *Partial Differential Equations*. Graduate studies in mathematics. American Mathematical Society, 2010. ISBN 9780821849743.

J. E. Ferrell Jr. Bistability, bifurcations, and waddington's epigenetic landscape. *Current biology*, 22(11):R458–R466, 2012.

S. Fischer and J. Gillis. How many markers are needed to robustly determine a cell's type? *BioRxiv*, 2021.

R. Fisher. *Statistical Methods For Research Workers*. Gyan Books, 2017. ISBN 9789351286585.

E. Fox Keller. Making sense of life. In *Fundamental Changes in Cellular Biology in the 20th Century. Biology of Development, Chemistry and Physics in the Life Sciences: Proceedings of the XXth International Congress of History of Science (Liège, 20-26 July 1997) Vol. III*, pages 173–178, 2002.

S. Freytag, L. Tian, I. Lönnstedt, M. Ng, and M. Bahlo. Comparison of clustering tools in r for medium-sized 10x genomics single-cell rna-sequencing data. *F1000Research*, 7, 2018.

J. Friedman, T. Hastie, R. Tibshirani, et al. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

L. L. Gao, J. Bien, and D. Witten. Selective inference for hierarchical clustering. *arXiv preprint arXiv:2012.02936*, 2020.

Gene Ontology Consortium. The gene ontology resource: enriching a gold mine. *Nucleic Acids Research*, 49(D1):D325–D334, 2021.

S. Goodwin, J. D. McPherson, and W. R. McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6): 333–351, 2016.

P. Greulich, R. Smith, and B. D. MacArthur. The physics of cell fate. In H. Levine, M. K. Jolly, P. Kulkarni, and V. Nanjundiah, editors, *Phenotypic Switching*, pages 189–206. Academic Press, 2020. ISBN 978-0-12-817996-3. .

D. Grün, L. Kester, and A. Van Oudenaarden. Validation of noise models for single-cell transcriptomics. *Nature methods*, 11(6):637–640, 2014.

A. Guillemin and M. P. Stumpf. Non-equilibrium statistical physics, transitory epigenetic landscapes, and cell fate decision dynamics. *arXiv preprint arXiv:2011.04252*, 2020.

S. P. Gygi, Y. Rochon, B. R. Franza, and R. Aebersold. Correlation between protein and mrna abundance in yeast. *Molecular and cellular biology*, 19(3): 1720–1730, 1999.

C. Hafemeister and R. Satija. Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression. *Genome Biology*, 20(1):1–15, 2019.

M. Hagemann-Jensen, C. Ziegenhain, P. Chen, D. Ramsköld, G.-J. Hendriks, A. J. Larsson, O. R. Faridani, and R. Sandberg. Single-cell rna counting at allele and isoform resolution using smart-seq3. *Nature Biotechnology*, 38(6): 708–714, 2020.

F. A. Haight. Handbook of the poisson distribution. 1967.

X. Han, Z. Zhou, L. Fei, H. Sun, R. Wang, Y. Chen, H. Chen, J. Wang, H. Tang, W. Ge, et al. Construction of a human cell landscape at single-cell level. *Nature*, 581(7808):303–309, 2020.

J. Hausser and K. Strimmer. Entropy inference and the james-stein estimator, with application to nonlinear gene association networks. *Journal of Machine Learning Research*, 10(7), 2009.

G. E. Hinton and S. T. Roweis. Stochastic neighbor embedding. In *Advances in neural information processing systems*, pages 857–864, 2003.

R. Hooke. *Micrographia: or some physiological descriptions of minute bodies made by magnifying glasses, with observations and inquiries thereupon*. Royal Society, 1665.

H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

S. Huang. The molecular and mathematical basis of waddington's epigenetic landscape: A framework for post-darwinian biology? *Bioessays*, 34(2): 149–157, 2012.

S. Huang, G. Eichler, Y. Bar-Yam, and D. E. Ingber. Cell fates as high-dimensional attractor states of a complex gene regulatory network. *Physical review letters*, 94(12):128701, 2005.

L. Hubert and P. Arabie. Comparing partitions. *Journal of classification*, 2(1): 193–218, 1985.

S. Islam, A. Zeisel, S. Joost, G. La Manno, P. Zajac, M. Kasper, P. Lönnerberg, and S. Linnarsson. Quantitative single-cell rna-seq with unique molecular identifiers. *Nature methods*, 11(2):163–166, 2014.

F. Jacob and J. Monod. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of molecular biology*, 3(3):318–356, 1961.

A. K. Jain. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666, 2010.

W. James and C. Stein. Estimation with quadratic loss. In *Breakthroughs in statistics*, pages 443–460. Springer, 1992.

E. T. Jaynes. Information theory and statistical mechanics. *Physical review*, 106 (4):620, 1957.

L. Jiang, H. Chen, L. Pinello, and G.-C. Yuan. Giniclust: detecting rare cell types from single-cell gene expression data with gini index. *Genome biology*, 17(1):1–13, 2016.

J. Jost and R. Mulas. Normalized laplace operators for hypergraphs with real coefficients. *Journal of Complex Networks*, 9(1):cnab009, 2021.

L. Junqueira, L. Junqueira, J. Carneiro, and R. Kelley. *Basic Histology*. Lange medical book. Prentice-Hall International, 1992. ISBN 9780838505793.

S. A. Kauffman. Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of theoretical biology*, 22(3):437–467, 1969.

J. K. Kim, A. A. Kolodziejczyk, T. Ilicic, S. A. Teichmann, and J. C. Marioni. Characterizing noise structure in single-cell rna-seq distinguishes genuine from technical stochastic allelic expression. *Nature communications*, 6(1):1–9, 2015.

T. Kim, I. R. Chen, Y. Lin, A. Y.-Y. Wang, J. Y. H. Yang, and P. Yang. Impact of similarity metrics on single-cell rna-seq data clustering. *Briefings in bioinformatics*, 20(6):2316–2326, 2019.

V. Y. Kiselev, K. Kirschner, M. T. Schaub, T. Andrews, A. Yiu, T. Chandra, K. N. Natarajan, W. Reik, M. Barahona, A. R. Green, et al. Sc3: consensus clustering of single-cell rna-seq data. *Nature methods*, 14(5):483–486, 2017.

V. Y. Kiselev, A. Yiu, and M. Hemberg. scmap: projection of single-cell rna-seq data across data sets. *Nature methods*, 15(5):359–362, 2018.

V. Y. Kiselev, T. S. Andrews, and M. Hemberg. Challenges in unsupervised clustering of single-cell rna-seq data. *Nature Reviews Genetics*, 20(5):273–282, 2019.

T. Kivioja, A. Vähärautio, K. Karlsson, M. Bonke, M. Enge, S. Linnarsson, and J. Taipale. Counting absolute numbers of molecules using unique molecular identifiers. *Nature methods*, 9(1):72–74, 2012.

S. Klamt, U.-U. Haus, and F. Theis. Hypergraphs and cellular networks. *PLoS computational biology*, 5(5):e1000385, 2009.

A. M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D. A. Weitz, and M. W. Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5): 1187–1201, 2015.

J. Krebs, E. Goldstein, and S. Kilpatrick. *Lewin's GENES XII*. G - Reference,Information and Interdisciplinary Subjects Series. Jones & Bartlett Learning, 2017. ISBN 9781284104493.

K. Kruse and F. Jülicher. Oscillations in cell biology. *Current opinion in cell biology*, 17(1):20–26, 2005.

S. Kullback and R. A. Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

R. Laskey and J. Gurdon. Genetic content of adult somatic cells tested by nuclear transplantation from cultured cells. *Nature*, 228(5278):1332, 1970.

J. Lause, P. Berens, and D. Kobak. Analytic pearson residuals for normalization of single-cell rna-seq umi data. *bioRxiv*, 2020.

R. E. Lewand. *Cryptological mathematics*, volume 16. American Mathematical Soc., 2000.

H. Li, J. Janssens, M. De Waegeneer, S. S. Kolluru, K. Davie, V. Gardeux, W. Sealens, F. David, M. Brbic, J. Leskovec, et al. Fly cell atlas: a single-cell transcriptomic atlas of the adult fruit fly. *bioRxiv*, 2021a.

T. Li, S. Ma, and M. Ogihara. Entropy-based criterion in categorical clustering. In *Proceedings of the twenty-first international conference on Machine learning*, page 68, 2004.

Y. Li, P. Luo, Y. Lu, and F.-X. Wu. Identifying cell types from single-cell data based on similarities and dissimilarities between cells. *BMC bioinformatics*, 22(3):1–18, 2021b.

B. Liu, C. Li, Z. Li, D. Wang, X. Ren, and Z. Zhang. An entropy-based metric for assessing the purity of single cell populations. *Nature communications*, 11 (1):1–13, 2020.

S. Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, and N. Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.

M. Lotfollahi, M. Naghipourfar, M. D. Luecken, M. Khajavi, M. Büttner, M. Wagenstetter, Ž. Avsec, A. Gayoso, N. Yosef, M. Interlandi, et al. Mapping single-cell data to reference atlases by transfer learning. *Nature Biotechnology*, pages 1–10, 2021.

M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):1–21, 2014.

M. D. Luecken and F. J. Theis. Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular systems biology*, 15(6):e8746, 2019.

A. Lun. Overcoming systematic errors caused by log-transformation of normalized single-cell rna sequencing data. *BioRxiv*, page 404962, 2018.

A. T. Lun, K. Bach, and J. C. Marioni. Pooling across cells to normalize single-cell rna sequencing data with many zero counts. *Genome biology*, 17(1): 1–14, 2016.

F.-J. Lv, R. S. Tuan, K. M. Cheung, and V. Y. Leung. Concise review: the surface markers and identity of human mesenchymal stem cells. *Stem cells*, 32(6): 1408–1419, 2014.

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

P. Mazzarello. A unifying concept: the history of cell theory. *Nature cell biology*, 1(1):E13–E15, 1999.

C. S. McGinnis, L. M. Murrow, and Z. J. Gartner. Doubletfinder: doublet detection in single-cell rna sequencing data using artificial nearest neighbors. *Cell systems*, 8(4):329–337, 2019.

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

M. Melé, P. G. Ferreira, F. Reverter, D. S. DeLuca, J. Monlong, M. Sammeth, T. R. Young, J. M. Goldmann, D. D. Pervouchine, T. J. Sullivan, et al. The human transcriptome across tissues and individuals. *Science*, 348(6235): 660–665, 2015.

A. Mescher. *Junqueira's Basic Histology: Text and Atlas, Fifteenth Edition*. McGraw-Hill Education, 2018. ISBN 9781260026177.

R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298 (5594):824–827, 2002.

K. Mitsui, Y. Tokuzawa, H. Itoh, K. Segawa, M. Murakami, K. Takahashi, M. Maruyama, M. Maeda, and S. Yamanaka. The homeoprotein nanog is required for maintenance of pluripotency in mouse epiblast and es cells. *Cell*, 113(5):631–642, 2003.

J. Monod and F. Jacob. General conclusions: teleonomic mechanisms in cellular metabolism, growth, and differentiation. In *Cold Spring Harbor symposia on quantitative biology*, volume 26, pages 389–401. Cold Spring Harbor Laboratory Press, 1961.

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

R. Mulas and M. J. Casey. Estimating cellular redundancy in networks of genetic expression. *Mathematical Biosciences*, page 108713, 2021. ISSN 0025-5564.

M. E. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

S. A. Newman. Cell differentiation: What have we learned in 50 years? *Journal of theoretical biology*, 485:110031, 2020.

S. Nurk, S. Koren, A. Rhie, M. Rautiainen, A. V. Bzikadze, A. Mikheenko, M. R. Vollger, N. Altemose, L. Uralsky, A. Gershman, et al. The complete sequence of a human genome. *bioRxiv*, 2021.

O. R. Oellermann and A. J. Schwenk. The laplacian spectrum of graphs. 1991.

E. Papalexi and R. Satija. Single-cell rna sequencing to explore immune cell heterogeneity. *Nature Reviews Immunology*, 18(1):35–45, 2018.

I. Peter and E. Davidson. *Genomic Control Process: Development and Evolution*. Elsevier Science, 2015. ISBN 9780124047297.

G. Peters, F. Crespo, P. Lingras, and R. Weber. Soft clustering–fuzzy and rough approaches and their extensions and derivatives. *International Journal of Approximate Reasoning*, 54(2):307–322, 2013.

S. Picelli, Å. K. Björklund, O. R. Faridani, S. Sagasser, G. Winberg, and R. Sandberg. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature methods*, 10(11):1096–1098, 2013.

H. A. Pliner, J. Shendure, and C. Trapnell. Supervised classification enables rapid annotation of cell atlases. *Nature methods*, 16(10):983–986, 2019.

M. Ptashne. *A Genetic Switch: Phage Lambda Revisited*. A Genetic Switch: Phage Lambda Revisited. Cold Spring Harbor Laboratory Press, 2004. ISBN 9780879697167.

R. Rabadan and A. Blumberg. *Topological Data Analysis for Genomics and Evolution: Topology in Biology*. Cambridge University Press, 2019. ISBN 9781107159549.

O. J. Rackham, J. Firas, H. Fang, M. E. Oates, M. L. Holmes, A. S. Knaupp, H. Suzuki, C. M. Nefzger, C. O. Daub, J. W. Shin, et al. A predictive computational framework for direct reprogramming between human cell types. *Nature genetics*, 48(3):331–335, 2016.

A. Raj, C. S. Peskin, D. Tranchina, D. Y. Vargas, and S. Tyagi. Stochastic mrna synthesis in mammalian cells. *PLoS Biol*, 4(10):e309, 2006.

D. Ramsköld, S. Luo, Y.-C. Wang, R. Li, Q. Deng, O. R. Faridani, G. A. Daniels, I. Khrebtukova, J. F. Loring, L. C. Laurent, et al. Full-length mrna-seq from single-cell levels of rna and individual circulating tumor cells. *Nature biotechnology*, 30(8):777–782, 2012.

W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

B. Ranjan, W. Sun, J. Park, K. Mishra, R. Xie, F. Alipour, V. Singhal, F. Schmidt, I. Joanito, N. A. Rayan, et al. Dubstepr: correlation-based feature selection for clustering single-cell rna sequencing data. *bioRxiv*, pages 2020–10, 2021.

A. Regev, S. A. Teichmann, E. S. Lander, I. Amit, C. Benoist, E. Birney, B. Bodenmiller, P. Campbell, P. Carninci, M. Clatworthy, et al. Science forum: the human cell atlas. *elife*, 6:e27041, 2017.

J. Reichardt and S. Bornholdt. Statistical mechanics of community detection. *Physical review E*, 74(1):016110, 2006.

D. Risso, F. Perraudeau, S. Gribkova, S. Dudoit, and J.-P. Vert. A general and flexible method for signal extraction from single-cell rna-seq data. *Nature communications*, 9(1):1–17, 2018.

A. H. Rizvi, P. G. Camara, E. K. Kandror, T. J. Roberts, I. Schieren, T. Maniatis, and R. Rabadan. Single-cell topological rna-seq analysis reveals insights into cellular differentiation and development. *Nature biotechnology*, 35(6):551–560, 2017.

S. J. Roberts, R. Everson, and I. Rezek. Maximum certainty data partitioning. *Pattern Recognition*, 33(5):833–839, 2000.

S. J. Roberts, C. Holmes, and D. Denison. Minimum-entropy data clustering using reversible jump markov chain monte carlo. In *International Conference on Artificial Neural Networks*, pages 103–110. Springer, 2001.

M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

O. Rozenblatt-Rosen, M. J. Stubbington, A. Regev, and S. A. Teichmann. The human cell atlas: from vision to reality. *Nature News*, 550(7677):451, 2017.

Y. Saeys, I. Inza, and P. Larranaga. A review of feature selection techniques in bioinformatics. *bioinformatics*, 23(19):2507–2517, 2007.

R. Sambasivan, B. Gayraud-Morel, G. Dumas, C. Cimper, S. Paisant, R. G. Kelly, and S. Tajbakhsh. Distinct regulatory cascades govern extraocular and pharyngeal arch muscle progenitor cell fates. *Developmental cell*, 16(6): 810–821, 2009.

A. Sarkar and M. Stephens. Separating measurement and expression models clarifies confusion in single-cell rna sequencing analysis. *Nature Genetics*, pages 1–8, 2021.

R. Satija and A. K. Shalek. Heterogeneity in immune responses: from populations to single cells. *Trends in immunology*, 35(5):219–229, 2014.

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

M. Schleiden. Arch anat physiol. *Wiss Med*, 13:137–176, 1838.

T. Schwann. Microscopic researches on the conformity in structure and growth between animals and plants. *Berlin, Germany*, 1839.

R. Sender, S. Fuchs, and R. Milo. Revised estimates for the number of human and bacteria cells in the body. *PLoS biology*, 14(8):e1002533, 2016.

C. E. Shannon. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948.

A. F. Shorrocks. The class of additively decomposable inequality measures. *Econometrica: Journal of the Econometric Society*, pages 613–625, 1980.

R. C. Smith and B. D. MacArthur. Information-theoretic approaches to understanding stem cell variability. *Current Stem Cell Reports*, 3(3):225–231, 2017.

R. C. Smith, P. S. Stumpf, S. J. Ridden, A. Sim, S. Filippi, H. A. Harrington, and B. D. MacArthur. Nanog fluctuations in embryonic stem cells highlight the problem of measurement in cell biology. *Biophysical journal*, 112(12): 2641–2652, 2017.

C. Soneson and M. D. Robinson. Bias, robustness and scalability in single-cell differential expression analysis. *Nature methods*, 15(4):255–261, 2018.

C. Soulé. Mathematical approaches to differentiation and gene regulation. *Comptes Rendus Biologies*, 329(1):13–20, 2006. ISSN 1631-0691. . Modélisation de systèmes complexes en agronomie et environnement.

B. Sparta, T. Hamilton, S. D. Aragones, and E. J. Deeds. Binomial models uncover biological variation during feature selection of droplet-based single-cell rna sequencing. *bioRxiv*, 2021.

J. W. Squair, M. Gautier, C. Kathe, M. A. Anderson, N. D. James, T. H. Hutson, R. Hudelle, T. Qaiser, K. J. Matson, Q. Barraud, et al. Confronting false discoveries in single-cell differential expression. *bioRxiv*, 2021.

S. H. Strogatz. *Nonlinear Dynamics and Chaos with Student Solutions Manual: With Applications to Physics, Biology, Chemistry, and Engineering*. CRC Press, 2018.

T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck III, Y. Hao, M. Stoeckius, P. Smibert, and R. Satija. Comprehensive integration of single-cell data. *Cell*, 2019.

P. S. Stumpf, X. Du, H. Imanishi, Y. Kunisaki, Y. Semba, T. Noble, R. C. Smith, M. Rose-Zerili, J. J. West, R. O. Oreffo, et al. Transfer learning efficiently maps bone marrow cell types from mouse to human using single-cell rna sequencing. *Communications biology*, 3(1):1–11, 2020.

J. E. Sulston, E. Schierenberg, J. G. White, and J. N. Thomson. The embryonic cell lineage of the nematode caenorhabditis elegans. *Developmental biology*, 100(1):64–119, 1983.

V. Svensson. Droplet scrna-seq is not zero-inflated. *Nature Biotechnology*, 38(2): 147–150, 2020.

V. Svensson, K. Natarajan, L.-H. Ly, R. Miragaia, C. Labalette, I. Macaulay, A. Cvejic, and S. Teichmann. Power analysis of single-cell rna-sequencing experiments. *Nature methods*, 14(4):381, 2017.

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

Tabula Muris Consortium et al. Single-cell transcriptomics of 20 mouse organs creates a tabula muris. *Nature*, 562(7727):367, 2018.

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

S. J. Tapscott, R. L. Davis, M. J. Thayer, P.-F. Cheng, H. Weintraub, and A. B. Lassar. Myod1: a nuclear phosphoprotein requiring a myc homology region to convert fibroblasts to myoblasts. *Science*, 242(4877):405–411, 1988.

B. Tasic, V. Menon, T. N. Nguyen, T. K. Kim, T. Jarsky, Z. Yao, B. Levi, L. T. Gray, S. A. Sorensen, T. Dolbeare, et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nature neuroscience*, 19(2):335–346, 2016.

A. E. Teschendorff and A. P. Feinberg. Statistical mechanics meets single-cell biology. *Nature Reviews Genetics*, pages 1–18, 2021.

The Tabula Sapiens Consortium and S. R. Quake. The tabula sapiens: a single cell transcriptomic atlas of multiple organs from individual human donors. *bioRxiv*, 2021. .

H. Theil. *Economics and Information Theory*. Studies in mathematical and managerial economics. North-Holland Publishing Company, 1967.

J. A. Thomson, J. Itskovitz-Eldor, S. S. Shapiro, M. A. Waknitz, J. J. Swiergiel, V. S. Marshall, and J. M. Jones. Embryonic stem cell lines derived from human blastocysts. *science*, 282(5391):1145–1147, 1998.

L. Tian, X. Dong, S. Freytag, K.-A. Le Cao, S. Su, A. JalalAbadi, D. Amann-Zalcenstein, T. S. Weber, A. Seidi, J. S. Jabbari, et al. Benchmarking single cell rna-sequencing analysis pipelines using mixture control experiments. *Nature methods*, 16(6):479–487, 2019.

R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

F. W. Townes, S. C. Hicks, M. J. Aryee, and R. A. Irizarry. Feature selection and dimension reduction for single-cell rna-seq based on a multinomial model. *Genome biology*, 20(1):1–16, 2019.

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

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

C. A. Vallejos, D. Risso, A. Scialdone, S. Dudoit, and J. C. Marioni. Normalizing single-cell rna sequencing data: challenges and opportunities. *Nature methods*, 14(6):565, 2017.

L. van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

G. Van Rossum and F. L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009. ISBN 1441412697.

B. Vieth, S. Parekh, C. Ziegenhain, W. Enard, and I. Hellmann. A systematic evaluation of single cell rna-seq analysis pipelines. *Nature communications*, 10 (1):1–11, 2019.

P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.

E. Voit. *A First Course in Systems Biology*. CRC Press, 2017. ISBN 9781351332941.

U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17 (4):395–416, 2007.

U. Von Luxburg, R. C. Williamson, and I. Guyon. Clustering: Science or art? In *Proceedings of ICML workshop on unsupervised and transfer learning*, pages 65–79. JMLR Workshop and Conference Proceedings, 2012.

C. H. Waddington. *The strategy of the genes*. Routledge, 2014.

C. H. Waddington et al. An introduction to modern genetics. *An introduction to modern genetics.*, 1939.

J. Wang, L. Xu, E. Wang, and S. Huang. The potential landscape of genetic circuits imposes the arrow of time in stem cell differentiation. *Biophysical journal*, 99(1):29–39, 2010.

T. Wang, B. Li, C. E. Nelson, and S. Nabavi. Comparative analysis of differential gene expression analysis tools for single-cell rna sequencing data. *BMC bioinformatics*, 20(1):1–16, 2019.

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

M. E. Wilkinson, C. Charenton, and K. Nagai. Rna splicing by the spliceosome. *Annual review of biochemistry*, 89:359–388, 2020.

B. Xia and I. Yanai. A periodic table of cell types. *Development*, 146(12): dev169854, 2019.

Y. Xue, T. C. Theisen, S. Rastogi, A. Ferrel, S. R. Quake, and J. C. Boothroyd. A single-parasite transcriptional atlas of toxoplasma gondii reveals novel control of antigen expression. *Elife*, 9:e54129, 2020.

S. H. Yip, P. C. Sham, and J. Wang. Evaluation of tools for highly variable gene discovery from single-cell rna-seq data. *Briefings in bioinformatics*, 20(4): 1583–1589, 2019.

L. A. Zadeh. Probability measures of fuzzy events. *Journal of mathematical analysis and applications*, 23(2):421–427, 1968.

L. Zappia and F. J. Theis. Over 1000 tools reveal trends in the single-cell rna-seq analysis landscape. *bioRxiv*, 2021.

G. X. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8(1): 1–12, 2017.

C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)*, 23(4):550–560, 1997.

C. Ziegenhain, B. Vieth, S. Parekh, B. Reinius, A. Guillaumet-Adkins, M. Smets, H. Leonhardt, H. Heyn, I. Hellmann, and W. Enard. Comparative analysis of single-cell rna sequencing methods. *Molecular cell*, 65(4):631–643, 2017.

A. J. Zomorodian. *Topology for Computing*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2005. .

Z. Zou, K. Hua, and X. Zhang. Hgc: fast hierarchical clustering for large-scale single-cell data. *bioRxiv*, 2021.