Decomposing scRNA-seq data using topic modelling

Alexandrina Pancheva

Submitted in fulfilment of the requirements for the
Degree of Doctor of Philosophy

Institute of Infection, Immunity and Inflammation
College of Medical, Veterinary and Life Sciences
University of Glasgow

July 2022

Abstract

In recent years, the development of single cell RNA-sequencing technologies has allowed scientists to study heterogeneity of cell populations, compare cells across conditions, analyse biological processes in development and disease, and infer cellular interactions. While single cell studies provide invaluable perspective in understanding disease and identifying therapeutic targets, such datasets are high-dimensional and pose unique challenges compared to earlier technologies. Machine learning techniques have become one of the most popular ways of overcoming those challenges. The work described here develops and applies interpretable models to single cell data. All methods described here are based on topic modelling, a popular technique within natural language processing. In this context, cells correspond to documents and genes to words. Firstly, we investigate the problem of doublet detection and assess the limitations of currently available methods. We propose an alternative approach based on topic modelling. While the proposed approach does not outperform state of the art methods, potential avenues for exploration are highlighted. Next, a topic modelling-based approach is used to detect genes that change as a result of cell-cell interactions in single cells. Experiments using synthetic and real datasets show that our approach is able to detect genes that change as a result of interaction, while also uncovering meaningful biological groups of genes that correspond to the latent topics which aids interpretation. The described approach also alleviates some of the prior information required by the previous methods, in particular ligand-receptor databases, clustering, and generation of synthetic doublets. Finally, the topic model formulation is extended to single cell data ordered in pseudotime. The dynamic topic modelling is able to capture groups of genes that change over time. This dynamic approach outperforms non-temporal topic models and standard differential expression as it detects more biologically relevant groups of genes. The final section outlines potential directions for future research.

Contents

Abstract i

Acknowledgements ix

Declaration xi

1 Introduction 1

1.1 Thesis aims . . . 2
1.2 Contributions . . . 2
1.3 Research output . . . 3
1.4 Code and data availability . . . 3
1.5 Thesis outline . . . 4

2 Single-cell RNA sequencing (scRNA-seq) 5

2.1 Introduction . . . 5
2.2 Data generation . . . 5
2.3 scRNA-seq computational analysis: overview . . . 7
2.3.1 Initial processing . . . 7
2.3.2 Quality control (QC) . . . 9
2.3.3 Normalisation . . . 9
2.3.4 Technical and biological covariates . . . 10
2.3.5 Feature selection, dimensionality reduction, and visualisation . . . 12
2.3.6 Clustering and cluster annotation . . . 13
2.3.7 Differential expression . . . 14
2.3.8 Pseudotime and trajectory inference . . . 14
2.3.9 Inferring cell-cell interaction . . . 17
2.3.10 Options for further analysis . . . 19
2.3.11 Summary . . . 20

3 Machine learning background 21

3.1 Mixture models . . . 22

3.1.1 Clustering revisited . . . 22
3.1.2 Mixture models as latent variable models . . . 22
3.2 Notes on inference . . . 24
3.3 Latent Dirichlet Allocation (LDA) and other topic modelling approaches . . . 25
3.3.1 Motivation and generative process . . . 25
3.3.2 Inference . . . 26
3.3.3 Extensions of the standard LDA . . . 27
3.3.4 Other topic modelling approaches . . . 28
3.3.5 Choosing the number of topics . . . 28
3.4 Gaussian Process (GP) . . . 29
3.4.1 Introduction to GPs . . . 29
3.4.2 Kernel functions . . . 31
3.4.3 GP regression . . . 33
3.4.4 Inference . . . 35
3.4.5 Gaussian Process Latent Variable Model (GPLVM) . . . 36
3.5 Other useful ML concepts . . . 37
3.5.1 Classification and support vector machines (SVM) . . . 37
3.5.2 Evaluating classifiers . . . 38
3.5.3 Entropy . . . 39
3.6 ML approaches within the scope of this thesis . . . 39

4 Investigating the potential of latent Dirichlet allocation for doublet detection 41
4.1 Introduction . . . 41
4.2 Computational methods for doublet detection . . . 43
4.2.1 DoubletFinder . . . 43
4.2.2 Scrublet . . . 45
4.2.3 DoubletDecon . . . 45
4.2.4 Summary . . . 46
4.3 Applications of LDA to scRNA-seq data . . . 47
4.4 Materials and methods . . . 48
4.4.1 LDA and entropy scoring . . . 48
4.4.2 Datasets . . . 48
4.4.3 Metrics . . . 50
4.4.4 Evaluating the predictive performance of different sets of features . . . 51
4.4.5 LDA and SVM implementation . . . 51
4.4.6 Analysis with DoubletFinder and DoubletDecon . . . 52
4.5 Results . . . 53
4.5.1 Demuxlet dataset . . . 53
4.5.2 Cell Hashing (PBMCs) . . . 56

4.5.3 Cell Hashing (cell lines) . . . 58
4.5.4 Downsampling and effect on performance of methods . . . 60
4.5.5 Validation of our assumptions . . . 61
4.6 Discussion and possible future directions . . . 61
4.6.1 Improving annotation of doublets . . . 61
4.6.2 Why is entropy not suitable for doublet annotation? . . . 62
4.6.3 Counts, housekeeping genes and doublets . . . 63
4.7 Conclusions . . . 64

5 Understanding cellular crosstalk in scRNA-seq using topic modelling 67
5.1 Introduction . . . 67
5.2 Materials and methods . . . 70
5.2.1 Latent Dirichlet Allocation . . . 70
5.2.2 Identifying topics linked to a cell type . . . 70
5.2.3 Choosing number of topics . . . 70
5.2.4 Motivating the need for new topics . . . 71
5.2.5 Ranking genes as potential candidates of interaction . . . 72
5.2.6 Evaluation datasets . . . 72
5.2.7 Pre-processing and analysis before LDA . . . 73
5.2.8 Running LDA . . . 74
5.3 Results and discussion . . . 74
5.3.1 Validation using synthetic doublets . . . 74
5.3.2 PIC-seq dataset . . . 76
5.3.3 BM dataset . . . 80
5.3.4 COVID-19 dataset . . . 80
5.4 Conclusion . . . 82

6 Using dynamic topic modelling to study temporal scRNA-seq data 84
6.1 Introduction . . . 84
6.1.1 Gene expression over time . . . 84
6.1.2 Adding a temporal dimension to scRNA-seq analysis . . . 85
6.1.3 Extensions of traditional LDA and applications to transcriptomics data . 85
6.1.4 GP methods in scRNA-seq . . . 86
6.1.5 Aims . . . 87
6.2 Materials and methods . . . 87
6.2.1 Dynamic Correlated Topic Model . . . 87
6.2.2 Model inference . . . 88
6.2.3 Relaxed LDA . . . 88
6.2.4 Autocorrelation of time-series . . . 89

6.2.5 Ranking genes in topics . . . 89
6.2.6 Topic interpretation . . . 89
6.2.7 Choosing interesting topics . . . 90
6.2.8 Choosing the number of topics . . . 90
6.2.9 Comparison with scRNA-seq analysis . . . 90
6.2.10 Datasets . . . 90
6.3 Results and discussion . . . 91
6.3.1 Randomised control and ordered data autocorrelation . . . 91
6.4 Model comparison . . . 92
6.4.1 Comparison with relaxed LDA and LDA . . . 92
6.4.2 Comparison with differential expression . . . 93
6.4.3 Malaria Cell Atlas . . . 95
6.5 Dendritic cells . . . 97
6.6 Model flexibility and practical considerations . . . 98
6.7 Conclusions and possible future directions . . . 98

7 Conclusions and Future Work 100
7.1 Doublet detection . . . 100
7.2 Cellular crosstalk . . . 102
7.3 Topic modelling for scRNA-seq ordered in pseudotime . . . 103
7.4 Summary . . . 104

A Investigating the potential of Latent Dirichlet Allocation for doublet detection 105

B Understanding cellular crosstalk in scRNA-seq using topic modelling 106

C Using dynamic topic modelling to study temporal scRNA-seq data 118

List of Tables

2.1 Comparative table of commonly used scRNA-seq protocols 7
2.2 Data integration methods overview 11
2.3 Overview of selected popular trajectory inference methods 16

4.1 Multiplet rate 42
4.2 Demuxlet cell annotations 49
4.3 PMBCs CellHashing doublet annotation 49
4.4 Cell lines Cell Hashing doublet annotation 50
4.5 Demuxlet results for all methods 54
4.6 DoubletFinder's performance with different doublet rates 55
4.7 Performance of methods for Cell Hashing PBMCs dataset 57
4.8 Methods' performance for Cell Hashing cell lines dataset 59
4.9 Comparing doublet types identified by each tool 59
4.10 Ratio of means between doublets and singlets 64
4.11 Ratio of means of two groups of singlets 65

6.1 Perplexity for different topic models 93

B.1 Filtering parameters COVID-19 data 106
B.2 Housekeeping and mitochondrial genes captured by second LDA 107
B.3 Genes from the reference topics 115
B.4 Top genes from the PIC-seq and the bone marrow datasets. 116

C.1 Unique GO terms in each method 119

List of Figures

2.1 Formation of a barcoded bead used in 10x Chromium. 6
2.2 Overview of the standard steps of scRNA-seq analysis. 8
2.3 Physical time vs pseudotime 15
2.4 Overview of ligand-receptor based methods 18

3.1 Mixture model at convergence. 24
3.2 Graphical model representation of LDA. 26
3.3 Univariate Gaussian. 29
3.4 Multivariate Gaussians. 30
3.5 Samples from multivariate Gaussian 30
3.6 Samples from 5-Dimensional Gaussian 31
3.7 From Gaussian distribution to a Gaussian process 31
3.8 Effect of lengthscale on the samples of a GP with RBF kernel 32
3.9 Samples from GP prior with zero mean and some popular kernel functions . . . 34
3.10 GP regression 35
3.11 Gaussian Process Latent Variable Model 36
3.12 Support vector machine illustration 37
3.13 SVM linearly and non-linearly separable data examples 38
3.14 Precision-Recall and ROC curves 39

4.1 Annotation based on HTO 42
4.2 Where do doublet detection tools fit in the single cell analysis pipeline? 44
4.3 DoubletFinder overview. 45
4.4 LDA and entropy scoring overview 49
4.5 Demuxlet PBMCs entropy and counts 54
4.6 Ground truth vs performance of each method in Demuxlet PBMCs 55
4.7 Effect of the entropy cutoff. 56
4.8 Precision-recall curves of different feature sets Demuxlet data 57
4.9 Counts and entropy in Cell Hashing PBMCs 57
4.10 Ground truth vs performance of each method in Cell Hashing PBMCs 58
4.11 Precision-recall curves of different feature sets for the HTO PBMCs 59

4.12 Comparing sequencing depth of two datasets: Demuxlet and Cell Hashing . . . 60
4.13 Downsampling of Demuxlet PBMCs data . . . 60
4.14 Entropy of synthetic doublets . . . 62
4.15 Ground truth Venn diagram . . . 63

5.1 Approach overview . . . 71
5.2 Perplexity analysis . . . 72
5.3 Synthetic data evaluation . . . 75
5.4 Synthetic data evaluation ROCs . . . 76
5.5 PIC-seq reference population topics . . . 77
5.6 Examples of top genes in PIC-seq reference . . . 78
5.8 Genes identified by the proposed approach in the PIC-seq data . . . 79
5.9 Sorted vs interacting gene expression BM data . . . 81
5.10 Doublets vs reference expression COVID-19 data . . . 82

6.1 Autocorrelations: ordered vs randomised . . . 92
6.2 DCTM vs relaxed LDA vs LDA . . . 94
6.3 DCTM vs DE genes . . . 94
6.4 Topic probabilities over time . . . 95
6.5 Top genes from time-varying topics . . . 96
6.6 Temporal correlations of topic 14 with topics 13 and 15. . . . 96
6.7 Time-varying genes in dendritic cells . . . 97

A.1 Ground truth vs performance of each method in Cell Hashing cell lines . . . 105

B.1 Cluster annotations in COVID-19 data . . . 107
B.2 Genes with unmodified expression for synthetic experiments . . . 108
B.3 Genes with modified expression for synthetic experiments . . . 109
B.4 Modified random genes synthetic experiment . . . 110
B.5 Unmodified random genes synthetic experiment . . . 111
B.7 1^{st}stage LDA with not enough topics . . . 112
B.8 ROCs for different number of topics . . . 113
B.9 AUC plot for number of topics . . . 114
B.10 Jesnsen-Shannon divergence and cosine similarity results . . . 114

Acknowledgements

First and foremost, I would like to thank my PhD supervisors, Thomas Otto, Simon Rogers, and Helen Wheadon. Thank you Thomas for always keeping me on my toes, introducing me to the wonderful world of single cell, and teaching me how to be pragmatic from time to time. Thank you Simon for introducing me to computational biology, back when I was a 4^{th} year undergraduate (and for putting up with me since 2016!). Safe to say I would have not made it here without your support, patience, and outstanding mentoring. Thank you Helen for all the cakes, coffee, and chats. Thank you also for taking me to the wet-lab for a few days! Despite what many might say, I found pipetting very therapeutic. Thank you all for all your scientific input and feedback.

I would also like to thank the MRC for making this PhD possible. Thank you for the generous funding for training and conferences.

Special thanks to Scott for being the best IT support and solving an issue or two or more. If anything, I have managed to break things in new and creative ways.

To the Otto lab: thank you for the interesting discussions.

To my PhD and post-doc pals: Fran, Fionnuala, Stephen, Ivaylo, thank you so much for listening to my rants, taking me out for coffees, and just being there for me, across the road or the other end of WhatsApp.

My PhD journey would have not been the same without my side projects, sometimes easing the stress of PhDing but also making it worse at times. To the School of Computing Science, thank you for giving me the opportunity to tutor. It has been a pleasure! To Computer Science Academy Africa (CSAA), thank you for giving me the opportunity to share my passion for machine learning and teaching. It was a privilege meeting our incredible participants, in Rwanda and Nigeria. To the Computational Biology group, thanks for the pre-pandemic regular sessions, thanks to some of you for joining writing retreat Mondays and reading over bits of this thesis, and finally for joining me in organising a conference! Not once but twice now! Wasn't it fun! So glad to see what was an idea turning into a great success. Special thanks to Vinny, Jake, Kieran, Olympia, Joe, Grimur, just to mention a few.

Thank you to some of my closest friends, Ruth and Daniel, for the gin, chats, walks, and encouragements.

To the furry family and friends: Liza, Mecho, Rory, Lexi, Misty, and Iris for providing cuddles

when needed. In particular, thank you Liza for being my longest study companion.

Thank you to Rob for being there for me, believing in me, helping me proofread all the words and just thank you for being amazing. Thank you for making this thesis sound fun.

To my mum, thank you for being the best mum and I do not want to hear anything less! I cannot imagine a more supportive parent and I am truly grateful for everything you have done and continue to do. Thank you for teaching me to stand up for myself! To my granny, thank you for teaching me to work hard and be the best version of myself. To my grandpa, miss you always and hope you are proud of me.

Declaration

I declare that, except where explicit reference is made to the contribution of others, that this dissertation is the result of my own work and has not been submitted for any other degree at the University of Glasgow or any other institution.

Alexandrina Pancheva, July 2022

Chapter 1

Introduction

Gene expression studies aim to detect and quantify the expression levels of messenger RNA (mRNA) of genes. Technologies that permit the measurement of gene expression have changed over time, with bulk RNA sequencing (RNA-seq) replacing microarray in the early 2000s, followed by single cell RNA-seq (scRNA-seq) which has been continuously gaining popularity (Emrich et al. 2007). Microarrays allow for profiling of predefined transcripts/genes (Slonim & Yanai 2009). Bulk RNA-seq generates an averaged expression for each transcript within a sample. Bulk transcriptomics have been applied to biological studies to identify differences between conditions. scRNA-seq enables the comparison of individual cells on transcriptomic level and it has been extensively used to study heterogeneity of cell populations since the first published study in 2009 (Tang et al. 2009).

Beyond studying heterogeneity of cell populations and identifying rare cell types, scRNA-seq can be used to study biological processes: cells can be ordered according to how much progress they have made through a process and assigned a pseudotime (Trapnell et al. 2014, Yang et al. 2019). In addition to the unique opportunities of single cell, some challenges known from bulk remain, in particular joint modelling of multiple omics approaches, studies of cell-cell interactions, and others. scRNA-seq datasets are often high-dimensional (Luecken & Theis 2019, Lähnemann et al. 2020). Depending on the sequencing protocol, datasets spanning millions of cells can be generated. The human genome contains over 20 000 protein coding genes. While not all genes are captured, the datasets still can have tens of thousands of samples (cells) and tens of thousands of features (genes).

Machine learning (ML) encapsulates a range of methods that allow us to learn from data. Machine learning has been applied to images, text, sounds, patients records, and other scenarios. Despite the range of data types, some of the tasks we would like to use machine learning for are similar: grouping similar objects together, such as text or images; predicting what the state of the object would be at a future timepoint having access to historical records; classifying objects based on features; reducing dimensionality, or others. Some of those problems are also shared by scRNA-seq data. Later sections will provide an overview of what further challenges beyond

data dimensionality exist for the analysis of scRNA-seq that make it an exciting avenue for the development and application of machine learning techniques.

With the increasing complexity of machine learning methods and the advent of deep learning, there is an ongoing interest in developing interpretable machine learning models. While "interpretability" can be considered a poorly defined concept with fluid definitions, here we define interpretability in terms of descriptive accuracy and relevancy (Murdoch et al. 2019). In the context of biology, can we extract biologically relevant information without sacrificing performance?

1.1 Thesis aims

As this thesis focuses on method development and addresses three distinct problems, the main aims can be formulated as follows:

Aim 1: Develop interpretable models for scRNA-seq data.

Aim 2: Relax some of the assumptions of state of the art methods that are either unrealistic or make them difficult to use in practice.

Aim 3: Use prior information where appropriate to improve biological insight.

Specifically, the first aim is addressed throughout this thesis. Aim 2 is addressed in the methods proposed in Chapters 4 and 5. Finally, aim 3 is the basis of the work proposed in Chapter 6.

1.2 Contributions

Overall thesis contributions are described below.

- A method based on topic modelling for doublet detection is proposed and evaluated based on synthetic and real data. The proposed approach is compared with state of the art methods in a comprehensive benchmarking study (**Chapter 4**). While the results of that chapter do not outperform current methods, it sets the context of using topic modelling in the area of single cell and sets the scene for the work in Chapter 5 which is based on situations when doublets are useful for studying cell-cell interaction.

- A novel method based on topic modelling for detecting genes that change their expression as a result of cell-cell interaction is discussed in **Chapter 5**. The proposed approach addresses the prior information requirements of previous work, such as clustering assignment and synthetic doublet creation.

- **Chapter 6** proposes the application of dynamic correlated topic models (DCTM) to temporal single cell data. This is the first application of dynamic topic models to single cell

data. Our results show that taking time into account allows for more interpretable topics and we identify further sets of GO terms.

1.3 Research output

Publications

Pancheva, A., Wheadon, H., Rogers, S. & Otto, T. D. (2022), 'Using topic modeling to detect cellular crosstalk in scRNA-seq', *PLOS Computational Biology* **18**(4), e1009975. Publisher: Public Library of Science.

Selected presentations

Poster and lightning talk *Using latent Dirichlet allocation for detecting doublets in scRNA-seq data* Single Cell Biology 2020

Poster and full talk *Using topic modeling to detect cellular crosstalk in scRNA-seq* ISMB/ECCB 2021

Full talk *scRNA-seq: opportunities and challenges* Biochemical Society Webinar 2021

Talk *Understanding cell-cell communication* RECOMB-SEQ and RECOMB-CCB 2022 Science communication session

Poster *Using dynamic topic modeling to study temporal scRNA-seq data* RECOMB2022

1.4 Code and data availability

All code is available on GitHub. All datasets used for analysis in this thesis are publicly available. Each chapter contains a section on datasets that provides links to where data have been acquired from.

- Chapter 4: https://github.com/alexpancheva/doubletsAnalysis
- Chapter 5:https://github.com/alexpancheva/ldapaper
- Chapter 6: https://github.com/alexpancheva/sc-DCTM

1.5 Thesis outline

The remainder of this thesis is structured as follows:

- **Chapter 2** introduces scRNA-seq and the computational analysis that follows data generation. The chapter describes commonly used scRNA-seq protocols. Next, the main steps of a traditional single cell analysis are outlined. Commonly established good practices for analysis of scRNA-seq data are highlighted. Challenges and open problems are discussed. If familiar with the field of scRNA-seq and state of the art technologies, the reader can go to Chapter 3.

- **Chapter 3** introduces the machine learning background required for this thesis. In particular, topic modelling and Gaussian processes are discussed. ML readers can go straight to the results chapters, Chapters 4, 5, 6.

- **Chapter 4** introduces the problem of doublet detection in single cell data. It explores the application of topic modelling to detecting doublets in scRNA-seq. The proposed approach is evaluated on simulated and real data with annotation available for some doublets. The proposed approach is compared with state of the art doublet detection methods.

- **Chapter 5** focuses on identifying cell-cell interactions without relying on existing databases. This chapter presents a method for detecting genes that change as a result of interaction in scRNA-seq data. Evaluation is performed on both simulated and real data. Real datasets cover protocols that enable the capture of interacting cells, such as PIC-seq, as well as the more widely used scRNA-seq protocol 10x Chromium.

- **Chapter 6** presents the application of an extension of the standard topic modelling framework, dynamic correlated topic model, to scRNA-seq data ordered in pseudotime. DCTM performance is compared with non-temporal topic models. DCTM is able to uncover more relevant gene groups compared to other topic models and standard differential expression analysis.

- **Chapter 7** summarises the work and contributions. It also highlights avenues for future work from both a machine learning perspective and single cell experimental setup.

Chapter 2

Single-cell RNA sequencing (scRNA-seq)

As this thesis focuses on method development for a particular type of data, scRNA-seq, this chapter covers data generation to illustrate to the reader some of the challenges when it comes to these datasets, specifically dimensionality, sparsity, and noise. Next, an overview of main analysis steps is presented. It will become evident from the next sections, scRNA-seq data science is a field with rapid tool development, and as such where appropriate commentary of benchmarking results is provided. Finally, open problems are discussed.

2.1 Introduction

Since the first scRNA-seq study in 2009, there has been a rapid increase in the popularity of measuring the transcriptomes of single cells (Tang et al. 2009). scRNA-seq allows for the comparison of individual cells on transcriptomic level. Therefore, one of the major applications of scRNA-seq has focused on studying the heterogeneity of cell populations: examples include immune cells, cancer cells, and transcriptional variation in parasites (Reid et al. 2018, Yang et al. 2019, Yeo et al. 2020). In addition to resolving heterogeneity in cell populations and discovering rare cell types, scRNA-seq has been applied to study developmental and disease processes and to infer cell-cell interactions. To date, there are over 1027 computational tools for scRNA-seq analysis, with that number expected to increase to over 5000 in the next few years (Zappia & Theis 2021).

In the next sections the protocols for generating the data are described, and the computational approaches for analysing scRNA-seq are outlined, covering: initial processing, quality control, clustering, and inference of cell-cell interactions.

2.2 Data generation

The scRNA-seq pipeline begins with isolation of single cells. Initially, cells were isolated by microdissection or pipetting, however high-throughput experiments use fluorescence-activated

cell sorting (FACS) or droplet emulsions. In SMART-seq2, following FACS cells are dropped into 96 or 384 well plates (Baran-Gale et al. 2018). For platforms such as Drop-seq, InDrop, and Chromium, flows of reagents and cells are combined. This combined flow is separated into droplets by adding oil at set intervals (Svensson et al. 2018). This process is shown in Figure 2.1. To ensure a single droplet contains only one cell, the flow is calibrated and the creation of droplets is controlled. However, sometimes based on the number of cells sequenced a single droplet can contain multiple cells or a cell and some ambient RNA (Svensson et al. 2018). Cell capture and isolation are followed by cell flysis. Poly(T) oligonucleotides allow for capture of poly(A)-tailed RNA which means other abundant RNA such as rRNA and tRNA are excluded. Post RNA capture, RNA is reverse-transcribed into complementary DNA (cDNA) (Haque et al. 2017, Svensson et al. 2018). This is when single-cell-specific barcodes are added to the poly(T) oligonucleotides, this process is known as multiplexing. With multiplexing, multiple samples can be pooled together in a cost-effective manner. The random sequences added to the poly(T) oligonucleotides serve as unique molecular identifiers (UMIs). UMIs are used to distinguish between copies of the same mRNA molecule and reads from separate mRNA molecules transcribed from the same gene. To increase the probability of measuring cDNA, it can be amplified by polymerase chain reaction (PCR) or *in vitro* transcription (IVT). UMIs can be used to correct for amplification bias and other technical noise (Haque et al. 2017, Svensson et al. 2018). Following amplification, the cDNA is fragmented prior to library preparation (Haque et al. 2017, Svensson et al. 2018). Following sequencing, read data go through quality control and alignment to produce count data (Luecken & Theis 2019). The computational analysis steps using the count data are described in Section 2.3.

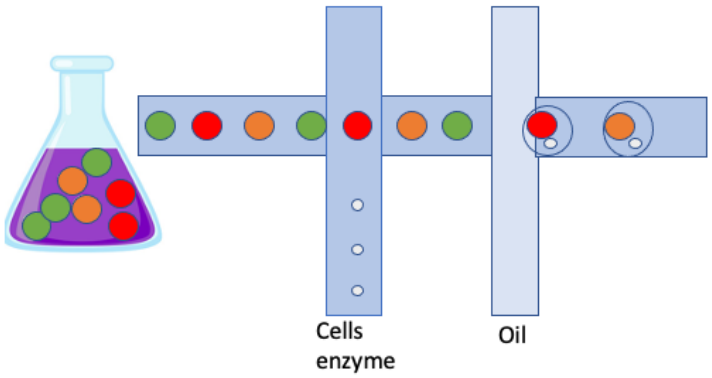


Figure 2.1: Formation of a barcoded bead used in 10x Chromium. Cells and reagents are combined in droplets. The oil droplets separate the flow of cells. Lysis and reverse transcription are performed inside the bead.

scRNA-seq protocols can be split into full-length and 3'/5'- end. Full-length sequencing provides full-length transcript data, while other methods count the 3'/5'- end. The choice depends on the goals of the experiment. Full-length sequencing allows for detection of alternative splicing

and understanding genetic alterations, such as single nucleotide polymorphisms (Baran-Gale et al. 2018). 3'- or 5'- end sequencing can be considered more cost-effective as full-length protocols do not allow for UMIs to be included, and as such library preparation is slower and more expensive (Baran-Gale et al. 2018). Table 2.1 summarises some of the popular scRNA-seq protocols and their characteristics.

Protocol	Transcript Data	Platform	Amplification	Throughput	Reference
C1 Fluidgm	full-length	Microfluidics	PCR	$10^2 - 10^3$	(Pollen et al. 2014)
Smart-seq2	full-length	Plate-based	PCR	$10^2 - 10^3$	(Picelli et al. 2013)
MARS-seq	3'-end	Plate-based	IVT	$10^2 - 10^3$	(Jaitin et al. 2014)
10x Chromium	3'/5'-end	Droplet	PCR	$10^3 - 10^4$	(Zheng et al. 2017)
Drop-seq	3'-end	Droplet	PCR	$10^3 - 10^4$	(Macosko et al. 2015)

Table 2.1: Comparative table of commonly used scRNA-seq protocols. Choice of protocol for analysis is dependent on the research question at hand.

With its ever-growing popularity, application of scRNA-seq goes beyond resolving heterogeneity in cell populations and uncovering new cell types. In recent years scRNA-seq has been applied to multiple patient samples to uncover differences between conditions: Is there an expansion of a particular cell type? How do signalling and cell-cell interaction change in the presence of disease? There is an increasing interest in understanding complex biological processes and what drives commitment to a particular lineage (Teo et al. 2019). The next section provides an overview of the computational methods that facilitate uncovering those biological insights from scRNA-seq data.

2.3 scRNA-seq computational analysis: overview

2.3.1 Initial processing

Raw data generated from sequencing machines (in the form of FASTQ files) need to be processed to obtain read count matrices or UMI counts, depending on the protocol that is used.

Pipelines such as CellRanger handle alignment, quantification, demultiplexing (assigning reads to the correct barcode) and quality control of reads. The resulting matrix is of the form of barcodes by number of transcripts. It is important to note that a single barcode does not always correspond to a single cell, as a barcode might correspond to an empty droplet or multiple cells.

While the noise levels differ in read counts data and UMI counts data, the steps we are going to discuss here are similar, and as such we will for simplicity refer to the matrix we will be using

in the next steps as counts data.

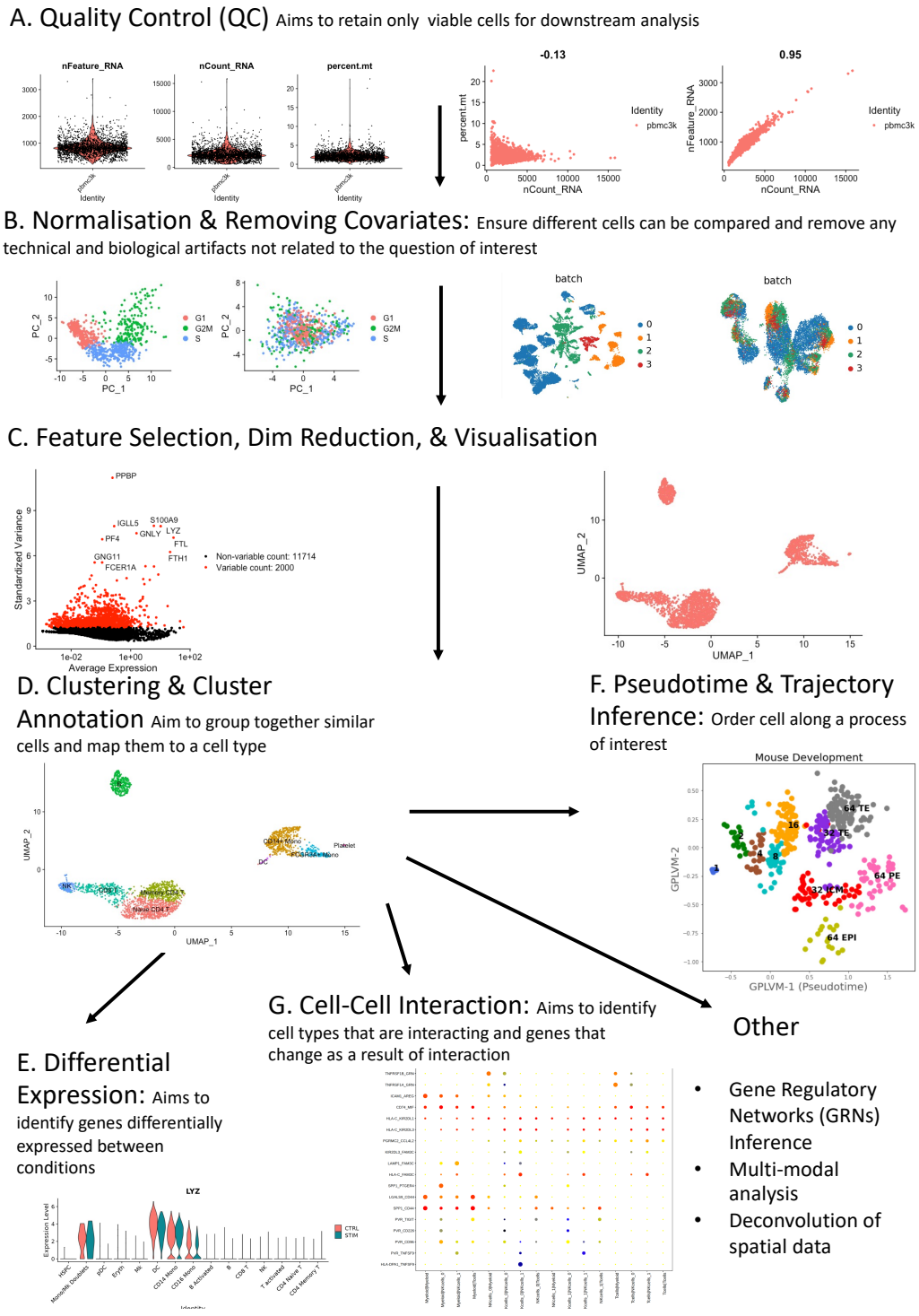


Figure 2.2: Overview of the standard steps of scRNA-seq analysis. **A.** QC plots based on Seurat analysis using total counts, total number of genes expressed, and percentage mitochondrial genes as metrics **B.** Covariates (biological): regressing cell cycle effect (Seurat). Batch correction done with BBKNN as part of scanpy. **C.** Feature selection, dimensionality reduction also via Seurat. **D.** Clustering and cluster annotation. **E.** Differential expression; **F.** Pseudotime & trajectory inference: GrandPrix; **G.** Cell-cell interaction: CellPhoneDB

2.3.2 Quality control (QC)

Before performing any downstream analysis, data should be QCed and only viable cells retained. Filtering is usually based on three metrics: total counts per barcode, number of genes per barcode, and percentage of mitochondrial genes, as shown in Figure 2.2A. Counts data are filtered by thresholding on those metrics. Specifically, a high percentage of mitochondrial genes might correspond to dying cells. On the other hand, cells with high total counts and high number of detected genes can correspond to doublets, which are multiple cells captured inside the same droplet (Luecken & Theis 2019). However, assuming high counts correspond to doublets is an oversimplification, and as such methods have been developed to avoid over-filtering that results in loss of cells. In-depth discussion of how doublet detection can be done without only relying on total counts per barcode and number of unique genes per barcode can be found in Chapter 4.

In addition to doublets, counts can be contaminated by mRNA released in the cell suspension, known as ambient RNA (Yang et al. 2020). Ambient RNA can be captured inside a droplet with a viable cell and it can be amplified together with the cell's own RNA. Ambient RNA causes problems in downstream analysis, especially in clustering, marker gene identification, and differential expression (Yang et al. 2020).

When filtering counts, all three metrics referred to above should be considered with care, as those metrics can potentially also have biological significance. For example, cells with low counts and low number of genes expressed can correspond to quiescent populations. High percentage of mitochondrial genes can be an artefact of sample quality (e.g. dying cells) or can be linked to specific respiratory processes, for example. Cells with high counts are not necessarily doublets, they might correspond to larger cells: for example, macrophages are much bigger than monocytes (Luecken & Theis 2019).

In addition to thresholding on the metrics mentioned above, filtering is often done based on how many cells express a gene, e.g. all genes expressed in fewer than 20 cells are filtered (Luecken & Theis 2019). This is another user-defined parameter which is dependent on dataset, similar to other QC metrics. QC-ing the data can be considered an iterative process and further steps of the analysis can indicate whether filtering needs adjusted, or perhaps doublets can be quantified independently of filtering thresholds (DePasquale et al. 2019, McGinnis et al. 2019, Wolock et al. 2019).

2.3.3 Normalisation

As differences can arise due to sampling cells, count depth can differ even in identical cells, so to ensure gene expression is comparable within the same sample, data should be normalised. A method adopted from bulk RNA-seq applied to scRNA-seq is CPM, counts per million: feature expression for each cell is normalised by the total expression and multiplied by a scaling factor, which is a power of 10. Another simple method for dealing with differences in counts is

downsampling the data to a pre-specified number of reads or counts (Luecken & Theis 2019). However, downsampling can increase the sparsity of the data and it does not take into account the heterogeneity in the cell population captured by scRNA-seq (Luecken & Theis 2019).

The single cell specific sources of variability have encouraged development of methods that take into account dropout in single cell (i.e., many zeros due to sampling), common in some scRNA-seq protocols such as Smart-seq2 (Cole et al. 2019). While we briefly mention dropout and how some methods aim to impute missing data in scRNA-seq in the next section, it is important to highlight a few things here. There is a common belief that droplet-based scRNA-seq data are zero-inflated, and some statistical models take the opportunity to model scRNA-seq data as a zero-inflated negative binomial (Li & Li 2018). However, Svensson has shown the number of zeros in the data is consistent with what is expected from distributional models of molecule sampling counts (Svensson 2020). Svensson performs experiments using negative-control data for several droplet protocols (Svensson 2020). While there is no negative-control data available for plate-based methods, a study looked into simulating scRNA-seq data and found that plate-based data required zero inflation to be modelled successfully, while negative binomial is a sufficient choice for droplet-based methods (Choi et al. 2020, Svensson 2020, Vieth et al. 2019).

In addition to the global scaling methods, non-linear methods might present a better alternative to normalising complex single cell data (Cole et al. 2019).

Normalisation can also be performed over genes, similarly to how it is performed to make cellular data comparable. However, currently there is no consensus whether normalisation over genes is necessary (Luecken & Theis 2019).

Gene expression data are often $log_e(x+1)$ transformed, which mitigates the mean-variance relationship in single cell and approximates the data to what the assumption of many downstream analysis methods is, i.e. normally distributed data (Luecken & Theis 2019).

2.3.4 Technical and biological covariates

A common technical covariate (variation due to technical factors that can confound biological insight) that needs to be specifically addressed when working with single cell data is batch effect due to handling cells in different environments, Figure 2.2B. Batch effects can occur between cells within the same experiment, between samples in the same lab, and samples across multiple labs (Luecken & Theis 2019). In the early days of scRNA-seq, when applying correction to cells within the same sample or several samples in the same experiment, methods adopted from bulk RNA-seq have been used, such as ComBat (Johnson et al. 2007). Some linear scRNA-seq specific methods have been developed, such as Harmony (Korsunsky et al. 2019). However, as it is becoming more popular to integrate multiple datasets which might not have the same composition, specific non-linear data integration methods have been applied to scRNA-seq data such as Canonical Correlation Analysis (CCA), Mutual Nearest Neighbour (MNN), and batch balanced k-nearest neighbours (BBKNN) (Butler et al. 2018, Haghverdi et al. 2018, Polański

et al. 2020). Due to the number of available tools for data integration, benchmarking studies have been performed to guide the community in their choice of suitable methods. One such example is the work of (Luecken et al. 2020). In their study integration methods are reviewed based on simulated and real datasets with different levels of complexity (multiple levels of batch effects) as well as scalability and usability. In this context usability evaluation is based on whether the tool is open source, response to GitHub issues, access to a tutorial, and whether the paper assessed robustness and accuracy (Luecken et al. 2020). Some of the findings are summarised below in Table 2.2 (Luecken et al. 2020).

Method	Approach	Benchmarking summary
CCA (Seurat v3)	Uses CCA to construct shared subspace between the batches; identifies mutual nearest neighbours across datasets (anchor points). Projection is inferred from the anchor points to integrated the datasets in a common hyperplane.	Good performance on simulations. Poor performance when it comes to conserving cell cycle variance and trajectory structure. Performs well on simpler tasks. High usability score.
MNN	Detects mutual nearest neighbours in two datasets and projects the second dataset onto the first one.	Conserves cell cycle variance and trajectory structure. Scalability problems for more than 100 000 cells.
BBKNN	Computes k-nearest neighbour (KNN) graph within each sample/batch, repeats the KNN computation for all cells between the different batches. Finally computes connectivity score between pairs of cells.	Performs well on real complex datasets. High scalability and usability scores.
Harmony	Takes PCA embedding of cells and their batch assignments. Until convergence, the method iterates over two stages: maximum diversity clustering and batch correction.	Good performance on simulations which display strong batch effect. Performs poorly when it comes to conserving cell cycle variance and trajectory structure. High usability score.

Table 2.2: Overview of some methods used for integrating samples from different experiments and a summary of how those methods performed when tested on different synthetic and real datasets (Luecken et al. 2020).

Another source of technical variation can be count depth which is often regressed out, similarly to biological covariates, discussed below. Handling differences in count depth can also be done at the normalisation step as discussed earlier. While dropout can also be considered a technical covariate and imputation methods have been developed. However, they can introduce false signal (Andrews & Hemberg 2019). Therefore in the datasets discussed in this thesis, imputation

methods have not been used as part of the processing. Discussion of available imputation methods for scRNA-seq is outside the scope of this thesis.

To focus on the biological signal of interest, other sources of biological variation should be isolated. A common example is the cell cycle which is often regressed out (linear regression against cell cycle score). Regressing out cell cycle can improve interpretation of other biological processes researchers might be interested in (Buettner et al. 2015). However, in some cases cell cycle can be informative of the underlying biology (Luecken & Theis 2019). Therefore, regression of biological covariates should be done with care.

While the focus of this section is to discuss computational methods for removing technical and biological covariates, it is important to also mention existing laboratory-based techniques which minimise batch effects. Examples of methods that allow for multiple samples to be sequenced together and then demultiplexed include Cell Hashing and Demuxlet (which uses genetic variation to determine the identity of each droplet). (Kang et al. 2018, Stoeckius et al. 2018).

2.3.5 Feature selection, dimensionality reduction, and visualisation

Due to the high dimensionality of single cell data, often only highly variable genes are retained for dimensionality reduction, Figure 2.2C. Highly variable genes are selected based on variance-mean ratio. Typically between 1000 and 5000 highly variable genes are selected depending on the complexity of the data (Luecken & Theis 2019).

Principal component analysis (PCA) is a linear dimensionality reduction technique where each principal component is a linear combination of variables in the original space. Unlike other dimensionality reduction methods, PCA is not as suitable for global visualisation of the data. However, it is often a fundamental step prior to clustering or trajectory inference methods (Luecken & Theis 2019). Generally, the top N principal components (PCs) are taken forward for downstream analysis tasks. Those are usually the PCs that preserve most of the variance. In practical analysis settings the way this is done is either by exploring an "elbow" plot (variance vs ranking of PCs plot) or by permutation-test-based jackstraw method (Chung & Storey 2015).

The two most commonly used dimensionality reduction techniques for visualisation of single cell data are t-distributed stochastic neighbour embedding (t-SNE) and Uniform Approximation and Projection (UMAP). UMAP scales well with sparse high-dimensional data and compared to t-SNE, it is better at preserving both global and local structures of the data (McInnes et al. 2020). As such, it is recommended and better practice to use UMAP for visual exploration of scRNA-seq data (McInnes et al. 2020).

An alternative to the methods mentioned above that might be suitable for dimensionality reduction and downstream tasks is an extension of the Gaussian process latent variable model (GPLVM) (Verma & Engelhardt 2020). The noise is modelled by Student's t-distribution and a weighted sum of non-smooth covariance functions is introduced. tGPLVM can be fit on raw

scRNA-seq data, improve clustering and cell type identification, and reconstruct an informative trajectory ordering (Verma & Engelhardt 2020).

2.3.6 Clustering and cluster annotation

Clustering is a common step when analysing single cell data to facilitate the identification of rare cells and understand heterogeneity of cell populations, Figure 2.2D (Reid et al. 2018, Yang et al. 2019, Yeo et al. 2020). The two most commonly-used and well-known tools for single cell analysis, Seurat (written in R) and scanpy (written in Python), both use detection of communities in K-nearest neighbour (KNN) graphs as their default clustering method. In this graph, cells are nodes and edges are computed based on Euclidean distance in PCA space. Edges are assigned weights based on Jaccard similarity, similar cells have high Jaccard similarity. Next, an iterative clustering algorithm is used, known as Louvain clustering. The Louvain algorithm is a modularity optimisation algorithm for clustering. Modularity describes the density of connections within a cluster. The algorithm finishes when the maximum modularity is reached.

Clustering is followed by marker genes, genes specific to a cell type, identification for each cluster and annotation. Sometimes the identified clustering solution might not reflect the underlying biology: data might be overclustered (more clusters than underlying cell types) or some cell subtypes might be put together when in fact heterogeneity is of interest. Cluster annotation is a tedious task as in most cases it is manual and relies on literature searches and user expertise (Abdelaal et al. 2019).

There have been attempts to automate cluster annotation. (Abdelaal et al. 2019) present a comprehensive overview of 22 single cell specific and general purpose classifiers and benchmark them using 27 publicly available scRNA-seq datasets. There also has been focus on developing well-annotated references for a range of tissues, for example the Human Cell Atlas (HCA), lung cell atlas, and mouse cell atlas (Almanzar et al. 2020, Regev et al. 2018, Travaglini et al. 2020). User data can be mapped onto the reference to facilitate annotation. Examples include supervised PCA and scArches, a transfer learning based approach to annotation (Barshan et al. 2011, Lotfollahi et al. 2020). The transfer learning approach by (Lotfollahi et al. 2020) relies on model sharing following training on a reference dataset. Fine tuning can be performed based on the query dataset (Lotfollahi et al. 2020).

A recent study done of published tools for scRNA-seq shows several trends in method development. The work of (Zappia & Theis 2021) demonstrates that there is an interest in developing methods for integration and classification to be able to bypass clustering and annotation steps. This also reflects the availability of scRNA-seq datasets that could be used for reference mapping.

2.3.7 Differential expression

A common question of interest in biological studies is to identify differentially expressed genes between conditions. Novel methods are being developed which are specifically designed to model the particular features of scRNA-seq data, as well as methods which are adopted from analysis of bulk RNA-seq studies (Soneson & Robinson 2018).

To better understand the strengths, weaknesses and potential bias in methods for differential expression, several studies have benchmarked methods derived from bulk and single cell specific approaches using synthetic and real datasets. The comparative analysis showed that methods designed for scRNA-seq data do not significantly outperform methods previously available from bulk RNA-seq (Soneson & Robinson 2018, Wang et al. 2019). Those studies also highlight the lack of agreement of the identified differentially expressed genes (Soneson & Robinson 2018, Wang et al. 2019).

With the increasing prevalence of multi-patient multi-condition studies, it is vital to account for variations between biological replicates to avoid false discoveries. However, the most widely used pipelines for scRNA-seq analysis utilise methods prone to false discoveries. In a recent study (Squair et al. 2021) compare methods that consider the gene expression of individual cells and methods that aggregate the cells within a biological replicate ("pseudobulk") before applying a statistical test. The study highlights that generally pseudobulk methods have better performance compared to single cell methods. Furthermore, single cell DE methods are biased towards highly expressed genes (Squair et al. 2021).

Using an unsuitable statistical method can lead to false discoveries and compromise biological insight. Along with developing methods that take biological variability into consideration, the field also requires real datasets with known grown truth and appropriate test suites that facilitate the evaluation of said methods. Generalising experimental results requires biological replicates. However, current state of the art differential expression approaches do not take sample variability into account. As there is an increasing availability of multi-sample studies across conditions, it is vital for more work to be done to address this open problem in the area (Zimmerman et al. 2021).

2.3.8 Pseudotime and trajectory inference

Single cell data allow for analysis of disease and developmental processes. However, to date, given the nature of the widely-used single cell quantification methods, the gene expression profile of the same cell over time cannot be tracked as cells are destroyed during library preparation.

One way of understanding gene expression over time is taking repeated measurements at different time points or sequencing a population sample of cells at different states. Even when capturing cells at the same time point, cells will not be in the same state and they can be more transcriptionally similar to cells at later points. This is where the idea of pseudotime fits: cells are ordered based on their progression through a biological process of interest. This continuous

representation of cells is described as a trajectory that can be linear or branching depending on the underlying biological process. Figure 2.3 illustrates the difference between physical time and pseudotime.

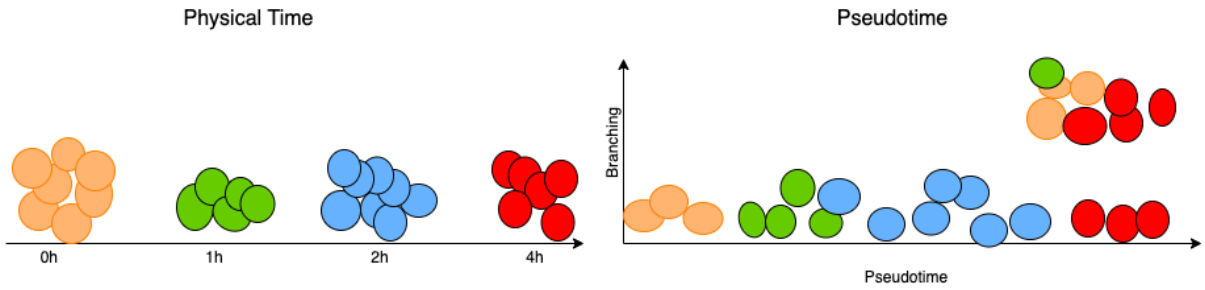


Figure 2.3: Illustrating the difference between physical time (cell capture time) and the notion of pseudotime. The pseudotime plot assumes the first dimension captures time and the second dimension can be interpreted as branching.

To date there are over 70 available pseudotime methods (Saelens et al. 2019). A recent benchmarking study compared 45 of those methods, using real and synthetic datasets (Saelens et al. 2019). Metrics used in this investigation include ordering of cells, reconstructed topology, scalability, and usability (Saelens et al. 2019). Usability scoring is based on a range of tool development criteria (e.g. open source, evaluation, testing), tutorial information, and quality of the documentation. The authors conclude there is no method that outperforms others for all types of trajectories. Some methods are more suitable for linear trajectories while others are better at reconstructing more complex, branching ones (Saelens et al. 2019). For example, the authors find that for simple linear trajectories Slingshot is the method that outperforms others, while PAGA is more suitable for complex trajectories (Street et al. 2018, Wolf et al. 2019). A summary of a selection of the methods reviewed by (Saelens et al. 2019) can be found in Table 2.3. Despite the emergence of numerous methods for pseudotime and trajectory inference, some challenges remain. For example, methods can underestimate or overestimate the complexity of the real biological trajectory. Furthermore, the challenge of scalability remains with the increasing sizes of scRNA-seq datasets. Finally, methods should be able to produce stable predictions.

Once cells are ordered in pseudotime and the trajectory is reconstructed there is interest in identifying genes that change their expression over the course of the biological process. Such genes are considered to be differentially expressed over time. Furthermore, in cases where there are multiple lineages in the reconstructed trajectory of the process of interest, genes specific to each lineage can be identified, or differential expression between or within lineages can be performed (Van den Berge et al. 2020). Furthermore, the reconstructed pseudotime can be the basis of identifying a branching dynamic of genes (Boukouvalas et al. 2018).

Method	Approach	Benchmarking summary
Monocle 2	Features can be selected by finding differentially expressed genes between clusters. A spanning tree is constructed using the centroids of the data. With the tree learned, the root node is specified and pseudotime for each cell is calculated based on distance from root.	Performs better on datasets with complex trajectories. Struggles with producing stable results.
Monocle 3	Can learn multiple disjoint trajectories. Prior knowledge is incorporated as the user specifies root nodes. The graph is split into number of subgroups with different subgroups not allowed to be part of the same trajectory.	Similarly to Monocle 2 performs better on more complex trajectories.
Slingshot	Does not require number of lineages to be pre-defined. Two step approach. Step 1 includes clustering and building a minimum spanning tree of clusters. Clusters are then ordered; all lineages share a root but have a unique terminal cluster. Step 2 infers the pseudotime for each lineage by fitting principal curves on each lineage. Prior knowledge can be incorporated in selecting the root and terminal states.	Generally better for dataset with simpler topologies. Excellent usability score. Good at placing cells on correct branch.
PAGA	A KNN graph is constructed using the UMAP representation. A degree of connectivity is calculated between different partitions, e.g clusters.	Broad trajectory type range. Results not as stable compared to Slingshot.

Table 2.3: Overview of selected popular methods for trajectory inference and summary of how they have been found to perform in a benchmarking study by (Saelens et al. 2019).

When discussing process dynamic and transitioning cells, the concept of RNA velocity cannot be omitted. So far, when considering the pseudotime and trajectory inference, cells are ordered in time and by connecting the cells in time process direction is inferred. However, standard pseudotime does not fully consider direction and speed of transition, and this is where RNA velocity comes in. RNA velocity is based on distinguishing newly transcribed RNA, i.e. unspliced mRNA (containing reads from introns) and mature mRNA, i.e. spliced. The difference in spliced and unspliced abundance allows for a metric for change in gene expression to be derived. Speed and direction of change are aggregated across all genes in a cell to arrive at the concept of RNA

velocity. Since the concept of RNA velocity has been introduced to single cell (La Manno et al. 2018), work has been done to improve on the modelling assumptions (Bergen et al. 2020). While current velocity frameworks have improved reliability, modelling of genes is still decoupled and they are considered to be independent. In biology this is not the case and a potential improvement would allow gene regulatory information to be included.

Finally, while the focus of this section has been computational methods for inferring pseudotime and trajectory inference, it is important also to mention advances in lab based techniques that aim to solve the problem. Techniques that combine scRNA-seq with metabolic RNA labelling and biochemical nucleoside conversion have been developed. There are several protocols able to deliver time-resolved scRNA-seq using metabolic labelling (Erhard et al. 2019, Qiu et al. 2020). Those approaches can be used to study transitions and perturbations. The RNA dynamics can be inferred by considering unspliced and spliced RNA. That is where the idea of metabolic labelling fits in: the old and new, nascent RNA are measured in a controlled fashion. The idea of chemical conversion relies on introducing T to C mutation. When data are analysed if there are no conversions of U-to-C that implies "old" RNA, while U-to-C conversions indicate "new" RNA. The metabolic labelling thus allows for a more controlled approach that overcomes some of the challenges of traditional quantification of spliced and unspliced RNA, specifically inaccuracies of intronic reads. Furthermore, the RNA velocity equations are scaled by the splicing rate and as such it lacks physical interpretation, molecules per hour (Qiu et al. 2022).

In addition to metabolic labelling, RNA timestamp presents a way for incorporating temporal information in standard scRNA-seq protocols. RNA timestamp is based on a recorder RNA motif where age is estimated based on accumulation of A-to-I edits (Rodriques et al. 2020).

While useful concepts, the previously described lab-based techniques do not scale well, and at present they are not widely used for generation of scRNA-seq data with a truly temporal aspect. As such, pseudotime and trajectory inference remain a common step in scRNA-seq analysis.

2.3.9 Inferring cell-cell interaction

Cell-cell interactions are vital for numerous biological processes including development, differentiation, and response to inflammation. Due to the nature of single cell and the general approach of sequencing protocols, as described in Section 2.2, a wide range of studies focus on communication that can be inferred from gene expression data. Such examples include autocrine (cell signals to itself) and paracrine (short-distances) signalling. As such interactions are mediated by ligands and receptors, some of the most commonly used approaches to understanding cell-cell communication take a ligand-receptor interaction-based strategy. Following the analysis steps outlined in Figure 2.2, taking the scRNA-seq data and existing databases like DLPR, iMEX, and Uniprot, methods like CellPhoneDB allow for identification of interacting ligand-receptor pairs (Vento-Tormo et al. 2018). Specifically, all possible pairs of interacting clusters are generated based on the initial clustering of the data. For each cell type a mean count is calculated. Means

are calculated after permuting labels corresponding to cell annotations 1000 times, and the p-value from the randomisation test (what proportion of the means are more extreme than the actual mean) allows for the identification of cell-type specific interacting ligand-receptor pairs. From this list, biologically relevant interactions can be selected (Vento-Tormo et al. 2018). The approach of using curated resources of ligands and receptors is common and explored by some further tools as well such as NicheNet, SingleCellSignalR and CellChat (Browaeys et al. 2020, Cabello-Aguilar et al. 2020, Jin et al. 2021). An overview of the ligand-receptor based strategy for inferring interactions can be seen in Figure 2.4. Due to the increasing amount of curated resources and available methods, a benchmarking study has been conducted to evaluate existing ligand-receptor based cell-cell communication approaches (Dimitrov et al. 2021). Dimitrov et al conclude that the choice of scoring method and the underlying curated resources affect the inferred interactions. Differences also stem from what each method assumes to be an interesting interaction, for example "most specifically-interacting cell types" rather than "most actively communicating ones" (Dimitrov et al. 2021). Additionally, the authors conclude that like any other prior knowledge, the resources used for understanding cell-cell communication are biased and only illustrate some biological actuality (Dimitrov et al. 2021).

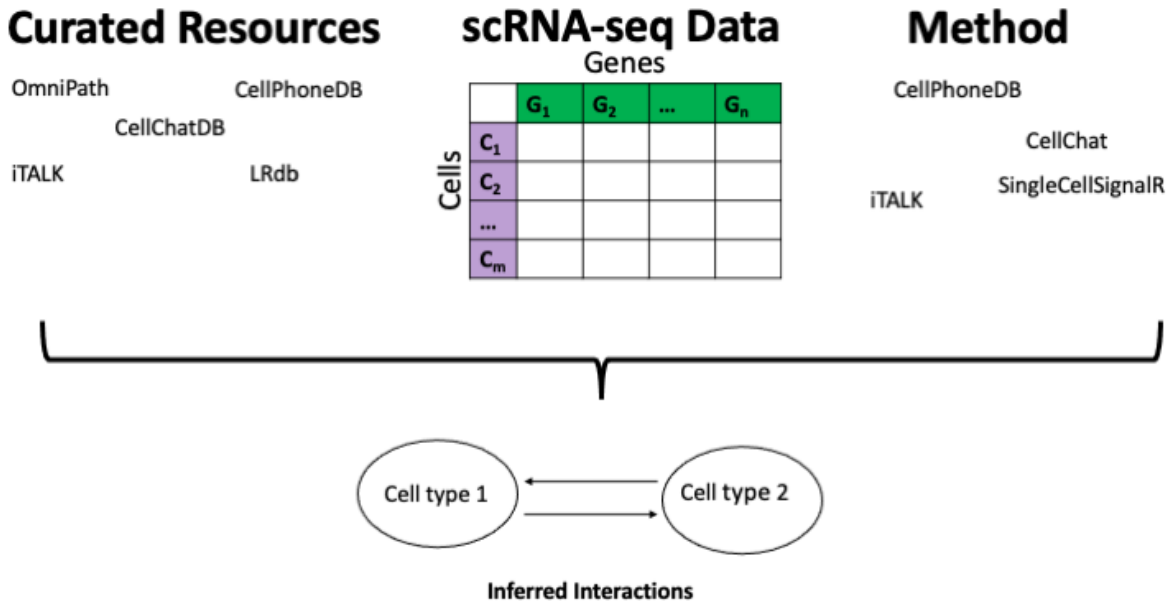


Figure 2.4: Some methods, for example CellPhoneDB, CellChat, SingleCellSignalR, rely on curated resources of ligands and receptors. Those resources are combined with a scoring function and using the scRNA-seq data of interest, interactions are inferred.

An alternative to the ligand-receptor interaction-based approach is what could be described as physically vicinal structure-based approach which relies on having access to physically interacting cells, and recent work has allowed sequencing cells involved in interactions (Boisset

et al. 2018, Giladi et al. 2020, Shao et al. 2020). One current example is PIC-seq, sequencing of physically interacting cells. PICs are isolated by a combination of tissue dissociation, staining for mutually exclusive markers, and flow cytometry sorting. Single positive and PIC populations are then sequenced. The capture of PICs potentially permits identification of novel interactions, beyond those described in curated resources. The computational side of their PIC-seq approach, firstly clusters mono-cultures, and the gene expression of each PIC is modelled as a doublet: $\alpha \times A + (1-\alpha) \times B$, where A and B are the two cell types that make the PIC and α is the mixing parameter. α is estimated by a linear regression model trained on synthetic PICs. This is followed by maximum likelihood estimation (MLE) of A and B. By identifying what the two subtypes that make the PIC are, expected expression can be computed. Expected and actual expression of the PIC are compared to identify changes as a result of interaction (Giladi et al. 2020).

Chapter 5 discusses the limitations of the computational methodology behind the PIC-seq approach, introduces a new method for detecting interactions, and presents the result of applying this method to synthetic and real data.

2.3.10 Options for further analysis

So far we have discussed some of the common steps for analysis of single cell data. While outside of the scope of this thesis, for completeness we would like to mention some further areas of research in scRNA-seq which we expect to gain even further popularity as more data become available.

As gene expression is tightly linked to networks of transcription factors and signalling molecules, understanding those networks is key to determining what drives transitions. Despite its inherent challenges, scRNA-seq data is a suitable base for developing methods to study gene regulatory networks (GRNs). While there are multiple approaches for inferring GRNs, this still remains a challenging problem, and perhaps a way to mitigate the shortcomings of current methods would be to include information from other modalities (Pratapa et al. 2020).

Single cell is an ever-growing field, currently in addition to capturing the transcriptomes of single cells, the expression of surface proteins (CITE-seq), assay for transposase-accessible chromatin (ATAC-seq), and chromatic accessibility (Hi-C) can also be measured (Buenrostro et al. 2015, Nagano et al. 2013, Stoeckius et al. 2017). As multiple modalities can be quantified for the same cell, it is of interest how multi-modal data can be analysed. For example, standard scRNA-seq might not be best-suited to capture and distinguish heterogeneity in T-cells. However, we can leverage the information that surface protein measurements provide and use the two modalities to perform joint clustering (Hao et al. 2020). The RNA expression will be invaluable when trying to resolve populations that do not have cell surface markers measured, and the protein expression will resolve the heterogeneity of some cells. Similar analysis can be performed using ATAC-seq data with scRNA-seq (Hao et al. 2020).

While scRNA-seq allows us to understand heterogeneity of cell populations, due to its nature

we lose spatial information about how the cells are organised in tissues. Even with the newly developed spatial transcriptomics methods where the size of a spot can be 50 microns, a single spot does not correspond to a single cell. Thus, scRNA-seq can be used to deconvolute the contents of those spots. Based on scRNA-seq data, cell type profiles are learned and then used to estimate proportions of cell types in each spot (Andersson et al. 2019, Elosua et al. 2020, Kleshchevnikov et al. 2020). Finally, in addition to deconvolution, scRNA-seq and spatial transcriptomics can be used to infer signaling relationships (Cang & Nie 2020).

2.3.11 Summary

In this chapter, the ways of generating scRNA-seq are described and common techniques for analysis of scRNA-seq data are outlined. Given current trends in scRNA-seq method development, the number of available tools is expected to rise to 3000 by the end of 2025 (Zappia & Theis 2021). The previous sections, while not providing an exhaustive list of the possible analysis and available tools, aimed to give an overview of the standard scRNA-seq analysis pipeline adhering to the currently established best practices in the field. Where appropriate, new developments and their effect on the field have been discussed.

Chapter 3

Machine learning background

Machine learning (ML) is a subset of artificial intelligence, focused on developing tools that can learn to solve problems by being trained on data. Depending on the application the goal might be to find similar objects and group them together, to learn a particular feature of those objects, or to predict something about them. Given the increasing complexity of generated data, it is of interest to create data-driven solutions to problems. With the advances in high-throughput sequencing and other technologies, high-dimensional omics datasets are generated and machine learning has been key to gaining insight from such datasets (Arjmand et al. 2022, Li et al. 2022).

As seen previously in Chapter 2, scRNA-seq datasets are very high dimensional, tens of thousands of genes across thousands or millions of cells. Given this high dimensionality, a manual approach to analysis is not feasible, and so automating the tasks of single cell exploration is necessary. One of the main scRNA-seq applications is studying cell types, grouping cells with similar expression profile together. This is equivalent to a problem, known as clustering, to find similar groups in the data. Another approach which has already been done is to reframe the problem of cell type annotation as a classification task, where classifiers have been trained on similar cell types and then used to predict the cell types of another dataset (Abdelaal et al. 2019). With the development of sequencing protocols that can scale to tens of thousands of cells, there is an advent of applying deep learning approaches to single cell data, from clustering to using pre-trained reference models on query datasets for annotation (Li et al. 2020, Lotfollahi et al. 2020). The application and development of ML methods to scRNA-seq is an area of dynamic research, of which the aforementioned examples are only scratching the surface.

In this chapter, the key machine learning concepts used throughout this thesis are presented. Mixture models are introduced and motivation is given for latent Dirichlet allocation, which is the basis of the work described in Chapters 4 and 5. The focus of Chapter 6 is an extension of a traditional topic model, one that takes into account dynamic and correlation. As dynamic is modelled through a Gaussian process, in Section 3.4 an overview of Gaussian processes is presented.

3.1 Mixture models

3.1.1 Clustering revisited

A common problem in machine learning is identifying latent groups in the data, known as clusters. A cluster contains a group of similar objects. Previously, in Chapter 2, we have seen clustering applied to scRNA-seq and as a fundamental step of scRNA-seq analysis. K-means is a common clustering approach, however here we are going to focus on mixture models. In K-means, a cluster is defined as the mean of all data points in that cluster, in the context of mixture models a cluster is a probability density. Let us assume the data came from K clusters or components. In this chapter we are going to use clusters and components interchangeably. Generating a data point, denoted $\boldsymbol{x}_n$, from a K component Gaussian mixture model is a two-step procedure:

- Select one of the K Gaussians with probability π_k where $\Sigma_k \pi_k = 1$
- Sample $\boldsymbol{x}_n$ from this Gaussian

However, in practice it is unknown which component generated the data and those can be treated as latent, unobserved, variables.

3.1.2 Mixture models as latent variable models

We assume each data point comes from one of several components (clusters) and the goal is to infer the distributions of the components. In this setting, the latent variables, denoted $\boldsymbol{z}_n$ are one-hot encoded and indicate which mixture component a data point was sampled from. For example, for a mixture model with $K = 4$ components, $\boldsymbol{z}_n = (0,1,0,0)$ indicates the data point belongs to the second mixture component. The generative process for a data point $\boldsymbol{x}_n$ can be defined as firstly sampling one of our mixture components and then sampling x_n from that component:

$$\begin{aligned} \boldsymbol{z}_n &\sim \text{Multinomial}(\boldsymbol{\pi}) \\ \boldsymbol{x}_n | z_{nk} = 1 &\sim \text{Gaussian}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \end{aligned} \tag{3.1}$$

where $\boldsymbol{\pi}$ denotes the mixing proportions and $\sum_k \pi_k = 1$. $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the mean and covariance of the Gaussian at index k. k indicates the mixture component sampling is done from $(k = 1...K)$. Here, we consider a mixture of Gaussians but this is not restrictive and other distributions can be used as well. To make notation easier, a parameter vector $\boldsymbol{\theta}$ can be introduced $\boldsymbol{\theta}$ $=\{\pi_1, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1 .... \pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$

The marginal distribution $p(\boldsymbol{x}_n|\boldsymbol{\theta})$ can be obtained by summing over z where $z_{nk} = 1$ indicates the data point belongs to component k.

$$p(\boldsymbol{x}_n|\boldsymbol{\theta}) = \sum_{k=1}^{K} p(\boldsymbol{x}_n, z_{nk} = 1|\boldsymbol{\theta_k}) \tag{3.2}$$

While the latent variable cannot be observed, their posterior distribution can be inferred. $\boldsymbol{X}$ denotes all observations. The log likelihood can be defined as:

$$logp(\boldsymbol{X}|\boldsymbol{\theta}) = \sum_{n=1}^{N} log \sum_{k=1}^{K} p(\boldsymbol{x}_n, z_{nk} = 1|\boldsymbol{\theta}) \tag{3.3}$$

Normally the expression in 3.3 will be differentiated and solved for 0. However, we cannot solve it analytically and EM (expectation-maximisation) is used instead. EM derives a lower bound on the likelihood. Let $q(\boldsymbol{z}_n)$ be a probability distribution over the latent variables. Using Jensen's inequality a lower bound can be defined on $logp(\boldsymbol{X}|\boldsymbol{\theta})$.

$$logp(\boldsymbol{X}|\boldsymbol{\theta}) = \sum_{n=1}^{N} log\,\mathbb{E}_q(\boldsymbol{z}_n)\frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta})}{q(\boldsymbol{z}_n)} \geq \sum_{n=1}^{N} \mathbb{E}\,q(\boldsymbol{z}_n) log \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta})}{q(\boldsymbol{z}_n)} = \mathscr{L}(\boldsymbol{\theta}, q) \tag{3.4}$$

The result of subtracting the lower bound from the log likelihood is non-negative and known as Kullback-Leibler (KL) divergence. $\boldsymbol{Z}$ is the collection of all latent variables.

$$\begin{aligned} &logp(\boldsymbol{X}|\boldsymbol{\theta}) - \mathscr{L}(\boldsymbol{\theta}, q) = logp(\boldsymbol{X}|\boldsymbol{\theta}) - \mathbb{E}q(\boldsymbol{Z}) log \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} \\ &= \mathbb{E}q(\boldsymbol{Z}) log \frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})} \\ &= KL(q(\boldsymbol{Z})||p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})) \end{aligned} \tag{3.5}$$

EM is an iterative maximum likelihood approach that consists of 2 steps:

- E-step: estimate the value for the latent variables
- M-step: optimise the parameters. In this particular case, update equations need to be derived for π_k, $\boldsymbol{\mu}_k$, and $\boldsymbol{\Sigma}_k$

In the E-step, the lower bound is maximised with respect to q and $\boldsymbol{\theta}$ remains fixed.

To obtain values for the model parameters, the derivatives of the lower bound of the log likelihood are taken with respect to π_k, $\boldsymbol{\mu}_k$, and $\boldsymbol{\Sigma}_k$, set to 0 and equations are derived. The final update equations for the parameters used in the M-step take the following form:

$$\begin{aligned} \pi_k &= \frac{1}{N}\sum_{n=1}^{N} q(z_{nk} = 1) \\ \boldsymbol{\mu}_k &= \frac{\sum_{n=1}^{N} q(z_{nk} = 1)\boldsymbol{x}_n}{\sum_{n=1}^{N} q(z_{nk} = 1)} \\ \boldsymbol{\Sigma}_k &= \frac{\sum_{n=1}^{N} q(z_{nk} = 1)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T}{\sum_{n=1}^{N} q(z_{nk} = 1)} \end{aligned} \tag{3.6}$$

An example of the mixture model at convergence following the EM algorithm can be seen in Figure 3.1. Each data point will have a probability for belonging to each of the clusters. The final assignment to a cluster is based on the highest probability.

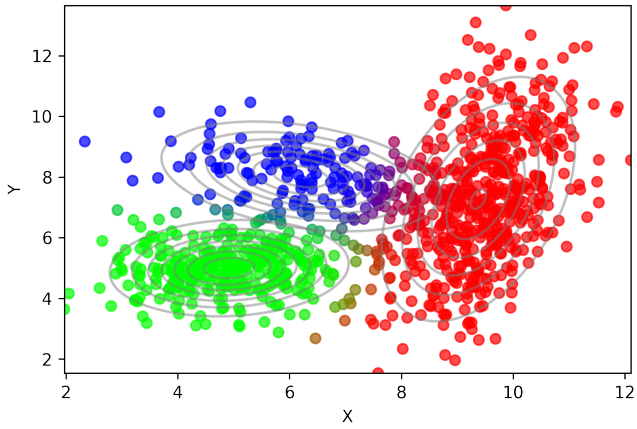


Figure 3.1: Three component mixture model at convergence following the EM algorithm. Data points are assigned to components based on the the highest probability

A Bayesian approach to the above outlined mixture model can be adopted as well. A prior distribution over the parameter values can be included. For example, a Dirichlet prior can be adopted for the mixing coefficients (conjugate to the multinomial). While in this thesis sampling is not used when a closed form for the posterior is not available, Gibbs sampling, a Markov chain Monte Carlo method, is a popular technique that can be used to obtain the posterior distribution and it has widely used for mixture models. Assuming conditional distributions can be computed, each parameter value is sampled from a distribution conditioned on all other parameters.

3.2 Notes on inference

In a Bayesian setting a posterior probability is required to make predictions. However, a closed form solution for the posterior is often unavailable; i.e., the posterior is not analytically tractable. A common way of overcoming the issue of an intractable posterior is variational inference. As the real posterior distribution cannot be derived, the aim is to approximate it by finding a variational distribution.

Variational inference can be considered an extension of EM, which was discussed in the context of mixture models. While under certain conditions EM and variational inference are equivalent, it is worth explicitly noting that EM provides a point estimate while variational inference results in a distribution. In the case of EM, the aim was to maximise the lower bound which is equivalent to minimising the KL divergence between $q(\boldsymbol{Z})$ and the true posterior. Here, the idea is similar: since there is no access to the target distribution, it needs to be approximated

in such way that the KL divergence between the two distributions is minimised. Specifically given two distributions p and q, the KL divergence can be formally defined as:

$$KL(q(\boldsymbol{z})||p(\boldsymbol{z}|\boldsymbol{x})) = \int q(\boldsymbol{z}) log \frac{q(\boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{x})} d\boldsymbol{z} = -\int q(\boldsymbol{z}) log \frac{p(\boldsymbol{z}|\boldsymbol{x})}{q(\boldsymbol{z})} d\boldsymbol{z} \tag{3.7}$$

It can be noted that the KL divergence is ≥ 0 and it is not symmetric, meaning the KL divergence between p and q is not the same as the KL divergence between q and p.

Previously, in the EM algorithm a lower bound was derived and the aim was to maximise this bound. In the setting of variational inference, as the KL divergence cannot be minimised exactly, a proxy needs to be used. This is known as the evidence lower bound (ELBO). Maximising the ELBO minimises the KL divergence. Optimisation of the ELBO is followed by iterative updating of the model parameters.

3.3 Latent Dirichlet Allocation (LDA) and other topic modelling approaches

3.3.1 Motivation and generative process

Previously we have introduced mixture models and assumed each data point to be generated from a single component. For example, consider a collection of documents. In the standard mixture model setting, each document from the collection is assumed to come from one mixture component, and each mixture component is a multinomial over words. In that setting, each mixture component can be thought of as a topic and the document is about that topic. For example, if a set of PhD dissertations are reviewed, some might come from the scRNA-seq cluster, some might be from the ML one, and some might be from the cluster that covers ML and scRNA-seq. However, instead of having three separate clusters for those documents, each dissertation can be considered as a contribution of the ML and the scRNA-seq clusters. This would result in a simpler model, and so not as many clusters would be required. Additionally, that would be a more accurate reflection of how documents are generated. For example, there will be some dissertations that are more ML-focused and apply novel methods to scRNA-seq data. Others use ML methods to uncover biological information. In the first case, the ML component will have higher contribution to the dissertation while in the second case the ML contribution will be lower compared to the contribution of scRNA-seq. Such dissertations can be modelled as different contributions of two main components.

LDA is a model that overcomes the limitations of standard mixture models. Instead of assuming that each document is generated by one topic, documents can be generated by multiple latent topics. Each topic is a multinomial distribution over a set vocabulary. Given D documents (indexed d = 1,...,D), N words (indexed n = 1,..., N), K topics, z_{dn} denotes the assignment of the

n-th word in the d-th document to the k-th topic. The generative process for a document can be defined as:

1. Sample a multinomial over K topics, $\boldsymbol{\theta}_d$, from a Dir($\boldsymbol{\alpha}$)
2. For each word in the document:

- Choose a topic from the K topics, $z_{dn} \sim$ Multinomial($\boldsymbol{\theta}_d$)
- Sample a word from the chosen topic, $w_{dn} \sim$ Multinomial($\boldsymbol{\beta}_{z_{dn}}$)

where $\boldsymbol{\beta_k} \sim Dir(\boldsymbol{\eta})$. $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$ are the parameters defining the Dirichlet priors over document-topic and topic-word multinomials. A graphical model of LDA can be seen below.

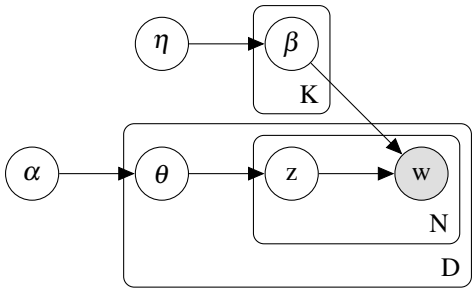


Figure 3.2: Graphical model representation of LDA.

In the context of scRNA-seq, genes are equivalent to words, cells are equivalent to documents and topics are groups of co-varying genes. Topics can be general or cell-type specific. As genes can contribute to multiple biological processes, they can be in multiple topics, similar to words. We discuss in more detail in Section 4.3 the application of LDA to scRNA-seq data.

3.3.2 Inference

Given the LDA formulation, the posterior takes the following form:

$$p(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\beta} | \boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\eta}) = \frac{p(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\beta} | \boldsymbol{\alpha}, \boldsymbol{\eta})}{p(\boldsymbol{w} | \boldsymbol{\alpha}, \boldsymbol{\eta})} \tag{3.8}$$

where $\boldsymbol{z}$ denotes the latent topic assignments. The learning task requires computing the posterior over $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, and latent variables $\boldsymbol{z}$ need to be marginalised. Marginalising $\boldsymbol{z}$ would entail averaging over all K^N configurations for z_{nd} per document, hence for a high-dimensional vocabulary and K, the posterior is intractable.

Given an intractable posterior, a possible solution is Gibbs sampling and indeed a collapsed version of Gibbs sampling is a popular option. However, within this thesis variational inference will be used. Similarly to the earlier section, we choose a variational distribution. For that arbitrary variational distribution a lower bound can be derived using Jensen's inequality. In

common with the previous sections, maximising the lower bound with respect to γ_d and ϕ_d is equivalent to minimising the KL divergence between the two distributions.

The variational inference algorithm can be described in two steps:

- E-step: For each document, optimise the variational parameters: γ, ϕ. To do this the bound is maximised with respect to each variational parameter, specifically ϕ_{nk} denotes the probability that the n^{th} word is generated by latent topic k. γ_k is the k^{th} component of the posterior Dirichlet parameter.

- M-step: For fixed values of the variational parameters, maximise the lower bound with respect to the model parameters, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$

The corresponding update equations as derived by (Blei et al. 2003) can be formulated as follows:

$$\begin{aligned}
\phi_{nk} &= \beta_{kn} \exp\left(\Psi(\gamma_k) - \Psi\left(\sum_{j=1}^{K} \gamma_j\right)\right) \\
\gamma_k &= \alpha_k + \sum_{n=1}^{N} \phi_{nk} \\
\boldsymbol{\alpha} &= \boldsymbol{\alpha} - H(\boldsymbol{\alpha})^{-1} g(\boldsymbol{\alpha}) \\
\beta_{ij} &= \eta + \sum_{d=1}^{D} \sum_{n=1}^{N_d} \phi_{dni} w_{dn}
\end{aligned} \tag{3.9}$$

where $\beta_{ij} = P(w^j = 1 | z^k = 1)$. Ψ is the first derivative of the $\log\Gamma$ function. $H(\alpha)$ and $g(\alpha)$ are respectively the Hessian and the gradient at the old value of α. In the M-step Blei (Blei et al. 2003) describes an efficient Newton-Raphson method for inverting the Hessian. Since $H(\alpha)$ is of specific form and satisfies the matrix inversion lemma, a Newton-Raphson algorithm with linear complexity can be derived. Further details can be found in the original publication by Blei et al. (2003).

3.3.3 Extensions of the standard LDA

So far we have discussed standard LDA where evolution of topics over time is not considered. However, certain collections of documents are dynamic and topics are expected to change over time. As such, incorporating the timestamp of the document can be more informative. Those are known as dynamic topic models, and probabilities of topics (or words) change over time. Additionally, as proposed by Blei (Lafferty & Blei 2006) some latent topics should not be considered independently, and topics are in fact correlated. An extension of LDA that takes into consideration both dynamic and correlations is discussed in the context of scRNA-seq in Chapter 6.

3.3.4 Other topic modelling approaches

Another widely used document model is probabilistic latent semantic indexing (pLSI) (Hofmann 1999). Similarly to LDA, pLSI models a document as a contribution of multiple topics. The challenge of pLSI arises when there is a need to assign probabilities to a previously unseen document (Blei et al. 2003, Hofmann 1999). Within the context of this thesis, it is important to make predictions for new cells, Chapters 5 and 6. As such, we consider LDA a more suitable approach.

In addition to pLSI, there are matrix factorisation approaches available like non-negative matrix factorisation (NMF). Given an $n \times m$ matrix $\mathbf{V}$ aims to find two non-negative matrices $\mathbf{W}$ ($n \times k$) and $\mathbf{H}$ ($k \times m$) such that:

$$\boldsymbol{V} \approx \boldsymbol{W}\boldsymbol{H}$$

The NMF algorithm starts with initialising the two matrices ($\boldsymbol{W}$ and $\boldsymbol{H}$), calculating their difference compared to $\boldsymbol{V}$, and minimising the error between $\boldsymbol{V}$ and their dot product. While NMF has been used for topic modelling, it is not a probabilistic method and similarly to pLSI cannot predict the topic contributions of a new document.

3.3.5 Choosing the number of topics

One of the inherent problems of topic models is choosing the number of topics. If the number of specified topics is too low, then the model might not be able to fully capture the complexity of the data. To select a suitable number of topics, there are metrics that can be computed for a range of values for the topic parameter of the model.

In information theory perplexity is a commonly used metric that evaluates how well a model describes the dataset, where the lower the perplexity the better fit the model is for the data. Specifically, the per-word-perplexity is computed as an exponent of the average negative ELBO per word. The average per-word perplexity is defined as:

$$\text{perplexity}_{pw} = exp\left(\frac{-\text{ELBO}}{\Sigma_{d \in D} N_d}\right)$$

Generally, as the number of topics increases, perplexity drops. A lower perplexity indicates a better model. We compute perplexity for a range of topics in Chapters 5 and 6, and then choose the most suitable range of values.

There are also further metrics available for evaluating the number of topics. For example, (Cao et al. 2009) proposed average cosine distance. A cosine distance is used to measure the correlation between topics. The average cosine distance is computed between all pairs to measure the stability of the topic structure. In the case of cosine distance, a lower value indicates a better performing model.

Alternatively, (Deveaud et al. 2014) proposed a metric based on Jensen-Shannon divergence (a symmetric version of KL divergence) which measures the information divergence between two distributions. Higher values indicate a better model.

In Chapter 5 we use primarily perplexity but we also show the results across perplexity, JS divergence, and cosine distance follow a similar pattern for the optimal number of topics.

3.4 Gaussian Process (GP)

3.4.1 Introduction to GPs

Here we are going to introduce Gaussian processes by starting with the Univariate Gaussian. The Univariate Gaussian is characterised by its mean and variance. In Figure 3.3 can be seen a Gaussian defined as $\mathcal{N}(0,1)$ with 0 mean and unit variance.

This can be generalised to a 2-dimensional Gaussian where the mean becomes a mean vector and the variance becomes a covariance matrix. To illustrate the effect of the covariance matrix, two examples of multivariate Gaussians are shown in Figure 3.4. In the first case, the two variables, x_1 and x_2 are independent and their correlation is 0. In the second plot, x_1 and x_2 have a correlation of 0.9.

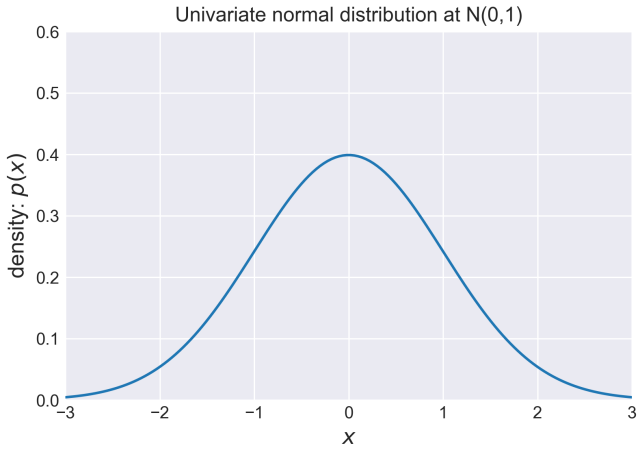


Figure 3.3: This one dimensional Gaussian can be characterised by its mean (0) and variance (1).

Given a 2-dimensional Gaussian, samples can be drawn as shown in Figure 3.5a. Plotting samples in the space of x_1 and x_2 is easy when working with 2-dimensional Gaussians, but as the dimensionality increases this type of visualisation becomes tricky. As such, a parallel coordinates plot can be used as an alternative for visualising samples with higher dimensionality. On the x-axis the plot is indexed by the dimensions of the Gaussian and a line is drawn between x values that came from the same sample. Figure 3.5b shows the same 20 samples from Figure 3.5a on a parallel coordinates plot.

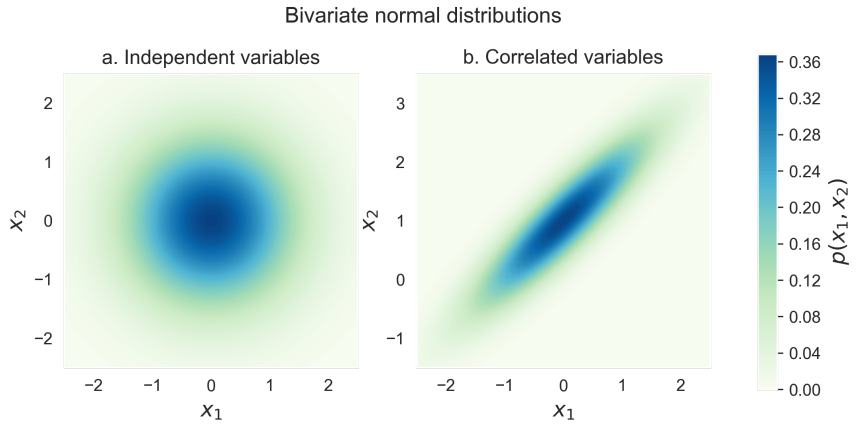
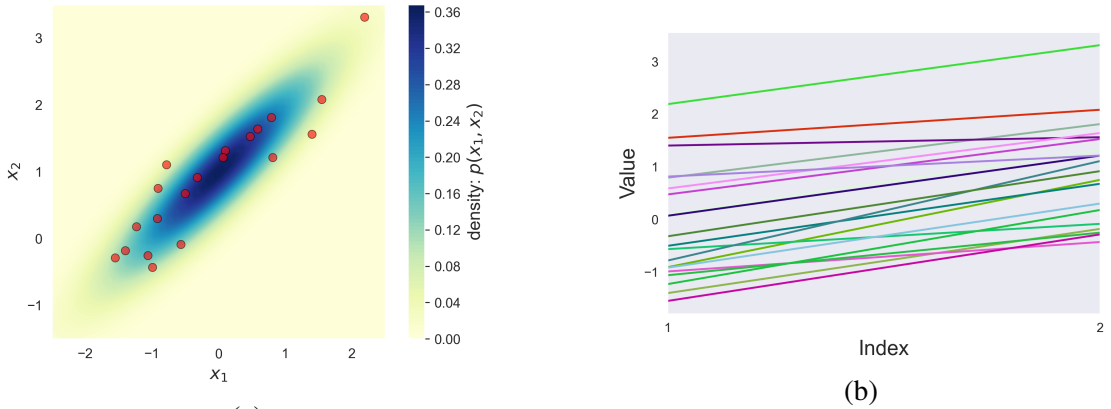


Figure 3.4: Two examples of multivatiate Gaussians. Covariance $\mathbf{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ of **a.** implies independent x_1 and x_2. In **b.** the two variables are correlated with covariance $\mathbf{\Sigma} = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}$



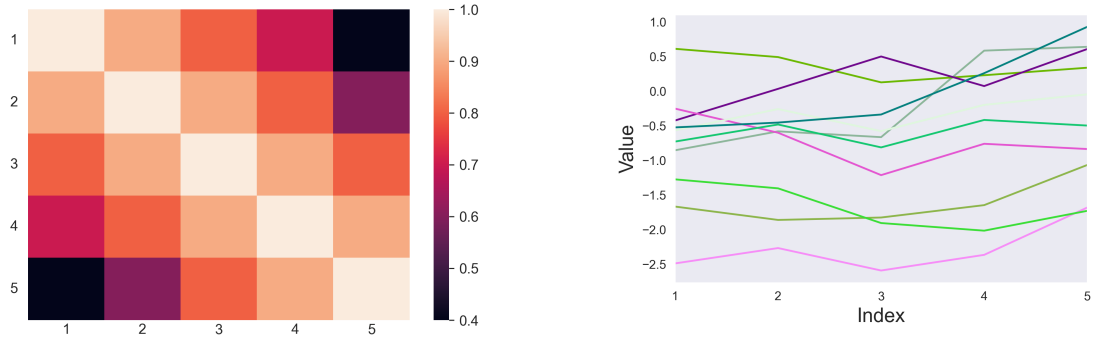
(a) (b)

Figure 3.5: (a) 20 samples drawn from that multivariate Gaussian distribution and visualised in the x_1 and x_2 space. The two variables are correlated with covariance matrix $\mathbf{\Sigma} = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}$. (b) Alternatively, the samples can be visualised on a parallel coordinates plot with x-axis indexed by the number of variables.

Equipped with the parallel coordinates plots, the 2-dimensional Gaussian can be extended to a 5-dimensional one as shown in Figure 3.6. Variables indexed at 1 and 5 have a correlation of 0.4, which is lower compared to the correlation of the other variables. This is also evident from the samples, Figure 3.6b. As more variables, i.e. dimensions, are added the samples from those multidimensional Gaussians start to look like functions, as seen in Figure 3.7.

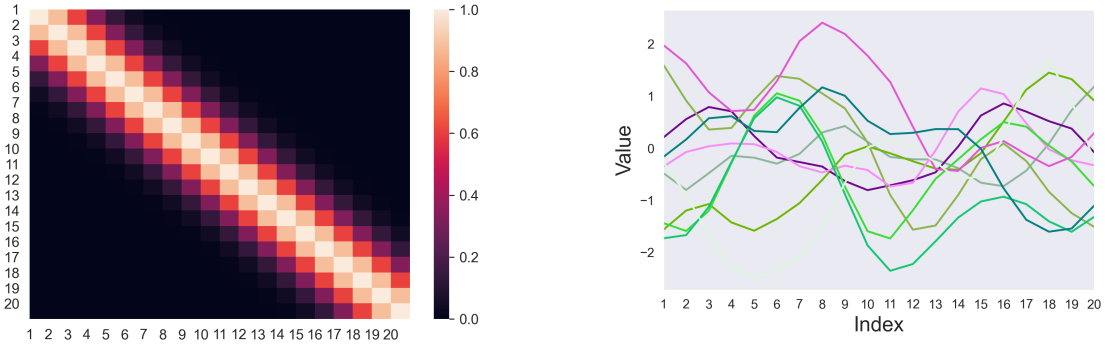
A GP can be defined as an infinite dimensional distribution over functions. Formally, a GP is a collection of random variables, any finite subset of which will be Gaussian distributed. GPs have several useful properties:

- closed under conditioning: Assume we are given $x_1 \ldots . x_n$ and they are Gaussian, then $x_1 | x_2 \ldots . x_n$ is also Gaussian. This property is very useful as it provides an analytical solution



(a) (b)

Figure 3.6: (a) Covariance matrix, showing lower correlation between variables indexed 1 and 5 compared to other variables. (b) 10 Samples from a 5-dimensional Gaussian.



(a) (b)

Figure 3.7: (a) 20x20 Covariance matrix. (b) 10 Samples from a 20-dimensional Gaussian.

to inference in GPs. An example is given in Section 3.4.3.

- closed under linear operations: for example two Gaussian processes can be added together and the result is still a Gaussian process.

A GP is typically specified through mean and covariance (kernel) functions, $GP(\mathbf{0}, \Sigma)$. Typically the mean is defined as zeros ,however this is not a limitation since the mean of the posterior is not fixed to be 0. (Rasmussen & Williams 2006). There are many different choices for the covariance function which will be discussed in the next section. A GP inherits its properties from the covariance function: e.g. smoothness, periodicity, and others.

3.4.2 Kernel functions

So far we have introduced GPs by motivating them from multidimensional Gaussian and we defined them as an infinite distribution over functions. From the properties of GPs we know that any finite subset of variables will be Gaussian distributed. However, depending on the problem domain some of those functions might be a better fit than others. For example, some data may

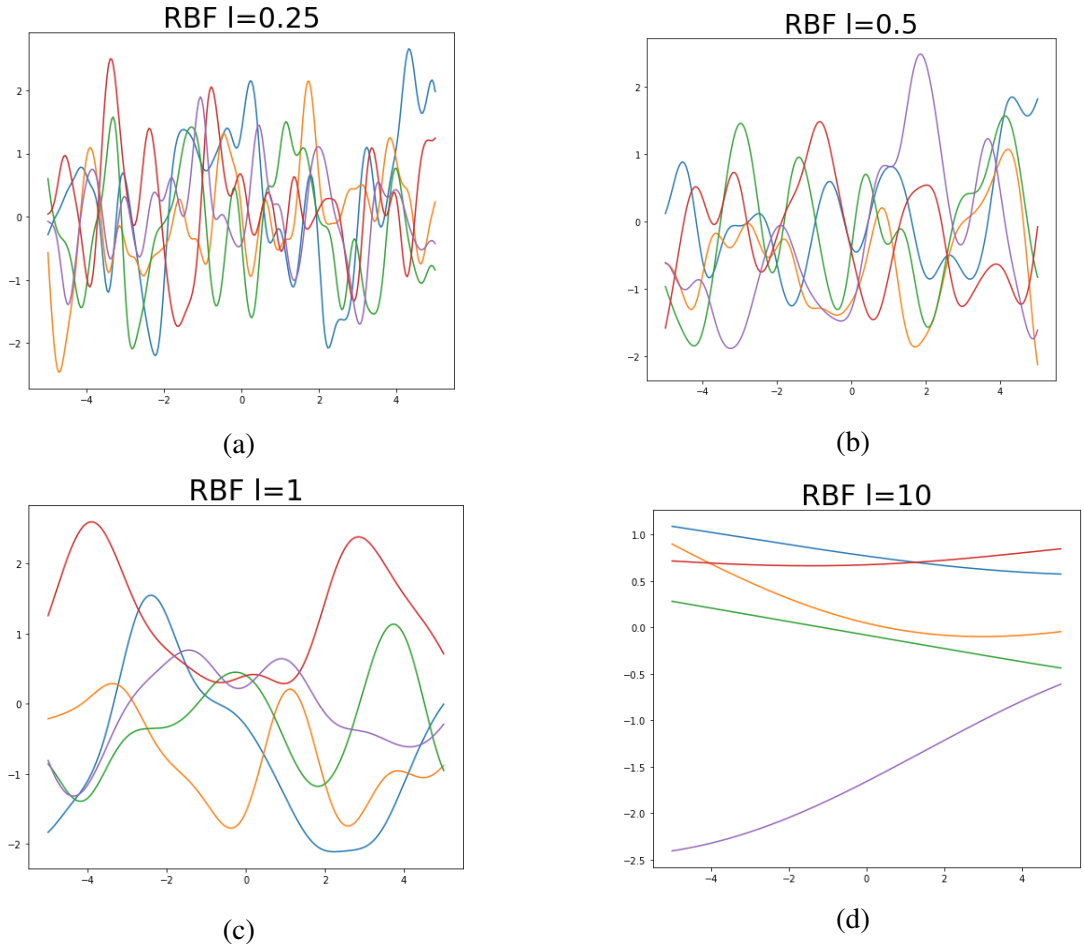
include periodicity. In a Bayesian setting of the problem, the more suitable family of functions can be included as prior information. This can be seen in the section discussing GP regression.

Choice of an appropriate kernel will depend on the application and can be made based on domain knowledge. The covariance describes how the data points correlate with each other. Given a set of input points $x_1, x_2 \ldots x_N$, the covariance matrix is computed by evaluating the kernel function for all pairs of x values.

A popular choice of kernel is RBF (Radial Basis Function), also known as squared exponential kernel:

$$k(x,x') = \alpha \exp\left(-\frac{(x-x')^2}{2l^2}\right) \tag{3.10}$$

The RBF covariance has two parameters: variance, α and lengthscale, l. As the RBF kernel is infinitely differentiable, it can model very smooth functions. For higher values of l, the GP samples approach straight lines, while for lower values they appear more like white noise or completely uncorrelated GPs. The effect of varying the lengthscale of the RBF kernel can be seen in Figure 3.8.



(a) (b) (c) (d)

Figure 3.8: Effect of the lengthscale, l, on the complexity of the samples from a GP with RBF kernel. Higher values of l result in GP samples that approach a straight line, while a lower value of l generates more complex functions, approaching white noise.

Another popular set of kernel functions, the Matérn class can be considered a generalisation of RBF. There is an additional parameter v which controls the smoothness of the function.

$$k(x,x') = \alpha \exp -\frac{\sqrt{2v}|x-x'|}{l} \frac{\Gamma(p+1)}{\Gamma(2p+1)} \sum_{i=0}^{p} \frac{(p+1)!}{i!(p-1)!} \left(\frac{\sqrt{8v}|x-x'|}{l} \right)^{p-i} \tag{3.11}$$

In particular, Matérn $3/2$ and Matérn $5/2$ have been used to model biological processes in gene expression data (Ahmed et al. 2019). Samples from GPs with those two Matérn kernels can be seen in Figures 3.9a and 3.9b.

Another type of kernel suitable for modelling periodic processes, such as cell cycle is the periodic kernel.

$$k(x,x') = \alpha \exp \left(\frac{-2\sin^2(\pi|x-x'|/p)}{l^2} \right) \tag{3.12}$$

The lengthscale behaves similarly to the lengthscale in the RBF. p is the period, determines the distance between function repetitions. A periodic kernel with period 2 can be seen in Figure 3.9c.

In addition to the family of kernels discussed earlier, custom kernels can also be created by summing kernels together, multiplying them, or even composing them with a function as all those operations would preserve the positive semi-definite requirement for the GP covariance (Rasmussen & Williams 2006).

3.4.3 GP regression

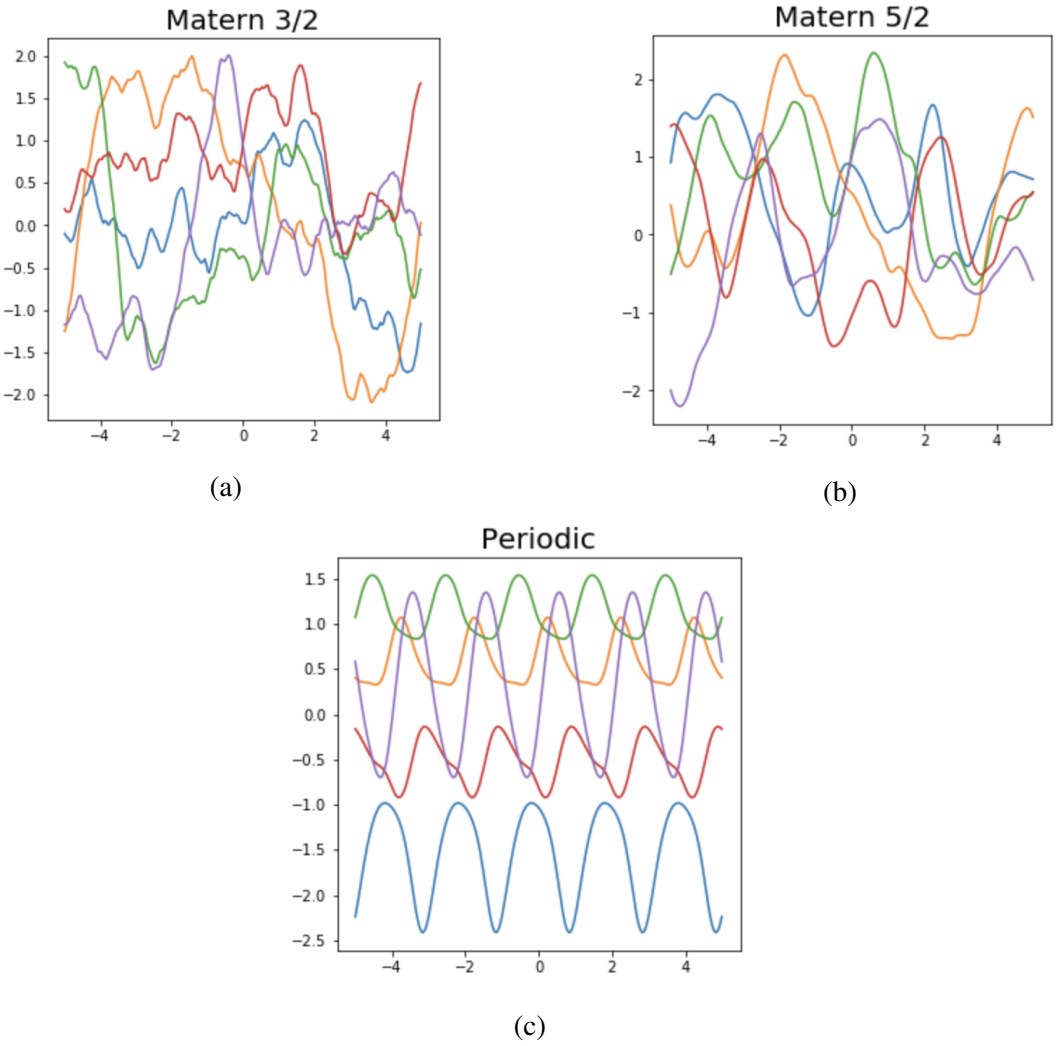
In a standard regression setting, given some training inputs we aim to fit a function on those inputs, so that we can make predictions at previously unseen values. While this can be done by choosing a parametric form for our function, here the function will be described by mean vector and covariance matrix.

Let $\boldsymbol{x} = \{x_n\}_{n=1}^N$ be variables with corresponding targets $\boldsymbol{y} = \{y_n\}_{n=1}^N$. For example, in the case of a time ordered data $x_n \geq x_{n-1}$. We assume that instead of observing true function values, we observe $\boldsymbol{y}$ where

$$\boldsymbol{y} = f(\boldsymbol{x}) + \varepsilon \tag{3.13}$$

ε is independently Gaussian distributed noise. Because the GP prior on the function f is Gaussian, the marginal likelihood can be obtained by integrating out $f(\boldsymbol{x})$.

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{x}) &= \int \mathcal{N}(\boldsymbol{y}|\boldsymbol{f}, \sigma^2 \boldsymbol{I}) \mathcal{N}(\boldsymbol{f}|\boldsymbol{0}, K_{NN}) d\boldsymbol{f} \\ &= \mathcal{N}(\boldsymbol{y}|0, K_{NN} + \sigma^2 \boldsymbol{I}) \end{aligned} \tag{3.14}$$



(a) (b) (c)

Figure 3.9: Samples from a GP prior with zero mean and some popular kernels such as Matérn 3/2, Matérn 5/2, and periodic. All those kernels have been applied to modelling biological processes in gene expression data.

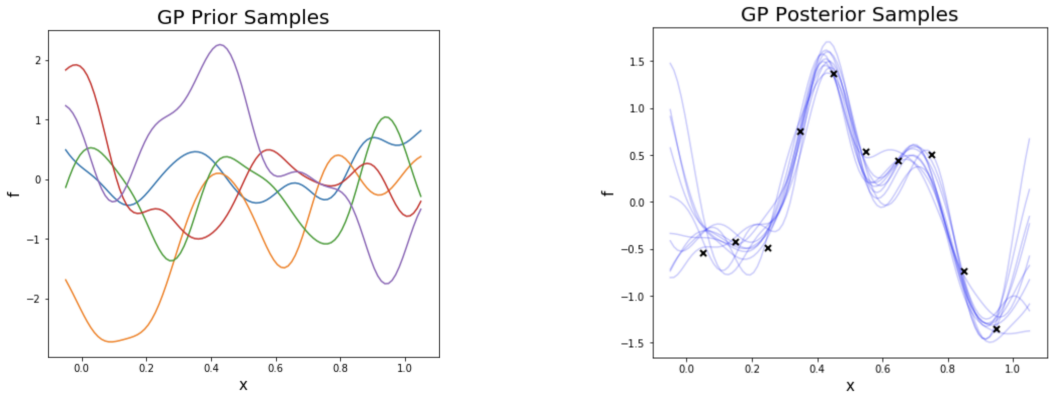
K_{NN} is the NxN covariance matrix, computed between all pairs of $\boldsymbol{x}$, and $\boldsymbol{f} = f(\boldsymbol{x})$. σ^2 is the variance. As in any standard regression model, we are interested in making predictions at a set of inputs $\boldsymbol{x}^*$. The posterior distribution $\boldsymbol{f}^*$ given the data is:

$$\begin{aligned} p(\boldsymbol{f}^*|\boldsymbol{y}) &\sim \mathcal{N}(\boldsymbol{\mu}^*, C^*) \\ \boldsymbol{\mu}^* &= k(\boldsymbol{x}, \boldsymbol{x}^*)^T (K_{NN} + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{y} \\ \boldsymbol{C}^* &= K(\boldsymbol{x}^*, \boldsymbol{x}^*) - K(\boldsymbol{x}, \boldsymbol{x}^*)(K_{NN} + \sigma^2 \boldsymbol{I})^{-1} K(\boldsymbol{x}, \boldsymbol{x}^*) \end{aligned} \tag{3.15}$$

where $K(\boldsymbol{x}, \boldsymbol{x}^*)$ is the covariance between the training set, $\boldsymbol{X}$, and the test points where we want to make a prediction. $K(\boldsymbol{x}^*, \boldsymbol{x}^*)$ is the test set covariance. The closed form solution observed here is a result of one of the properties we have discussed earlier: GPs are closed under conditioning.

An example with RBF covariance can be seen in Figure 3.10. Figures 3.10a and 3.10b show samples from the GP prior and the GP posterior respectively. In the noisy regression case, the

samples as seen from Figure 3.10b are not required to pass through the training data. As we move away from the training points, the mean predictions become closer to the GP prior mean.



(a) (b)

Figure 3.10: GP Regression: (a) Samples from a GP prior with RBF covariance. (b) Adding observations and sampling from the GP posterior in a noisy GP regression. As we move away from the training data, the GP is returning to the mean.

Given a dataset with N data points, the GP time complexity is $O(N^3)$ due to the matrix inversions and memory demands of $O(N^2)$. This is considered a practical limitation, and in order to improve scalability of GPs to large datasets a range of approximation techniques have been proposed.

3.4.4 Inference

An exact inference of a GP has time complexity of $O(N^3)$ and requires $O(N^2)$ memory where N is the number of training points. To permit for wider use of GPs, sparse approximation strategies have been developed. Those approximations reduce the complexity of the GP to $O(NM^2)$ where M is the number of inducing points. Inducing points can either be selected at random from the initial training set, or they can be optimised via gradient optimisation. The two major inference strategies are FITC (fully independent training conditional), and VFE (variational free energy) (Snelson & Ghahramani 2005, Titsias 2009). The originally proposed FITC has been reformulated several times over the years (Bauer et al. 2016).

In the case of FITC, the original model is approximated to a simpler one. The alternative to model approximation is approximate inference in the case of VFE which approximates the GP posterior to another Gaussian distribution. By maximising the ELBO, a variational approximation of the posterior is constructed, and the inducing points and kernel hyperparameters are learned. The variational inference approximation is done in a similar fashion to the approach described in section 3.2.

3.4.5 Gaussian Process Latent Variable Model (GPLVM)

Previously, $\boldsymbol{x}$ values have been known. However, in the setting of GPLVM we are going to treat them as latent variables, and their positions will need to be optimised. GPLVM can be considered as the unsupervised alternative to GPs (Lawrence 2005). For example, function values are shown in Figure3.11a. However, there are multiple functions that can explain the data (see Figure 3.11b, 3.11c, 3.11d). The GP prior narrows down the choice of functions, and the kernel family can be chosen using domain knowledge as previously explained.

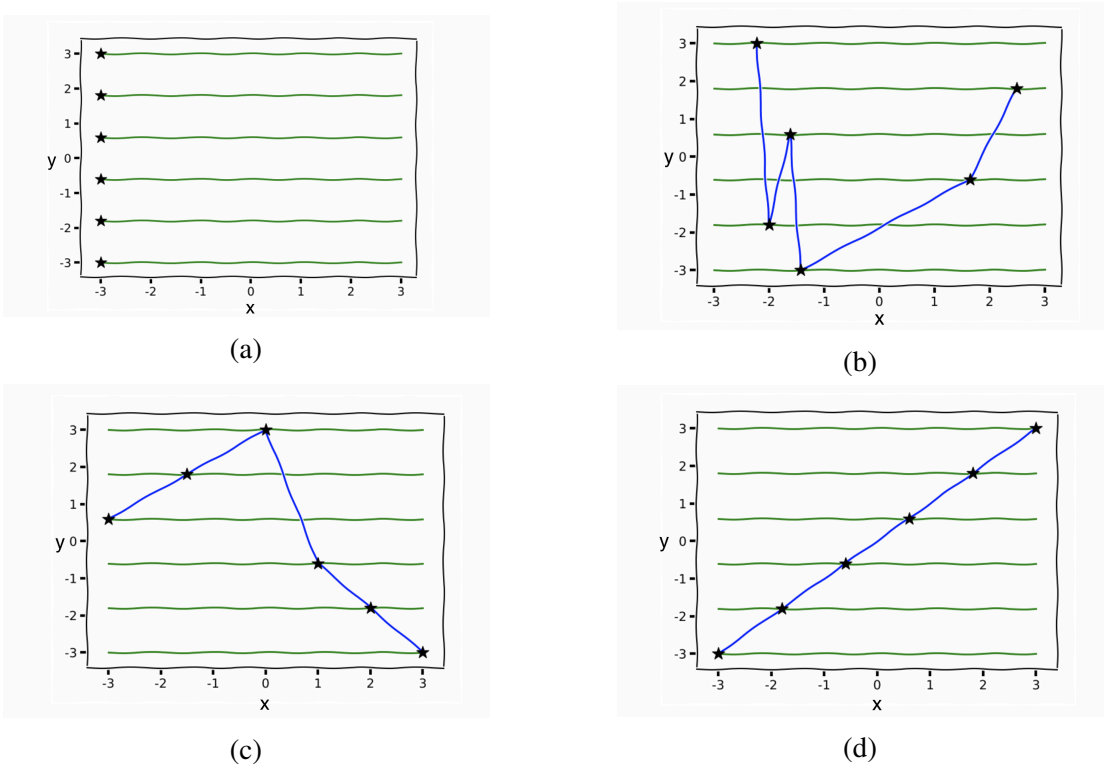


Figure 3.11: (a). Under the formulation of GPLVM, we know the $\boldsymbol{Y}$ values but not the $\boldsymbol{X}$ which is considered latent. (b), (c), (d): Different functions that could have generated the data. The choice of functions as previously seen is narrowed down by the choice of the GP prior (e.g. smooth functions, periodic functions)

Similarly to standard GPs, GPLVM is also affected by the same complexity and scalability issues, and similar approaches have been developed to improve the inference procedure (Titsias & Lawrence 2010).

3.5 Other useful ML concepts

In this section some further machine learning concepts are introduced as they aid results evaluation and interpretation in later chapters.

3.5.1 Classification and support vector machines (SVM)

Classification is a supervised machine learning approach that learns a mapping between a set of input variables $\boldsymbol{x}_1, \boldsymbol{x}_2, ...\boldsymbol{x}_n$ and output labels, y where $y = \{0, 1\}$ for binary classification or $y = \{0, 1, 2...N\}$ for a multi-class example.

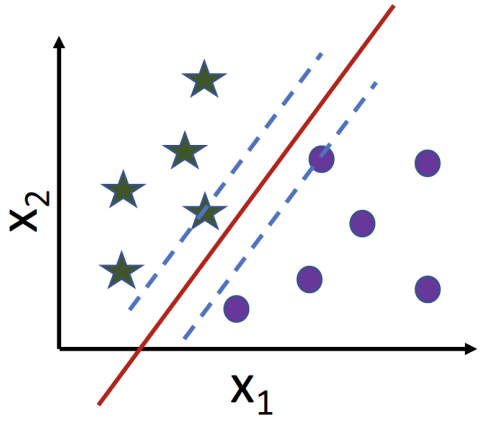
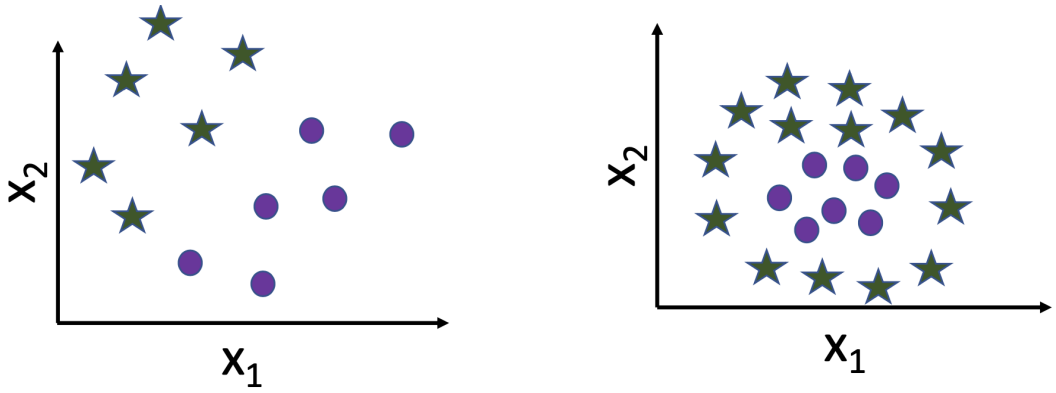


Figure 3.12: A representation of SVM in 2D where the two classes can be separated linearly in their original space. The two classes are indicated in purple and green, and the red line is the decision boundary. The two dashed line specify the margin which is defined by the support vectors, the closest points to the decision boundary.

In Chapter 4 we use support vector machine (SVM). SVM has been chosen as it has demonstrated excellent empirical performance and it is usually one of the best performing classifiers. Given a set of N training objects $(\boldsymbol{x}_1, \boldsymbol{x}_2, ...\boldsymbol{x}_n)$ and their corresponding labels $y \in \{-1, 1\}$, the label of a new data point is $\text{sign}(\boldsymbol{w}^T \boldsymbol{x}_{new} + b)$. In this case, the learning task is finding suitable values for $\boldsymbol{w}$ and b by maximising a quantity known as the margin.

While in some cases linear decision boundaries in the original input space can separate the classes, this is often not the case. In order to make the data linearly separable, a transformation can be applied to the data to make it classifiable with a linear decision boundary. For example, in Figure 3.13b the two classes are not linearly separable. In this setting we are going to make use of the kernels, introduced earlier in this chapter.



(a) (b)

Figure 3.13: (a) An example of a dataset that is linearly separable. (b) In the original space, the two classes cannot be separated by a straight line. In this case we are going to use the "kernel trick", map the data to a space where the classes can be linearly separable.

3.5.2 Evaluating classifiers

In Chapter 4 the performance of different methods for labelling cells as doublets or singlets will be evaluated based on sensitivity and specificity. Sensitivity and specificity are defined based on true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

$$\text{Specificity} = \frac{TN}{TN+FP}$$

To summarise the predictive performance of different set of features we use Precision-Recall curves in Chapter 4.

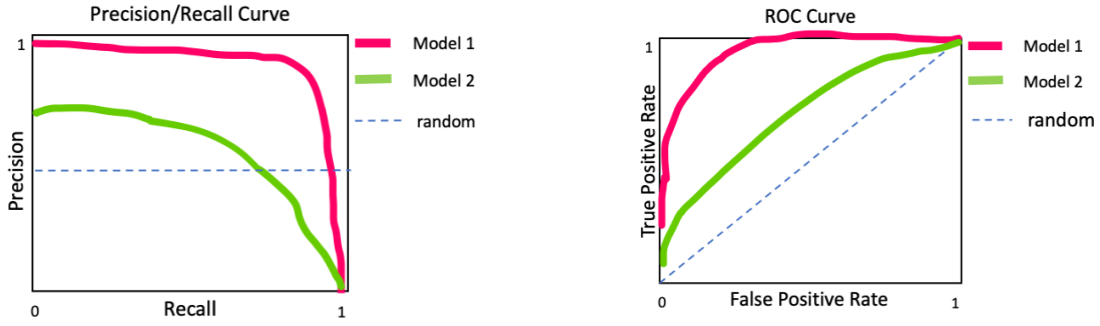
$$\text{Recall} = \text{Sensitivity} = \frac{TP}{TP+FN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

As it can be seen from the formulation of precision and recall, the calculation does not make use of true negatives. The focus is the prediction in the case of the minority class, as such precision-recall curves are good for summarising the performance of classifiers when applied to a problem where there is a class imbalance.

Another way of visualising the performance of a classifier is using a receiver operating characteristic (ROC) curve. In a ROC curve, we plot on the x-axis the false positive rate (1 - specificity) against the true positive rate (also known as sensitivity or recall).

Both precision-recall and ROC curves allow us to see how performance changes as we vary a threshold. This threshold is often a probability, as obtained from a probabilistic classifier for



(a) (b)

Figure 3.14: (a) Precision-Recall curve where Model 1 shows better performance. A better model has both high precision and high recall, making the curve closer to the top right. (b) ROC curve with Model 1 having better performance. A better model has low false positive rate and high true positive rate, making the curve closer to the top left.

example. As the threshold is varied, precision and recall (or true positive and false positive rates) are computed. In the case of precision-recall curves, we aim to have high precision and high recall, the closer the curve is to the right hand top corner the better. Similarly, in the case of the ROC curve, we aim to have low false positive rate and high true positive rate aka the curve should be closer to the top left.

The performance can be quantified by computing area under the curve (AUC). A classifier that is able to perfectly classify the data will have an AUC of 1.

3.5.3 Entropy

In information theory, entropy allows us to measure the heterogeneity or uncertainty of a probability distribution. Let p be the probability distribution of interest, then entropy can be computed as follows:

$$\text{entropy} = -\sum_{i=1}^{i}(p_i * log(p_i))$$

In Chapter 4 we compute entropy per document, in our case cells, following the LDA fit.

3.6 ML approaches within the scope of this thesis

In Chapter 4 the standard LDA formulation (as implemented in scikit-learn) is used to evaluate the suitability of this approach for detecting doublets in scRNA-seq. An LDA is fit on a cell by gene matrix, and the inferred topics per cell are used to compute an entropy score which is in turn used to determine doublets. Furthermore, the proposed approach is then compared with state of the art doublet detection methods.

In Chapter 5 a 2-step LDA is used to identify genes that change as a result of interaction in scRNA-seq. Firstly, an LDA is fit on a reference population of cells that are not considered to be

interacting. Secondly, topics are fixed before fitting another LDA on the interacting population. The topics of the second LDA are used to identify genes that change as a result of interaction. A figure describing the approach and evaluation of using both real and synthetic datasets can be found in Chapter 5.

The focus of Chapter 6 is understanding process dynamics, and an extension of the traditional topic model is applied to pseudotime-ordered scRNA-seq data. Under the proposed approach, topic probabilities change over time and the topic and word dynamics are modelled as GPs.

Chapter 4

Investigating the potential of latent Dirichlet allocation for doublet detection

This chapter presents an application of an LDA-based approach to doublet detection in scRNA-seq data. We hypothesise that if each cell is described as contribution of topics, a doublet will contain the topics of the singlets it has been formed by. We found this method to be unsuitable for detecting doublets in single cells. However, the chapter also includes comprehensive benchmarking results that reflect on the state-of-the-art doublet detection approaches.

4.1 Introduction

Doublets (or multiplets) in scRNA-seq are a result of two (or more) cells being mistaken for a single cell due to being captured within the same droplet in a microfluidic device (AlJanahi et al. 2018, Lareau et al. 2020). Such cells should be accounted for as they might introduce false signal in downstream analysis, for example in differential expression (Ilicic et al. 2016). Furthermore as scRNA-seq has been used to study processes, for example to better understand hematopoietic progenitors or the life cycle of Plasmodium parasites (Haque et al. 2017, Howick et al. 2019, Pellin et al. 2019), doublets can confound trajectory and pseudotime inference (DePasquale et al. 2019). The amount of doublets varies between single cell protocols and with the amount and type of cells sequenced. For example, the rate of the C1 Fluidigm protocol is about 4% while the 10x Chromium percentage of doublets in relation to the amount of sequenced cells can be found in Table 4.1. The C1 system allows for visualising captured cells and users can filter doublets, empty wells, and wells containing cell debris (See et al. 2018). However, in 10x and other droplet based protocols doublets cannot be filtered prior to library preparation. Multiplets can occur in other types of sequencing data as well, for example a recent study of scATAC-seq has shown a multiplet rate of $\sim$ 13-21% (Lareau et al. 2020). While doublets are common, other multiplets can also appear in the data. However, multiplets of more than two cells are considered rare events (DePasquale et al. 2019). Doublets can form within the same cell type or between cell types,

known as homotypic or heterotypic doublets respectively.

Number of Cells Loaded	Number of Cells Recovered	Multiplet Rate
~ 870	~ 500	~ 0.4%
~ 1700	~ 1000	~ 0.8%
~ 3500	~ 2000	~ 1.6%
~ 5300	~ 3000	~ 2.3%
~ 7000	~ 4000	~ 3.1%
~ 8700	~ 5000	~ 3.9%
~ 10500	~ 6000	~ 4.6%
~ 12200	~ 7000	~ 5.4%
~ 14000	~ 8000	~ 6.1%
~ 15700	~ 9000	~ 6.9%
~ 17400	~ 10000	~ 7.6%

Table 4.1: The multiplet rate depends on the number of cells and increases linearly with the number of cells loaded (*10X Genomics* 2020)

A standard step in single cell analysis is quality control (QC), ensuring only viable cells are considered. This QC is usually done based on examining the number of counts per cell, number of genes per cell, and fraction of mitochondrial counts. Cells with high counts are often filtered out as they can represent doublets (Luecken & Theis 2019). However, as can be seen in Figure 4.1 since there is an overlap in the distribution of counts for singlets and doublets, filtering based on arbitrary counts threshold can be insufficient, and a more systematic approach may be preferable.

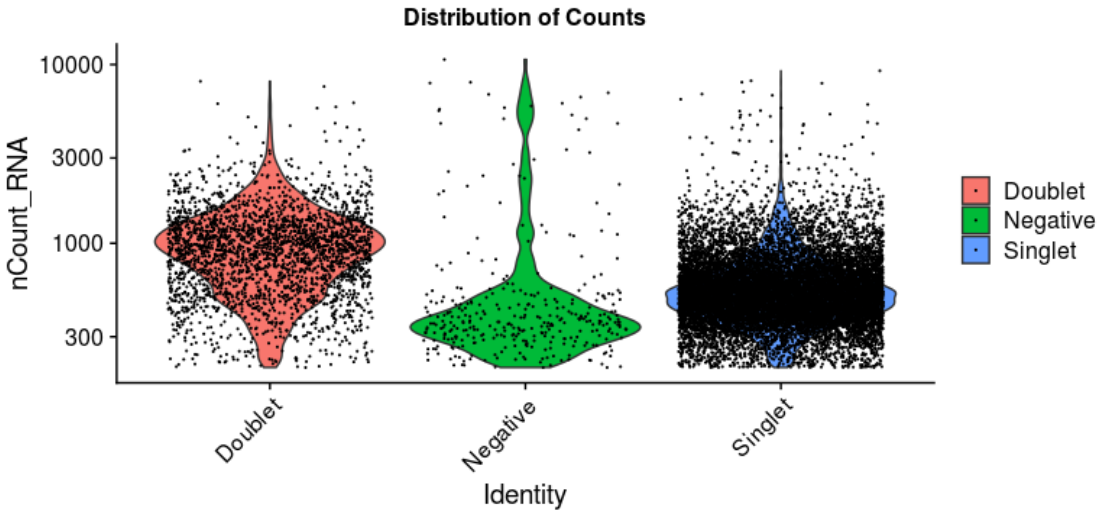


Figure 4.1: Cells annotated as doublets, singlets, and negative (ambiguous) cells based on HTO (hashtag oligonucleotide). Cells positive for more than one HTO are annotated as doublets. As shown on the plot, not all doublets have higher counts than singlets and as such filtering based on counts is not sufficient.

There are lab-based techniques available that mitigate the problem of overloading sequencing machines without increasing the probability of doublet creation. These techniques minimise

batch effects as they allow samples from multiple donors to be sequenced together. One such example is multiplexing techniques. A set of monoclonal antibodies are chosen and combined into identical pools. Each pool is conjugated to a distinct hashtag oligonucleotide (HTO). The HTOs contain a 12 basepair (bp) barcode that can be sequenced alongside the cellular transcriptome (Stoeckius et al. 2018). Demuxlet is another approach for sample multiplexing where doublets can be identified based on sample specific single nucleotide polymorphisms (SNPs) (Kang et al. 2018). Both CellHashing and Demuxlet have been used to sequence samples from multiple donors together. Both methods can identify homotypic and heterotypic doublets between donors, however neither can identify intra-donor doublets.

In addition to the multiplexing techniques, there are also computational methods developed to tackle the problem of doublet detection. Those methods, along with their assumptions and limitations, are discussed in Section 4.2.

4.2 Computational methods for doublet detection

While demultiplexing techniques are available, often samples are run in isolation. As such, doublet detection techniques can be used to avoid introducing false signal during data analysis. Methods for doublet detection generally rely on a similar idea: create artificial or *in silico* doublets by merging different proportions of cells in the data. Variations of how doublet annotation is done will be covered in this section. Generally, doublet detection should be performed on a single sample as artificial doublets might be generated between cells from different samples and in reality those cells do not appear together in the same experiment (McGinnis et al. 2019). Doublet identification could be done after normalising the data for tools that do not require prior clustering (DoubletFinder and Scrublet) or following clustering (DoubletDecon). Figure 4.2 presents a schematic overview of the part of the pipeline in which those tools fit.

4.2.1 DoubletFinder

DoubletFinder randomly samples cells, combines cell profiles, and generates artificial doublets that are a 50/50 contribution of the randomly sampled cells. Those simulated doublets are then added back to the original datasets, normalisation is performed, and real data and artificial doublets are projected in PCA space. The proportion of artificial doublets when merged with the real data (*pN*) is 0.25. However, based on the original publication's experiment where *pN* is varied between 0.05 and 0.3, the performance seems proportion invariant (McGinnis et al. 2019). Annotation of cells as doublets is based on looking at the artificial nearest neighbours. The artificial nearest neighbours proportion (*pANN*) is computed for each cell (McGinnis et al. 2019). Finally, real doublets are identified by taking the top n *pANN* values where n is the number of expected doublets. DoubletFinder aims to identify parameters that produce non-unimodal *pANN* distributions to separate doublets and singlets. Every *pANN* distribution is tested to

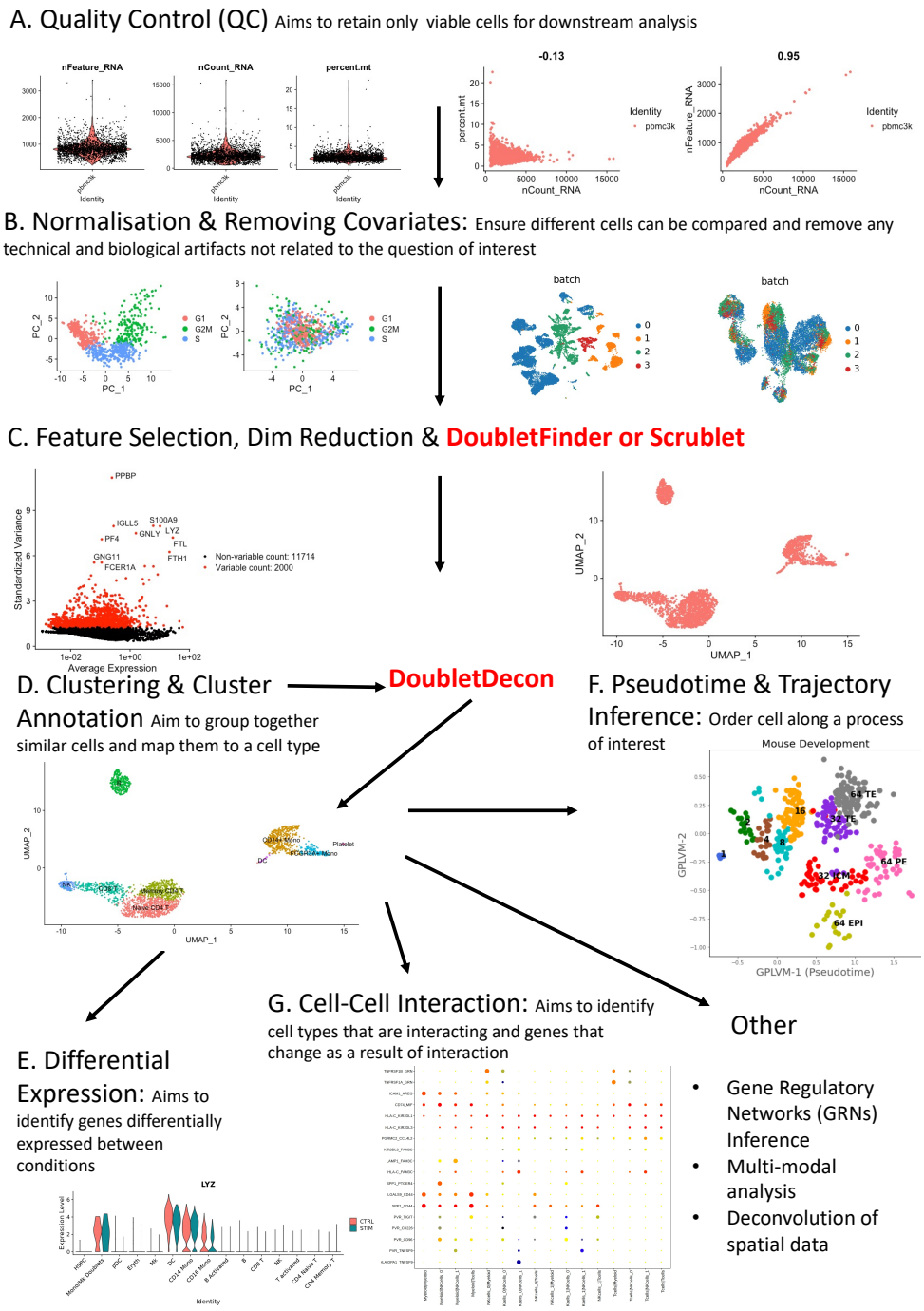


Figure 4.2: Single cell analysis pipeline overview illustrating where doublet detection tools fit. For example, DoubletFinder and Scrublet do not require data to be clustered beforehand. Both methods rely on adding artificially created doublets to the data and then performing single cell analysis steps. DoubletDecon uses clustering to infer the expression profile of each cluster.

identify those with high bimodality coefficient (BC) values, as BC measures deviations from unimodality. The size of the neighbourhood (pK) is dataset specific and is determined using mean-variance-normalised bimodality coefficient ($BCmv$) maximisation. An overview of the steps taken by DoubletFinder can be seen in Figure 4.3.

DoubletFinder simplifies the concept of doublets as it assumes doublets are equal contribution

of two cells. However, in practice a doublet might contain a singlet and a broken cell (Ilicic et al. 2016). Furthermore, it requires a doublet rate to perform the parameter sweep. However, in practice the doublet rate is often unknown. The method assumes homotypic doublets are benign and focuses on detecting heterotypic ones. Lastly, the method cannot distinguish doublets from transitioning cells (McGinnis et al. 2019).

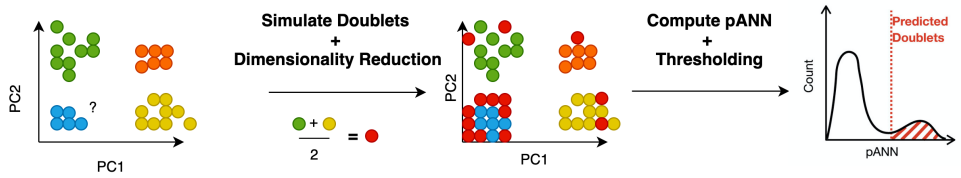


Figure 4.3: Overview of how DoubletFinder works. Figure adapted from (McGinnis et al. 2019)

4.2.2 Scrublet

Similarly to DoubletFinder, Scrublet also relies on simulated doublets and nearest neighbours for labelling cells as doublets or singlets. Synthetic doublets are created by combining the unnormalised counts of randomly sampled pairs of cells. Those synthetic doublets are then projected in the same PCA space as the original data. A KNN graph is then constructed from the union of observed and simulated doublets and a doublet score is calculated as a fraction of neighbours that are simulated doublets (Wolock et al. 2019). Doublet scores are computed for the real cells and the simulated doublets. A doublet threshold is selected based on the score distribution.

Scrublet assumes two main types of doublets, "embedded" and "neotypic". Embedded doublets, similar to homotypic, arise if cells with similar transcriptomes are combined and the impact of such errors is typically small. As such, when labelling doublets Scrublet aims to identify "neotypic" ones as the combination of those transcriptomes can result in new clusters or they can be mistaken for transitioning cells. Scrublet assumes both cell states that contributed to the doublet formation are present in the dataset. Scrublet's performance decreases if applied to complex continuous manifolds, e.g. processes where transitions or branching are captured (Wolock et al. 2019).

4.2.3 DoubletDecon

Compared to the previously mentioned methods, DoubletDecon requires prior clustering to label cells as doublets or singlets, as seen in Figure 4.2. Unlike DoubletFinder and Scrublet, the synthetic doublets can be created as 50/50 or a weighted average of 30/70 and 70/30 contributions of singlets sampled from two distinct clusters. The number of created doublets depend on the dataset, but 10% has been identified as a value that works well (DePasquale et al. 2019). The profile of each cell is deconvoluted as a contribution of cluster centroids. Next, the profile of

each cell in the data is compared with the profiles of cells in each cluster and the profile of the synthetic doublets cluster. If the profile of the cell is highly correlated to the doublet cluster profile, it is labelled as a doublet. Cells are then reclustered based on their deconvoluted profiles with annotated doublets removed from their original clusters. In the last step of the DoubletDecon algorithm, some cells previously annotated as doublets can be "saved" based on unique gene expression profiles. In this stage, gene by gene comparison is performed and if the number of unique genes in a cluster exceeds a threshold, cells initially labelled as doublets are reannotated as "singlets" and returned to the dataset (DePasquale et al. 2019).

In common with the tools discussed already, DoubletDecon assumes homotypic doublets are benign and aims to identify heterotypic ones. DoubletDecon makes the following assumptions about the clustering input: (1) transcriptionally similar cells are merged in the same cluster and no two clusters have similar transcriptomic profiles, (2) for a doublet to be detected, a cluster for each of the two contributing cell types must be present in the data, and finally (3) the dataset should not contain a doublet cluster. Unlike other tools, DoubletDecon can distinguish doublets from transitioning cells if the transition is defined by a unique set of genes compared to the end states.

4.2.4 Summary

All tools discussed aim to identify doublets in the data by creating artificial doublets that are added to the original dataset. Sampling of singlets is done at random or from already defined clusters. However, while some tools generate synthetic doublets as an equal contribution of two cells, a multiplet might contain a high-quality singlet and a broken cell (Ilicic et al. 2016). All tools assume homotypic doublets are benign and aim to identify heterotypic ones. Furthermore, all tools have poor performance on datasets of transcriptionally similar cells, e.g. cell subtypes.

Only one of the discussed mentioned methods (DoubletDecon) is able to distinguish between doublets and transitioning cells, as long as the transition is defined by a unique set of genes. As scRNA-seq is used to study disease and developmental processes, this is a vital feature.

Furthermore, some of the assumptions make those methods difficult to use in practice. For example, DoubletFinder requires a known doublet rate, which sometimes cannot be determined for a sequencing experiment. DoubletDecon makes the assumption there is no doublet cluster in the data which does not conform with the output of some clustering algorithms.

In order to address some of those issues, this chapter presents a method based on latent Dirichlet allocation. The proposed approach does not require generation of synthetic doublets, prior knowledge of doublet rate, or clustering assignment. The proposed approach is applied to real and synthetic data and benchmarked against DoubletFinder and DoubletDecon. Scrublet is not taken forward in the analysis due to its similarity with DoubletFinder. It is worth noting that while the real datasets used for evaluation contain multiple donor samples and while demultiplexing techniques can distinguish between donors, doublets within the same donor remain undetected by

such methods.

4.3 Applications of LDA to scRNA-seq data

LDA has been applied to different types of omics data (Liu et al. 2016, Rogers et al. 2005), with particular applications of RNA-seq, ATAC-seq, and Hi-C in single cell. In the context of LDA, cells are equivalent to documents and genes (regions in ATAC-seq and locus-pairs in Hi-C) are words. Topics can be described as groups of genes whose expression co-varies (Bravo González-Blas et al. 2019, Kim et al. 2020, Kotliar et al. 2019). The identified topics can be interpreted as general, cell type specific, or linked to the technical quality of the samples. For example, ribosomal or mitochondrial-dominated topics might correspond to dying cells. In addition to the standard implementation of LDA, work has been completed to allow for simultaneous topic identification and cell clustering (Campbell et al. 2020). Furthermore, CellTree (duVerle et al. 2016) uses LDA for trajectory inference: the method takes LDA in its standard form but computes chi-square distance between cells, and uses the distance to build a tree to describe a branching process. Most recently, a modified version of LDA has been used to decontaminate counts from ambient RNA: DecontX assumes counts come from two topics, native counts and ambient RNA. Using only native counts improves clustering and downstream analysis (Yang et al. 2020).

Chapter 3 describes LDA, how it extends standard mixture models, the generative process and inference in the context of documents. This section focuses on LDA in the context of scRNA-seq data.

In the context of scRNA-seq, cells correspond to documents and genes to words. Word frequencies are replaced by counts. We obtain a set of topic distributions over cells and a per-topic gene distribution. Given D cells (indexed $d = 1,..., D$), N genes (indexed $n = 1,..., N$), K topics, z_{dn} denotes the assignment of the n-th gene in the d-th cell to the k-th topic. We can define the generative process as follows:

$$\begin{aligned}
\boldsymbol{\phi}_k &\sim \text{Dir}(\boldsymbol{\beta}), k = 1...K \\
\boldsymbol{\theta}_d &\sim \text{Dir}(\boldsymbol{\alpha}), d = 1...D \\
z_{dn} &\sim \text{Multinomial}(\boldsymbol{\theta}_d) \\
w_{dn} &\sim \text{Multinomial}(\boldsymbol{\phi}_{z_{dn}})
\end{aligned} \tag{4.1}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are vectors of lengths K and V, where V is the size of the vocabulary (all genes in the dataset). $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the parameters defining the Dirichlet priors over document-topic and topic-word multinomials. These parameters control the sparsity of the model.

4.4 Materials and methods

4.4.1 LDA and entropy scoring

The formulation of LDA for scRNA-seq data has been presented in the previous section. Here, the first step of the analysis is to fit LDA on the cell by gene expression matrix as seen in Figure 4.4. LDA allows us to obtain probabilities over topics for each cell and probabilities over genes for each topic.

The topic-cell probabilities matrix, denoted $\boldsymbol{\theta}$, containing a row for each cell d with contribution of all k topics, is taken forward for entropy scoring (Step 2), see Figure 4.4. In a heterotypic doublet, we expect to find the topics appearing in the singlets of the cell types that contributed to the doublet creation. However, for a homotypic doublet the topics contributing to the two cells will be the same or very similar. To quantify this we compute entropy for each cell.

In Chapter 3 we have introduced the concept of entropy for a probability distribution. Specifically here, we compute the entropy score per cell d based on the probability distribution over topics. We hypothesise that multiplets will have higher entropy as they will be explained by the topics that contribute to the cell types they are made of. However, this is only the case for heterotypic doublets as they will have contribution from different cell types, while homotypic doublets will have entropy similar to singlets as they are made up of one cell type and described by the topics of that cell type. Let $\boldsymbol{\theta}_{dk}$ be probabilities over topics for each cell, then the entropy for a cell can be computed as:

$$\text{entropy} = -\sum_{k=1}^{k}(\boldsymbol{\theta}_{dk} * log(\boldsymbol{\theta}_{dk}))$$

It is this entropy scoring that can be subsequently used to label cells as doublets or singlets based on a proposed cutoff. This entropy cutoff is user-defined and based on examining the distribution of entropy scores for each cell in the dataset. If a high entropy cutoff is chosen, the specificity score will improve but sensitivity will suffer. However, if the entropy cutoff is relaxed, additional singlets will be labelled as doublets, and clusters might be lost. The effect of using different entropy scores can be seen in Figure 4.7.

4.4.2 Datasets

Demuxlet dataset

The pre-processed dataset, originally available from the Demuxlet paper, is available on Gene Expression Omnibus (GEO) under accession number GSE96583, GSM2560248 (Kang et al. 2018). The dataset consists of peripheral blood mononuclear cells (PBMCs) from eight donors. The ground truth is determined using Demuxlet, a demultiplexing technique based on genetic variation between donors. Cells annotated as ambiguous by Demuxlet are not considered as part of the evaluation. Cell annotations are available at https://github.com/yelabucsf/demuxlet_paper_code.

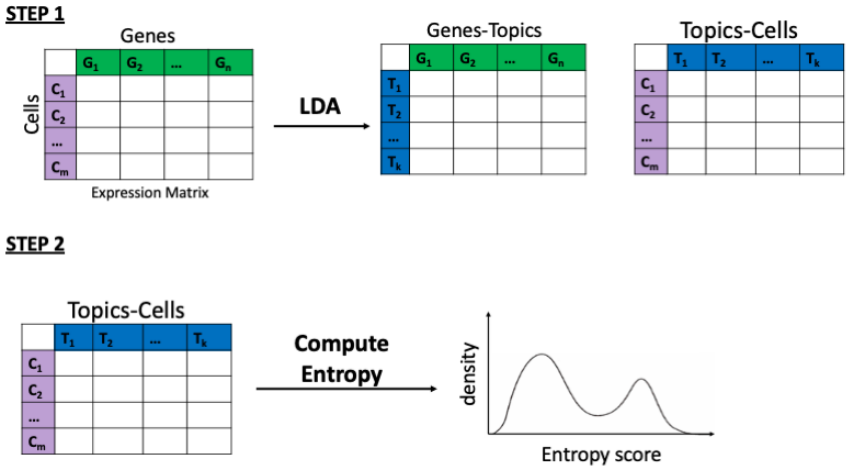


Figure 4.4: Firstly, in step 1 LDA is used to decompose the original expression matrix. Next, in step 2 using the topics-cells distributions entropy is computed, and an entropy cut-off is selected to determine whether cells are doublets or singlets.

Singlets	Doublets	Ambiguous
13030	1565	24

Table 4.2: Cells annotated by Demuxlet using donor SNP information.

Cell Hashing dataset (PBMCs)

A pre-processed dataset was downloaded from Seurat 3.1; FASTQ files and processed data are also available from GEO under accession number: GSE108313. Singlet or doublet labels, used as ground truth in the next sections, were obtained by using the HTODemux() function from Seurat v.3.1. Each cell has RNA counts and HTO counts. Demultiplexing is done by performing k-means clustering on the normalised HTO values. For each HTO, a negative distribution is calculated and the cluster with the lowest average value is used as the negative group. A negative binomial distribution is fitted to the negative cluster. The 0.99 quantile is used as a threshold and based on this threshold a cell is classified as positive or negative for a particular HTO. If a cell is positive for more than one HTO, it is classified as doublet. Cells that cannot be labelled as singlets or doublets, have been labelled as negative and have been removed for the purposes of the benchmarking analysis in this chapter.

This dataset consists of PBMCs from eight different donors. Cell labels as determined by the HTO demultiplexing:

Singlets	Doublets	Negative
13972	2598	346

Table 4.3: Cells annotations based on Cell Hashing, PBMC data

Cell Hashing dataset (cell lines)

A pre-processed dataset was downloaded from Seurat 3.1 and FASTQ files and processed data are also available from GEO under accession number: GSE108313. Demultiplexing was performed as described above. The dataset contains single cells from four cell lines: human embryonic kidney (HEK), human blast crisis chronic myeloid leukaemia (K562), human acute myeloid leukaemia (KG1), and mixed lineage leukaemia (THP1). Each cell line was split into three samples and doublets were detected both across and within cell types. Ground truth includes homotypic doublets between samples but not within the same sample. As negative cells cannot be classified as doublets or singlets, they have been removed from the benchmarking analysis.

Singlets	Doublets	Negative
6489	1465	239

Table 4.4: Cells annotations based on Cell Hashing, cell lines data

Synthetic data

Synthetic doublets were created on the basis of a real dataset of 2639 PBMCs (cells left after filtering). Following the standard steps of scRNA-seq analysis, the main clusters of the data were identified. Doublets were assumed to have equal contribution from the two cell types that generated them, and the cells they consist of were sampled from clusters with distinct transcriptional profiles. The following doublets were generated:

- 250 doublets between B-cells and CD14+ monocytes
- 200 doublets between memory CD4 T-cells and CD14+ monocytes
- 300 doublets between B-cells and CD4 T-cells

4.4.3 Metrics

Doublets identified by our proposed approach and the tools summarised earlier (DoubletFinder and DoubletDecon) are compared to the ground truth as defined for each dataset. The following metrics are computed for each dataset and approach: TP (true positive), TN (true negative), FP (false positive), FN (false negative), sensitivity, and specificity. Sensitivity and specificity we defined in Chapter 3 with respect to evaluating classifiers. High sensitivity means more possible doublets are removed from the analysis, although a low FP rate is also desirable as additional FPs would mean loss of cells, with a corresponding loss of clusters in the downstream analysis.

When computing TP, TN, FP, and FN for our cell hashing or SNP demultiplexed datasets, it should be noted those methods can also annotate homotypic doublets, but doublets of the same donor (heterotypic and homotypic) remain in the data. As a consequence, some FPs can in fact

correspond to true doublets not annotated by the protocol. Additionally, there are undetected homotypic doublets that appear in the FNs.

4.4.4 Evaluating the predictive performance of different sets of features

While entropy can be a useful feature as described earlier, cut-offs can affect results. Additionally, entropy in isolation cannot be used for homotypic doublets and in the standard QC analysis total counts per cell are often used for filtering analysis. Finally, while UMAP is often used primarily for visualisation of scRNA-seq data, as a non-linear dimensionality reduction technique, dimensions can be interpreted (similarly to what PCA components capture). Thus UMAP coordinates can be treated as features, as they preserve global and local distances. It is expected that heterotypic doublets will be in proximity to both cell types that created them.

Next, features like entropy, total counts per cell, and UMAP coordinates are evaluated for whether they are predictive of whether a cell is a doublet or a singlet. Those four features are taken forward and evaluated if each, all or some combination of them can facilitate doublet annotation. To achieve this, we use those features as an input to a classification method.

In this particular instance we are going to use support vector machines (SVMs) as binary classifiers to evaluate whether we can correctly classify cells as doublets or singlets. SVM has been introduced in Chapter 3, Section 3.5.1. In the case of classifying cells as doublets or singlets, we use SVM with an RBF kernel.

Since doublets are a smaller proportion of the total number of cells in each dataset, we ensure the training set is representative of the proportions of singlets and doublets that are expected. Results are displayed as recall-precision plots. Recall-precision curves have been introduced and discussed in Chapter 3.

4.4.5 LDA and SVM implementation

Experiments are performed in Python 3.6.5. The LDA implementation in this chapter is based on scikit-learn (sklearn.decomposition.LatentDirichletAllocation). For each of the datasets 30 topics (n_components) are used. From the LDA model, topic probabilities over documents and word probabilities over topics are obtained. While we have not performed an exhaustive search for the most appropriate number of topics in this chapter compared to Chapters 5 and 6, we have shown in Chapter 5 that under-specifying the number of topics can be more detrimental to performance than using too many topics. Here we use 30 topics for each of the datasets as we have several different cell types in the PBMCs datasets (different types of T-cells, B-cells, monocytes, and others). We expect to find cell type specific topics but also topics that correspond to general biological processes. In this case we use the default number of iterations. A good practice for determining if the model has converged would be to compute perplexity for a number of different iterations, similarly to how the number of topics evaluation is done.

To compute the entropy for each cell, entropy is used from scipy (version 1.4.1), where *pk* (distribution for computing entropy over) is the topic contributions for each cell. Entropy cut-off values for each dataset have been indicated in the corresponding results tables. The effects of the entropy cut-off are shown in Figure 4.7. While for the Demuxlet dataset, we observe a shift in the distributions of doublets and singlets in terms of entropy, this is not the case for the Cell Hashing data. Entropy cut-offs are chosen in order to facilitate the split of the two distributions.

For SVM, the scikit-learn implementation of non-linear binary SVM is used (svm.NuSVC). RBF kernel is used as the kernel parameter. The probability parameter is set to True. This ensures we can use the predict_proba function for the classification of points in the test set. This is later used for plotting recall-precision curves. All other parameters are kept as defaults. To ensure separation of data for cross-validation similar to the underlying proportions of doublets and singlets in the data, stratified k-fold (StratifiedKFold) from scikit-learn is used and data are split into two folds. To generate the precision-recall curves, we use the precision_recall_curve function with parameters the true labels of the test set and the predicted probabilities for those points.

4.4.6 Analysis with DoubletFinder and DoubletDecon

Both packages were run in R 3.6.0. Both packages were installed from the GitHub project page from the original publications:

- https://github.com/chris-mcginnis-ucsf/DoubletFinder
- https://github.com/EDePasquale/DoubletDecon

DoubletFinder To run DoubletFinder, functions compatible with Seurat v3 were used, paramSweep_v3 and doubletFinder_v3. The first 10 principal components were used with paramSweep_v3. The number of generated artificial doublets, *pN* was kept to default, 25%. Both the Demuxlet dataset and the Cell Hashing PBMCs have been analysed in the original publication with PCs and *pN* set to defaults (McGinnis et al. 2019). For both Demuxlet and Cell Hashing PBMC datasets we achieve results similar to the original publication. The output of the parameter sweep was used to determine the neighbourhood for doublet scoring, *pK*. This value is dataset specific.

To illustrate the effect of the doublet rate on the results, different values were used for the Demuxlet PBMCs: 8% (under-specified doublet rate), 10.9% (estimated doublet rate from the original Demuxlet publication), 11.5% (sub-optimal high value for doublet rate), 12.5% (DoubletFinder estimated doublet rate).

DoubletDecon The MainDoubletDecon function was used for labelling the doublets. Species was set to "hsa". As DoubletDecon requires clustering assignments, for each dataset clustering was performed in Seurat. Different proportions of doublets were allowed to be set with the only50 parameter set to False.

Complete analysis can be found in https://github.com/alexpancheva/doubletsAnalysis

4.5 Results

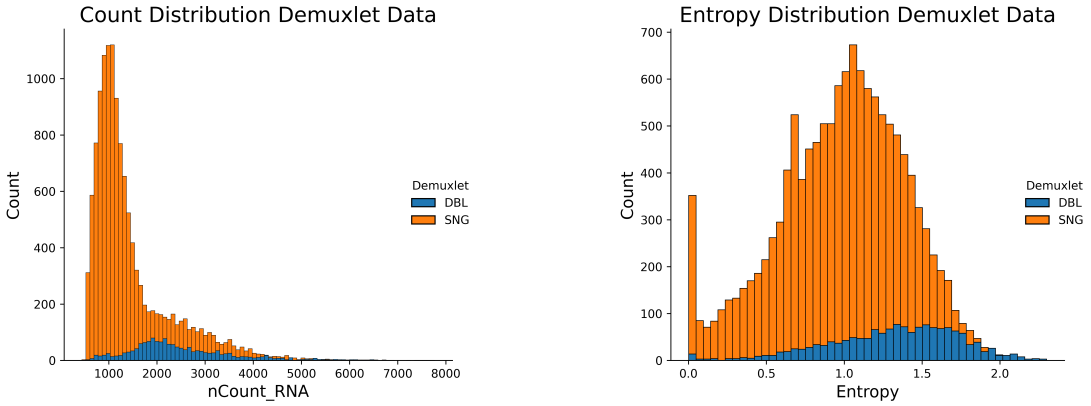
In order to benchmark the proposed LDA and entropy scoring approach against existing methods, three real datasets are used for evaluation. The first dataset, referred to as Demuxlet, has been annotated based on donor SNP information and contains PBMCs. Another PBMC was annotated based on HTOs (Cell Hashing), and finally a cell lines dataset was also annotated with HTOs. The performance of all methods when applied to transcriptionally similar cells is demonstrated in the case of the HTOs cell lines data. Furthermore, the sensitivity of each method to lower number of UMI counts and unique genes is demonstrated by downsampling the Demuxlet data. Finally, the predictive performance of different sets of features is evaluated for Demuxlet and Cell Hashing datasets.

The results below show the performance of only DoubletFinder and DoubletDecon as Scrublet is similar to DoubletFinder in terms of doublet generation, no prior clustering requirement, and use of nearest neighbours for doublet annotation and as such it is not included in the results.

4.5.1 Demuxlet dataset

Earlier when describing entropy scoring, it was hypothesised that doublets would have higher entropy than singlets as they would have contribution from the topics characterising each cell type. As such a shift in the entropy distribution of doublets compared to singlets is expected. However, this would only be the case for heterotypic doublets as they will have contributions from multiple cell types. To evaluate to what extent this holds true for real data, entropy and counts distributions for singlets and doublets are plotted based on Demuxlet annotations. Figure 4.5 demonstrates that there is a shift in the entropy distribution of the doublets, as expected. However, there are also cells annotated as doublets, but which have low entropy (around 0.5). Those are potentially some homotypic doublets that were annotated by Demuxlet based on the donor SNP information. However, annotating such doublets solely using gene expression is not feasible.

While the original Demuxlet paper estimates a doublet rate of 10.9% this number only accounts for inter-donor doublets and therefore, the doublet rate was adjusted to 12.5%. This results in a sensitivity of 73.35% for DoubletFinder which is similar to the value reported in the original publication (Kang et al. 2018, McGinnis et al. 2019). Both DoubletDecon and the entropy analysis have lower sensitivity scores (about 56%). However, using the entropy for labelling doublets results in higher specificity score compared to DoubletDecon. While achieving high sensitivity is important, reducing specificity and over-filtering FPs might result in losing a population of cells. Sensitivity and specificity scores for this dataset can be found in Table 4.5.



(a) (b)

Figure 4.5: (a) Counts distribution for doublets and singlets as annotated by Demuxlet based on SNPs. (b) Entropy distribution for doublets and singlets. There is an evident shift in the distribution of entropy scores in the case of doublets.

Method	TP	FP	FN	TN	Sensitivity	Specificity
DoubletFinder doublet rate= 12.5%	1148	679	417	12351	73.35%	94.79%
DoubletDecon	877	6067	688	6963	56.03%	53.43%
Entropy cutoff = 1.3	879	3027	686	10003	56.17%	76.77%

Table 4.5: Performance of methods for the Demuxlet dataset

This dataset contains labelled homotypic inter-donor doublets as they can be distinguished based on donor SNP information. However, as those doublets are generally made up of the same cell type, they cluster with the singlets and the only potential way of distinguishing them is based on counts if they do have higher counts. On the UMAPs, Figure 4.6 those are the doublets that appear surrounded by singlets. As a consequence, the sensitivity of all methods is affected. Additionally, some of the heterotypic doublets within donors are not annotated and as a result some of the FPs might correspond to doublets not detected by Demuxlet.

While in Table 4.5 a single entropy cutoff value is chosen, it is important to understand the effect of this cutoff on the performance of the precision and recall, as illustrated by Figure 4.7. Figure 4.7 demonstrates that choosing an entropy cutoff for labelling cells as doublets comes with a trade-off: low entropy results in high recall but too many singlets are discarded which can affect downstream analysis. High entropy leads to potential doublets remaining in the data. For comparison with the other available methods, an entropy cut-off value of 1.3 was chosen in an attempt to separate the entropy distributions of doublets and singlets, the entropy distribution of doublets peaks at entropy of 1.5.

While DoubletFinder performed well, it is important to see to what extent the doublet rate parameter affects results, as in a standard sequencing experiment the exact value might not be available. In order to evaluate the effect of doublet rate on sensitivity, DoubletFinder is used with different doublet rate values. A multiplet rate of 8% is chosen as a lower bound based

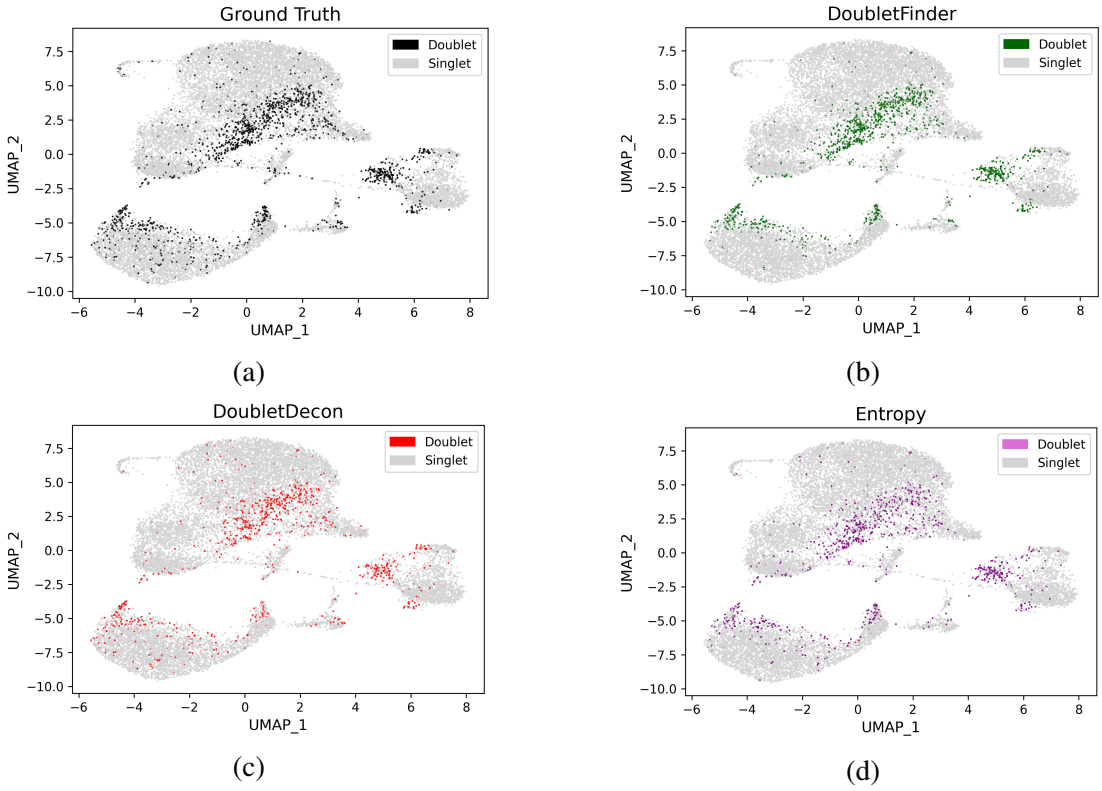


Figure 4.6: Demuxlet PBMCs ground truth and doublets correctly identified by the different methods. (a) UMAP projection of ground truth doublet as identified by the Demuxlet SNP annotation. Some of the doublets shown are potentially homotypic doublets between different donors. (b) Correctly annotated doublets by DoubletFinder. (c) True positive doublets annotated by DoubletDecon and (d) LDA and entropy scoring true positives

on what can be expected as doublet rate in a 10x Chromium experiment. 10.9% is the doublet rate the Demuxlet paper reports, however it does not account for intra-donor doublets. Table 4.6 illustrates that unsuitable doublet rate can reduce the sensitivity significantly. Using an unsuitable doublet rate of 8% results in worse performance of DoubletFinder compared to both entropy scoring and DoubletDecon. While the values of 10.9% and 11.5% result in sub-optimal sensitivity performance, DoubletFinder still has the highest specificity. DoubletFinder also offers the opportunity to adjust the doublet rate for the presence of homotypic doublets. Results of 12.5% doublet rate plus adjustment for homotypic doublets are shown in Table 4.6.

Doublet rate	TP	FP	FN	TN	Sensitivity	Specificity
8%	822	348	743	12682	52.52%	97.33%
10.9%	1042	551	523	12479	66.58%	95.77%
11.5%	1083	598	482	12432	69.2%	95.41%
12.5% + adj	1148	679	417	12351	73.35%	94.79%

Table 4.6: DoubletFinder's performance with different doublet rates. The last value 12.5% + adj refers to adjusting the doublet rate for the presence of homotypic doublets.

Finally, we evaluate the predictive performance of entropy, counts, and UMAP coordinates in

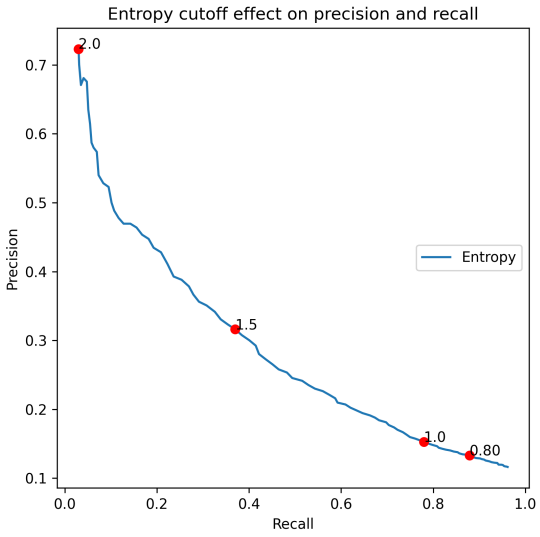


Figure 4.7: Entropy values equally spaced between 0.5 and 2. Low entropy cutoff results in high recall but too many singlets end up labelled as doublets. The plot illustrates the trade-off when choosing an entropy value.

the context of a classification problem. We fit SVM on the data using those features. Precision-recall results can be seen in Figure 4.8. As illustrated from the UMAPs of the annotated doublets (see Figure 4.6), most doublets appear to be around the edges of the clusters. The best-performing classifier's features include entropy and UMAP coordinates. However, the total counts on their own or combined with other features do not perform well. When the total counts per cell are included in the features, the classifier performs worse. This is possibly due to the fact that there is an overlap between the counts distributions for doublets and singlet as evident from Figure 4.9a. From Figure 4.5 count distributions of singlets and doublets are mostly overlapping, so this is potentially a noisy feature that affects SVM performance in some regions. Even the best-performing set of features for this dataset only achieves AUC of 0.47 (AUC closer to 1 indicates a better performing model). AUC in the context of classification was introduced and discussed in Chapter 3.

4.5.2 Cell Hashing (PBMCs)

Next, we evaluated performance based on a Cell Hashing dataset of PBMCs. DoubletFinder again appears to have the highest sensitivity (66.58%), consistent with previous results. This sensitivity score is similar to the result reported by the DoubletFinder paper (McGinnis et al. 2019). However, this sensitivity score is improved by the authors as they remove homotypic doublets and sensitivity becomes 82%. Compared to the previous dataset of PBMCs, DoubletDecon is the worst affected, with sensitivity 26.75%.

Similarly to previous results, methods are not able to identify homotypic doublets as they are

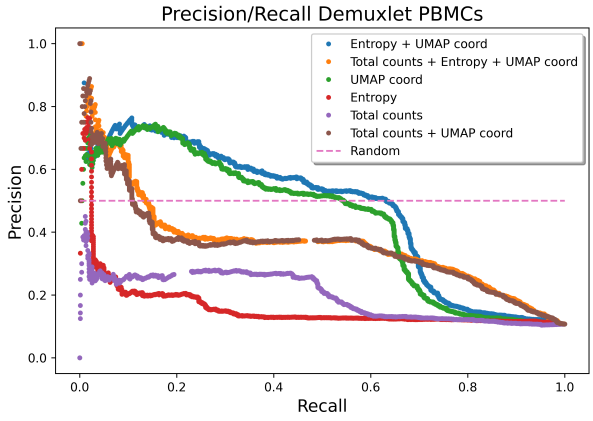
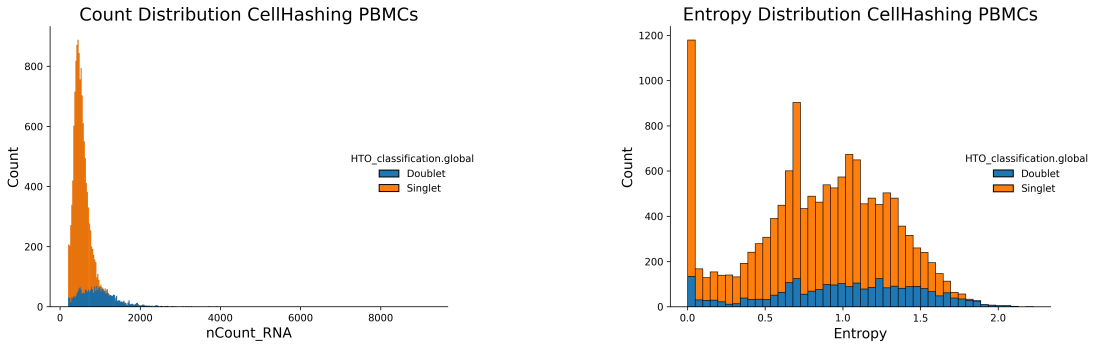


Figure 4.8: Precision-recall curves using different sets of features with non-linear SVM. Similarly to our analysis before, it is evident that solely using the entropy score results in a precision-recall trade-off. UMAP coordinates combined with entropy appear to be the best-performing classification features. The AUC (area under the curve) for the best-performing set of features for this dataset is 0.47.



(a) (b)

Figure 4.9: (a) Counts distributions of singlets and doublets based on the HTO annotation. (b) Entropy distributions for singlets and doublets as annotated by HTOs. In this dataset, there does not appear to be a shift in the entropy distribution.

Method	TP	FP	FN	TN	Sensitivity	Specificity
DoubletFinder	1730	968	868	13004	66.58%	93.07%
DoubletDecon	695	1692	1903	12280	26.75%	87.89%
Entropy=1.2	999	3368	1599	10604	38.45%	75.89%

Table 4.7: Performance of methods for Cell Hashing PBMCs dataset

transcriptionally similar to the singlets and as it can be seen from the UMAPs those doublets are most likely to be projected alongside singlets.

As sensitivity and specificity scores are computed against ground truth that only detects inter-donor doublets (both homotypic and heterotypic) whereas intra-donor doublets are not annotated, it is likely some of the FPs correspond to real doublets.

The predictive performance of features is also evaluated. Compared to our Demuxlet results,

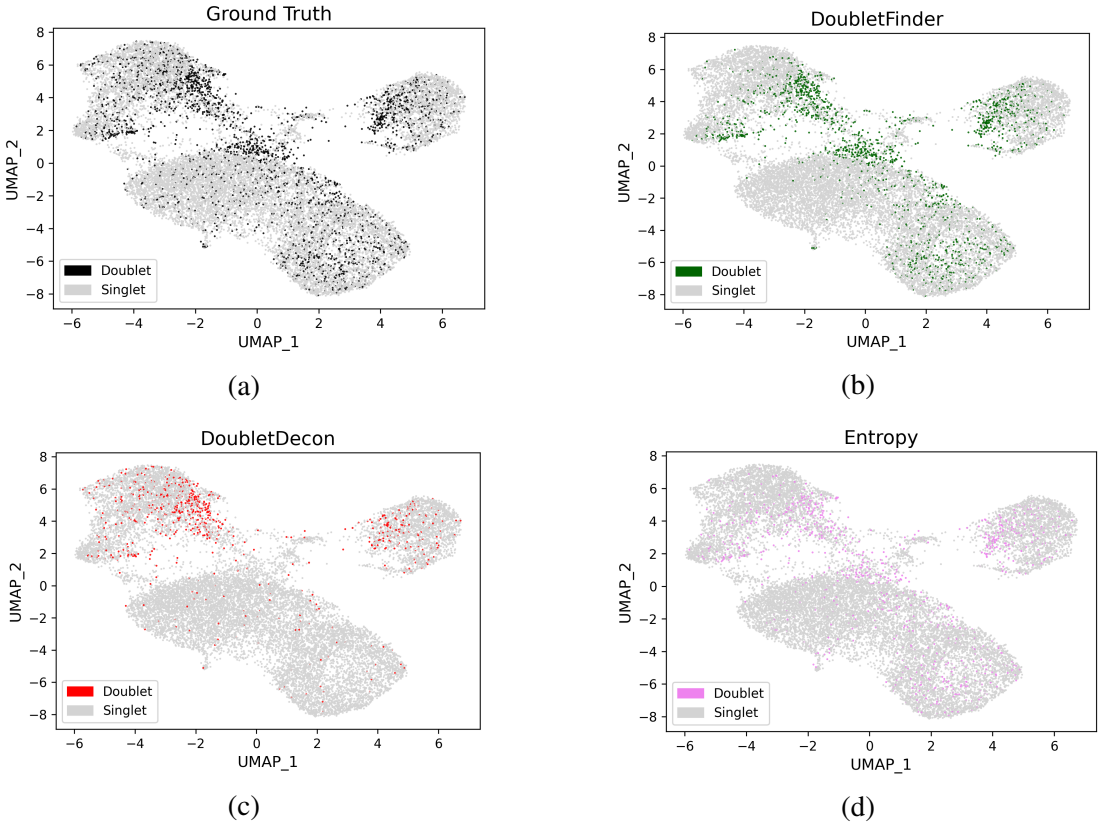


Figure 4.10: Ground truth (a) and doublets identified correctly by each method: (b) DoubletFinder, (c) DoubletDecon, and (d) LDA with entropy scoring.

here UMAP coordinates do not prove to be a very predictive feature because doublets are no longer mostly projected along the edges (see Figure 4.10). Entropy on its own performs better than the UMAP coordinates. However, even the best-performing classifier achieves only 0.30 AUC compared to the results on the Demuxlet dataset with AUC of 0.47. The overlap of the singlets and doublets distributions (see Figure 4.5) makes the entropy not a very predictive feature. Interestingly, in some regions counts appear to be performing well, this is perhaps because counts are generally lower for some cells, as shown in Figure 4.9.

While both the Demuxlet dataset discussed earlier and this Cell Hashing dataset use PBMCs, this dataset differs in total counts and number of features. We evaluate the effects of downsampling in Section 4.5.4.

4.5.3 Cell Hashing (cell lines)

The last dataset used for evaluation contains cells from 4 cell lines and 3 labelled replicates for each cell line, for a total of 12 HTOs. Compared to previous experiments, all tools suffer a drop in sensitivity with DoubletFinder having highest sensitivity of 34.06%. Previous experiments have shown that combining LDA and entropy cutoff generally performs better than DoubletDecon, but in this particular setting DoubletDecon comes second with a sensitivity of 14.47%.

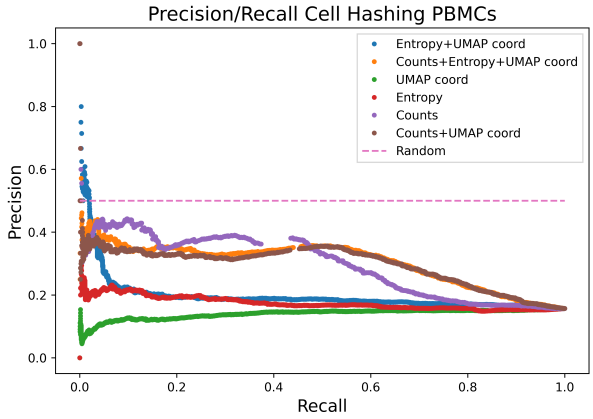


Figure 4.11: UMAP coordinates and entropy are not predictive features. Feature sets involving counts seem to have better predictive performance. However, the highest AUC score is 0.30 compared to 0.47 for the Demuxlet dataset.

However, accounting for the specificity scores, our method has highest specificity compared to DoubletFinder and DoubletDecon.

Method	TP	FP	FN	TN	Sensitivity	Specificity
DoubletFinder	499	1092	966	5397	34.06%	83.17%
DoubletDecon	242	696	1223	5793	16.51%	89.27%
Entropy=1.25	212	492	1253	5997	14.47%	92.41%

Table 4.8: Methods' performance for Cell Hashing cell lines dataset

As this dataset contains annotated homotypic and heterotypic doublets, the TP discovered by each method can themselves be split into homotypic and heterotypic, see Table 4.9. The dataset contains 299 homotypic doublets and 1166 heterotypic doublets. As expected, the majority of the doublets all methods discover are heterotypic.

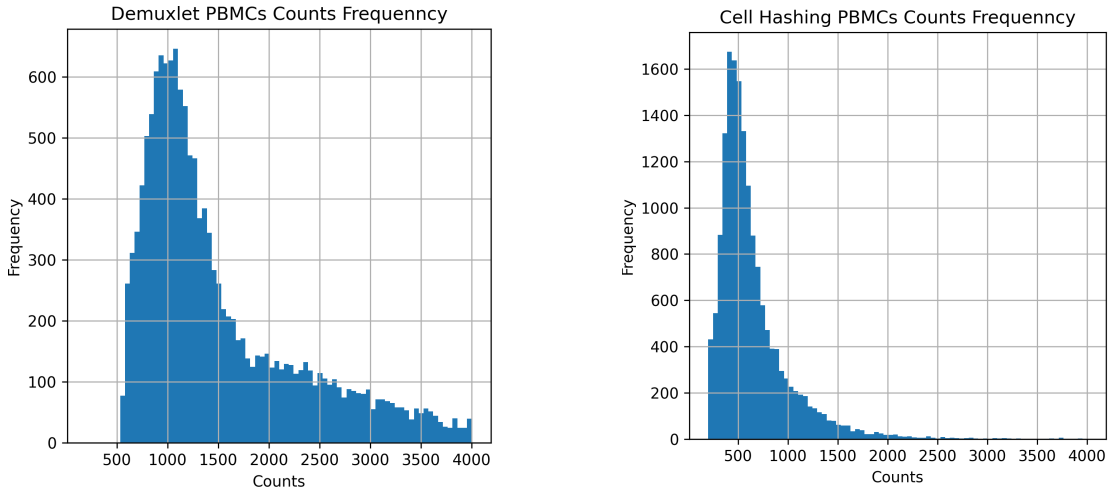
Method	Total Identified	Homotypic	Heterotypic
DoubletFinder	499	72	427
DoubletDecon	242	33	209
Entropy=1.25	212	16	196

Table 4.9: Comparing doublet types identified by each tool

Overall, the poor performance of all methods on this dataset could be explained by the transcriptional similarity of the cell types. For example, two of the cell lines used are myeloid and as such, it is expected that DoubletDecon and LDA combined with entropy might not be able to identify the doublets between those cell lines.

4.5.4 Downsampling and effect on performance of methods

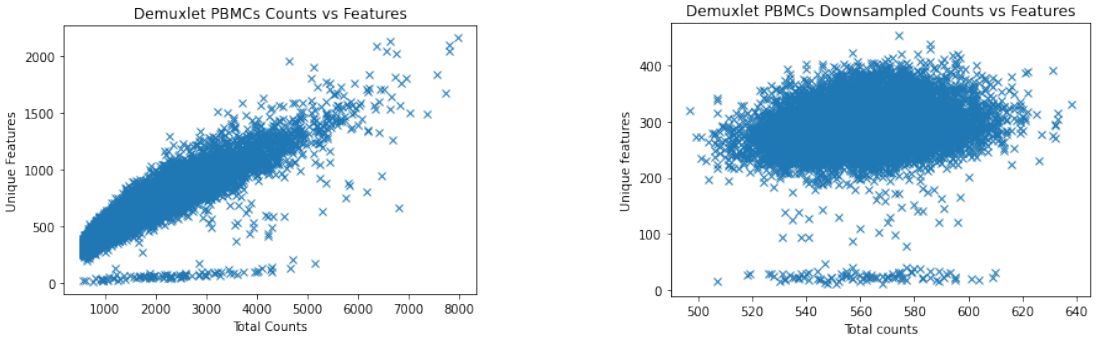
Across the two datasets of PBMCs, the performance of DoubletFinder was relatively stable, with sensitivity scores of 73.35% and 66.58% when applied to the Demuxlet and the Cell Hashing datasets respectively. However, DoubletDecon's sensitivity decreased almost by half between the two datasets. As seen in Figure 4.12, the Demuxlet dataset has higher sequencing depth (distribution peak around 1000) compared to the Cell Hashing data of PBMCs, with the peak of the distribution around 500. DoubletDecon is affected by the lower sequencing depth as it relies on clustering information and number of unique genes for the annotation of putative doublets and the "rescue" step at the end ((DePasquale et al. 2019)).



(a) (b)

Figure 4.12: Comparing sequencing depth of two datasets of PBMCs. (a) Demuxlet counts distribution with peak near 1000 counts. (b) Cell Hashing data with peak near 500 counts.

To investigate the effect of sequencing depth on method performance, we sub-sample the Demuxlet dataset. The results of downsampling the data are shown in Figure 4.13.



(a) (b)

Figure 4.13: Effect of downsampling on number of unique features and total counts (a) Counts vs Unique Features on the initial Demuxlet datasets. (b) Downsampled Demuxlet dataset.

DoubletFinder resulted in the highest sensitivity and specificity scores, the doublet rate was set to 12.5% for the downsampling experiment. DoubletFinder identifies 262 out of 1565 doublets while labelling 1555 cells falsely as doublets. Additionally, DoubletDecon does not identify any doublets. DoubletFinder relies on a dimensionality reduction for generating doublets and computing nearest neighbours. Therefore, it is challenging to determine doublets and singlets due to the lower number of features and potentially low number of informative genes. Similarly, DoubletDecon relies on computing expression profiles for each cluster, and the lower number of features could have easily impacted the similarity of the profiles for the different cell types. The same argument would apply to the LDA and entropy scoring approach. How many genes are necessary to result in the creation of a new topic?

So far it was assumed doublets will have higher entropy than singlets. However, this statement does not seem to hold when evaluated on real data. In the next section, the validity of such assumptions is evaluated.

4.5.5 Validation of our assumptions

While previous analysis was based on real datasets with doublets annotated either via Cell Hashing or donor SNP information, we use synthetically generated doublets to check what assumptions affect the performance of our proposed approach. Similar to previously described experiments, an LDA is fit on the data and entropy score is computed for each cell. This dataset consists of PBMCs and 750 doublets (B-cells and monocytes, T-cells and monocytes, and B-cells and T-cells). We assume that doublets will have higher entropy than the singlets that have generated them, and here we test this assumption by plotting the entropy of the clusters that generated the doublets and the actual doublets. Entropy distributions can be seen in Figure 4.14. In the case of doublets between T-cells and B-cells, there is a clear evidence of a shift in the entropy distributions when comparing the two clusters of singlets and the doublets. A similar shift in entropy is observed between T-cells and monocytes doublets, albeit smaller. However, the entropy of B-cells and monocytes doublets completely overlaps with the entropy of the two clusters of singlets.

While for doublets of two cell types, entropy can indeed be higher for doublets, that is not always the case and entropy scores can overlap between doublets and singlets. As such, the assumption that doublets have higher entropy than singlets does not always hold.

4.6 Discussion and possible future directions

4.6.1 Improving annotation of doublets

As the performance summary of the methods shows (see Tables 4.5, 4.7, 4.8), DoubletFinder is consistently the best-performing doublet detection method, assuming well-selected doublet rate.

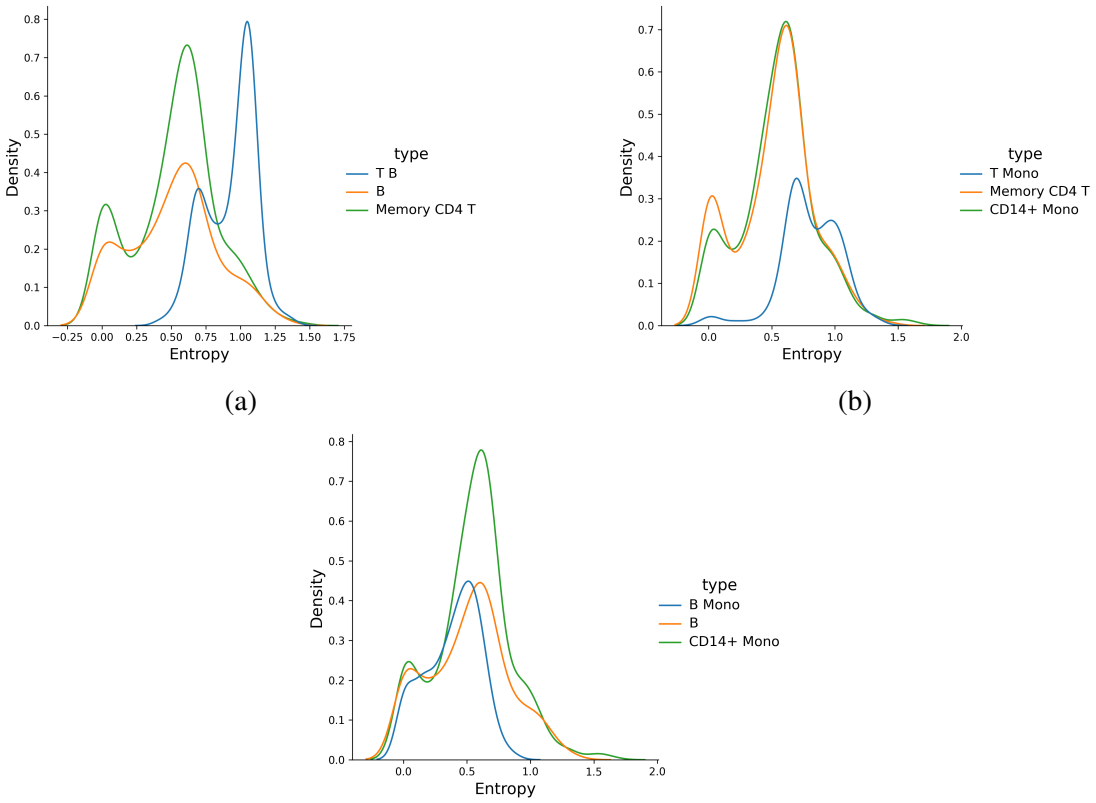


Figure 4.14: Entropy distributions for the 3 sets of doublets created using real data. (a) A shift in the distribution of the the entropy for doublets can be observed for doublets between T-cells and B-cells. (b) There is a slight shift in entropy, but also entropy distribution overlaps. (c) There is a clear overlap in the distributions between the clusters of singlets and the doublets generated between B-cells and monocytes.

However, across the three datasets the best-performing method still only achieves about 73% sensitivity. This score can be potentially improved by taking a union of all tools as each of the above discussed methods identifies unique doublets, see Figure 4.15.

4.6.2 Why is entropy not suitable for doublet annotation?

In some situations entropy does seem to have a degree of predictive performance, such as in the case of the Demuxlet dataset. However, performance is dependent on the number of UMI counts and features detected, as in the cases of the Cell Hashing dataset and the downsampled Demuxlet dataset. Furthermore, as shown in Figure 4.7, choosing an entropy cutoff means deciding whether it is more important to filter out more doublets or to reduce the number of FPs.

Furthermore, the distributions of entropy values for doublet and singlets show that there is an overlap, and that some singlets have high entropy scores. This is possible if a group of cells is both from a particular cell type and proliferating, for example. They will have additional topics capturing this process. Moreover, different cell types have different entropy value distributions,

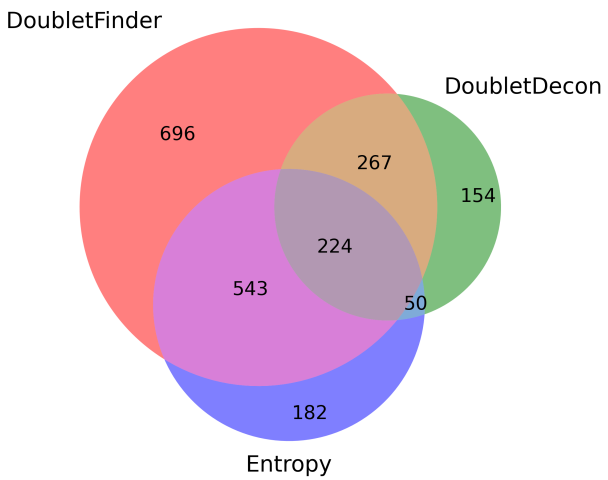


Figure 4.15: True positives across DoubletFinder, DoubletDecon, and entropy scoring in the Cell Hashing PBMCs dataset.

as shown in Figure 4.14. In addition, certain cell types might require the same topics to model their counts, and therefore it is worth exploring how much difference can be captured by the topics. Entropy analysis in its current state could be improved by taking the cell identity into account, performing initial clustering, and obtaining a reference entropy per cell type.

4.6.3 Counts, housekeeping genes and doublets

Housekeeping genes (HKGs) are genes considered to be stably expressed in different cell types, tissues, and developmental stages (Eisenberg & Levanon 2013). Initially, the commonly used set of HKGs has been derived from microarray studies. However, with the advent of single cell it is of interest to determine whether such patterns can be identified on single cell level and how stable they are. A recent study aimed to answer those questions by analysing 11 scRNA-seq datasets generated from diverse tissues and biological systems (Lin et al. 2019). Can housekeeping genes be used to inform the annotation of doublets? A housekeeping gene is assumed to have uniform counts across a dataset. If this is not the case for some cells, then they are potentially doublets. To evaluate this assumption, the PBMCs of Demuxlet and Cell Hashing datasets are used.

If we sample two groups of random cells from singlets populations, we expect the ratio of means for a stably expressed gene to be approximately 1. While the ratio of means of doublets and singlets should be higher than 1. Tables 4.10 and 4.11 demonstrate that the ratio of means of doublets and singlets is higher than 1.5 for the Demuxlet dataset and higher than 1.4 for the Cell Hashing dataset, while the ratio of randomly sampled groups of singlets in both datasets is around 1, but again slightly lower for the Cell Hashing dataset.

The counts information and more specifically the ratio of means can be used as the basis of a new model for identifying doublets. Specifically, a possible avenue to explore would be to begin by clustering the data, then the counts within each cluster for a pre-selected set of genes can

Gene	Ratio of means Demuxlet	Ratio of means Cell Hashing
SNRPD3	1.74	1.42
PFN1	1.63	1.73
SRRM1	1.85	1.54
HNRNPA2B1	1.73	1.68
YWHAB	1.85	1.95
RPL36	1.84	1.65
GDI2	1.77	1.72
NCL	1.99	1.65
RPL8	1.82	1.70
CSDE1	1.80	1.98
C14ORF2	-	-
ARF1	1.76	1.76
TARDBP	1.78	1.73
STOML2	1.59	1.54
RPS5	1.92	1.70
THRAP3	1.90	1.71
HNRNPM	1.71	1.64
POLR2E	1.50	1.46
SRSF3	1.89	1.76
CKS1B	1.76	-

Table 4.10: Ratio of means (doublets divided by singlets). Rounded to two decimal places. In this table "−" denotes a gene that hasn't been measured. Genes are based on the list of genes by (Lin et al. 2019).

be modelled with a mixture model where the mixture components correspond to doublets and singlets.

4.7 Conclusions

This chapter proposed a novel approach for doublet annotation, which combines LDA and entropy scoring, and compared that method with state of the art doublet detection approaches across three real datasets containing doublet annotations. In its current formulation this approach does not require prior clustering or knowledge of doublet rate. We have benchmarked our approach against state of the art doublet detection methods, DoubletFinder and DoubletDecon. Benchmarking was done using three datasets with varying sequencing depth and similarity of the included cell types. We have shown that no existing method can annotate doublets with high specificity and sensitivity. All methods are limited to identifying heterotypic doublets. As shown by the downsampling experiment, the total UMI counts affect the performance of all methods. Finally, as the analysis of the cell hashing dataset of cell lines has shown all methods suffer poor performance when applied to transcriptionally similar cells. Shortcomings of assumptions were identified by analysing a synthetic doublets dataset.

Gene	Ratio of means Demuxlet	Ratio of means Cell Hashing
SNRPD3	0.97	0.84
PFN1	1.03	0.91
SRRM1	1.02	0.93
HNRNPA2B1	1.06	1.04
YWHAB	1.02	0.99
RPL36	0.95	0.95
GDI2	1.01	0.86
NCL	0.89	1.00
RPL8	1.00	0.96
CSDE1	1.02	1.03
C14ORF2	-	-
ARF1	1.02	0.83
TARDBP	0.95	1.09
STOML2	0.96	0.81
RPS5	1.99	0.92
THRAP3	1.03	0.98
HNRNPM	0.96	0.87
POLR2E	1.03	1.03
SRSF3	0.99	0.91
CKS1B	1.34	-

Table 4.11: Ratio of means between randomly sampled singlets, rounded to 2 decimal places.

The discussed methods were not applied to transitioning cells, as to date DoubletDecon is the only tool that can distinguish between transitions and at present there are no well-annotated datasets that allow for such analysis. As trajectory inference is becoming a standard step in single cell analysis, ensuring false signal has been removed is vital for results interpretation (Luecken & Theis 2019).

While the LDA-based approach does not outperform other methods, a method based on LDA has been used successfully to remove ambient RNA from counts data. In DecontX, one such method based on LDA, counts are assumed to come from two topics, real data and ambient RNA (Yang et al. 2020). While the LDA-based approach for doublet detection does not result in superior performance, it will become evident from the results of the next chapters that LDA is still a useful and suitable method for analysis of single cell data.

Given the demand for sequencing greater number of cells and the limitations of demultiplexing techniques and doublet detection methods, there is a need to develop better methods for doublet detection with higher accuracy that will not be affected by the range of assumptions of current methods. A possible area of exploration would be to take into account the ratio of means, and include some prior knowledge in the form of stably expressed genes.

While computational methods can enable the identification of doublets and improve downstream analysis, improvements in scRNA-seq protocols can prevent doublets forming in the data

in the first place. For example, recently a platform that combines double emulsion encapsulation and phenotyping via FACS (Dropception) was developed. The authors compare multiplet rate with state of the art droplet-based methods. In the case of Dropception, the multiplet rate is very low, less than 2% (Brower et al. 2020).

Chapter 5

Understanding cellular crosstalk in scRNA-seq using topic modelling

This chapter is based largely on a paper accepted at PLoS Computational Biology
https://doi.org/10.1371/journal.pcbi.1009975
Some sections have been rewritten to improve flow. Paper supplementary information has been included as Appendix B

5.1 Introduction

Cell-cell communication is vital for most biological processes, from maintaining homeostasis to determining specific immune responses (Shao et al. 2020). In disease states, malfunctioning cells can induce changes in cell-cell interactions and secondary changes in their micro-environment, which leads to reprogramming of the niche to their advantage (Scadden 2014). Improving understanding of essential interactions has the potential to aid discovery of novel therapeutic targets (Song et al. 2019).

One way to study interactions between cell types, widely used in scRNA-seq studies, relies on ligand-receptor pairs screenings. Examples of such methods, using a priori curated interactions include: CellPhoneDB, NicheNet, and SingleCellSignalR (Browaeys et al. 2020, Cabello-Aguilar et al. 2020, Vento-Tormo et al. 2018). CellPhoneDB or variations of their method have been applied in practice to answer questions about intercellular communication between cell types in a range of tissues. For example, Cohen et al (Cohen et al. 2018) consider the interaction of lung basophils with the immune and non-immune compartment by examining known ligand-receptor pairs and how those potentially link to development. As these methods are based on databases of curated resources, they do not allow for new genes that change as a result of interaction between cell types to be identified, so results are limited to known biology. Furthermore, most curated resources of ligand-receptor pairs are only available for humans or mouse orthologs (Vento-Tormo et al. 2018).

As discussed in detail in Chapter 4, in scRNA-seq, it is possible for two cells to be sequenced together, known as “doublets". Often doublets are a result of errors in cell sorting or capture, but recently two studies have shown that doublets can capture two physically interacting cells (PICs), offering a valuable method to measure the transcription pattern of interaction, without relying on prior knowledge. Boisset et al (Boisset et al. 2018) used mouse bone marrow (BM) to demonstrate that cell-cell interaction can be studied by dissociating physically interacting doublets. The two interacting cells are separated by needles and sequenced. Further experiments also managed to infer interactions by sequencing intact doublets which were then deconvoluted based on the gene expression of the sequenced singlets. Giladi et al (Giladi et al. 2020) developed a method for sequencing PICs, known as PIC-seq. With other single cell technologies, information about cell-cell interactions are lost due to cell dissociation while PIC-seq captures pairs of interacting cells. PICs are isolated by a combination of tissue dissociation, staining for mutually exclusive markers, and flow cytometry sorting. Single positive and PIC populations are then sequenced. The ability to capture PICs allows Giladi et al (Giladi et al. 2020) to study physical interactions between cells and potentially capture a novel set of genes that might be changing as a result of physical proximity. On the computational side of their PIC-seq approach, they (Giladi et al. 2020) cluster the mono-cultures and from these the gene expression of each PIC is modelled as a doublet: $\alpha \times A + (1-\alpha) \times B$, where A and B are the two cell types that make the PIC and α is the mixing parameter. α is estimated by a linear regression model trained on synthetic PICs. This is followed by maximum likelihood estimation (MLE) of A and B. By identifying the two subtypes that comprise the PIC, expected expression can be computed. Expected and actual expression of the PIC are compared to identify changes as a result of interaction (Giladi et al. 2020). There are several potential limitations of the outlined approach. Since the PIC-seq algorithm relies on deconvoluting doublets, it cannot be applied to transcriptionally similar cells, such as subtypes or the same cell type. Furthermore, for the training of the linear regression, synthetic PICs are created by pairing pooled cells from A and B that are then downsampled to a predefined total number of unique molecular identifiers (UMIs). While the approach of combining cell profiles to create a doublet has been used with some modifications in a range of studies (DePasquale et al. 2019, McGinnis et al. 2019), it simplifies how a doublet arises in practice (Ilicic et al. 2016). Additionally, the method described in PIC-seq requires prior clustering of cells before simulating artificial PICs and deconvolution.

As discussed previously in Chapter 4.3, LDA has been applied to different types of omics data (Liu et al. 2016). In the context of LDA, cells are equivalent to documents and genes are words. Topics can be described as groups of genes whose expression co-varies (Bravo González-Blas et al. 2019, Kim et al. 2020, Kotliar et al. 2019). The identified topics can be interpreted as general, cell type specific, or linked to technical quality of the samples. For example, ribosomal or mitochondrial-dominated topics might correspond to dying cells. In addition to the standard implementation of LDA, work has been completed to allow for simultaneous topic identification

and cell clustering (Campbell et al. 2020). Furthermore, CellTree (duVerle et al. 2016) uses LDA for trajectory inference: the method takes LDA in its standard form but computes chi-square distance between cells, and uses the distance to build a tree to describe a branching process. Most recently, a modified version of LDA has been used to decontaminate counts from ambient RNA: DecontX assumes counts come from two topics, native counts and ambient RNA. Using only native counts improves clustering and downstream analysis (Yang et al. 2020).

In this chapter we propose a novel method based on LDA that allows for identification of genes that change their expression as a result of cell-cell interaction. Once trained on a reference population we can fit LDA on an interacting population and capture changes that cannot be explained by the initially learned topics. Firstly, we show how the proposed model behaves when fit on synthetic doublets with some upregulated genes. We also show new topics are needed to model the counts of genes related to interaction, even if they are not expressed in all interacting cells. We fit LDA as described by (Blei et al. 2003) on a population of singlets or sorted cells. Then we fix the topics from the singlets reference population and fit another LDA on the interacting cells population. The second LDA allows us to rank the genes that have high probabilities in the new topics. We apply our method to two datasets containing PICs and identify genes that change their expression as a result of interaction between cell types. Examples of genes include adhesion and co-stimulatory molecules, which are direct evidence of physical interaction between cells. Finally, we demonstrate the challenges of applying our method to a 10x Chromium dataset bronchoalveolar lavage fluid (BALF) of patients with COVID-19 (Liao et al. 2020). We link our findings to how well the sequencing protocol can preserve interaction, and to what extent we can identify reference populations. However, as the work of (Boisset et al. 2018) and (Giladi et al. 2020) has shown, there is a need to modify currently available scRNA-seq protocols to allow physical interactions to be captured. To our knowledge, this is the first paper that models interaction using an LDA-based approach. Furthermore, our approach does not require prior clustering or synthetic generation of doublets compared to the computational approach previously used to identify genes related to interaction in the work of (Giladi et al. 2020). We use the genes identified by (Giladi et al. 2020) as the ground truth and show how the number of top genes we select affects true positive and false positive rates. In addition to identifying genes discussed by the original paper, we provide a comprehensive ranking of further genes that might change as a result of interaction, such as ones involved in cellular response and adhesion. Taking the top 5 genes per topic in the PIC-seq data allows us to identify 20 further known genes related to cell adhesion and immunity. Additionally, our ranked list of genes includes genes lacking comprehensive annotation and as such allows us to go beyond known interactions. While the analysis of Boisset et al (Boisset et al. 2018) does not consider specific genes that would change as a result of interaction but focuses on cell types known to interact, we perform a literature survey to verify whether we can identify known genes related to interaction in the bone marrow. Specifically we consider the top 25 genes per topic and we identify over 90 genes linked to cell

adhesion and response.

5.2 Materials and methods

5.2.1 Latent Dirichlet Allocation

Back in Chapter 3, the traditional LDA formulation was introduced. In Chapter 4, the formulation was redefined for single cell data. Specifically, cells correspond to documents, genes are equivalent to words, and word frequencies are counts. The generative process is similar to the explanation in Chapter 4, section 4.3.

In our proposed approach, we initially fit LDA on a reference population: co-cultures of the cell types in the PIC-seq dataset, sorted BM cells in Boisset's dataset (Boisset et al. 2018), and what we identify as singlets in the COVID-19 BALF data (Liao et al. 2020). The assumptions of LDA fit well in the context of scRNA-seq as at any given point we can observe multiple processes in a cell. A cell can be described as a contribution from multiple topics, some specific to a cell type and some shared across all cells. Those processes can be described as genes that co-vary. As words can be in multiple topics, genes can be part of multiple processes. By fitting LDA on the co-culture of T-cells and the co-culture of dendritic cells (DCs), we obtain for each topic a distribution over genes that we then fix before we fit another LDA on the population of PICs, dissociated BM doublets, or DoubletFinder identified doublets respectively for the three datasets discussed in the results. The initial LDA captures a reference state of cells, a state without interaction. Fixing topic-gene probabilities learned from the reference, not interacting populations, allows us to capture in the new topics any changes as a result of interaction due to the setting of the datasets analysed. Our LDA approach is shown in Figure 5.1.

5.2.2 Identifying topics linked to a cell type

In order to aid interpretation of the identified topics, we link topics to cell types. For each topic we group together cells from the same cell type as annotated in the reference, and perform a Mann–Whitney U test on the topic-cell vectors. Under the null hypothesis, we assume a topic has the same probability in the two cell types. To correct for multiple testing, we use the Benjamini-Hochberg procedure with α set to 0.05.

5.2.3 Choosing number of topics

To select a suitable range of topics, we compute perplexity, defined in Chapter 3, for a range of topics starting with 2. A lower perplexity score is an indication of a better model. We note that perplexity decreases rapidly up to $K = 10$ and then flattens out. We also measure cosine and JS which show very similar patterns (Figure B.10). Such behaviour is common when fitting

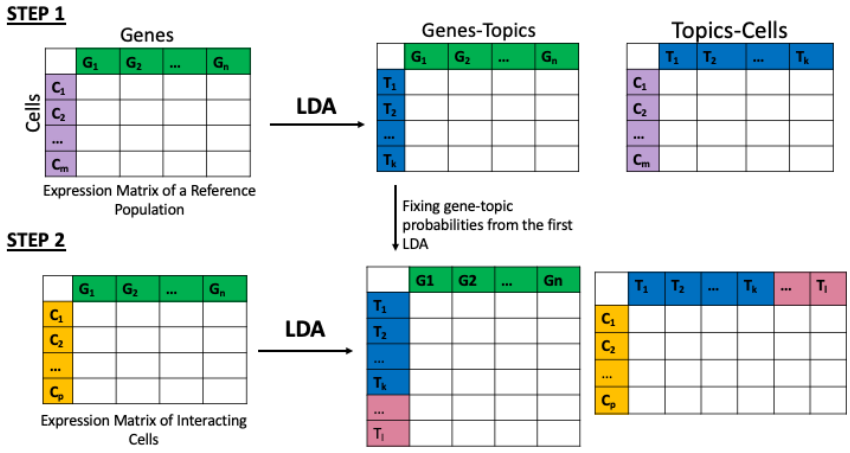


Figure 5.1: We begin by fitting LDA on a reference population. Depending on the setting of the experiment the reference could be sorted cells (Boisset's dataset) or co-cultures of the cell types involved in the interaction (PIC-seq dataset). Using the topic-gene probability vectors we learn from this first LDA, we fit another LDA with some topic-gene probabilities fixed on the interacting population.

LDA-style models and it is typical to choose the smallest value of K that is able to explain the data: i.e. the value of K at the point in which the perplexity flattens out. In Figure 5.2, for example 10 would be a suitable number of topics for that dataset. Our goal is to recover interacting genes. To demonstrate that this strategy for choosing the number of topics is appropriate for that final goal, we show the effect the number of topics has on the Area Under the ROC curve (AUC). A plot of number of topics versus AUC can be seen in Figure B.9 and shows agreement with the perplexity plot: optimal results are observed for K = 10, the value at which the perplexity flattens out. While performance is relatively consistent, the ROC curves with higher number of topics show some decay.

5.2.4 Motivating the need for new topics

Consider a cell, n, being one of the PICs. Cell n decomposes into $\boldsymbol{\theta_n}$ where $\sum_k \theta_{nk} = 1$ (probability distribution for cell n over all k topics). Some of the topics come from the reference LDA fit, which are fixed before fitting the second LDA, and some topics come from the PICs. We want to compare a fit with all topics with a fit where we do not use any new topics. Let $\Delta_{nk} = \theta_{nk}$ but we set the contribution of all topics that come from the PICs to 0 and re-normalise, so that $\sum_k \Delta_{nk} = 1$. Let $\boldsymbol{\beta_k}$ = topic probability for topic k and $\sum_m \beta_{km} = 1$. If we are only interested in the probability of picking the counts for one gene versus all other genes, the multinomial distribution reduces to a binomial distribution. For each topic distribution, we compute the probability $P(X >= x)$ for a binomial distribution defined as $X \sim Binomial(n, p)$ where n is the total counts for cell n, p is the probability for that gene in that topic, and x is the count for a particular gene in the current cell. Once we have computed the probability for a gene for each topic, we multiply them by θ_{nk}

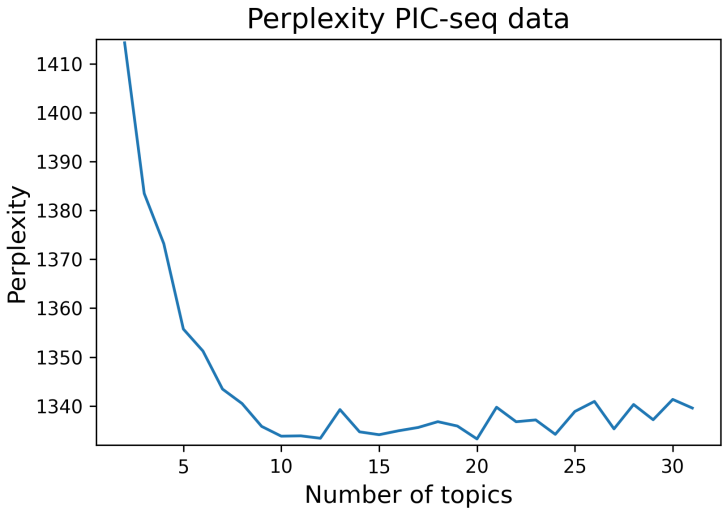


Figure 5.2: Illustrating the drop of the perplexity as the number of topics increases. In this particular case, a suitable number of topics would be 10. After that as the number of topics increases, the perplexity values plateau with some fluctuations

and Δ_{nk}, topic weightings for cell n, and sum on k. We expect the probability of observing counts greater than or equal to the actual count for a gene changing as a result of interaction to be higher under θ_{nk} where all topics are included compared to Δ_{nk} which is based only on the initial topics. Probabilities should be similar for genes not involved in interaction.

5.2.5 Ranking genes as potential candidates of interaction

One of the outputs from our LDA is a probability distribution for a gene in a cell across all topics. Let N be the total number of cells in our doublets/interacting population of interest, then for a gene G and a topic k we find how many cells require topic k to explain the expression of gene G. For a newly identified topic k, for each gene we count how many cells have highest probability for this gene in topic k. For each topic, we produce a ranking of genes based on the number of cells that require this topic to explain the expression of that gene. The rankings for each newly identified topic can then be analysed. We choose a different number of top genes from each topic in subsequent experiments. For our synthetic experiments, we plot how those values affect true and false positive rates.

5.2.6 Evaluation datasets

- PIC-seq of T-cells and DCs: The count matrix and the metadata were downloaded from GEO under accession number GSE135382. The metadata file was used to filter for co-cultures of the same cell type and co-culture of T-cells and DCs. The reference population consists of cells tagged in the metadata as: Co-culture $TCRb+$ (T-cells) and Co-culture $CD11c+$ (DCs). All three timepoints 3h, 20h, and 48h were used. PICs were selected from

the metadata as: Co-culture, $TCRb+CD11c+$, all three timepoints 3h, 20h, and 48h.

- BM dataset: The count data was acquired from GEO under accession number GSE89379. The sorted cells were used as a reference and the dissected doublets were used for analysing interactions. Cells prefixed JC20 to JC47 denote micro-dissected cells. Cells with prefix JC4 denote sorted hematopoietic stem cells (used as reference).

- COVID-19 BALF dataset: The data were downloaded from GEO, under accession number GSE145926, in the form of h5 files, CellRanger output. The next subsection describes how the reference and the population of potentially interacting cells are identified.

5.2.7 Pre-processing and analysis before LDA

PIC-seq and BM dataset

PIC-seq: Since the PIC-seq dataset was generated using the MARS-seq platform which has higher sequencing depth than the standard 10x Chromium, we set higher filtering cutoffs for the number of unique features per cell. As we are not relying on clustering, we can also set a higher cutoff for the number of cells in which a gene is captured. Genes appearing in fewer than 200 cells were filtered out. Cells with more than 500 features were retained for downstream analysis. Similar to the original publication we exclude ribosomal genes.
BM dataset: This dataset has been sequenced using CEL-seq. Only genes expressed in more than 10 cells were considered for downstream analysis, resulting in over 10 000 total genes. 369 sorted cells and 1546 dissected doublets were used. No other pre-processing was performed before fitting LDA.

For both datasets filtering steps are performed independently of any scRNA-seq pipeline.

COVID-19 BALF

Quality control, filtering, normalisation, integration, and clustering were done in R, using Seurat, version 3.1.2. Filtering decisions are dataset dependent and are based on three main metrics: number of genes per cell, number of cells a gene is expressed in, and fraction of mitochondrial genes. It is typical to filter for cells with a very high number of genes expressed to prevent including doublets in the data. For example, cells with low counts and high mitochondrial fraction indicate the mRNA has leaked out through a broken membrane. As such, samples were filtered for cells with fewer than 500 genes. Since we are interested in doublets, which are often assumed to have higher counts than singlets, the maximum cut-off was relaxed (Luecken & Theis 2019). To exclude dying cells we also set a mitochondrial gene expression cut-off to 25. The full list of filtering cut-offs for the different COVID-19 samples can be found in Table B.1.

Seurat's NormalizeData and FindVariableFeatures functions were used before integration. Samples were integrated first by condition, and then all conditions were integrated using FindIn-

tegrationAnchors. Clustering resolution was set to 0.5 to identify general populations. Based on the cluster identification, a subset of the BALF cells were taken forward for LDA analysis.

To confirm cells we identified as doublets, we use DoubletFinder. Since DoubletFinder can only be used on a single sample and not integrated data, patient samples C143 and C145 were analysed separately (McGinnis et al. 2019). Those two samples were chosen based on the amount of cells in what we defined as a doublet cluster. Figure B.1 shows the annotated clusters for sample C145.

5.2.8 Running LDA

In the case of the PIC-seq dataset, we use 10 topics for the reference population, see Section 5.2.3. We use 20 topics for the interacting population. In each case, we run LDA for 500 iterations. The same setting is used for the Boisset's BM dataset and the COVID data.

5.3 Results and discussion

5.3.1 Validation using synthetic doublets

Before testing our method on a real dataset of interacting cells, we apply it to a dataset containing synthetic doublets in order to show that we are able to detect genes that change in interacting cells. We simulate doublets by merging the expression profiles of singlets using different ratios: 50/50 (equal contribution of each cell type), 60/40, and 30/70. In order to obtain a ground truth for genes that change as a result of interaction, we modify the expression of some genes by adding 1.5, 3 and 10 to their total counts. Results are shown in Figures 5.3, 5.4 and the supplementary information, Figures. B.2, B.3 and B.6. The value of 1.5 increase was chosen as it represents a typical count for a gene. We chose to modify the following genes in the synthetic doublets: *Sell, Mif, Bcl2l1, Cd40, Myc, Ncl, Cst3, Ly6a, Ctla4, Ccl22, Cd69, Dll4, Lgals1*. We trained our first LDA on a randomly sampled subset of T-cells and DCs mono-culture from the (Giladi et al. 2020) paper. After fixing the topic from the reference, we fit a second LDA on the doublets that were created as different contributions of the singlets and upregulation of some genes. We expect those genes to require contributions from the new topics to model their counts.

We applied our approach to the datasets of upregulated doublets. We expect the probability of observing counts greater than or equal to the actual counts of the list of upregulated genes (e.g. *Ly6a*, *Sell*) to be higher when we use all topics and we compute the probability under θ_{nk}, where k is topics learned from singlets and simulated interacting doublets. For the genes we chose not to upregulate, the probabilities under Δ_{nk} and θ_{nk} should be similar. This is shown in Fig 5.3, where we plot the probability of observing counts greater than or equal to the actual counts in doublets with modified expression of the previously listed genes. *Sell*'s counts in the modified doublets can be explained better if the new topics are included. However, in the case of *mt-Cytb*,

a gene we have not modified, probabilities are similar under the two models, using all topics or the initial ones. Further examples of genes we have modified and genes with counts that can be modelled by the original topics can be seen in the supplementary information Figures B.2 and B.3. Results are similar to the genes shown in Figure 5.3.

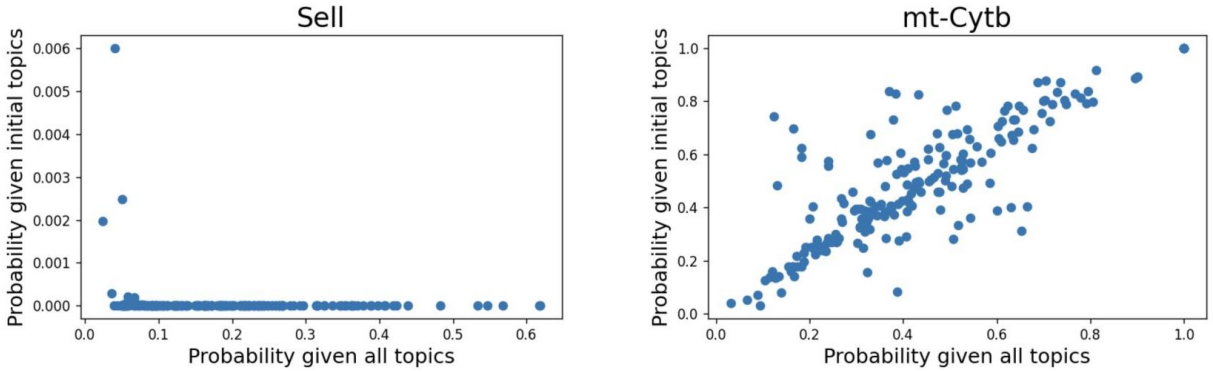


Figure 5.3: Using the synthetic data, we evaluate how likely it is to observe counts for some genes under θ and Δ. For 200 upregulated doublets, we plot the probability of observing the counts of a gene that has been upregulated, *Sell*, and a gene that has not, *mt-Cytb*, under a model using all topics or using only the topics learned from the singlets populations. In the case of *Sell*, a gene with modified expression, the probability of observing the actual counts or greater than in the upregulated doublets using all topics is higher compared to the probability of observing those counts if we only use the initial topics. However, for *mt-Cytb* that we have not upregulated the probabilities under the two models of observing the actual counts or greater than are similar. Thus, we can conclude that the additional topics are required to model the genes that change.

For each of our simulated doublets experiments, we rank the genes based on whether they require contribution from the new topics to explain their expression. We plot true positive rate vs false positive rate using different cutoff values for the top ranked genes. The ROCs in Figure 5.4 show how the results are affected by picking a different number of top genes per topic. A further experiment was performed where we upregulated a random set of genes *Gcfc2, Wdsub1, AU040320, Pank3, Dcaf12, Gm26669, Ehd2, Bag3, Rpl10-ps2, Notch1, Ppm1g, Oxsr1, Nrarp, Ppp3ca, Rpl28-ps1, Stbd1, Srgap2, Cpped1, Gm10420, St6galnac3*. Results can be seen in the supplementary B.4 and B.5.

Additionally, we evaluate whether we can identify genes that change in a subset of cells by upregulating some genes in 10% of the total PICs population. We upregulate the expression of *Gbp4, Gbp7, Gzmb, Il2ra, Psma4* in 20 cells (10% of the total PICs). In all four experiments, using up to 15 top genes per topic resolves the list of upregulated genes that we use as ground truth. We note that even if a gene is upregulated in as few as 20 cells that gene can still appear in the ranking, and we recommend also exploring genes which change in few cells when analysing results.

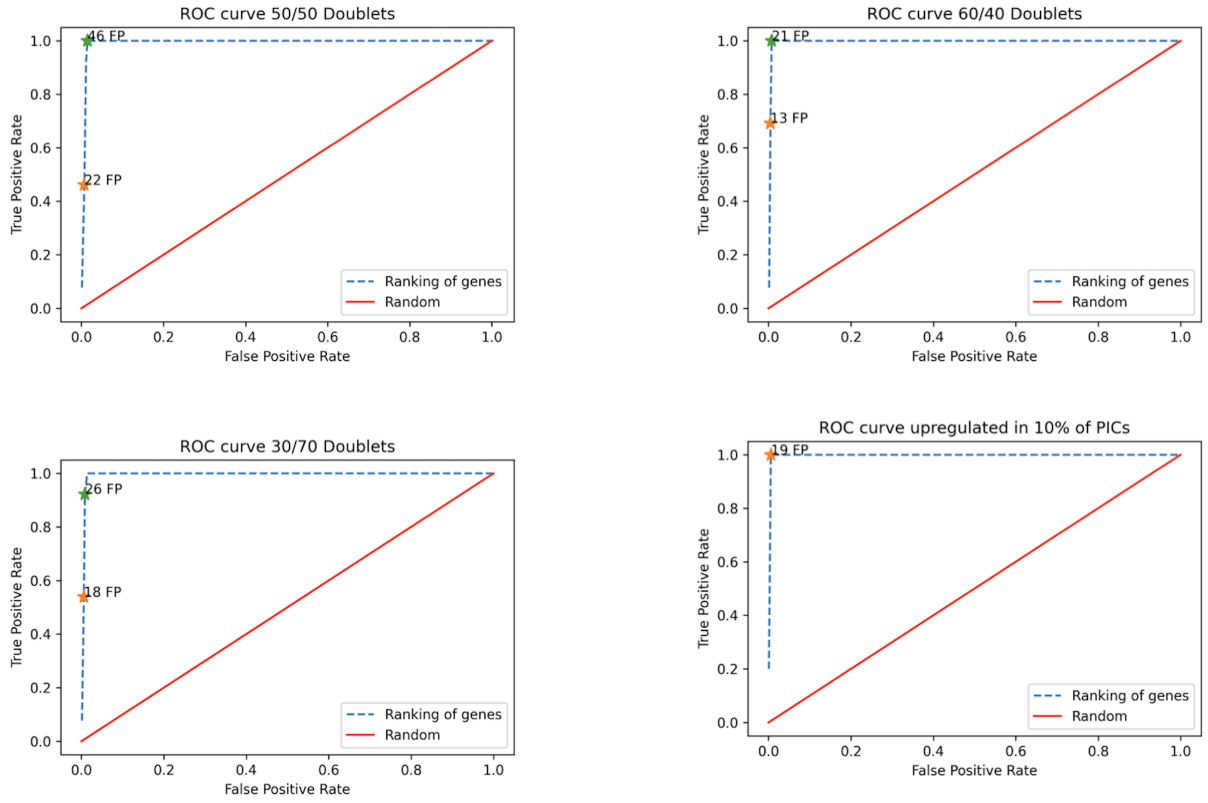


Figure 5.4: For each of our 4 synthetic experiments, we plot a ROC curve using a different number of top genes based on our ranking. In all cases the genes we have modified the expression of appear at the top of the newly identified topics. For all synthetic experiments, using up to the top 15 genes per topic resolves the list of upregulated genes we consider as ground truth. However, as seen from the plots even if a slightly higher number of top genes are used, the false positives are gradually increasing. In each plot we have indicated the number of false positive genes ranked for each experiment while the full truth set has not been identified. Total list of ranked genes for those experiments is over 2500.

5.3.2 PIC-seq dataset

The first real dataset we use for evaluation has been generated by PIC-seq and includes interacting T-cells and DCs (gated for TCRβ^+CD11c^+) as well as two co-cultures of a single cell type (T-cells: TCRβ^+ and DCs: CD11c^+) across three different timepoints, 3h, 20h, and 48h. The original Giladi et al. (2020) work uses a metacell model to cluster the cultures of a single cell type. Then each PIC is modelled as a combination of metacells, and a mixing proportion, α, is estimated by a linear regression model trained on synthetic PICs. The metacells are identified using MLE. Genes of interest are identified by comparing expected expression, based on the inferred cell types contributing to the interacting pair and actual expression of the PICs.

Our model does not require prior clustering and generation of synthetic reference profiles. As a first step we train one LDA on the co-cultures of T-cells and DCs, using those as a reference. With the first LDA we manage to capture topics specific to T-cells and DCs (groups of genes co-varying in one cell type over the other) and specific time-points. As seen in Figure 5.5 topics 0

and 1 are specific to DCs. We identify topics specific to a cell type as described in Section 5.2.2. In addition to having high probability for a cell type, some of those topics have higher probability over the different timepoints. For example, topic 5 has high probability for T-cells during the 48h window. Similarly topic 1 is higher DCs in the 3h time period, while topic 0 is DC specific for the 20h time period, see Figure 5.5 and Table B.3. To explore what the top genes are in some of those topics, we pick topics 0 and 8 and order the genes in those topics by probability. We see DC specific genes *Fscn1, Ccl22, Tmem123, and Cd74* in topic 0. Similarly, some of the genes with highest probability in the T-cell specific topic include *Mif, Ncl, Nmp1* (see Figure 5.6). Further examples of genes with high probabilities in some of those topics can be found in Table B.3.

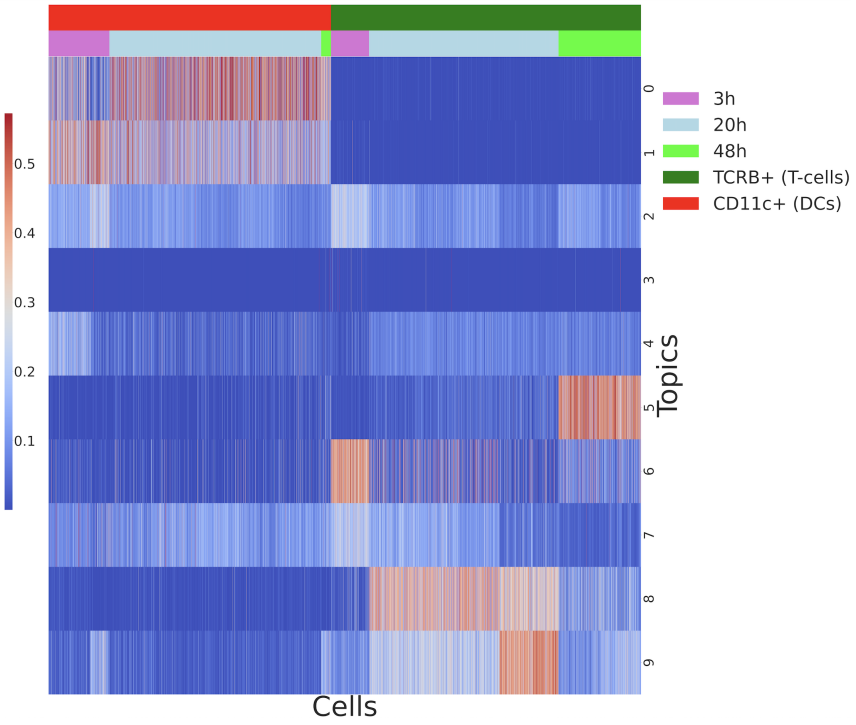


Figure 5.5: Heatmap of topics expression in the reference populations of T-cells and DCs. To map topics to a cell type, we group together cells from the same cell type and perform a Mann–Whitney U test for each topic. Results are corrected using the Benjamini-Hochberg procedure for multiple testing.

We fix the topics we learned from the co-cultures of the two cell types, T-cell co-culture and DCs co-culture, before we fit another LDA on the physically interacting populations of PICs. As described earlier, in order to rank interesting genes, for each topic we learned from the PICs, we count how many cells use this topic for a particular gene. Then we rank the genes based on the number of cells. Our PICs population contains over 3000 cells and we only consider in our final ranking genes that require a particular topic for more than 10 cells.

To validate our findings, we check whether the top genes in each of the newly learned topics have also been highlighted by Giladi et al. (2020) in Supplementary figure 4 of their paper. Due to differences in filtering, we have not retained 10 of the genes they identify to change as a result

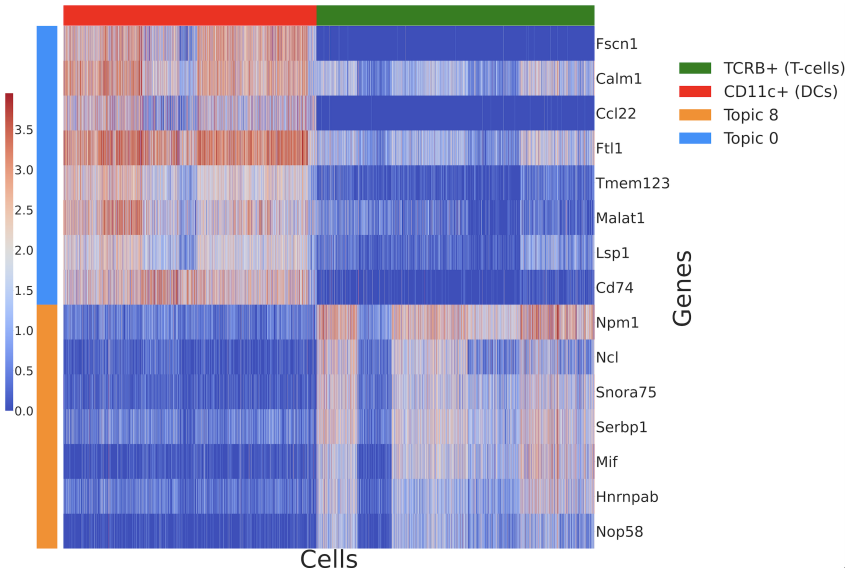


Figure 5.6: Heatmap of top genes based on probability in that topic (for each topic we can obtain the probability of the genes in that topic) from topics 0 and 8 (log-transformed). Topic 0 contains genes that tend to have higher expression in DCs compared to T-cells. Topic 8 contains genes co-varying in the T-cells co-culture. Results are similar to Figure 1 from the Giladi et al. (2020) manuscript

.

of interaction and we take the remaining 81 genes present in our data as a ground truth. In order to evaluate how results are affected by the number of top genes per topic we select, we plot true positive rate vs false positive rate (see Figure 5.7).

While the analysis done in the original publication (Giladi et al. 2020) groups cells by the types of the singlets involved in the interaction and the timepoint of capture before performing log fold change (results in Supplementary figure 4d of the original paper) we show that some of the new topics we identify correspond to the timepoints of capture and reveal genes with temporal patterns as shown in Figure 5.8. For example, *Ldha, Ptma, Pcna, Trac, Dut* are needed by a subset of cells and captured in the same topics. Their pattern of expression is higher after the first 3h. *Tnfrsf4, Tnfrsf9, Tnfrsf18* seem to be expressed across all timepoints and the shift of their expression is captured by the same topics. The expression of new topics across the cells can be seen in Figure B.11.

While for the purposes of the ROC analysis we are considering genes that are not amongst the ones discussed by Giladi et al (Giladi et al. 2020) as false positives, for some of those genes there is evidence they could be involved in interaction. Taking the top 5 genes per topic considered as false negatives previously, we find genes related to immunity and cell adhesion, some of which are ligand-receptor pairs (over 20 genes in the first 100 ranked). Examples include *Cd2, Cd74, Il2ra, Il2rb*. Additionally, while some known genes appear high in the ranking, some of the genes in our list are not well-annotated. This makes them potential targets for further analysis to elucidate their role. Genes can be found in Table B.4.

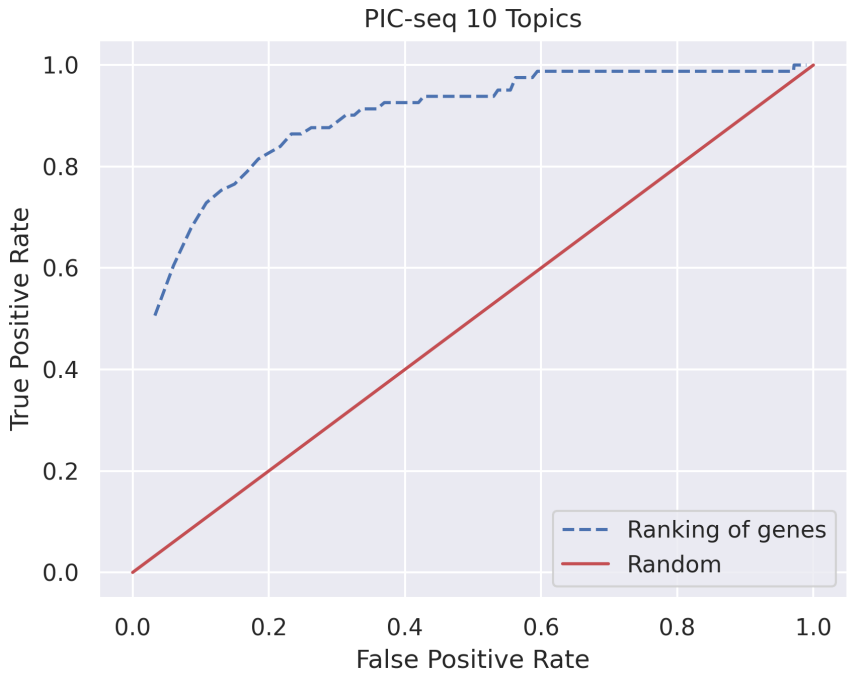


Figure 5.7: ROC curve using genes from (Giladi et al. 2020) as ground truth based on top genes cutoff for each topic. It is important to note that some of our false positive values correspond to true interacting genes that have not been presented in the (Giladi et al. 2020) paper amongst their Supplementary fig 4 genes.

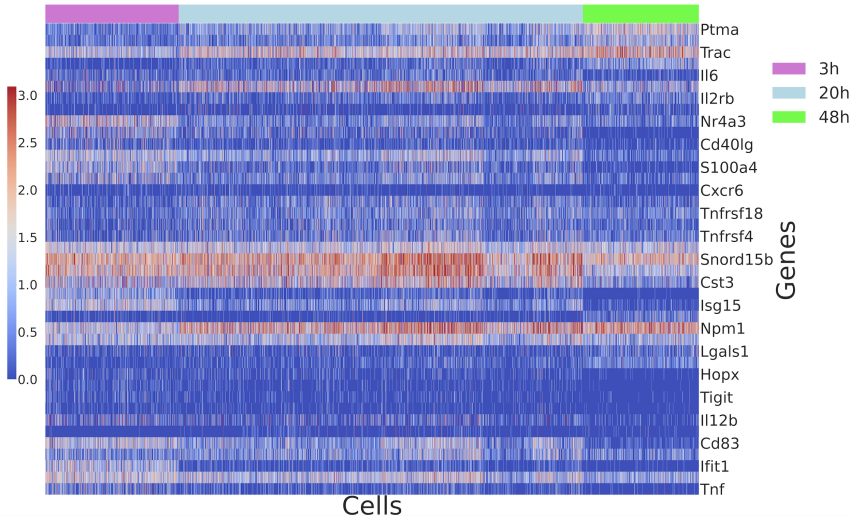


Figure 5.8: Log transformed expression of genes identified by both (Giladi et al. 2020) and our ranking approach (top 20) and showing the possible temporal expression pattern. For example, genes *Ldha, Ptma, Pcna, Trac, Dut* have the highest probability in the same new topics and their expression increases after the first 3h, while *Tnfrsf4, Tnfrsf9, Tnfrsf18* do not seem to show a temporal pattern and the shift in their expression compared to the single cell type co-cultures is captured by the same topics.

5.3.3 BM dataset

The original work by Boisset et al (Boisset et al. 2018) is focused on identifying significant interactions between cell types, using sorted BM cells and needle dissected doublets. Firstly, we fit our LDA on the sorted BM cells and fix the learned topics. Next, we fit the second LDA on the needle dissected doublets.

We hypothesised our approach would be able to identify genes involved in the main interactions discussed by (Boisset et al. 2018). Their work considered three specific interacting pairs: macrophages and erythroblasts, plasma cells and myeloblasts/promyelocytes, and megakaryocytes and neutrophils. Macrophages and erythroblasts have been known to interact, and erythroblastic islands are considered an important niche for the maturation of red blood cells. In addition to anchoring erythroblasts within island niches, macrophages also provide interactions which are important for erythroid proliferation and differentiation (Chasis & Mohandas 2008). When describing physical interactions, adhesion molecules are of particular interest. In our analysis we identify *Vcam-1* and *Itgam*, which are known to support adhesive interactions in macrophages.

Boisset et al (Boisset et al. 2018) also identify and validate the interaction between megakaryocytes and neutrophils. Their findings support other studies that have looked at emperipolesis (whereby neutrophils are engulfed by BM megakaryocytes) as a process mediated by both lineages. This interaction is important for production of platelets. (Cunin et al. 2019) identified that emperipolesis is mediated by $\beta 2$ integrin Cd18 and Icam-1 interaction. Blocking $\beta 2$ integrin *Cd18 (Itgb2)* impairs emperipolesis (Cunin et al. 2019). This is another integrin we identify in our analysis. *Elane* and *Igj* are two other genes discussed by Boisset et al (Boisset et al. 2018) that we identify to require additional topics to model their expression. The genes shown in Fig 5.9 are identified by taking the top 25 genes from each topic. Overall, based on top 25 genes ranking per new topic (over 300 genes in total), we identify genes linked to cell adhesion, innate immunity, and immune response. The full list of genes can be found in the supplementary information, B.4. While some of the genes are known to be linked to neutrophils (*Cd177, Prtn3, Serpinb1a, Lsp1*), other genes are less well-annotated in terms of function, and as such this demonstrates the benefits of using an approach that is not based on curated resources of known interactions.

5.3.4 COVID-19 dataset

Previously, we used datasets generated by modifying standard protocols to allow for PICs to be generated. However, here we explore the potential of our method to identify genes that change as a result of interaction in datasets generated with the 10x Chromium platform, which does not have the ability to preserve interacting cells as there is no specific way of capturing doublets. We took a recently published COVID-19 BALF dataset containing several cell types like T-cells, macrophages, B-cells, DCs, and neutrophills. There are samples from patients with moderate COVID-19, severe infection, and healthy controls. During cluster annotation, the authors labeled

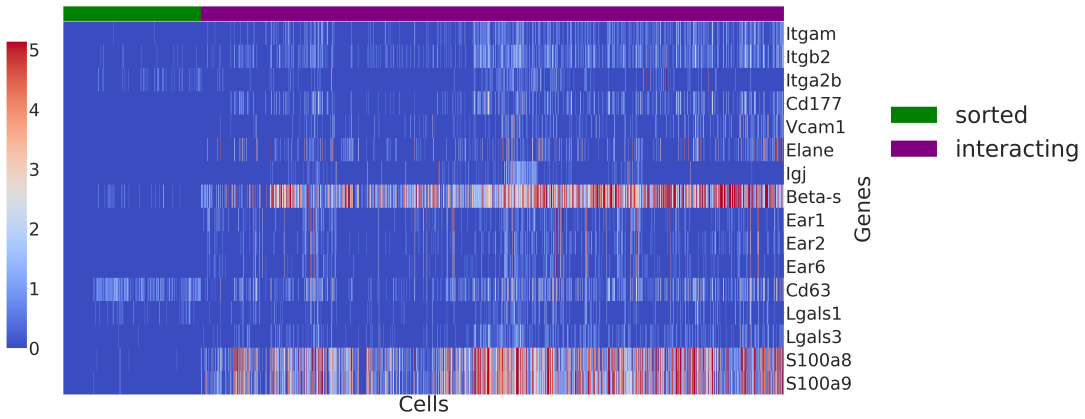


Figure 5.9: Heatmap of genes chosen based on their ranking in the new topics and evidence from literature that they are linked to interactions in the BM (log-transformed). Amongst the examples, we see genes linked to neutrophil adhesion such as *S100a8*, *S100a9*, and *Cd177*. *Elane* and *Igj* are two genes confirmed by (Boisset et al. 2018).

several clusters as doublets Liao et al. (2020). We hypothesised that some of those doublets might represent interaction, as macrophages are known to interact with T-cells. To confirm the identity of the doublet cluster in the severe illness patient samples, we looked at marker gene expression followed by analysis with DoubletFinder.

We use the populations labeled as singlets by DoubletFinder as a reference for LDA. We fix those topics and fit a second LDA on the doublets population. Identification of potential interacting genes was performed similarly to the datasets analysed earlier, by ranking genes within each new topic based on how many cells require this particular topic to explain the gene expression. Additionally, as we only have just over 200 doublets, we only consider genes using a specific topic in at least 10 cells. As we can see from Fig 5.10, some of the genes that require contribution from the new topics to model their expression include cytokines and chemokines, which might suggest interaction. A subset of the cells also seem to require new topics for genes related to interaction. We refer to work by Takada et al (Takada et al. 2007) and Magee et al (Magee et al. 2012) that discuss physical interactions to identify suitable candidates. As not all doublets require new topics, it is possible we have a mix of interacting and technical doublets. While we are capturing a shift in the expression of certain genes, our results are inconclusive, potentially due to the quality of our reference population as the reference is constructed based on the cells DoubletFinder annotated as singlets, and as such the reference might contain some interacting cells. Depending on the amount of cells loaded, a standard 10x Chromium experiment can result in the formation of over 0.7% technical multiplets with doublets being predominant. While with DoubletFinder we have managed to label some of the doublets, computational tools for doublet detection do not achieve perfect sensitivity and specificity scores, so it is possible the reference population contains cells that exhibit signs of interaction and changes in the expression

pattern of some potentially interesting genes. Our approach can identify genes that change as a result of interaction and is suitable for datasets where the reference population is clearly labeled as in the PIC-seq and dissociated BM examples discussed earlier.

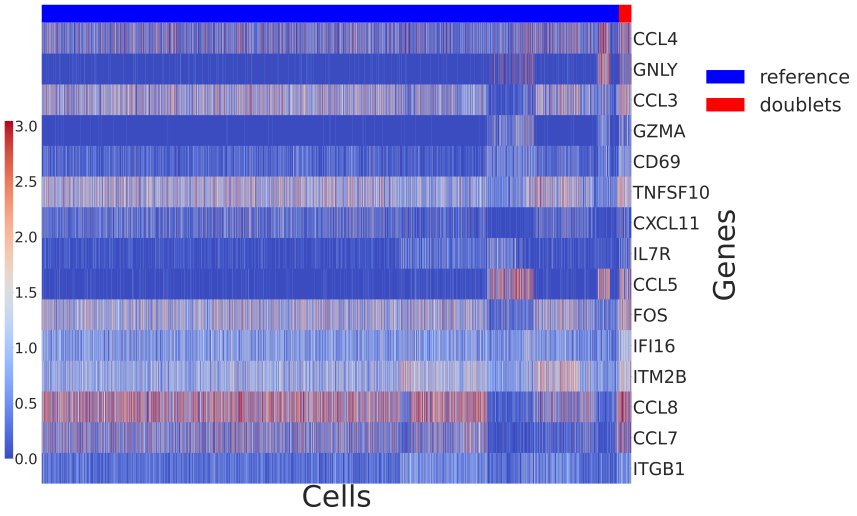


Figure 5.10: Heatmap of the COVID-19 data showing the expression of top ranked genes needing new topics for a subset of the doublets population (log-transformed). As none of the genes are uniquely expressed in the doublets population, we are capturing a shift in the expression of those genes. Some doublets can represent biologically interacting cells as the top ranking genes include cytokines and chemokines. However, due to the quality of the reference population our results are inconclusive.

5.4 Conclusion

In this chapter we have demonstrated the suitability of LDA for analysing PICs. We have shown that our model is sensitive to changes in gene expression that cannot be explained by the non interacting populations and thus new topics are needed to model the expression of genes that change as a result of interaction. Our model has been applied to two datasets of sequenced PICs and a dataset generated by standard 10x Chromium. Our approach assumes there is a reference population that can be used to fit the first LDA; for example this could be populations before an interaction has occurred. In addition to genes known to be involved in interaction and discussed by (Boisset et al. 2018) and (Giladi et al. 2020), we also rank further candidates for interaction that might be of interest for validation. We demonstrate the challenges of applying our approach to a dataset where a reference population cannot be clearly labeled in the case of the COVID-19 BALF analysis using the standard 10x Chromium protocol. However, amongst the top genes we rank there are cytokines and chemokines, genes known to be regulated by physically interacting cells, which might suggest the doublet population includes both technical and interacting doublets. While this is informative, the current setup of the 10x Chromium protocol is not fully suitable for

studying cellular crosstalk of physically interacting cells, and as such the likes of PIC-seq should be considered when studying interactions.

An approach that models jointly the reference and interacting population might be a better fit. However, such a model would introduce additional complexity. In practice this might result in a model that scales poorly.As such here we have focused on our 2-step LDA procedure that is able to capture genes that change as a result of interaction.

As seen from the datasets discussed here, modelling interactions based on doublets can be very useful. As such, distinguishing technical from biologically significant doublets poses an interesting challenge. While we have applied our LDA approach to PICs, there is potential for our work to allow for distinguishing technical doublets from transitioning cells as long as the transition is described by a unique set of genes, so that a new topic can be defined. While there are cell hashing methods that allow for mitigation of batch effects and overloading of sequencers, those methods help identify doublets between different samples/patients while intra-donor doublets remain undiscovered. On the computational side of doublet detection, methods make a range of assumptions that pose challenges to using them in practice. For example, DoubletFinder requires doublet rate which is not available in all experiments. The accuracy of all methods is affected when applied to transcriptionally similar cells, and DoubletDecon would not allow a doublet cluster to be present in the data. DoubletDecon is the only method able to distinguish technical doublets from transitioning cells. As such, there is a clear need for methods that can eliminate technical noise, but not at the expense of biological significance (DePasquale et al. 2019, Kang et al. 2018, McGinnis et al. 2019, Stoeckius et al. 2018, Wolock et al. 2019).

As PIC-seq is a very powerful approach, it can potentially be used to generate data including physically interacting doublets as well as singlets. It would be of interest to identify whether some of the singlets in fact show signs of interaction. Are they cells that have interacted but separated? Or maybe they have not interacted at all? Such datasets could easily be analysed following the methodology described earlier. We would expect to see signs of interacting topics in some singlets but maybe not all and as such we should be able to distinguish between singlets and singlets that have interacted before.

As there is a demand for understanding PICs, we believe methods like PIC-seq will be used more often in future and further work will be done to develop sequencing protocols that allow for capturing physical interactions that have the potential to become therapeutic targets. When such datasets are generated, there should be techniques that allow for their analysis and are not limited to knowledge captured in biological databases. The method described here is one such example that does not require any prior information such as clustering of the cell types involved and generation of synthetic reference profiles. As further datasets are generated, fields that would benefit from more in depth understanding of interactions include: understanding parasite-host interactions, crosstalk between immune cells and other lineages, and effect of cell-cell interaction in cancer progression.

Chapter 6

Using dynamic topic modelling to study temporal scRNA-seq data

6.1 Introduction

6.1.1 Gene expression over time

Previously, in Section 2.3.8, methods for pseudotime and trajectory inference were discussed. However, in practice estimating the pseudotemporal ordering is not the last step. Once cells are ordered in time, there are options for further analysis that aim to gain understanding of the underlying biological process, be it a disease or developmental one. It is assumed that if a gene changes its expression over the course of the pseudotime (also referred to as differentially expressed over pseudotime), then this gene is vital to the process of interest. As such, there are options for identifying such genes that change their temporal pattern across pseudotime or between different branches of a trajectory. There are also options for genes with similar expression patterns to be clustered over time. Of course, there are challenges in performing this analysis such as uncertainty of pseudotime inference, considering genes in isolation, and others. Additionally, as there is an increasing interest in comparing healthy and disease conditions over time, trajectories across conditions need to be aligned to determine where they start differing for the first time (Alpert et al. 2018). The next section provides an overview of methods developed to analyse gene expression dynamics.

One method that allows identifying differentially expressed genes over lineages or between two lineages is tradeSeq (Van den Berge et al. 2020). The method is independent of the previous steps for dimensionality reduction and pseudotime inference. tradeSeq uses a generalised additive model (GAM) where there is a separate smoothing spline for each lineage. The smoothing coefficients of each lineage are then used to assess the differential expression within or between lineages. tradeSeq implements several statistical tests. For example, in the case of testing of genes within the same lineage, that could be start versus end point gene expression or association

of genes to specific lineage. The p-values are used as a numerical summary for ranking of genes for further analysis. Additionally, the GAM can be used for clustering of gene expression patterns which are then plotted for further exploration.

To tackle the issue of pseudotime uncertainty when it comes to statistical tests for differential expression PseudotimeDE was developed (Song & Li 2021). PseudotimeDE subsamples 80% of the cells 1000 times (default parameter value). For each subsample, the same pseudotime inference (with the same parameters) is performed. The pseudotime of each subsample is then permuted. Similarly to tradeSeq, PseudotimeDE fits negative binomial-GAM to every gene in the original data. The same model is then fitted to each subsample to approximate null values for the test statistics. Finally, PseudotimeDE calculates a p-value from the gene's statistic in the original dataset and the approximate null distribution. The current implementation only handles DE genes within the same lineage. Furthermore, due to the subsampling hierarchical topologies cannot be analysed.

Most existing methods perform gene expression clustering as a two-step process: cells are firstly ordered in pseudotime and then clustering is performed. Pseudotime orderings are often subject to uncertainty, and one method that quantifies the uncertainty in both by jointly inferring pseudotemporal ordering and gene clusters is GPseudoClust (Strauss et al. 2020). Since GPseudoClust uses MCMC to sample from the complex posterior distribution, it is computationally expensive, does not scale well with high dimensional single cell datasets, and relies on pre-selecting genes across the different time points (Strauss et al. 2020).

6.1.2 Adding a temporal dimension to scRNA-seq analysis

One of the key strengths that comes with single cell data is the ability to recover a cell's progress through a process of interest. By taking into account the underlying dynamic of biological systems, we can gain an insight into which genes are co-expressed over time; which genes change across time and which genes are not time-dependent. Exploration of co-expressed genes can then lead to better understanding of gene regulation.

Furthermore, single cell data are inherently noisy and biased towards highly expressed genes. As such, being able to smooth the noise by adding an ordering dimension, time specifically, can allow us to identify genes that otherwise would be missed by traditional methods such as differential expression.

6.1.3 Extensions of traditional LDA and applications to transcriptomics data

The work discussed in Chapters 4 and 5 was based on LDA. However, under the standard topic model formulation documents are independent. In the original topic model as proposed by Blei (Blei et al. 2003), each topic is defined as a distribution over a vocabulary and the words in a

document come from a mixture of topics. However, for certain collections of documents there might be an underlying dependency which can be captured by evolving topics. An example of such a dependency is documents across time. Dynamic topic models allow for two ways of modelling dynamics: the document-topic probabilities can be changing over time and/or the topic-word probabilities can be changing over time. Additionally, topics should not be considered in isolation which led to another proposed extension by Blei (Lafferty & Blei 2006). Recently, (Tomasi et al. 2020) propose a scalable dynamic correlated topic model (DCTM) which models the evolution of topics, words in those topics, and their correlations.

To help obtain an intuition about dynamic correlated topic models before describing them formally in the next section, we will consider the following example applicable to text. Our corpus consists of scientific publications about omics analysis. We can consider if the topics of *gene expression*, *proteomics*, and *metabolomics* have changed over time. Is there one that has gained more popularity? Taking the *gene expression* topic as an example, we can note that the words distributions have been changing over time. In particular, before 2009 (when the first scRNA-seq papers appeared) microarrays and bulk RNA-seq had higher probabilities. Furthermore, we can look into the correlation with other topics. The topic of *gene expression* is increasingly correlated with the topic of *machine learning* as gene expression datasets are becoming more high dimensional and require machine learning method development to analyse the results.

Chapter 4 discusses applications of topic modelling to scRNA-seq, from an alternative of clustering approaches to an approach of removing ambient RNA. While dynamic topic models (DTMs) have not been applied to single cell data, they have been used to study gene expression of time-series toxicogenomics microarray data. This dataset contains 3144 microarrays treated with 132 compounds across 4 timepoints. (Lee et al. 2016) consider the same up and down regulated gene as two different words. They explore the topics linked to the different conditions and match topics to functional pathways. Finally, they assess the evolution of genes over time. The study highlights the suitability of dynamic topic modelling to study biological systems as they manage to capture the complexity of the data and provide insights into gene regulation (Lee et al. 2016).

6.1.4 GP methods in scRNA-seq

GPs were introduced in Chapter 3. Variations of the Gaussian processes framework have been used to model gene expression data. Examples include clustering gene expression time series, ranking differentially expressed genes in a temporal dataset using GP regression, and removing technical and cell cycle noise by using GPLVM for dimensionality reduction (Buettner & Theis 2012, Kalaitzis & Lawrence 2011, McDowell et al. 2018) In the field of scRNA-seq, GPLVM has been used to infer pseudotime ordering (Ahmed et al. 2019). Additionally, given pseudotemporal ordering, a modification of the standard GP framework has been developed to infer the gene-specific branching dynamic (Boukouvalas et al. 2018). Finally, GPLVM has been combined with

clustering of time-series to allow for simultaneous ordering and identification of gene clusters in scRNA-seq data (Strauss et al. 2020).

6.1.5 Aims

The currently available methods for studying gene expression changes over time do not model processes together, thus not taking into consideration correlations of gene expression. Specifically, clustering genes based on pseudotime is often done in isolation. A model that takes into consideration process dynamic and correlation between processes over time can offer insight into co-expression and regulation. Additionally, correlations and ordering might enable the capture of genes with lower signal-to-noise ratio. This chapter aims to 1) evaluate the suitability of DCTM for modelling temporally ordered scRNA-seq data, 2) identify dynamic and correlated gene expression modules. Studying dynamic and correlated gene modules can allow for extrapolation of data at missing timepoints and can also enhance understanding of gene functions by improving available annotations.

6.2 Materials and methods

6.2.1 Dynamic Correlated Topic Model

As previously described, in the context of scRNA-seq: cells are equivalent to documents, genes to words, a topic is a group of genes that co-vary, and counts are word frequencies.

In the setting of DCTM, topic probabilities and words probabilities in a topic can change over time. Both topic and word dynamics are modelled by Gaussian processes. Gaussian processes have been introduced in Chapter 3. Let D be the timepoints, K is the number of topics, and N the number of words in each document. The generative process for a document d at timepoint t_d can be described as follows:

1. Draw a mixture of topics $\boldsymbol{\eta}_d \sim \mathcal{N}(\boldsymbol{\mu}_{t_d}, \boldsymbol{\Sigma}_{t_d})$
2. For each word $n = 1, ..., N$:
    - Draw a topic assignment $z_n|\boldsymbol{\eta}_d$ from a multinomial distribution with the parameter $\sigma(\boldsymbol{\eta}_d)$
    - Draw a word $w_n|z_n, \beta$ from a multinomial distribution with the parameter $\sigma(\boldsymbol{\beta}_{z_n})$

$\boldsymbol{\beta}_{z_n}$ is the word probabilities for a topic. σ is the softmax function defined as $\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$ which allows probabilities to be obtained for $\boldsymbol{\eta}_d$ and $\boldsymbol{\beta}_{z_n}$. Topic probabilities and the distribution of words over topics are modelled as zero-mean GPs, specifically $p(\boldsymbol{\mu}) = \text{GP}(\mathbf{0}, \boldsymbol{k}_\mu)$ and $p(\boldsymbol{\beta}) = \text{GP}(\mathbf{0}, \boldsymbol{k}_\beta)$.

Σ_{t_d} is modelled using a generalised Wishart process (GWP), derived from a set of GPs. GWP is a collection of positive semi-definite matrices indexed by an arbitrary input variable. This input variable can originate from any arbitrary set and can also represent time.

The marginal likelihood becomes:

$$p(W|\boldsymbol{\mu},\Sigma,\boldsymbol{\beta}) = \prod_{d=1}^{D} \int (\sum_{z_n=1}^{k} p(W_d|z_n,\boldsymbol{\beta}_{t_d}) p(z_n|\boldsymbol{\eta}_d)) p(\boldsymbol{\eta}_d|\boldsymbol{\mu}_{t_d},\Sigma_{t_d}) \mathrm{d}\boldsymbol{\eta}_d$$

(6.1)

6.2.2 Model inference

Sometimes the posterior distribution is intractable, there is not a closed-form solution, and as such approximation strategies need to be used. Given the real posterior distribution cannot be derived, the aim is to approximate it by finding a variational distribution. This approximation becomes an optimisation problem where the aim is to minimise the Kullback–Leibler (KL) divergence to the exact posterior. As we cannot compute the KL divergence, we instead maximise the evidence lower bound (ELBO). The overall idea and approach have been discussed in Chapter 3. The variational inference procedure for the DCTM is derived by assembling the lower bounds of the document-topic proportion inference, the GPs inference, and the Wishart process inference.

6.2.3 Relaxed LDA

Previously the generative process of LDA has been discussed and some of the sampling was done from Dirichlet distributions. However, DCTM and similar variations, like CTM and DTM, rely on logistic normal distribution obtained by drawing the document-topic proportions from a logistic normal. In order to compare DCTM with a simpler model that does not take time into account but also uses softmaxed multinomial for probabilities, a different version of LDA is presented.

The generative process of the relaxed or logistic normal topic model can be described as follows:

1. Draw a mixture of topics $\boldsymbol{\eta}_d \sim \mathcal{N}(\boldsymbol{\mu}_{t_d}, \boldsymbol{\Sigma}_{t_d})$
2. For each word $n = 1,...,N$:
    - Draw a topic assignment $z_n \sim \text{Multinomial}(\sigma(\boldsymbol{\eta}_d))$
    - Draw a word $w_n \sim \text{Multinomial}(\sigma(\boldsymbol{\beta}_{z_n}))$

Here, if Σ is a diagonal covariance, the model will exhibit some of the characteristics of LDA, specifically uncorrelated topics. While a non-diagonal covariance will introduce correlations between the topics and thus this logistic normal topic model will result in a correlated topic model (Mimno et al. 2008). Similarly to the DCTM, a softmax is used to compute the multinomial probabilities.

DCTM and relaxed LDA experiments are based on the implementation of the original publication (Tomasi et al. 2020). The GitHub for this chapter: https://github.com/alexpancheva/sc-DCTM

6.2.4 Autocorrelation of time-series

Analysing the autocorrelation of time-series data allows for the evaluation of how predictive earlier timepoints are of future data. Since topic and word probabilities over time are modelled as Gaussian processes, smooth functions can be expected. In order to evaluate the ability of the model to find meaningful patterns over time, we compute autocorrelations in the cases of a randomised pseudotime dataset and a dataset where cells have been ordered in pseudotime. Once we fit the DCTM on the data, we obtain a distribution of topics over time. We compute autocorrelation based on the topics by cell matrix after fitting the model on the two datasets.

6.2.5 Ranking genes in topics

While genes per topic can be ranked based on probability, such ranking might rank highly, across all topics, potential "background" genes, expressed highly in all cells. As such, in order to assign high ranking to genes that distinguish a topic compared to all other topics, the extractTopFeatures function from the R package CountClust, version 1.16, is used. Provided with a topics by genes probability matrix, for each topic the distinctiveness of each gene g is measured with respect to any other topic using KL divergence. Genes are ranked per topic based on maximisation of the min KL divergence with other topics. This ranking of genes per topic is used for all topic models.

6.2.6 Topic interpretation

In order to identify the biological significance of topics gene ontology (GO) terms are used. For the Malaria Cell Atlas, the gene association file (gaf) file is downloaded from PlasmoDB. For each topic, all genes are ranked as described in the previous section. Once ranked, AUC score is computed per GO term:

- For each gene in the topic ranking, assign 1 (if gene is in the GO term) and 0 if not
- Compute AUC score using inverse ranking of genes (highest gene ranked at position N where N is the total genes in the data) and GO binary score

Only GO terms with AUC>0.8 are used for comparative analysis. GO terms containing fewer than two genes are excluded.

6.2.7 Choosing interesting topics

While for the purposes of the comparative analysis all topics are used, in practice it might be of interest to prioritise which topics to explore. As the focus of this chapter is temporal processes, topics that are time-varying are of interest. To eliminate topics corresponding to noise or background processes, we perform Durbin-Watson test for autocorrelation in the residuals, as implemented in the Python statsmodel package using the topic distribution over time matrix. A score of 2 indicates no correlation, while 0 and 4 correspond to positive or negative correlation respectively.

6.2.8 Choosing the number of topics

To choose the most suitable number of topics across all topic models, we use perplexity for a range of topics. Perplexity was defined in Section 3.3.5.

6.2.9 Comparison with scRNA-seq analysis

To compare the insight from DCTM to standard scRNA-seq analysis, data are clustered in the same number of clusters as topics used for LDA, relaxed LDA, and DCTM, in the case of the Malaria Cell Atlas, we use 20 topics. All clustering analysis was performed in Seurat 3.1. To obtain 20 clusters, the resolution parameter of FindClusters in Seurat is set to 1.12. Next, for each cluster all genes are ranked using Wilcoxon rank sum test, as implemented in the R package presto, version 1.0.0 using wilcoxauc function.

For each cluster, genes are ranked based on their adjusted p-value, based on Bonferroni. To obtain AUC score per GO term, for each topic for each GO term ranked genes are assigned 0 or 1, depending on whether the gene is absent or present in that GO term. Then the sklearn AUC function is used as described earlier, see Section 6.2.6.

6.2.10 Datasets

Malaria Cell Atlas

The dataset discussed here has been generated as part of the Malaria Cell Atlas (Howick et al. 2019) and is deposited at the European Nucleotide Archive at European Molecular Biology Laboratory European Bioinformatics Institute with accession number ERP110344. Samples have been generated using 10x Chromium and contain 4763 cells and 4890 genes following all pre-processing. Data with assigned pseudotime has been downloaded from the first author's

GitHub https://github.com/vhowick/MalariaCellAtlas. The authors order cells in pseudotime by fitting an ellipse to the first two principal components (PCs) and calculating the angle relative to a start cell. The results of this pseudotime ordering the authors of the original publication correlate with published bulk data (Howick et al. 2019).

This dataset was selected as it is expected that it will contain groups of genes with changing expression depending on the lifecycle stage. Furthermore, there are multiple cells per timepoint and the data and current knowledge of the lifecycle suggest it is a continuous process.

Dendritic cells

This dataset consists of dendritic cells stimulated with LPS, taken over several timepoints, 1, 2, 4, and 6 hours, originally generated by (Shalek et al. 2014). 390 LPS stimulated cells and 4016 genes are used. Data have been downloaded from (Song & Li 2021) and cells have been ordered in pseudotime using Slingshot. The original publication identifies genes that show time-dependent behaviour.

6.3 Results and discussion

6.3.1 Randomised control and ordered data autocorrelation

While pseudotime analysis can be considered a useful approach for analysing single cell data, it is uncertain, often this uncertainty is not quantified by the pseudotime method and above all there will always be an ordering produced by the method, sometimes independent of the underlying biology. Assuming there is a pseudotemporal ordering that indeed is modelling an underlying biological process, can DCTM identify this process and pick up the temporal signal when it exists?

To investigate this question and to demonstrate the sensitivity of DCTM to temporal data, DCTM is fitted on both randomised and ordered data. This experiment used the Malaria Cell Atlas 10x *P. berghei* ordered in pseudotime and completely randomised. Using the topics over time probabilities for DCTM, autocorrelation is computed. This allows us to evaluate how predictive earlier timepoints are of later data: in the case of temporal signal, autocorrelations should be predictive while in the case of randomised data autocorrelation should be low.

Figure 6.1 illustrates autocorrelations of the topics over time are higher in the case of the ordered data compared to the randomised experiment. In the case of the randomised experiment, even for closer timepoints the correlation is close to 0, suggesting time-series forecasting is not possible.

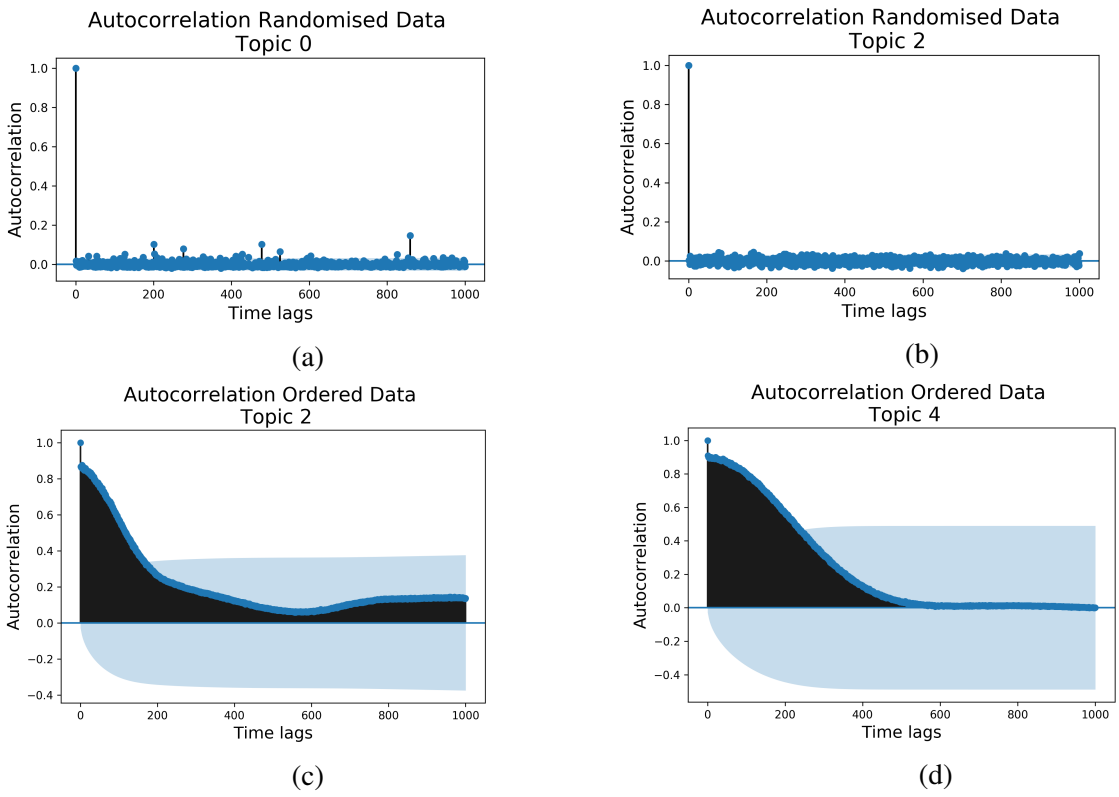


Figure 6.1: Autocorrelations for some topics in the randomised and ordered Malaria Cell Atlas data. (a) (b) Autocorrelation in randomised experiment. Autocorrelation is close to 0. (c) (d) As expected in the ordered data, we obtain higher autocorrelation and the earlier timepoints are predictive of later data

6.4 Model comparison

So far it has been demonstrated that DCTM captures temporal signal based on pseudotime ordering. Next, we evaluate whether adding temporal dimension results in a model that outperforms simpler topic models and standard scRNA-seq analysis. This is evaluated first by comparing perplexity across models and then performing GO analysis.

6.4.1 Comparison with relaxed LDA and LDA

We evaluate how the results from the DCTM compare with other topic models applied on the same dataset.

Firstly, we use LDA with relaxation in order to make the models more comparable. In the setting of relaxed LDA, η is drawn from a normal distribution and not from a Dirichlet distribution. While for relaxed LDA and DCTM normal distributions are used, in the original LDA the topic-word and the document-topic priors come from a Dirichlet distribution. While a comparison between relaxed LDA and DCTM is more appropriate due to base similarities between the two models, here we also include a comparison with LDA for completeness and as Dirichlet is a more natural distribution for a probability simplex.

In order to assess the suitability of the different LDA models, we compute perplexity for a varying number of topics for each model. Perplexity is computed on a subset of data that has not been used for training to avoid overfitting. Lower perplexity values are preferred. Results can be seen in Table 6.1.

Number of topics	Perplexity Relaxed LDA	Perplexity DCTM	Perplexity LDA
5	2187.90	1174.08	1156.48
10	2030.55	1167.71	1154.20
20	2187.24	1169.23	1214.66
30	2521.22	1261.49	1262.13
40	2779.64	1276.83	1291.79
50	2820.04	1290.09	1338.16
60	2673.77	1305.72	1362.24

Table 6.1: Comparing the perplexity values of three topic models for a range of topics. Perplexity is computed using a previously unseen random subset of the data to prevent overfitting. Lower perplexity indicates that the model is a better fit for the data.

DCTM and LDA generally result in lower perplexity and as such can be considered a better fit for the data compared to relaxed LDA. However, DCTM and LDA have similar perplexities for a range of topics.

To further evaluate how good the fit of DCTM, relaxed LDA, and LDA are, we perform GO analysis for each model to evaluate how those topics map to biological insight, Section 6.2. We fit all models with 20 topics, rank all genes and identify GO terms with AUC>0.8.

While there is some overlap between DCTM and the relaxed LDA, many GO terms are only unique to the DCTM. Examples include ribosomal biogenesis processes, gene expression, processes related to cell motility, and microtubule-based processes. Those GO terms have been highlighted by either (Howick et al. 2019) or (Caldelari et al. 2019).

Compared to the overlap between DCTM and relaxed LDA, there are additional GO terms only shared between DCTM and the standard LDA, 61 GO terms. While there is a temporal component to the data, this temporal component also corresponds to cell types and as such LDA is also able to capture those terms.

Unique to DCTM for example are processes related to cell motility, regulation of gene expression, microtubule-based processes, protein folding, ribosomal biogenesis, and others. For completeness, all GO terms unique to each method can be found in Table C.1.

6.4.2 Comparison with differential expression

While LDA is a model gaining popularity in scRNA-seq and has been applied to a range of problems, the state of the art single cell analysis for GO enrichment generally relies on differential expression. As described in Section 6.2.9, the data are clustered in 20 clusters, the same as

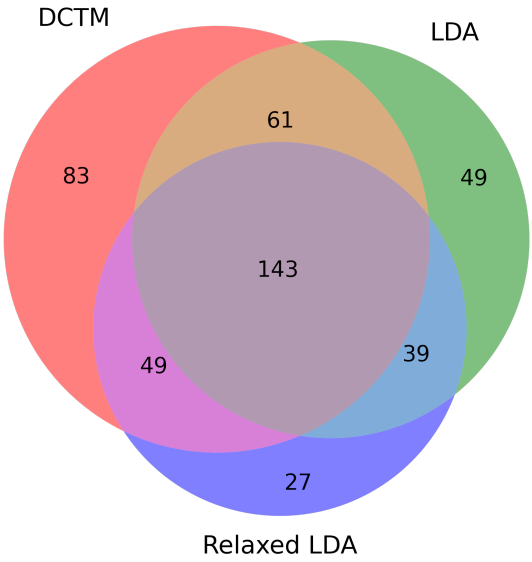


Figure 6.2: DCTM and LDA with Dirichlet priors have better overlap of GO terms compared to relaxed LDA. There are 61 terms shared only between DCTM and standard LDA.

number of topics, using Seurat 3.1. Then Wilcoxon rank sum test is used to rank the genes for each cluster. Complete details can be found in Section 6.2.9.

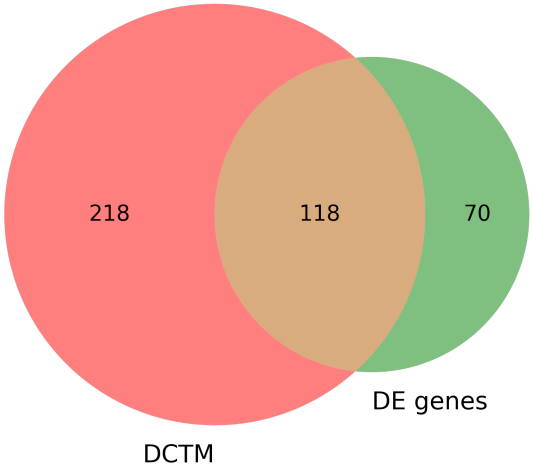


Figure 6.3: There is high overlap between DCTM and standard DE analysis potentially due to the fine clustering granularity. However, DCTM uncovers temporal GO terms not captured by DE.

Potentially due to the fine granularity of clustering, some of the overlapping GO terms between DCTM and DE analysis include temporal terms which the other topic models did not identify for microtubule-based process, some biosynthetic processes, regulation of protein catabolic process, and metabolic processes. Here again DCTM has a higher number of unique GO terms (218) including protein folding, RNA methylation, and gene expression.

6.4.3 Malaria Cell Atlas

Following the initial experiments, the application of DCTM to the Malaria Cell Atlas can be explored in more detail. The topic-word probabilities are modelled as a GP with Matérn 1/2 kernel with amplitude 1 and length-scale 7.5 to allow for word probabilities that are not changing quickly, and the document-topic probabilities are modelled as a GP with exponential quadratic kernel with amplitude 1 and lengthscale 0.5. Those are then learned in the model. This would mean we are allowing for "flatter" gene probabilities and topics which change more rapidly, which is expected due to the presence of multiple cell types over the timeframe involved.

As shown in Table 6.1, the perplexity scores for this dataset for 10 and 20 topics are very similar, and so a model with 20 topics is preferred to allow for potentially further groups of interesting genes to be captured. We expect to find topics that are changing over time but also topics that are relatively stable, corresponding to housekeeping processes involved in all cells. Furthermore, as we have shown earlier in Chapter 5 under-specifying the number is problematic as the complexity of the data is not captured.

Figure 6.4 illustrates that topics are found with higher probability over certain stages of the lifecycle. For example, topic 13 has higher probability between 0 and 1. Topic 6 has high probability between -0.25 and 0.5. Topics 14 and 15 have high probability between -1 and -0.5. There are also topics with low and similar expression over the lifecycle, for example topics 0, 9 and 10.

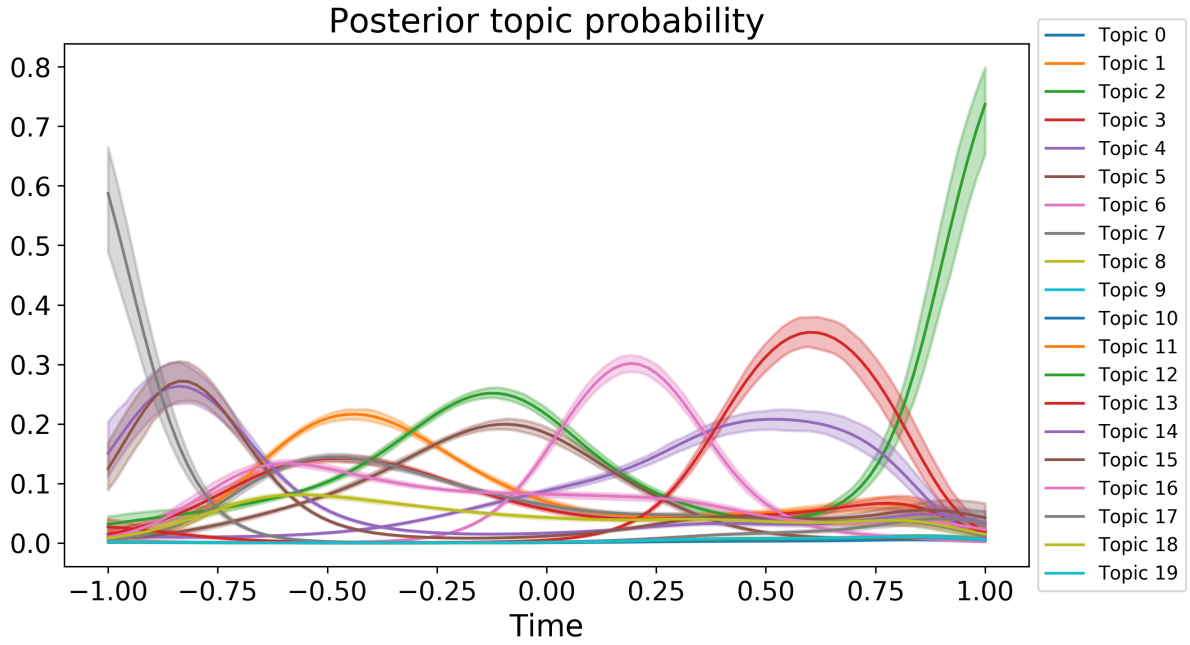


Figure 6.4: Posterior topic probabilities of 20 topics. While some topics have lower and relatively constant probabilities over time, some topics are highly expressed in particular timeframes.

Using the topic selection approach described earlier, we take the first 100 genes from some of

the topics that have positive autocorrelation, examples include 17, 14, 1, 5, 6. Results are seen in Figure 6.6. We are able to capture genes that co-vary within a specific timeframe of the lifecycle. While DCTM enables the identification of such groups of genes, it is important to evaluate the biological significance of these topics. To interpret the topics, we identify the GO terms based on the first 200 genes per topic.

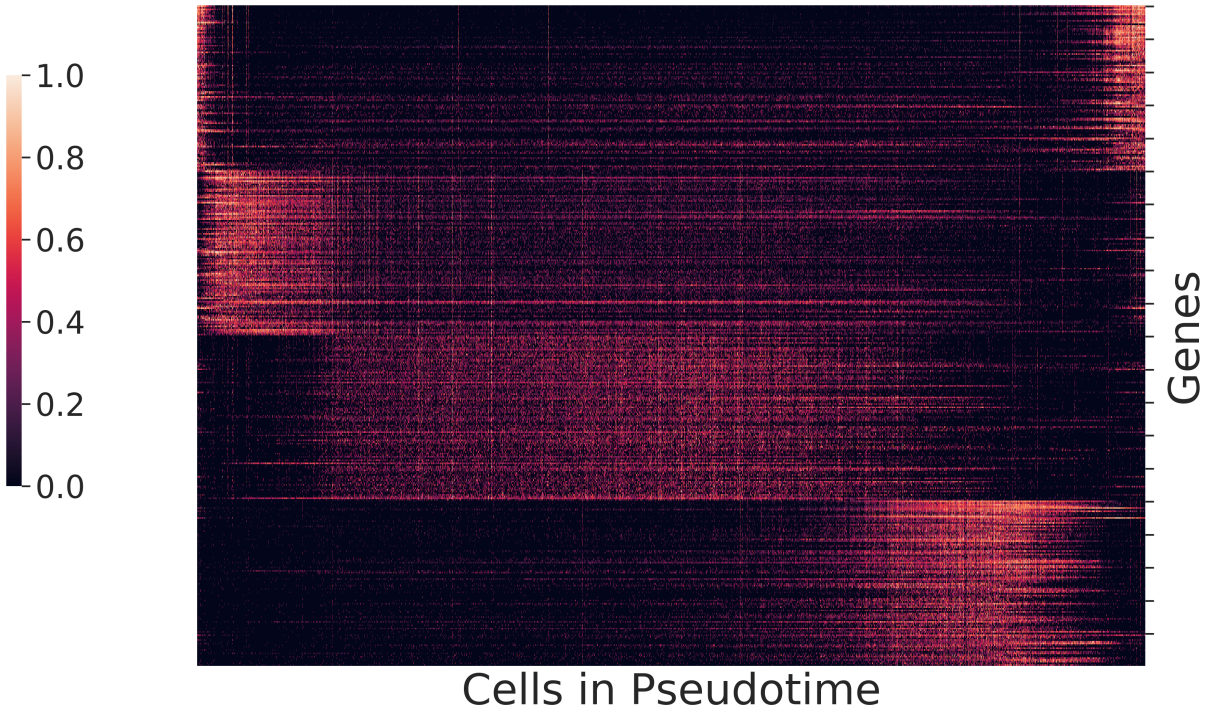


Figure 6.5: Top 100 genes from topics 17, 14, 1, and 6. We choose topics with positive autocorrelation as they will express temporal effect. Some of the other topics we capture are expressed at particular life-stages as well, e.g., 13 and 15. However, we also capture topics that are fairly constant along the lifecycle, topics 9 and 10. Log1p expression heatmap with standard scaling on rows.

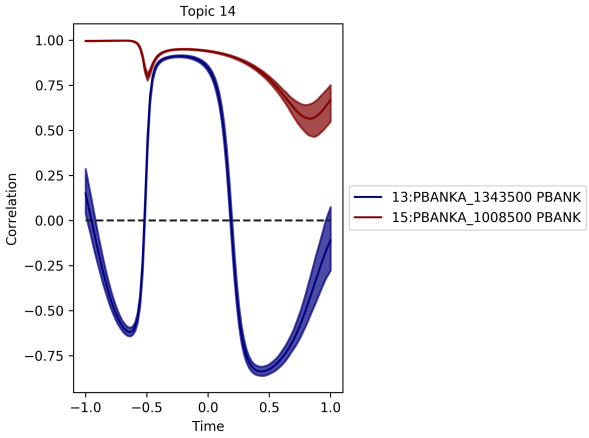


Figure 6.6: Temporal correlations of topic 14 with topics 13 and 15.

Topics 14 and 15 are both expressed between -1 and -0.5 timepoints with significant GO terms like ribosome biogenesis, RNA metabolic process and processing, and gene expression. Topic 13 contains genes associated with protein phosphorylation. The identified GO terms are linked with findings of bulk studies of the parasite. Caldelari et al show high expression of genes associated with protein phosphorylation in the schizonts stages, which correspond to timepoints between 0.33 and 1 and topic 13 (Caldelari et al. 2019). Finally, topics 9 and 10 that appear fairly constant over the lifecycle include ribosomal genes and PIR genes which have low expression across all stages as also shown by the Malaria Cell Atlas (Howick et al. 2019).

6.5 Dendritic cells

Previous analysis covered a lifecycle dataset with different cell types, where topics changed more rapidly due to differences in the cell types. Here DCTM is applied to dendritic cells, a single cell type, which is stimulated over time causing particular genes to change over the timecourse. Unlike the previous dataset, here there are fewer cells (in some cases 1) per pseudotime point and the timeframe is shorter. As the dataset consists of one cell type with some genes changing over the pseudotime in response to stimulation, we initialise the lengthscale for the $\boldsymbol{\eta}$ to 100 to allow for "flatter" topics or topics that remain constant during the pseudotime. A model with 10 topics, topic probabilities that are not changing over time, and genes changing results in a better perplexity compared to a model with changing topic probabilities.

The original publication by (Shalek et al. 2014) identifies genes related to antiviral and inflammatory response peaking towards the later time-points. The posterior probabilities of a selection of those genes are plotted in Figure 6.7.

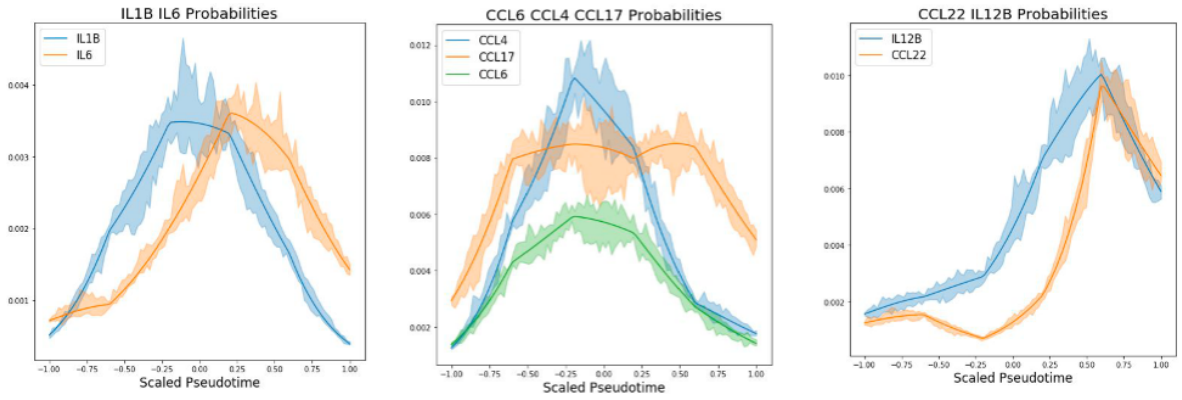


Figure 6.7: Examples of genes related to inflammation that change their expression over the pseudotime course following stimulation with LPS.

6.6 Model flexibility and practical considerations

DCTM is a very flexible model that can be used to model dynamics in several different ways: in either topics and words or both; stable topics across the timeframe and changing gene probabilities; or fairly constant gene probabilities and changing topic probabilities. While DCTM can adapt well to datasets given optimisation of kernel hyperparameters, here it is also important to note that interpretability is key when it comes to the outlined biological context. If the data are modelled with both changing topic and word probabilities, results will be more difficult to interpret in some situations. As such, here we choose where to place the complexity of the model, either in the changing topic probabilities (Malaria Cell Atlas) or the changing gene probabilities (Dendritic cells). Two examples have been selected to illustrate that depending on the type of data and nature of experiment, complexity can be captured by either the topics or gene probabilities.

6.7 Conclusions and possible future directions

Once data are ordered in pseudotime, further analysis can be performed, examples include gene clustering, identification of differentially expressed genes over pseudotime, or co-expression analysis. However, current work considers genes in isolation and does not account for temporal correlations. This chapter proposes applying DCTM to pseudotime ordered scRNA-seq data and evaluates the suitability of the model for understanding gene expression changes over time.

We have shown that DCTM uncovers meaningful temporal patterns in the data. Additionally, adding a time component improves the biological interpretation and with DCTM we find more GO terms compared to LDA models as time allows for detection of potentially noisier but interesting genes. Finally, we have discussed practical considerations when using this model on scRNA-seq data ordered in pseudotime and the different ways of modelling dynamics. In the case of the Malaria Cell Atlas due to the changing cell types, changing topic probabilities are more suitable while in the case of one cell type over a short period of time changing gene probabilities result in a more interpretable model.

While it has been shown that adding temporal information is a better alternative to non-temporal topic models, there are some limitations to this approach and some interesting extensions and further work that could be considered. Specifically:

- Adapt the current model to more complex trajectory structures. At present it is possible to fit DCTM on a linear or circular trajectory as long as there is no branching involved. However, often biological processes are more complex and include multiple branches where different groups of genes switch on and off. As such adding some branching to the topics should be considered.

- Currently, the analysis relies on data correctly ordered in pseudotime and the uncertainty of pseudotime ordering is not taken into consideration. A model that infers the cell orderings

and topics jointly might be a better fit. However such a model could be very complex and potentially scale poorly.

Chapter 7

Conclusions and Future Work

In this thesis, three methods based on topic modelling have been described and applied to scRNA-seq data. One of the main aims of this thesis was to develop interpretable models for scRNA-seq, and topic modelling is one such example. In this context, cells correspond to documents, words correspond to genes, and a topic is a group of co-varying genes. Due to its assumptions of multiple topics being expressed in a document and a word being part of multiple topics, topic models are suitable for biological problems. The models are able to identify interpretable topics that reflect both housekeeping processes and genes specific to cell types. Next, the work aimed to relax some of the assumptions of existing methods that affect their ability to be applied in practice, for example known doublet rate, artificial simulation of doublets at specific proportions, and initial clustering of data. Examples of that are the doublet detection approach proposed in Chapter 4, and the proposed LDA-based approach for detecting genes that change as a result of interaction in Chapter 5. In some cases incorporating another layer of information can also allow for obtaining more interpretable results and one such example is the addition of a temporal dimension to the topic model.

7.1 Doublet detection

In Chapter 4, LDA was combined with entropy scoring aiming to identify doublets in scRNA-seq data while making fewer assumptions compared to state of the art approaches, e.g. known doublet rate, doublets being 50/50 or 30/70 contribution of two cells, no doublet cluster in the data, and others. The approach was evaluated on synthetic and real data with annotated doublets. The proposed approach does not achieve better sensitivity and specificity compared to existing methods, and all compared methods cannot identify homotypic doublets which are generally considered benign as they do not affect downstream analysis. Furthermore, all methods are sensitive to count depth as shown in the downsampling experiment. A potential way of improving doublet annotation would be to take the union of all predicted doublets by different approaches. Following results evaluation, the limitations of the proposed approach are recognised and entropy

scoring in its current state is considered inappropriate for doublet detection. The current setting can be modified to consider entropy based on cluster level as some cell types might have higher entropy than others. Furthermore, in Chapter 4 housekeeping genes expression was explored as a potential way of detecting doublets.

Since the experiments of Chapter 4 have been performed, novel methods have emerged aiming to remove doublets computationally. Examples include a neural network based approach called Solo (Bernstein et al. 2020). Similarly to previously described methods, DoubletFinder and DoubletDecon, Solo also creates simulated doublets. Next, the model is trained to distinguish *in silico* doublets from observed data. Unlike previous methods, Solo embeds cells in latent space using a variational autoencoder. The final step adds a classifier at the end of the encoder (Bernstein et al. 2020). Another recent method, cxds, takes a different approach to doublet detection as it does not generate artificial doublets. Instead, it relies on the assumption that a heterotypic doublet would express the marker genes of multiple cell types. Gene pairs are ranked based on how often they are co-expressed (Bais & Kostka 2020). Furthermore, a comprehensive benchmarking study now exists that compares all available doublet detection methods (Xi & Li 2021). Similarly to the analysis described here, the study considers datasets with variable sequencing depth and heterogeneity. Their findings echo the results of this thesis that the performance of most methods is affected by low sequencing depth, and doublets in a homogeneous populations are more difficult to identify. Finally, as suggested by this thesis and the work of Xi et al (Xi & Li 2021) the best performing method, DoubletFinder, relies on known doublet rate which makes it difficult to use in practice as doublet rate is often unknown. However, while cxds is the mostly scalable and computationally efficient method, it demonstrates unstable performance. Xi and colleagues confirm our observations that a potential avenue to be explored is an ensemble method for doublet detection (Xi & Li 2021). The option for doublet identification based on an ensemble method has been recently explored in a method called Chord (Xiong et al. 2022). Chord firstly removed doublets with cxds and DoubletFinder. Simulated doublets are created and added to the data. Then a generalised boosted regression model (GBM) is fitted on the training data. The GBM integrates and weights the predictions of the doublet detection methods. In the final step, the GBM is used to predict doublets in the original data (Xiong et al. 2022).

Studying differences between conditions has been the basis of multiple studies with detection power and replicates being of considerable interest. To achieve this, HTO tagging or donor SNP information can be used for the samples in the same run, which also simplifies doublet detection tasks. In addition to the antibody-based demultiplexing, there are now protocols available that enable lipid-based demultiplexing (Mylka et al. 2022). Removing doublets from scRNA-seq data allows for high concentration loading of experiments without imposing strict filtering constraints, however with the increasing affordability of sequencing and some of the advances of cell type identification (e.g. mapping approaches instead of unsupervised clustering), the question of doublet detection should be reframed to distinguishing technical from biological doublets as

further exploration of biological doublets might be a way of studying interactions or transitions.

While the approach proposed in Chapter 4 based on LDA and entropy scoring does not outperform existing methods, this chapter benchmarked currently available methods in the field at the time, highlighted important issues, and outlined potential avenues for exploration.

7.2 Cellular crosstalk

While it is generally considered that doublets are a technical artifact, sometimes they are an indication of interacting cells. This forms the basis of the work described in Chapter 5, where a 2-step LDA procedure to identify genes that change as a result of interaction is used. The proposed method does not require prior clustering or formation of artificial doublets. Our approach has been tested on protocols specifically designed to capture interaction, PIC-seq and isolation of interacting cells, and a standard sequencing protocol, 10x Chromium. In the case of the specialised protocols able to capture interactions, the reference population of singlets, cells before interactions have occurred, is clearly labelled. However, the setting of a standard 10x protocol is different. While some genes potentially linked to physically interacting cells are captured, results are inconclusive due to uncertainty in the reference population as it might already contain interacting cells.

In the case of PIC-seq and the needle dissociated bone marrow data, genes were identified to change as a result of interaction, however the interacting population is only made up of double-positive cells, cells that are currently interacting. An interesting future experimental setup would consider access to interacting doublets as well as cells that have separated following interaction. It would be of interest to explore not only the differences between cells before and during interaction but following interactions as well. However, to date there is a missing temporal dimension to such interacting datasets, specifically examining how cells change over the course of a disease or what happens once cells separate.

Furthermore, to gain a complete understanding of cell-cell communication adding a spatial dimension can be of interest, and such opportunities are now more accessible with the advent of spatial transcriptomics. For example, traditional single cell methods based on ligand-receptor knowledge lack the long-range diffusion aspect which can be uncovered in spatial data. Furthermore, when such ligand-receptor interactions are inferred in scRNA-seq they lack spatial context and as such spatial transcriptomics can aid interpretation. Leveraging the strengths of PIC-seq—like approaches and spatial data provides new opportunities to study interactions in data-driven fashion.

7.3 Topic modelling for scRNA-seq ordered in pseudotime

Finally, in Chapter 6 we apply an extension of the standard LDA to scRNA-seq data ordered in pseudotime. While there are a plethora of methods covering pseudotime ordering, challenges remain in the steps that follow. It might be argued that pseudotime as an approach is not here to stay as lab-based techniques have been developed to facilitate the exploration of temporal processes, for example metabolic labelling. However, those lab-based techniques are still not widely used and pseudotime is still a common step in single cell analysis. To alleviate some of the challenges, with DCTM we take into account both time and correlation. We demonstrate the flexibility of the model by applying it to distinct biological scenarios. Furthermore, DCTM outperforms both relaxed and standard LDA as it can better detect signal from noise and such uncovers further biological insight.

In its current state, the model has several limitations, such as only being able to model linear or circular trajectories. A potential extension of DCTM can take into account branching of pseudotemporal trajectories. There are several possibilities for doing this. One way is to implement a branching Gaussian kernel; this has been done in scRNA-seq to identify when a gene branches in pseudotime. The current implementation models a branching event as an intersection of three latent functions (Boukouvalas et al. 2018). Another possibility is modelling the branching that corresponds to two cell types at the same pseudotime as mixtures. The approach described in Chapter 6 relies on data that has been already ordered in pseudotime. However, considering the uncertainty in pseudotime, implementing a joint model might be another useful extension. It is worth noting that such a model might be very complex and not scale well.

While Chapter 6 explores gene dynamics in one dataset, there are interesting applications that would benefit from joint analysis of multiple datasets: what groups of genes share the same temporal patterns across datasets? For example, if orthologs are considered across species: which orthologs behave the same way over time? Which orthologs follow different temporal patterns? Can we learn more about gene regulation and dynamics by considering genes that are not orthologs but cluster with the orthologs? Assuming multiple datasets are ordered in time, the aim is to identify groups of genes that cluster together across different datasets. Instead of clustering each dataset independently and then trying to map results across datasets, datasets are modelled jointly and the allocation of genes to clusters in one dataset affects the cluster allocation in another. Similar methods have been proposed previously by (Rogers et al. 2008) and (Kirk et al. 2012). Both methods have been used for simultaneous clustering of datasets across multiple modalities. The approach proposed by (Kirk et al. 2012) allows for time dynamics to be modelled by GPs.

7.4 Summary

The complexity of scRNA-seq data makes it an excellent avenue for applying and developing machine learning algorithms. This thesis described and applied three models based on topic modelling to scRNA-seq data. In conclusion, topic modelling is an interpretable approach for analysing scRNA-seq data. In this context, cells correspond to documents, genes to words, and the latent topics are co-varying groups of genes. The identified topics correspond to cell type-specific and general biological processes. This thesis has shown the ability of topic modelling to detect genes that change as a result of interaction in Chapter 5. Additionally, where appropriate assumptions of previous methods were relaxed, for example the 2-step LDA procedure in Chapter 5 does not require prior clustering or generation of synthetic doublets. While the approach described in Chapter 4 does not outperform existing methods, Chapter 4 outlines issues with doublet detection and potential future avenues for exploration. Finally, adding a temporal dimension to the topic modelling analysis of scRNA-seq adds further biological insight to the analysis as shown in Chapter 6. The development and application of interpretable models is of vital importance in a field where ground truth is limited. This thesis demonstrated the potential of such models while also relaxing assumptions of previously described methods, making them easier to use in practice and adding prior information that enhances biological signal where appropriate.

Appendix A

Investigating the potential of Latent Dirichlet Allocation for doublet detection

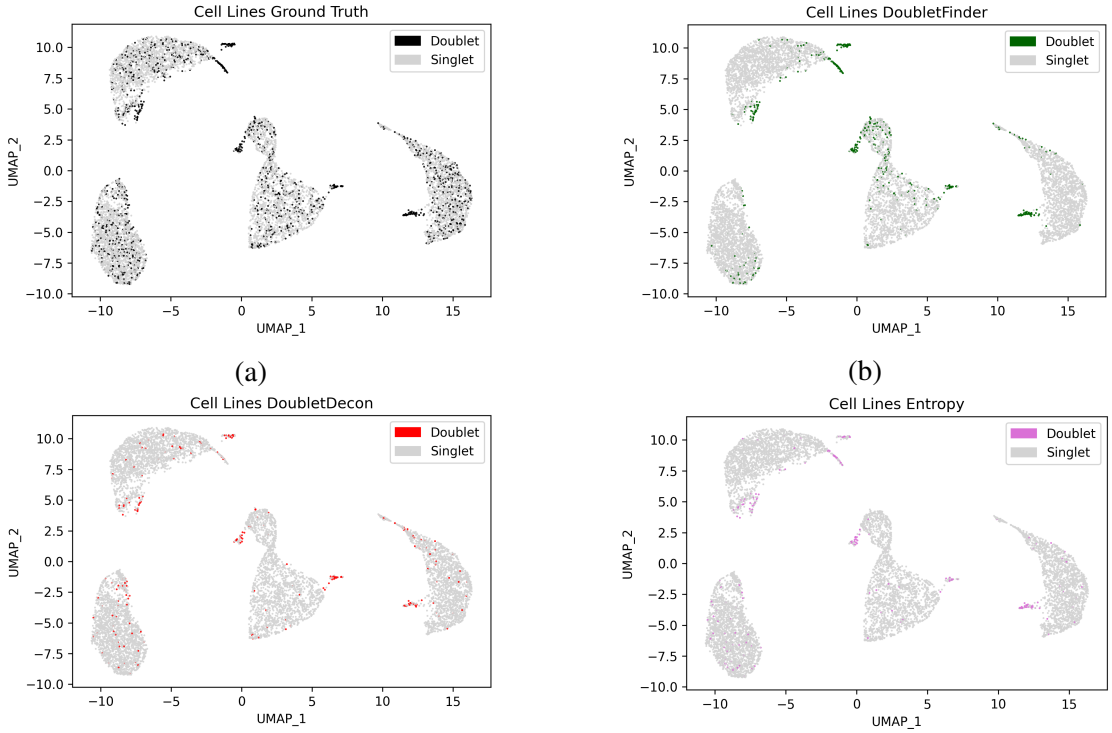


Figure A.1: Ground truth (a) and doublets identified correctly by each method: (b) DoubletFinder, (c) DoubletDecon, and (d) LDA with entropy scoring.

Appendix B

Understanding cellular crosstalk in scRNA-seq using topic modelling

Sample Name	features lower cutoff	features upper cutoff	% mt
C51	500	3000	25
C52	500	2000	20
C100	500	6000	25
C141	500	7000	25
C142	500	6500	25
C144	500	4000	25
C143	500	6000	25
C145	500	4500	20
C146	500	4500	25
C148	500	7500	25
C149	500	7000	25
C152	500	6000	25

Table B.1: Filtering parameters used for each sample in the COVID-19 dataset. Upper cutoffs for nFeatures has been set to relatively high values as we are interested in potential doublets. The percentage of mitochondrial genes (% mt) cutoff allows us to exclude dying cells.

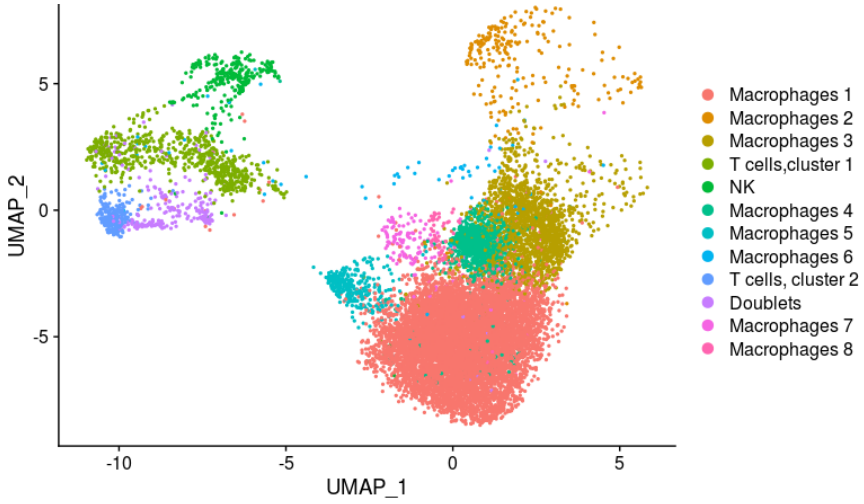


Figure B.1: 145 cluster annotation: COVID-19 BALF, UMAP projection of patient sample C145. We have identified the cluster containing doublets based on expression of marker genes and annotation by DoubletFinder.

Topic ID	Genes	Notes
7	mt-Rnr2, mt-Co1, mt-Rnr1, mt-Nd5, mt-Nd1	appearing in over 1000 cells
8	H2-Aa,H2-Ab1, H2-Eb1, Fth1	appearing in over 800 cells
9	B2m, Eef1a1, Snord35nm Fth1	

Table B.2: We perform stage 2 by fixing these 5 topics and then fitting the second LDA on the interacting DCs and T-cells. We observe genes related to housekeeping and mitochondrial processes. These processes also exist in the reference population but, due to the low number of topics that we specified initially, it seems that they are only picked up in the second stage. Indeed, when 10 topics are used for the first stage, these genes appear at that stage (see Section 5.3.2 and Table B.3)

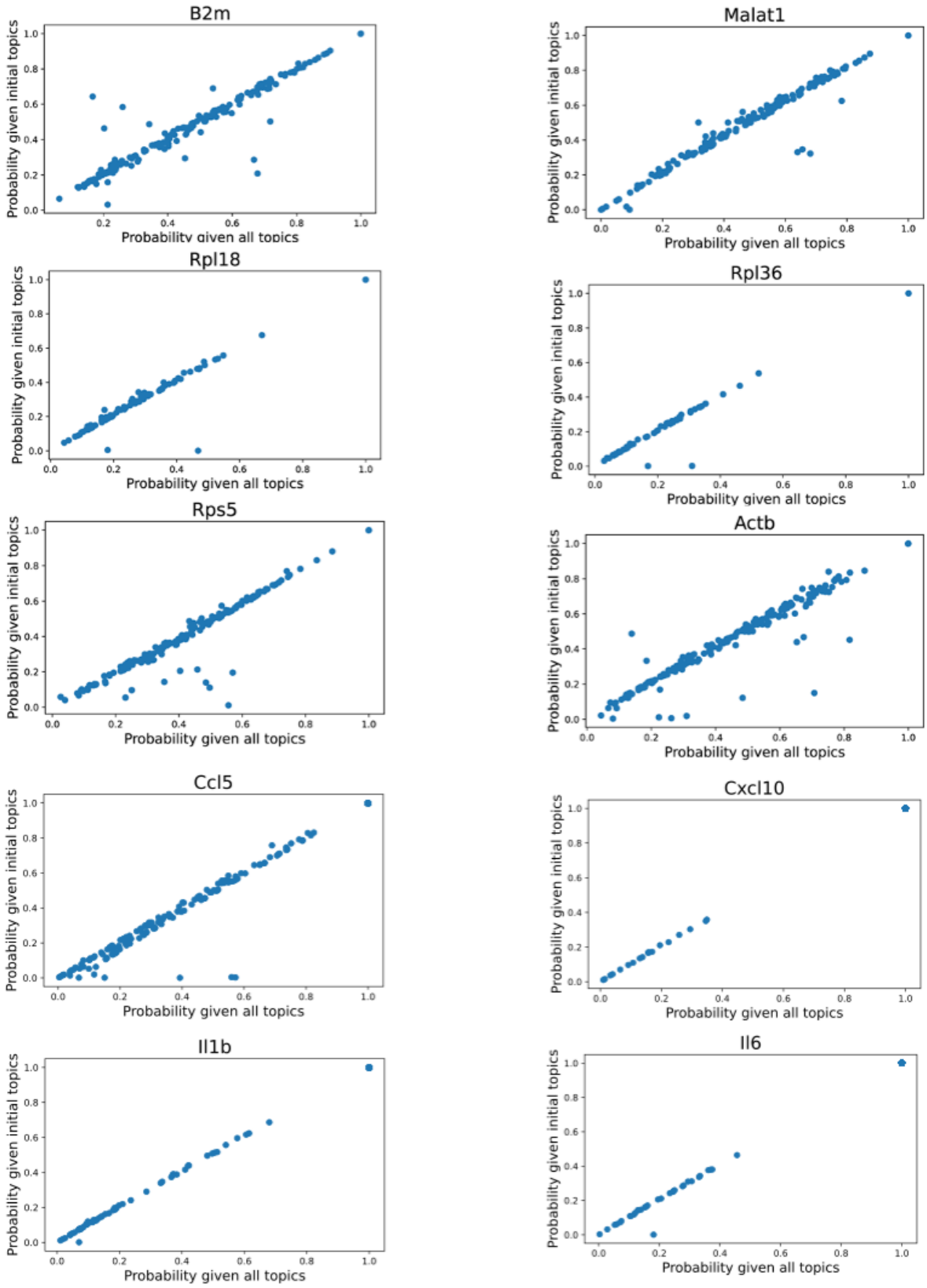


Figure B.2: Examples of genes for which we have not modified the expression. As expected, the probability of observing their counts is similar under the two models.

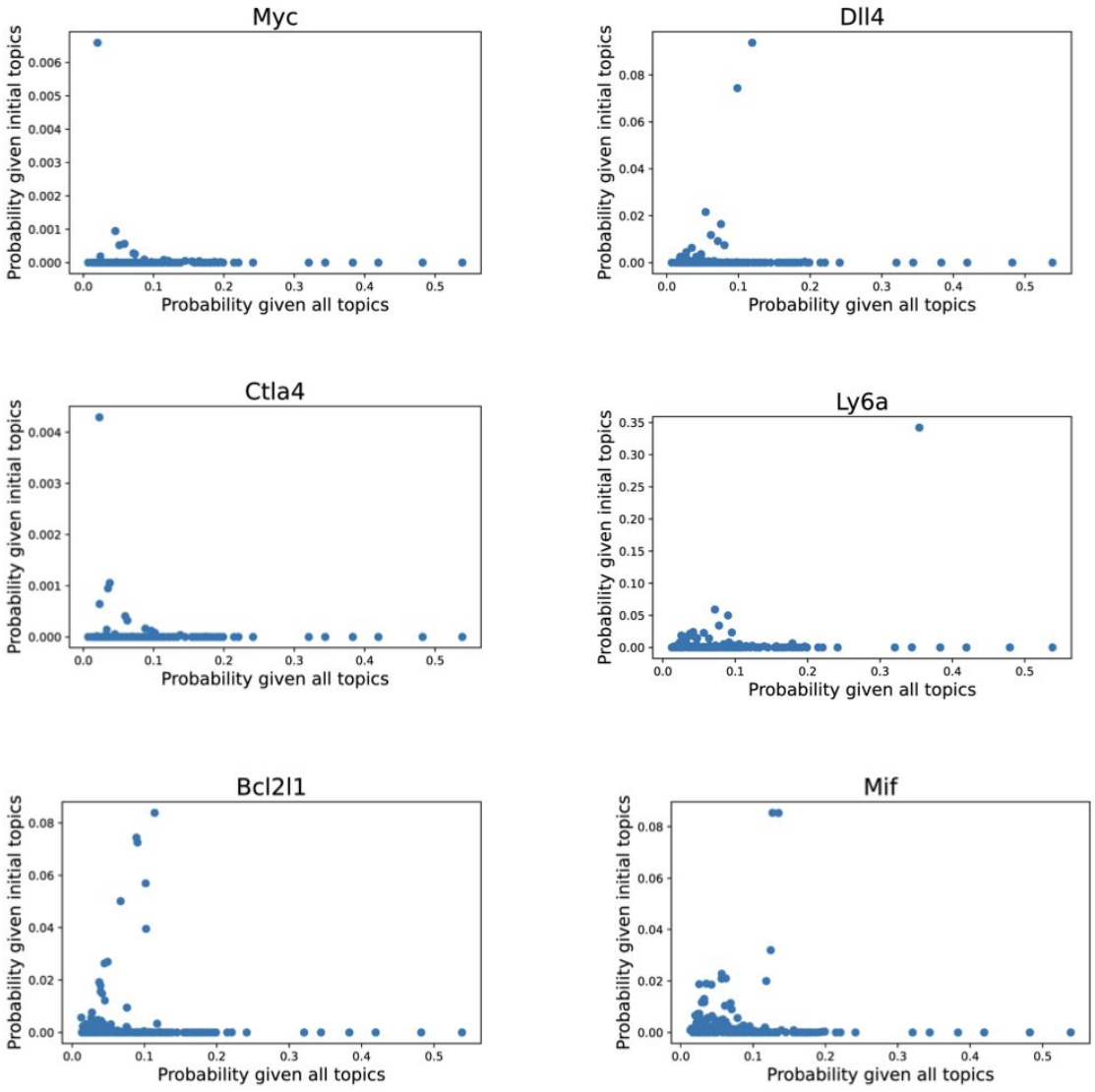


Figure B.3: Examples of genes for which we have modified the expression. Their counts are observed with higher probability under the more complex model with new topics.

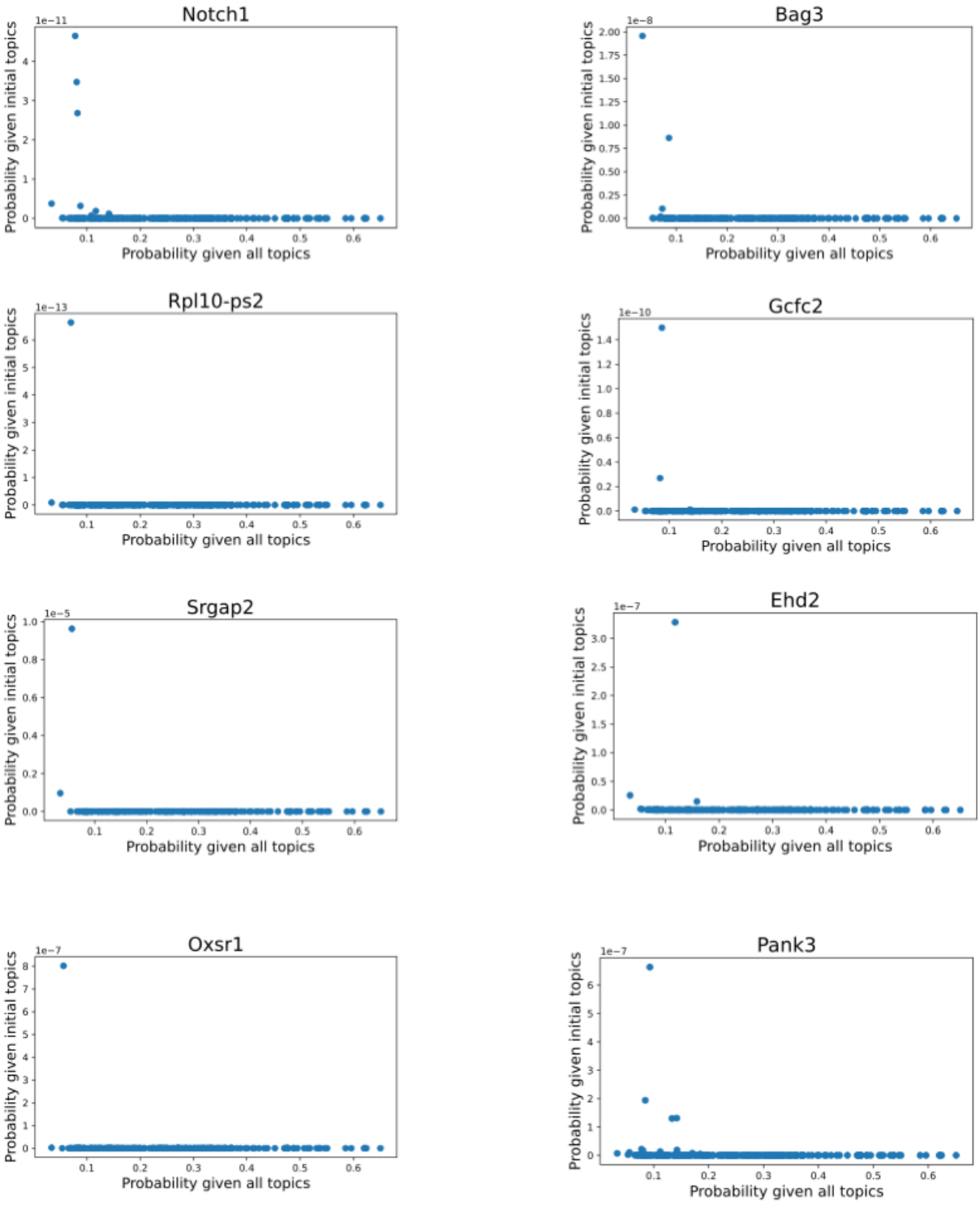


Figure B.4: Additional synthetic experiment with a different set of randomly sampled genes modified

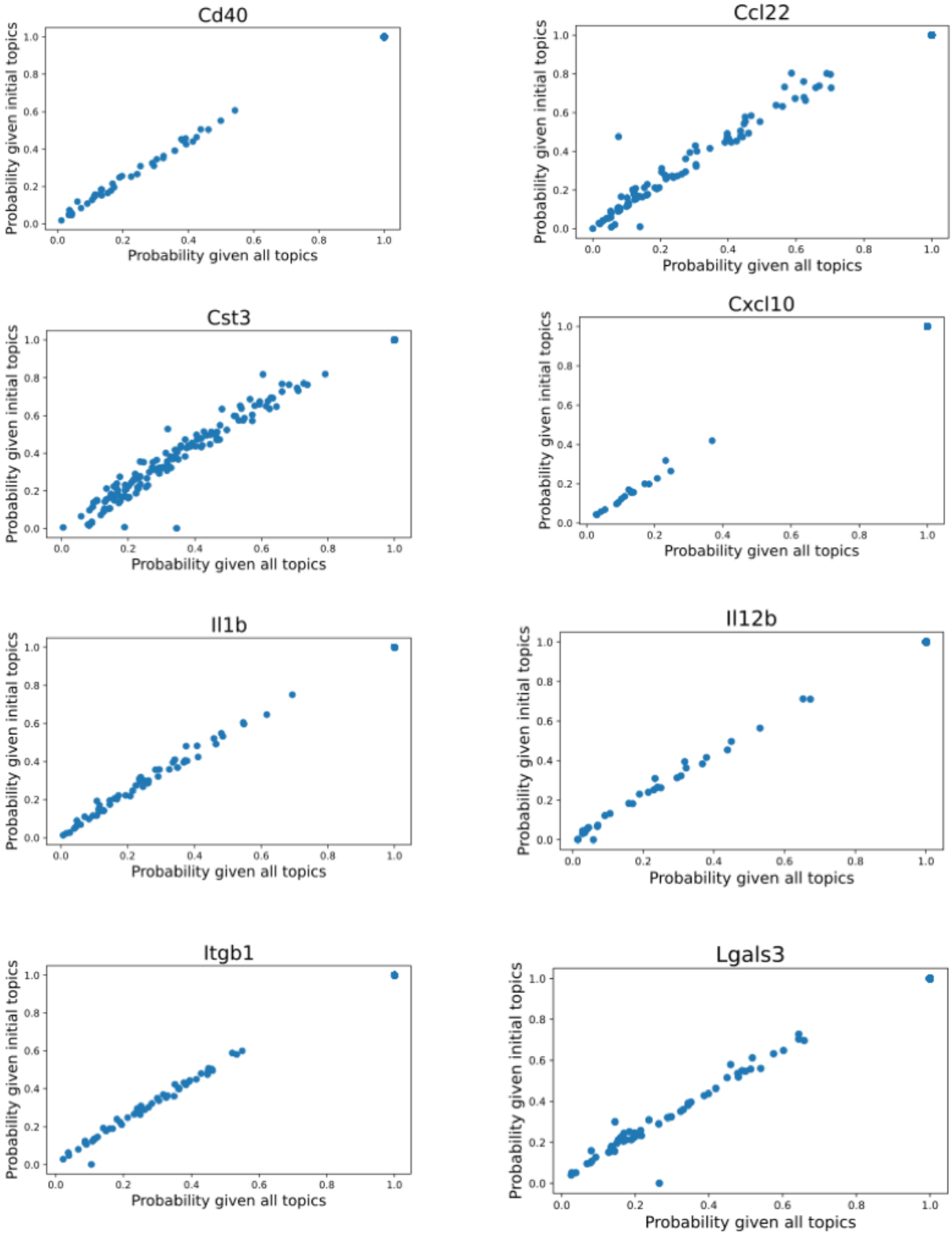


Figure B.5: Additional synthetic experiment showing genes with unmodified expression. Similar probabilities can be observed under the two models.

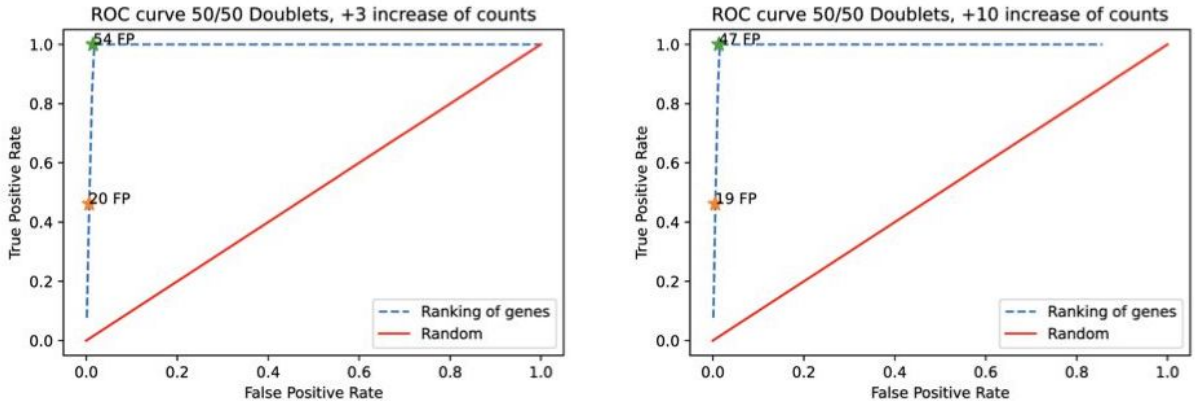


Figure B.6: ROC curves for 50/50 doublets if the increase of counts for a set of genes is 3 or 10 respectively.

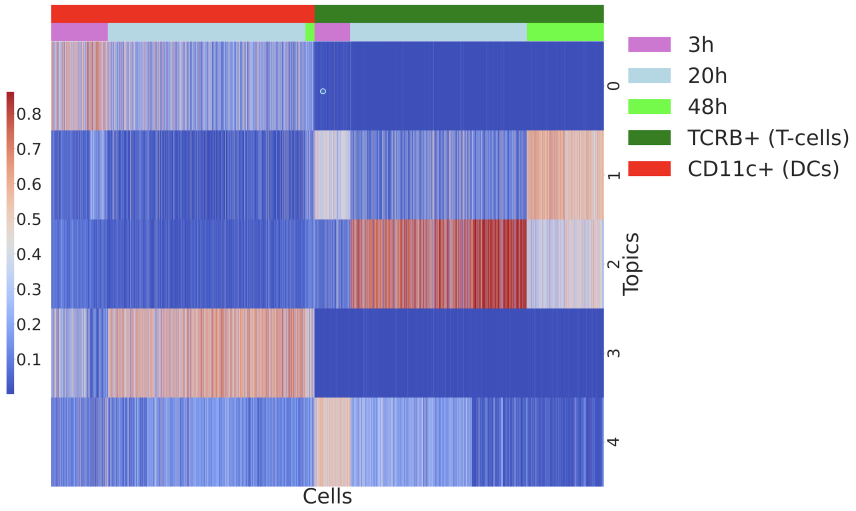


Figure B.7: 1st stage LDA with 5 topics, capturing topics specific to T-cells and DCs during the different timepoints. No topic is shared across all cells. We fit a model with 5 topics on the initial reference population, co-culture of DCs and co-culture of T-cells. As can be seen from the figure, we are capturing topics that are unique to DCs and T-cells. For example, topic 2, seems to be expressed in T-cells at 20h, while topic 3 is expressed in DCs at 20h. However, we would expect at least some genes to be expressed across both T-cells and DCs, for example housekeeping or mitochondrial ones, and these do not appear to be represented by any topics. It seems possible therefore that one result of under-specifying is that some processes that ought to be captured in stage 1 are actually captured in stage 2.

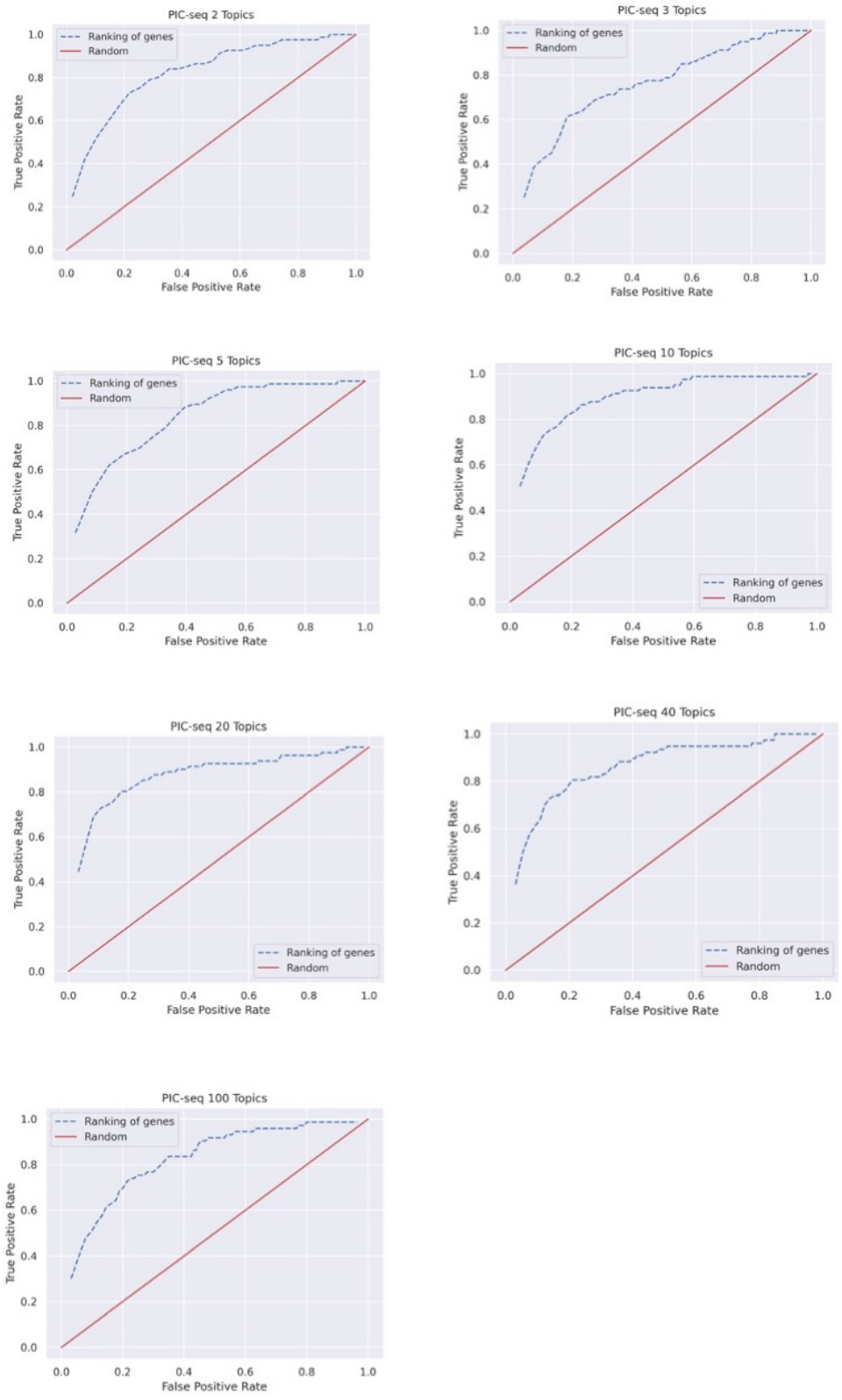


Figure B.8: ROC curves for a range of topics. Ground truth is considered the genes identified by Giladi et al (Giladi et al. 2020). Following 10 topics, which can be considered the optimum for this dataset, the performance starts to drop and the ROC curves for 40 and 100 topics show decay

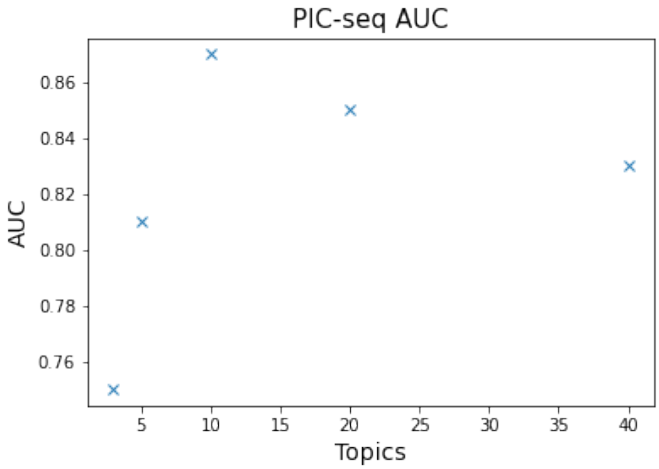


Figure B.9: Area under the curve (AUC) is high for the optimal value of topics (10) and decays slowly after.

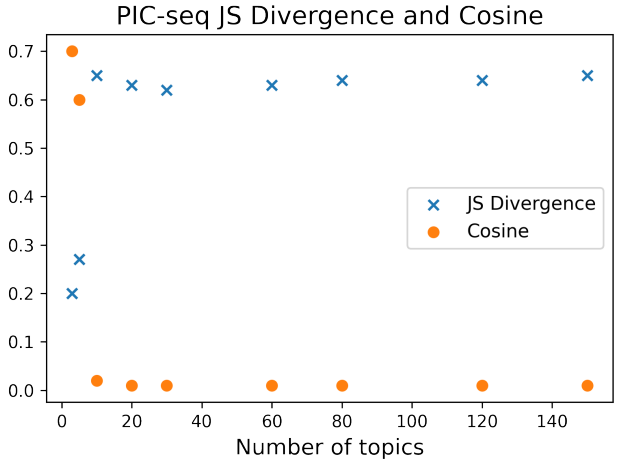


Figure B.10: Jensen-Shannon divergence and cosine follow a similar pattern to the perplexity. JS increases from 10 (higher is better) and cosine decreases from 10 (lower is better).

Topic ID	Genes	Notes
0	Fscn1, Calm1, Tmem123, Cd74, Malat1, Ftl1	Generally high expression in DCs, particularly 20h and 48h
1	Cst3, Ccl5, Cd74	DC specific genes, higher in 3h and some 20
2	mt-Rnr2, mt-Rnr1, mt-Cytb, mt-Nd4, mt-Nd1	Mitochondrial genes; not specific to a cell type topic
3	Igkc, Ighm, Igha, Gm42418, Gm26917, Jchain, Iglj1, B2m	Not specific to a cell type
4	Cdkna1a, Hspa5, Nfkbia, Ubc, Esd, Nr4a3	Similar expression across but some slightly higher in some T-cells
5	Snord32a, Ldha, Npm1, Ly6a, Eef2, Eef1a1, Trac	High in 48h T-cells
6	Ly6e, Trac, Stat1, Cd52, Gbp2	High in 3h T-cells
7	Gm42418, Gm26917, Hsp90ab1, Actb, Calr, Lars2, Myh9	Similar expression but slightly lower in some T-cells
8	Npm1, Ncl, Ldha, Ddx21, Nop58, Ccnd2	High in 20h T-cells
9	AC117232.5, Snord32a, Snord15b, Gm15710, Gm9794, Gm13456, Snord55, Rack1, Mir3091	High in 20h T-cells subset

Table B.3: Genes with high probabilities appearing in the topics identified in the reference

Boisset	'Lrg1', 'Ube2s', 'Ly6c2', 'Fcer1g', 'Cpox', 'Ctss', 'Il16', 'Hspe1', 'Lig1', 'Gmfg', 'Gm10845', 'Acta1', 'Cdk1', 'Klf6', 'Gstm1', 'Ear1', 'Tpm1', 'Cyp4f18', 'Ltb4r1', 'Hp', 'Mxd1', 'Slbp', 'Ubac1', 'Plk1', 'Itgam', 'Tmed10', 'Myl1', 'Eif5b', 'B630005N14Rik', 'Sertm1', 'Itga2b', 'Selplg', 'Snca', 'Hmox1', 'Slc25a4', 'Pf4', 'Tomm7', 'Gnai3', 'Tmsb10', 'Ngp', 'A130077B15Rik', 'Alox15', 'Mpeg1', 'Tceb2', 'Anxa1', 'Srgn', 'Eno3', 'Ptprd', 'Lsp1', 'Blvrb', 'Tnnt3', 'Ear6', 'Plek', 'Slc4a1', 'Gapt', 'Gp5', 'Gm12504', 'Smc2', 'Lpl', 'Zfp71-rs1', 'Camp', 'Mmp8', 'Atp5k', 'Rrm2', 'Phb2', 'Actn3', 'Ckm', 'Mrc1', 'Mylpf', 'Birc5', 'Igj', 'Cebpe', 'Slx1b', 'Nrgn', 'Tusc1', 'Coro1a', 'Lgals1', 'Ifitm6', 'Lasp1', 'Tuba1c', 'Sec61b', 'Ctsh', 'Tmem14c', 'Rgcc', 'Cask', 'S100a11', 'Banf1', 'S100a6', 'Ube2c', 'Clec5a', 'Myh4', 'Adpgk', 'Cd79b', 'Mkrn1', 'H2afx', 'Pdia6', 'Mmp9', 'Epx', 'C3', 'Cd63', 'Dusp22', 'H2-K1', 'Ssr2', 'Slc40a1', 'Shfm1', 'Lars2', 'Mpo', 'Npm1', 'Ppp1r15a', 'Mki67', 'Marcks', 'Hnr-pdl', '2810417H13Rik', 'Atp6ap2', 'Fech', 'Prdx5', 'Lmnb1', 'Prtn3', 'Cd164', 'Ear2', 'Erp29', 'Casq1', 'Pnp', 'A630089N07Rik', 'Cxcl12', 'Igf2bp2', 'Clta', 'Alox12', 'Cd177', 'Alas2', 'Ndufs3', 'Car1', 'Parvb', 'Pvalb', 'Ckap4', 'Cdca8', 'Itgb3', 'Gm20594', 'Cdca3', 'Cmtm7', 'Syne1', 'Prpf19', 'Aqp1', 'Abcb10', 'Minpp1', 'Ahdc1', 'Msn', 'Rhd', 'Myeov2', 'Epb4.1', 'Comt', 'Gp1bb', 'Snrpb2', 'Ptgfrn', 'Uhrf1', 'Nol7', 'Gas5', 'Cdkn3', 'Fam101b', 'Zc3hav1l', 'Gm6525', 'Gna11', 'Retnlg', 'Plac8', 'Cpne3', 'Sepp1', 'Car2', 'Eef1g', 'Alkbh5', 'Ccl6', 'Ctse', 'Mtus1', 'Serpinb1a', 'Tnni2', 'Cd9', 'Prg2', 'Fcna', 'Ms4a3', 'Cdkn2d', 'Prss57', 'Zyx', 'Vcam1', 'Snrpf', 'Neb', 'Fam132a', 'Grn', 'Ccnb1', 'Tmed9', 'Tsc22d1', 'C1qc', 'Hmgn5', 'Smc4', 'Mfsd10', 'Hbb-b1', 'Bhlhe41', 'Eif3g', 'Prdx2', 'C1qa', 'Mybpc2', 'Hmgcr', 'Slc25a37', 'S100a9', 'Nfia', 'Gyg', 'Gypa', 'Mtdh', 'Cd52', 'Fpr2', 'Mcm7', 'Ube2l6', 'Clu', 'Hdc', 'Thbs1', 'Glul', 'Mgst2', 'Beta-s', 'Prc1', 'Isca1', 'Igsf6', 'Slpi', 'H2afy', 'Cd53', 'Prdx1', 'Pgk1', 'Cxcr2', 'Eef1b2', 'Rrm1', 'Clec12a', 'Nfkbia', 'Prg3', 'Lcn2', 'Tagln2', 'Elane', 'Ctsg', 'Ttn', 'Lgals3', 'Hbb-b2', 'Ogfrl1', 'Pygm', 'Pglyrp1', 'Myl9', '5830416I19Rik', 'Stmn1', 'Alad', 'Rgs2', 'Lta4h', 'Hba-a1', 'Kif11', 'Tpm4', 'Map1a', 'Wnt4', 'Impdh2', 'Ccnb2', 'Atpif1', 'Anxa3', 'Eif3e', 'Apoe', 'Isg20', 'Atp5g1', 'Hmbs', 'Sdpr', 'Chi3l3', 'Ccna2', 'Tnnc2', 'Psma5', 'Ltf', 'Alas1', 'Fcnb', 'S100a8', 'Rcc2', 'Csf3r', 'Rorb', 'Msrb1', 'Lyz2', 'Ctla2a', 'Lyz1', 'Gda', 'Nkg7', 'F5', '1100001G20Rik', 'Xbp1', 'Snrpd1', 'Fam46a', 'Arsb', 'C1qb', 'Des', 'Anxa2', 'Crip1', 'Gm17821', 'Grina', 'Ncf1', 'Mtmr3', 'Ppp1cb', 'Ncl', 'Glrx5', 'Grk4', 'Ptchd1', 'Ppbp', 'Igfbp4', 'Axl', 'Atp2a1', 'Pcna', 'Cst7', 'Kcna2', 'Tmpo'
PIC-seq	Itm2b Creg1 4930542C12Rik Gm27390 Gm10288 Ptma Pcna Dut Mcm4 Mcm3 Mcm6 Tubb5 Lgals1 Gzmb Vim Cd52 Cxcr6 Tcf7 Il12b Cst3 Il6 Ccr7 Id2 Irf8 Ccl5 Psmb1 Dusp5 Npc2 Gnas Ccl22 Tnfrsf4 Hopx Tnfrsf8 Tnfrsf18 Ifi2712a Il2ra Cd74 Lgals3 Ldha Il2ra Cd2 Il2rb Cd69 Dusp2 Tnf Cxcl10 Ifit1 Ifih1 Isg15 Ifi209 S100a4 Il4il Cd53 Gbp2 Gm11263

Table B.4: Top genes from the PIC-seq and the bone marrow datasets.

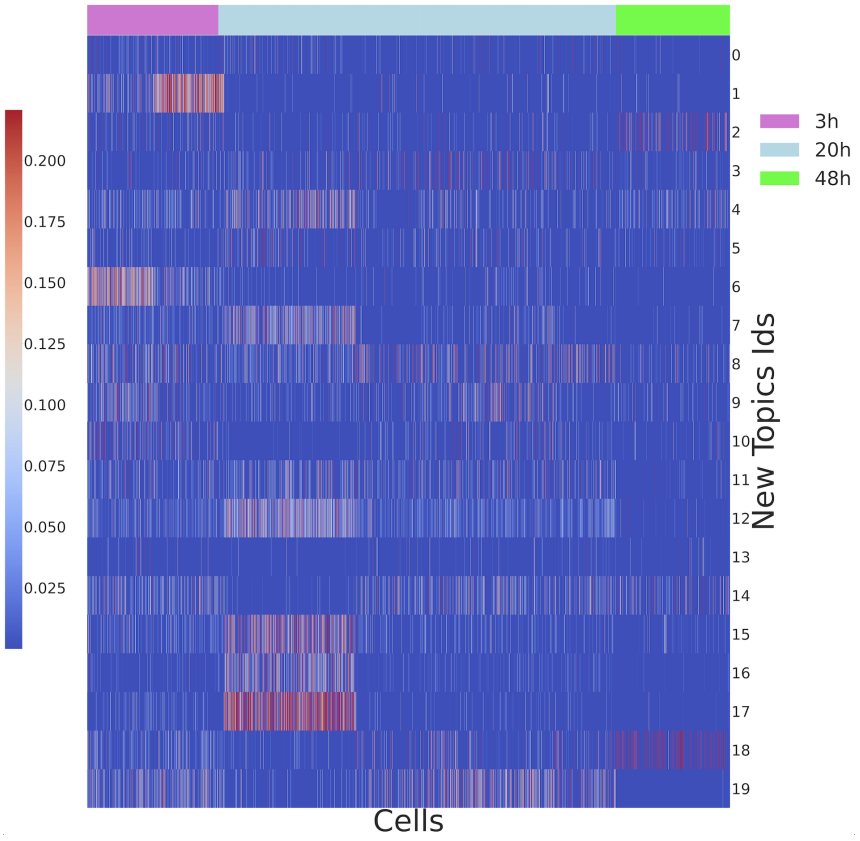


Figure B.11: New topics and PICs heatmap. While some topics have higher probabilities associated with a timepoint, other topics appear expressed across all timepoints, so the gene expression shift is linked to the cell types that contribute to the PICs.

Appendix C

Using dynamic topic modelling to study temporal scRNA-seq data

Method	GO terms
DCTM	GO:0071976 GO:0007017 GO:0044310 GO:0031122 GO:0031225 GO:0017176 GO:0003714 GO:0071704 GO:0016747 GO:0051082 GO:0030145 GO:0004579 GO:0006221 GO:0005315 GO:0042393 GO:0016209 GO:0000339 GO:0004392 GO:0000350 GO:0006457 GO:0046654 GO:0006094 GO:0042823 GO:0016462 GO:0009116 GO:0032958 GO:0032266 GO:0018024 GO:0045048 GO:0035556 GO:0009190 GO:0000381 GO:0000796 GO:0006415 GO:0032508 GO:0042765 GO:0006379 GO:0006378 GO:0006338 GO:0005847 GO:0043486 GO:0051603 GO:0005839 GO:0004298 GO:0004577 GO:0002161 GO:0003860 GO:0051920 GO:0006241 GO:0016972 GO:0005680 GO:0004865 GO:0043039 GO:0005665 GO:0005347 GO:0034511 GO:0007186 GO:0019236 GO:0006807 GO:0016255 GO:0070569 GO:0005869 GO:0006488 GO:0000266 GO:0022900 GO:0044311 GO:0042273 GO:0006352 GO:0070084 GO:0008235 GO:0010468 GO:0006364 GO:0047429 GO:0051015 GO:0004725 GO:0009408 GO:0006164 GO:0030014 GO:0030015 GO:0009536 GO:0030008 GO:0045039 GO:0000932
LDA	GO:0032447 GO:0005744 GO:0032543 GO:0016829 GO:0030433 GO:0020036 GO:0006812 GO:0016624 GO:0005543 GO:0140326 GO:0009405 GO:0019288 GO:0032515 GO:0042578 GO:0005471 GO:0051016 GO:0008290 GO:0031204 GO:0031207 GO:0005938 GO:0008081 GO:0046068 GO:0000956 GO:0046658 GO:0004359 GO:0003887 GO:0045273 GO:0006825 GO:0020002 GO:0005507 GO:0006779 GO:0006782 GO:0051087 GO:0006096 GO:0030544 GO:0008569 GO:0008289 GO:0006897 GO:0003774 GO:0005742 GO:0030276 GO:0006207 GO:0016627 GO:0043657 GO:0042254 GO:0042274 GO:0004553 GO:0016836 GO:0008320
Relaxed LDA	GO:0006261 GO:0007033 GO:0071949 GO:0045892 GO:0003951 GO:0031201 GO:0004402 GO:0046488 GO:0046854 GO:0048015 GO:0048500 GO:0008312 GO:0009298 GO:0070682 GO:0016668 GO:0019205 GO:0006750 GO:0051028 GO:0006298 GO:0030983 GO:0015078 GO:0046034 GO:0044538 GO:0004427 GO:0009678 GO:0000213 GO:0003684

Table C.1: GO terms unique to each topic modelling approach.

Bibliography

10X Genomics (2020).
URL: *https://kb.10xgenomics.com/hc/en-us/articles/360001378811*

Abdelaal, T., Michielsen, L., Cats, D., Hoogduin, D., Mei, H., Reinders, M. J. T. & Mahfouz, A. (2019), 'A comparison of automatic cell identification methods for single-cell RNA sequencing data', *Genome Biology* **20**(1), 194.
URL: *https://doi.org/10.1186/s13059-019-1795-z*

Ahmed, S., Rattray, M. & Boukouvalas, A. (2019), 'GrandPrix: scaling up the Bayesian GPLVM for single-cell data', *Bioinformatics* **35**(1), 47–54.
URL: *https://doi.org/10.1093/bioinformatics/bty533*

AlJanahi, A. A., Danielsen, M. & Dunbar, C. E. (2018), 'An Introduction to the Analysis of Single-Cell RNA-Sequencing Data', *Molecular Therapy - Methods & Clinical Development* **10**, 189–196. Publisher: Elsevier.
URL: *https://www.cell.com/molecular-therapy-family/methods/abstract/S2329-0501(18)30066-4*

Almanzar, N., Antony, J., Baghel, A. S., Bakerman, I., Bansal, I., Barres, B. A., Beachy, P. A., Berdnik, D., Bilen, B., Brownfield, D., Cain, C., Chan, C. K. F., Chen, M. B., Clarke, M. F., Conley, S. D., Darmanis, S., Demers, A., Demir, K., de Morree, A., Divita, T., du Bois, H., Ebadi, H., Espinoza, F. H., Fish, M., Gan, Q., George, B. M., Gillich, A., Gòmez-Sjöberg, R., Green, F., Genetiano, G., Gu, X., Gulati, G. S., Hahn, O., Haney, M. S., Hang, Y., Harris, L., He, M., Hosseinzadeh, S., Huang, A., Huang, K. C., Iram, T., Isobe, T., Ives, F., Jones, R., Kao, K. S., Karkanias, J., Karnam, G., Keller, A., Kershner, A. M., Khoury, N., Kim, S. K., Kiss, B. M., Kong, W., Krasnow, M. A., Kumar, M. E., Kuo, C. S., Lam, J., Lee, D. P., Lee, S. E., Lehallier, B., Leventhal, O., Li, G., Li, Q., Liu, L., Lo, A., Lu, W.-J., Lugo-Fagundo, M. F., Manjunath, A., May, A. P., Maynard, A., McGeever, A., McKay, M., McNerney, M. W., Merrill, B., Metzger, R. J., Mignardi, M., Min, D., Nabhan, A. N., Neff, N. F., Ng, K. M., Nguyen, P. K., Noh, J., Nusse, R., Pálovics, R., Patkar, R., Peng, W. C., Penland, L., Pisco, A. O., Pollard, K., Puccinelli, R., Qi, Z., Quake, S. R., Rando, T. A., Rulifson, E. J., Schaum, N., Segal, J. M., Sikandar, S. S., Sinha, R., Sit, R. V., Sonnenburg, J., Staehli, D., Szade,

K., Tan, M., Tan, W., Tato, C., Tellez, K., Dulgeroff, L. B. T., Travaglini, K. J., Tropini, C., Tsui, M., Waldburger, L., Wang, B. M., van Weele, L. J., Weinberg, K., Weissman, I. L., Wosczyna, M. N., Wu, S. M., Wyss-Coray, T., Xiang, J., Xue, S., Yamauchi, K. A., Yang, A. C., Yerra, L. P., Youngyunpipatkul, J., Yu, B., Zanini, F., Zardeneta, M. E., Zee, A., Zhao, C., Zhang, F., Zhang, H., Zhang, M. J., Zhou, L., Zou, J. & The Tabula Muris Consortium (2020), 'A single-cell transcriptomic atlas characterizes ageing tissues in the mouse', *Nature* **583**(7817), 590–595. Number: 7817 Publisher: Nature Publishing Group.
URL: *https://www.nature.com/articles/s41586-020-2496-1*

Alpert, A., Moore, L. S., Dubovik, T. & Shen-Orr, S. S. (2018), 'Alignment of single-cell trajectories to compare cellular expression dynamics', *Nature Methods* **15**(4), 267–270. Number: 4 Publisher: Nature Publishing Group.
URL: *https://www.nature.com/articles/nmeth.4628*

Andersson, A., Bergenstråhle, J., Asp, M., Bergenstråhle, L., Jurek, A., Navarro, J. F. & Lundeberg, J. (2019), 'Spatial mapping of cell types by integration of transcriptomics data', *bioRxiv* p. 2019.12.13.874495. Publisher: Cold Spring Harbor Laboratory Section: New Results.
URL: *https://www.biorxiv.org/content/10.1101/2019.12.13.874495v1*

Andrews, T. S. & Hemberg, M. (2019), 'False signals induced by single-cell imputation', *F1000Research* **7**.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6415334/*

Arjmand, B., Hamidpour, S. K., Tayanloo-Beik, A., Goodarzi, P., Aghayan, H. R., Adibi, H. & Larijani, B. (2022), 'Machine Learning: A New Prospect in Multi-Omics Data Analysis of Cancer', *Frontiers in Genetics* **13**.
URL: *https://www.frontiersin.org/article/10.3389/fgene.2022.824451*

Bais, A. S. & Kostka, D. (2020), 'scds: computational annotation of doublets in single-cell RNA sequencing data', *Bioinformatics* **36**(4), 1150–1158.
URL: *https://doi.org/10.1093/bioinformatics/btz698*

Baran-Gale, J., Chandra, T. & Kirschner, K. (2018), 'Experimental design for single-cell RNA sequencing', *Briefings in Functional Genomics* **17**(4), 233–239.
URL: *https://doi.org/10.1093/bfgp/elx035*

Barshan, E., Ghodsi, A., Azimifar, Z. & Zolghadri Jahromi, M. (2011), 'Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds', *Pattern Recognition* **44**(7), 1357–1371.
URL: *https://www.sciencedirect.com/science/article/pii/S0031320310005819*

Bauer, M., van der Wilk, M. & Rasmussen, C. E. (2016), Understanding Probabilistic Sparse Gaussian Process Approximations, *in* 'Advances in Neural Information Processing Systems',

Vol. 29, Curran Associates, Inc.
URL: *https://proceedings.neurips.cc/paper/2016/hash/7250eb93b3c18cc9daa29cf58af7a004-Abstract.html*

Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. (2020), 'Generalizing RNA velocity to transient cell states through dynamical modeling', *Nature Biotechnology* **38**(12), 1408–1414. Number: 12 Publisher: Nature Publishing Group.
URL: *https://www.nature.com/articles/s41587-020-0591-3*

Bernstein, N. J., Fong, N. L., Lam, I., Roy, M. A., Hendrickson, D. G. & Kelley, D. R. (2020), 'Solo: Doublet Identification in Single-Cell RNA-Seq via Semi-Supervised Deep Learning', *Cell Systems* **11**(1), 95–101.e5.
URL: *https://www.sciencedirect.com/science/article/pii/S2405471220301952*

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003), 'Latent Dirichlet Allocation', p. 30.

Boisset, J.-C., Vivié, J., Grün, D., Muraro, M. J., Lyubimova, A. & van Oudenaarden, A. (2018), 'Mapping the physical network of cellular interactions', *Nature Methods* **15**(7), 547–553. Number: 7 Publisher: Nature Publishing Group.
URL: *https://www.nature.com/articles/s41592-018-0009-z*

Boukouvalas, A., Hensman, J. & Rattray, M. (2018), 'BGP: identifying gene-specific branching dynamics from single-cell data with a branching Gaussian process', *Genome Biology* **19**(1), 65.
URL: *https://doi.org/10.1186/s13059-018-1440-2*

Bravo González-Blas, C., Minnoye, L., Papasokrati, D., Aibar, S., Hulselmans, G., Christiaens, V., Davie, K., Wouters, J. & Aerts, S. (2019), 'cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data', *Nature Methods* **16**(5), 397–400. Number: 5 Publisher: Nature Publishing Group.
URL: *https://www.nature.com/articles/s41592-019-0367-1*

Browaeys, R., Saelens, W. & Saeys, Y. (2020), 'NicheNet: modeling intercellular communication by linking ligands to target genes', *Nature Methods* **17**(2), 159–162. Number: 2 Publisher: Nature Publishing Group.
URL: *https://www.nature.com/articles/s41592-019-0667-5*

Brower, K. K., Khariton, M., Suzuki, P. H., Still, C., Kim, G., Calhoun, S. G. K., Qi, L. S., Wang, B. & Fordyce, P. M. (2020), 'Double Emulsion Picoreactors for High-Throughput Single-Cell Encapsulation and Phenotyping via FACS', *Analytical Chemistry* **92**(19), 13262–13270. Publisher: American Chemical Society.
URL: *https://doi.org/10.1021/acs.analchem.0c02499*

Buenrostro, J., Wu, B., Chang, H. & Greenleaf, W. (2015), 'ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide', *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]* **109**, 21.29.1–21.29.9.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4374986/*

Buettner, F., Natarajan, K. N., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J., Teichmann, S. A., Marioni, J. C. & Stegle, O. (2015), 'Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells', *Nature Biotechnology* **33**(2), 155–160. Number: 2 Publisher: Nature Publishing Group.
URL: *https://www.nature.com/articles/nbt.3102*

Buettner, F. & Theis, F. J. (2012), 'A novel approach for resolving differences in single-cell gene expression patterns from zygote to blastocyst', *Bioinformatics* **28**(18), i626–i632.
URL: *https://doi.org/10.1093/bioinformatics/bts385*

Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. (2018), 'Integrating single-cell transcriptomic data across different conditions, technologies, and species', *Nature Biotechnology* **36**(5), 411–420. Number: 5 Publisher: Nature Publishing Group.
URL: *https://www.nature.com/articles/nbt.4096*

Cabello-Aguilar, S., Alame, M., Kon-Sun-Tack, F., Fau, C., Lacroix, M. & Colinge, J. (2020), 'SingleCellSignalR: inference of intercellular networks from single-cell transcriptomics', *Nucleic Acids Research* **48**(10), e55–e55. Publisher: Oxford Academic.
URL: *https://academic.oup.com/nar/article/48/10/e55/5810485*

Caldelari, R., Dogga, S., Schmid, M. W., Franke-Fayard, B., Janse, C. J., Soldati-Favre, D. & Heussler, V. (2019), 'Transcriptome analysis of Plasmodium berghei during exo-erythrocytic development', *Malaria Journal* **18**(1), 330.
URL: *https://doi.org/10.1186/s12936-019-2968-7*

Campbell, J., Corbett, S., Koga, Y., Yang, S., Reed, E. & Wang, Z. (2020), 'celda: CEllular Latent Dirichlet Allocation'.
URL: *https://bioconductor.org/packages/celda/*

Cang, Z. & Nie, Q. (2020), 'Inferring spatial and signaling relationships between cells from single cell transcriptomic data', *Nature Communications* **11**(1), 2084. Number: 1 Publisher: Nature Publishing Group.
URL: *https://www.nature.com/articles/s41467-020-15968-5*

Cao, J., Xia, T., Li, J., Zhang, Y. & Tang, S. (2009), 'A density-based method for adaptive LDA model selection', *Neurocomputing* **72**(7), 1775–1781.
URL: *https://www.sciencedirect.com/science/article/pii/S092523120800372X*

Chasis, J. A. & Mohandas, N. (2008), 'Erythroblastic islands: niches for erythropoiesis', *Blood* **112**(3), 470–478. Publisher: American Society of Hematology.
URL: *https://ashpublications.org/blood/article/112/3/470/25296/Erythroblastic-islands-niches-for-erythropoiesis*

Choi, K., Chen, Y., Skelly, D. A. & Churchill, G. A. (2020), 'Bayesian model selection reveals biological origins of zero inflation in single-cell transcriptomics', *Genome Biology* **21**(1), 183.
URL: *https://doi.org/10.1186/s13059-020-02103-2*

Chung, N. C. & Storey, J. D. (2015), 'Statistical significance of variables driving systematic variation in high-dimensional data', *Bioinformatics* **31**(4), 545–554.
URL: *https://doi.org/10.1093/bioinformatics/btu674*

Cohen, M., Giladi, A., Gorki, A.-D., Solodkin, D. G., Zada, M., Hladik, A., Miklosi, A., Salame, T.-M., Halpern, K. B., David, E., Itzkovitz, S., Harkany, T., Knapp, S. & Amit, I. (2018), 'Lung Single-Cell Signaling Interaction Map Reveals Basophil Role in Macrophage Imprinting', *Cell* **175**(4), 1031–1044.e18. Publisher: Elsevier.
URL: *https://www.cell.com/cell/abstract/S0092-8674(18)31181-4*

Cole, M. B., Risso, D., Wagner, A., DeTomaso, D., Ngai, J., Purdom, E., Dudoit, S. & Yosef, N. (2019), 'Performance Assessment and Selection of Normalization Procedures for Single-Cell RNA-Seq', *Cell Systems* **8**(4), 315–328.e8. Publisher: Elsevier.
URL: *https://www.cell.com/cell-systems/abstract/S2405-4712(19)30080-8*

Cunin, P., Bouslama, R., Machlus, K. R., Martínez-Bonet, M., Lee, P. Y., Wactor, A., Nelson-Maney, N., Morris, A., Guo, L., Weyrich, A., Sola-Visner, M., Boilard, E., Italiano, J. E. & Nigrovic, P. A. (2019), 'Megakaryocyte emperipolesis mediates membrane transfer from intracytoplasmic neutrophils to platelets', *eLife* **8**.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6494422/*

DePasquale, E. A. K., Schnell, D. J., Van Camp, P.-J., Valiente-Alandí, , Blaxall, B. C., Grimes, H. L., Singh, H. & Salomonis, N. (2019), 'DoubletDecon: Deconvoluting Doublets from Single-Cell RNA-Sequencing Data', *Cell Reports* **29**(6), 1718–1727.e8.
URL: *http://www.sciencedirect.com/science/article/pii/S2211124719312860*

Deveaud, R., SanJuan, E. & Bellot, P. (2014), 'Accurate and effective latent concept modeling for ad hoc information retrieval', *Document numérique* **17**(1), 61–84.
URL: *http://dn.revuesonline.com/article.jsp?articleId=19419*

Dimitrov, D., Türei, D., Boys, C., Nagai, J. S., Flores, R. O. R., Kim, H., Szalai, B., Costa, I. G., Dugourd, A., Valdeolivas, A. & Saez-Rodriguez, J. (2021), Comparison of Resources and Methods to infer Cell-Cell Communication from Single-cell RNA Data, Technical report,

bioRxiv. Section: New Results Type: article.
URL: *https://www.biorxiv.org/content/10.1101/2021.05.21.445160v1*

duVerle, D. A., Yotsukura, S., Nomura, S., Aburatani, H. & Tsuda, K. (2016), 'CellTree: an R/bioconductor package to infer the hierarchical structure of cell populations from single-cell RNA-seq data', *BMC Bioinformatics* **17**(1), 363.
URL: *https://doi.org/10.1186/s12859-016-1175-6*

Eisenberg, E. & Levanon, E. Y. (2013), 'Human housekeeping genes, revisited', *Trends in Genetics* **29**(10), 569–574. Publisher: Elsevier.
URL: *https://www.cell.com/trends/genetics/abstract/S0168-9525(13)00089-9*

Elosua, M., Nieto, P., Mereu, E., Gut, I. & Heyn, H. (2020), 'SPOTlight: Seeded NMF regression to Deconvolute Spatial Transcriptomics Spots with Single-Cell Transcriptomes', *bioRxiv* p. 2020.06.03.131334. Publisher: Cold Spring Harbor Laboratory Section: New Results.
URL: *https://www.biorxiv.org/content/10.1101/2020.06.03.131334v1*

Emrich, S. J., Barbazuk, W. B., Li, L. & Schnable, P. S. (2007), 'Gene discovery and annotation using LCM-454 transcriptome sequencing', *Genome Research* **17**(1), 69–73.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1716268/*

Erhard, F., Baptista, M. A. P., Krammer, T., Hennig, T., Lange, M., Arampatzi, P., Jürges, C. S., Theis, F. J., Saliba, A.-E. & Dölken, L. (2019), 'scSLAM-seq reveals core features of transcription dynamics in single cells', *Nature* **571**(7765), 419–423. Number: 7765 Publisher: Nature Publishing Group.
URL: *https://www.nature.com/articles/s41586-019-1369-y*

Giladi, A., Cohen, M., Medaglia, C., Baran, Y., Li, B., Zada, M., Bost, P., Blecher-Gonen, R., Salame, T.-M., Mayer, J. U., David, E., Ronchese, F., Tanay, A. & Amit, I. (2020), 'Dissecting cellular crosstalk by sequencing physically interacting cells', *Nature Biotechnology* **38**(5), 629–637. Number: 5 Publisher: Nature Publishing Group.
URL: *https://www.nature.com/articles/s41587-020-0442-2*

Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. (2018), 'Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors', *Nature Biotechnology* **36**(5), 421–427. Number: 5 Publisher: Nature Publishing Group.
URL: *https://www.nature.com/articles/nbt.4091*

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zagar, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E. P., Jain, J., Srivastava, A., Stuart, T., Fleming, L. B., Yeung, B., Rogers, A. J., McElrath, J. M., Blish, C. A., Gottardo, R., Smibert, P. & Satija, R. (2020), 'Integrated analysis of multimodal single-cell

data', *bioRxiv* p. 2020.10.12.335331. Publisher: Cold Spring Harbor Laboratory Section: New Results.
URL: *https://www.biorxiv.org/content/10.1101/2020.10.12.335331v1*

Haque, A., Engel, J., Teichmann, S. A. & Lönnberg, T. (2017), 'A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications', *Genome Medicine* **9**(1), 75.
URL: *https://doi.org/10.1186/s13073-017-0467-4*

Hofmann, T. (1999), 'Probabilistic Latent Semantic Indexing', *ACM SIGIR Forum* **51**(2), 8.

Howick, V. M., Russell, A. J. C., Andrews, T., Heaton, H., Reid, A. J., Natarajan, K., Butungi, H., Metcalf, T., Verzier, L. H., Rayner, J. C., Berriman, M., Herren, J. K., Billker, O., Hemberg, M., Talman, A. M. & Lawniczak, M. K. N. (2019), 'The Malaria Cell Atlas: Single parasite transcriptomes across the complete Plasmodium life cycle', *Science (New York, N.Y.)* **365**(6455).

Ilicic, T., Kim, J. K., Kolodziejczyk, A. A., Bagger, F. O., McCarthy, D. J., Marioni, J. C. & Teichmann, S. A. (2016), 'Classification of low quality cells from single-cell RNA-seq data', *Genome Biology* **17**(1), 29.
URL: *https://doi.org/10.1186/s13059-016-0888-1*

Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A. & Amit, I. (2014), 'Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types', *Science* **343**(6172), 776–779. Publisher: American Association for the Advancement of Science Section: Report.
URL: *https://science.sciencemag.org/content/343/6172/776*

Jin, S., Guerrero-Juarez, C. F., Zhang, L., Chang, I., Ramos, R., Kuan, C.-H., Myung, P., Plikus, M. V. & Nie, Q. (2021), 'Inference and analysis of cell-cell communication using CellChat', *Nature Communications* **12**(1), 1088. Number: 1 Publisher: Nature Publishing Group.
URL: *https://www.nature.com/articles/s41467-021-21246-9*

Johnson, W. E., Li, C. & Rabinovic, A. (2007), 'Adjusting batch effects in microarray expression data using empirical Bayes methods', *Biostatistics* **8**(1), 118–127.
URL: *https://doi.org/10.1093/biostatistics/kxj037*

Kalaitzis, A. A. & Lawrence, N. D. (2011), 'A Simple Approach to Ranking Differentially Expressed Gene Expression Time Courses through Gaussian Process Regression', *BMC Bioinformatics* **12**(1), 180.
URL: *https://doi.org/10.1186/1471-2105-12-180*

Kang, H. M., Subramaniam, M., Targ, S., Nguyen, M., Maliskova, L., McCarthy, E., Wan, E., Wong, S., Byrnes, L., Lanata, C. M., Gate, R. E., Mostafavi, S., Marson, A., Zaitlen, N., Criswell, L. A. & Ye, C. J. (2018), 'Multiplexed droplet single-cell RNA-sequencing using natural genetic variation', *Nature Biotechnology* **36**(1), 89–94. Number: 1 Publisher: Nature Publishing Group.
URL: *https://www.nature.com/articles/nbt.4042*

Kim, H.-J., Yardımcı, G. G., Bonora, G., Ramani, V., Liu, J., Qiu, R., Lee, C., Hesson, J., Ware, C. B., Shendure, J., Duan, Z. & Noble, W. S. (2020), 'Capturing cell type-specific chromatin compartment patterns by applying topic modeling to single-cell Hi-C data', *PLOS Computational Biology* **16**(9), e1008173. Publisher: Public Library of Science.
URL: *https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008173*

Kirk, P., Griffin, J. E., Savage, R. S., Ghahramani, Z. & Wild, D. L. (2012), 'Bayesian correlated clustering to integrate multiple datasets', *Bioinformatics* **28**(24), 3290–3297.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3519452/*

Kleshchevnikov, V., Shmatko, A., Dann, E., Aivazidis, A., King, H. W., Li, T., Lomakin, A., Kedlian, V., Jain, M. S., Park, J. S., Ramona, L., Tuck, E., Arutyunyan, A., Vento-Tormo, R., Gerstung, M., James, L., Stegle, O. & Bayraktar, O. A. (2020), 'Comprehensive mapping of tissue cell architecture via integrated single cell and spatial transcriptomics', *bioRxiv* p. 2020.11.15.378125. Publisher: Cold Spring Harbor Laboratory Section: New Results.
URL: *https://www.biorxiv.org/content/10.1101/2020.11.15.378125v1*

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.-r. & Raychaudhuri, S. (2019), 'Fast, sensitive and accurate integration of single-cell data with Harmony', *Nature Methods* **16**(12), 1289–1296. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 12 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Computational models;Data integration;Statistical methods Subject_term_id: computational-models;data-integration;statistical-methods.
URL: *https://www.nature.com/articles/s41592-019-0619-0*

Kotliar, D., Veres, A., Nagy, M. A., Tabrizi, S., Hodis, E., Melton, D. A. & Sabeti, P. C. (2019), 'Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq', *eLife* **8**, e43803. Publisher: eLife Sciences Publications, Ltd.
URL: *https://doi.org/10.7554/eLife.43803*

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

Number: 7719 Publisher: Nature Publishing Group.
URL: *https://www.nature.com/articles/s41586-018-0414-6*

Lafferty, J. D. & Blei, D. M. (2006), 'Correlated Topic Models', p. 8.

Lareau, C. A., Ma, S., Duarte, F. M. & Buenrostro, J. D. (2020), 'Inference and effects of barcode multiplets in droplet-based single-cell assays', *Nature Communications* **11**(1), 866. Number: 1 Publisher: Nature Publishing Group.
URL: *https://www.nature.com/articles/s41467-020-14667-5*

Lawrence, N. (2005), 'Probabilistic Non-linear Principal Component Analysis with Gaussian Process Latent Variable Models', p. 34.

Lee, M., Liu, Z., Huang, R. & Tong, W. (2016), 'Application of dynamic topic models to toxicogenomics data', *BMC Bioinformatics* **17**(13), 368.
URL: *https://doi.org/10.1186/s12859-016-1225-0*

Li, R., Li, L., Xu, Y. & Yang, J. (2022), 'Machine learning meets omics: applications and perspectives', *Briefings in Bioinformatics* **23**(1), bbab460.
URL: *https://doi.org/10.1093/bib/bbab460*

Li, W. V. & Li, J. J. (2018), 'An accurate and robust imputation method scImpute for single-cell RNA-seq data', *Nature Communications* **9**(1), 997. Number: 1 Publisher: Nature Publishing Group.
URL: *https://www.nature.com/articles/s41467-018-03405-7*

Li, X., Wang, K., Lyu, Y., Pan, H., Zhang, J., Stambolian, D., Susztak, K., Reilly, M. P., Hu, G. & Li, M. (2020), 'Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis', *Nature Communications* **11**(1), 2338. Number: 1 Publisher: Nature Publishing Group.
URL: *https://www.nature.com/articles/s41467-020-15851-3*

Liao, M., Liu, Y., Yuan, J., Wen, Y., Xu, G., Zhao, J., Cheng, L., Li, J., Wang, X., Wang, F., Liu, L., Amit, I., Zhang, S. & Zhang, Z. (2020), 'Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19', *Nature Medicine* **26**(6), 842–844. Number: 6 Publisher: Nature Publishing Group.
URL: *https://www.nature.com/articles/s41591-020-0901-9*

Lin, Y., Ghazanfar, S., Strbenac, D., Wang, A., Patrick, E., Lin, D. M., Speed, T., Yang, J. Y. H. & Yang, P. (2019), 'Evaluating stably expressed genes in single cells', *GigaScience* **8**(9). Publisher: Oxford Academic.
URL: *https://academic.oup.com/gigascience/article/8/9/giz106/5570567*

Liu, L., Tang, L., Dong, W., Yao, S. & Zhou, W. (2016), 'An overview of topic modeling and its current applications in bioinformatics', *SpringerPlus* **5**(1), 1608.
URL: *https://doi.org/10.1186/s40064-016-3252-8*

Lotfollahi, M., Naghipourfar, M., Luecken, M. D., Khajavi, M., Büttner, M., Avsec, Z., Misharin, A. V. & Theis, F. J. (2020), 'Query to reference single-cell integration with transfer learning', *bioRxiv* p. 2020.07.16.205997. Publisher: Cold Spring Harbor Laboratory Section: New Results.
URL: *https://www.biorxiv.org/content/10.1101/2020.07.16.205997v1*

Luecken, M. D., Büttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Mueller, M. F., Strobl, D. C., Zappia, L., Dugas, M., Colomé-Tatché, M. & Theis, F. J. (2020), 'Benchmarking atlas-level data integration in single-cell genomics', *bioRxiv* p. 2020.05.22.111161. Publisher: Cold Spring Harbor Laboratory Section: New Results.
URL: *https://www.biorxiv.org/content/10.1101/2020.05.22.111161v1*

Luecken, M. D. & Theis, F. J. (2019), 'Current best practices in single-cell RNA-seq analysis: a tutorial', *Molecular Systems Biology* **15**(6), e8746. Publisher: John Wiley & Sons, Ltd.
URL: *https://www.embopress.org/doi/full/10.15252/msb.20188746*

Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., Vallejos, C. A., Campbell, K. R., Beerenwinkel, N., Mahfouz, A., Pinello, L., Skums, P., Stamatakis, A., Attolini, C. S.-O., Aparicio, S., Baaijens, J., Balvert, M., Barbanson, B. d., Cappuccio, A., Corleone, G., Dutilh, B. E., Florescu, M., Guryev, V., Holmer, R., Jahn, K., Lobo, T. J., Keizer, E. M., Khatri, I., Kielbasa, S. M., Korbel, J. O., Kozlov, A. M., Kuo, T.-H., Lelieveldt, B. P., Mandoiu, I. I., Marioni, J. C., Marschall, T., Mölder, F., Niknejad, A., Raczkowski, L., Reinders, M., Ridder, J. d., Saliba, A.-E., Somarakis, A., Stegle, O., Theis, F. J., Yang, H., Zelikovsky, A., McHardy, A. C., Raphael, B. J., Shah, S. P. & Schönhuth, A. (2020), 'Eleven grand challenges in single-cell data science', *Genome Biology* **21**, 31.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7007675/*

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., Trombetta, J. J., Weitz, D. A., Sanes, J. R., Shalek, A. K., Regev, A. & McCarroll, S. A. (2015), 'Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets', *Cell* **161**(5), 1202–1214. Publisher: Elsevier.
URL: *https://www.cell.com/cell/abstract/S0092-8674(15)00549-8*

Magee, C. N., Boenisch, O. & Najafian, N. (2012), 'THE ROLE OF CO-STIMULATORY MOLECULES IN DIRECTING THE FUNCTIONAL DIFFERENTIATION OF ALLO-REACTIVE T HELPER CELLS', *American journal of transplantation : official journal*

of the American Society of Transplantation and the American Society of Transplant Surgeons **12**(10), 2588–2600.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3459149/*

McDowell, I. C., Manandhar, D., Vockley, C. M., Schmid, A. K., Reddy, T. E. & Engelhardt, B. E. (2018), 'Clustering gene expression time series data using an infinite Gaussian process mixture model', *PLOS Computational Biology* **14**(1), e1005896. Publisher: Public Library of Science.
URL: *https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005896*

McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. (2019), 'DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors', *Cell Systems* **8**(4), 329–337.e4.
URL: *http://www.sciencedirect.com/science/article/pii/S2405471219300730*

McInnes, L., Healy, J. & Melville, J. (2020), 'UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction'. Number: arXiv:1802.03426 arXiv:1802.03426 [cs, stat].
URL: *http://arxiv.org/abs/1802.03426*

Mimno, D., Wallach, H. M. & McCallum, A. (2008), 'Gibbs Sampling for Logistic Normal Topic Models with Graph-Based Priors', p. 8.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R. & Yu, B. (2019), 'Definitions, methods, and applications in interpretable machine learning', *Proceedings of the National Academy of Sciences* **116**(44), 22071–22080. Publisher: Proceedings of the National Academy of Sciences.
URL: *https://www.pnas.org/doi/10.1073/pnas.1900654116*

Mylka, V., Matetovici, I., Poovathingal, S., Aerts, J., Vandamme, N., Seurinck, R., Verstaen, K., Hulselmans, G., Van den Hoecke, S., Scheyltjens, I., Movahedi, K., Wils, H., Reumers, J., Van Houdt, J., Aerts, S. & Saeys, Y. (2022), 'Comparative analysis of antibody- and lipid-based multiplexing methods for single-cell RNA-seq', *Genome Biology* **23**(1), 55.
URL: *https://doi.org/10.1186/s13059-022-02628-8*

Nagano, T., Lubling, Y., Stevens, T. J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E. D., Tanay, A. & Fraser, P. (2013), 'Single-cell Hi-C reveals cell-to-cell variability in chromosome structure', *Nature* **502**(7469), 59–64. Number: 7469 Publisher: Nature Publishing Group.
URL: *https://www.nature.com/articles/nature12593*

Pellin, D., Loperfido, M., Baricordi, C., Wolock, S. L., Montepeloso, A., Weinberg, O. K., Biffi, A., Klein, A. M. & Biasco, L. (2019), 'A comprehensive single cell transcriptional landscape of human hematopoietic progenitors', *Nature Communications* **10**(1), 2395. Number: 1 Publisher: Nature Publishing Group.
URL: *https://www.nature.com/articles/s41467-019-10291-0*

Picelli, S., Björklund, K., Faridani, O. R., Sagasser, S., Winberg, G. & Sandberg, R. (2013), 'Smart-seq2 for sensitive full-length transcriptome profiling in single cells', *Nature Methods* **10**(11), 1096–1098. Number: 11 Publisher: Nature Publishing Group.
URL: *https://www.nature.com/articles/nmeth.2639*

Polański, K., Young, M. D., Miao, Z., Meyer, K. B., Teichmann, S. A. & Park, J.-E. (2020), 'BBKNN: fast batch alignment of single cell transcriptomes', *Bioinformatics* **36**(3), 964–965.
URL: *https://doi.org/10.1093/bioinformatics/btz625*

Pollen, A. A., Nowakowski, T. J., Shuga, J., Wang, X., Leyrat, A. A., Lui, J. H., Li, N., Szpankowski, L., Fowler, B., Chen, P., Ramalingam, N., Sun, G., Thu, M., Norris, M., Lebofsky, R., Toppani, D., Kemp, D. W., Wong, M., Clerkson, B., Jones, B. N., Wu, S., Knutsson, L., Alvarado, B., Wang, J., Weaver, L. S., May, A. P., Jones, R. C., Unger, M. A., Kriegstein, A. R. & West, J. A. A. (2014), 'Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex', *Nature Biotechnology* **32**(10), 1053–1058. Number: 10 Publisher: Nature Publishing Group.
URL: *https://www.nature.com/articles/nbt.2967*

Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A. & Murali, T. M. (2020), 'Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data', *Nature Methods* **17**(2), 147–154. Number: 2 Publisher: Nature Publishing Group.
URL: *https://www.nature.com/articles/s41592-019-0690-6*

Qiu, Q., Hu, P., Qiu, X., Govek, K. W., Cámara, P. G. & Wu, H. (2020), 'Massively parallel and time-resolved RNA sequencing in single cells with scNT-seq', *Nature Methods* **17**(10), 991–1001. Number: 10 Publisher: Nature Publishing Group.
URL: *https://www.nature.com/articles/s41592-020-0935-4*

Qiu, X., Zhang, Y., Martin-Rufino, J. D., Weng, C., Hosseinzadeh, S., Yang, D., Pogson, A. N., Hein, M. Y., Hoi (Joseph) Min, K., Wang, L., Grody, E. I., Shurtleff, M. J., Yuan, R., Xu, S., Ma, Y., Replogle, J. M., Lander, E. S., Darmanis, S., Bahar, I., Sankaran, V. G., Xing, J. & Weissman, J. S. (2022), 'Mapping transcriptomic vector fields of single cells', *Cell* **185**(4), 690–711.e45.
URL: *https://www.sciencedirect.com/science/article/pii/S0092867421015774*

Rasmussen, C. E. & Williams, C. K. I. (2006), *Gaussian processes for machine learning*, Adaptive computation and machine learning, MIT Press, Cambridge, Mass. OCLC: ocm61285753.

Regev, A., Teichmann, S., Rozenblatt-Rosen, O., Stubbington, M., Ardlie, K., Amit, I., Arlotta, P., Bader, G., Benoist, C., Biton, M., Bodenmiller, B., Bruneau, B., Campbell, P., Carmichael, M., Carninci, P., Castelo-Soccio, L., Clatworthy, M., Clevers, H., Conrad, C., Eils, R., Freeman, J., Fugger, L., Goettgens, B., Graham, D., Greka, A., Hacohen, N., Haniffa, M., Helbig, I.,

Heuckeroth, R., Kathiresan, S., Kim, S., Klein, A., Knoppers, B., Kriegstein, A., Lander, E., Lee, J., Lein, E., Linnarsson, S., Macosko, E., MacParland, S., Majovski, R., Majumder, P., Marioni, J., McGilvray, I., Merad, M., Mhlanga, M., Naik, S., Nawijn, M., Nolan, G., Paten, B., Pe'er, D., Philippakis, A., Ponting, C., Quake, S., Rajagopal, J., Rajewsky, N., Reik, W., Rood, J., Saeb-Parsy, K., Schiller, H., Scott, S., Shalek, A., Shapiro, E., Shin, J., Skeldon, K., Stratton, M., Streicher, J., Stunnenberg, H., Tan, K., Taylor, D., Thorogood, A., Vallier, L., van Oudenaarden, A., Watt, F., Weicher, W., Weissman, J., Wells, A., Wold, B., Xavier, R., Zhuang, X. & Committee, H. C. A. O. (2018), 'The Human Cell Atlas White Paper', *arXiv:1810.05192 [q-bio]* . arXiv: 1810.05192.
URL: *http://arxiv.org/abs/1810.05192*

Reid, A. J., Talman, A. M., Bennett, H. M., Gomes, A. R., Sanders, M. J., Illingworth, C. J. R., Billker, O., Berriman, M. & Lawniczak, M. K. (2018), 'Single-cell RNA-seq reveals hidden transcriptional variation in malaria parasites', *eLife* **7**, e33105. Publisher: eLife Sciences Publications, Ltd.
URL: *https://doi.org/10.7554/eLife.33105*

Rodriques, S. G., Chen, L. M., Liu, S., Zhong, E. D., Scherrer, J. R., Boyden, E. S. & Chen, F. (2020), 'RNA timestamps identify the age of single molecules in RNA sequencing', *Nature Biotechnology* pp. 1–6. Publisher: Nature Publishing Group.
URL: *https://www.nature.com/articles/s41587-020-0704-z*

Rogers, S., Girolami, M., Campbell, C. & Breitling, R. (2005), 'The latent process decomposition of cDNA microarray data sets', *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **2**(2), 143–156. Conference Name: IEEE/ACM Transactions on Computational Biology and Bioinformatics.

Rogers, S., Girolami, M., Kolch, W., Waters, K. M., Liu, T., Thrall, B. & Wiley, H. S. (2008), 'Investigating the correspondence between transcriptomic and proteomic expression profiles using coupled cluster models', *Bioinformatics* **24**(24), 2894–2900.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4141638/*

Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. (2019), 'A comparison of single-cell trajectory inference methods', *Nature Biotechnology* **37**(5), 547–554. Number: 5 Publisher: Nature Publishing Group.
URL: *https://www.nature.com/articles/s41587-019-0071-9*

Scadden, D. T. (2014), 'Nice Neighborhood: Emerging Concepts of the Stem Cell Niche', *Cell* **157**(1), 41–50. Publisher: Elsevier.
URL: *https://www.cell.com/cell/abstract/S0092-8674(14)00205-0*

See, P., Lum, J., Chen, J. & Ginhoux, F. (2018), 'A Single-Cell Sequencing Guide for Immunologists', *Frontiers in Immunology* **9**. Publisher: Frontiers.
URL: *https://www.frontiersin.org/articles/10.3389/fimmu.2018.02425/full*

Shalek, A. K., Satija, R., Shuga, J., Trombetta, J. J., Gennert, D., Lu, D., Chen, P., Gertner, R. S., Gaublomme, J. T., Yosef, N., Schwartz, S., Fowler, B., Weaver, S., Wang, J., Wang, X., Ding, R., Raychowdhury, R., Friedman, N., Hacohen, N., Park, H., May, A. P. & Regev, A. (2014), 'Single-cell RNA-seq reveals dynamic paracrine control of cellular variation', *Nature* **510**(7505), 363–369. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 7505 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Biotechnology;Engineering;Gene expression analysis;Gene regulation in immune cells Subject_term_id: biotechnology;engineering;gene-expression-analysis;gene-regulation-in-immune-cells.
URL: *https://www.nature.com/articles/nature13437*

Shao, X., Lu, X., Liao, J., Chen, H. & Fan, X. (2020), 'New avenues for systematically inferring cell-cell communication: through single-cell transcriptomics data', *Protein & Cell* .
URL: *http://link.springer.com/10.1007/s13238-020-00727-5*

Slonim, D. K. & Yanai, I. (2009), 'Getting Started in Gene Expression Microarray Analysis', *PLoS Computational Biology* **5**(10), e1000543.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2762517/*

Snelson, E. & Ghahramani, Z. (2005), 'Sparse Gaussian Processes using Pseudo-inputs', p. 8.

Soneson, C. & Robinson, M. D. (2018), 'Bias, robustness and scalability in single-cell differential expression analysis', *Nature Methods* **15**(4), 255–261. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 4 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Databases;Gene expression;RNA sequencing;Statistical methods Subject_term_id: databases;gene-expression;rna-sequencing;statistical-methods.
URL: *https://www.nature.com/articles/nmeth.4612*

Song, D. & Li, J. J. (2021), 'PseudotimeDE: inference of differential gene expression along cell pseudotime with well-calibrated p-values from single-cell RNA sequencing data', *Genome Biology* **22**(1), 124.
URL: *https://doi.org/10.1186/s13059-021-02341-y*

Song, D., Yang, D., Powell, C. A. & Wang, X. (2019), 'Cell–cell communication: old mystery and new opportunity', *Cell Biology and Toxicology* **35**(2), 89–93.
URL: *https://doi.org/10.1007/s10565-019-09470-y*

Squair, J. W., Gautier, M., Kathe, C., Anderson, M. A., James, N. D., Hutson, T. H., Hudelle, R., Qaiser, T., Matson, K. J. E., Barraud, Q., Levine, A. J., La Manno, G., Skinnider, M. A.

& Courtine, G. (2021), 'Confronting false discoveries in single-cell differential expression', *Nature Communications* **12**(1), 5692. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Computational biology and bioinformatics;Functional genomics;Gene expression analysis;Spinal cord injury;Statistics Subject_term_id: computational-biology-and-bioinformatics;functional-genomics;gene-expression-analysis;spinal-cord-injury;statistics.
URL: *https://www.nature.com/articles/s41467-021-25960-2*

Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P. K., Swerdlow, H., Satija, R. & Smibert, P. (2017), 'Simultaneous epitope and transcriptome measurement in single cells', *Nature Methods* **14**(9), 865–868. Number: 9 Publisher: Nature Publishing Group.
URL: *https://www.nature.com/articles/nmeth.4380*

Stoeckius, M., Zheng, S., Houck-Loomis, B., Hao, S., Yeung, B. Z., Mauck, W. M., Smibert, P. & Satija, R. (2018), 'Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics', *Genome Biology* **19**(1), 224.
URL: *https://doi.org/10.1186/s13059-018-1603-1*

Strauss, M. E., Kirk, P. D. W., Reid, J. E. & Wernisch, L. (2020), 'GPseudoClust: deconvolution of shared pseudo-profiles at single-cell resolution', *Bioinformatics* **36**(5), 1484–1491.
URL: *https://doi.org/10.1093/bioinformatics/btz778*

Street, K., Risso, D., Fletcher, R. B., Das, D., Ngai, J., Yosef, N., Purdom, E. & Dudoit, S. (2018), 'Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics', *BMC Genomics* **19**(1), 477.
URL: *https://doi.org/10.1186/s12864-018-4772-0*

Svensson, V. (2020), 'Droplet scRNA-seq is not zero-inflated', *Nature Biotechnology* **38**(2), 147–150. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 2 Primary_atype: Correspondence Publisher: Nature Publishing Group Subject_term: Bioinformatics;Gene expression analysis;Transcriptomics Subject_term_id: bioinformatics;gene-expression-analysis;transcriptomics.
URL: *https://www.nature.com/articles/s41587-019-0379-5*

Svensson, V., Vento-Tormo, R. & Teichmann, S. A. (2018), 'Exponential scaling of single-cell RNA-seq in the past decade', *Nature Protocols* **13**(4), 599–604. Number: 4 Publisher: Nature Publishing Group.
URL: *https://www.nature.com/articles/nprot.2017.149*

Takada, Y., Ye, X. & Simon, S. (2007), 'The integrins', *Genome Biology* **8**(5), 215.
URL: *https://doi.org/10.1186/gb-2007-8-5-215*

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., Lao, K. & Surani, M. A. (2009), 'mRNA-Seq whole-transcriptome analysis of a single cell', *Nature Methods* **6**(5), 377–382. Number: 5 Publisher: Nature Publishing Group.
URL: *https://www.nature.com/articles/nmeth.1315*

Teo, Y. V., Rattanavirotkul, N., Olova, N., Salzano, A., Quintanilla, A., Tarrats, N., Kiourtis, C., Müller, M., Green, A. R., Adams, P. D., Acosta, J.-C., Bird, T. G., Kirschner, K., Neretti, N. & Chandra, T. (2019), 'Notch Signaling Mediates Secondary Senescence', *Cell Reports* **27**(4), 997–1007.e5. Publisher: Elsevier.
URL: *https://www.cell.com/cell-reports/abstract/S2211-1247(19)30451-6*

Titsias, M. (2009), Variational Learning of Inducing Variables in Sparse Gaussian Processes, *in* 'Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics', PMLR, pp. 567–574. ISSN: 1938-7228.
URL: *https://proceedings.mlr.press/v5/titsias09a.html*

Titsias, M. & Lawrence, N. D. (2010), Bayesian Gaussian Process Latent Variable Model, *in* 'Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics', JMLR Workshop and Conference Proceedings, pp. 844–851. ISSN: 1938-7228.
URL: *https://proceedings.mlr.press/v9/titsias10a.html*

Tomasi, F., Ravichandran, P., Levy-Fix, G., Lalmas, M. & Dai, Z. (2020), 'Stochastic Variational Inference for Dynamic Correlated Topic Models', p. 10.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S. & Rinn, J. L. (2014), 'Pseudo-temporal ordering of individual cells reveals dynamics and regulators of cell fate decisions', *Nature biotechnology* **32**(4), 381–386.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4122333/*

Travaglini, K. J., Nabhan, A. N., Penland, L., Sinha, R., Gillich, A., Sit, R. V., Chang, S., Conley, S. D., Mori, Y., Seita, J., Berry, G. J., Shrager, J. B., Metzger, R. J., Kuo, C. S., Neff, N., Weissman, I. L., Quake, S. R. & Krasnow, M. A. (2020), 'A molecular cell atlas of the human lung from single-cell RNA sequencing', *Nature* **587**(7835), 619–625. Number: 7835 Publisher: Nature Publishing Group.
URL: *https://www.nature.com/articles/s41586-020-2922-4*

Van den Berge, K., Roux de Bézieux, H., Street, K., Saelens, W., Cannoodt, R., Saeys, Y., Dudoit, S. & Clement, L. (2020), 'Trajectory-based differential expression analysis for single-cell sequencing data', *Nature Communications* **11**(1), 1201. Number: 1 Publisher: Nature Publishing Group.
URL: *https://www.nature.com/articles/s41467-020-14766-3*

Vento-Tormo, R., Efremova, M., Botting, R. A., Turco, M. Y., Vento-Tormo, M., Meyer, K. B., Park, J.-E., Stephenson, E., Polański, K., Goncalves, A., Gardner, L., Holmqvist, S., Henriksson, J., Zou, A., Sharkey, A. M., Millar, B., Innes, B., Wood, L., Wilbrey-Clark, A., Payne, R. P., Ivarsson, M. A., Lisgo, S., Filby, A., Rowitch, D. H., Bulmer, J. N., Wright, G. J., Stubbington, M. J. T., Haniffa, M., Moffett, A. & Teichmann, S. A. (2018), 'Single-cell reconstruction of the early maternal–fetal interface in humans', *Nature* **563**(7731), 347–353. Number: 7731 Publisher: Nature Publishing Group.
URL: *https://www.nature.com/articles/s41586-018-0698-6*

Verma, A. & Engelhardt, B. E. (2020), 'A robust nonlinear low-dimensional manifold for single cell RNA-seq data', *BMC Bioinformatics* **21**(1), 324.
URL: *https://doi.org/10.1186/s12859-020-03625-z*

Vieth, B., Parekh, S., Ziegenhain, C., Enard, W. & Hellmann, I. (2019), 'A systematic evaluation of single cell RNA-seq analysis pipelines', *Nature Communications* **10**(1), 4667. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Data processing;Transcriptomics Subject_term_id: data-processing;transcriptomics.
URL: *https://www.nature.com/articles/s41467-019-12266-7*

Wang, T., Li, B., Nelson, C. E. & Nabavi, S. (2019), 'Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data', *BMC Bioinformatics* **20**(1), 40.
URL: *https://doi.org/10.1186/s12859-019-2599-6*

Wolf, F. A., Hamey, F. K., Plass, M., Solana, J., Dahlin, J. S., Göttgens, B., Rajewsky, N., Simon, L. & Theis, F. J. (2019), 'PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells', *Genome Biology* **20**(1), 59.
URL: *https://doi.org/10.1186/s13059-019-1663-x*

Wolock, S. L., Lopez, R. & Klein, A. M. (2019), 'Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data', *Cell Systems* **8**(4), 281–291.e9.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S2405471218304745*

Xi, N. M. & Li, J. J. (2021), 'Benchmarking Computational Doublet-Detection Methods for Single-Cell RNA Sequencing Data', *Cell Systems* **12**(2), 176–194.e6.
URL: *https://www.sciencedirect.com/science/article/pii/S2405471220304592*

Xiong, K.-X., Zhou, H.-L., Lin, C., Yin, J.-H., Kristiansen, K., Yang, H.-M. & Li, G.-B. (2022), 'Chord: an ensemble machine learning algorithm to identify doublets in single-cell RNA sequencing data', *Communications Biology* **5**(1), 1–11. Number: 1 Publisher: Nature Publishing Group.
URL: *https://www.nature.com/articles/s42003-022-03476-9*

Yang, C., Siebert, J. R., Burns, R., Gerbec, Z. J., Bonacci, B., Rymaszewski, A., Rau, M., Riese, M. J., Rao, S., Carlson, K.-S., Routes, J. M., Verbsky, J. W., Thakar, M. S. & Malarkannan, S. (2019), 'Heterogeneity of human bone marrow and blood natural killer cells defined by single-cell transcriptome', *Nature Communications* **10**(1), 3931. Number: 1 Publisher: Nature Publishing Group.
URL: *https://www.nature.com/articles/s41467-019-11947-7*

Yang, S., Corbett, S. E., Koga, Y., Wang, Z., Johnson, W. E., Yajima, M. & Campbell, J. D. (2020), 'Decontamination of ambient RNA in single-cell RNA-seq with DecontX', *Genome Biology* **21**(1), 57.
URL: *https://doi.org/10.1186/s13059-020-1950-6*

Yeo, S. K., Zhu, X., Okamoto, T., Hao, M., Wang, C., Lu, P., Lu, L. J. & Guan, J.-L. (2020), 'Single-cell RNA-sequencing reveals distinct patterns of cell state heterogeneity in mouse models of breast cancer', *eLife* **9**, e58810. Publisher: eLife Sciences Publications, Ltd.
URL: *https://doi.org/10.7554/eLife.58810*

Zappia, L. & Theis, F. J. (2021), Over 1000 tools reveal trends in the single-cell RNA-seq analysis landscape, Technical report. Company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Results Type: article.
URL: *https://www.biorxiv.org/content/10.1101/2021.08.13.456196v2*

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., Bharadwaj, R., Wong, A., Ness, K. D., Beppu, L. W., Deeg, H. J., McFarland, C., Loeb, K. R., Valente, W. J., Ericson, N. G., Stevens, E. A., Radich, J. P., Mikkelsen, T. S., Hindson, B. J. & Bielas, J. H. (2017), 'Massively parallel digital transcriptional profiling of single cells', *Nature Communications* **8**(1), 14049. Number: 1 Publisher: Nature Publishing Group.
URL: *https://www.nature.com/articles/ncomms14049*

Zimmerman, K. D., Espeland, M. A. & Langefeld, C. D. (2021), 'A practical solution to pseudoreplication bias in single-cell studies', *Nature Communications* **12**(1), 738. Number: 1 Publisher: Nature Publishing Group.
URL: *https://www.nature.com/articles/s41467-021-21038-1*