Imperial College London
Department of Mathematics

Single Cell Mitochondrial DNA: Bioinformatics And Statistical Approaches

Aidan. S. Marshall

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy in Mathematics
March 2024

Declaration of Originality

I certify that this thesis, and the research to which it refers, are the product of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline

Copyright

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC). Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose. When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes. Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Abstract

New technologies drive scientific discoveries. Recently, genomics sequencing has rapidly advanced, opening numerous doors for understanding human health and disease. Sequencing techniques are capable of producing prodigious amounts of raw sequencing data routinely, and their output is only increasing. Developments in computational and statistical methods for processing and analyzing this data have striven to keep pace with this data growth. This thesis builds on these developments in the context of studying mitochondrial DNA (mtDNA) and understanding the role it plays in health and disease made possible through nascent *single-cell* genomics technologies.

We begin by detailing the landscape of single-cell sequencing techniques and existing work applying these techniques to study cellular mtDNA heterogeneity. Multitudes of single-cell datasets are publicly available and current analyses largely ignore the abundant mitochondrial content of their data. Importantly, single-cell datasets allow us to gain an understanding of the role mtDNA plays that is missing in bulk genomic datasets. Previous research has linked mtDNA mutations to ageing pathology. Using the Moran model, we show that post-mitotic cells should accumulate mtDNA mutations, reaching a pathological threshold on the timescale of a human lifespan. Such mutational burden is only seen at the single-cell level, which we term *cryptic* as it is unobservable in bulk datasets. A pipeline we developed leveraged available single-cell data, confirming that empirical results corroborate theoretical expectations. Moreover, we link the presence of cryptic mutations to many classic hallmarks of ageing.

We then use novel scATAC data to explore the landscape of widespread mtDNA mutations in the developing brain. Such samples are frozen, limiting the mtDNA content that can be captured from each cell. To overcome this, we take advantage of tools from Bayesian inference. This approach allows us to acquire compelling functional insight from low-depth data, whilst faithfully representing the inherent uncertainty.

Acknowledgements

I must begin by expressing my deep gratitude to my PhD supervisor, Nick. S. Jones. I am thankful for the opportunity to have pursued a PhD at all, but all the more so for having had the pleasure of doing it under Nick's supervision. I am in debt to him for his patience, critical questioning, and helpful suggestions. Above all, I am appreciative for the joy and scientific curiosity that Nick exudes and how this is reflected in the excellent culture of the Systems and Signals group.

I am thankful to the many members of the Systems and Signals group, past and present. I begin with those with whom I have worked most closely at the beginning of my PhD: Florian Klimm and Ali Green. Our collaboration on somatic cryptic mtDNA mutational ageing, a daunting first PhD project, was made manageable largely through your mutual reassurance and support. More than this, once the Pandemic hit, our work on that project was an island of stability in an otherwise chaotic world. I must also thank Ferdinando Insalata, Marco Esposito, Ben Ingledow, Johannes Happenhofer, Rein Leetma, Sarab Sethi, Hethvi Jethwani, and Daniel Kornai. Not only is the group constantly abuzz with stimulating ideas, but laughter abounds. It was a pleasure to have spent time in all of your company.

I reserve a special space for two members of the group in particular. Juvid Aryaman has been a mentor to me far beyond his time in Systems and Signals. The deepest parts of the pandemic were made enjoyable through our conversations, half-baked ideas, and quirky shared interests. Our deep forays into the underbelly of Python were, somehow, the most exciting of learning experiences. I hope life allows me the same pleasure in another domain.

If posed with a problem, be it a mathematical problem or ethical dilemma, a good first step in solving it would be to ask "What would Sahil Loomba do?" Thank you for being patient with all my academic intrigues and intellectual curiosities, no matter how small or large. More than this, thank you for being a dear friend when it was most needed.

One member of Imperial not in our group has had a deep impact on this thesis, though he may not know it. Andy Thomas, the Research Computing Manager in the Mathematics Department has been nothing short of fantastic. The work you put into keeping the computing resources reliable and available is understated, to say the least. Your responsiveness is unmatched and

the University is profoundly lucky to have someone as dedicated to their work as you. Without you, this thesis would not have gone anywhere near as far as it did.

Many thanks to my collaborators at the MBU. In particular, Patrick. F. Chinnery, Nell Nie and Dane Wolf. Patrick's insight, critical questioning, and supportive demeanor are an exemplar of excellence in scientific leadership. I am thankful for all the time you offered and your deep knowledge of mitochondrial biology and medicine. Nell, you have been not only a wonderful collaborator, but kind and supportive during my time in Cambridge. Dane's data and questioning have been a great source of motivation in thinking about the physical behaviour of mtDNA.

A thesis is in particular accompanied by the support of many friends and family. Without my Mam and Dad, none of this would have been conceivable. I am beyond lucky to have been born to parents who have been so supportive of me achieving my goals. I am grateful to my father for believing in the power of education and always encouraging me, despite us being from somewhere where this is seldom the case.

In particular, I am thankful to my excellent teachers at Queen Elizabeth Sixth Form College. If I achieve anything in life, it will be because of the truly excellent *state education* I received there. Any failings are my own. Sarah McAdam saw me through my maths A-Level. I owe you a great debt. Brian Wardle, you were the best of all possible physics teachers. Your passion for physics still resonates today. Jane and Allen provided the best chemistry education possible. Beyond education, I owe a lot to the best of friends who came all the way from that fine institution. Jones, Morgan, and Shaya. I am beyond thankful for our endless calls and the best of times. You each helped me get to, and complete this PhD in your own way.

I must give two final notable mentions. To Keir, my oldest friend. You have seen it all with me. This destination seemed so improbable and the ride has indeed taken us down many long and meandering roads. I look forward to its continuation for many years to come.

Cheryl, you have been beyond patient. I am unworthy of the deep support you have given me over these demanding last few months. Words cannot express my gratitude. You, more than anyone, helped me see this through to the end.

Dedication

For you, forever ago

‘We are drowning in information and starving for knowledge.’

- *Rutherford D. Rogers*

Contents

Abstract **ii**

Acknowledgements **iii**

1 Introduction - Mitochondria **1**

1.1 Summary 2

1.2 Mitochondrial Biology 3

1.3 Thesis Overview 6

1.4 Publications And Contributions 10

2 Discovering Cellular Mitochondrial Heteroplasmy Heterogeneity with Single Cell RNA And ATAC Sequencing **11**

2.1 Summary 11

2.2 Introduction 13

2.3 Background On Single-Cell Omics 15

2.4 Sequencing Protocols & mtDNA 17

2.4.1 scRNA-seq 17

2.4.2 scATAC-seq 19

2.4.3 The quality of single-cell omics mtDNA data 19

2.4.4 High-Throughput scATAC-seq 26

2.5 Heteroplasmy In Single-Cell Data 28

2.5.1 Mitochondrial Heterogeneity With scRNA-seq And scATAC-seq 28

2.5.2 EMBLEM 31

2.5.3 mtscATAC-seq 32

2.5.4 MAESTER 34

2.6 Future Directions And Open Problems In Assessing Heteroplasmy At The Single-Cell Level 36

2.6.1 PCR and heteroplasmy 36

2.6.2 Photobleaching And Strand Concordance 37

2.6.3 Mitochondrial Cellular Transfer 38

2.6.4 Ambient RNA In Droplet-Based Approaches 38

2.6.5 Mitochondrial Gene Expression Proportion Quality Control 39

2.6.6 NUMTs And Heteroplasmy 40

2.6.7 High Throughput Droplet Data 41

2.7 Conclusion 43

3 Somatic Mitochondrial Mutations In Ageing 44

3.1 Summary 45

3.2 Introduction 46

3.3 The Moran Model 50

3.4 Single Cell mtDNA Variant Calling Pipeline 55

3.4.1 Background On Sequencing Protocols 56

3.4.2 Data Downloading 59
3.4.3 Alignment, Demultiplexing, and UMI Counting 59
3.4.4 Single Cell Variant Calling 61
3.4.5 Classifying Detected Mutations 62
3.4.6 Expression Analysis 63
3.5 Data Selection 64
3.6 Results 65
3.6.1 Coverage And Variant Calls 65
3.6.2 Accumulation Of Cryptic Mutations Through Time 67
3.6.3 Impact On Brown Adipose Tissue And Liver In A Rat Comparing Caloric Restriction To Normal Ageing 70
3.6.4 Linking Cryptic Mutations To Pathophysiology 72
3.7 Conclusion 74

4 Bayesian Inference As A Tool For Low Depth Data 76

4.1 Summary 77
4.2 Introduction 79
4.3 Background On Bayesian Inference 84
4.3.1 Motivating The Bayesian Approach 84
4.3.2 Bayesian Networks 85
4.3.3 Probabilistic Programming Languages 87
4.3.4 MCMC 87
4.3.5 Concluding Remarks On The Advantages Of The Bayesian Approach . . 97

4.4 Representing Heteroplasmy Uncertainty . . . 99

4.5 Modelling Gene Activity . . . 102

4.5.1 Negative Binomial Regression . . . 106

4.5.2 Negative binomial regression with heteroplasmy uncertainty . . . 108

4.6 Simulation Studies . . . 111

4.6.1 Low Depth Destroys Parameter Estimates In Conventional Models . . . 111

4.6.2 Simulation Based Calibration . . . 115

4.6.3 Calibration Under Misspecified Priors . . . 120

4.7 Multiple Testing Concerns . . . 123

4.7.1 The Bonferroni Correction . . . 124

4.7.2 Benjamini-Hochberg Procedure . . . 125

4.7.3 Benjamini-Hochberg Procedure — A Bayesian Interpretation . . . 126

4.7.4 The Benjamini-Hochberg procedure applied to Bayesian Posteriors . . . 130

4.7.5 Verifying Benjamini-Hochberg Controls The FDR *In Silico* . . . 131

4.8 Conclusion . . . 136

5 Mitochondrial Mutations In The Developing Brain 139

5.1 Summary . . . 140

5.2 Introduction . . . 141

5.3 Sample Collection . . . 144

5.4 Computational Pipeline . . . 145

5.4.1 Snakemake Pipeline . . . 146

5.4.2 Large Scale Pipeline . . . 149

5.4.3 Notebooks 150

5.5 Single-Cell Analysis 151

5.5.1 Clustering 151

5.6 Cell Type Annotations 155

5.6.1 The Forebrain 155

5.6.2 The Cerebellum And The Medulla 157

5.6.3 The Medulla 159

5.7 The Landscape Of Developmental Mutations 161

5.7.1 Shallow Depth In Frozen Samples 161

5.7.2 Mutatbility Of Mitochondrial Genome 161

5.7.3 Identifying Candidate Mutations For Further Analysis 163

5.7.4 Properties Of Identified Mutations 166

5.8 Association Between Heteroplasmy And Gene Activity 173

5.8.1 Negative Binomial Regression Including Heteroplasmy Uncertainty 173

5.8.2 Bayesian Inference At A Massive Scale: Software Engineering Challanges 173

5.8.3 Posterior Predictive Checks 175

5.8.4 Bayesian Inference At A Massive Scale: Validating Model Behaviour . . . 178

5.9 Functional Relationships 183

5.9.1 Functional Effects Across All Mutations 186

5.10 Conclusion 188

5.11 Future Directions 192

6 Addendum — Observing mtDNA Motion **195**

6.1 Summary 195

6.2 Tracking mtDNA 196

6.3 Simulating Brownian Motions In Finite Boxes 197

6.4 Cristae Distort Distance Measurements 204

6.5 Conclusion 207

7 Conclusion **208**

7.1 Thesis Summary 208

7.2 Closing Remarks 213

A Supplementary Material For Chapter 4 **218**

A.1 Stan Code 219

B Supplementary Material For Chapter 6 **220**

B.1 Stan Code 221

Bibliography **222**

List of Tables

3.1 Summary of analysed publicly available datasets 64

4.1 Simulation shows that applying the Benjamini-Hochberg procedure to posterior quantiles recovers the theoretical false discovery rate 134

5.1 Posterior predictive quantile coverage 177

5.2 Problematic sampling is caused by inactive genes 180

5.3 GO terms for genes associated with 3010G>A 186

5.4 Recurrent GO terms for widespread developmental mutations 187

List of Figures

2.1 Overview of central dogma, mitochondrial DNA, heteroplasmy and basic transcriptomic sequencing modalities 18

2.2 Breadth and depth of coverage in single-cell data 22

2.3 Breadth and depth of coverage across the mitochondrial genome 27

3.1 The site-frequency spectrum over human lifespan 52

3.2 DAG of single-cell RNA heteroplasmy pipeline 55

3.3 Coverage quality of different sequencing modalities 66

3.4 Single-cell sequencing reveals the accumulation of cryptic mutations with age . . 68

3.5 Empirical ageing of the human site-frequency spectrum 70

3.6 Caloric restriction inhibits the accumulation of pathological mutations with age 71

3.7 Functional analyses link the presence of mtDNA mutations with the hallmarks of ageing 73

4.1 DAG representing $p(x, y, z) = p(x)p(y|x)p(z|x, y)$. 85

4.2 DAG representing $p(x, y, z) = p(y)p(z|y)p(x|y, z)$. 86

4.3 DAG representing $p(x, y, z) = p(x)p(y|x)p(z|x, y) = p(x)p(y|x)p(z|y)$. 86

4.4 Traceplots distinguishing identifiable and non-identifiable models 96

4.5 DAG representing heteroplasmy uncertainty model 100

4.6 Posterior heteroplasmy varies with depth 101

4.7 The Gamma-Poisson model adequately describes gene activity 106

4.8 DAG showing dependencies in the negative binomial regression model. 108

4.9 DAG representing negative binomial regression including heteroplasmy uncertainty 110

4.10 Naive heteroplasmy estimate in low-depth data introduces discretisation effects . 112

4.11 Bayesian model including full parameter uncertainty recover true parameter values where naive MLE fails 114

4.12 Trace plots for full Bayesian negative binomial heteroplasmy uncertainty model . 115

4.13 Illustration of posterior quantiles used in simulation-based calibration 118

4.14 Uniform quantiles in simulation-based calibration demonstrates validity of our implementation of negative binomial heteroplasmy uncertainty model 119

4.15 Simulation-based calibration of shows that even mispecified models become calibrated with enough cells 122

4.16 Posterior quantiles of regression coefficients under the null are uniformly for large numbers of cells 133

4.17 Visualisation of the values in table 4.1 135

5.1 Snakemake pipeline 147

5.2 Expanded snakemake pipeline 150

5.3 UMAP Structure of all ATAC data combined 153

5.4 UMAPs for data aggregated by Carnegie stage 154

5.5 Forebrain UMAPs for different Carnegie Stages 156

5.6 Forebrain marker genes 157

5.7 Cerebellum UMAPs for different Carnegie Stages 157

5.8 Cerebellum marker genes 159

5.9 Medulla UMAPs for different Carnegie Stages 159

5.10 Medulla marker genes 160

5.11 Mean mtDNA depth of coverage in embryonic brain data 162

5.12 Mutatbility of different mtDNA loci 163

5.13 The profile of identified mtDNA mutations 167

5.14 The landscape of early developmental mtDNA mutations. 172

5.15 Posterior uncertainty of mtDNA heteroplasmy-gene activity associations 175

5.16 Traceplots of negative binomial heteroplasmy uncertainty model fit to real data 181

5.17 Posterior predictive checks for the negative binomial heteroplasmy uncertainty model applied to real gene activity data 182

5.18 Shifts in posterior mean attenuated by depth of coverage 182

5.19 Analysing Bayesian mtDNA heteroplasmy-gene activity associations at a massive scale 185

6.1 HMC samples of diffusion proposals in a finite domain 199

6.2 Reflected Gaussian samples of diffusion proposals in a finite domain 203

Chapter 1

Introduction - Mitochondria

"Powerhouse Of The Cell" – Philip Siekevitz [1]

1.1 Summary

Chapter 2 reproduces material from a review article I authored, providing a comprehensive background of single sequencing regarding mitochondrial and covers much of the introductory technical material for this thesis. As such, this introductory chapter is kept brief, giving a short overview of the biology of mitochondria and their DNA, highlighting the important role that these organelles play in the regular function of the cell. We then give an outline of the thesis structure and how all the pieces tie together before summarising publications and contributions.

1.2 Mitochondrial Biology

Mitochondria are organelles inhabiting the interiors of eukaryotic cells. Frequently, they are dubbed *the powerhouse of the cell*, for their crucial role in the production of adenosine triphosphate (ATP) [1]. ATP is the most widely used activated carrier molecule in the cell that can be transported to locations in need of energy. Energy is then supplied through the hydrolysis of ATP into adenosine diphosphate [2]. In order to maintain the supply of this vital molecule, mitochondria synthesise ATP through the phosphorylation of ADP. To achieve this remarkable feat, the boundaries of mitochondria are comprised of double membranes. Protons are channelled across the inner membrane to maintain an electrochemical gradient between the intermembrane space and the mitochondrial matrix, the mitochondria's interior. As protons travel back down this gradient, from the intermembrane space back into the mitochondrial matrix, they do so by way of ATP-synthase which extracts work to synthesise ATP [3]. To establish this gradient in the first place, the inner membrane houses the series of proteins that form the *electron transport chain* (ETC). The ETC takes high-energy electrons produced by numerous other oxidative processes within the cell and extracts energy from them to pump protons against the gradient [4]. Taken together, these processes allow mitochondria to fulfil their energetic role of converting energy derived from numerous processes into the universal currency of ATP. However, the role that mitochondria play in the functioning of the cell far exceeds the energetic, including but not limited to: maintenance of the cellular membrane; regulation of cellular metabolism; immune and hormonal signalling; and cell death [5, 6, 7, 8, 9].

A characteristic feature of mitochondria, compared to other organelles, is that they harbour their own DNA, mitochondrial DNA (mtDNA), which is distinct from the DNA in the nucleus and typically comes in 100-1,000s of copies in every cell. Evolutionarily, this additional DNA is thought to have originated through an endosymbiotic event in early evolutionary history in which an archaeon and a bacterium merged [10, 11]. Since this ancient event, most genes are thought to have migrated into the nucleus, with evidence suggesting this process is still continuing today [12, 13]. This has resulted in a human mitochondrial genome approximately 16kb long, containing 13 protein-coding genes, 22 tRNAs genes, and 2 rRNA genes. Importantly,

these genes play a key role in the electron transport chain and the regular functioning of the mitochondria. Disruption to mtDNA in the form of mutations thus has the potential to be pathological to the normal functioning of the cell. Unfortunately, a range of mitochondrial diseases, such as MERRF, MELAS, and Leber Hereditary Optic Neuropathy (LHON), are known to be caused by mitochondrial mutations [14, 15, 16, 17]. Nuclear mutations occurring at some position either occur on neither, one or both copies of a chromosome inside a cell. The numerous copies of the mtDNA in each cell complicate matters relative to nuclear mutations. Some proportion, between 0 and 1, of the mtDNA in a cell or a tissue sample can carry the mutation and is termed *heteroplasmy*. Moreover, the heteroplasmy of mutations is known to be a key determinant of the severity of known diseases and is an active area of study [18]. Importantly, some mutations are known to become pathological only once their heteroplasmy has surpassed a certain threshold [19].

This thesis explores the role of mitochondrial DNA at the level of the single cell. Primarily, this centres around analysing mitochondrial DNA sequence information at a cellular level and attempting to understand the functional consequences this might have. Secondarily, at the end of the thesis, an addendum chapter summarizes my contribution to ongoing work exploring the physical motion of mtDNA nucleoids through mitochondria, as revealed by high-resolution microscopy, is explored.

Most of the work contained in this thesis has only been made possible through technological innovation that has occurred in the last few years. The relevant technologies can be divided into two areas: 1) the revolution in single-cell sequencing techniques, which allows DNA samples from tens of thousands of single cells to be sequenced simultaneously, and 2) advances in Bayesian computation and computing power, enabling advanced yet resource-hungry Bayesian model implementations to become viable on a large scale.

The breakthrough in high-throughput single-cell genomics has allowed cell to cell variation to be explored in fine granularity at a massive scale. Researchers are routinely producing studies of thousands of cells with up to around 20,000 features for each cell, with the literature ambitiously aiming to generate a human-cell atlas [20]. Such vast quantities of data pose challenges about

how it can be intelligently probed to gain tangible and robust insights, motivating the epigraph at the start of this thesis — *We are drowning in information and starving for knowledge.* This thesis aims to satiate some of this hunger by presenting some novel insights regarding the important role mitochondrial DNA mutations play in cellular function in developing and ageing human tissues. The main focus of the single-cell literature regarding mtDNA is as a quality control metric, in which cells with an mtDNA content exceeding a certain threshold are excluded from analysis [21, 22]. We aim to make much greater use of this data for understanding cellular biology. With mitochondrial DNA being present in 1,000s of copies in each cell and also being the most abundant form of RNA in the cell, many single-cell data sets are overflowing with *genotype* information about mtDNA and *phenotype* information about the rest of the cell. As such, this thesis views the mitochondrial content of single-cell data as a primary piece of information for understanding the health of the cell, and not simply a by-product of sequencing to be used for quality control purposes.

Towards this goal of extracting insight from biological sequence data, we turn to Bayesian inference later in the thesis for measuring the association between the heteroplasmy of mutations and gene activity. The substantive developments in Bayesian analysis in recent decades have allowed ever more sophisticated models to be developed and fit to data on manageable timescales. Though Bayesian modelling allows for bespoke models to be developed to solve specific problems, the sampling algorithms used to approximate posterior distributions are resource-hungry. The continued growth in computing power has allowed increasingly complex, high-dimensional models to be easily fit to data over the last decade. Probabilistic programming languages such as `stan` implement efficient sampling algorithms like Hamiltonian Monte Carlo in compiled code that can be run quickly [23]. In these languages, sampling implementations are abstracted away, allowing practitioners to focus on modelling details. Though software and hardware advances enable researchers to build and fit ever more complicated models, not every model is valid or well-defined for every set of data it is fit to. Recent methodological developments under the term *Bayesian Workflow* allow us to guard against these scenarios ensuring our models are robust, identifiable, and reasonably describe our data [24].

1.3 Thesis Overview

Chapter 2 is derived from the contents of a published review article I authored, providing the background material and context required for understanding the thesis as a whole. The chapter serves as a primer on single-cell genomics, differentiating the different sequencing techniques and aims to elucidate many of the technical aspects concerning the mtDNA content of single-cell data. The terminology surrounding these techniques and their technical analyses is scattered throughout the literature, especially components regarding mitochondrial content. This includes impressive previous work of other labs using single-cell techniques using mitochondria DNA as cellular barcodes for lineage tracing and exploring cellular heterogeneity. By connecting disparate sources, we synthesise the relevant pieces of the literature, clarifying the concepts of mitochondrial data quality relating to depth, and highlighting recurrent shortcomings and open questions relating to mitochondrial DNA in the single-cell literature.

Chapter 3 builds directly upon the understanding gained in Chapter 2 and applies it to the study of single-cell mtDNA ageing. A simple model from population genetics, the Moran Model, indicates how mitochondrial mutations should accumulate in fixed-size populations of DNA like the mtDNA in post-mitotic cells. Using literature values for population size, mtDNA turnover, and mutation rates, the Moran model indicates that the heteroplasmy of mitochondrial mutations in single cells would accumulate to pathological levels on the timescale consonant with a human life.

This accumulation of mtDNA mutations has gone unobserved in previous bulk tissue studies and observing them would only be possible in single-cell data. We thus dub these post-mitotic, somatic, single-cell mutations, *cryptic mutations*. Chapter 3 seeks to answer if there is evidence that such an accumulation of high heteroplasmy cryptic mutations occurs with age, and if so, what effects do they have? As part of a team with collaborators Alistair Green and Florian Klimm, we set out to find such evidence. We developed a new computational pipeline capable of going from raw, single-cell sequencing data, to heteroplasmy estimates of identified mitochondrial variants, alongside the expression date for each single cell in the data.

Using this pipeline, we analysed publicly available from 3 different species, and several different tissues. My primary contribution was the development, testing and deployment of the pipeline on publicly available data which I was also responsible for obtaining. To highlight the relevance of my work, I present analysis by my colleagues using the outputs of the pipeline, showing that the accumulation of cryptic mutations does indeed occur at a level previously unobserved in bulk data. Furthermore, we show the presence of these mutations could be linked to many of the hallmarks of ageing [25]. I conducted a further analysis of the brown adipose tissues in rats, some of which were calorically restricted, and others which were not. This analysis reveals that caloric restriction, a known longevity therapeutic, attenuates the accumulation of high heteroplasmy mutations with age.

Chapter 4 and **Chapter 5** shift attention from human ageing, back to life's beginning. Our collaborators, Patrick Chinnery and Nell Nie, from Cambridge University's Mitochondrial Biology Unit supplied us with single-cell ATAC data of embryonic human data covering 5 Carnegie Stages from three different brain regions. This project set out with a view to exploring the landscape of mitochondrial mutations in human brain development and the functional effects that they may play. At the early stages of development, the problematic cryptic mutations of late life, studied in Chapter 3 are less relevant. At this early time point, cryptic mutations are both rare and exceedingly unlikely to be found at a heteroplasmy of consequence, thus we keep our focus on widespread mutations that will have either been transmitted through the germline or occurred very early in development, which are to be found in many cells.

The tissue samples sequenced by our collaborators were frozen before they were sequenced, having the undesirable effect of curtailing the rate at which mtDNA could be captured during sample preparation and sequencing. Freezing samples limits the number of mtDNA sequencing reads that could be obtained in each cell. Such limited depth poses a challenge for obtaining reliable cellular heteroplasmy estimates. In order to retain the uncertainty in this estimation, we turn to Bayesian inference. Variations in the heteroplasmy of mutations is known to impact the degree of pathology in mitochondrial disease. As such, it is reasonable to suppose that heteroplasmy has an impact on cellular gene activity. Therefore, measuring the association between heteroplasmy and gene activity is desirable for understanding the role mtDNA mutations play

during development, even in the context of low-depth data. In Chapter 4, I develop a Bayesian regression model with uncertain heteroplasmy predictors capable of measuring the association between heteroplasmy and gene activity, even in the context of low-depth data. By taking account of the uncertainty of heteroplasmy in low-depth data, I show, *in silico*, that the model can recover true parameter values where models using naive point estimates of heteroplasmy fail.

Importantly, to acquire the most understanding, it is necessary to measure the association between the heteroplasmy of a mutation and the activity of thousands of genes, naturally giving rise to multiple testing concerns. In the frequentist literature these can be addressed using standard procedures such as the Bonferroni-Correction or the Benjamini-Hochberg(BH) procedure [26]. Chapter 4 further gives an approach for controlling the false discovery rate (FDR) that can be applied to Bayesian models under restricted conditions. I demonstrate that if the posterior quantile of the null, under data generated from the null follows a $Unif(0, 1)$ distribution, the classic BH procedure can be applied to null posterior quantiles, rather than p-values, and control the FDR at the theoretical level. *In silico*, I demonstrate that our Bayesian heteroplasmy uncertainty regression model fits this requirement with data from enough cells. The BH procedure can then be used to control the FDR of measured associations between heteroplasmy and gene activity.

With the methodology of Chapter 4 established, Chapter 5 goes on to apply these techniques to the data. I developed a new computational pipeline to meet the demands of new data-types. Clustering analysis of the scATAC data gives an overview of the general pattern of how cells behave during development. An approach for robustly differentiating widespread mitochondrial mutations from sequencing errors in low-depth data is developed and gives a set of mutations in line with the characteristic profile of mtDNA mutations. The methods of Chapter 4 are then applied to this set of mutations to measure the association between activity and heteroplasmy for this set of mutations. To measure the association between tens of mutations and the activity of approximately 20,000 genes, in data from separate brain regions and multiple Carnegie stages, the model has to be applied *millions* of times. To approximate these posteriors, trillions of MCMC samples were obtained. Finally, the sets of confident genes

obtained from these posteriors suggest the widespread mutations found in the developing brain are associated with changes to mitochondrial network dynamics and alterations to regular neuronal functioning.

Chapter 6 is a brief addendum chapter showing some simple mathematical modelling of the motion of mtDNA inside the mitochondrion. These models are used to guide the interpretation of an on-going piece of work exploring the motion of mtDNA using high-resolution microscopy done in collaboration with researchers at the University of Cambridge. The description is qualitative in nature, describing the observed motion of mtDNA within the intercristae space, using a microscopy with a limited frame rate. Naively, observations measuring the distance travelled between sequential frames indicate that the distance mtDNA traverse over one minute increases linearly with the area of the inter-cristae space. The mathematical model is able to intuitively and adequately explain these observations as illusory by describing mtDNA as undergoing normal Brownian motion in a constrained physical region.

1.4 Publications And Contributions

A. S. Marshall and N. S. Jones, "Discovering Cellular Mitochondrial Heteroplasmy Heterogeneity with Single Cell RNA and ATAC Sequencing," Biology, vol. 10, no. 6, p. 503, Jun. 2021. Available: https://doi.org/10.3390/biology10060503.

Contribution: Researched and wrote the paper with feedback from the supervisor.

A. P. Green, F. Klimm, **A. S. Marshall**, R. Leetmaa, J. Aryaman, A. G´omez-Dur´an, P. F. Chinnery, and N. S. Jones, "Cryptic mitochondrial ageing coincides with mid-late life and is pathophysiologically informative in single cells across tissues and species," bioRxiv, 2023. Available: https://doi.org/10.1101/2023.07.04.547509 2023.07.04.547509. (Submitted to a large-audience multidisciplinary scientific journal)

Contribution: Wrote, tested and ensured the bulk of bioinformatics pipeline used in the project. Specifically, components used to obtain, align, and de-multiplex publicly available sequence data from several different sequencing modalities. Ran the pipeline end to end on the human retinal and brown adipose rate 10x3'datasets. Additionally, adapted the pipeline to handle CEL-Seq2 which I then also produced. Additionally statistical analyses comparing SFS from brown adipose tissues were conducted by myself.

Material for chapters 4, and 5 contributes to an ongoing project and was done in collaboration with Nell Nie and Patrick Chinnery, of which my analysis forms a part. All of the analysis presented is my own.

Material for chapter 6 contributes to an ongoing project and was done in collaboration with Dane Wolf and Patrick Chinnery of which my analysis forms a part. All of the analysis presented is my own.

Chapter 2

Discovering Cellular Mitochondrial Heteroplasmy Heterogeneity with Single Cell RNA And ATAC Sequencing

2.1 Summary

In this section, we reproduce material from a published review article authored by my supervisor and me. I conducted the necessary literature review and wrote all the material in the article under the supervision of my supervisor [27]. This work fits suitably at the beginning of the thesis as it gives context to the reader for much of the material that follows in later chapters. The range of technologies and the associated terminology can be daunting to the unfamiliar. Before the article, few materials existed that new researchers in the field could use turn to understand the place of mtDNA in the context of single-cell genomics and the technical challenges associated. The article aims to provide this context, serving both as a primer for newcomers and to flesh out some of the nuances surrounding mtDNA in sequencing technologies that might be of benefit to those fluent in the language of mitochondrial genomics or single-cell

genomics, but perhaps not both. It is hoped the reader finds this synthesis to be a meaningful first step to understanding what follows it. The contents of this chapter is summarised using the abstract from the article.

Next-generation sequencing technologies have revolutionised the study of biological systems by enabling the examination of a broad range of tissues. Its application to single-cell genomics has generated a dynamic and evolving field with a vast amount of research highlighting heterogeneity in transcriptional, genetic and epigenomic state between cells. However, compared to these aspects of cellular heterogeneity, relatively little has been gleaned from single-cell datasets regarding cellular mitochondrial heterogeneity. Single-cell sequencing techniques can provide coverage of the mitochondrial genome which allows researchers to probe heteroplasmies at the level of the single-cell, and observe interactions with cellular function. In this review, we give an overview of two popular single-cell modalities—single-cell RNA sequencing and single-cell ATAC sequencing—whose throughput and widespread usage offer researchers the chance to probe heteroplasmy combined with cell state in detailed resolution across thousands of cells. After summarising these technologies in the context of mitochondrial research, we give an overview of recent methods that have used these approaches for discovering mitochondrial heterogeneity. We conclude by highlighting the current limitations of these approaches and open problems for future consideration.

2.2 Introduction

Mitochondria play a central role in most human cell types, and by extension, ensuring their healthy function is critical. When such mitochondrial function is perturbed, substantial pathology can ensue, resulting in debilitating mitochondrial diseases [28]. Unlike other organelles found in animal cells, mitochondria possess their own DNA, referred to as *mtDNA*, which exists as 100s to 1,000s of copies per cell. Individual humans universally possess a mixture of mitochondrial genomes, a state referred to as *heteroplasmy* [29]. The precise allele fraction of this mixture at any given locus can vary from , 0%, all the way up to 100%, the extremes being called *homoplasmy*, in which a mutant variant has taken over the population of mtDNA. The state of heteroplasmy can originate through inheritance down the maternal line or via somatic mutations that occur during development or later in life. Originally mitochondrial mutations were hypothesised to have arisen due the exposure of mtDNA to large amounts of reactive oxygen species, however the pattern of mutations is more consistent with errors that take place during replication [30]. Numerous diseases, as well as aspects of ageing, have been linked with mitochondrial mutations and the extent to which these mutations are pathological can greatly depend on their heteroplasmy [31, 32].

The level of heteroplasmy is heterogeneous at the population, individual, tissue, and cellular level [33, 34]. Whilst some studies focus on specific mitochondrial mutations which are pathological or have been associated with mitochondrial diseases [35, 36], the focus of the current review is on methods which discover mutations with high throughput to uncover deeper heterogeneity. Next-generation sequencing (NGS) has been a key technology for investigating many aspects of heteroplasmy, however many early studies were conducted using bulk sequencing techniques, which entails homogenising cells from a tissue placing an inherent limit on the resolution of these approaches, preventing analysis of cellular variation [37, 29]. In recent years, single-cell omics approaches have revolutionised biological research, enabling levels of granularity that were previously not possible, with several recent studies making use of these technologies to study the heterogeneity of heteroplasmy within a tissue. To give one example, lineage tracing is the process of identifying a cell's progeny and grouping the descendent cells

together through the use of a labelling strategy. Lineage tracing has applications in fields including stem cell biology for understanding cell fate decisions, and in cancer biology to infer how tumour heterogeneity arises with implications for treatment resistance [38]. A body of recent work has demonstrated that mitochondrial mutations are projected down cell lineages, and that this can be observed using single-cell data, making the mitochondrial genome an effective *in vivo* barcode. Beyond lineage tracing, researchers should in principle be able to obtain vital mitochondrial features at the single-cell level: both the presence of mutations, and the heteroplasmy of those mutations. Single-cell approaches are further advantageous in that observed mutations can be cross-validated in other cells to protect against sequencing error. However, they also offer the prospect of detecting mutations which would be unobservable in homogenised mixtures, which would be expected of some somatic mutations [39]. Together with features of cell state, this data would facilitate research in understanding the mechanisms by which the accumulation of heteroplasmy leads to pathology.

In this review we begin by reporting what methodological progress has been made in exploring aspects of the mitochondrial genome in single-cell data thus far, what its limitations are, and what the future scope is for using this data to further explore heterogeneity in heteroplasmy at the single-cell level. A particular emphasis will be placed on scRNA-seq and scATAC-seq, two of the most popular single-cell techniques currently in use with the capability of obtaining information about cell state as well as mitochondrial genotype. Though we provide a succinct overview of scRNA-seq and scATAC-seq to contextualise the discussion with regards to mitochondrial DNA, we direct readers to the numerous excellent reviews covering the experimental and computational methods associated with these protocols [40, 41, 42, 22].

2.3 Background On Single-Cell Omics

Single-cell omics data is omics data derived from individual cells rather than tissue aggregates [43]. These techniques encompass single-cell RNA sequencing, epigenomics, immune profiling, copy number variation, and proteomics, among others. RNA sequencing quantifies the expression level of various genes and is among the most prominent of these techniques. RNA transcripts from a sample are captured, sequenced, mapped to the genome and then counted (see Figure. 2.1). Another prominent technique is Assay for Transposase-Accessible Chromatin (ATAC) sequencing, which measures which regions of the genome are accessible to regulatory elements, allowing for measurement of a component of the epigenome. We discuss these techniques in more detail below. Additionally, these techniques can be combined to obtain two or more sets of features from the same cells, referred to as *multi-omics* [44]. Prior to the development of high throughput single-cell techniques, most studies were conducted in bulk analyses, where numerous cells were homogenised and processed at once. Whilst such analyses have yielded useful insights [37], they are inherently unable to detect heterogeneity at the single-cell level meaning cellular states relevant to pathology went unobserved. The breakthrough of single-cell RNA sequencing changed this. The original inception of scRNA-seq was confined to studying 5 cells, though recent developments have greatly expanded throughput by several orders of magnitude [45, 46]. This has been enabled through the use of innovations which simplified and automated isolation of cells, as well as intelligent strategies for barcoding individual RNA molecules from different cells enabling multiplexed sequencing of RNA from numerous cells at once [47]. This development has resulted in massively parallel, commercially available techniques such as those offered by 10x genomics, which can allow for >100,000s of cells to have their transcriptomes sequenced simultaneously in a single experiment [46]. Other single-cell omics are following suit with developments in obtaining high throughput [48]. By leveraging these recent technological breakthroughs, large consortia such as the *Human Cell Atlas* and *LifeTime* aim to use sequences from 1,000,000s of cells to further understand human pathology at the single-cell level [20, 49].

The granularity combined with the scale of these techniques has had a major impact on many

areas of biology, such as immunology where scRNA-seq has been used to help define cellular states, observe heterogeneity in genomic sequences of cells derived from the same donor, and single-cell immune profiling can be used to measure T-Cell Receptor (TCR) sequences to perform lineage tracing and measure TCR repertoires [50]. A similar impact has been in tumour research where single-cell techniques have been used to probe heterogeneity throughout the tumour micro environment, and to perform lineage tracing of sub-clonal populations within the tumour [51, 52]. In contrast to these areas of biology, the use of single-cell omics data to probe single-cell mitochondrial heterogeneity and how it interacts with cellular phenotype remains relatively underdeveloped, even though they can provide coverage of the mitochondrial genome [53].

2.4 Sequencing Protocols & mtDNA

For researchers interested in discovering heterogeneity in mitochondrial heteroplasmy with single-cell data two primary tasks must be performed: 1) Identify real mitochondrial mutations; 2) Accurately assess the heteroplasmy of these mutations. This is achievable as single-cell sequencing data obtains sequences derived from the mitochondrial genome in individual cells. We focus on two popular modalities for obtaining such sequences, scRNA-seq and scATAC-seq due to their wide availability and because of their potential to combine observations of cell state with mitochondrial genotype. After giving a brief overview of these two techniques, we describe how the quality of the mtDNA data recovered from different modalities can be assessed and compared.

2.4.1 scRNA-seq

RNA sequencing has been around in excess of a decade, and is now ubiquitous in biological research. Figure 2.1 (i)-(l) illustrates some of the keys steps of this procedure in the context of single-cell genomics. In this technique, RNA is extracted from a biological tissue, complementary cDNA is synthesised by reverse transcription, then PCR amplified resulting in what is referred to as a *library of cDNA reads*. This library is then sequenced on a high throughput sequencer. Many variations of this basic procedure exist, and more have arisen since the emergence of *single-cell* RNA sequencing (scRNA-seq). Instead of aggregating and homogenizing many cells from a tissue together, scRNA-seq is capable of capturing, amplifying, sequencing, and then quantifying reads of individual cells separately, allowing for the the expression of each gene to be quantified on a per cell basis. This is done by obtaining tissue samples and separating individual cells (see Figure 2.1 (f)-(h)). The membranes of these cells can then be lysed, exposing the mRNA transcripts within. The identity of the cell from which a transcript originates is persisted throughout the sequencing protocol for later identification. As is discussed below, different protocols employ different methods for doing this, but appending sequence barcodes has proven to be among the most effective. The transcript counts per cell for each cell in a sample is typically then compiled into an array of data referred to as the

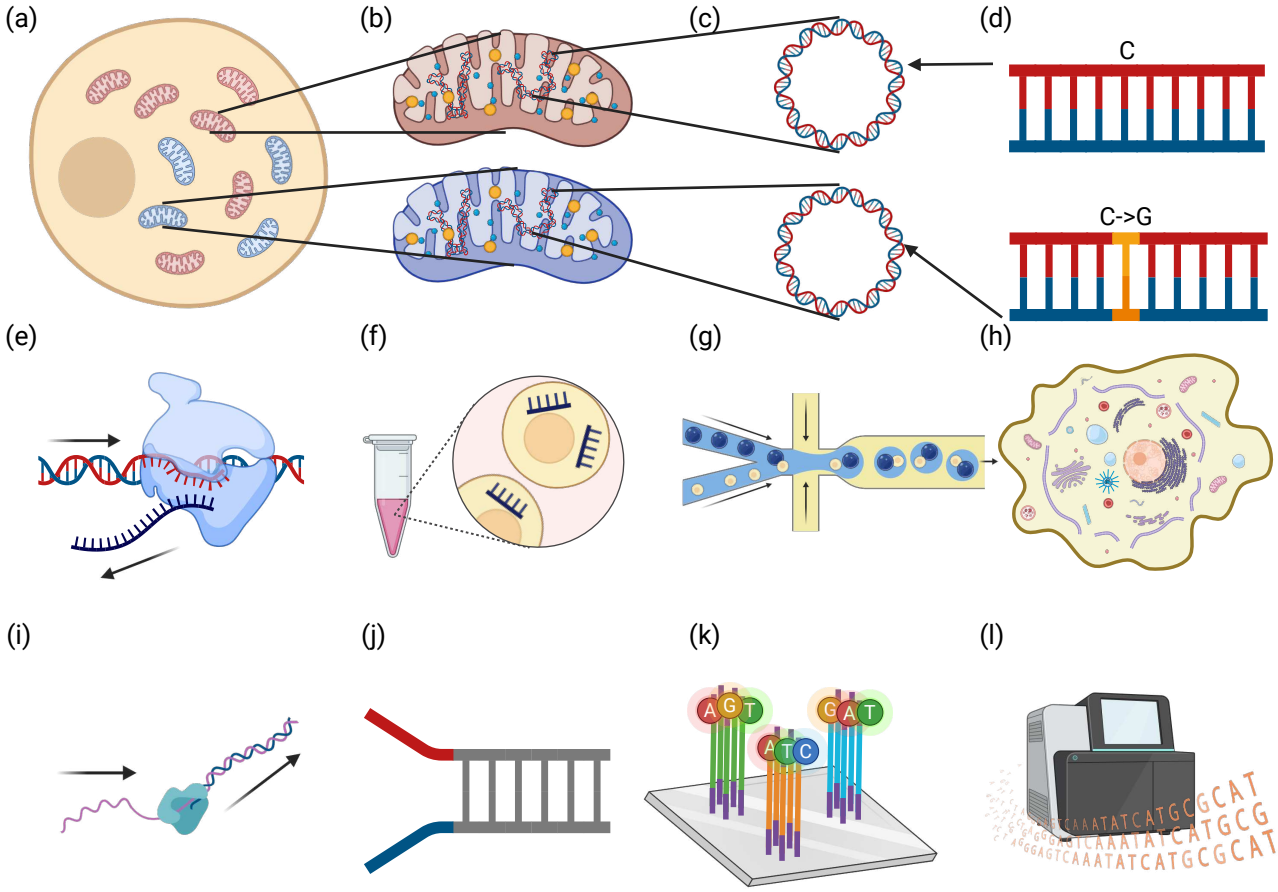


Figure 2.1: (a) The nucleus of the cell, depicted by the darkened yellow circle, contains the majority of a cell's DNA, inherited from both of the organism's parents. Additionally, many animal cells contain 100s-1,000s of organelles called *Mitochondria* which are crucial for the production of adenosine tri-phosphate (ATP), a vital energy transport molecule. (b) Mitochondria have their own distinct DNA, *mtDNA*, which is exclusively maternally inherited. mtDNA is typically bound up in aggregates called nucleoids inside the mitochondria (c) mtDNA is circular, and short compared to nuclear chromosomes at only 16,569 bps in humans. (d) Sequence variation can exist at a given position across different copies of mtDNA within a single cell. This is illustrated here by displaying a base having a mutation from C->G in one sequence that is absent in another. The percentage of mtDNA within a cell carrying a mutation at a given position is termed *heteroplasmy*. (e) The *central dogma of molecular biology* states that DNA is used to encode information for RNA, and RNA is then used to synthesise proteins. As part of this process, DNA is *transcribed* into messenger RNA (mRNA) transcripts by an enzyme called RNA polymerase. This mRNA is then transported throughout the cell for the later stages of protein synthesis. (f) To measure how active various genes are in this process, *transcriptomics* counts the number of transcripts originating from each gene. Single-cell transcriptomics allows this to be done individually for each single cell in a tissue sample. (g) Micro-fluidics techniques allow individual cells to be sorted and inserted into separate gel beads in large quantities, enabling high throughput sequencing [46]. (h) The gel beads contain enzymes for lysing a cell's membrane allowing the mRNA to become accessible. (i) mRNA is then *reverse-transcribed* into complementary DNA (cDNA). (j) In this process, cellular barcodes and unique molecular identifiers, made from artificially constructed DNA sequences, are attached to the ends of the cDNA. The next step of the sequencing process aggregates and replicates the cDNA molecules to allow for more effective sequencing, and such barcodes allow for individual molecules and their cells to be identified. Replication is achieved by the Polymerase Chain Reaction (PCR), a standard DNA amplification technique. The resulting collection of DNA molecules is called a sequencing *library*. (k) The library is sequenced by sequentially illuminating base pairs of DNA from each strand of DNA. (l) This data is then recorded digitally to be analysed downstream.

expression matrix which is then used for downstream analysis [22].

2.4.2 scATAC-seq

Single Cell Assay for Transposase Accessible Chromatin using sequencing (scATAC-seq) is another popular technique capable of obtaining sequence data from individual cells. Briefly, scATAC profiling methods make use of a hyperactive transposase Tn5 — an enzyme used to cut DNA — which binds preferentially to regions of accessible chromatin. Accessible regions of chromatin are regions of DNA that are not tightly bound by histones or other proteins, allowing transcription factors and other regulatory proteins to bind and initiate or regulate gene expression. The Tn5 enzymes efficiently cut DNA and perform the simultaneous ligation of adapters to the ends of the associated DNA molecules which can then be PCR amplified [42]. Regions of the genome which have accessible chromatin will then typically tend to have greater coverage than other areas of the genome. These differences in coverage can then be used to perform *peak calling* in which regions which have a far larger number of aligned reads than background are then used to infer the presence of accessible regions [42]. These accessible regions are associated with cis-regulatory elements which play a key role in transcriptional regulation, which scATAC-seq therefore measures.

2.4.3 The quality of single-cell omics mtDNA data

scRNA and scATAC are among the two most common forms of published single-cell sequencing data. Both modalities include sequenced data derived from the mitochondrial genome. Though RNA does not measure heteroplasmy directly, heteroplasmy in DNA should broadly be reflected in RNA sequences. Additionally, RNA modifications in mitochondrial genes could be another axis of variation which could be phenotypically relevant within single-cells [54]. The primary purpose of these two distinct approaches is to quantify gene expression and find accessible regions of the genome to quantify DNA regulation respectively. RNA sequences from scRNA-seq and DNA sequences for scATAC-seq enable the discovery of mitochondrial heteroplasmy heterogeneity by finding mismatches in these sequences from the reference genome, called *variant*

discovery.

A challenge posed by single-cell data however is the variation in data quality as measured by the metrics relevant to identifying mtDNA mutations and measuring their heteroplasmy at the single cell level. Due to inconsistent terminology surrounding quality metrics in the literature, we clarify and standardise our use of language surrounding the metrics used throughout this review. As is described in more detail below, sequence data comes in the form of many separate sequences called a *library* with each of the constituent sequences referred to as a *read*. Within a library, any given locus on the genome can typically be represented multiple times in separate reads. Having more sequence reads at any given position on the genome is desirable as it makes mutation detection more reliable by protecting against sequencing errors and enables more accurate estimation of the heteroplasmy of mutation at a given position. Building on the clear terminology from Sims. *et al* (2014) [55], we use the following definitions: *Depth of coverage* is a property of a single site on the genome and is the number of times a base in the reference genome is covered by a high quality read; *Breadth of coverage* is a property of a target sequence on the genome and is the proportion of bases on the target sequenced to a desired depth of coverage. In this chapter, the primary target genome is the mtDNA. Moreover, as we are interested in cellular heterogeneity, we are interested in distribution of these quantities *per cell*, rather than aggregated at the tissue level. Both quantities also depend on parameters such as sequencing base quality thresholds, in which some positions from a read are neglected due to low quality sequencing.

Figure 2.2. illustrates these definitions with data from two contrasting scRNA-seq data types, showing the breadth and depth of coverage of the mt-ND1 gene in two human data sets. Using mt-ND1 as the sequence of interest, the top figure shows the number of reads in the data which map to each position in the gene. 10x data (blue) shows high mean depth of coverage per cell for positions > 4000. For positions less than this, 10x data shows shallow depth of coverage, and can therefore be described as having narrow breadth of coverage restricted to the 3' end of the transcript. Smart-Seq2 data in contrast to this has superior mean depth of coverage across almost the entire length of mt-ND1 and can therefore be said to have superior breadth of coverage to 10x sequencing data. The intervals around the mean emphasise that

these definitions are made per cell, with considerable variation in both depth and breadth of coverage which can be attributable to biological signal, such as transcriptional variation, as well as technical stochasticity in detection.

Figure 2.2 (b) shows how the breadth of coverage per cell declines across the whole mitochondrial genome as the demand for depth of coverage increases. Almost every base in the mitochondrial genome is covered by at least a low numbers of reads in some cells. 10x data shows a rapid decline in the breadth of coverage as the depth threshold is increased, whilst Smart-seq2 data shows greater stability.

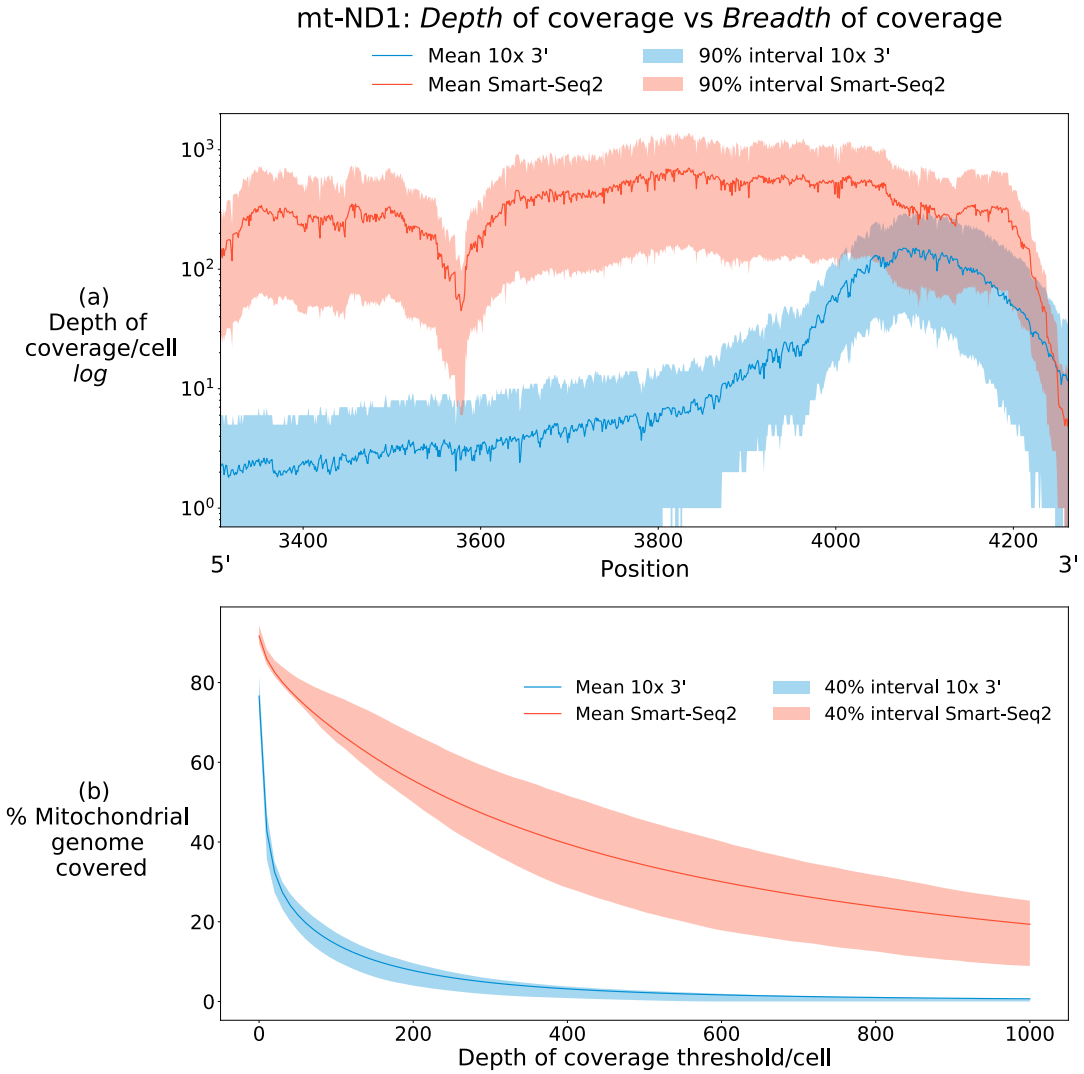


Figure 2.2: Breadth of coverage and depth of coverage differ between sequencing types. (**a**) Here, both meanings of coverage are illustrated for the mt-ND1 gene from two scRNA-seq data types [56, 57]. It can be seen that 10x 3' data (blue) has a high mean depth of coverage at the 3' end of the transcript, with many sequenced reads aligning to this position. The mean depth of coverage rapidly declines at bases with positions roughly $\leq 4,000$ moving in the 5' direction, behaviour typical of 10x 3' data (see 2.4.1). 10x data is therefore said to have a narrow *breadth* of coverage of the mt-ND1 transcript. Smart-Seq2 (red) in contrast has a higher mean depth of coverage across almost the entire length of the transcript and can therefore be described as having both high depth and breadth of coverage. Central 90% quantiles demonstrate the heterogeneity in both coverage measures across cells, with some cells possessing far less depth of coverage in certain regions, posing a challenge for robustly identifying mutations across the mitochondrial genome in all cells. (**b**) Smart-seq2 shows superior breadth of coverage of the mitochondrial genome at all depth thresholds. This difference reflects the 3' bias of 10x scRNA-sequencing data, limiting its breadth of coverage at modest depth thresholds, where Smart-Seq2 maintains much more stable coverage up to high thresholds. Both methodologies however show reductions in the proportion of cells attaining large breadth of coverage as depth thresholds increase. Central quantiles illustrate the distribution around the mean. 40% quantiles are selected for a more informative visualisation, as lowest quantile cells typically have a very low proportion of the genome covered.

Full length transcriptomics

Full length protocols sequence the entire length of a transcript. There are two typical approaches to full length transcriptomics. One method, followed by popular approaches such as the Smart-Seq2 protocol [58], is to fragment a library of full length cDNAs into shorter molecules followed by short read sequencing on a short read Illumina sequencer. Another approach is to use long read sequencing platforms such as Oxford Nanopore or Pacific Biosciences. Long read length sequencing is advantageous in that the full sequence of a gene's RNA is obtained, which is particularly useful for detecting variants across the length of a gene. The primary drawback of direct long read protocols is that they are low throughput and costly to perform [40]. This is because cells are typically placed into separate wells on a plate and sequenced separately, bounding their throughput. Importantly, the long read length sequencers offered by Oxford Nanopore or Pacific Biosciences suffer from higher error rates than short read Illumina sequencers, posing problems for analysis at the single nucleotide level as might be done to investigate mitochondrial heteroplasmy [59]. The ability to sequence the entire length of transcripts however is attractive as it has the potential for discovering variants across most of the mitochondrial genome at a cellular level.

3' and 5' droplet based transcriptomics

Alternatives to standard full length approaches are 3'and 5' methods in which reads are reverse transcribed from the 3' or 5' end of a transcript — 3' and 5' being the standard labels for each end of a DNA molecule. These approaches do not make use of fragmentation in the same manner as Smart-seq2, thereby resulting in only one cDNA molecule per transcript prior to PCR. In contrast to full length approaches, 3' and 5' protocols have been able to achieve immense throughput, and high accuracy, being sequenced on an Illumina sequencer. This has largely been achieved by microfluidics technologies which capture single-cells in oil droplets containing cellular barcodes which are attached to each read (see Figure. 2.1 (f)-(j)). These barcodes are randomised, artificial sequences of DNA that are appended to the cDNA reads which can then be later sequenced along with the rest of the DNA read. This barcoding protocol

allows reads from many separate cells to be pooled together (referred to as *multiplexing*), PCR amplified and then sequenced, improving throughput. Reads can then be de-multiplexed by using software that exploits the pre-appended barcodes identifying each cell. In other words, intelligent barcoding strategies for cell identification enable massive scaling up of throughput. Such an approach is routinely used by three of the most popular scRNA-sequencing platforms—inDrop, Drop-seq, and 10x Chromium—and can be easily used to sequence 100,000s of cells in a single experiment [60, 61, 46]. Additionally short read technologies enable the use of *unique molecular identifiers*.

Unique molecular identifiers

In protocols using unique molecular identifiers (UMIs), each individual cDNA read has a random nucleotide sequence attached during the reverse transcription process, with such random nucleotide sequences being referred to as UMIs. Each resulting cDNA molecule therefore has both a cell barcode and a unique molecular identifier attached. Attachment takes place during the first step of library preparation, and importantly before PCR amplification. cDNA libraries undergo multiple rounds of PCR amplification, resulting in duplicated copies of every original molecule from the cell, and has been shown to result in PCR bias, in which some molecules are amplified more than others [62]. This is significant for researchers who wish to estimate the proportion of sequence molecules (RNA or otherwise) carrying a certain variant as it is possible that variation induced by PCR could be a source of error in the resulting values of variant allele fractions. This could be particularly relevant for mitochondrial researchers, for whom accurate heteroplasmy measurements are important. By allowing groups of PCR amplified reads to be grouped together based on their UMI, reads can be de-duplicated to enable robust estimation of the original number of transcripts from each gene, and is the basis of the output from tools like STARSolo and CellRanger which give a UMI counts based expression matrix to quantify gene expression [63]. Non-UMI approaches rely on read counts without any correction [64]. Though the use of UMIs to form robust counts has been addressed by software tools for sequence counting such as UMI-tools [65], the ability to form a consensus read sequence based on all the copies of a molecule remains nascent, with numerous computationally intensive approaches

having been suggested, but few used broadly.

Such UMI approaches are typically limited to short read sequencing, in which only positions within 100s of bases of the 3' or 5' end of a transcript are covered, whilst full length approaches have typically been missing UMIs. Smart-Seq2 like approaches prevent UMI usage due to the fragmentation step which would require linking every fragment to the source UMI, something which remains challenging. Direct full length sequencing approaches using Oxford Nanopore or PacBio also inhibit this approach due to the high error rate of long read sequencers (nearly 50% of UMIs can have at least one error in these approaches) which inhibits robust UMI grouping as sequence errors can create spurious UMIs which did not tag any original molecule. Additionally, long read sequencers typically obtain fewer reads per UMI than short read sequencers further compounding the issue of accurate UMI usage. Efforts are being made to address these obstacles [66, 67, 68]. Smart-seq3 has recently been developed as a full length UMI protocol, though it has not yet attained widespread usage [69]. Such problems are not associated with short read Illumina sequencers which have a low error rate enabling accurate barcode-UMI sequencing. Whilst this is the conventional approach to sequencing libraries prepared with 10x technologies [46], it does entail fragmenting prepared cDNA reads into two short sequences—one containing the barcode and UMI sequences, one containing the genomic sequence—of around 100—200 bps resulting in substantially less breadth of coverage of the genome than a full length approach.

Comparing full length and 3'/5' for mitochondrial heterogeneity discovery

We can understand the quality of these two approaches for discovering mitochondrial heterogeneity by using data from Smart-Seq2 and 10x Chromium platforms as exemplars [57, 56]. Figure 2.3 shows the depth of coverage across the mitochondrial genome on a per cell level for each of these approaches. Smart-Seq2 displays consistently high depth of coverage across the mitochondrial genome. In comparison, 10x data shows far lower breadth of coverage with consistent enhancements in depth of coverage near the 3' end of each gene. Whilst this non-uniform depth of coverage might prevent identification of variants across the mitochondrial genome variants should be consistently identifiable within close proximity to the 3' end.

We summarise the trade-off between popular full length sequencing protocols and those which use 3'/5' droplet based approaches as follows: Full length approaches typically maintain a superior breadth of coverage, but lack UMIs and can have a higher sequencing error rate when used in conjunction with long read sequencing. In contrast, 3'/5' droplet based approaches incorporate UMIs into their analysis, meaning that biases and errors arising from PCR can in principle be ameliorated, and sampled transcripts can be accurately counted. Cells sequenced with these 3'/5' library approaches however show great heterogeneity in precisely which bases attain a sufficiently high depth of coverage for accurate variant calling in any given cell. Furthermore, the breadth of coverage of 3'/5' droplet based approaches is circumscribed, with high depth of coverage localised around the corresponding transcript end due to the end bias of these approaches, leaving a lot of the genome with insufficient depth of coverage to inspect heteroplasmy [70]. Comparing variants across different cells using such approaches should be possible, but requires care to account for coverage variation. Comparing variants across different cells using such approaches should be possible, but requires care to account for coverage variation.

2.4.4 High-Throughput scATAC-seq

The use of cellular barcoding strategies has enabled the development of high throughput droplet methodologies similar to those used for RNA sequencing, and there are also commercially available platforms such as those offered by BioRad and 10x Genomics [71, 48]. Unlike scRNA-approaches, UMI methodologies are not as mature and are not as widely used. scATAC UMI based approaches are however being developed to mitigate amplification biases. Standard approaches de-duplicate reads based on the start positions of where reads align to the genome are established [72], however this can induce further biases. By making use of UMIs, one study was able to quantify how many Tn5 insertions may independently occur at the same position [73]. This showed that up to 20% of the reads discarded based on their alignment positions, a standard method step in many protocols, were done so erroneously and actually arose due to independent Tn5 insertion events. As de-duplication alters the counts of reads which carry variants or not, this approach could bias heteroplasmy estimates.

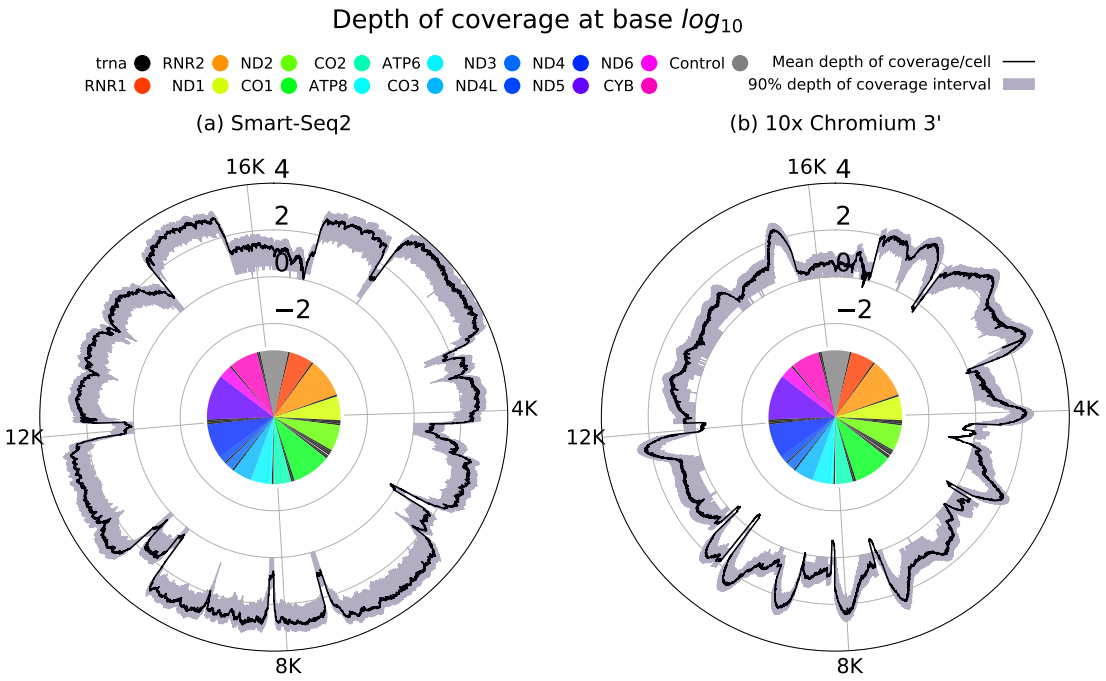


Figure 2.3: Sequencing protocols demonstrate a high degree of variation in both depth and breadth of coverage they attain across the mitochondrial genome. (**a**) Full length Smart-Seq2 data [57] demonstrates broadly uniform coverage across transcript lengths. 10x 3' protocols achieve a far higher throughput of cells, than full length approaches. (**b**) However, 10x 3' data [56] has a poorer breadth of coverage of the mitochondrial genome due to reads preferentially capturing the 3' poly A tails of transcripts. This results in the pronounced depth of coverage which can be observed in the 3' end of each gene. High throughput combined with consistent 3' coverage could enable 3' data to facilitate the discovery of mitochondrial heterogeneity. Such coverage plots can be generated by counting the reads aligned to each position in the mitochondrial genome in each individual cell.

In early ATAC-seq approaches mitochondrial reads accounted for an overwhelming proportion of aligned reads. This is because nuclear DNA regions only come in 2 copies in any cell, whilst mtDNA comes in 1,000s of copies, and each mtDNA is accessible as they are not bound up in a nucleosome. Consequently, a lot of the signal from the nuclear DNA was considered to be drowned by the "noise" of mtDNA reads, and efforts were taken to suppress the coverage of the mtDNA in these sequencing approaches. More recent high throughput single-cell protocols for ATAC sequencing typically have reduced mitochondrial content as these protocols isolate individual nuclei, discarding the rest of the cell [48]. Despite this, these approaches still offer the potential to probe mitochondrial heterogeneity as mitochondria have been closely associated with the peri-nuclear sheath, meaning total isolation is unlikely to occur [74].

2.5 Heteroplasmy In Single-Cell Data

In this section we review previous work that has been undertaken to probe mitochondrial heteroplasmy in single-cell data. This centres on the work of three previous papers, one that analysed conventional scRNA-sequencing protocols to perform lineage tracing, and two that developed new protocols to gain greater insight into the mitochondrial variants at the single-cell level.

Whilst some studies have investigated mitochondria in single-cells their throughput has been limited, eg [75]. Others studies seek to inspect mutants known to be pathogenic, or have required producing clonal populations of cells [35, 76]. The focus of the current chapter is on approaches which offer both high throughput and scope for discovering previously unobserved heteroplasmy from *in vivo* samples. Each of the following papers focuses on using mitochondria for lineage tracing, and all demonstrate how mitochondrial heterogeneity in both the mutations found and their range of heteroplasmies can be discovered using these techniques.

2.5.1 Mitochondrial Heterogeneity With scRNA-seq And scATAC-seq

In vivo lineage tracing using single-cells conventionally relied upon the detection of nuclear somatic variants. By performing variant calling in scRNA-seq and scATAC-seq data, Ludwig. *et al* (2019) showed that mitochondrial sequences could in certain circumstances be used as natural genetic barcodes for tracing cellular lineages [77]. Lineage tracing is the task of identifying cells descended from a single progenitor. If an ancestral cell harbored a particular mutation, its progeny would then share this mutation. The authors argue that mitochondrial sequences are short enough to allow for cost effective sequencing, yet are long enough to harbour numerous such mutations. As such, mitochondrial sequences offer enough genetic diversity to serve as an effective barcode that can differentiate between cellular lineages long enough to be present in multiple cells and aided by the higher mutation rate of mitochondrial sequences to display the genetic diversity—i.e. enough available sites for mutations to occur—required to

serve as an effective barcode that can differentiate between cellular lineages. Grouping cells by the mitochondrial mutations that are shared between them should enable the inference of lineage.

They compare the percentage (i.e. breadth of coverage) of the mitochondrial genome, attaining different levels of depth of coverage for a range of single-cell RNA sequencing techniques. Short read, 3' approaches show considerably less of the mitochondrial genome being sequenced at any given level of depth of coverage, as would be intuitively expected (see section 2.4.3). Smart-seq2 has the greatest breadth of coverage at any fixed level of depth compared to the other approaches they analysed. By using a scATAC-seq derived from whole cells, rather than nuclei, using an earlier approach [78] they show that this modality gives uniform depth and breadth of read coverage across the mitochondrial genome, making it a suitable candidate for obtaining reliable heteroplasmy data from single-cells data.

Interestingly, the authors also present a comparison between the heteroplasmies observed in joint WGS and scRNA-seq data from single-cells. For the variants found in both datasets there appears to be a high level of agreement, demonstrating the utility of scRNA-sequencing approaches to indirectly discover mutations in the mtDNA as well as their heteroplasmy. However there are some variants which appear to be specific to the RNA modality. The authors claim that some of these are attributable to RNA editing, and transcription errors. However, some of the low heteroplasmy RNA variants are likely attributable to errors in RNA sequencing arising from a combination of reverse transcription errors and errors occurring during early PCR. The former errors are absent for WGS data, highlighting a source of false positive variants in scRNA data. High heteroplasmy RNA variants are unlikely to be due to these systemic errors as the chance of reverse transcription errors, PCR artifacts, or sequencing errors independently producing errors at the same position in numerous reads, as would be required for an observed high heteroplasmy, is unlikely to occur by chance. High heteroplasmy variants are therefore more robust to multiple sources of error.

In vitro, the authors performed variant calling across cells from 64 different erythroid and myeloid colonies— two cell types that arise from hematopoietic stem cells (HSCs) in the bone

marrow during the process of hematopoiesis, the production of blood cells — profiling between 8 and 16 cells in each colony. The donor and colony of origin was known, enabling a supervised analysis to be performed. Variants observed at near homoplasmy were capable of separating cells into their donor of origin as well as singling out individual colonies. It was observed that unique clonal mutations were found in many of the colonies, at a range of heteroplasmies capable of differentiating many of the colonies. They made similar observations using a scATAC-seq dataset published by Buenrostro. *et al* [78]. Both insights demonstrate how heteroplasmic mutations are propagated through colonies. Using publicly available T-cell, Smart-Seq2 [58] scRNA-seq data and TCR-seq data [79, 80], Ludwig *et al* were also able to demonstrate similar findings *in vivo*. T-cell receptors (TCRs) are specialized proteins found on the surface of T-cells. Each T-cell has a unique TCR sequence. TCRs are also highly variable, differing from cell to cell. This diversity is generated through a process called V(D)J recombination, which rearranges gene segments coding for the TCR. Upon activation by an immune stimulus, which the TCR is responsible for detecting, T-cells with the appropriate TCR divide and proliferate, forming a clonal population that shares the same TCR. Grouping T-cells by their TCR-receptor allowed for a similar supervised analysis to take place, in which mitochondrial mutations were shared among cells sharing TCR sequences. Some mtDNA mutations were capable of further refining the clusters identified by identifying sub-populations within TCR groups, whilst some mtDNA mutations were shared across TCR groups, suggesting a common ancestor before V(D)J rearrangement took place.

In addition, by using somatic mitochondrial mutations first identified in bulk RNA-seq Ludwig. *et al* further elucidate clustering of cells derived from colorectal adenocarcinoma primary tumors when using scRNA-seq to further resolve which cells did and did not possess these mutations highlighting the ability of mtDNA mutations to elaborate intratumoral clonal heterogeneity. A dimensionality reduction technique called t-distributed stochastic neighbor embedding (t-SNE), a dimensionality reduction technique commonly used for visualizing high-dimensional data, showed nearly perfect separation of cells into their originating donors when clusters were based on mitochondrial genotypes as features. By using publicly available chronic myelogenous leukemia scRNA-seq dataset [81], t-SNE clusters based on mitochondrial genotypes as features

showed nearly perfect separation of cells into their originating donors. Finally analysis was performed of 3' 10x chromium using the PBMC data in the original 10x chromium publication [46]. These PBMCs were taken from a recipient before and after transplantation of HSCs. Two homoplasmic mutations were observed to distinguish the donor and recipient, and post transplant PBMCs from the recipient were found to be overwhelmingly derived from the donor, demonstrating the potential of 3' approaches to detect heterogeneity in spite of their narrow breadth of coverage. These observations highlight the potential for finding biological insight regarding heterogeneity in heteroplasmy at the cellular level through the use of single-cell data.

2.5.2 EMBLEM

Epigenome and Mitochondrial Barcode of Lineage from Endogenous Mutations (EMBLEM) which focuses on single-cell ATAC sequencing to perform lineage tracing using mitochondrial variants, was a strategy presented by Xu. *et al* [82]. Just as lineage tracing in scRNA-seq allows users to get cellular lineage information and cell state information from the same cells, EMBLEM aims to let users obtain epigenomic state alongside clonal genotypes.

As noted by the authors, observing a mutation at $x\%$ heteroplasmy in bulk could arise due to $x\%$ of cells harbouring that mutation at homoplasmy or all cells separately harbouring that mutation at $x\%$, as well as other combinations between these extremes. Inspecting the variant allele fraction of different mutations in single-cells using using scATAC-seq, Xu. *et al* saw that different mutations occupying different positions along this spectrum with some mutations being present in many cells at low heteroplasmy, whilst other mutations are found at high heteroplasmies in single-cells.

Leukemic stem cells (LSCs) — stem cells responsible for the initiation and maintenance of Leukemia — and blast cells are both found in the human bone marrow. By sequencing LSC and blast cell populations Xu. *et al showed* that they shared mitochondrial mutations at similar variant allele fractions. This is coherent with the established hierarchical lineages in which LSCs give rise to blast cells and should therefore be expected to share mitochondrial mutations [83]. Taking a sample from a single patient they found several cells that shared the

same four mutations, with subsets of these mutations contained in many other cells. Combining this observation with the fact that all cells are derived from the same progenitor cell and should therefore share the same mutations, the authors aimed to quantify the proportion of cells for which a mutation would be observed in as function of mitochondrial read depth. Mutations with moderate $\sim 20\%$ heteroplasmies could be detected in $\sim 90\%$ of cells with only 20 reads covering the position of that mutation. Conversely, when a mutation has heteroplasmy $< 1\%$, more than 100 read depth can be required to obtain a detection rate exceeding 90%, reflecting the need for greater sampling to detect low heteroplasmy variants. Moreover, without deep sampling, such low heteroplasmy mutations would be challenging to distinguish from sequencing error.

These observations highlight the level of confidence which mitochondrial variants can be called with accuracy, showing that near homoplasmic variants can be called robustly and discovered even when read depth is low at a position. This has implications for detecting mtDNA mutations in other sequencing types, such as 10x 3', where depth of coverage at different bases of the mitochondrial genome is heterogeneous both between cells and across the genome due to the variability of 10x coverage at the cellular level (see Figure 2.3 (b)). In such sequencing types, high heteroplasmy mutations should be observable in regions with low depth of coverage away from transcript ends, whilst regions of the genome with high depth of coverage should be able to discover mutations with low heteroplasmy.

2.5.3 mtscATAC-seq

Building on the earlier work by Ludwig. *et al* (2019) [77], a new scATAC-seq protocol was developed by Lareau. *et al* (2021) to perform mitochondrial genotyping in single-cells [84]. Ludwig. *et al* (2019) showed the limited breadth of coverage of the mitochondrial genome in high throughput 3' sequencing assays such as Drop-seq. They show similar shortcomings for the massively parallel and massively popular 10x chromium system when compared to a full length, low throughput assay like Smart-Seq2 [77]. In order to capture the best of both of these approaches, that is to say, a high throughput approach with uniform coverage of the mitochondrial genome, Lareau developed mitochondrial single-cell assay for transposase-

accessible chromatin with sequencing (mtscATAC-seq). As mentioned previously, most scATAC protocols have been developed to work with single nuclei, which depletes the mitochondrial coverage. By modifying scATAC-seq the authors develop an assay that works with whole cells.

This protocol is a modification of the scATAC-seq system offered by 10x genomics. Usually this uses pooled nuclei before adding the Tn5 enzyme. Instead of this Lareau. *et al* use "mild lysis" of whole cells in combination with fixation. This enables cells to take up the Tn5 enzyme whilst preventing the leakage and cross contamination of mitochondria between between cells in suspension. After masking NUMTs (see section 2.6.6 below) they find that their approach shows remarkable uniformity across the mitochondrial genome, with residual coverage variation reflecting stochasticity in PCR and Tn5 insertion variability. They show that their approach leads to an approximately 20 fold increase in the mean depth of coverage of the mtDNA genome compared to the original protocol achieving an mean read depth of 191 across the mitochondrial genome, across cells. Their method in general aimed to have around 20 reads after PCR duplicate removal.

One of the key features of their work is the development of Mitochondrial Genome Analysis Toolkit (mgatk). In contrast to some variant callers which have aimed at genotyping single-cells, mgatk aims to analyse clonal mutations. CellRanger is a popular computational package used to process the outputs of 10x sequences, with a Binary Alignment/Map (`.bam`) being a standard outputs. The `.bam` file stores information about how sequencing reads map to a reference genome. Taking the `.bam` file output from CellRanger, mgatk produces a matrix of de-duplicated per-cell, per-strand count of all alleles at all positions in the genome. Here, de-duplication groups all reads by their starting position all the reads sharing a starting position, the read with the highest mean base quality is then used as a representative read. This de-duplication procedure should enhance heteroplasmy estimation by counteracting PCR amplification biases. ATAC-seq obtains reads from both the forward and reverse DNA strands. For each mutation mgatk measures the correlation between the allele counts on both strands across the dataset of cells and is called *strand concordance* which helps to protect against photobleaching effects (see section 2.6.2). Only mutations with a high strand concordance and variance mean ratio in their heteroplasmies are retained for further analysis. mgatk then counts

all cells where the variant was detected on both strands in at least two reads. Lineage variants are considered as high confidence for lineage tracing if the variant is detected in 5 or more cells.

Using this variant calling approach, Lareau. *et al* were able to observe mitochondrial heterogeneity in a range of phenomena including identification of clonal subgroups using mitochondrial mutations in TF1 cells — a cell line derived from a patient with Erythroleukemia. Furthermore, they were able to refine clonal substructure in PBMCs and measure an association between cell's chromatin profiles and their clonal sub-grouping, thereby linking genotype and phenotype. Additionally, observations were made of the level of heteroplasmy ranging from 0-100% in the 8344A>G mitochondrial mutation among cells from an individual suffering from the chronic mitochondrial disease, myclonic epilepsy with red ragged fibers (MERRF) [85, 86], which this mutation is associated with. In adjacent work by Walker. *et al* (2020), the mtscATAC and mgatk protocol was applied to PBMCs from 3 patients suffering from mitochondrial encephalomyopathy with lactic acidosis and stroke-like episodes (MELAS), commonly associated with the 3243A>G mutation [36]. Of all the cell types present, T-cells consistently had this mutation at lowest heteroplasmy across all 3 donors, showing how single-cell approaches can discover mitochondrial heterogeneity among different cell types. These examples highlight the potential of high throughput approaches to discovering mitochondrial heterogeneity.

2.5.4 MAESTER

A recent pre-print by Miller. *et al* [87] develops a high throughput scRNA-seq approach. As previously discussed, Smart-seq2 is able to offer far superior breadth of coverage to the 10x 3' platforms, making it superior for discovering variants that can be used for lineage tracing, yet is limited by its throughput. The authors present Mitochondrial Alteration Enrichment from Singlecell Transcriptomes to Establish Relatedness (MAESTER), which seeks to address this by enriching full length mitochondrial cDNA which is produced in an incallediate step by platforms like 10x but are usually only sequenced to partial length. Complementing this is the Mitochondrial Alteration Enrichment and Genome Analysis Toolkit (maegatk), an extension of mgatk. This exploits UMIs more extensively than previous approaches. By grouping all reads

sharing the same UMI together, maegatk creates a consensus read by using the most common nucleotide at every position in conjunction with the base's quality. This should ultimately protect against modified sequencing and PCR errors as well enhancing the accuracy heteroplasmy estimates.

Though early in its development, MAESTER offers the exciting potential to discover mitochondrial variants across a high proportion of the mitochondrial genome, along with accurate heteroplasmy estimates combined with transcriptional cell state.

2.6 Future Directions And Open Problems In Assessing Heteroplasmy At The Single-Cell Level

Having explored recent developments using single-cell to explore mitochondrial heterogeneity, we now provide a survey of challenges, limitations, and open problems in these approaches.

2.6.1 PCR and heteroplasmy

As we have noted throughout this review, the heteroplasmy of different mitochondrial mutations at the single-cell level form a key source of variation, with potential implications for cell state. As such, obtaining accurate estimations of heteroplasmy is important to be able to fully understand this. A simple approach to quantify heteroplasmy for each cell is to go through each position in the mitochondrial genome, count the number of reads which display an alternative allele at that position, and divide this by the total number of reads at that position [77, 88]. This simple approach has proven capable of detecting real biological variation, as corroborated by orthogonal pieces data such as TCR recepetors discussed earlier. However, as sequencing reads are the result of PCR, the heteroplasmy of input biological molecules could be distorted by variation in the degree of amplification of different input molecules.

Approaches to address this amplification bias have used de-duplication of reads, in which all PCR duplicates are grouped together and one read is picked to represent those reads. mgatk does this by using GATK's picard tools [72] to group reads by their starting position, as it is thought that reads derived from different biological molecules are unlikely to share a starting position by chance. Among reads sharing this starting position, the one with the highest mean base quality is chosen as a representative. Recent work has however shown that different reads all derived from the same starting biological molecule may not share the identical starting position due to *PCR stutter* (see Sena *et, al* (2018) for more details [89]). Grouping reads by there starting position could therefore also bias heteroplasmy quantification.

Building on mgatk, a recent pre-print by Miller *et, al* (2021) developed the Mitochondrial Alteration Enrichment and Genome Analysis Toolkit (maegatk) [87]. This toolkit is designed

to work with UMI data by grouping all reads sharing the same UMI together and forming a consensus call at every position along the transcript. This approach should both mitigate sequencing errors and help identify accurately the heteroplasmic proportions of input cells.

2.6.2 Photobleaching And Strand Concordance

Intuitively, mitochondrial mutations are present on both DNA strands and should be present at comparable levels in DNA based sequencing data. Lareau. *et al* suggest that strand concordance breaks down due to a technical confound which recurrently generates sequencing errors in certain regions of the mtDNA genome on one DNA strand due to the effect of surrounding G's on successive cycles, a form of sequencing bias [84, 90]. The complementary strand would instead having surrounding Cs. This bias is therefore reflected and corrected by observing strand concordance, a measure of agreement between the heteroplasmy observed on both strands of a sequence across cells in a dataset. Mutations with concordance below the threshold are then discarded to protect against the photobleaching effect.

As a default, mgatk uses these statistics of variants found across multiple cells to identify likely sub-clonal variants. Lareau. *et al* note that this approach is valid not only for mtscATAC, but also full length scRNA-seq approaches such as Smart-Seq2 in which the strandedness of RNA transcripts is lost during double stranded cDNA synthesis [58]. Therefore observing stand concordance should still counteract photobleaching in these approaches. In reanalysing some of the data from their previous study [77], they in fact show that one variant which they had previously identified actually demonstrated strand discordance, only being present in one strand. However, they do show that a strong degree of overlap between the variants that they called using their old approach and with mgatk factoring in strand concordance and VMR, suggesting that variants may still be called with some degree of accuracy without factoring in strand concordance given other rigerous quality control thresholds.

As 3' approaches retain the complementary strandedness of the original transcript molecule, strand concordance cannot be calculated, thus the authors state that mgatk unsuitable for identifying high quality lineage variants in this datatype. The software does however still offer

the option of genotyping single-cells obtained from droplet based UMI sequencing approaches. In conjunction with suitable quality control thresholds, this approach could yield heterogeneity at the per cell level regardless of the ability to measure strand concordance in 3' data. This underscores the technical considerations that have to be factored into answering any research question that may arise from variant calling scRNA-seq data.

2.6.3 Mitochondrial Cellular Transfer

It has been observed that mitochondrial transfer between cells is a common phenomenon, occurring in multiple contexts [91, 92, 93]. Such observations potentially undermine the use of endogenous mitochondrial mutations as *in vivo* barcodes for cellular lineage tracing. Xu. *et al* applied EMBLEM to scATAC-seq data from a previous experiment which entailed the mixing of human and mouse T-cells [94]. They found that species-specific mtDNA and nuclear DNA always paired appropriately to the correct species. This, they argue, demonstrates that inter-cellular mtDNA transfer does not occur universally between cells. However, it remains unclear the extent to which non-observations of mitochondrial transfer between T-Cells of different species generalises to eliminate mitochondrial transfer as a potential confounder for lineage tracing studies in other contexts.

Simulations by Ludwig. *et al* suggest that rates of intercellular transfer between cells would need to be high in order to for the transferred mutations to confound the analysis, as variant calls are typically done at a level to exclude low heteroplasmy variants [77]. Furthermore, the success of these lineage tracing approaches in a supervised context is also suggestive of the limited extent of mitochondrial transfer in some settings. However, more work needs to be done to conclusively eliminate mitochondrial transfer as a potential source of error in lineage tracing studies in other tissues.

2.6.4 Ambient RNA In Droplet-Based Approaches

Droplet based approaches to single-cell data entail pooling cells together in a single suspension [46]. Cells are then separated into their separate droplets. Standard subsequent analyses

assume that all reads originate from a single-cell. Common violations of this assumption are empty droplets, in which no cell is found and doublets in which two cells made it into the same droplet. Popular alignment tools such as STARSolo and CellRanger have approaches for identifying likely empty droplets and software is being developed to handle doublets [95, 96]. A less well known violation of the standard assumption is *ambient RNA*, which is cell free RNA in the original cell suspension which can accompany individual cells into their droplets. Such ambient RNA could arise from other cells that were lysed or damaged within the suspension. Tools have been developed to correct the expression matrix in light of such observations [97, 98]. Species mixing data suggests that 1-2% of observed transcripts within a cell could in fact be ambient RNA. This suggests a technical limit on attempts to elucidate clonal substructure of tissues by using low heteroplasmy variant calls, which could mistake ambient RNA shared in the suspension for a clonal variant.

2.6.5 Mitochondrial Gene Expression Proportion Quality Control

In the analysis of single cell expression data, best-practice recommendations include a quality control step to filter cells based on the proportion of mitochondrial reads or transcripts they have [22]. Both the tutorials for Seurat and scanpy, two popular single-cell analysis software packages, use a 5% threshold as a default [99, 100]. This originated as a standard procedure for distinguishing low quality cells from high quality cells, where it was observed that broken cells with a loss of cytoplasmic content were observed to have a much higher proportion of mitochondrial reads of cells than obviously intact cells [101]. Whilst claiming to be robust across cell types, such observations were predicated on on T-cells, dendritic cells, and mouse embryonic stem cells. Bulk studies of the much more energy intensive tissue of the heart have found mitochondrial read proportions in excess of 30% of all reads, with considerable heterogeneity found across tissue types [102]. Moreover, a systematic study of single-cell data comparing mouse and human data suggests that different thresholds might be appropriate for different species and tissues [103]. Therefore researchers should take care to adjust their thresholds according to the subject of their study.

Whilst establishing reliable quality control thresholds might be challenging, they pose another problem for researchers interested in exploring the heterogeneity of heteroplasmy in single-cells. Beyond lysis, another rationale for the removal of cells with high mitochondrial count proportions is the removal of potentially *stressed* or pre-apoptotic cells, as retaining such cells would result in clustering cells by their stress state which researchers are commonly uninterested in [104]. As such they are often removed from the analysis by default. However mitochondrial mutations [105], as well as mitochondrial proliferation and copy number [106, 107, 108] have been associated with apoptosis. When paired with mitochondrial threshold effects [31] it is plausible that the default processing of scRNA-seq data in which discarding highly stressed are a source of heterogeneity might systematically bias conclusions researchers draw in multiple ways. The assumptions underlying such a quality control process are not fully explicated in standard tutorials, and the extent to which the standard approach to these analyses might be a source of confounding in is not well understood.

2.6.6 NUMTs And Heteroplasmy

Nuclear sequences of mitochondrial origin (NUMT) are sequences found within the nuclear genome which have a homologous relationship and therefore high sequence similarity to sequences of the mitochondrial genome. NUMT sequences can be found in humans as well as other species [109, 110]. Such sequence similarity poses a problem for short read sequencing experiments in which aligners map reads to positions of the genome based on sequence similarity and can results in false positive and false negative variants being called. False negatives can arise due to mtDNA reads being mapped to the NUMT positions in the nuclear genome, and false positives can arise due to NUMT reads being mapped to the nuclear genome. Such effects have the potential to confound inferred variant allele frequencies, skewing the heteroplasmy estimates of single-cells. Computational study has shown these misalignments result in erroneous loss of coverage for the mitochondrial genome for reads derived from the mitochondrial reference genome *in silico*, and more severe loss of coverage can result when reads harbour variants that differ from the reference sequence. Some variants are even capable of completely removing all alignment from the mitochondrial genome [111]. By causing mutated

reads to align to homologs in the nuclear genome, a greater proportion of wild type reads align to the mitochondrial genome, thereby reducing the measured heteroplasmy. Low heteroplasmy variants may simply go unobserved as a result. For more details regarding this, we highly recommend Maude. *et al* (2019) [111].

The degree to which NUMTs confound mitochondrial heteroplasmy analysis will vary between modalities and approaches to library preparation. In their ATAC study, Lareau. *et al* addressed the issue of NUMTs by masking nuclear genomic regions which *in silico* mitochondrial derived reads mapped to. This was accomplished by simulating reads 20bp long from the mitochondrial genome with a 1-base error included. These simulated reads were then mapped to the reference genome with the mitochondrial genome excluded. The regions of the nuclear genome to which these simulated reads aligned were then masked from a re-alignment of the dataset. As a result all reads of mitochondrial origin would map to the mitochondrial genome. The supplementary material in their paper demonstrates the corresponding increase in coverage that results. The high copy number of mtDNA in each cell relative to nuclear sequences should result in a low error rate of NUMTs being assigned to the mitochondrial genome. Others have found that by aligning exclusively to the mitochondrial genome scRNA achieves superior breadth coverage of the mitochondrial genome in mice, which possess NUMTs that affect 6 mtDNA genes [88].

2.6.7 High Throughput Droplet Data

High throughput droplet based approaches to single-cell sequencing are the dominant modality with 10x chromium being one of the leading methods of library preparation available. Such short read approaches are biased with reads being sequenced from the 3' or 5' end of a transcript preventing consistent full mitochondrial per cell. Recent work however has shown that in spite of this end bias, 10x 3' and 5' libraries have low levels of coverage up to 10kbp from the 3' and 5' ends of the transcripts respectively [112]. Under certain circumstances, these regions could allow for high heteroplasmy variants to be called as they have a high detection rate even at low depth of coverage, and are less likely to be confused with errors. Whilst such coverage would be heterogeneous on per cell level, the large throughput of these techniques implies that even if

a small proportion of cells within a dataset yield coverage at points far from the transcript end, enough cells covering identical regions could still enable discovery of heterogeneity. Regardless of this, high throughput approaches yield consistent coverage near transcript ends, which should enable for reliable variant detection in those regions of the genome.

2.7 Conclusion

Studying heteroplasmy at the granular level of the single-cell has important applications from lineage tracing cells in clonal populations to potentially assessing the pathological impact of such mutations on gene expression. Here, we surveyed several recent approaches in the literature and the questions to which they were applied. Subsequently we highlighted several technical limitations of current technology, both in terms of experimental approach such as read depth, ambient RNA and end bias, and potential software induced errors like quality control filters and NUMTs. Whilst such technical issues can confound analyses, biological heterogeneity can and has been detected with such approaches. As sequencing technology advances, finer levels of granularity will be observable, and more robustly. However, given the volume of data already published covering many different tissue types and the successful explorations of mitochondrial variation exhibited in these approaches we believe that current data could be further leveraged to explore mitochondrial heterogeneity at the level of the single-cell.

Chapter 3

Somatic Mitochondrial Mutations In Ageing

"The race is not to the swift or the battle to the strong, nor does food come to the wise or wealth to the brilliant or favor to the learned; but time and chance happen to them all." – Ecclesiastes 9:11

3.1 Summary

The preceding chapter outlined the technical background for understanding single-cell genomics, and mtDNA variant calling in this context. This chapter employs this understanding in the context of single-cell mitochondrial ageing. In this chapter, I outline why mitochondria are expected to play a role in human ageing based on previous literature regarding tissue aggregates. I go on to describe a simple model that illustrates how mitochondrial mutations are expected to arise in single cells throughout human life, and how the heteroplasmy of these single-cell mutations is expected to reach pathological levels in middle to late-aged humans. Single-cell genomics data allows for these claims to be tested empirically and gain understanding about the role such mutations might play in the ageing process.

As no standard protocols existed for obtaining mtDNA variants from single-cell data, the first project of my PhD was devoted to developing a bioinformatics pipeline that would also go on to serve as a template for a variety of related research projects. The developed pipeline also served as a template for the pipeline used for processing data in Chapter 5 of this thesis.

Though I was primarily responsible for the development of the bioinformatics, the project as a whole was done in collaboration with Alistair Green and Florian Klimm. As well as my own analysis, I present some of the results of analysis they produced based on this pipeline, further highlighting the relevance of my work and placing it in its broader context.

Collectively, our research team was able to deploy the bioinformatics pipeline I developed to empirically test a variety of research hypotheses regarding this previously unappreciated aspect of human ageing. We go on to probe the claims based on this model empirically using single-cell genomics data in a manner that has largely gone unnoticed by the wider research field.

The results presented here are currently in peer-review.

3.2 Introduction

Numerous societies are facing declining birth rates, with most global regions forecasted to experience substantial increases in the proportion of their population over 65 years of age by 2050 [113]. The societal challenges implied by these trends are daunting [114]. This underscores the importance of uncovering novel approaches to understanding and improving human healthspan.

Mitochondrial dysfunction has been identified as one of the major hallmarks of human ageing [115]. As one of the critical organelles in maintaining the functioning of the cell, it would not be surprising that degradation in the regular operation of the mitochondria would lead to pathological consequences. Why might such degradation arise?

There is a long history of research linking mitochondria to ageing. This history started with a theory centred around the "free-radical theory of ageing", in which reactive oxygen species (ROS) are thought to create damage to cells, mitochondria being one of the main sources of ROS [116, 117]. Contrary to such initial research hopes, the evidence of ROS as a causal driver is scant. This point has been further reinforced by comparisons between mice and naked mole rats, which have far more oxidative damage despite their far longer life spans [118]. Moreover, mice with substantial modifications to their oxidative damage have lifespans comparable to unaffected mice [119]. In recent years, however, evidence has emerged that point mutations in mitochondrial DNA could play a functional role in humans as they age [32]. Modern sequencing technologies have made it clear that individual humans possess heterogeneous populations of mitochondrial DNA at the cellular and tissue level [31, 34]. Some of this variation may be passed in the germline from mother to child, undergoing various forms of segregation during development. These mutations, however, are inherited and are not present in large numbers in any individual. On the other hand, *de novo*, somatic mutations, which take place during life rather than being inherited from a parent, are also known to occur. A Mechanism for producing such mutations, which would accumulate over a human lifespan, could be shared among all humans. What could such a mechanism look like?

Whilst it used to be believed that mitochondrial point mutations are largely generated by ROS production, this is unlikely based on the biochemistry of mtDNA. A common form of ROS damage result in a lesion on G bases, referred to as 8-oxo-dG. Such lesions are mutagenic because they are liable to bond with adenine, rather than the expected cytosine bases, resulting in G-T *transversion* in the DNA. Mutations of this type are observed to be rare in mtDNA, undermining one of the main predictions of the ROS causing DNA damage theory. Conversely, *transition* mutations, consistent with replication errors, are observed more commonly and are reported in tumors in aged tissues [32]. A further line of evidence supporting the claim that replication errors contribute to mammalian ageing is the phenomena of polymerase γ (Polgγ) mutator mice. Polgγ is a subcomponent of the mitochondrial polymerase enzyme responsible for performing mitochondrial DNA replication. Laboratory mice which have modified versions of Polgγ are error-prone, and accumulate elevated levels of somatic mitochondrial mutations relative to wild type controls, leading them to be dubbed "mutator mice". The primary result of interest with mutator mice is that they display an ageing phenotype without increased ROS production or oxidative stress [120, 121].

Previous studies have shown how mitochondrial mutations accumulate with age in cellular and tissue aggregates [122, 123], again primarily from replication errors [124]. Such mutations, however, must first occur within single cells, with "bulk" aggregate analyses missing detail that is intrinsic to the underlying cellular process. In post-mitotic tissues in particular, the mitochondria in each cell can be thought of as separate populations of replicating mtDNA. A reasonable simplifying assumption is that in post-mitotic cells mtDNA mutations would originate and proliferate more or less independently in each single cell. As such, the spectrum of mutations present would vary from cell to cell, with many cells harboring *de novo* mutations, not shared by neighboring cells in the aggregate. Furthermore, the set of *de novo* mutations acquired by single cells would increase during a mammalian organism's life, as more erroneous mtDNA replication occurs. As we will show below, the heteroplasmy of mutations acquired in early life will rise to the extent of being able to have a physiological impact. Standard bulk analyses miss such salient features of the mtDNA dynamics, as cellular identity is lost in such analyses and only mutations present in many cells would be capable of entering any

downstream analysis. As such, we argue that most post-mitotic mtDNA mutations that occur with ageing are *cryptic*, being missed from standard bulk analyses. To make this more concrete, we state that single-cell analyses allow us to distinguish three classes of mutations on a per donor basis: *cryptic mutations* which are found within a single-cell in the sample; *partially prevalent* mutations which are found in multiple but not every cell in the sample; and *fully prevalent* mutations which are found in every cell in the sample.

We develop an understanding of this process focused on the accumulation of mtDNA mutations in post-mitotic tissues. In this context, cryptic mutations defined in the strict sense above — those found in only one cell in the sample — would correspond to *de novo* mutations occurring post-mitotically within a single cell. As is detailed below, a simple population genetics model, the Moran Model, is capable of capturing the behaviour of mtDNA populations in single cells [125]. Whilst the mechanisms of mtDNA damage accumulation in replicating cell types are of interest, they are beyond the scope of the research presented in the current chapter, as models describing replicating populations of cells possess different dynamics. With the aid of this simple model, we can make tentative predictions about cryptic mutations and their properties as we age. In the next section, we show, *in silico*, using previous literature values that somatic mitochondrial mutations are expected to reach pathological levels on the scale of a human life-span. Given an understanding of these dynamics, we set out to test these observations empirically and begin to look for links between cryptic mutations and their functional consequences.

As chapter 2 outlined, in recent years, single-cell datasets have become abundant, and continue to be published at a rapid pace. Such datasets afford us opportunities to empirically test our theories describing cryptic mtDNA accumulation. Many such datasets are publicly available, and are stored in repositories such as the Gene Expression Omnibus (GEO) [126]. We make extensive use of these available datasets. In the current chapter, we primarily focus on scRNA sequencing. The reasons for this are two-fold: 1) Single Cell Transcriptomics data is typically far more abundant, affording us the flexibility in searching for relevant datasets; 2) Single cell transcriptomics data allows us to link the presence of cryptic mtDNA mutations to observed phenotypic differences. This follows previous work by others using scRNA data to perform lineage tracing with mtDNA variants [77, 82]. Such mutations are, however, by definition, present

in the many cells which form a lineage. Here, by contrast, we are interested in post-mitotic mutations which arose *de novo* in single cells, which are the *cryptic* mutations identifiable in single-cell data.

Chapter 2 detailed the technical aspects of single-cell genomics in the lab, and introduced the relevant terminology. In this chapter, we detail the bioinformatics pipeline and tools used to process the data output by sequencers after library preparation has taken place. The bulk of my contribution to this project was in the construction of this pipeline. The simple aim of the pipeline is to take the raw sequencing data, and produce expression data, and mitochondrial genotyping of every cell in the data. Standard, single, end-to-end tools for this analysis do not exist, so we built them. This gives us the capability to identify previously unseen cryptic mutations. The details of the pipeline are explored at length in section 3.4. As we will be analysing various single-cell data sets throughout the remainder of the thesis, all of which are processed in a similar way, the presentation of the pipeline in this chapter will serve the reader as a guide for understanding later material.

In section 3.5, we give an overview of which data sets went into our analyses and the rationale for their inclusion regarding the ageing process. The data covers human, rats, and mice at a variety of different age, allowing us to test the behavior of our model in a range of different settings.

We round off the chapter by showing an application of the pipeline to a variety of datasets with a series of results about cryptic mtDNA mutations in human ageing. We show that mtDNA mutations accumulate with age in post-mitotic single cells in a manner consonant with theory, and notably, these mutations reach a pathological level on a time scale approximating human lifespan. After demonstrating this plausible role played by mtDNA mutations, we then extend our analysis to ask whether the burden of cryptic mtDNA mutations could be controlled through a previously discovered anti-ageing technique, *caloric restriction*. Finally, by exploiting the RNA modality of these data sets, we link the presence of cryptic mutations to their likely functional consequences. This analysis links the presence of mtDNA mutations to five of the nine established hallmarks of ageing [115].

3.3 The Moran Model

To model the evolution of mitochondrial DNA sequences in post-mitotic cells we use a well-established model from statistical genetics, the *Moran Model* [125]. The Moran Model used here is a continuous—time, birth—death process, in which the population size, N, is held fixed. In its simplest version, each event of the stochastic process involves one member of the population, labelled x, being randomly chosen to die equiprobably [127]. One of member of the population is then chosen to replace x, including x itself [127]. The time between these birth-death events is modeled by an exponential distribution. This leads to the inter-arrival time of birth-death events to be distributed according to $t_N \sim \exp(\frac{N\ln(2)}{t_{1/2}})$, with $t_{1/2}$ as the half-life of mtDNA, and t_N as the inter-event time in a population of N mtDNA. To see this, consider that the lifetime of an individual mtDNA, indexed by i, is $t_i \sim \exp(\frac{\ln(2)}{t_{1/2}})$. At the start of the process, the time until the first birth-death event is equivalent to the first degradation and replacement event of a single mtDNA. It can be seen that t_N is the minimum order statistic of N, IID, exponential random variables. Exponential random variables are, however, memoryless. Once the first birth-death event takes place, the process resets and the time between the first and second birth-death event will also be $\exp(\frac{N\ln(2)}{t_{1/2}})$. The process is therefore a sequence of birth-death events driven by such exponential random variables.

The mutation process is then incorporated into the model by allowing a mutation to occur with probability $P(\text{Mutation}) = \theta$ every time that a molecule is chosen to replicate. A more detailed version of the model would have m mutations occurring at every time step, where $m \sim \text{Binom}(L_{mtDNA}, \theta')$, where θ' is the error rate of Polγ. With θ' sufficiently small, the probability of more than one mutation occurring within an mtDNA in a single replication is small, and we are left with the result that $P(\text{Mutation}) \approx L_{mtDNA}\theta' = \theta$. As a simplifying assumption, all mutations are chosen to be unique. This is not an unreasonable assumption given that the chances of a mutation occurring at the same position on two different molecules or a new mtDNA mutation reverting a previous mutation back to the reference code is vanishingly small for large DNA molecules [127].

The Moran model is just one among many standard models encountered in population genetics.

An alternative such as the Wright-Fisher model is similar, but inappropriate in that all members of the population are replaced in each generation, whilst the Moran model allows us to capture the behaviour of *overlapping generations*, a behaviour more reasonable for the mtDNA population of a single cell. In its simplest form, the Moran model we are using also assumes a fixed population size, a simplifying assumption appropriate to post-mitotic cells. In contrast, populations of mtDNA in mitotic cells would experience fluctuating population sizes, adding complexity to the model.

The *site frequency spectrum* (SFS) is the distribution of allele frequencies at individual genetic loci across a sampled population. Figure 3.1 shows the SFS for an isolated, fixed-size population of evolving mtDNA through time, as modeled by the Moran model. This was simulated using fixed population sizes of 200, 500, and 1,000. We set $t_{1/2} = 30$ days, and a Polgγ per base error rate of $\theta' = 10^{-7}$ based on previous literature values [128, 129, 130]. The model shows that at around the age of 5, any *de novo* mutations present would be typically be below 10% heteroplasmy, with few mutations exceeding 30% heteroplasmy, even in cells with a population of 200 mtDNA (note the logarithmic scale). As we proceed through to later life, mutations are more likely to be found at higher heteroplasmies. At the extremes, it can be seen that homoplasmic mutations accumulate, as some mutations reach fixation in cells, dominating the entire cellular population. This even occurs in cells with 1,000s of mtDNA, by the ages of 75, with a more pronounced effect in cells with smaller mtDNA populations. This highlights how on the scale of a human life span, cells will be present in which most of the their mtDNA harbours potentially pathological mtDNA mutations at high heteroplasmies.

The Moran model gives a straightforward description of how mitochondrial mutations are persistently introduced as we age, how they accumulate, and how the distribution of the heteroplasmy of these mutations increases over time, in post-mitotic cells. Moreover, it shows how increased population size attenuates the increase of cellular heteroplasmy. Population size differences underscore how different tissues are expected to undergo evolution of the site frequency spectrum at different rates, reflecting copy number differences. The arrival of appreciable numbers of homoplasmic mutations is found to be approximately the time scale of human life in tissues with hundreds of copies of mtDNA. Previous research has shown that mitochondrial

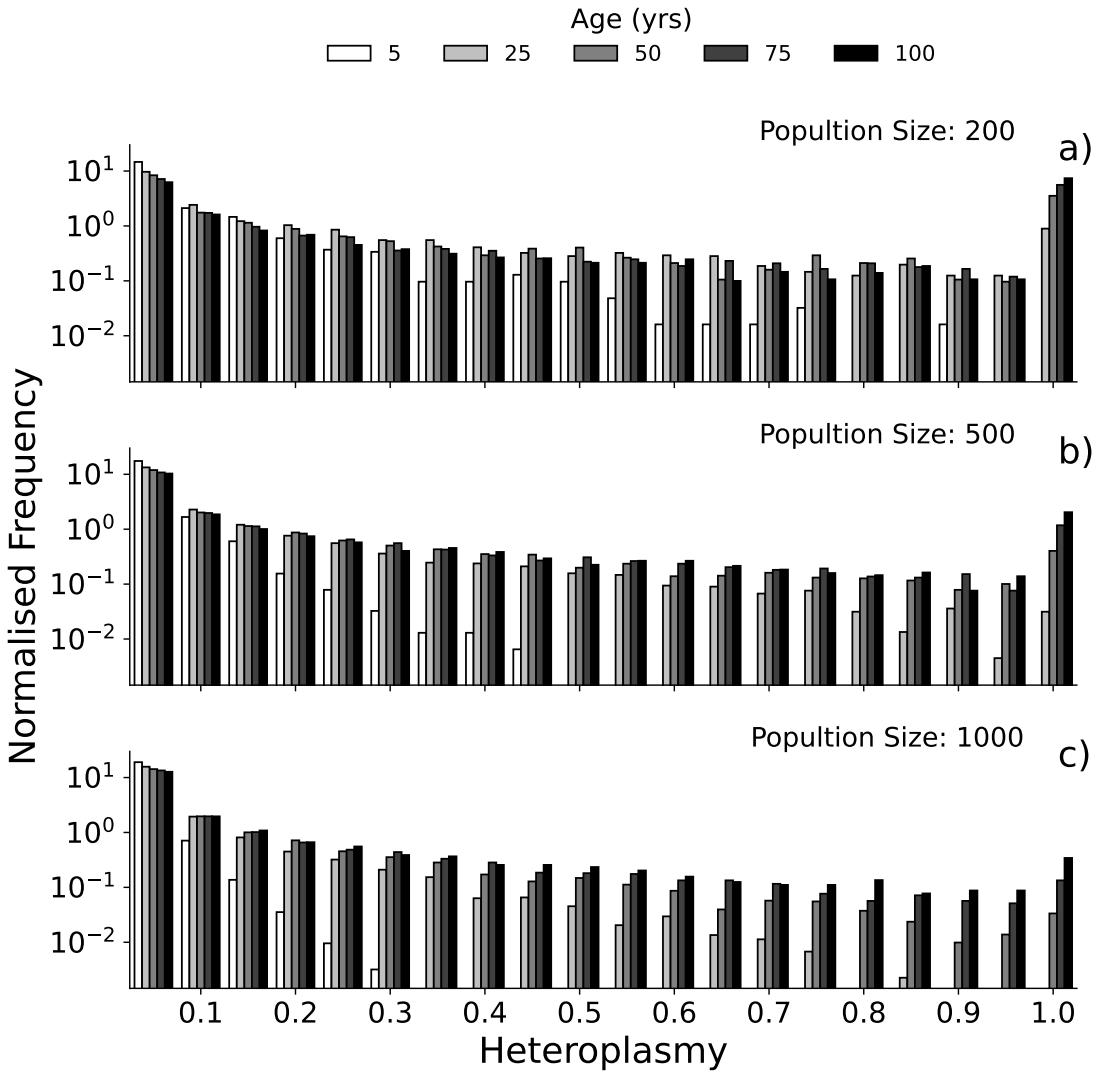


Figure 3.1: Simulations from the forwards-in-time Moran model shows how the heteroplasmy of somatic mutations increases as humans age. The Moran model was simulated with with a fixed population size, with population sizes of 200, 500, and 1,000 mtDNA. For mtDNA turnover, we set the half life at $t_{1/2} = 30$ days, and a Polgγ per base error rate of $\theta' = 10^{-7}$. These parameter values for data are based on previous literature [128, 129, 130]. The simulation was run for $1,000$ cells aging over a 100-year period. As we progress through human life, we begin to see that heteroplasmy of single-cell mtDNA mutations progressively increases with an accumulation of heteroplasmic mutations reaching fixation.

mutations typically need to be present at high heteroplasmies, typically in excess of 50% to exhibit a phenotypic effect [19]. It is an intriguing observation that the times at which such single-cell mutations begin to accumulate at heteroplasmies of physiological effect is around the time that human health begins to decline. In particular, the time scale of these accumulations is controlled by the particular combination of mtDNA copy number and turnover exhibited in

mammalian cells.

To empirically test how well this model describes the evolution of single-cell mitochondrial DNA and their potential effects, we probe single-cell genomics sequencing data. We now turn to the details of the bioinformatics pipeline that we developed to process such sequencing data and test these claims about human mitochondrial ageing.

3.4 Single Cell mtDNA Variant Calling Pipeline

Figure 3.2: Directed acyclic graph representing the data processing pipeline. After downloading the rawscRNA-seq data as `.fastq` files, we align them to a reference genome, which yields aligned reads in a `.bam` file and an expression matrix. We were able to obtain the heteroplasmy of mitochondrial mutations in each single cell in the datasets produced by M.J. Muraro, *et al.*, A.P. Voigt, *et al.* S. Ma, *et al.* [131, 56, 132]. Custom `python` scripts were also used to implement standard steps of single-cell expression analysis [22]. Then using custom Jupyter notebooks we perform quality control on the mitochondrial content of cells followed by assigning mutant classes to every mutation based on the number of cells from a donor it is found in.

Tools exist for identifying mtDNA mutations and their heteroplasmies in bulk samples [133]. However, software tools for analysing identifying mtDNA mutations in single-cell data are nascent and no tools exist for identifying cryptic mutations specifically. In order to overcome this limitation, we developed our own bioinformatics pipeline, capable of taking raw sequencing data in the form of `.fastq` files, right through to identifying putative single-cell mtDNA variants, alongside the expression levels of these cells (see Figure. 3.2). The details of the bioinformatics pipeline given here serves as a template for subsequent analyses throughout this thesis, with some variations depending on sequencing modality and the method of library preparation that has been used. Deviations from this template will be highlighted where appropriate.

I was responsible for conducting the research into identifying the main tools used throughout the pipeline. Writing such pipelines is a time consuming task, with research on the kinds of input files used and the variety of tools available to process them taking considerable effort. This effort is spent in: developing familiarity with the broad range of sequencing modalities and the output file types associated along with nuances in how they are formatted; evaluating alternative approaches to reaching the desired output with different combinations of tools; assessing what was in feasible in terms of the compute resources available; and testing the pipeline to ensure that each step performed as intended in pre-defined situations whilst still being extensible to new use cases. This was time costly, requiring us to program in `bash`, `python`, and `R`. Once finally developed, the pipeline was run on a variety of datasets which is discussed in section 3.5.

In Figure. 3.2, the pipeline is represented as a directed acyclic graph (DAG) for simple visualisation. Code and scripts are represented by ovals, with rectangular nodes representing the input and output file of the various scripts. We now give a detailed explanation of this pipeline and some of the technical details of sequencing protocols that informed the pipeline structure.

3.4.1 Background On Sequencing Protocols

In section 2.4, an outline of various sequencing protocols was given. Here, the specific details of a bioinformatics pipeline used to analyse the data produced by these sequencing protocols

is elaborated and clarified. We further elucidate the steps taken in library preparation and sequencing to facilitate an understanding of how the data were generated, and why the pipeline functions the way that it does.

The protocols of primary concern are 3' single-cell RNA sequencing data sets described earlier (see 2.4.3). These modalities offer the advantage of allowing massive numbers of cells to be sequenced together through the use of microfluidics technologies like those offered by 10x genomics [46]. The crux of this approach is that many thousands of cells can be captured separately in individual oil droplets that contain oligonucleotides unique to the droplet. In the case of 10x sequencing chemistries, these oligonucleotides contain 4 important features:

1. *Sequencing adapters and primers* containing the DNA sequences necessary for the DNA reads produced to bind to sequencing plates once they are finally sent for sequencing.

2. *Cellular barcodes* unique to each oil droplet, and thus to the unique cell that the oil droplet captures.

3. *Unique molecular identifiers* (UMI) which are unique to each oligonuceliotide, allowing captured RNA molecules to be uniquely identified.

4. *Oligo-dT primers* which are annealed to the polyadenylated (poly-A) tails of RNA transcripts. Poly-A tails are long adenine sequences that are appended to transcripts prior to translation in eurkaryptoc cells. Oligo-dT primers primers allow for the efficient initiation of reverse transcription.

Once cells are captured in oil droplets they are lysed and the RNA they contain exits the cell membrane, entering the oil droplet. The gel-beads release their oligonucleotides which then bind poly-adenalated tails of the RNA molecules commencing reverse transcription, producing a complementary DNA molecule (cDNA). Consequently, each cDNA molecule in the sample has a barcode, identifying which cell the molecule came from, and a UMI distinct to each cDNA molecule. These identifiers are key in what follows. All the cDNA produced within all of the separate oil drops are now pooled together, a process termed *multiplexing* and amplified by

PCR to produce a larger pool of molecules, facilitating later sequencing with the resulting pool having many copies of each original cDNA molecule. The entire pool, often called a *sequencing library*, can then be sequenced together, allowing all of the sequencing reads from different cells to be digitised at once.

As the oligonucleotides originally incorporated into the reads at the start of the process contain the cell barcodes and UMIs, we can identify which cell and molecule each PCR amplified cDNA originated from once it has been sequenced. The process of identifying which cell a sequenced read a cell comes from is termed *de-multiplexing*. Separately, as each cDNA molecule is PCR amplified, we can use the UMIs to group all sequencing reads by their original cDNA molecule, and choose a single candidate amongst all amplicons to represent the original molecule in a process termed *deduplication*.

Once prepared, sequencing libraries are processed on a sequencer, often an Illumina sequencer. The cDNA to be sequenced are placed onto a *flow cell*, binding to oligonucleotides on the flow cell's surface which are complementary to the oligonucleotides previously attached to the cDNA. Each cDNA fragment is then amplified in a process called *bridge amplification* to produce a cluster of identical reads at a physical location on the sequencer.

Once the amplified clusters of cDNA has been formed, a process termed *sequencing by synthesis* is used to digitise the information encoded in the cDNA. During this process, the cDNA template strand is used as a starting point, and a DNA polymerase enzyme adds complementary nucleotides one at a time. Importantly, each sequencing nucleotide added is labeled with a fluorescent marker specific to the base (A, C, G, or T) it represents. Additionally, the fluorescence markers also act as a blocking group, such that once a complementary base has been incorporated to the synthesised strand, it prevents further synthesis from taking place until the fluorophore is removed. The colour of the fluorophore is recorded by an optical sensor and recorded into a computer. The fluorophore is then removed, allowing the next cycle of sequencing by synthesis to continue, with the next complementary base being incorporated into the synthetic sequence with its associated fluorophore. This process is repeated until the entire length of the cDNA sequence has been obtained [134]. This is done for all clusters on the flow

cell separately allowing the entire library to be sequenced in parallel. Having pooled the cDNA for all cells together in the library, this process obtains the transcriptome of thousands of cells simultaneously.

The standard output of these sequencers are `.fastq` files containing sequence information of all bases on the sequencer with each bases corresponding quality score [135]. The quality score, also called the *phred score* of these bases is typically defined by $Q = -10 \cdot \log_{10}(P)$, where p is the probability that the corresponding base call is incorrect. Errors can occur in sequencing by synthesis in which not all strands within a cluster on the flow cell incorporate a new base at each step, leading some to lag behind others in the process. This leads to conflicting fluorophore signals emerging within a single cluster, leading to uncertainty about what the true base in the original sequence was, as captured by the phred score.

3.4.2 Data Downloading

Alongside publications, `.fastq` files are typically deposited onto the Gene Expression Omnibus (GEO) and made publicly available [126]. The exact datasets that we used are described in table 3.6. Publicly accessible data was downloaded from the Gene Expression Omnibus (GEO) using SRA toolkit (version 2.10.0) to retrieve the data [126].

3.4.3 Alignment, Demultiplexing, and UMI Counting

In the context of single-cell RNA sequencing, we would like to achieve the following from our sequencing reads: Align reads to the reference genome; associate the aligned reads with particular genes on the genome, perform demultiplexing to associate each read to its cell of origin; and perform UMI deduplication to count the unique RNA molecules detected in the experiment. Aligned reads are used to obtain two key pieces of information for our interest. Firstly, we would like to know whether the aligned sequences of the reads deviate from the reference genome, reflecting any mutations that were present in the cell. In particular, we are interested in locating and finding mitochondrial mutations within each cell to estimate its heteroplasmy. This can then be used to empirically test claims about the evolution of single-

cell mtDNA mutations through human life. Secondly, we would like to count all reads from the same gene in each cell, and use that to obtain expression data for that cell. Combined, these two pieces of information will be used later in the chapter to find associations between heteroplasmy and cellular gene expression.

To align raw reads to a reference genome, we use the `STAR` aligner (version 2.7.5c) [63]. `STAR` offers the advantage of being widely used for RNA sequencing data, and having functionality called `STARsolo` for specifically handling multiplexed data [136]. `STARsolo` takes `.fastq` files containing all cells sequencing information as input and returns (1) all aligned reads in a `.bam` file and also (2) a deduplicated UMI counts matrix `.tsv`. In the process, `STARsolo` demultiplexes the sequencing information, using the barcode which forms the first bases on each sequenced read. The resulting `.bam` file contains the cellular barcode as metadata along side every read's alignment information. For non-multiplexed datasets (such as Smart-Seq2) `STARsolo` appends the sample name to the reads of each cells `.fastq` file and then creates an expression matrix for all cells simultaneously, as well as a `.bam` file. We can then split by sample name for variant calling on each cell.

For full-length datasets, we constructed an expression matrix for the whole experiment with `STARsolo`, and we created the aligned `.bam` file using STAR for each cell separately. We processed `CEL-seq2` `.fastq` files using UMI-Tools (version 1.0.1) [65]. Cellular barcodes of reads were extracted from barcode `.fastq` files and placed into the read names of genomic reads using the software package `umi tools`. We then aligned reads to the reference genome using STAR with default settings, outputting an aligned `.bam` file. The package `featureCounts` was then used to tag reads with the gene they map to. We processed this `.bam` file by using the `umi tools` count function to produce an expression matrix. We used the expression matrix and aligned `.bam` files for DGE analysis and variant calling, respectively, in the downstream analysis.

`.bam` files contain mapped reads, indicating the position of each read on the genome. A typical output from `STARSolo` is a single, large `.bam` file, containing all of the aligned reads from all cells in the data. For the purposes of calling variants on the mitochondrial genome, only reads

that align to the mitochondrial genome are required. Therefore, we extract all reads that align to the mitochondrial genome using `samtools` (version 1.11) [137]. This is achieved by first indexing the `.bam` file. Indexing a `.bam` file essentially means creating a data structure that allows for quick random access to the contents of the `.bam` file without needing to scan the whole file. This index, usually a separate file with a `.bai` extension (for `.bam` Index), stores a map of the positions of the aligned reads in the `.bam` file. The process works by dividing the reference genome into small intervals or bins and then keeping a record of which reads fall into which bins. By doing this, the index allows software to quickly locate and retrieve all reads mapping to a particular region of the genome (specified as a range of positions) without having to scan the entire file.

Once the `.bam` has been indexed, it can be queried for reads that only align to the mitochondrial genome using `samtools`. We can then sort reads by the cell barcode tag in the file. Reads are then further filtered to retain only those reads that contain putative cellular barcodes found within a whitelist, a predefined list detailing all the possible barcodes that could be used in the experiment. This eliminates any barcodes that may have been spuriously sequenced. Finally, the mitochondrial `.bam` file is split by cell barcode resulting in a separate mitochondrial `.bam` file for each cell in the data. This typically results in 1,000s of separate `.bam` files. Each of these can then be separately supplied to the main variant caller script to identify putative cellular variants. This step can be done in parallel by making use of multiprocessing tool on separate clusters on an high performance computer's (hpc) compute node, making per cell variant calling a relatively rapid procedure.

3.4.4 Single Cell Variant Calling

We used a custom variant calling script written using `pysam` (v0.15.3) in order to call variants on each single cell's mitochondrial `.bam` file. In the process of doing this, we first drop reads which are not uniquely aligned to the mitochondrial genome to avoid heteroplasmy calls caused by Nuclear sequences of mitochondrial origin (NUMTs) — sequences found within the nuclear genome that have a homologous relationship and therefore high sequence similarity to sequences

of the mitochondrial genome. We only considered genomic positions with over 200 reads align to the position with a base quality score above 30, and only retain cells for further analysis if the number of positions passing quality control was above 300. Any mismatches on a candidate read between the reference genome is then considered to be a potential mutation. At any given position of the genome, there are four possible bases, $i \in \{A, C, G, T\}$, with NM_i, the number of mutated reads supporting base i. We, therefore, define the heteroplasmy of the mutation in a single-cell to be:

$$h_i = \frac{NM_i}{\sum_{j \in [A,C,G,T]} NM_j} \tag{3.1}$$

An obvious corollary of this definition is that a given position could have up to three different mutations, though this seldom occurs, and is not considered further [138]. Insertions and deletions are not considered in this work due to the technical difficulties of accurately identifying these mutations.

3.4.5 Classifying Detected Mutations

The result of running the pipeline for a given library is a list of mutations observed within every single-cell in the dataset. As was highlighted in the previous chapter, depth of coverage across the genome can vary substantially between different cells in the data. In some cells, some positions of the genome will be absent entirely. In the introduction we stated the aim of the current research is to focus on *de novo*, somatic mutations, which will appear as cryptic mutations in single-cell data. These mutations would be localized to the single-cell which they occur in. This adds a complication in that it is conceivable, though unlikely, that some mutations might be present in multiple cells in the data, but only a sub-set of cells had the requisite quality at that position to detect the mutation. As we stated earlier, there are three types of mutations we consider in our analysis, which we define concretely in terms of processing single-cell data:

1. *Maximally prevalent*: Those found in every cell from a donor for which that position of the genome passed quality control.

2. *Partially prevalent*: Those found in multiple but not all cells in which the position of the mutation passed quality control.

3. *Cryptic*: Those found to be in only one cell in a donor's data.

To ensure further protection from errors that arose in library preparation and sequencing, we retain only those mutations that are present at $h \geq 10\%$.

The combined set of stringent quality control thresholds and an appropriate classification mechanism ensure that only somatic mutations are carried for further analysis.

3.4.6 Expression Analysis

Whilst the steps of the pipeline outlined above were responsible for performing mitochondrial genotyping on single cells, the other side of the pipeline is responsible for giving us phenotypic transcription level data. Using the count matrix produced from `.STARsolo`, we exclude cells as recommended by current best practices [22]. For the analysis of the expression matrices, we use SCANPY. We filter each expression matrix using three variables: the total UMI counts per cell, total genes per cell, and the percentage of reads aligned to the mitochondrial genome. These quality variables are examined for outlier peaks that are filtered out by thresholding. We determined the threshold separately for each data set to account for quality differences. This quality control procedure allows us to establish cells with unexpectedly low read depth or high fraction of mitochondrial content, indicating that mRNA leaked from the cell during membrane permeabilization, or those with high read depth which could be doublets. We only keep cells, which pass both filtering steps — the variant calling filtering (as discussed in subsection 'Variant calling') and the expression matrix filtering (as discussed in this subsection).

3.5 Data Selection

We analysed publicly available datasets data from human and rat samples, covering 4 tissue types. All were chosen because of their potential roles in the ageing process. The datasets used are described in table 3.1.

In particular, pancreatic decline is known to link to diabetes and human retinal epithelium has been linked to macular degeneration, another ageing-related disease [139, 140]. The low cellular turnover rates in cell types associated with these tissues ensure correspondence with the Moran model developed earlier [141]. Additionally, another opportunity was presented in a publicly available rat scRNA sequencing dataset which sequenced young and old rats, as well as calorically restricted (CR) old rats [132]. CR is a known longevity intervention [142]. We investigated the brown adipose tissues due to its high mitochondria content [143]. We hypothesised that this would likely increase the chance of obtaining high quality sequencing data. CR is also known to increase mtDNA copy number in rat livers [144]. All else equal, the Moran model that we use to understand mtDNA accumulation in single cells indicates that high heteroplasmy mtDNA mutations would accumulate more slowly in larger mtDNA populations. The increase in mtDNA copy number in CR gives a plausible mechanism through which caloric restriction might operate to alleviate some of the ageing phenotype.

Reference	Species	Tissues	Age range	No. of cells	Accession number	Protocol
M. Enge, *et al.* [57]	Human	Pancreas	1 mo-54 yrs	2,544	GSE81547	Smart-seq2
M.J. Muraro, *et al.* [131]	Human	Pancreas	23-59 yrs	12,200	GSE85241	CEL-seq2
A.P. Voigt, *et al.* [56]	Human	Retinal Epithelium and Choroid	54-92 yrs	8,217	GSE135922	10x
N. Almanzar, *et al.* [25]	Mouse	Pancreas, Liver	3-24 mo	9,380	AWS	Smart-seq2
J. Buenrostro, *et al.* [78]	Human	Cell lines	N/A	1,344	GSE65360	scATAC-seq
S. Ma, *et al.* [132]	Rat	Brown Adipose Tissue, Liver	5-27 mo	29,165	GSE137869	10x
				$\sum = 62,850$		

Table 3.1: **Summary of the public data sets analysed in this study.** Overall, we analyse sequencing data of more than 62,850 cells in three different species and five different tissues, as well as, cell-line data.

3.6 Results

3.6.1 Coverage And Variant Calls

We analysed data from three different scRNA modalities: CEL-seq2, Smart-seq2, and 10x 3' sequencing. Figure 3.3 compares the output of these modalities. It can be seen that the average depth is the most consistent in the full length Smart-seq 2 data, followed by CEL-seq2, and finally by 10x 3' data which characteristically only possesses high depth at the 3' end of the protein-coding genes. This gives some understanding of the heterogeneity and limitations of current single-cell technologies. As different modalities clearly possess different coverage profiles, care needs to be taken when making comparisons between data sets. For instance, comparing the number of mutations per cell observed in Smart-seq2 and 10x data would be problematic as only small proportions of the mitochondrial genome make it past quality control filters in the 10x data compared to Smart-seq2 data. In all data sets, however, mutations are observed at a range of heteroplasmies in regions where the depth of coverage is sufficient to pass quality control thresholds. These regions can still be used to shed light on the effect of cryptic heteroplasmy on cell function. The fact that such observations can be made in a wide range of sequencing modalities serves to highlight the relevance and ubiquity of previously unnoticed cryptic mutations, as well as their potential use as an explanatory variable in single-cell genomics.

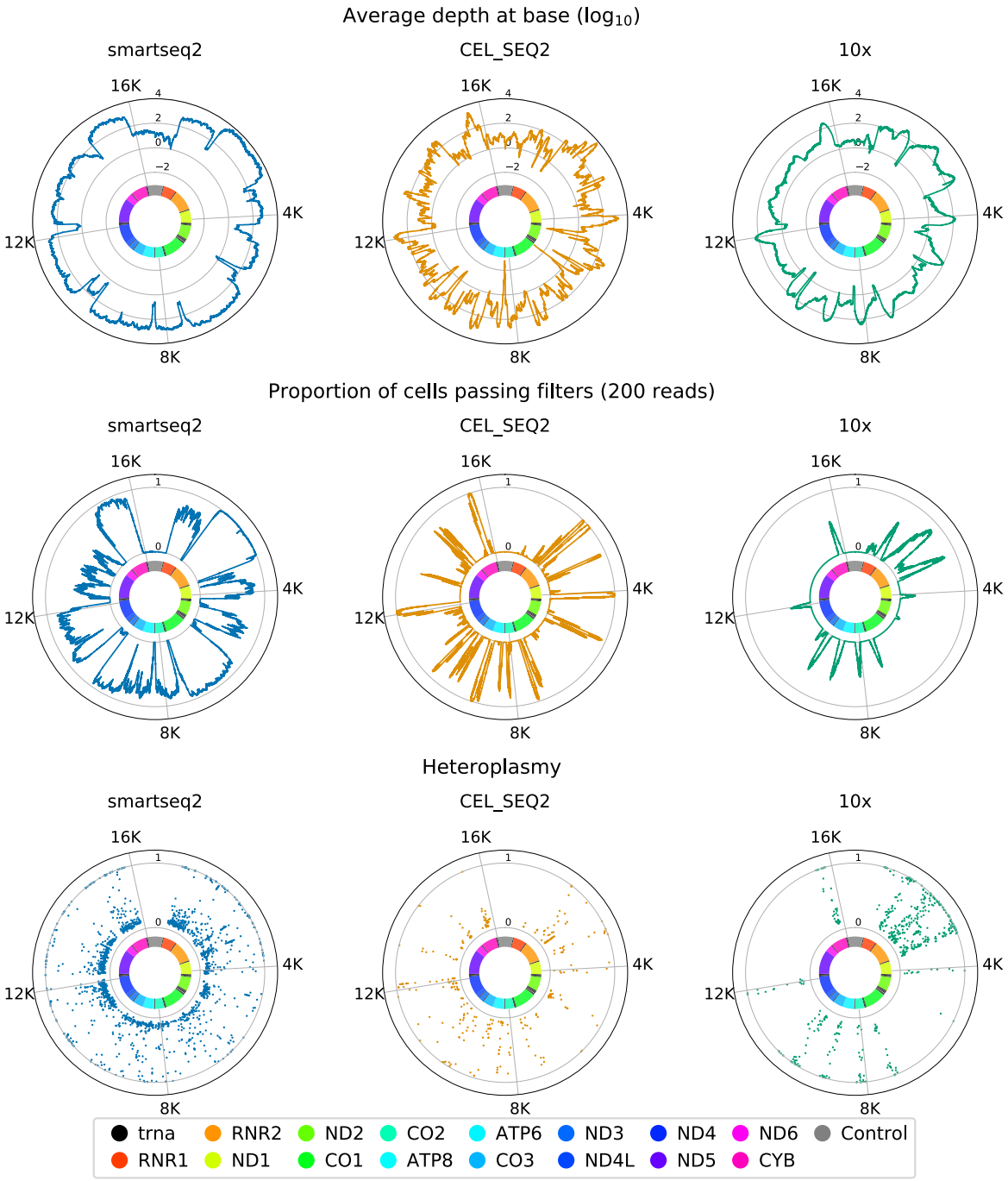


Figure 3.3: Throughout this research we use data from three different scRNA sequencing technique, both full length and 3'. We show how the coverage of the mitochondrial genome differs for each technique, and how this affects the number of potential mutations we can identify. As expected full length sequencing techniques provide the most coverage across the mitochondrial genome, but for all techniques we observe that mutations are spread evenly across the sequenced regions. No coverage of the D-Loop is observed as no transcripts are derived from this region of the genome

3.6.2 Accumulation Of Cryptic Mutations Through Time

This first analysis uses five single-cell RNA-seq datasets from two species, three tissue types, produced using three sequencing techniques to demonstrate the wide relevance of our results [57, 131, 56, 25]. Alistair Green, produced the following analysis showing the accumulation of cryptic mutations. I retrieved the raw publicly available sequencing data and processed it using the pipeline developed above. I also suggested the rank-biserial correlation as the appropriate measure of the effect size for comparing two empirical site frequency spectra. With permission, I present his analysis for the completeness of results in this chapter.

Bulk analyses typically constrain themselves to look at mitochondrial mutations that are present at $h \geq 0.5\%$. By analysing single-cell data, we found that approximately 90 % of mutations that are at $h < 0.5\,\%$ using a bulk heteroplasmy analysis. Of these low bulk-heteroplasmy mutations, approximately 95 % are only found in single cells in the single cell analysis. Figure 3.4 a) shows how the majority of cryptic, single-cell mutations detected by our pipeline would be excluded from a standard bulk analysis. This is done by combining reads from all cells in the single-cell analysis to create a *pseudo-bulk* data, and observing what bulk heteroplasmy these mutations would create. Previous work, therefore, fails to detect the huge mitochondrial mutant load present in many tissue which is only observable at the level of the single-cell.

After establishing that our pipeline is capable of detecting cryptic mutations, we sought to assess whether the general form of the dynamics indicated by the Moran model were born out. Primarily, this focuses on establishing that the heteroplasmy of cryptic mutations should increase over a human lifetime in a manner coherent with the site frequency spectrum dynamics we laid out earlier in this chapter. A detailed analysis of the correspondence between the Moran model and the empirical data was performed by Doctor Green and is explored further elsewhere [138, 145] We define $\mu^{t\,\%} = \sum_{i=1}^{m} h_i \mathbf{1}_{(h_i > t)}(h_i)$ as a measure of mutant load in each single cell, where the indicator function is used to only include mutations with a heteroplasmy beyond a certain level. We choose $\mu^{10\,\%}$ to protect against potential confusion with sequencing and PCR errors as a source of fake mutations, which would likely introduce low heteroplasmy mutations. To avoid confounding issues related to variation in breadth of coverage associated with different

modalities mentioned in the previous section, we focused on data from a single modality, Smart-seq2. This modality was chosen for its superior coverage of the mitochondrial genome. This data had pancreatic tissue samples from 8 subjects with ages ranging from 1 month to 54 years of age [57]. Figure 3.4 b) shows how mutant load increases with age, displaying the first basic insight from the Moran model, that mutant load should increase with age.

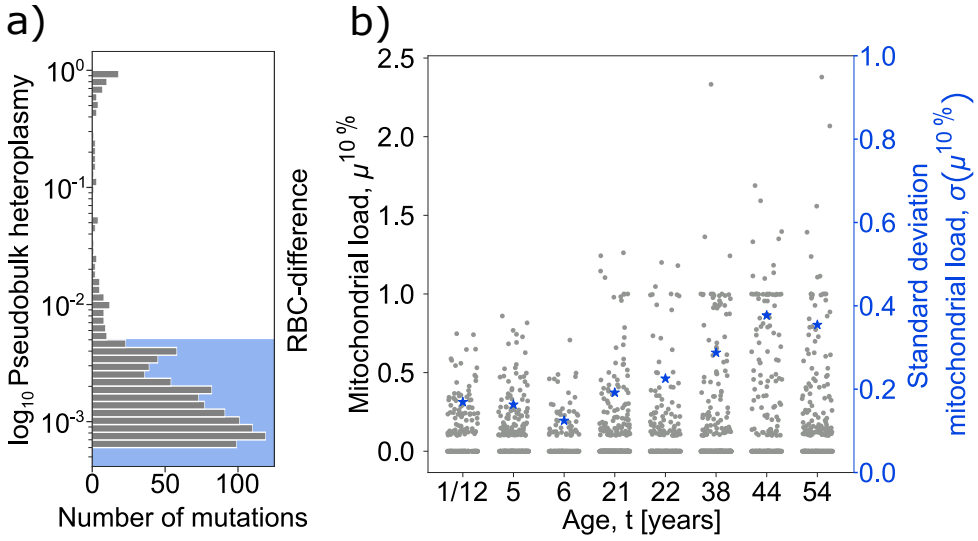


Figure 3.4: a) Single-cell sequencing is necessary to reveal true mutant load of a tissue (reproduced from [145] with permission of collaborators). 90.8 % of mtDNA mutations are at a pseudobulk heteroplasmy $h < 0.5\,\%$ (marked in blue) and so would not be reliably detected in most bulk experiments (Data taken from all mutations found in [57]) This indicates that most mitochondrial mutations are cryptic in nature b) The mitochondrial load $\mu^{10\,\%}$ of potentially pathogenic, cryptic mutations increases with age in human pancreas cells. We show the mitochondrial load $\mu^{10\,\%}$ for all eight donors from [57] and observe an increase with age (Spearman correlation $\rho \approx 0.04$ and $p < 0.049$). The standard deviation $\sigma(\mu^{10\,\%})$ of the mitochondrial load also increases with age (blue asterisks; Spearman correlation $\rho \approx 0.88$ and $p < 0.005$) [145].

Beyond the empirical observations of mitochondrial load, the Moran model with mutations implies that the SFS of cryptic mutations should have more mass at increasing heteroplasmy as individuals age. This prediction is born out systematically across our datasets. As an illustration, figure 3.5 b) shows the site frequency of a 1-month-old and 54 years old, both obtained from the same tissue and sequencing modality [57]. It can be seen that the 54-year-old has considerably more probability mass at higher heteroplasmy, with concentration at homoplasmic levels, whilst the 1-month-old has no observed homoplasmic mutations. To show that this difference is systematic, we make comparisons between the cSFS of all individual

donors, within and across all datasets that we analysed. Cross dataset comparison is legitimate in this case, as the site frequency spectrum concerns the heteroplasmy distribution of mutations, which is not affected by variations in depth of coverage, unlike the total mutant load.

To make comparisons between the cSFS of two individuals we measure the rank biserial correlation (RBC), r, a kind of rank correlation. The rank biserial correlation is a statistical method used to measure the strength and direction of the relationship between a continuous variable and a dichotomous (binary) variable.

Consider two groups. One group is represented by the label G_1 and the other by the label G_2. In our setting, G_1 and G_2 are two sets of heteroplasmic single-cell mtDNA mutations from donors of different ages. We also have a set of continuous data, in our case, the heteroplasmy of these single-cell mutations. The key question here is: How differently does this continuous data behave between the two groups?

Let f be the proportion of all such pairs of mutations in which mutations in G_1 have a greater heteroplasmy than mutations from G_2, and u be the proportion in which G_2 mutations have a greater heteroplasmy than the mutation from G_1. The rank biserial correlation is simply defined as $r = f - u$, and can range from -1 to 1. We posit the hypothesis that mutations from G_1 are typically greater than those in G_2. $r = 1$, indicates a positive effect size for the hypothesis indicating that all mutations in G_1 are greater in heteroplasmy than those in G_2, with $r = -1$ indicating the opposite. $r = 0$ or near 0 indicates that mutations selected from G_1 are equally likely to be higher or lower than those in G_2 [146].

We compute the RBC between the cSFS of all pairwise donors in our data, displayed in figure 3.5 a). Visually, we see a this results in in a strong correlation between the age difference of two donors and the RBC of their cSFS. We recognise that these samples are not all independent, but display the simple Spearman correlation as indicative of the positive association present in the data — $\rho \approx 0.70$ and $p < 10^{-26}$. A more comprehensive analysis using hierarchical modelling which does account for the association between these data points has been performed by Dr Alistair Green. This analysis reaffirms this association between the SFS and donor age. We refer the reader there for a more in-depth analysis [138, 145]. This bears out one of the major

hypotheses of the theory; that the heteroplasmy of single-cell mutations increases with age.

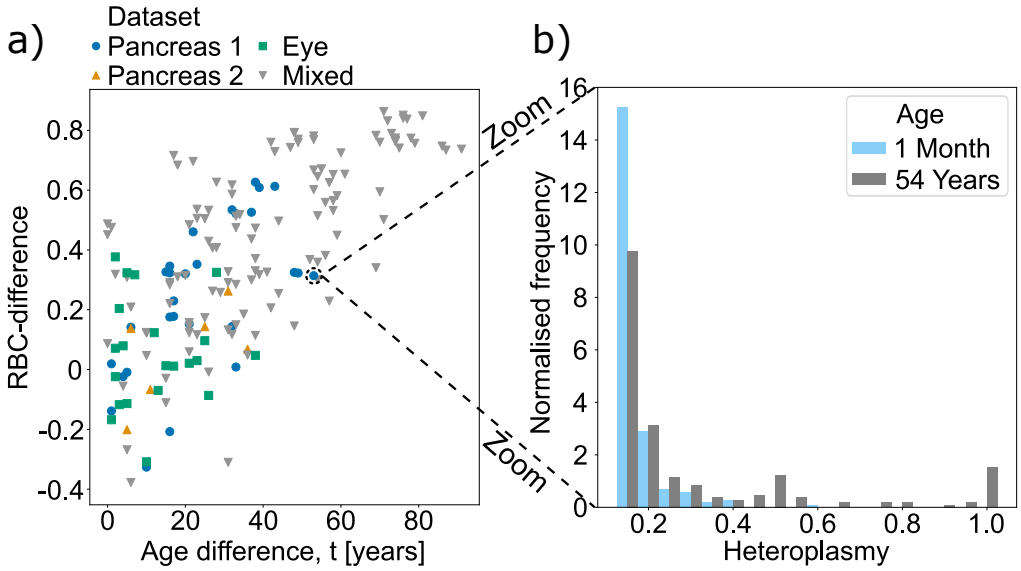
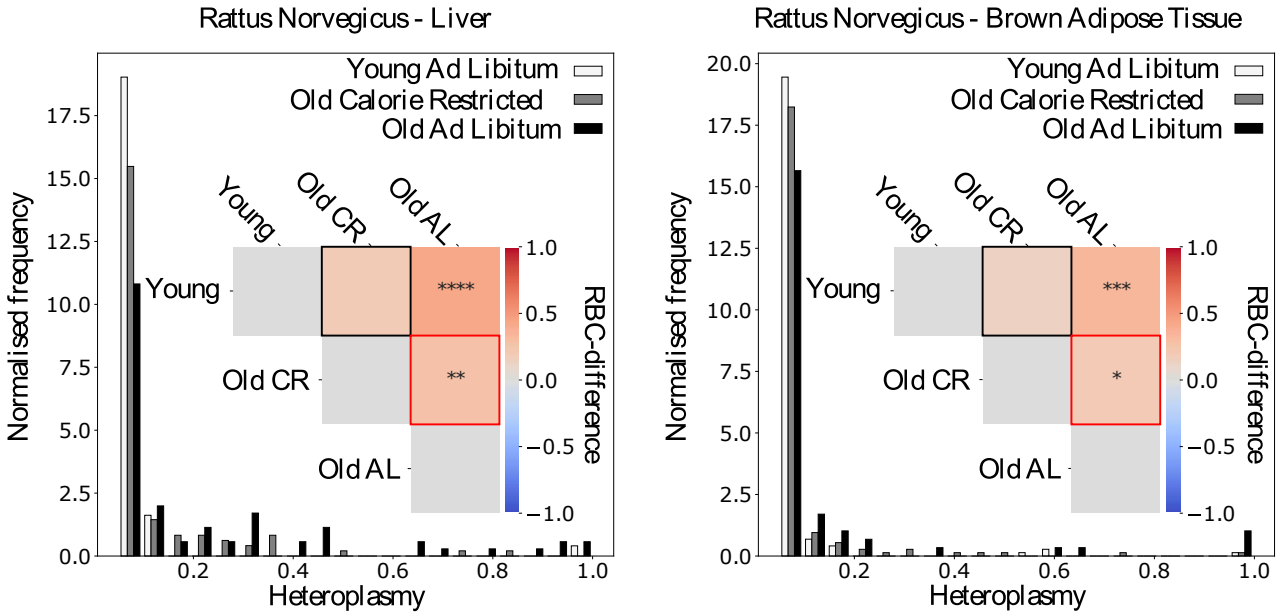


Figure 3.5: The cryptic site frequency spectrum (cSFS) evolves with time (reproduced from [145] with permission of collaborators). We see that for 19 individuals across 3 human datasets taken from different tissues and differing sequencing techniques, the rank biserial correlation distance (RBC-difference, a measure of how likely a mutation sampled from one spectrum will have a higher heteroplasmy than a mutation sampled from another) between two spectra increases with the age difference between the individuals (Spearman correlation $\rho \approx 0.70$ and $p < 10^{-26}$). By looking directly at an example pair of spectra (b) we see that, as predicted by our theory, there is an accumulation of mutants at high heteroplasmies (for a breakdown of the number of cells and mutations represented in the cSFS see table 3.1).

3.6.3 Impact On Brown Adipose Tissue And Liver In A Rat Comparing Caloric Restriction To Normal Ageing

I further retrieved, processed, and analysed data to give the following results exploring the role that caloric restriction might play on the accumulation of cryptic mutations. Caloric restriction has been recognised as one of the most effective interventions to promote longevity, and combat ageing [142]. We use scRNA-seq data from young and old *ad libitum* fed rats (Y-AL and O-AL, respectively) and old calorically restricted rats (O-CR) to obtain cSFS of each group's liver and brown adipose tissue [132]. Using the RBC-difference we observed that (as anticipated in Figure. 3.5 b)) cryptic mutations are of higher heteroplasmy in the O-AL than Y-AL (Figure. 3.6), in both liver ($p < 0.0001$, Bonferroni corrected) and brown adipose tissue ($p < 0.001$, Bonferroni corrected). Likewise, mutations in O-AL are higher in

heteroplasmy than in O-CR, in both liver ($p < 0.01$, Bonferroni corrected) and brown adipose tissue ($p < 0.05$, Bonferroni corrected). By contrast, the mutations in O-CR rats are not significantly different in heteroplasmy to those found in Y-AL rats. These findings suggest that caloric restriction slows the rate of increase of average cryptic heteroplasmy, this mechanism might be an explanatory factor for the observed longevity in calorically restricted organisms. Caloric restriction can increase the number of mtDNA molecules in rat livers [144]: our theory suggests that increasing mtDNA-copy-number slows the rate of increase of mean heteroplasmy.



(a) (b)

Figure 3.6: We find the cSFS for rats in three groups, young *ad libitum* (Y-AL), old *ad libitum* (O-AL) and old calorically restricted (O-CR), for variants called at a 5% heteroplasmy. This is displayed for liver in (a) and in (b) for brown adipose tissue. Pairwise RBC-difference between each group's cSFS indicates that caloric restriction slows the accumulation of cryptic mutation. This comparison was done between all pairings of the three groups and is displayed in the inset in (a) for liver and in (b) for brown adipose tissue. Mutations in the O-CR are marginally more likely to be at a higher heteroplasmy than the Y-AL in both the liver and the brown adipose tissue ($p > 0.05$). Compared to this, mutations in the O-AL are far more likely to be at a higher heteroplasmy than Y-AL mice (Mann-Whitney U Test $p < 0.0001$ & $p < 0.001$ Bonferroni corrected in liver and brown adipose tissue respectively). Critically, mutations in O-AL rats are statistically significantly more likely to be at a higher heteroplasmy than mutations found in O-CR (Mann-Whitney U test $p < 0.01$ & $p < 0.05$ Bonferroni corrected in liver and brown adipose tissue respectively). Mean heteroplasmies of observed liver mutations in Y-AL, O-CR, O-AL respectively are 0.0846 ± 0.0025, 0.115 ± 0.0013, 0.228 ± 0.0035. In BAT the means are 0.0796 ± 0.00068, 0.0909 ± 0.00072, 0.144 ± 0.0035 [145].

3.6.4 Linking Cryptic Mutations To Pathophysiology

We round off our results by showing evidence linking the presence of cryptic mutations to the hallmarks of ageing. The following analysis includes an additional mouse dataset that was processed by Alistair Greene using the same pipeline described in this paper, with further statistical analysis conducted by Dr Florian Klimm as part of this project [145]. These results are presented for completeness of the present thesis chapter. The additional data included single-cell RNA mouse data from liver and pancreas tissues from a publicly available dataset [25]. This further broadened the generality of the claims we were able to make to an additional species.

All cells found with cryptic mutations that were not-synonymous, above 10% heteroplasmy were compared to cells without in a differential gene expression (DGE) analysis. We initially performed this test in the high-quality full-length scRNA dataset, finding mitochondrially-encoded (e.g., MT-CO2, MT-RNR1) and nuclearly-encoded (e.g, NDUFA1, NDUFA13) OXPHOS genes upregulated, which is an established response to an impaired energy production [57, 147]

We perform a DGE analysis for seven scRNA-seq data sets from three different mammals (human, mice, and rat) and four different tissues, and identify in each of them that the cryptic mutations are linked to gene expression changes. Biological pathways linked to the presence of cryptic mutations across organisms we identified by performing a GO-term enrichment analysis with PANTHER [148]. We display terms that were present in at least six of the seven datasets in Figure. 3.7 b). Cryptic mutations we observed to coincide with a perturbation to the regulation of biological quality, response to stress, ER-stress, viral response, leukocyte activation, apoptosis, hypoxia and proteolysis, and protein-folding. Combined with an enrichment of immune effector process these terms are consistent with an immune response triggered by cryptic mutations: in line with recent findings linking neo-epitopes to *de novo* mtDNA-mutations [149, 150, 151]. An enrichment of response to nutrient levels indicates that these processes might interplay with dietary interventions, a hypothesis that we test in the following section. Beyond genomic instability (a hallmark of ageing) the transcriptional discrepancies between cells with cryptic mtDNA load and those without are consonant with four further hall-

marks of ageing (loss of proteostasis, deregulated nutrient-sensing, mitochondrial dysfunction, and altered intercellular communication).

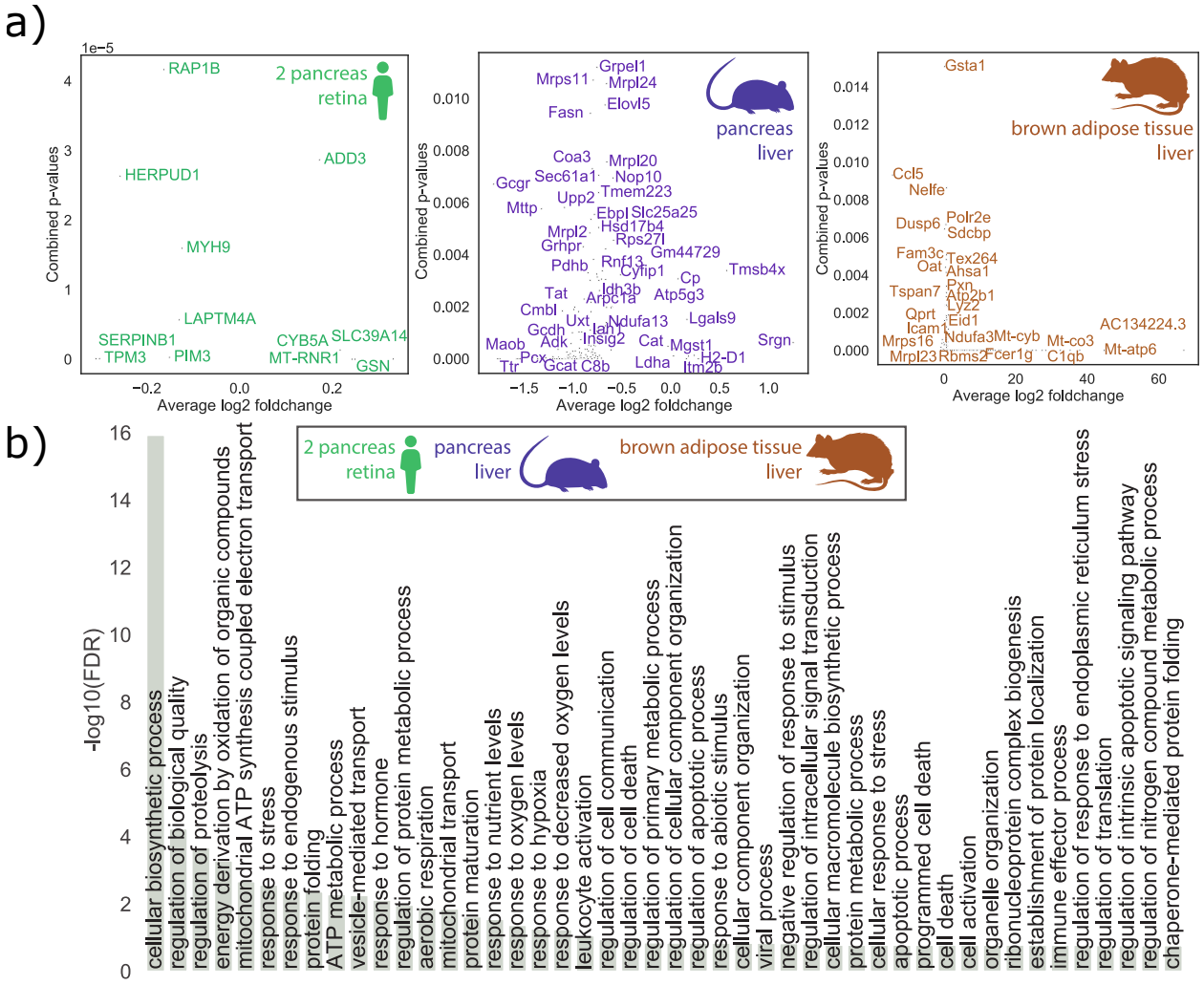


Figure 3.7: Cryptic mtDNA mutations are associate with numerous functional effects (reproduced from [145] with permission of collaborators). a) Differential expression analysis was performed on each data set separately. We display the volcano plot for differential gene expression analysis in human, mouse, and rat datasets respectively. b) GO-term enrichment analysis was performed using PANTHER on each of the seven datasets separately [148, 145]. GO terms present in at least six of the seven datasets are displayed, ordered by their False Discovery Rate as found by the analysis of data from [57]

3.7 Conclusion

We have developed a sophisticated, end to end, bioinformatics pipeline that is capable of taking raw, single cell, sequencing data in the form of `.fastq` files, aligning the data to the reference genome to produce a `.bam` file. Then, unlike previous bulk pipelines, we demultiplex reads in the `.bam` file to assign them to different cells, identify mutations within them through variant calling, and then classify them based on their prevalence within the data. Ultimately, this allowed us to robustly identify *cryptic* heteroplasmic mutations within single cells. Moreover, because this was expression data, our pipeline is able to give both cellular phenotype and mitochondrial genotypes, allowing us to identify potential links between cryptic mitochondrial mutations and cellular function. The pipeline can be applied to a range of single-cell datasets with only minor modifications. Using a simple model from population genetics, we were able to make and test high-level predictions about how the distribution of cryptic mutations — the cryptic site frequency spectrum — would evolve through a human life course. The dynamics of cellular populations are controlled by mtDNA half-life, cellular population size, and Polgγ per base error rates. Values based on previous literature showed that cryptic mutations with heteroplasmy exceeding a pathological threshold would start to emerge on a timescale consistent with the decline of human health. If this description is correct, such mutations ought to be visible in single-cell datasets, alongside their impact on cellular function.

Predictions based on the Moran model were then tested by application of the pipeline we developed. When applied to different single-cell modalities we did uncover the limitations of some approaches which fail to cover large proportions of the genome at sufficient depth to pass sensible quality control filters. However, the regions that were covered could still be used to investigate our research questions in the rise of mtDNA heteroplasmy with age. The pipeline was applied to single-cell RNA sequencing datasets from a variety of different tissues, in individuals of a broad range of ages. Focusing on data from a single source modality allowed us to observe how mutant load systematically increased with age, inline with model predictions. We showed that cryptic, single-cell mutations are typically missed in bulk sequencing approaches, yet the majority of mitochondrial mutant load is indeed cryptic in nature, highlighting the contribution

of this work to the literature, and how the development of the pipeline allowed this to be discovered. Furthermore, the empirically derived, RNA based, cSFS did show an age-dependent accretion in mtDNA heteroplasmy, with cells from old individuals harbouring homoplasmic mtDNA mutations absent in younger individuals. The difference in the cSFS between young and old was systematically demonstrated by comparing all pairs of individuals analysed within our data.

Next, through the application of our pipeline, we gave indicative evidence that mtDNA copy number, which increases in calorically restricted rats, could be used to control the increasing heteroplasmy of cryptic mitochondrial mutations, in line with theory. Caloric restriction is also known to attenuate the onset of ageing. Combined, these two observations are suggestive that impeding the accumulation of somatically acquired mtDNA mutations may play a role in the ageing process, and that this can, at least partially, be ameliorated.

Finally, I summarised an analysis conducted by my colleague, Dr Florian Klimm, investigating the functional implications of mtDNA mutations. This analysis showed that previously unseen cryptic mutations are consistent with effects contributing to many of the functional consequences of ageing. Moreover, many functional effects were recapitulated across species and tissues in the study.

Combined, these results paint an image of accumulating somatic mtDNA mutations as a form of mosaicism of mitochondrial diseases at the single-cell level, consistent with many of the established hallmarks of ageing. Through the use of a novel bioinformatics pipeline, we have been able to leverage existing single-cell genomics data, shedding light on this previously unexplored facet of mammalian ageing.

Chapter 4

Bayesian Inference As A Tool For Low Depth Data

"... no models are [true]—not even the Newtonian laws. When you construct a model you leave out all the details which you, with the knowledge at your disposal, consider inessential.... Models should not be true, but it is important that they are applicable, and whether they are applicable for any given purpose must of course be investigated. This also means that a model is never accepted finally, only on trial."
– Georg Rasch [152]

4.1 Summary

Thus far, we have shown how mutations in mitochondrial DNA play a role in health and disease. In chapter 2, we detailed how nascent single-cell technologies give the opportunity to explore the impact of mitochondrial mutations at the fine-grained resolution of the single cell. In chapter 3, we leveraged publicly available single-cell data to gain novel insight into the role that mitochondrial mutations play in ageing. This was achieved through the development of a bioinformatics pipeline that allowed us to extract cell-specific heteroplasmy estimates. Using this data, we demonstrated that: single-cell technologies allow us to identify *cryptic* mutations, only found within single cells; that most mitochondrial mutations are missing from bulk analyses and are cryptic in nature, which accumulate with age; and that the rise of these mutations is linked with several hallmarks of ageing [115].

Beyond these successes, we have also seen some of the limitations imposed by current single-cell technologies. In particular, we have seen how *depth of coverage* varies from cell to cell, and how the typical *breadth of coverage* varies between technologies (see sections 2.4 and 3.6.1). More generally, low depth of coverage in a given cell, at a particular position poses a limit on our ability to accurately estimate its heteroplasmy.

Collaborators at the University of Cambridge's Mitochondrial Biology Unit presented us with a novel dataset that poses this particular challenge. Human embryonic brain data was sequenced using 10x's ATAC sequencing modality. Such samples were frozen, limiting the ability to extract high quantities of mitochondrial reads.

In this chapter, we address the limitations imposed by low-depth. By employing Bayesian inference, we seek to quantify the uncertainty of heteroplasmy estimates in any given cell. Moreover, we shift our attention away from cryptic mutations to those found in multiple cells, with a view to measuring the association between the heteroplasmy of a particular mitochondrial mutation, and the activity of each gene. Through the use of modern computational Bayesian inference, we demonstrate *in silico* how we can recover these associations, even in very low-depth data, and that the posterior estimates we obtain are well calibrated. Finally, we address multiple

testing concerns of our procedure through an extended application of the Benjamini-Hochberg procedure to Bayesian posterior quantiles of the "null" parameter value. This is a novel extension of the Benjamini-Hochberg procedure, and we give justification for application to our model. We further demonstrate that this procedure works, and controls the false discovery rate *in silico*.

4.2 Introduction

Single-cell sequencing modalities vary in their capacity to capture information about the mitochondrial genome, with the depth and breadth of coverage varying dramatically. Frustratingly, high-throughput sequencing modalities, such as 10x 3' sequencing techniques (see section 2.4), which are widely available and capable of sequencing 1,000s of cells, have some of the most restricted breadth of coverage of the mitochondrial genome, with only $\sim$ 150 bps at the 3' end of a gene typically getting sequenced to any meaningful depth of coverage. A further nuance is that the depth of coverage — the number of sequenced reads — at a particular position of the mitochondrial genome can vary dramatically across cells within a single sample.

In previous chapters, we estimated heteroplasmy within a single cell by the simple formula

$$H_{i,j} = \frac{NM_{ij}}{N_{ij}} \tag{4.1}$$

Where

$$\begin{aligned} H_{ij} &:= \text{The heteroplasmy for cell } i \text{ at position } j \\ NM_{ij} &:= \text{The number of mutated mtDNA reads sequenced in a cell } i \text{ at position } j \\ N_{ij} &:= \text{The total number of sequenced mtDNA reads within cell } i \text{ at position } j \end{aligned}$$

Under ideal conditions, a complete census of all the mtDNA in the cell, covering each position of every strand of mtDNA would be performed. In real laboratory conditions, however, this aspiration is far from realised. Good quality datasets leave us with 100-1,000s of reads per cell. Under such conditions, formula 4.1 can reliably be used without much further consideration as the sampling error will be small. Section 4.6.1 explores this further. However, even in these high-quality data sets, there will be positions of the genome in particular cells which fail to

attain a high level of depth. Thus far, we have simply dispensed with these cells and positions from analysis.

In lower-quality datasets, few positions or cells may reach this desirable standard. In some settings, as will be seen in the next chapter, ubiquitous shallow depth is unavoidable. This includes where tissues are frozen as part of sample preparation, reducing the capture rate of mtDNA in standard single-cell procedures. Under these settings, having few reads at any position is the norm. When the depth of coverage gets to be as low as ten reads per cell, information about the heteroplasmy of a mutation in a cell is present, though it is deeply uncertain. For instance, if $NM_{ij} = 1$ and $N_{ij} = 2$, it makes little sense to proceed with an analysis predicated on $H_{ij} = \frac{1}{2}$, which fails to incorporate any of the uncertainty in that estimate.

This chapter outlines a statistical and computational methodology to address this problem in anticipation of the next chapter, where the methodology developed here will be deployed on real data. We investigate the nuances and resolve issues posed by the variation in single-cell sequencing data. We begin with an overview of the tools used in modern Bayesian inference. Though the basic elements of Bayesian statistical analysis are straightforwardly embodied in Bayes' theorem, the computational toolkit essential to modern Bayesian analysis, not to mention the associated jargon, can be daunting, even to those well-versed in the statistical literature. As such, we aim to give the reader a basic outline of these modern tools. We give a brief overview of the rudiments of Bayesian inference, immediately followed by a simple explanation of *Bayesian Networks*, which are immensely useful in representing these models and will be leaned on heavily throughout this chapter. Computational tools, especially Markov Chain Monte Carlo (MCMC), and a particular flavour *Hamiltonian Monte Carlo* will be described in detail. These algorithms are the bread and butter of modern Bayesian analysis and are essential for performing Bayesian inference on all but the most rudimentary models [153, 154, 23]. The use of these algorithms to fit a variety of complicated models has been made simpler through the development of freely available probabilistic programming languages (PPLs). A comprehension of these algorithms, and particularly their computational demands is essential for grasping the technical challenge of deploying Bayesian inference 100,000-1,000,000 of times. This is the scale demanded in the

context of genetics where associations must be measured to the approximately 20,000 genes in the genome, in multiple datasets. Running even a single model can be computationally demanding, so appropriate approaches to software engineering are required to address inference at this scale.

After an outline of the basic toolkit of modern Bayesian analysis, we proceed to solve our problem: the estimation of the association of heteroplasmy on gene activity in low-depth single-cell sequencing data. This starts with a simple use of the binomial model to estimate the heteroplasmy in any given cell, illustrating how the Bayesian approach naturally allows us to represent the uncertainty for each given cell. We follow this by showing how gene activity can be modeled with a negative binomial distribution, giving a generative justification for this choice of distribution. The estimation of uncertain heteroplasmy and gene activity are then combined into a single negative binomial regression model, with heteroplasmy as a latent predictor variable.

Throughout this process, we go through the steps of model development and strive to abide by modern standards of principled *Bayesian workflow* [24, 155, 154, 153]. *Bayesian inference* can be considered to simply be an application of Bayes' rule to a statistical model, $P(\theta|D) = \frac{P(D|\theta) \times P(\theta)}{P(D)}$, in which the *posterior* distribution, $P(\theta|D)$ over model parameters is obtained by conditioning the model on the data (detailed further below). *Bayesian workflow*, however, refers to an expanded set of concepts related to the application of Bayesian inference. This includes: the choice and construction of a statistical model; assessment of a model's adequacy to solve the problem at hand, and if possible performing model comparison; understanding of any particular model's limitations, such as when it might be expected to fail; verification that the model is *identifiable*; rigorous testing of the computational procedures used to fit the model and approximate its posteriors; ensuring the PPL code used in implementation is error free; and finally assessing what ways a model might feasibly be improved to solve the problem at hand [24].

Finally, we round the chapter off by demonstrating how we can address methodological concerns surrounding repeated use of the model. The model we develop in this chapter measures the

association between the heteroplasmy of a particular mutation and the activity of a particular gene. In itself, this is an innovation, allowing us to extract insight from data that would otherwise be considered unusable. There are, however, approximately 20,000 genes (including encoded in the nucleus and mtDNA together), all of which the heteroplasmy of a mutation could plausibly affect. Developing a single, plausible Bayesian model to capture all of these interactions is an elusive objective, especially in conjunction with the computationally daunting task of obtaining posterior samples in such a high dimensional space. In the face of these constraints, repeated application of the same model to mutation-gene activity pairs to measure their association is a pragmatic approach. Issues surrounding multiple hypothesis correction in frequentist approaches can be dealt with in several ways, with popular approaches such as the Bonferroni correction, controlling the family-wise error rate (FWER), or the Benjamini-Hochberg (BH) procedure, which aims to control the false discovery rate (FDR) [26]. In the Bayesian literature, no equivalent, default procedure fulfills the same role. We bridge this gap in our research by application of the Benjamini-Hochberg procedure to posterior quantiles of the null parameter value.

The reader might, at this point, be concerned about the fallacious transfer of a method from one school of statistics to another. As such, we are careful to demonstrate the validity of this procedure, verifying its results *in silico*. Some Bayesian models possess the property that when they are repeatedly fit to data generated by the null, the posterior quantiles of the null value are uniformly distributed. In particular, this behavior often occurs asymptotically, when the model is fit to large amounts of data. In so far as the posterior quantiles of the null are uniform, they behave analogously to p-values, which are uniform under the null, by definition. As the BH procedure relies on this uniformity to control the FDR, we show that it can be readily applied to Bayesian posterior quantiles which also possess this quality to control the FDR. Furthermore, we show that the regression model we develop to measure the association between heteroplasmy and gene activity displays this quality. Though the understanding of this approach is nuanced, it is straightforward to apply and enables us to run 1,000s of Bayesian models in parallel on multiple disconnected CPUs. A single model aspiring to represent all genes and their interactions would not offer the possibility of such parallelism.

Once all this machinery is in place, we are able to proceed to the next chapter, which applies this material to frozen, embryonic, brain single-cell sequencing data. Low depth is inherent to the preparation of such frozen samples. This approach bears fruit by allowing us to obtain an understanding of the functional effects of mitochondrial mutations in early development.

4.3 Background On Bayesian Inference

4.3.1 Motivating The Bayesian Approach

$$P(H|D) = \frac{p(D|H)p(H)}{P(D)} \tag{4.2}$$

The heart of Bayesian Inference is embodied in equation (4.2). The terms in the equation are as follows:

- The *Posterior* Probability of a hypothesis given the data, $P(H|D)$
- The *likelihood* of the data given the hypothesis $p(D|H)$
- The *Prior* probability of the hypothesis before seeing the data $p(H)$

In the context of Bayesian modeling, the set of hypotheses are typically parameter values in a parametric model. In this setting, one specifies a likelihood of data given a parameter value, $p(y|\theta)$, with a prior distribution on parameter values, $p(\theta)$. As such, Bayesian models are specified by the joint distribution of parameters and data $p(y, \theta)$. We then condition on observed data to obtain posterior distributions over parameter values within the model. According to Bayesians, probability in this context is interpreted as the *credibility* placed on different parameter values.

The virtues of the Bayesian approach to statistical inference are numerous. For the current context, however, the value arises in allowing us to specify sophisticated generative models of our data, and quantify the uncertainty in the parameter values of that generative model by conditioning on the observed data. Qualitatively, the generative mechanism can be conceptualised as each cell having a true heteroplasmy. This heteroplasmy, through some complicated set of unobserved causal mechanisms, has an influence on the activity of many genes, inducing a form of association between heteroplasmy and gene activity. Independent of this, the true cell heteroplasmy is also responsible for generating the number of observed mitochondrial sequencing reads, which are seen to carry a mitochondrial mutation. We can make the structure

of this generative mechanism more concrete through the use of Bayesian networks, to which, we now turn our attention.

4.3.2 Bayesian Networks

In order to both state our models and aid interpretation, we make extensive use of probabilistic graphical models. Joint probability distributions of the form $p(X_1, ..., X_n) = p(\underline{\mathbf{X}}); X \in \mathbf{R}^n$ often lend themselves to different forms of representation. All probability distributions may be factorized as $p(\underline{\mathbf{X}}) = \prod_i^n p(X_i|X_1, ..., X_{i-1})$, known as the *chain rule* of probability theory. Such a factorisation may be performed in any order, yielding alternative factorisations of the same distribution. Any particular, valid, factorisation can be represented as a directed acyclic graph (DAG). The nodes of the graph represent components of a random vector $(X_1, ..., X_n)$. The edges are used to represent conditional dependencies in the factorised distribution. This is done such that directed edges feed into node X_i from all other variables upon which it depends in a particular factorisation. To illustrate, we can use (X, Y, Z) as our vector and factorise the distribution, which we write as $p(x, y, z) = p(x)p(y|x)p(z|x, y)$. This particular factorisation can be represented by a DAG with nodes labeled X, Y, and Z. X has no incoming edges as it is not dependent on any other variables in this particular factorisation. Y has an incoming edge from X due to its dependence on X seen in the $p(y|x)$ term of the factorisation. Similarly, the term $p(z|x, y)$ tells us that edges feed from X, and Y into Z. Put together, this yields the following DAG:

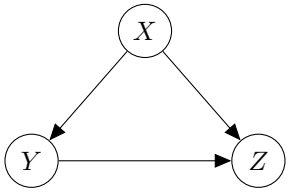


Figure 4.1: DAG representing $p(x, y, z) = p(x)p(y|x)p(z|x, y)$.

In the connection between two nodes, the node from which the edge emanates is referred to as the *parent*, and the node it enters is referred to as the *child* node. For instance, in the DAG in Figure. 4.1, X is said to be a parent of Y, and also a parent of Z. Similarly, Y is said to be child of X, and Z is a child of both X and Y.

Alternatively, we could instead factorise the distribution as $p(x, y, z) = p(y)p(z|y)p(x|y, z)$. This particular representation would then be represented by this DAG:

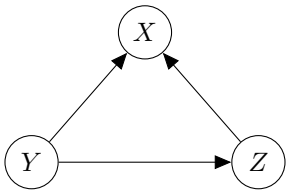


Figure 4.2: DAG representing $p(x, y, z) = p(y)p(z|y)p(x|y, z)$.

Factorisations of the form $p(\underline{\mathbf{X}}) = \prod_i^n p(X_i|X_1, ..., X_{i-1})$ always imply fully connected DAGs in which each node connects to all other nodes in the network. Some distributions however have conditional independencies between some variables . For instance, it might be possible to write $p(x, y, z) = p(x)p(y|x)p(z|x, y) = p(x)p(y|x)p(z|y)$, in which Z is independent of X given Y. Such a conditional independence relationship is denoted by $(Z \perp\!\!\!\perp X \,|\, Y)$ [1]. The corresponding DAG for a distribution that permitted such a factoristion would be as follows:

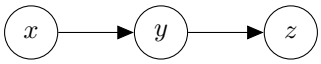


Figure 4.3: DAG representing $p(x, y, z) = p(x)p(y|x)p(z|x, y) = p(x)p(y|x)p(z|y)$.

Just as a particular factorisation implies a DAG, so does a DAG imply a particular factorisation of a joint distribution, with all of its associated conditional independencies. Representing dependencies in distributions as DAGs is immensely useful for being able to interpret the conditional dependence relations in a multivariate distribution at a glance.

Bayesian models are represented as multivariate probability distributions over un-observable latent parameters $\underline{\theta}$, and observable data $\underline{y}$. We can then visualise conditional dependencies in Bayesian models by use of DAGs, which facilitate comprehension of our models. We will make use of Bayesian networks throughout the rest of this text, using white coloured nodes for parameters, and grey nodes to for nodes representing data. In the context of Bayesian inference, we condition on the data, and thus shade the grey nodes in the model. DAGs will allow us to think clearly about the data-generating mechanism by showing the dependence relationships between model parameters and produced data.

[1] The converse statement of dependence would be written as $(Z \not\perp\!\!\!\perp X \,|\, Y)$

4.3.3 Probabilistic Programming Languages

Many posterior distributions either cannot be obtained analytically, or involve performing computationally prohibitive, high-dimensional integrals. As such, a variety of computational techniques have been developed that allow us to make approximations of the posterior distribution over parameters in our models. We primarily employ one such approximation method throughout this work — Markov Chain Monte Carlo (MCMC), specifically a flavour of MCMC referred to Hamiltonian Monte Carlo (HMC). Computational frameworks exist that allow us to specify complicated probability distributions and obtain samples from them, typically either through MCMC or variational inference. Such frameworks are referred to as probabilistic programming languages, (PPLs). A popular example of such a PPL, which we will be making extensive use of, is `stan` [23].

The utility of many PPLs stems from automating many of the steps between model specification, (ie, writing down the formula of the joint probability model) and obtaining samples from the posterior distribution of parameters conditioned on the data. In particular, many PPLs employ auto-differentiation libraries to compute gradients with respect to the parameters of a model's log probability distribution. When you define a probabilistic model in a PPL, you implicitly specify a function that computes the log probability for a given set of parameter values and observed data. As such, PPLs can efficiently obtain gradients for complex models without requiring users to manually derive and implement these gradients, which can be error-prone and time-consuming. The precise reasons for needing the gradients of the log-probability distribution will become clearer once we have provided a sketch of how HMC works, which we now turn to.

4.3.4 MCMC

One of the primary tasks in statistical computing is to be able to obtain samples from a broad class of distributions. A popular family of approaches to this problem are Markov Chain Monte Carlo Methods. The archetypal example of such an algorithm is the Metropolis-

Hastings algorithm [156]. The problem can be posed as follows: we have a function, that when normalised appropriately, forms a probability distribution that we wish to obtain numerical samples from $h(x)$. In general, a valid probability distribution can be obtained by dividing it by $\int h(x)$, though evaluating this integral might be prohibitively costly in high dimensions, so we need not normalise. This algorithm proceeds as follows:

1. Give x some initial state, X_i.
2. Draw a sample $p(x_{i+1}|x_i)$ using a conditional distribution that we are already capable of drawing random samples from.
3. Compute $a(x_i, x_{i+1}) = min(\frac{p(x_{i+1}|x_i)h(x_i)}{p(x_i|x_{i+1})h(x_{i+1})}, 1)$
4. Sample with $I \sim Bernoulli(a(x_i, x_{i+1}))$.
5. If $I = 1$, *accept*, setting $X_{i+1} = x_{i+1}$ to be the proposed value. Else, *reject*, once again setting $X_{i+1} = x_i$ to its current value

By repeating this procedure, we can build up a *Markov Chain* $X_1, X_2, ..., X_{i+1}, .., X_n$. It can be shown that the *stationary distribution* of a Markov chain produced in this manner will be $p(x) = \frac{h(x)}{\int h(x)}$, the normalised distribution we were initially interested in. Another way of stating this is that given a conditional distribution for proposing transitions that we can sample from,$p(y|x)$, we can create random samples from other target distributions, $h(x)$, of interest. This is particularly helpful in the context of Bayesian inference where many of our posterior distributions are often written as $P(\theta|D) = \frac{p(D|\theta)p(\theta)}{\int p(D|\theta)p(\theta)d\theta} \propto p(D|\theta)p(\theta)$, as the integrals in the denominator are either analytically intractable or have prohibitive computational costs associated. In such situations, we can make use of the Metropolis-Hastings algorithm to generate Markov Chains with stationary distributions identical to the posterior distribution obtained for our problem [157, 158, 159, 153].

One of the problems posed by such algorithms is the form of the proposal distribution used to construct the Markov Chain, $p(x_{i+1}|x_i)$. For instance, consider that the proposal distribution was decided to have the form, $p(x_{i+1}|x_i) = Norm(x_i, \sigma)$. What is an appropriate value for

σ? Such parameters for the proposal distribution alter the rate at which the Markov Chain will converge to its stationary distribution. Intuitively, this is reflected in the acceptance rate of our proposals. Large proposed steps sizes often have lower acceptance rates, as regions far from a modal region typically have a lower probability density than regions near to them. A such, large values of sigma result in more rejected proposals, and consequently more time in the same position. Conversely, reduced step sizes tend to have greater acceptance rates, though this comes at the expense of slowly exploring the space the distribution lives in. Either of these two extremes reduces the rate at which the distribution converges to the stationary distribution. With these considerations in mind, a desirable proposal distribution has both a high acceptance rate and explores large areas of the state space. Random walk methods which neglect the geometry of a distribution are limited in this regard.

Hamiltonian Monte Carlo (HMC) is one such approach for attaining higher acceptance rates when sampling whilst exploring the parameter space. Rather than attempting random walks like those found in the Metropolis-Hastings algorithm, HMC performs a complex set of operations to intelligently explore the posterior distribution. HMC borrows ideas from statistical mechanics in physics. For some physical system with states $\underline{X}$, and an Energy function for those states, $E(\underline{X})$, the canonical probability distribution over the system's states is given by $P(\underline{X}) = \frac{1}{Z} exp\left(\frac{-E(\underline{X})}{T}\right)$. In physics, a system with momenta and position in a potential energy field has its energy function specified in the form of a Hamiltonian. For many problems, this takes the form $H(\underline{p}, \underline{q}) = U(\underline{q}) + k(\underline{p})$ where $\underline{q}$ characterises the system's physical coordinates and $\underline{p}$ characterises its momenta. $U(\underline{q})$ characterises the systems potential energy and $k(\underline{p})$ the particle's kinetic energy. This yields the canonical distribution for the system as being

$$P(\underline{q}, \underline{p}) = \frac{1}{Z} exp\left(\frac{-H(\underline{q}, \underline{p})}{T}\right) = \frac{1}{Z} exp\left(\frac{-U(\underline{q}) - V(\underline{p})}{T}\right) = \frac{1}{Z} exp\left(\frac{-U(\underline{q})}{T}\right) exp\left(-\frac{V(\underline{p})}{T}\right)$$

with factorisation of the kinetic and potential energy terms implying that the position and momenta variables are independent random variables. If we are given a generic probability distribution, we infer a corresponding energy function to be $E(x) = -logP(x) - logZ$. In Bayesian inference, the corresponding potential energy of the system is given by the unnor-

malised posterior

$$U(\underline{\theta}) = -log(p(D|\underline{\theta})p(\underline{\theta})) \tag{4.3}$$

with the $\underline{\theta}$ now playing the role of the coordinates of the physical system. For each variable in the distribution θ_i, corresponding momentum variables ϕ_i, are added as auxillary variables, which is acceptable due to the independence of the energy and momenta terms in the canonical distribution. Momenta $\underline{\phi}$ are drawn from the multivariate normal distribution with diagonal covariance and zero mean. The logarithm of this momenta distribution yields a corresponding kinetic energy function in the canonical distribution, formed from the quadratic form in the Gaussian distribution's exponent

$$V(\underline{\phi}) = \sum_i \frac{\phi_i^2}{2m_i} \tag{4.4}$$

The mass terms, m_i, correspond to the standard deviations found in the covariance matrix.

Putting together equations (4.3) and (4.4), gives us a Hamiltonian for the posterior distribution obtained from Bayesian inference.

$$H(\underline{\theta}, \underline{\phi}) = U(\underline{\theta}) + V(\underline{\phi}) = -log(p(D|\underline{\theta})p(\underline{\theta})) + \sum_i \frac{\phi_i^2}{2m_i} \tag{4.5}$$

The HMC algorithm updates coordinates as follows: Firstly $\underline{\phi}$ is sampled from the MVN with a diagonal covariance matrix. Secondly, given coordinates $(\underline{\theta}, \underline{\phi})$, Hamiltonian dynamics are performed on the system using *leapfrog* updates. Hamiltonian dynamics have a long history in physics. Given a Hamiltonian, a physical system's position and momentum coordinates evolve according to *Hamilton's equations*:

$$\frac{dq_i}{dt} = \frac{\partial H}{\partial p_i}, \tag{4.6}$$

$$\frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i} \tag{4.7}$$

Solving these equations analytically is seldom possible. As such, numerical approximation schemes are often deployed, such as the commonly used Leapfrog approach. Leapfrog updates approximate the evolution of the system according to the following equations:

$$p_i(t + \Delta t/2) = p_i(t) - \frac{\Delta t}{2}\frac{\partial H(q(t))}{\partial q_i} \tag{4.8}$$

$$q_i(t + \Delta t) = q_i(t) + \Delta t\frac{p_i(t + \Delta t/2)}{m_i} \tag{4.9}$$

$$p_i(t + \Delta t) = p_i(t + \Delta t/2) - \frac{\Delta t}{2}\frac{\partial H(q(t) + \Delta t)}{\partial q_i} \tag{4.10}$$

In HMC, these leapfrog steps are performed for L steps, which in conjunction with the step size, Δt, are the parameters to be tuned by the algorithm to obtain good performance. The position coordinates at the end of this L-step path, along with the reversed the momenta coordinates at the end of this trajectory yield a proposal state $(\underline{\theta}*, \underline{\phi}*)$. In an analogous way to step 3 of the Metropolis-Hastings algorithm, proposals are accepted according to the ratio of probabilities at the inital and proposal states, equivalent to:

$$min(exp(-H(\underline{\theta}*, \underline{\phi}*) + H(\underline{\theta}, \underline{\phi})), 1) \tag{4.11}$$

Throughout Hamiltonian dynamics, $H(\underline{\theta}*, \underline{\phi}*)$ is a conserved quantity, meaning the Hamiltonian update steps move around iso-probability contours of the posterior (ie, the energy function/log posterior, is held constant in the second step). Different contours are only explored

by the randomised updates of the $\underline{\phi}$ terms. As such, the HMC algorithm yields much higher acceptance rates than the standard Metropolis-Hastings procedure. This is particularly relevant to high-dimensional problems. It can be shown that the two steps of the HMC algorithm yield Markov Chains with stationary distributions identical to those of the canonical distribution being employed. The main reason these results work is due to the reversible nature of Hamiltonian dynamics, which ultimately leads to detailed balance being established[2] [156].

Beyond the efficiency of sampling in high-dimensional probability spaces, HMC offers further benefits that we will bring to bear in our analyses. As the leapfrog steps are performed, the Energy/Hamiltonian of the system should be preserved throughout the trajectory. When regions of high curvature are encountered, however, the discretization inherent in the leapfrog procedure can fail to keep the energy constant, resulting in *divergent transitions* as the leapfrog steps proceed. Whilst such divergent transitions might seem like shortcomings of the discretization procedure, they can be inherently helpful in identifying broader issues with our ability to obtain samples from the model's posterior that are often associated with such regions of high curvature [160]. Benign regions of high curvature can be explored effectively by tuning the step size and the number of leapfrog updates in the Hamiltonian update steps. More pathological regions of extreme curvature can result from degeneracies in the model's specification, with regions of small volume in the parameter space possessing an extremely high probability density. Typical Markov transitions fail to explore these narrow regions of parameter space effectively. Such pathologies often occur in hierarchical Bayesian models, where sharp funnels in parameter space often arise [161]. Such problems are flagged by divergent transitions in HMC samplers relying on leapfrog steps, with the sampler making the user aware of its limitations. Samplers such as the Metropolis-Hastings fail to alert the user that they are not effectively exploring the parameter space in such instances. In some models, such regions of high curvature in the posterior can simply result from a certain choice of parameterization of the model. This problem can then be overcome by reparameterizing the model [153, 154, 160]. The ability of numeri-

[2]Detailed balance ensures that the Markov chain satisfies a key condition for reversibility. If π_i is the stationary probability of being in a state, i, and $p_{i,j}$ is the probability of transition from state i to j, then the detailed balance condition reads as $\pi_i p_{i,j} = \pi_j p_{j,i}$. Along with some other minor assumptions, showing that π_i and $p_{i,j}$ satisfy the detailed balance condition is sufficient to show that π_i is the stationary distribution of a Markov chain produced by proposals $p_{i,j}$. This relationship can be used in proving the convergence of many MCMC algorithms [156].

cal approximations to raise errors in such contexts should therefore be seen as advantageous, especially in the context of model development, where modelers may unintentionally produce novel models that possess unforeseen degeneracies.

In addition to this, there are other standard MCMC diagnostics to sanity check our models behaviour. Effective sample size (ESS), is a measure of the number of independent samples obtained from the MCMC chains, adjusting for auto-correlation present in the Markov chain [153]. It is used to evaluate the efficiency of the sampling process and is defined as:

$$\text{ESS} = \frac{N}{1 + 2\sum_{t=1}^{T} \rho_t} \tag{4.12}$$

where

- N: the total number of samples,
- T: the time-lag where the autocorrelation function becomes negligible,
- ρ_t: the autocorrelation at lag t.

$\hat{R}$, also known as the Gelman-Rubin statistic, another diagnostic statistic, relies on running multiple MCMC chains in parallel [162]. It then compares the between-chain variance and the within-chain variance to determine if the chains have mixed well and are exploring the same target distribution, meaning they have converged to the stationary distribution [153]. It is defined as:

$$\hat{R} = \sqrt{\frac{\frac{n-1}{n}W + \frac{1}{n}B}{W}} \tag{4.13}$$

where

- n: the number of iterations per chain,
- B: the sample estimate of between-chain variance,
- W: the sample estimate of within-chain variance.

As is clear from the definition, $\hat{R}$ measures what proportion of the total variance in an MCMC run is determined by the the within chain variance. If the MCMC has converged to the stationary distribution, then $\hat{R}$ should be very close to one. If the chains have yet to converge, then $\hat{R} > 1$, implying that any estimates of posterior parameters based on these MCMC samples would be unreliable.

To illustrate how we can incorporate this diagnostic information, we generate data from a standard linear regression model of the form:

$$x_i \sim Normal(0, 1)$$
$$y_i \sim Normal(1 + 2x_i, 1)$$

We specify an identifiable Bayesian model to fit this data:

$$y_i \sim Normal(\alpha + \beta x_i, \sigma^2) \tag{4.14}$$
$$\tag{4.15}$$

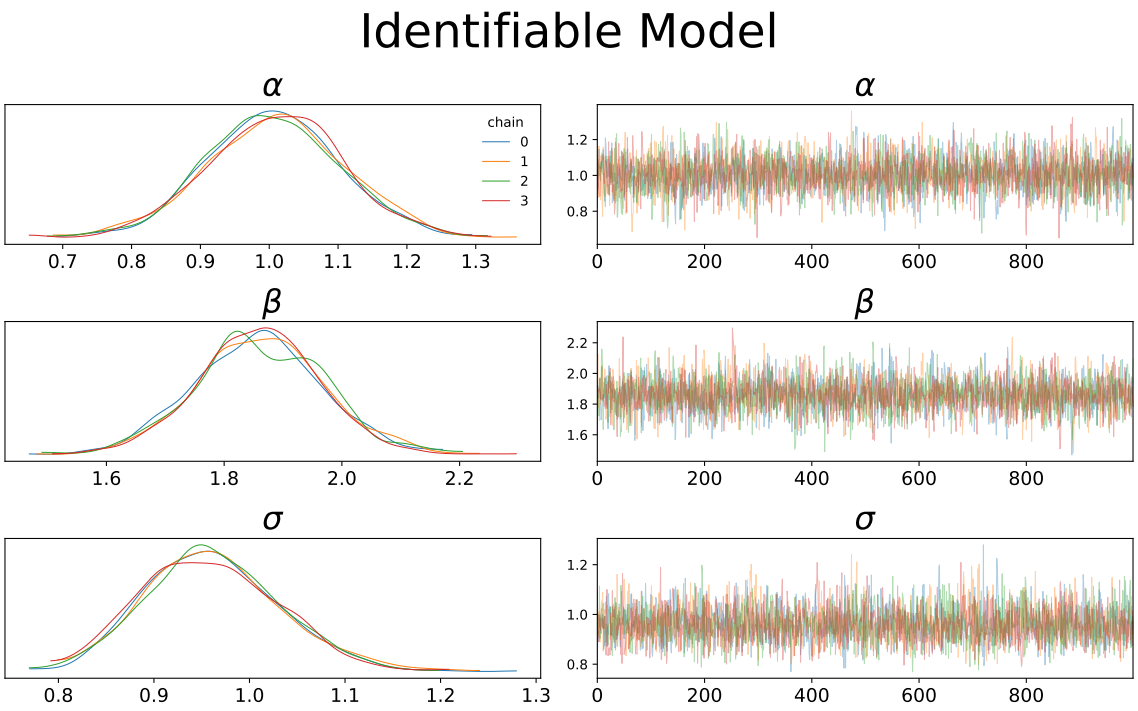
We can specify a similar, non-identifiable Bayesian model with two intercept terms in the model:

$$y_i \sim Normal(\alpha_1 + \alpha_2 + \beta x_i, \sigma^2) \tag{4.16}$$
$$\tag{4.17}$$

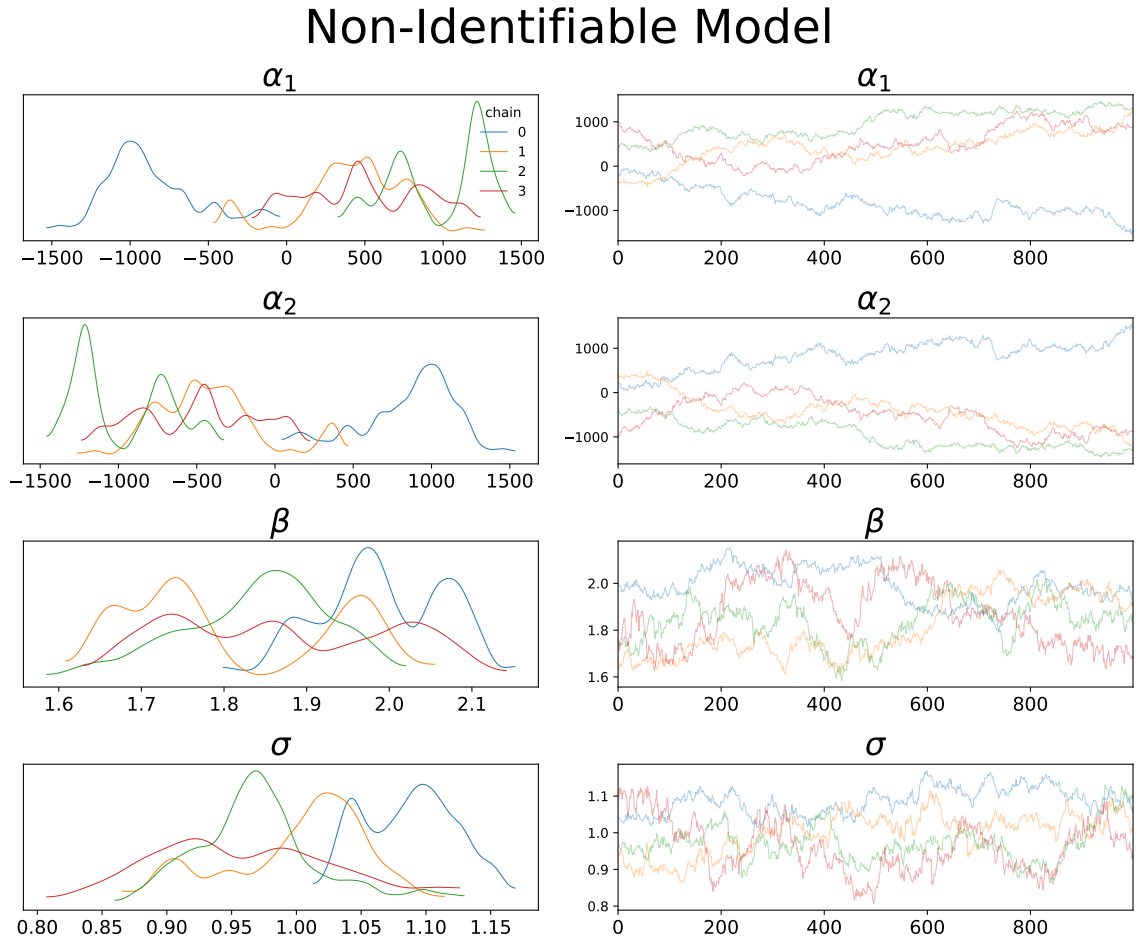
The model is pathological as there is an infinitely long line $\alpha_1 + \alpha_2 = c$ for which the likelihood is constant, with no way to prefer one point over another. The model is therefore *unidentifiable*.

We fit both of these models to identical simulated data, running four separate HMC chains for each model. The trace plots are visualized in Figure. 4.4. We ran these chains for 1,000 iterations each. In the left column of the figure, we have the distribution of α, β, σ parameter values from the sample run. We can see clearly from the identifiable model how the distribution of samples for all chains is more or less indistinguishable. In the right-hand column, we can see the path of the chains through time. The chains exhibiting *mixing*, as each chain appears the thread through one another, each being barely distinguishable from any other. Contrast this with the non-identifiable model. In the left-hand column, we can see that the distribution of samples from each chain shows almost no agreement. Upon inspecting each chain's path through the sample space in the right-hand column, we can see why. Each chain got stuck in a local region of the parameter space surrounding constant values of each α parameter. That the sampler attempts to follow the $\alpha_1 + \alpha_2 = c$ line can be seen in the fact that the chains in $\alpha 1$ and $\alpha 2$ are almost perfect mirror images of one another about the horizontal axis of the page, whilst the other parameters show no such clear relationship.

The computed ESS was in excess of 2,000 for each parameter in the identifiable model, whilst it was below 50 in the non-identifiable model, indicating the high auto-correlation present in the chains, as the sampler fails to obtain independent samples from the latter posterior. Additionally, the $\hat{R}$ statistic is 1.0 for each of the parameters in the identifiable model, whilst the non-identifiable model had $\hat{R}$ statistics in excess of 1.2, for all parameters, picking up on the poor mixing observed between chains. By making use of statistics like $\hat{R}$, we can check for mixing in the MCMC chains, to ensure they are all exploring a well-defined target distribution appropriately. We will therefore make use of HMC in our model development to act as a sanity check, lending greater confidence that our models are well-identified, and do not exhibit any major pathologies.



(a)



(b)

Figure 4.4: (a) A properly identifiable Bayesian model exhibits mixing and stationarity in its Markov chains. (b) Non-identifiable models typically exhibit pathological behaviour in many ways. Here it can be seen that the posteriors for the two degenerate alpha terms are on a massive scale compared to the true values. In general, chains for all parameters did not mix, with intercept terms in the model seldom even overlapping.

4.3.5 Concluding Remarks On The Advantages Of The Bayesian Approach

Having given an outline of Bayesian inference and the computational tools widely used in its implementation, I conclude this subsection with further remarks about why the Bayesian approach was chosen over other seemingly simpler approaches such as MLE. To begin with, the model we construct in the following section will be high dimensional to incorporate the uncertainty in the heteroplasmy of each cell in the data. At the same time, each heterosplasmy will be used as a predictor variable in a regression model. We will be primarily interested in quantifying the uncertainty of a subset of model parameters, namely the regression coefficients, and not in all of the latenty heteroplasmy components. Bayesian inference offers a simple machinery for quantifying the uncertainty in a subset of components of a high dimensional model's parameter vector, $\underline{\theta}$, by marginalizing over nuisance parameters. Mathematically, this operation is simply integrating over all parameters that are not of interest: $p(\theta_i) = \int_{\underline{\theta}_{(-i)}} P(\underline{\theta}) d\underline{\theta}_{(-i)}$ where $\underline{\theta}_{(-i)}$ is the vector of parameters excluding component θ_i. This is simple to obtain numerically as we can approximate the marginal distribution θ_i by taking the corresponding component of the sampled parameter vectors that were obtained through MCMC.

An additional reason for using Bayesian inference is that the model we construct in the following section is non-standard. When constructing probabilistic models, especially high-dimensional ones, model developers can unintentionally create structures on which it is difficult to perform inference or valid inference is not actually possible. The computational framework used to perform Bayesian inference can detect these nuances, flagging problems in the model. This is reflected in the *folk theorem of statistical computing* which states: "When you have computational problems, often there's a problem with your model" [24]. Optimisation based approaches to inference do not in general highlight problems with a model in the same way. For valid models which might possess multimodality or other complex geometries, Bayesian inference with HMC can obtain samples from the full space and valid inference can be performed. Gradient-based inference methodologies such as MLE fail to capture this salient information in the model by getting stuck at a single local optimum. Furthermore, in invalid models that

contain non-identifiable parameters, gradient based approaches will keep iterating until their exit criteria are met and return a vector of parameter estimates. For example, MLE algorithms using iterative solvers will traverse the likelihood surface by updating parameter values until the change in the value of the likelihood is below some pre-specified threshold. As such, the user will not in general be aware of such problems in their model by a simple deployment of MLE. In contrast to this, when the same model is sampled using HMC, different chains of the MCMC run will not mix and convergence in estimators will be slow, resulting in an inflated $\hat{R}$ statistics allowing the user to identify these problems.

With these considerations in mind, we rely on the Bayesian inference framework as a natural tool for quantifying uncertainty and its computational framework to test the validity of our models. As we will be deploying the model many times, we will be consistently quantifying these diagnostic MCMC statistics to ensure the model's validity is maintained when the model is conditioned on different datasets.

4.4 Representing Heteroplasmy Uncertainty

Shallow depth constrained the number of mtDNA reads that we are able to obtain in many cells within the embryonic brain data our collaborators provided. Heteroplasmy is defined as the proportion of mtDNA molecules within a cell that harbour a mutation. As we cannot perform a census of all the mtDNA reads in a cell, and often obtain very few sequencing reads, there will be some uncertainty in the true value of the heteroplasmy of any given cell. This makes heteroplasmy a latent parameter, about which, we are uncertain. The literal definition of heteroplasmy a mutation would take on discrete values defined by $H_{true} = \frac{NumMut}{Total}$ where

$$H_{true-i} := \text{The true heteroplasmy for cell } i$$
$$NumMut_i := \text{The number of mtDNA mutated in cell } i$$
$$Total_i := \text{The total number of mtDNA within cell } i$$

We however do not have access to this data in a cell. Instead, we have access to

$$NM_i := \text{The Number of sampled mtDNA reads from cell carrying the mutation } i$$
$$D_i := \text{The total sequencing depth from cell } i \text{ at the position of the mutation}$$

Typical cells have of $Total_i$ $100s - 1,000s$, indicating that $D_i << Total_i$ for our embryonic brain data. As such, we can consider a cell to be a large reservoir of mtDNA relative to our sample size, which can be treated as sampling with replacement, with negligible detriment. This naturally leads us to model the number of mutated reads, NM_i, out of the D_i as being sampled from the binomial model with success parameter H_i. The corresponding Bayesian network can be seen in (Figure. 4.5). Choosing a uniform prior on heteroplasmy to reflect our uncertainty leads to the following structural model equations:

$$H_i \sim Beta(1, 1) \tag{4.18}$$

$$NM_i \sim Binom(H_i, D_i) \tag{4.19}$$

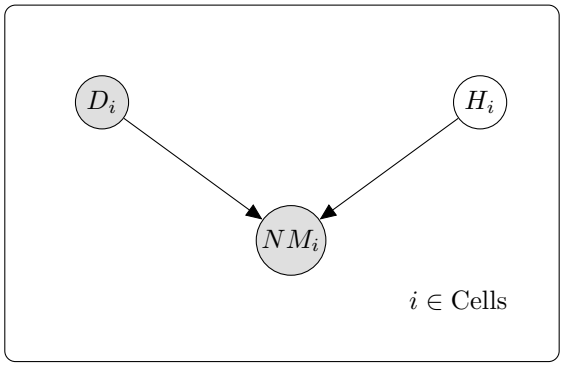


Figure 4.5: Bayesian Network representing the data generating process of the number of reads at a position (NM_i) carrying a mutation, given the true heteroplasmy of a cell (H_i), and the total number of reads (depth) at that site (D_i).

In this case, this is equivalent to using a beta-binomial model for each separate cell in the data, yielding as many posterior distributions as there are cells in the data. The beta-prior representation is convenient, as this is *conjugate prior*. We display the posterior distribution in this case for a variety of cells in figure 4.6, illustrating how posterior uncertainty varies with depth.

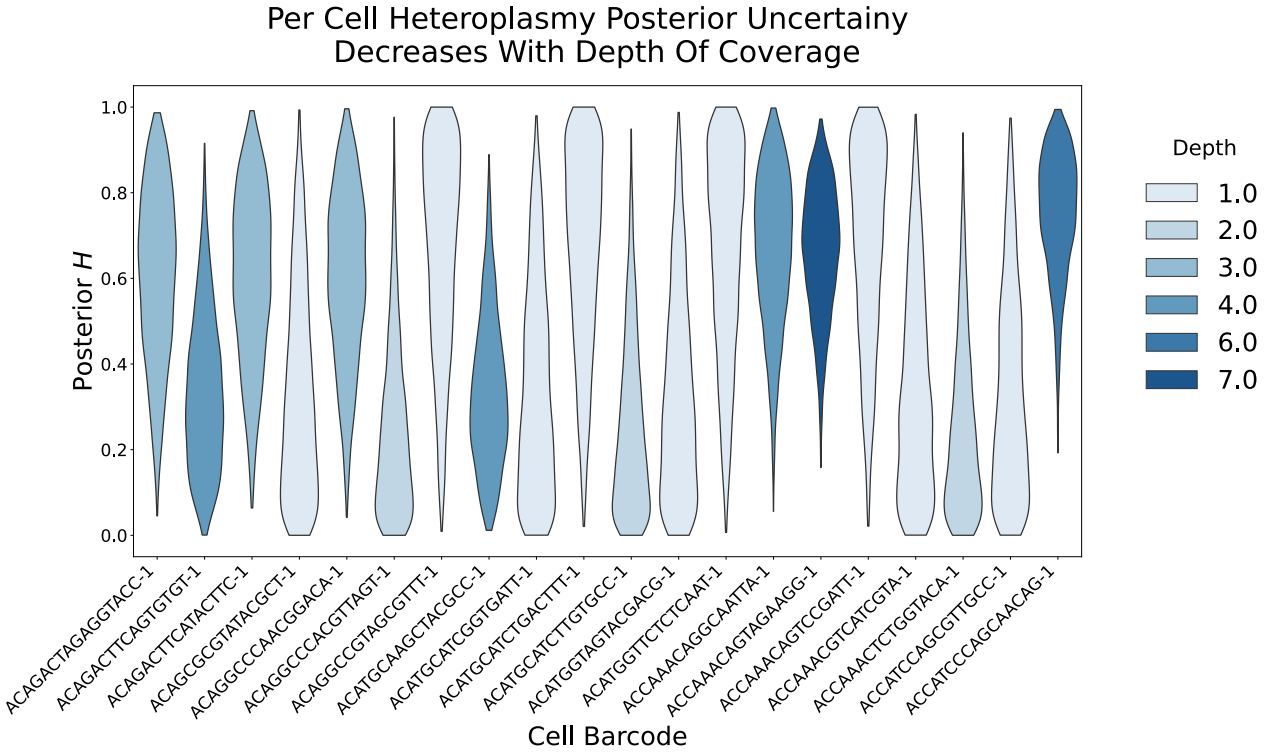


Figure 4.6: The posterior uncertainty of a particular cell's heteroplasmy can be seen to decrease with the number of reads observed at the site of the mutation in the cell. The posteriors effectively capture the naive estimates of the heteroplasmy as expected. We took some representative cell's reads, from data produced in the next chapter. At a single position on the mtDNA, different cells will have varying amounts of depth of coverage. It is important to be able to capture this variation in the depth of coverage between different cells and measure the uncertainty accordingly. The most straightforward model is to use the beta-binomial model, with beta parameters $\alpha, \beta = 1$, equivalent to the uniform prior. This prior can be interpreted at the modeler stating they are completely uncertain about a cell's heteroplasmy before any data has been produced. We display the posteriors for different cells, as represented by violin plots, all of which are beta distributions. It can be clearly seen how this simple model concentrates to a smaller region of the parameter space as the depth of each cell increases.

4.5 Modelling Gene Activity

Before we begin to assess the relationship between heteroplasmy and gene activity, it is worth considering what form the marginal distribution of gene activity might take. *Gene activity* is distinct from *gene expression*. At the single-cell level, gene expression is typically quantified by the number of RNA transcripts from a gene in that cell. On the other hand, gene activity is concerned with how accessible the DNA is along the length of a gene. A more detailed description of gene activity is given in chapter 5. Here, we give the reader a simplified picture, enough to understand the model and justify its suitability.

DNA is typically coiled up as chromatin. This structure is held in place by histones. In section 2.4.2, we described how some regions of the genome are more accessible, as the chromatin becomes unbound, how histones are displaced from certain regions, and how this accessibility is associated with cis-regulatory elements. Each gene has many potentially accessible sites along the length of its body. In a given cell, however, the chromatin will only be open at some of these locations. We compute the *gene activity*, the total number of accessible regions for a particular gene, by summing the number of accessible regions along the gene's length, in a given cell [163]. We are principally interested in learning how the heteroplasmy of mtDNA associates with this number. Before going on to measure that association, we develop a simple model for the marginal distribution of gene activity. From there, we develop an approach to link heteroplasmy to shifts in that distribution.

If we took an individual gene, j in cell i, in a given context, it seems reasonable to suppose the activity of that gene, $y_{i,j}$ would be constrained to have a mean $\lambda_{i,j}$. If we consider a specific gene, it has multiple candidate regions along its length that might be accessible. Let's index these sites with index k. Then each of these k sites would have a corresponding probability p_k of being occupied. It is reasonable to suppose the $p_k \to 0$ and that $n_j \to \infty$, where n_j is the total number of potentially accessible regions in gene j. The total number of these accessible sites is the gene activity for gene j in cell i, $Y_{i,j} = \sum_k^{n_j} X_{i,j,k}$, $X_{i,j,k}$ is the indicator variable stating whether region k is accessible in gene j, in cell i. It can be shown that under these assumptions $Y_{i,j}$ would be a Poisson random variable characterised by the mean parameter $\lambda_{i,j}$

[164]. In turn, we might also suppose that the rate gene j is active would vary from cell to cell, due to the different contexts and demands placed upon each cell. At a microscopic level, one might suppose that the values of p_k would vary from cell to cell. If we consider drawing cells at random from the population as a whole, we can ask what the distribution of activity, y_j (where we have dropped index i) is for gene j. If all cells expressed gene j at the same rate, y_j would also be a Poisson random variable, with an identical rate parameter governing each cell.

$$p(y_i|\lambda_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}$$

However, as it is reasonable to imagine that the mean rate of activity for any given gene might vary from cell to cell, this would lead to a mixture distribution. An established manner of introducing heterogeneity into the rate of the Poisson is to express the rate as $\lambda' = \lambda u_i$, with u_i being derived from some family of distribution. This leads to a mixture distribution of the form:

$$p(y_i|\lambda_i) = \int \frac{e^{-\lambda_i u_i}(\lambda_i u_i)^{y}{}_i}{y_i!} p(u_i) du_i$$

The question now is to ask what form $p(u_i)$ is to take. A simple choice is to suppose that $u_i \sim \Gamma(\phi, \nu)$ which is a flexible choice of distribution for model parameters which will largely be informed by data. This has the added benefit of being a conjugate distribution to the Poisson distribution, a mathematical convenience when it comes to evaluating the integral. Furthermore, the resultant distribution is quite flexible. Integrating over latent variables in the following integral leads to the Gamma-Poisson distribution, otherwise known as the negative binomial distribution.

$$p(y_i|\lambda_i) = \int \frac{e^{-\lambda_i u_i}(\lambda_i u_i)_i^y}{y_i!} \frac{\nu^\phi}{\Gamma(\phi)} u_i^{\phi-1} e^{-\nu u_i} du_i$$

Models of this variety exhibit *over-dispersion* [165]. Under a Poisson model, $X \sim Poisson(\lambda)$, the following standard properties hold

$$\mathbf{E}[X] = \lambda$$
$$\mathbf{V}[X] = \lambda$$

A distribution is said to be over-dispersed when $\mathbf{V}[X] > \mathbf{E}[X]$.

By making use of the tower properties of expectation and variance, we can see that the first two moments of the Gamma-Poisson model as:

$$\begin{aligned}
\mathbf{E}[y_i] &= \mathbf{E}[\mathbf{E}[y_i|u_i]] = \mathbf{E}[\lambda_i u_i] = \lambda_i \mathbf{E}[u_i] \\
&= \lambda_i \frac{\phi}{\nu} = \mu_i \qquad (4.19) \\
\mathbf{V}[y_i] &= \mathbf{E}[Var[y_i|u_i]] + \mathbf{V}[\mathbf{E}[y_i|u_i]] \\
&= \mathbf{E}[\lambda_i u_i] + \mathbf{V}[\lambda_i u_i] \\
&= \lambda_i \mathbf{E}[u_i] + \lambda_i^2 \mathbf{V}[u_i] \\
&= \lambda_i \frac{\phi}{\nu} + \lambda_i^2 \frac{\phi}{\nu^2} \\
\mathbf{V}[y_i] &= \mu_i + \frac{\mu_i^2}{\phi} \qquad (4.19)
\end{aligned}$$

Using this new parameterisation of the first two moments, we can write:

$$
\begin{aligned}
p(y_i) &= \int p(y_i|u_i)p(u_i)du_i \\
p(y_i) &= \int \frac{e^{-\lambda_i u_i}(\lambda_i u_i)^{y_i}}{y!}\frac{\nu^{\phi}u_i^{\phi-1}e^{-\nu u_i}}{\Gamma(\phi)}du_i \\
&= \frac{\lambda_i^{y_i}}{y_i!}\frac{\nu^{\phi}}{\Gamma(\phi)}\int e^{-(\lambda_i+\nu)u_i}u_i^{\phi+y_i-1}du_i \\
&= \binom{y_i+\phi-1}{y_i}\left(\frac{\nu}{\nu+\lambda}\right)^{\phi}\left(\frac{\lambda}{\nu+\lambda}\right)^{y_i'}
\end{aligned} \tag{4.19}
$$

From equation 4.19, we can substitute $\lambda_i = \frac{\nu}{\phi}\mu_i$ into equation 4.19 to arrive at a standard representation of the negative binomial distribution in terms of its mean parameter:

$$
p(y|\phi,\mu_i) = \binom{y_i+\phi-1}{y_i}\left(\frac{\phi}{\phi+\mu_i}\right)^{\phi}\left(\frac{\mu}{\phi+\mu}\right)^{y_i} \tag{4.19}
$$

Over-dispersion manifests itself as greater variation in count data than would be expected under a Poisson model assumption, ie, over dispersion implies $\mathbf{V}[X] > \mathbf{E}[X]$. The generative description of gene activity laid out accounts for this over-dispersion as being due to the range of different activity rates that each cell has for a particular gene, and yields a testable model in the form of the negative binomial distribution. We fit the negative binomial model, without any predictor variable, to real gene activity data with 2,943 cells, using MLE. The results are displayed in Figure. 4.7. This model offers a remarkably good fit of marginal gene activity data. The activity distribution for each gene varies considerably, and it can be seen that the negative binomial model is able to capture the counts of gene activity for each gene well.

We have given a plausible account of negative binomial distribution being appropriate to gene activity and seen that it captures our data well. Now we turn to linking external predictor variables to this distribution.

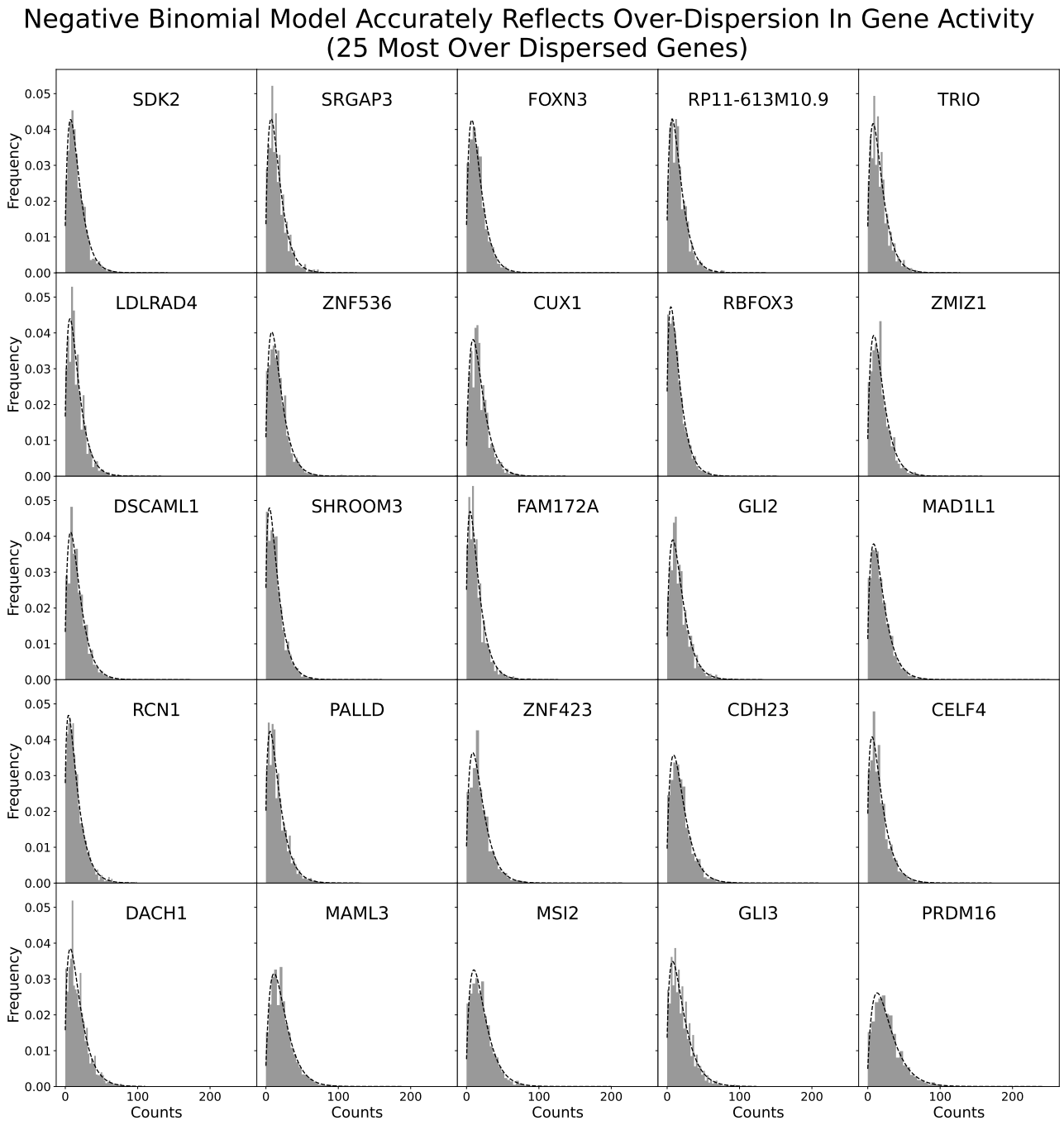


Figure 4.7: MLE fits of the NB2 model, offer compelling fits to the observed gene activity data. This lends credibility to the mechanistic explanation of gene activity laid out in the text. This was fit to a single-cell ATAC data set, discussed more extensively in the next chapter.

4.5.1 Negative Binomial Regression

Linking changes in the parameters of a distribution is often achieved with a generalised linear model (GLM). In a standard Poisson GLM, the log link is used as $ln(\lambda_i) = \alpha + \beta x_i$ to connect the mean rate to a predictor variable x, with parameters α, β to be determined as part of the

regression. In the Gamma-Poisson, an additional Gamma noise term is added to the model to allow for over-dispersion, variation in the Poisson rate arising under identical values of the predictor variable. In the derivation in the previous section, this variation in the Poisson rate is represented by u_i. Conditioned on the value of u_i, the rate term in the model becomes $ln(\lambda_i u_i) = \alpha + \beta x_i + ln(u_i)$ with $ln(\lambda_i u_i) = \alpha + \beta x_i + ln(u_i)$ in a manner similar to the Poisson regression link function.

If we use our reparameterization from 4.19, we can model the unconditional rate directly as $ln(\mu_i) = ln(\lambda_i \frac{\phi}{\nu}) = \alpha + \beta x_i + ln(\frac{\phi}{\nu}) = (\alpha + ln(\frac{\phi}{\nu})) + \beta x_i = \alpha' + \beta x_i$ We can simply absorb the additional parameters into the intercept term, re-written as α' for clarification. As such, the link function originally inserted into the mean rate of the original Poisson likelihood, is now used to model the mean of the negative binomial distribution. An additional ϕ parameter is used to control the degree of observed over-dispersion within the model, as we have seen in 4.19. This formulation of the negative binomial model is sometimes referred to as the NB2, being based on the Poisson-Gamma mixture derivation [3] [165].

Above, we have given a generative explanation of why gene activity would follow a negative binomial distribution as a Gamma-Poisson mixture of gene activities. We would now like to explain why those rates of activity vary from cell to cell. In particular, we would like to understand what part of this variation can be explained as a function of heteroplasmy. Mitochondrial genes encode some of the key metabolic genes within the cell and are essential to the regular functioning of the cell, so it is to be expected that cellular heteroplasmy would be associated with activity in at least some genes. In order to proceed with this line of inquiry, we use the log link function on the mean of the negative binomial distribution as described above, with heteroplasmy as the predictor variable.

When conducting the regression for a particular mutation, index i denotes a cell, and index j denotes the gene for which the regression between activity and heteroplasmy was performed.

[3] This is different from the NB canonical regression model based on GLM derivations. The standard derivations of GLMs model the means of the sufficient statistics of distributions belonging to the exponential family [165]. The derivation we offer is more in line with the generative process as we have outlined it when describing the sequencing procedure

Now, simply preferring to call the intercept α and denoting equation 4.19 by $NegBinom(\mu_{i,j}, \phi_j)$, we can write this regression as:

$$ln(\mu_{i,j}) = \alpha_j + H_i\beta_j$$
$$y_{i,j}|\mu_{i,j}, \phi_j \sim NegBinom(\mu_{i,j}, \phi_j)$$

The following DAG in figure 4.8 helps us to understand the relationship between variables in this simple model. However, as we stated in section 4.4, we are additionally uncertain about the true values of the heteroplasmies used as the regressor in the model. We now turn our attention to combining heteroplasmy uncertainty with the Negative binomial regression model.

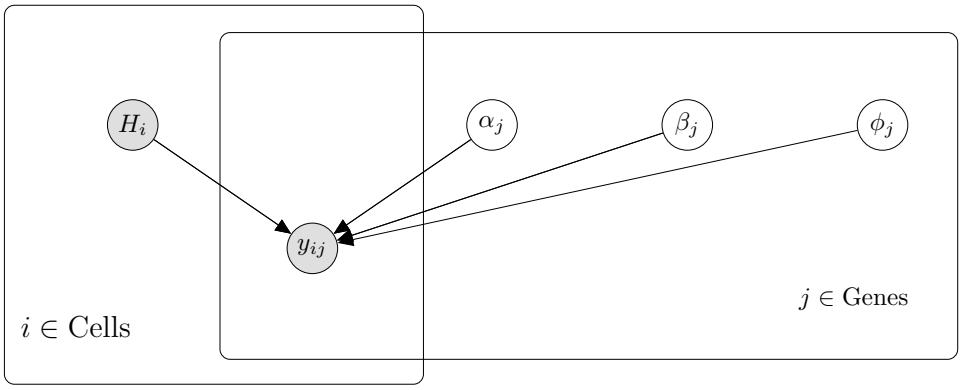


Figure 4.8: DAG showing dependencies in the negative binomial regression model.

4.5.2 Negative binomial regression with heteroplasmy uncertainty

As stated earlier, measurement error is a major source of bias and can dramatically alter the estimated parameters of a statistical model if not accounted for appropriately. In the sections leading up to this one, we have shown how heteroplasmy is subject to considerable measurement error due to the limited depth observed within many of our samples. In addition to this, we gave a generative account of why we might assign a negative binomial distribution to gene activity, a line of reasoning reinforced by the quality of its fit to the marginal distribution of gene activity. This follows a tradition of work in the RNA sequencing literature that has often made use of the negative binomial distribution, in one form or another, to model transcript

counts [166, 167, 168]. We followed this by showing how one could in turn link predictor variables to gene activity through the use of the negative binomial regression, specifically, using the NB2 parameterisation with the following link function $log(\mu_i) \sim \alpha + \beta H_i$, where H_i is the heteroplasmy of a cell as a predictor of gene activity. Now we combine the measurement error in the heteroplasmy estimates with our negative binomial regression model. By combining the models seen in Figs. 4.5 and 4.8, we arrive at the DAG displayed in 4.9. In terms of equations:

$$
\begin{aligned}
H_i &\sim Beta(1,1) \\
NM_i|H_i, D_i &\sim Binom(H_i, D_i) \\
\alpha_j &\sim Normal(0,1) \\
\beta_j &\sim Normal(0,1) \\
ln(\mu_{i,j}) &= \alpha_j + H_i\beta_j \\
\phi_j &\sim Exp(2) \\
y_{i,j}|\mu_{i,j}, \phi_j &\sim NegBinom(\mu_{i,j}, \phi_j)
\end{aligned}
$$

This is a fully Bayesian model with prior distributions over the model parameters, over which, we can obtain posterior distributions once we condition on the data. We chose largely uninformative priors. As will be seen later, in the limit of large numbers of cells, as is typical in single-cell data, the role the prior plays in informing the posterior distribution is negligible.

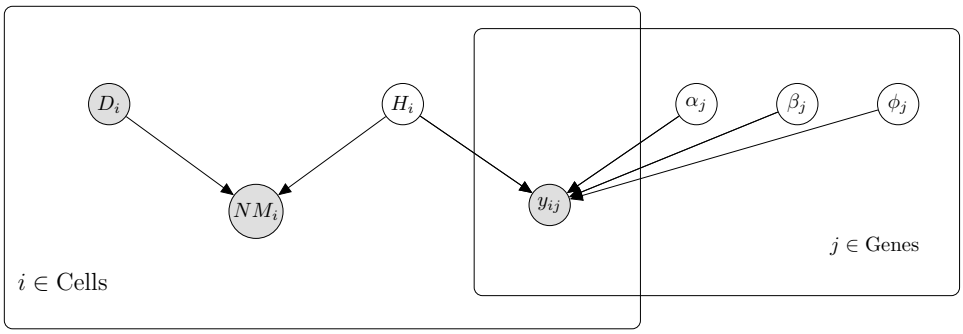


Figure 4.9: Negative binomial Regression model (NB2) including heteroplasmy predictor uncertainty. In this model, only the depth, number of mutated reads, and the activity of a corresponding gene are observed. The true heteroplasmy is unknown and is modeled as a latent variable within the model. The heteroplasmy, in conjunction with the negative binomial regression parameters, predicts the gene activity.

4.6 Simulation Studies

Thus far, we have taken a major step towards our aim in quantifying the association between heteroplasmy and gene activity in the low-depth regime. The strength of this relationship is embodied in the regression coefficient β, and we seek to apply this model to real-world data. In the meantime, however, we have to establish the model's properties, its strengths, its weaknesses, and understand any potential failings. It is not simply enough to write down the equations of a model and let it loose upon data. The main source of concern arises from using a model with 1,000s of latent variables and many skeptical questions may be asked of the model. Can the model be fit in an acceptable amount of time? Can the model, in fact, fit at all without problems? Does the model add benefit over a simple MLE using the naive parameter values? What influence do the priors play in influencing our conclusions? To what extent does low depth matter, and can this be ameliorated by having a large number of cells? Can the model successfully regain the parameter values in a low depth regime?

We perform a simulation study to answer these questions, fully demonstrating the utility of the model. We show that the model is capable of fitting the data and recovering simulated parameter values under a range of settings. In addition to this, we show how failing to factor in predictor uncertainty, as would be the case when a naive point estimate of heteroplasmy is used, fails to recover the true strength of associations in the data.

4.6.1 Low Depth Destroys Parameter Estimates In Conventional Models

To begin with, we simulate data showing how low depth destroys the ability of models using naive estimates of heteroplasmy to recover true parameter values from the data. We simulated data for a single mutation in 1,000 cells. To each cell we randomly draw a heteroplasmy value from the uniform distribution. We sampled gene activity values for each cell according to $y_i \sim NegBinom(\mu_i, \phi)$ such that $\mu_i = \alpha + H_i\beta$, with $\alpha = 1$, $\beta = 1.5$, and $\phi = 2$, relatively strong effect sizes that a model ought to be able to detect under reasonable conditions. We then

generate two sets of data, equivalent to two different data sets with different amounts of depth. For one data set we simulate depth, D_i from a geometric distribution with mean $1,000$ and the other we draw depth samples from a geometric distribution with mean 5. We then sample the number of mutated reads as $NM_i \sim Binom(H_i, D_i)$. Standard approaches to analysing the impact of heteroplasmy typically measure the heteroplasmy as $\frac{NM_i}{Depth_i}$, and use these point estimates as predictors in various regression models. When depth is low, we expect such naive estimates to be very noisy and to lose inferential value.

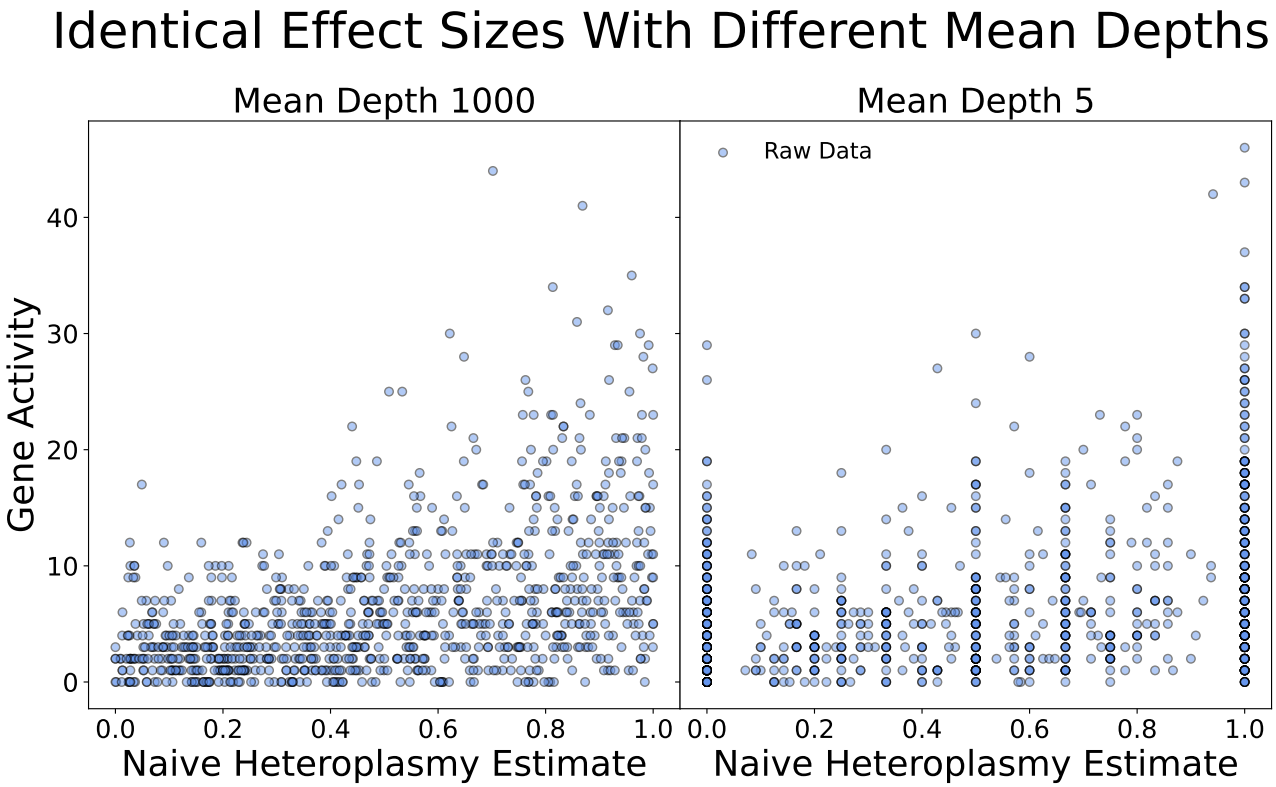


Figure 4.10: Naive heteroplasmy point estimates per cell and the corresponding gene activity. Naive heteroplasmy point estimates are calculated as the fraction of reads harbouring a mutation to the read-depth at the site of the mutation $\frac{NM_i}{D_i}$. We generated two different data sets with equal effect sizes, only varying the depth of coverage in the mitochondrial genome that we attained per cell.

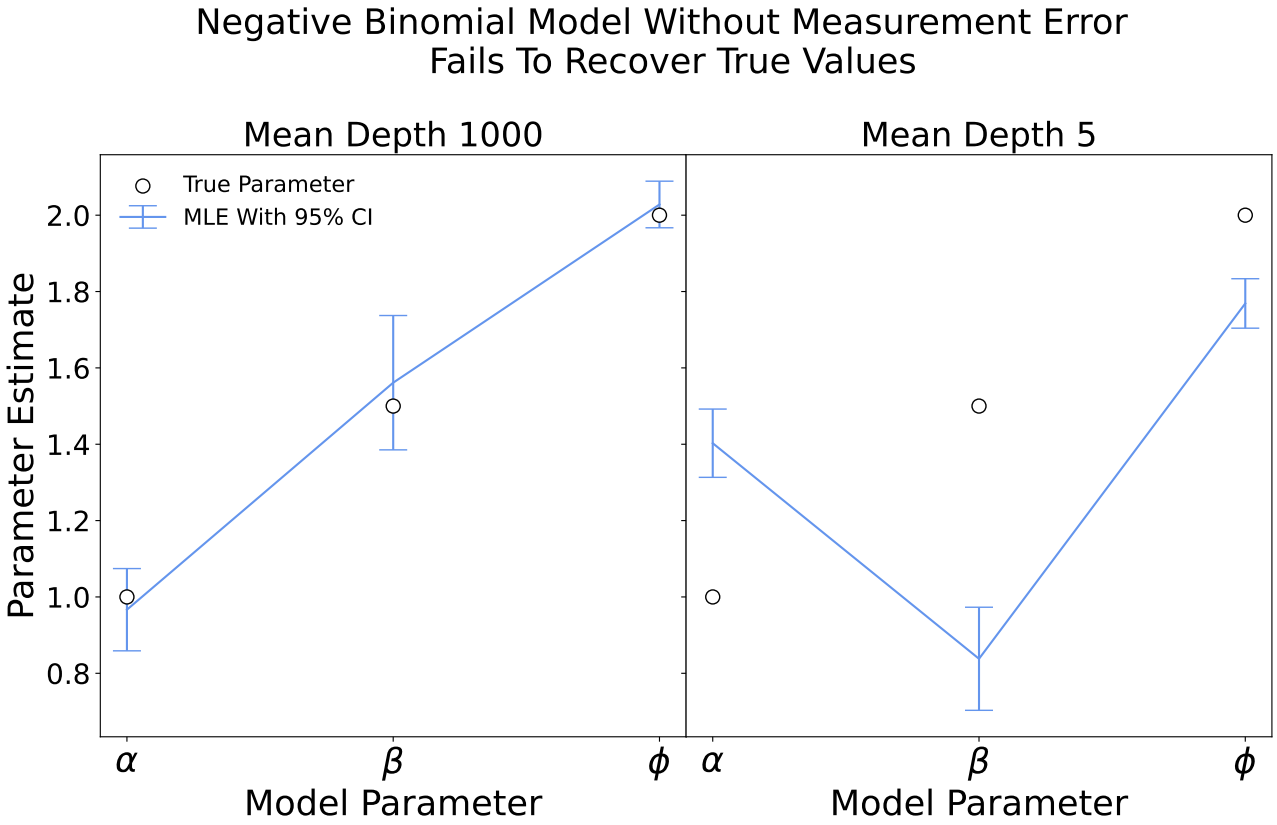
Figure 4.10 shows the data simulated under these two conditions. We plot the naive estimates of heteroplasmy against the value of gene activity, which is associated with the true heteroplasmy through the β parameter. We can see that in the low-depth regime, discretisation artifacts show up in the heteroplasmy estimates as vertical lines in the scatter plot. For example, if there were only 5 reads in a cell, the naive heteroplasmy estimate could only be at 0, 0.2, 0.4, 0.6, 0.8, 1. As such, the observed heteroplasmy estimate is largely determined by the lack

of sequencing depth and is only partially reflective of the true heteroplasmy of a cell. High depth data however is expected to be far less noisy and to yield reasonably good estimates of parameter values. We run negative binomial regression on these two data sets by using the naive heteroplasmy estimates as predictors. We use MLE for simplicity. Figure 4.4a shows the results of this procedure. It can be seen that when the mean depth is high, use of the naive heteroplasmy estimates is capable of recovering heteroplasmy estimates. However, when depth is depleted, true parameter values lie far outside of the confidence interval. Intuitively this is what is expected to occur, as when depth diminishes the mutual information between the true heteroplasmy and the point estimate, $I(\frac{NM_i}{Depth_i} : H)$, diminishes.

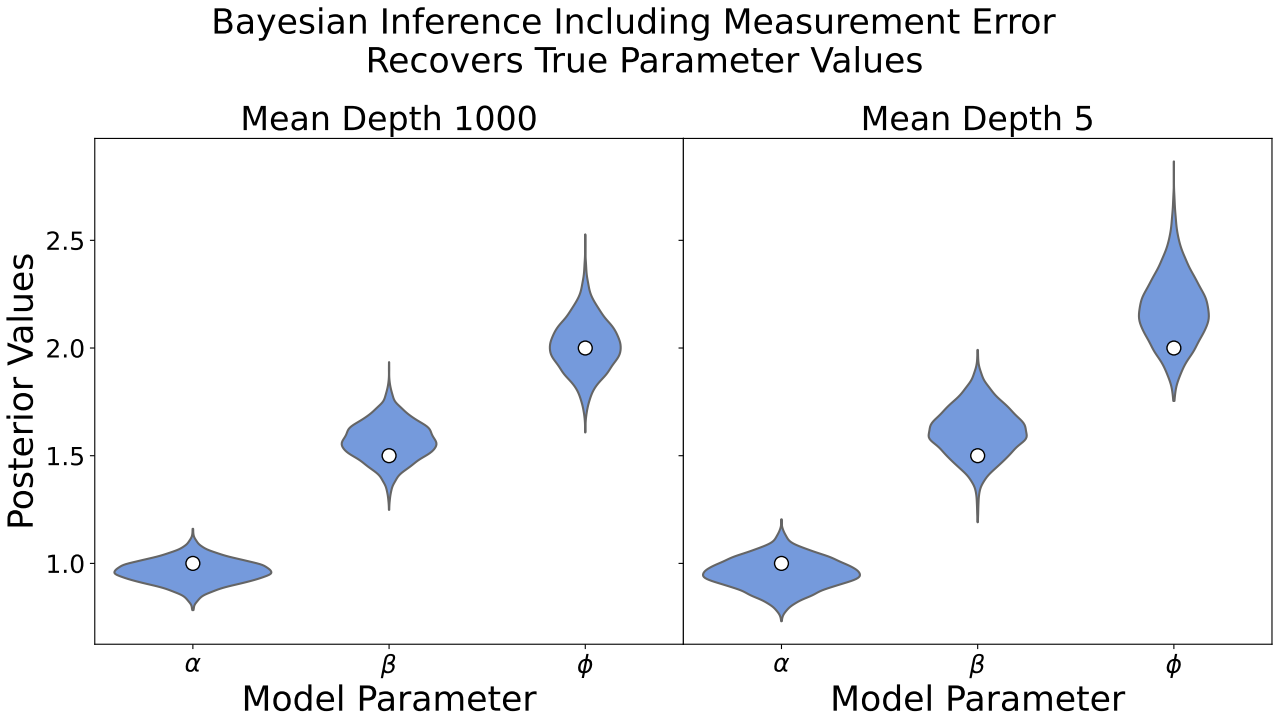
To demonstrate the utility of our approach, we now fit our fully Bayesian, measurement error model to the same data. The models were fit using the no-U-turn sampler (NUTS) flavour of HMC using `Stan` [23]. The code used to implement the negative binomial heteroplasmy uncertainty regression model is shown in appendix A. The recovered posteriors are displayed in 4.11 (b). We represent these as violin plots, emphasising how we get a distribution over parameter values, unlike fig 4.11 (a), in which MLE only yields confidence intervals. As can be seen, the posteriors cover the true values in both cases and are not perturbed by the lack of depth in the data. This is because the Bayesian measurement error model is capable of making use of all information available, including how many reads were available for each heteroplasmy estimate. Intuitively, the model will give more weight to high-depth data points rather than low-depth data points.

It is worth noting that the Bayesian posteriors are wider in the low-depth data. This makes intuitive sense, in line with low depth of coverage destroying how much information the model has about each cell's heteroplasmy. Nevertheless, the model can still appropriately locate this parameter value, whilst accurately reflecting its greater uncertainty.

No divergent transitions were present for the HMC simulations in either the high or low-depth regimes, which might otherwise be expected to arise if the model had serious degeneracies, unidentifiable parameters, or was poorly parameterised. We display the trace plots of MCMC chains of the parameters of interest in figure 4.12. All chains appear to mix and display

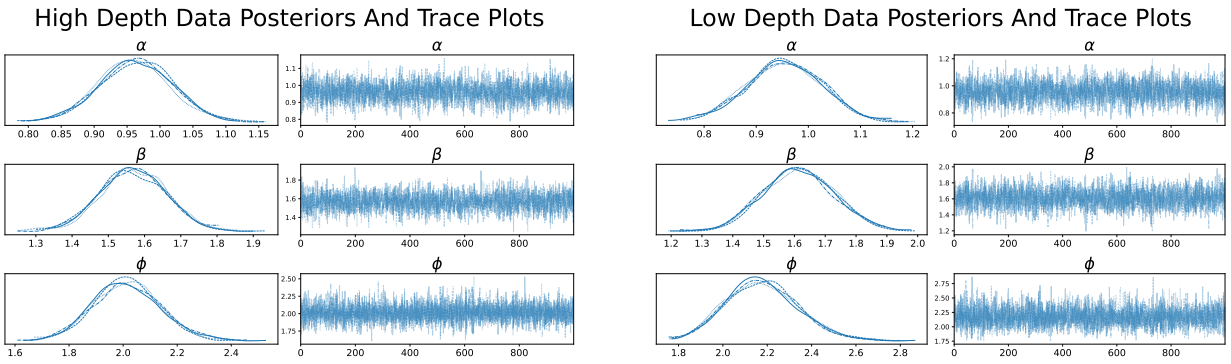


(a)



(b)

Figure 4.11: (a) Maximum likelihood estimates using naive point estimates of heteroplasmy. It can be seen that when the mean depth per cell is low, the point estimate does not precisely reflect the true heteroplasmy, destroying the strength of the effect size observed when using maximum likelihood on the negative binomial model. (b) The fully Bayesian negative binomial model which does factor in heteroplasmy uncertainty does not suffer from the same issues, recovering the true effect sizes.



(a) (b)

Figure 4.12: Traceplots for the α, β, and ϕ parameters of the negative binomial regression model with uncertain predictors. It can be seen that all chains converge and mix well, indicative of good sampling. Lack of divergences indicates no obvious problems occurred with the model.

stationarity, with $\hat{r}$ values were very near 1 for all parameters in the model.

4.6.2 Simulation Based Calibration

In the previous section, we outlined how our measurement error model is capable of recovering parameter values in low-depth settings, whilst models that do not account for heteroplasmy uncertainty fail. Parameter recovery, however, was only done for a single set of parameter values. Though compelling and reassuring to see that our model is capable of performing as intended in this situation, this is not entirely sufficient to demonstrate the efficacy of our model. Firstly, parameter recovery could have worked by chance — perhaps the sample data simply "worked out" in this situation. Secondly, because in general, Bayesian inference is only *calibrated* when averaged over the prior distribution, not when performed for any particular single set of parameter values. As such we need to simulate data from a wider set of parameters. A component of Bayesian workflow is to simulate data from the model's joint distribution multiple times and fit the model to the data generated in this manner. This is referred to as *Simulation Based Callibration* (SBC) [169, 24, 170]. We can create sample data using a *generator* through simple ancestral sampling: data samples are obtained from a generator by sampling from the prior distribution $p_\theta(\tilde{\theta})$, to yield a single value $\tilde{\theta}$; we then sample from the likelihood $p(y|\tilde{\theta})$ to obtain a data set, y; once the generator has produced data, Bayesian inference is performed on the data set y, using a *probabilistic program* to yield posterior samples

from $p(\theta|y)$.

The posterior quantiles of the true parameter originally sampled are defined as:

$$q(\tilde{\theta}) = \int d\theta \; p(\theta|y)\mathbf{I}[\theta < \tilde{\theta}] \tag{4.19}$$

The main theorem of SBC states that if the software being used is actually generating samples from the posterior distribution, then the quantiles in 4.19 will be $q(\tilde{\theta}) \sim Uniform(0,1)$ [169]. The way 4.19 is written is suggestive of how to compute this quantile numerically. Simply generate data using the generator, draw samples from the posterior resulting from that data, and compute the proportion of posterior draws larger than true parameter $\tilde{\theta}$.

To see why this result holds, first, we have to note a property of the joint distribution of the SBC draws. We use indicative subscripts for clarity:

$$\begin{aligned} p_{SBC}(\tilde{\theta}, y, \theta) &= p_{prior}(\tilde{\theta})p_{data}(y|\tilde{\theta})p_{post}(\theta|y) \\ &= p_{marg}(y)p_{post}(\tilde{\theta}|y)p_{post}(\theta|y) \\ \implies p_{SBC}(\tilde{\theta}, \theta|y) &= p_{post}(\tilde{\theta}|y)p_{post}(\theta|y) \end{aligned}$$

The crux of this result is that conditional on the data, the true parameter value, $\tilde{\theta}$, and a simulated posterior value, θ, *should* follow identical distributions [169, 171]. We sketch how this can be used to imply the uniform distribution of posterior quantiles. First, note we can write the quantiles as $q(\tilde{\theta}) = F_{\theta|y}(\tilde{\theta})$, and assume that this is an invertible function.

$$
\begin{aligned}
Pr(q(\tilde{\theta}) \leq x|y) &= Pr(F_{\theta|y}(\tilde{\theta}) \leq x) \\
&= Pr(\tilde{\theta} \leq F^{-1}_{\theta|y}(x)) \\
&= F_{\tilde{\theta}|y}(F^{-1}_{\theta|y}(x)) \\
&= F_{\theta|y}(F^{-1}_{\theta|y}(x)) \\
&= x
\end{aligned}
$$

where the fourth line relies upon the identical conditional distributions shown previously. If the distribution of quantiles computed in this manner is not uniform, this would imply that the sampler was not faithfully capturing the posterior distribution.

Figure 4.13 demonstrates this visually for our negative binomial regression model. The "true" parameter value β, was sampled from the prior, its value represented by the vertical line. Data were then generated, and the Bayesian model fit to the data, with samples from the fitted posterior distribution shown in blue. The posterior CDF is then shown by the orange line. Where the vertical line intersects the posterior cdf gives the posterior quantile of the "true" β parameter. If the computational method is well implemented, and the data provide enough information to identify the model's parameters, then these quantiles should follow a uniform distribution. Deviations from the uniformity in this procedure would indicate some failure of implementation in our model.

We implemented simulation-based calibration for the negative binomial model regression model including predictor uncertainty. The procedure was run for 200 iterations, yielding 200 posterior quantiles of the true parameter values. In figure 4.14 we display the QQ plots of the posterior quantiles vs the theoretical quantiles of the uniform distribution. The QQ plots show that the posterior quantiles are well described by a uniform distribution with little meaningful deviation, indicating our model is well calibrated and that our computational implementation is functioning properly.

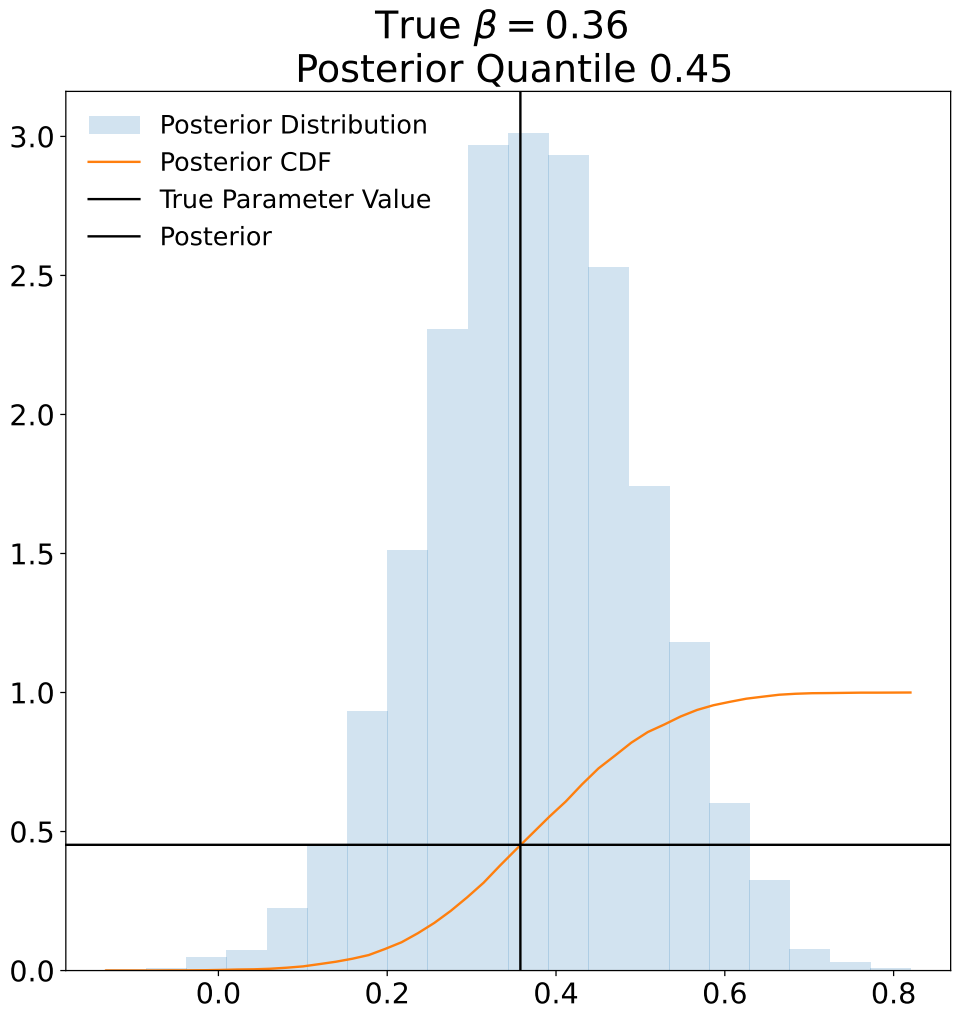


Figure 4.13: In simulation-based calibration, a parameter value is sampled from the prior distribution. This value is shown as the vertical line in the figure. Data are in turn sampled from the likelihood using this drawn parameter value. Bayesian inference is then performed on this dataset to obtain a posterior distribution over the prior. We obtained an approximation to this posterior by using HMC in stan to obtain many samples and visualised them as a light blue histogram. We can then compute the quantile of the posterior distribution at which the true parameter value is found. This is visualised here by showing the value CDF of the posterior distribution (orange), at the true parameter value, showing the quantile as the horizontal line in the plot. If this whole procedure is repeated many times, the quantile values should follow a uniform distribution for a well-calibrated simulation procedure [24, 169].

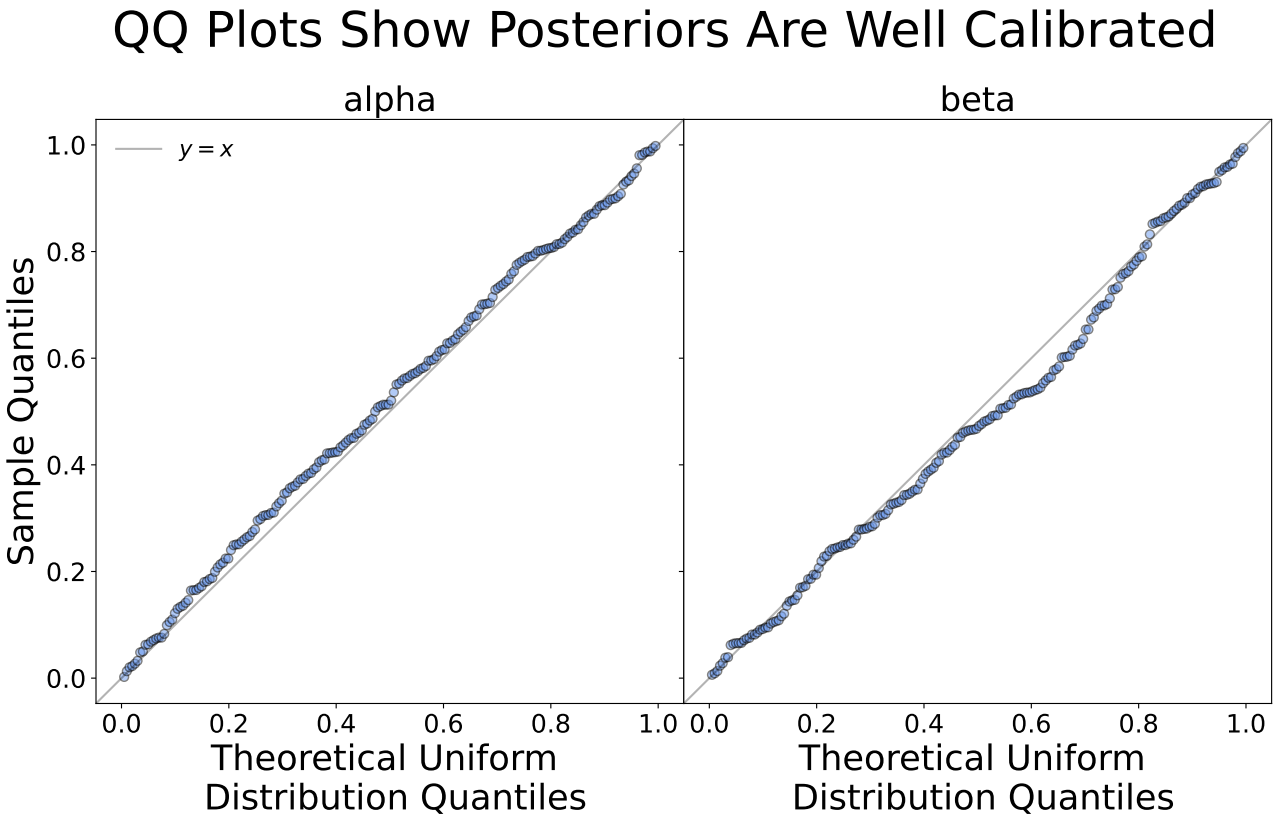


Figure 4.14: QQ plots showing the posterior quantiles and the quantiles of the theoretical uniform distribution cohere. This shows that our computational inference procedure is able to accurately recover the true parameter values and that the procedure is well calibrated.

4.6.3 Calibration Under Misspecified Priors

Simulation based calibration, as implemented above, is primarily intended as a tool for testing that the model at least fulfills the self-consistency requirement, is capable of recovering true parameters across a range of different parameter values, and that the computational procedure used to implement the inference performs as intended. This is all part of good practice in the Bayesian workflow [24, 169]. We now develop and extend its usage, moving beyond this simple form of SBC to stretch, strain, and test how well our model works under a variety of different settings.

One of the key challenges to Bayesian models is the setting of the prior. Of particular interest is how the model would behave if the parameters of the prior were somehow misspecified. For instance, in our model, we say that $\beta \sim Normal(0, \sigma_\beta)$, with $\sigma_\beta = 1$. What if another value were more appropriate? We can test the performance of our model when the data were produced using a generator for which $p(\tilde{\theta})$ differed from the prior distribution of the model being fit to the data.

Figure 4.15 shows the consequences of this misspecification for SBC. The first thing to note is that when the prior is correct, $\sigma_\beta = 1$, the posterior quantiles are uniform for the model across all cell counts at all levels of depth. This even applies to datasets with mean depth as low as 2 and only 10 cells, an extreme condition. This shows that posterior inferences are well-calibrated and that no pathologies in the model occur, even at this extreme. In such low-depth conditions, however, it can be seen that when the prior is misspecified in either direction relative to the generator, the model is poorly calibrated and severely distorted. As cell counts run higher, these distortions are attenuated, and once 1,000s of cells are present, the posteriors are calibrated once again. The restoration of calibration at high cell counts reflects that for large data samples, priors become less relevant, and posteriors become well calibrated, even for fixed parameter values with a large magnitude [153]. Far large amounts of data it would be expected that under repeated sampling, $\alpha\%$ intervals of the posterior distribution would indeed cover the true value $\alpha\%$ of the time, regardless of the parameter values we chose for our prior.

These results serve to highlight that in the realm of large amounts of data, the model is well-behaved, coverage properties are restored, and our choice of prior does not influence our inferences in a meaningful way. Fortunately, this is the situation we typically find in single-cell experiments, with 1,000s of cells being the norm. It also sheds light on the regime that we would expect the model to fail without deep consideration of priors.

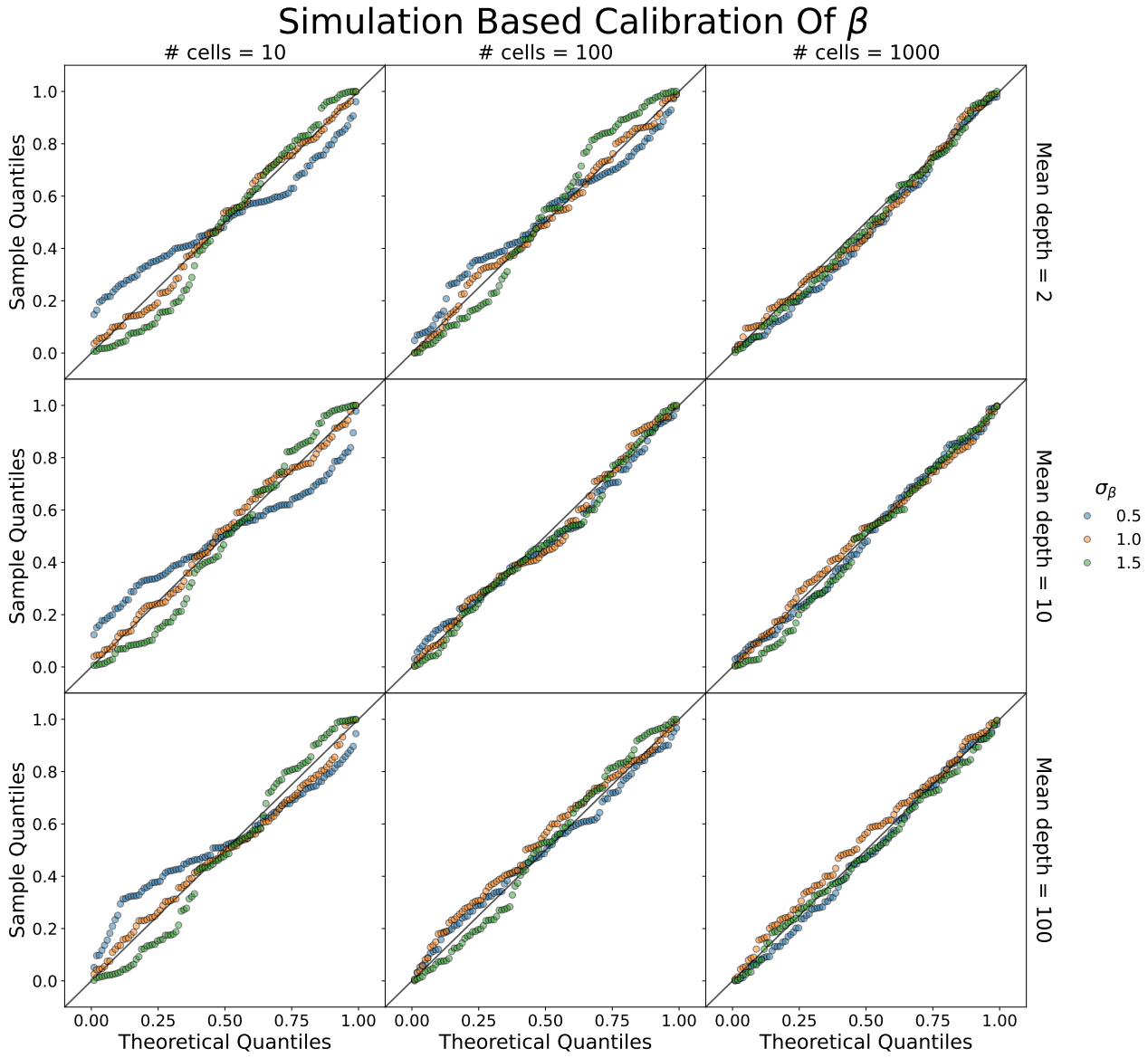


Figure 4.15: QQ-plots showing posterior calibration when the SBC data generator has a misspecified prior distribution. The data generator matches the model for $\sigma_\beta = 1$ and is calibrated under this circumstance, being uniform under extreme datasets with mean depth 2 and only 10 cells present. When the priors are misspecified with different values of σ_β, we observe distortions to the posterior calibration. However, even under low-depth regimes, large cell counts can attenuate the damage restoring calibration to the posteriors. This is reflective of large amounts of data restoring coverage properties of the posterior distribution [153]. This range of results demonstrates that for the cell counts encountered in typical datasets, prior considerations are not of paramount concern.

4.7 Multiple Testing Concerns

We have now discussed the possibility of model misspecification. An extreme form of model misspecification would be that there is, in reality, no association between heteroplasmy and gene activity. A continuous prior on the regression coefficient does not allow for this precise possibility. In repeated applications of the model, it is plausible that many of the associations are effectively null. A fully Bayesian approach to solving this problem could involve a hierarchical model in which the prior itself is learned from the data [172]. We made efforts in this direction but found that attempting to fit a model with ∼20,000 parameters was computationally prohibitive when attempting to sample the posterior. Instead, we model each relationship separately, (and assume independence). This offers the important benefit of being pleasingly parallelisable onto multiple compute cores, which is outlined in further detail in chapter 5. Presumably, some proportion of the time when we apply our model, β is in reality zero, contradicting the structure of the prior. This repeated application of a model carries a resemblance to conventional frequentist multiple hypothesis concerns.

Loosely stated, standard frequentist null hypothesis significance testing (NHST) revolves around a "null-hypothesis", often denoted H_0, with a corresponding null distribution for some test statistic. This is to say, the distribution of test statistic values that would arise for outcomes produced by chance, or no-effect conditions, under repeated sampling. A p-value is then the probability of seeing a test statistic as, or more, extreme as that observed, under this null distribution. By definition, the distribution of p-values is uniform, under the null. Typically, a "significance level", α, is specified before the test is carried out. For all tests under which the p-value, $p \leq \alpha$, the null hypothesis H_0 is rejected [173].

In many settings, we have many hypotheses, indexed as H_{0i}, some of which are null. Assuming that all these null hypotheses are true, we would in fact expect to reject 5% of these null hypotheses if we were to blindly apply an NHST testing procedure. In general, presumably, some of these null hypotheses are true, and others are false. How can we identify which null hypotheses to reject among potentially many null hypotheses? Two popular approaches are often called upon in this situation: the Bonferroni correction and the Benjamini-Hochberg (BH)

procedure. We describe these procedures in sections 4.7.1 and 4.7.2.

Some Bayesian models can be repeatedly fit to distinct data which is repeatedly generated from a single *fixed* parameter value and have the posterior quantiles of the true value be uniformly distributed, including when that fixed parameter value is the null value [153]. This situation is distinct to SBC in sections 4.6.2 and 4.6.3, where a parameter value is generated from the prior (misspecified or otherwise) before sampling from the likelihood using that parameter value. Typically, one would expect uniformity in posterior quantiles when fit to fixed-parameter data to arise asymptotically, when the model is fit to large amounts of data [153]. If, when data is generated from the null, posterior quantiles of the null value are uniform under repeated sampling, one can be justified in applying standard frequentist multiple-testing procedures to posterior quantiles due to their resemblance to p-values.

We give an explanation of the Bayesian interpretation of the Benjamini-Hochberg procedure, which relies on p-values being uniformly distributed under the null. We show that posterior quantiles of the null, when behaving in the manner just described, can also have the Benjamini-Hochberg procedure applied to them to control the false discovery rate.

We show, numerically, that uniformity of posterior quantiles for null generated data is observed for the regression coefficient of the negative binomial heteroplasmy uncertainty model when cell count is sufficiently large. This is the case despite the growing number of parameters in the model, ie, another latent heteroplasmy for every cell. As such, we are justified in applying the BH procedure to the posterior quantiles of our models in the large datasets that we typically observe in single-cell studies. We demonstrate *in silico* that this application of the BH procedure to the negative binomial heteroplasmy uncertainty model does, in fact, control the FDR at the theoretical rate, even for low-depth data.

4.7.1 The Bonferroni Correction

A popular, but conservative approach is the *Bonferroni Correction*. For a particular test, the significance level, α, is the probability of incorrectly rejecting the null hypothesis, that is, rejecting the null, when the null hypothesis is actually true. For a collection of hypotheses, the

family-wise error rate (FWER), is the probability of making 1 or more incorrect rejections of the null hypothesis. The Bonferroni correction controls the FWER by only rejecting the null hypothesis for each test when $p_i \leq \frac{\alpha}{N}$, where N is the total number of hypotheses tested. To see how this controls the FWER, suppose that out of the collection of N hypotheses, a set I of size N_0 are actually true null hypotheses. Then by only rejecting tests at the level of $p_i \leq \frac{\alpha}{N}$, we can use the union bound to see that:

$$\begin{aligned} \text{FWER} &= P\left(\bigcup_{i \in I}\left(p_i \leq \frac{\alpha}{N}\right)\right) \leq \sum_{i \in I} P\left(p_i \leq \frac{\alpha}{N}\right) \\ &= N_0 \frac{\alpha}{N} \leq \alpha \end{aligned}$$

4.7.2 Benjamini-Hochberg Procedure

The Bonferroni correction is very conservative in that it aspires to prevent even one false test from making it through to the set of candidate results. A more recent innovation, the false discovery rate (FDR), by contrast, is more permissive [26, 174]. Given some decision rule, D, which we use to decide if a test is null or not, the test will reject N_R of the N total hypotheses. Of these rejections, N_{RT} will actually be true null hypotheses. The false discovery proportion is defined to be $\frac{N_{RT}}{N_R}$. The false discovery rate is the *expected* false discovery proportion, over different realisations of the experiment. The aspiration is to come up with decision rules that control the FDR to an acceptable level.

The Benjamini-Hochberg procedure is straightforward, assuming each of the tests are independent [26]. Compute the list of uncorrected p-values for each test separately. Sort the p-values into order, from lowest to highest, letting $p_{(i)}$ denote the ith ordered p-value. Specify a desired false discovery threshold, q_{FDR}. If $p_{(i)} \leq \frac{q_{FDR}}{N} i$, reject the hypothesis producing that p-value. Then, proceed onto the next ordered p-value, repeating these steps, until i_{max} is reached, the index of the final ranked p-value for which $p_{(i)} \leq \frac{q_{FDR}}{N} i$ is satisfied[4]

[4]In particular we focus on the pFDR, or positive false discovery rate, which is conditional on rejecting at least one hypothesis. See [175] for more discussion on this point.

Remarkably, this procedure controls the FDR to the level of $\pi_0 q_{FDR} \leq q_{FDR}$ (as $0 \leq \pi_0 \leq 1$) π_0 being the proportion of hypotheses among those tested that are actually known. Of course, we do not *know* π_0. But it can be estimated, which we discuss below.

4.7.3 Benjamini-Hochberg Procedure — A Bayesian Interpretation

Having laid out the simple mechanics of the Benjamini-Hochberg procedure, we give a sketch of an established Bayesian interpretation of the protocol [26, 175, 174]. In the next section, we leverage some of the insights produced by this chapter to identify scenarios where the procedure can be applied to Bayesian posterior quantiles, rather than p-values.

The Benjamini-Hochberg procedure can be given a form of Bayesian interpretation based on *empirical Bayes* methods. Empirical Bayes (EB) methods describe a loose set of related methods that aim to estimate the prior distribution from the data [176, 174]. Bayesian models typically have parameters on the prior distribution. EB methods typically get a point estimate of the parameters governing the prior based on the data, rather than using computationally intensive fully Bayesian approach [176].

In the context of the FDR, empirical Bayes gives a very simple and compelling interpretation. Of the hypotheses tested, a proportion π_0 will be truly null and $\pi_1 = 1 - \pi_0$ will be non-null. Denote the test statistic by the random variable, Z. The distribution of the test statistic assuming that the hypothesis is null is $f_0(z)$, and $f_1(z)$ if it is not. The null distribution, $f_0(z)$, as highlighted above, is known, whilst $f_1(z)$ will not be known — there are infinitely many conceivable ways for the data-producing mechanism to be non-null, and we do not, in general, know what that distribution is.

We can denote the corresponding CDFs to these two distributions as $F_0(z)$, and $F_1(z)$, defining the associated survival curves to be $S_0(z) = 1 - F_0(z)$ and $S_1(z) = 1 - F_1(z)$. Straightforwardly, the survival curve is the probability of the test statistic being in excess of the value z, ie $P(Z > z)$. The observed distribution of test statistics will then be given by the mixture distribution:

$$\begin{aligned} f(z) &= p(z|H_0)p(H_0) + p(z|H_1)p(H_1) \\ &= f_0(z)\pi_0 + f_1(z)\pi_1 \end{aligned}$$

where in the second line we substituted the defined terms above. Doing the straightforward subtraction and integration gives the mixture survival curve to be $S(z_0) = \int_{z_0}^{\infty} f(z)dz$. This is the probability of the test statistic being as or more extreme than z_0, under the mixture distribution. If we take exceeding this threshold as our decision rule, D, the FDR is then the expected proportion of false rejections of the null hypothesis, given a test statistic exceeds z_0. Following (Storey 2003) We show that the FDR using this decision rule is equivalent to the posterior probability that the hypothesis is null, given that test exceeds the threshold, $P\left(H_0|Z \geq z_0\right)$ [175]. First we, define

$$H_{0i} := \begin{cases} 1 & \text{if the } i\text{th considered hypothesis is null} \\ 0 & \text{otherwise} \end{cases}$$

Using the definition of the pFDR gives the following [175]

$$
\begin{aligned}
FDR(z_0) &= \mathbf{E}\left[\frac{N_{RT}}{N_R}\middle| N_R \geq 1, D\right] \\
&= \sum_{k=1}^{k=N} \mathbf{E}\left[\frac{N_{RT}}{N_R}\middle| N_R = k, D\right] P(N_R = k | k \geq 1, D) \\
&= \sum_{k=1}^{k=N} \mathbf{E}\left[\frac{N_{RT}}{k}\middle| N_R = k, D\right] P(N_R = k | k \geq 1, D) \\
&= \sum_{k=1}^{k=N} \frac{1}{k}\mathbf{E}\left[N_{RT} | N_R = k, D\right] P(N_R = k | k \geq 1, D)
\end{aligned}
$$

Under the assumption that all tests are independent, it can be shown that this is equivalent to

$$
\begin{aligned}
FDR(z_0) &= \sum_{k=1}^{k=N} \frac{1}{k}\mathbf{E}\left[\sum_{i=1}^{i=k} H_{i0}\middle| N_R = k, D\right] P(N_R = k | k \geq 1, D) \\
&= \sum_{k=1}^{k=N} \frac{1}{k}\sum_{i=1}^{i=k} \mathbf{E}\left[H_{i0} | Z_i \geq z_0\right] P(N_R = k | k \geq 1, D) \\
&= \sum_{k=1}^{k=N} \frac{1}{k} k P\left(H_{i0} | Z \geq z_0\right) P(N_R = k | k \geq 1) \\
&= \sum_{k=1}^{k=N} P\left(H_{i0} | Z \geq z_0\right) P(N_R = k | k \geq 1) \\
&= P\left(H_{i0} | Z \geq z_0\right)
\end{aligned}
$$

Now that we have shown that the FDR can be interpreted as a posterior probability, we can use this to show, in conjunction with EB, that the procedure in section 4.7.2 controls the FDR at the desired level. A simple application of Bayes' theorem gives:

$$FDR(z_0) = \frac{P(Z_i \geq z_0 | H_{0i}) p(H_{0i})}{p(Z_i \geq z_0)} \tag{4.20}$$

$$= \frac{S_0(z_0)\pi_0}{S(z_0)} \tag{4.21}$$

In general, we know $S_0(z)$. It is simply an integral of the null distribution. The empirical component of this EB procedure comes from using the empirical CDF of test statistics $\hat{S}(z_0)$ as a plug-in for the true $S(z)$, the mixture survival curve. Furthermore, we can finally link this back to the results of the previous section. In terms of a 1-tailed test, we have $p_i = S_0(z_i)$. Additionally, $\hat{S}(z_{(i)}) = \frac{i}{N}$, where $z_{(i)}$ is the ith largest observed test statistic.

The previous section stated the Benjamini-Hochberg procedure to reject all hypotheses based on $p_{(i)} \leq \frac{q_{FDR}}{N} i$ until i_{max} was attained. We can substitute in $S_0(z_{(i)}) \leq \hat{S}(z_{(i)})q$. Rearranging, we get that, $\frac{S_0(z_{(i)})}{\hat{S}(z_{(i)})}\pi_0 \leq \frac{S_0(z_{(i)})}{\hat{S}(z_{(i)})} \leq q$. If we set out the decision rule to be that of the Benjamini-Hochberg procedure, rejecting all those hypotheses for which $p_{(i)} \leq (i/N)q_{FDR}$ is equivalent to rejecting those hypotheses for which

$$\widehat{FDR} = \frac{S_0(z_{(i)})}{\hat{S}(z_{(i)})}\pi_0$$

$$\leq q_{FDR}\pi_0$$

In other words, the Benjamini-Hochberg procedure can be interpreted as only rejecting those cases for which the probability of being null is deemed to be less than $q\pi_0$, and equivalently the FDR, as estimated by the EB procedure [174].

4.7.4 The Benjamini-Hochberg procedure applied to Bayesian Posteriors

Until now, everything we laid out is established and widely known within the literature. Now, we identify standout features of the Benjamini-Hochberg procedure that allow it to be applied to Bayesian posteriors under certain settings. The insight we offer here, has, to the best of our knowledge not been laid out explicitly in other literature. In any case, the formulation we give here is useful in the next chapter, where it is deployed on low-depth frozen brain data.

The first thing to note with the Bayesian interpretation of the Benjamini-Hochberg procedure is that the test statistics are generated by two mechanisms: in proportion π_0, which is not known, the statistics come from the null distribution $f_0(z)$, which is known (typically being associated with data generated from the null value); and the non-null distribution $f_1(z)$ in proportion π_1, neither of which are known. By way of the null survival function $S_0(z)$, test statistics, are massaged into p-values, which are uniformly distributed under the null. Upon ranking those p-values, and retaining only those for which $p_{(i)} \leq \frac{q_{FDR}}{N} i$, we have a list for which the FDR is bounded below $\pi_0 q_{FDR}$.

Though widely applied to p-values, to distinguish null from non-null, this procedure applies far more generally. One simply needs to have a data-generating mechanism with two modes, for which test statistics can be gathered. One mechanism will be operative with rate π_0, and the other $\pi_1 = 1 - \pi_0$. Additionally, one requires that test statistics produced under the π_0 mechanism can be transformed into a uniform distribution. The distribution under the other mechanism need not be known at all. As long as these criteria are satisfied, we can identify which test statistics were produced under π_1 with an arbitrarily controlled FDR.

In general, posterior quantiles for a continuous parameter, θ, are not uniformly distributed, as sections 4.6.2 and 4.6.3 highlighted. In general, if the data generating mechanism does correspond corresponds to the fitted model, posterior quantiles will be uniformly distributed. We also observed that when the model is fit to large amounts of data, posterior calibration could be resorted. In general, we would not expect calibration to be the case when data are

generated exclusively for fixed $\theta = 0$. Asymptotically, under the limit of large amounts of data, posterior intervals have the correct coverage properties under repeated samples of data from the same parameter value. This is to say that, a $\alpha\%$ interval of the posterior will cover the true parameter value $\alpha\%$ of the time when sample sizes are large. As such, in the large sample regime, the quantiles of the true parameter value will be uniformly distributed, even when data is sampled from a fixed θ, not just when θ is sampled from the prior [153]. *As long as we can establish that posterior quantiles of the parameter of interest are uniform when $\theta = 0$, the procedure can be used to obtain a list of non-null applications of the model with a controlled FDR.*

4.7.5 Verifying Benjamini-Hochberg Controls The FDR *In Silico*

We demonstrate this computationally for our negative binomial heteroplasmy uncertainty model in fig 4.16. We generated data exclusively for $\beta = 0$. For each combination of 10, 100, and 1,000 cells, and mean depth of 2, 10, and 100, we generated 50 datasets. We fit our model, which has prior $\beta \sim Normal(0, 1)$. We then computed the posterior quantile of $\beta = 0$, the true data-generating parameter, for each dataset. With only 10 cells, we can see that the posterior quantiles are far from uniformly distributed. As we increase to 100 cells, higher depth data sets are seen to be approximately uniform, with all depth data sets being more or less uniform with 1,000 cells, as demonstrated by the QQ-plots tightly hugging the $y = x$ line.

This test allows us to establish that we can apply the Benjamini-Hochberg procedure to our negative binomial Heteroplasmy uncertainty model when we have 1,000s of cells in our data set. Due to the uniformity of posterior quantiles in the large data regime when the data are produced by $\beta = 0$, we satisfy the conditions of the Benjamini-Hochberg procedure, to distinguish between null and non-null associations, as laid out in section 4.7.3. When there are only 10s or 100s of cells, then straightforward applications of this procedure will fail, as the uniformity of the null quantile is violated. Though this may seem trivial, it was not *a priori* obvious that the model would satisfy this condition with increasing numbers of cells or the number of cells that would be required to obtain this behaviour. As the number of parameters increases with every

new cell, we have an additional latent variable for each cell in the dataset. In some models, this growth of parameters with data size prevents the convergence to uniformity of quantiles [153]. To reiterate, in order to apply the Benjamini-Hochberg procedure to distinguish null from not, with a Bayesian model, one needs to demonstrate the posterior quantiles of the null are uniformly distributed when fit to data generated by the null.

Having demonstrated the uniformity of the negative binomial heteroplasmy uncertainty model under the null, for large numbers of cells, we now demonstrate, *in silico*, that the Benjamini-Hochberg procedure does indeed control the FDR of the model. We generate true parameter values according to $\beta \sim N(0, \sigma_\beta)$. We denote setting $\beta = 0$ as corresponding to $\sigma_\beta = 0$. We draw parameter values from the normal with $\sigma_\beta = 0, 1, 2$ in fixed ratios, with the largest component of the ratio being sampled 500 times. For instance, if we sample $\sigma_\beta = 0 : \sigma_\beta = 1 : \sigma_\beta = 2$ in ratio $10 : 1 : 1$, we produce 500 samples of $\beta = 0$, and 50 samples from each of $N(0, \sigma_\beta = 1)$ and $N(0, \sigma_\beta = 2)$. We then sampled a data set for each of these β parameter values, and fit the model to each dataset separately. The motivation for the additional use of $\sigma_\beta = 2$ is to emphasise that the non-null parameter values do not need to be generated from a distribution that exactly matches the model being fit to the data, which in this case is $\sigma_\beta = 1$.

Using the quantiles of the posterior distribution, we applied the Benjamini-Hochberg procedure. To compute the posterior quantile of the null, $\tilde{q}(0)$, we sampled from the posterior 4,000 times, and estimated the quantile of the null by the proportion of samples for which $\beta \leq 0$ [5]. For every regression we perform, we have $\tilde{q}(0)$. We can rank these and reject all hypotheses for which $\tilde{q}(0)_{(i)} \leq \frac{q_{FDR}}{N} i$, where q_{FDR} is the threshold at which we would like to control the FDR, as in section 4.7.2. This results in a list of models where we arrive at the decision that the parameter is not-null. We can compare to the underlying, simulated truth and see the proportion of false discoveries in the simulated batch. This is done for datasets with mean depth 2, and mean depth 10. The results are displayed in table 4.1.

We can see that the observed rate of false discoveries is kept below the $\pi_0 q_{FDR}$ bound. As such,

[5] To make this equivalent to a two-tailed test, we transformed the posterior quantiles, $\tilde{q}(0)$, into $\tilde{q}'(0)$ by $\tilde{q}'(0) = (2\tilde{q}(0) - 1)(2\mathbf{I}_{(\tilde{q}(0) \leq 0.5)} - 1) + 1$, where $\mathbf{I}$ is the indicator function. If $\tilde{q}(0) \sim Unif(0, 1)$ when the model is fit to data generated from the null, so will be the distribution for $\tilde{q}'(0)$. $\tilde{q}'(0)$ captures how extreme $\beta = 0$ is in either tail of the distribution. As the uniformity assumption holds, we can apply BH procedure to $\tilde{q}'(0)$

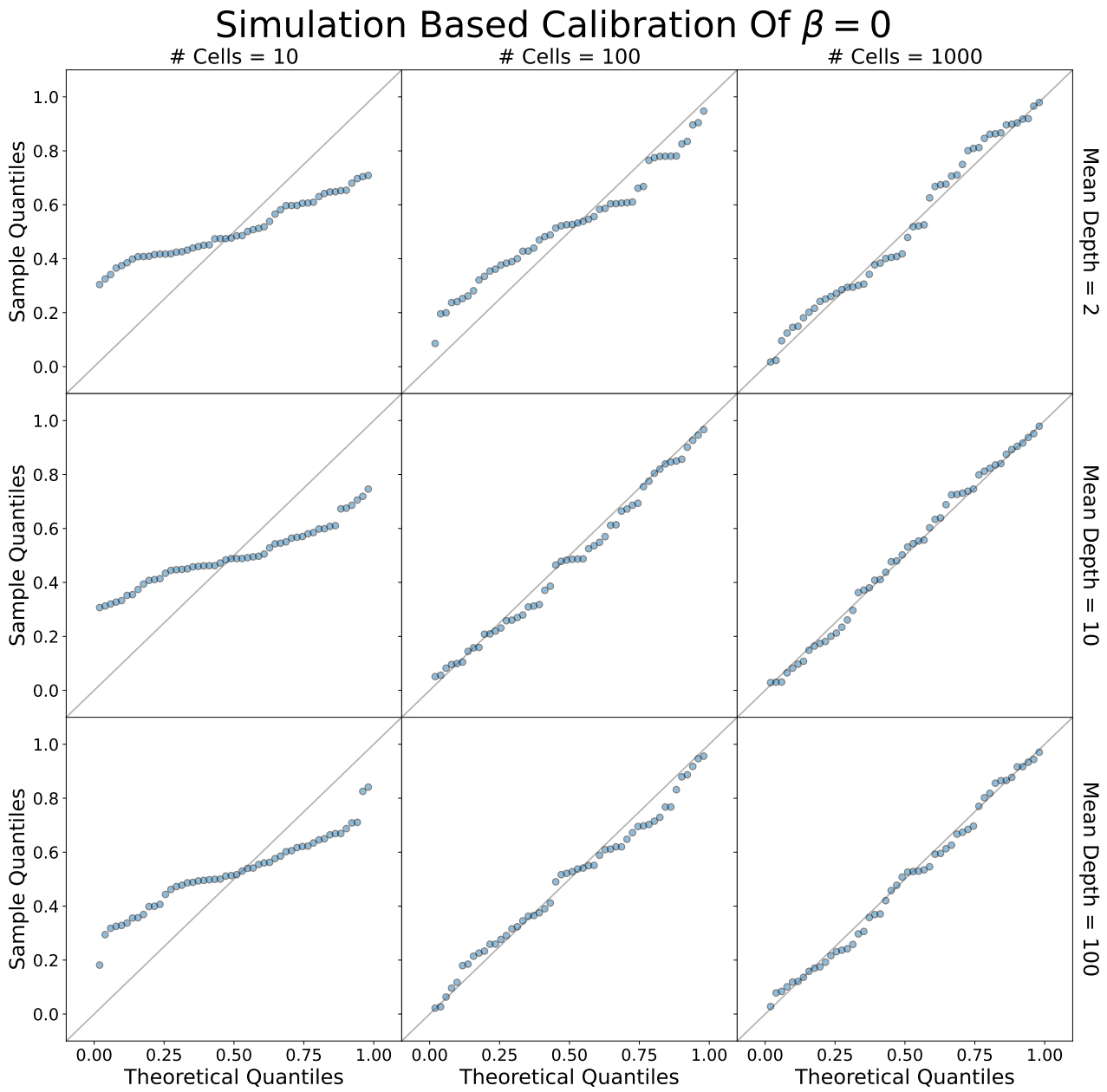


Figure 4.16: For large numbers of cells, posterior quantiles of the true β become uniformly distributed for when data is exclusively generated from the null $\beta = 0$. QQ-plots compare the sampled quantiles of distribution posterior quantiles to the theoretical quantiles of the $Unif(0, 1)$ distribution. We can see that for small numbers of cells, substantive departures from uniformity are present. Under such conditions, applications of the BH procedure to posterior quantiles of $\beta = 0$ would be inappropriate as they lack the required uniformity. As more data is gathered convergence to uniformity is observed, even with low-depth data. Application of the BH procedure to Posterior quantiles will control the FDR at a specified level of q_{FDR} as they behave sufficiently like p-values.

these simulations serve to demonstrate how the Benjamini-Hochberg procedure successfully controls the False Discovery Rate.

Statistic	$\pi_0 q_{FDR}$	False Discovery Rate		Acceptance Rate	
Mean Depth		**2**	**10**	**2**	**10**
10:1:1	0.042	0.000	0.043	0.097	0.115
2:1:1	0.025	0.019	0.022	0.321	0.371
1:1:1	0.017	0.010	0.014	0.445	0.491
1:2:2	0.010	0.007	0.009	0.545	0.594
1:10:10	0.004	0.007	0.002	0.692	0.692

Table 4.1: We simulated data with mean depth 2 and mean depth 10. We then sampled "true" values according to $\beta \sim$Normal $(0, \sigma_\beta)$, with, $\sigma_\beta = 0, 1, 2$ in different ratios, with $\sigma_\beta = 0$ denoting samples coming from the null. Different ratios of these values were used to give samples with various proportions of β being null and variations of distributions of non-null values. The largest component of the ratio was sampled 500 times, with corresponding numbers of samples coming from σ_β with smaller values. For example $\sigma_\beta = 0 : \sigma_\beta = 1 : \sigma_\beta = 2$ in ratio $10 : 1 : 1$ indicates 500 samples were set to the null, with 50 values of β being sampled from each of $Normal(0, \sigma_\beta = 1)$ and $Normal(0, \sigma_\beta = 2)$. The theoretical bound of the BH procedure is $\pi_0 q$, though π_0 is unknown in real applications. The BH procedure applied to posterior quantiles has realised false discovery rates tightly corresponding to the theoretical expectations, even with mean depth 2. We can see that as the ratio of non-null values increases, the observed acceptance rate also increases.

Strictly, the FDR, is an expectation, thus an individual numerical sample might occasionally observe proportions higher than the control level, as is the case for the sample with σ_β ratio 10:1:1 and mean depth 10, being at 0.043 compared to the theoretical expectation. By the central limit theorem, using larger sample sizes, or repeating the procedure and averaging would bring the observation in line with the theoretical bound, at the expense of computational effort. As it stands, these results are easily attributed to simulation noise. The results in the table are displayed in a bar chart in figure 4.17, displaying the monotonic trends with a decreasing ratio of null to non-null samples. Of further note is that the number of results with a positive decision for ratio 10:1:1 with mean depth 2, was zero, which is understandable given the larger uncertainty that would be observed with mean depth 2 samples, and that returning no results with extreme enough values to merit rejection is plausible for this ratio.

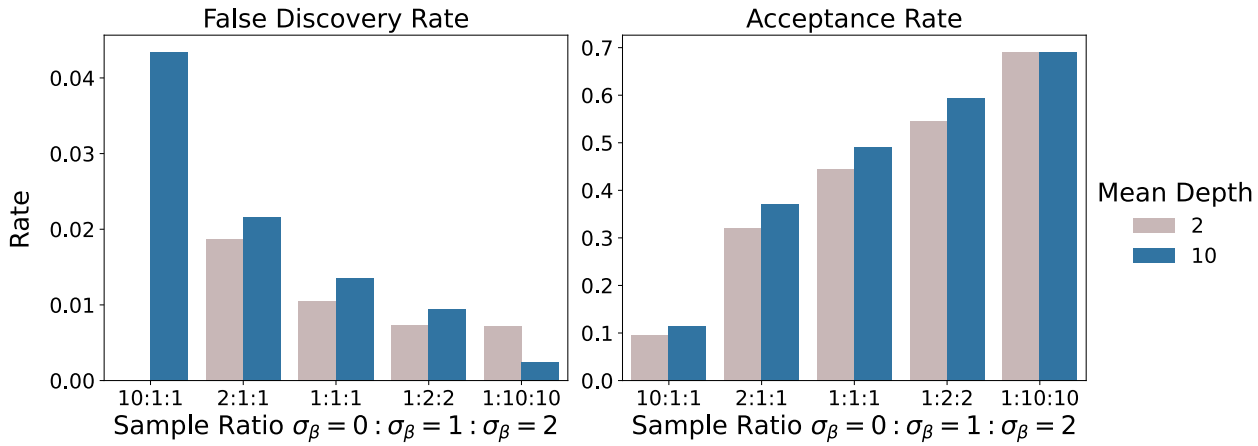


Figure 4.17: Visualisation of the values in table 4.1. One can see clearly how the BH procedure succeeds in controlling the false discovery rate near the theoretical expectations, $\pi_0 q_{FDR}$. The FDR of the low depth data is 0 for the 10:1:1 as no models were confidently identified as being "significant" by the procedure. The acceptance rate of the data set with mean depth 2 marginally lags behind that of mean depth 10. Intuitively, this is due to the broader posteriors with lower-depth data, making it more difficult for the model to produce posterior quantiles which exceed the significance threshold.

4.8 Conclusion

In this chapter, we set out with a single aim: to construct a model capable of measuring the association between heteroplasmy and gene activity at the cellular level in low-depth data. Given the high volume of low mitochondrial depth single-cell data sets, how could we possibly extract meaningful biological insight from the data? In low-depth data, the problem arises that the heteroplasmy of each cell within the data is deeply uncertain and that honest analyses of the data ought to reflect this fact.

Due to the specific nature of our problem, we turned to Bayesian inference to arrive at a solution. As a framework, it naturally aims to quantify the uncertainty in any inferences. Furthermore, its flexibility allows practitioners to develop models appropriate to the situation at hand. At the start, it is not obvious which model ought to be deployed and whether the model one decides to deploy will be reasonable. Usefully however, Bayesian inference can allow practitioners to think *generatively* about the data at one's disposal, rather than pulling an inappropriate method off the shelf. By deploying tools from Bayesian inference, we were able to develop a simple, interpretable model, capable of leveraging what information is available in the data.

In the development of the model, its computational instantiation, and the application of this model to our data, we strove to adhere to modern approaches to Bayesian workflow [24, 155]. We started by deriving a simple model of heteroplasmy uncertainty based on single-cell read data, and a straightforward, mechanistically grounded, model of gene activity. The structure of both of these models was captured by directed acyclic graphs, capturing the conditional dependencies between variables. We then combined these two models in a single negative binomial heteroplasmy uncertainty regression model. As an illustration, we demonstrated how a maximum-likelihood approach of the negative binomial regression model that used naive point estimates of heteroplasmy failed to recover true parameter values in low-depth data, whilst our Bayesian model succeeded. When viewed from the angle of Bayesian workflow, this was a powerful demonstration of the utility and capability of our model derived from first principles, compared to conventional approaches which might ignore the uncertainty in

heteroplasmy estimates. Furthermore, it demonstrated large step to achieving the primary aims of this chapter: to accurately quantify the relationship between heteroplasmy and gene-activity in low-depth data.

Establishing a model capable of recovering a parameter value from a single dataset using a probabilistic programming language is not enough to demonstrate its efficacy or its validity. One must demonstrate that one can recover the true parameter in a range of values in the parameter space. A good computational test of both the model and the computational implementation is *Simulation Based Calibration.* Simple theoretical arguments imply that when the model is fit to parameters generated from the prior, posterior quantiles of the true parameter ought to be uniformly distributed. Our model and the HMC implementation satisfied this requirement.

Furthermore, we tested the model's ability to recover true parameters under a range of settings. We first did this by using the model to evaluate data from a very low depth regime, with cells on average only having, on average, two reads, with which the model was still identifiable, and passed the SBC checks. Secondly, we test the model by fitting it to adversarial data, generated in a way that was at odds with the model's prior. As expected, the model was not calibrated under this setting when few cells were present in the data. However, by increasing the cell count to numbers realistically obtained in single-cell experiments, we were able to show that the model's calibration was restored. This is a feature of some Bayesian models, which is that as the amount of data increases, the posterior becomes well calibrated, in the sense that $\alpha\%$ credible intervals do in fact cover the true parameter value $\alpha\%$ of the time [153]. This is not guaranteed, however, in models in which the number of parameters increases with the size of the data, such as ours. Using computation to demonstrate calibration in this regime serves as a strong indication that our model is well-behaved in this limit, despite each new cell adding a parameter to the model. It further tests the behaviour of the model and the computational algorithm approximating it. Variational inference for instance will not produce samples from the exact posterior in the manner guaranteed by HMC. As such, samples from variational inference will not in general exhibit the uniformity in their posterior quantiles during SBC [177]

We wrapped up by establishing a framework for deploying the model in a large-scale, multiple-testing framework, which further protects us from concerns around a potential preponderance of null hypotheses in the underlying data-generating mechanism. By leveraging the structure of the Empirical Bayes interpretation of the Benjamini-Hochberg procedure, and deploying it in a novel manner to Bayesian posterior quantiles rather than p-values, we demonstrated that this application was not a naive misuse of statistics, a blunder akin to taking the average of p-values or some other such fallacy. *If the posterior quantiles of the null are uniformly distributed when fit to data generated by the null, then the Benjamini-Hochberg procedure can be applied.* In general, this is not the case, as a wide class of Bayesian models and the data they're fit to do not satisfy this requirement, with examples seen in sections 4.6.3 and 4.7.4. With large amounts of data, however, some do. We demonstrated, via simulation, that our negative binomial heteroplasmy regression model behaved in the desired manner. Furthermore, *in silico*, we put our claim to the test and saw that, indeed, when the samples were generated from the null in different proportions, the BH procedure did in fact control the false discovery rate at the theoretical threshold.

This chapter required an accumulation of linked concepts to achieve its primary goal of measuring the association between heteroplasmy and gene activity in low-depth data. We have a model capable of quantifying the association between genes and heteroplasmy in low-depth data, and we have demonstrated a valid method of controlling the false discovery rate that applies to this model. The domain of validity was strenuously tested, whilst pragmatically addressing concerns surrounding compute time and interpretability. In doing so, we have unlocked an approach to analysing datasets that might otherwise have been deemed unworkable. This chapter has set the methodological foundations for the analysis that lies ahead in the next chapter. There, we get to see the results of this chapter in put into action on real-world data.

Chapter 5

Mitochondrial Mutations In The Developing Brain

"The brain is a wonderful organ; it starts working the moment you get up in the morning and does not stop until you get into the office." – Robert Frost

5.1 Summary

Earlier chapters sketched the potential of single-cell genomics datasets to make novel discoveries in a range of datasets. Then, by exploring publicly available single-cell datasets, I presented empirical evidence of a previously unnoticed component of human ageing: the accumulation of *cryptic* mtDNA mutations reaching pathological levels coincident with the decline of human health. At the same time, it was observed that limitations in the breadth and depth of coverage can prevent us from probing the full length of the mitochondrial genome. In the previous chapter, I developed methodological tools to surmount these limitations for mutations that were present in many cells with low depth of coverage at the position of the mutation. The model allows us to measure the association between the heteroplasmy of a mutation and gene activity. Now I apply this model to real data.

The brain is one of the most energetically demanding organs in the body. Mitochondria, and by extension the mtDNA, play a key role in its healthy functioning. Our collaborators at the Mitochondrial Biology Unit in Cambridge sequenced embryonic human brain cells to explore the landscape of mitochondrial mutations in early brain development using frozen tissue samples. Supplied with the raw data, I performed the standard steps of single-cell analysis, using an improved pipeline to the one developed in Chapter 3. I showed that the depth of coverage was heavily restricted, explained due to the inability to capture mitochondrial DNA effectively from frozen tissue samples. I show that through a simple model of sequencing error and a conservative use of statistics, widespread mutations found in early developmental cells can be identified, despite severely limited depth of coverage in each cell. The limited depth spurred me to make the methodological developments explored in detail in the previous chapter. With these tools in hand, we were able to understand some of the mutations present in early brain development and identify a plausible set their of functional effects on the developing brain.

Outside of data collection and sequencing, performed by Nell Nie, all the work discussed in this chapter was performed by myself.

5.2 Introduction

Thus far in the thesis, we have laid out how recent developments in single-cell genomics sequencing technologies offer the potential for discovery. We gained novel insight into mammalian ageing through the use of a sophisticated computational pipeline that leveraged publicly available datasets. The accumulation and progressive increase in the heteroplasmy of *cryptic* mtDNA mutations was seen to coincide with the onset of human ageing. Moreover, we identified associations between the presence of these mutations and changes in gene expression which are consistent with multiple of the hallmarks of aging [115].

In this chapter, we shift our focus from the end of life to the beginning. In particular, we focus on the development of the human brain. Mitochondria are integral to the regular function of neuronal cells, and mitochondrial decline has been linked with neurological disorders. Moreover, mtDNA variation as both deletions and point mutations have been linked with dysfunctional brain regions, Parkinson's disease, and ageing in neurons [178, 179, 180, 181, 182]. In spite of such substantive evidence for the role of mtDNA in the healthy functioning of the brain, very little is known about the landscape of mtDNA point mutations present during brain development. Here, we pioneer the cartography of this landscape using single-cell ATAC data to probe brain development at five Carnegie stages across 3 different brain regions. This was made possible through a collaboration with Nell Nie and Patrick Chinnery at Cambridge University's Mitochondrial Biology Unit. They obtained the samples and sequenced the data.

With the shift from ageing to development, our emphasis on cryptic mutations diminishes. At the earliest phases of life, cryptic mutations have had little time to emerge. As the Moran model simulations showed in Chapter 3, there is a negligible chance of mtDNA mutations reaching functional thresholds in early life. Furthermore, due to the low heteroplasmy of the few cryptic mutations that would be present, detecting them, even in single-cell datasets with deep coverage, would be challenging. This is due to the difficulty of distinguishing between true mutations and errors that might arise in the sequencing process or during PCR. With these technical and biological considerations in mind, we focus our attention on widespread mtDNA mutations that are found in multiple cells in each tissue sample. Mutations that are

this widespread so early on in development are likely to be those that have been passed on through the germline or that took place very early in development.

Standard ATAC sequencing techniques are single-nuclei techniques, which do not enrich for their mitochondrial content. Others have been able to modify these methods to enrich the mitochondrial content [84]. The samples we analysed were frozen. Whilst freezing was essential to the preservation of the samples, this placed further constraints on the quality of the mtDNA data that could be obtained. A standard step in performing library preparation is cell lysis, allowing the mtDNA to be extracted. Frozen cells prevent this meaning that cell nuclei had to be extracted from the sample. As a consequence of this, only mtDNA in the presence of the peri-nuclear sheath were extracted and then sequenced, greatly limiting the number of biological mtDNA samples that could be sequenced per cell. Such restricted sampling placed strong limitations on both the breadth and depth of sequencing that was attained within this data, consequently limiting the power to identify cryptic mutations.

After describing the sample collection, performed by Nell Nie (section 5.3), we outline an improved bioinformatics pipeline that is adapted to handle the structure of 10x scATAC data (section 5.4.1). This pipeline is built using `snakemake`, making manageable a more complicated set of processing steps to be run on substantially more data than earlier chapters. The new pipeline also has additional steps to overcome obstacles associated with low-depth data. We use gene activity data to perform a clustering analysis, revealing broad structures and categories of cell types in this data (section 5.6). Low-depth makes identifying mutations challenging as sequence mis-matches from the reference genome inevitably occur in the data due to sequencing errors. We develop a conservative statistical test for identifying widespread single-cell mutations from sequencing noise (section 5.7). We go on to apply this method to each of our data sets separately and recover a consistent pattern of mutations across all three brain regions for each individual subject in the study. Moreover, the set of identified mutations display the classic transversion profile of mtDNA mutations. Having confidently identified a set of widespread mutations, we go on to deploy the methodology developed in Chapter 4 (section 5.8). We discuss some of the software engineering challenges posed by trying to repeatedly apply a high-dimensional Bayesian model *millions* of times, before performing posterior predictive checking

of some of the fitted models. Finally, we end by looking at some of the functional effects that negative binomial regression with heteroplasmy uncertainty was able to uncover (section 5.9). This gives preliminary evidence that the heteroplasmy of mtDNA mutations present in early developmental is associated with genes playing a role in mitochondrial network structure and that this could have disruptive effects on neuronal structure and functioning.

5.3 Sample Collection

Our experimental collaborators obtained frozen, embryonic human brain. For the interested reader, the following library preparation steps were performed by Nell Nie to isolate cell nuclei:

Human tissue samples were processed in the safety cabinet at containment level 2. Homogenization buffer2 and Percoll solution were made fresh and filtered before use. Fresh frozen brain material was sliced and leave on dry ice. In a pre-chill clean Dounce homogenizer, brain tissues were processed in homogenization buffer2 gently using 7-14 strokes with pestle B, performing on ice and avoid foaming. The homogenate was filtered through a $40\mu m$ strainer and centrifuged at 3000 x g to enrich nuclei. The nuclei pellet was then resuspended in homogenization buffer and mixed with Percoll solution to create a 27% Percoll-nuclei solution and centrifuged at 20,000 x g for 15 minutes at 4 degrees. The purified nuclei layer was kept, spun down, and re-suspended in buffer compatible with 10x Genomics sequencing. The intact nuclei were stained using trypan blue and quantified by Countess cell counter.

Once libraries had been prepared, samples were sent for sequencing. Raw `.fastq` files containing all sequencing reads were then returned to the lab so that data analysis could begin.

5.4 Computational Pipeline

The structure of the data in this chapter has a resemblance to the data from Chapter 3. However, the sequencing modality here, primarily 10x scATAC data, is a more complicated form of data to analyse. This is due to the number of additional steps required in the analysis (described below). Due to the low-depth of the mitochondrial genome in the data analysed in this chapter, additional steps were required to implement the negative binomial regression heteroplasmy uncertainty model from Chapter 4. Often, fitting a single high-dimensional Bayesian model can be considered computationally demanding. To measure the association between many genes and mutations required performing Bayesian inference millions of times on models with 1,000s of parameters and is discussed further in section 5.8.2. These extra processing and analytic steps necessitated adjustments to the structure of the pipeline. Furthermore, the number of datasets analysed in this chapter far exceeds the number processed and analysed in the collaborative work Chapter 3. To make this additional complexity and scale manageable as an individual researcher requires more sophisticated workflow management tools.

For this, we turn to `snakemake` [183]. The essence of `snakemake` is enabled by decomposing processing pipelines into "*rules*". Generally, each rule specifies how to produce a set of output files from a set of input files. Each step in a pipeline takes the form of a `python`, `R`, or `bash` script, that reads input files, and writes the corresponding output files to disk. A complete workflow or pipeline can thus be composed of rules, in which the output of one rule is used as the input to a downstream rule. In this way, the standard workflows encountered in computational research can be represented as Directed Acyclic Graphs (DAGs), with nodes representing rules, and edges representing connecting rules, where the output of one rule serves as the input to another (see Figure.5.1). `snakemake`, therefore, offers a natural framework for writing and extending complex computational workflows. By making use of a "snakefile", a researcher can specify a set of rules with the corresponding inputs and outputs, linking them together. Additionally, snakemake allows for conda environments to be employed, allowing specific versions of `python` and `R` packages to be bundled into the same environment, facilitating the reproducibility [184]. The utility of the pipeline I wrote is emphasised by its use by others in our group [185].

5.4.1 Snakemake Pipeline

MakeRDS `CellRanger-ATAC` outputs a tabular file where each row corresponds to a unique fragment captured by the assay. This fragment was created by a transposition event, one occurring at each end of the fragment. Due to a limitation of file formats on available on our compute cluster at the level of the operating system, we have to convert the data to the `.rds` format as an intermediate step in the analysis. This is done using `Seurat`, one of the most popular single cell genomics libraries [163, 186].

Process ATAC Using the files output from `CellRanger-ATAC` and the previously created `.rds` file, we perform peak calling using `MACs2` using the `R` package, `Signac`, as an interface [187, 163]. In order to visualise cells, *Latent semantic indexing* (LSI) was performed on the peaks matrix [188]. UMAP, and nearest neighbour inference was then performed on the dimension 2-30 of the LSI matrix, as the first LSI dimension typically correlates strongly with the sequencing depth of cells. The first two UMAP dimensions were then used to visualise the dataset. As a simple way quantify the activity associated with each gene and its promoter, the number of fragments within the gene body and up to 2kb upstream (the area near the promoter) are summed. This is performed for each gene and cell in the data set to yield a *gene activity matrix*, which serves as a functional analog of the *gene expression matrix* found in scRNA sequencing experiments. This gene activity matrix is used down stream analysis to look for active genes and idenfity putative cell types within the data. The gene activity matrix for each sample is then saved to the `.rds` format for later re-use in `R`.

Activity RDS To CSV An additional step in the pipeline is to read in the gene activity `.rds` file and convert it to a `.csv` file for subsequent use in python.

Get Mitochondrial `.bam` File

An output of running `CellRanger-ATAC` is an aligned `.bam` file. `.bam` files are a standard format for alignment data and are the binary cousins of Sequence Alignment/Mapping (`.sam`) files [189]. For the purposes of calling variants on the mitochondrial genome, only reads which

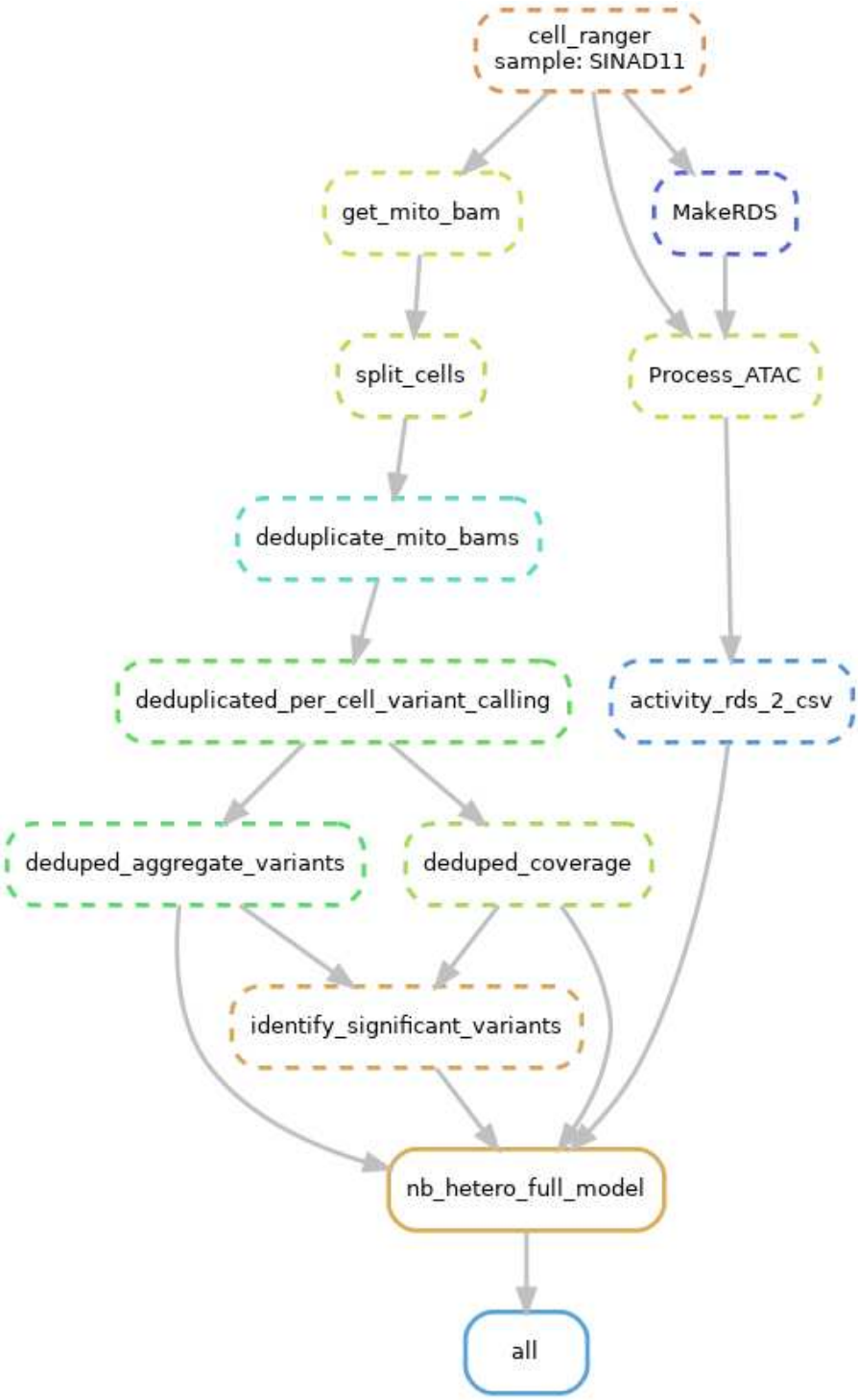


Figure 5.1: Snakemake computational workflow currently deployed for this project, represented as a Directed Acyclic Graph (DAG). Each node in the graph represents a *rule*, each rule specifying how to to obtain a set of output files from a set of input files. A pipeline can be thought of as a set of sequential rules, the inputs of one rule being generated by the outputs of another.

align to the mitochondrial genome are required. Therefore, we extract all reads that align to the mitochondrial genome using `samtools` (version 1.11) [137].

Split Cells

The mitochondrial `.bam` file resulting from the previous step contains only reads align to the mitochondrial genome as well as various tags (descriptive metadata fields) associated with each read. In order to estimate heteroplasmy at the cellular level, we separate the mitochondrial `.bam` file into 1,000s of separate mitochondrial `.bam` files, one for each cell. This is done using a custom python script which uses `pysam`, a python wrapper for samtools. Reads are separated into different `.bam` files using the CB (CellBarcode) tag in the `.bam` files.

Deduplicate Mito`.bams`

The cellular `.bam` files produced up to this point contain PCR duplicates. Due to the way DNA is fragmented, the odds of separate sequencing reads stemming from different biological reads sharing these coordinates by chance is small. Any two reads that share the same coordinate are thus presumed to arise from PCR amplification steps. To eliminate these duplicates, samtools is deployed to deduplicate all reads that share the same start and end points. This results in a deduplicated mitochondrial `.bam` file for each cell.

Deduplicated Per Cell Variant Calling

Variant calling is then performed using another custom `python` script deploying `pysam`. This is the same as Chapter 3. Each cellular `.bam` file has variants called separately allowing the number of reads harbouring a SNP or the reference allele to be counted. Simultaneously, we count the number of reads covering every position of the genome allowing us to assess the depth of coverage across the mitochondrial genome at the per cell level. Together, these values are used as predictors for the negative binomial heteroplasmy uncertainty model developed in Chapter 4.

Deduplicated Aggregate Variants & Coverage

With the mitochondrial genome annotated in the previous steps, we can aggregate all of the variants from all the cells into a single file. Likewise, we can aggregate all the depth of coverage

data into a single file. These files are then used in downstream analysis to assess data quality and as an input to our Bayesian inference procedure.

Identify Significant Variants

Sequencing and PCR error can be a source of false variants in sequencing data. In order to overcome this we performed a statistical test to identify variants that were significantly different from a conservative error threshold (see. 5.7.3).

Convergence In The Pipeline

Once all the previous steps of the pipeline have run, the two strands of the pipeline converge. The cellular heteroplasmy data obtained in strand two is used to try and predict the gene activity information output by strand one. The rest of pipeline thus follows a single, sequential chain of execution.

Negative Binomial Heteroplasmy Full Model

We measure the strength of the relationship between heteroplasmy and gene activity. This is done using a Bayesian Negative Binomial model which allows for measurement error in the heteroplasmy predictors. This is the implementation of the model developed in Chapter 4. In addition to quantifying the strength of this relationship, this Bayesian model allows us to quantify the uncertainty strength of this relationship, flagging relationships between genes and heteroplasmy at different levels of confidence. Section 5.8.2 discusses this in more detail.

5.4.2 Large Scale Pipeline

The workflow just outlined highlights the pipelines use for a single sequencing dataset. The scATAC data studied in this chapter were all produced in an analogous manner. As such, the same pipeline can be reapplied to different data sets in an identical manner. `snakemake` makes this scaling straightforward by specifying the set of analysis files we would like to generate, and inferring all the samples to do this for by using the range of existing file paths. In this

manner, the full computational workflow is automatically configured. The results of this scale are visualised in Fig 5.2.

Figure 5.2: In a lot of computational research the same set of analytic steps are re-applied to different datasets. `snakemake` naturally allows for this generalisation by simply expanding the list of all output files to be generated, leading to broader dag comprised of identical computational steps. `snakemake` keeps track of which files have been generated and which have not. By knowing which files are currently missing, snakemake will deploy the appropriate rules to generate them. This allows for efficient use of University HPC resources where compute time is finite. If a job fails due to timeouts, `snakemake` will track the dependency. The job can then be re-submitted, allowing the pipeline to pick up from where it left off.

5.4.3 Notebooks

Many of the exploratory analyses were performed manually. These can not straightforwardly be included in an automated pipeline. These analyses were conducted in jupyter notebooks running `R` and `python`, and were used to visualise the results displayed in this chapter.

5.5 Single-Cell Analysis

Before investigating the association between heteroplasmy and gene activity in early development, we conduct a singe-cell analysis in line with standard practice [22, 163, 186]. This brief exploratory analysis allows us to perform quality control and reveals structure in the data consonant with developmental tissues in the brain.

Single-cell ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) is a technique used to study the epigenetic landscape of individual cells. Section 5.3, described the procedure we used to extract nuclei from the cell. The transposase tn5 cuts open chromatin regions, those segments of DNA that are not wrapped around histones, cutting out fragments of DNA at these locations. The fragments originating from these open regions then have sequencing adapters appended before the subsequent steps of library preparation are performed, followed by sequencing. Sequencing reads are then aligned to the reference genome. Regions of the genome where reads are enriched are identified as peaks using a peak calling algorithm. These enriched regions, or "peaks," indicate areas of the genome that were accessible to the transposase and are therefore likely to be functionally important for gene regulation.

5.5.1 Clustering

To understand the structure of the data, we integrated the peaks matrices from all data that we analysed. First, Seurat objects were created from the outputs of `cell ranger`. These Seurat objects we then combined using the `signac` package [163]. On the combined object, peaks were called using the `MACS2` [187]. Cells with a nucleosome signal ≤ 4 and tss enrichment ≥ 2 were retained for further analysis in line with standard practise [163]. Latent semantic indexing (LSI) was performed on the peaks matrix. The UMAP algorithm was run on dimensions 2-30 of the LSI matrix, as the first dimension is typically strongly correlated with sequencing depth and is thus excluded from the analysis. KNN clustering is then performed on these LSI dimensions with the results displayed using the first two UMAP dimensions [190, 163]. We apply default parameter settings in all following analysis, preventing additional bias from emerging in between sample comparison. As these algorithms are used for a brief exploration of

the data, we do not explore them in technical detail. We note however, that UMAP is a tool that is sensitive to hyperparameter values and controversy surrounds the application of non-linear dimensionality reduction tools in the analysis of single-cell data [191]. Any reduction of data by four orders of magnitude to a 2-dimensional visual representation inevitably dispenses with meaningful structure in the data. As such, we try to ground the following observations about gene activity in what is previously known from literature on brain development and without implausibly claiming to identify cell types at a fine granularity. We show that the findings are largely repeated across the data in a manner that is consonant with existing knowledge of the developing brain.

We compute the gene activity by summing the total number of fragments intersecting the gene body and up to 2,000 base pairs upstream of the transcript start site to include promoter region. This computation was accomplished using `signac`. Gene activity in this manner can be thought of as the total number of accessible regions along the length of the gene and in the promoter region. Gene activity scores of this kind are known to be positively correlated with expression levels [192].

The results for peaks aggregated across all samples are displayed in figure 5.3. When all data are combined, we can see that distinct clusters emerge. When coloured by Carnegie Stage, data from different individuals are observed to cluster together, though there is substantive overlap between samples. The right panel of the figure clarifies the situation showing that the data clusters most strongly by brain region. This suggests that the cell types found within each brain region are stable through development, and that brain region explains the pattern gene accessibility in early brain neurodevelopment.

Data from different Carnegie stages were sequenced separately precluding batch effects from explaining the strongest form of clustering, which is by brain region. We repeated the analysis for each Carnegie stage separately. We excluded CS16 from the aggregation due to the poor quality of the cerebellum data in this sample (see section 5.4). At this finer level, clustering by region is still observed within samples from the same individual. We subsequently analyse data sets separately for each brain region and Carnegie stage. An additional observation is

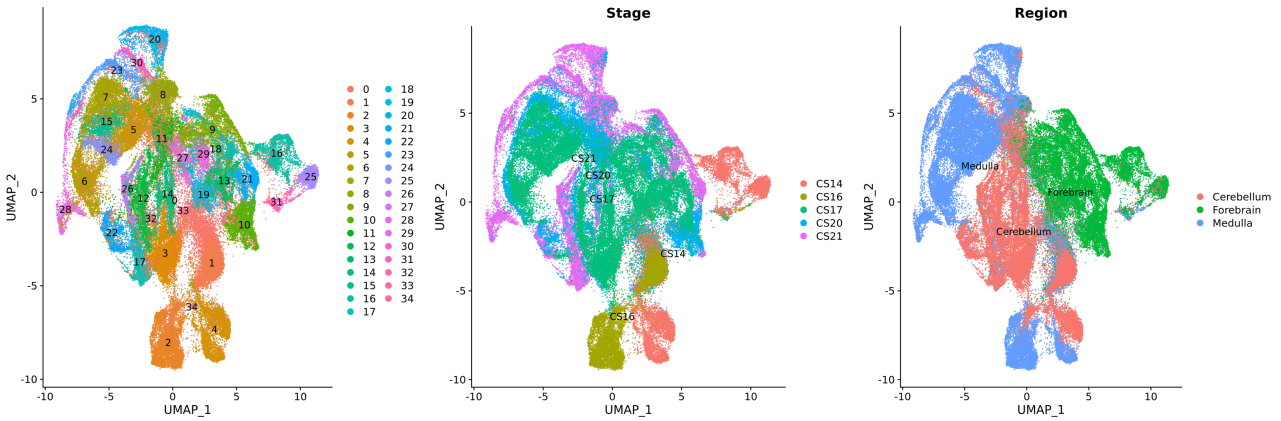


Figure 5.3: We integrated ATAC data from across all Carnegie stages and regions. (Left) shows all the data with clusters obtained using the default parameter values in `Seurat`. (Middle) The UMAP indicates that the data clusters by individual Carnegie states. (Right) Importantly and in line with expectations, the data most strongly clusters by brain region. This behaviour is anticipated as similar cell types are expected to be present in identical brain regions from different samples, though the composition will change as development progresses. Importantly, data from different Carnegie stages were sequenced separately, precluding batch effects from explaining the strongest form of clustering, by brain region.

that UMAPs start in a globular structure, becoming more dispersed through development from CS14 to CS21. This is consonant with cell type differentiation, in which the different expression profiles of different cell types become more pronounced. Nevertheless, we once again emphasise that there exists evidence to dispute claims that non-linear dimensionality reduction techniques such as UMAP or tSNE are faithful representation of local and/or global structure in the data. As such, we do not lean on this interpretation later in this chapter when it comes to mapping the landscape of mitochondrial mutations and assessing the roles that such mutations play during development. We present a brief exploration of the data with evidence that various expected cell types are present in the data based on expression markers found in the experimental literature.

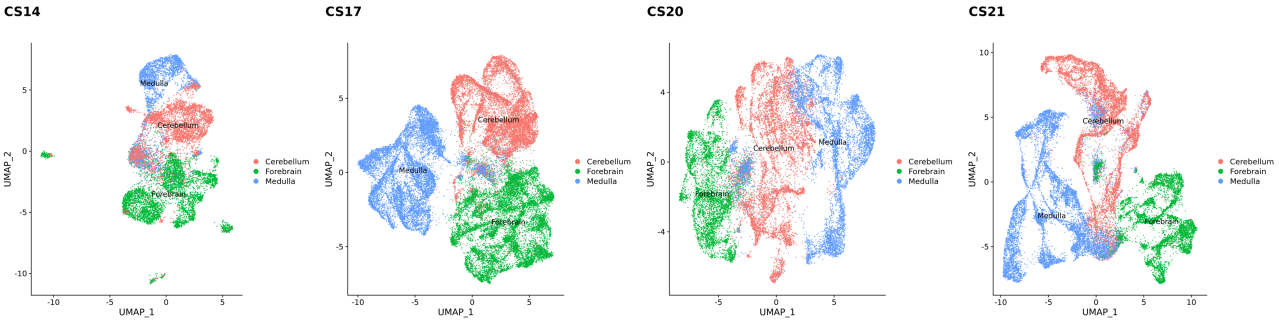


Figure 5.4: UMAPs for data aggregated by Carnegie stage display clustering by brain region. This is in line with the composition of cell types from different brain regions being distinct. The pattern of globular structures becoming more dispersed through development is observed when data are aggregated at the level of Carnegie stage.

5.6 Cell Type Annotations

Embryology standardises the periods of development into 23 *Carnegie Stages*, based on morphology and the formation of certain structures, rather than chronology. As such, it is possible to make comparisons between different mammalian species using Carnegie stages [193]. In the following, we draw on previous understanding of cell type formation during neurogenesis in other species to guide our cell type annotations.

At these early stages of development, cell type differentiation is still ongoing. Gene activity data measures the accessibility of different regions of the genome, though is distinct in meaning from gene expression datais positively correlated with gene expression, though is derivative [192]. Nevertheless, due to nature of the data, identifying meaningful cell types at a fine grained level is an unreasonable aspiration. As such we sought coarse-grained categories of cell types that would allow us to gain insight into the structure present in the data.

Much about human brain development is unknown, but much has been learned from rats/mice, and chickens. It is possible to find corresponding Carnegie stages of development between different mammals and find the developmental time points that they correspond to. With this, we can take what is known about cell type development in other species to make inferences about the cell types that we would expect to see in our human samples [194, 195].

5.6.1 The Forebrain

For the developmental stages displayed in the data, neurogenesis begins [194]. SOX2 is known to be a marker of stem cells and plays a key role in neuronal differentiation [196, 197]. Stem cells expressing neurogenic genes at this early stage of development are indicative of radial glial cells, among the most common cells in the brain [195]. In the forebrain, PAX6 begins to stimulate asymmetric, neurogenic cell divisions, determining cell fate [198]. In conjunction PAX6 and SOX2 expression are indicative of radial glial neuroprogenitors. In Fig 5.6, we display the activity of these genes for the forebrain data. It can be seen that SOX2 and PAX6 activity are present in overlapping clusters, with the overall level of PAX6 activity being greater. We can

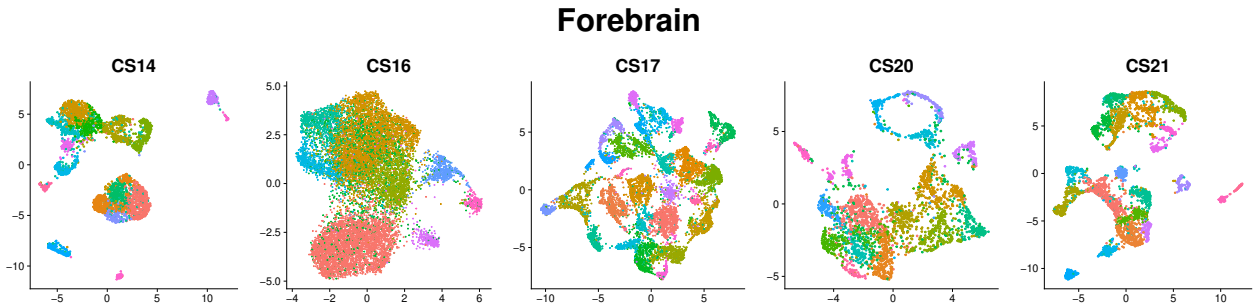


Figure 5.5: UMAPs of gene activity in the forebrain for separate Carnegie Stages. For each Carnegie stage, the UMAP algorithm was run on dimensions 2-30 of the cellular LSI matrix of the forebrain data. The first dimension is typically highly correlated with sequencing depth and is therefore excluded from the analysis. KNN clustering is then performed on these LSI dimensions. Gene activity is then visualsed on the first two UMAP dimensions, with identified clusters annotated by colour.

see in the CS17 plots and later that GAD1 activity is largest in clusters where SOX2 is lowest. GAD1 is linked to the presence of GABAergic neurones in the forebrain [199]. SHH is a major signalling molecule in brain development and has been linked with late neuronal differentiation [200]. The lack of SOX2 in clusters where these genes are most active suggests these clusters are differentiated neurones. Cluster structure suggests that most cells in the forebrain data are radial glial cells, in line with previous similar findings [195]. In CS14 we find a large cluster with activated ALX1. This suggests there are large numbers forebrain mesenchyme, which has previously been observed to be present in large numbers up till CS15, at which point their numbers begin to rapidly decline. Another cluster not present in later stages is MICAL2, an endothelial gene [201]. Activity of this gene could be linked to the presence of the blood-brain barrier, congruent with previous observations in another study [201].

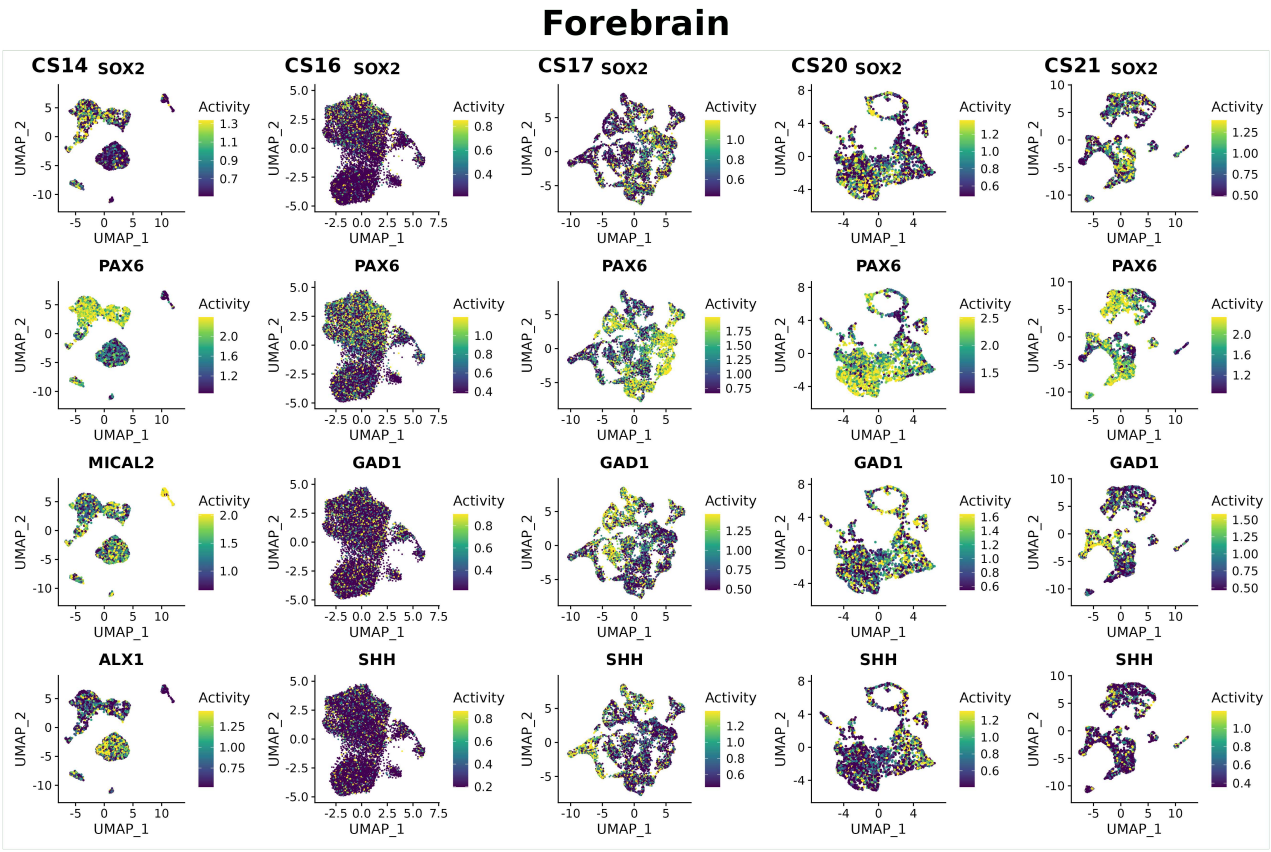


Figure 5.6: Gene activities of established genes in forebrain brain development. Different markers were selected at different phases of neurodevelopmen

5.6.2 The Cerebellum And The Medulla

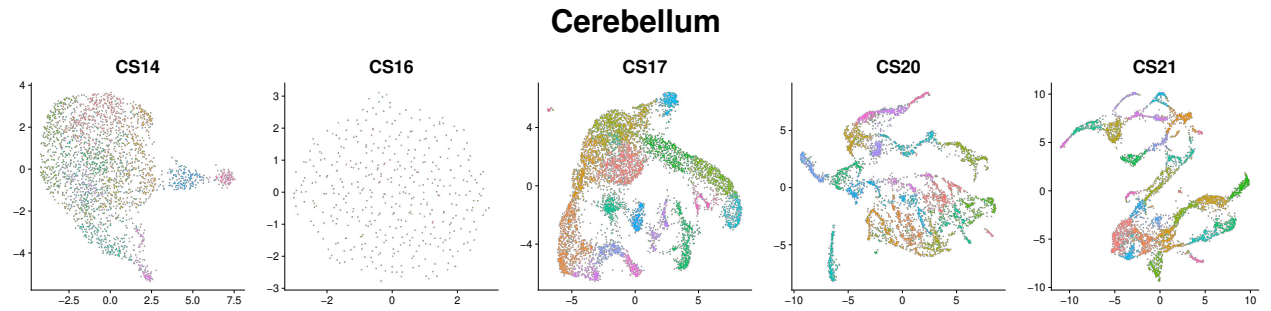


Figure 5.7: UMAPs of gene activity in the cerebellum for separate Carnegie Stages. For each Carnegie stage, the UMAP algorithm was run on dimensions 2-30 of the cellular LSI matrix of the cerebellum data. The first dimension is typically highly correlated with sequencing depth and is therefore excluded from the analysis. KNN clustering is then performed on these LSI dimensions. Gene activity is then visualsed on the first two UMAP dimensions, with identified clusters annotated by colour.

Firstly, the CS16 data from the cerebellum has a small cell count and a defined cluster structure is not well observed. It is included in the figure for completeness. CS14-CS21 are early stages

in development [193]. At these early time points, the predominant cell types can be divided into *GABAergic* and *GLUTAmergic* cell types [202, 203]. Paired box (PAX) genes are well established in patterning genes in the cerebellum [204]. In particular, PAX2 and PAX5 have long been established in their role as defining this patterning and classifying early precursors as GABAergic [205, 206]. Expression of PAX6, by contrast, is associated with GLUTAmergic precursors [207].

The cerebellum and the brain stem of physically near each other in the brain. The brain stem contains the medulla/pons, the third brain region sequenced in this analysis. In early development, the medulla and the cerebellum all form part of the hindbrain. A similar role for PAX2/PAX5 as GABAergic and PAX6 and GLUTAmergic neurogenisis has also been detailed in the literature around the development of the spinal cord and brain stem [208, 209, 210]. As such,we analyse these paired box genes in the cerebellum and medulla data in the same way in Figs. 5.8, 5.10. The activity of PAX6 occurs most strongly in regions where PAX2 and PAX5 activity is lowest. This is difficult to see in the data from early development at CS14, however, the distinction becomes more pronounced at CS21 in both cerebellar and medulla data. This image is consonant with differentiation progressing through development and is manifest in the gene activity of established genes.

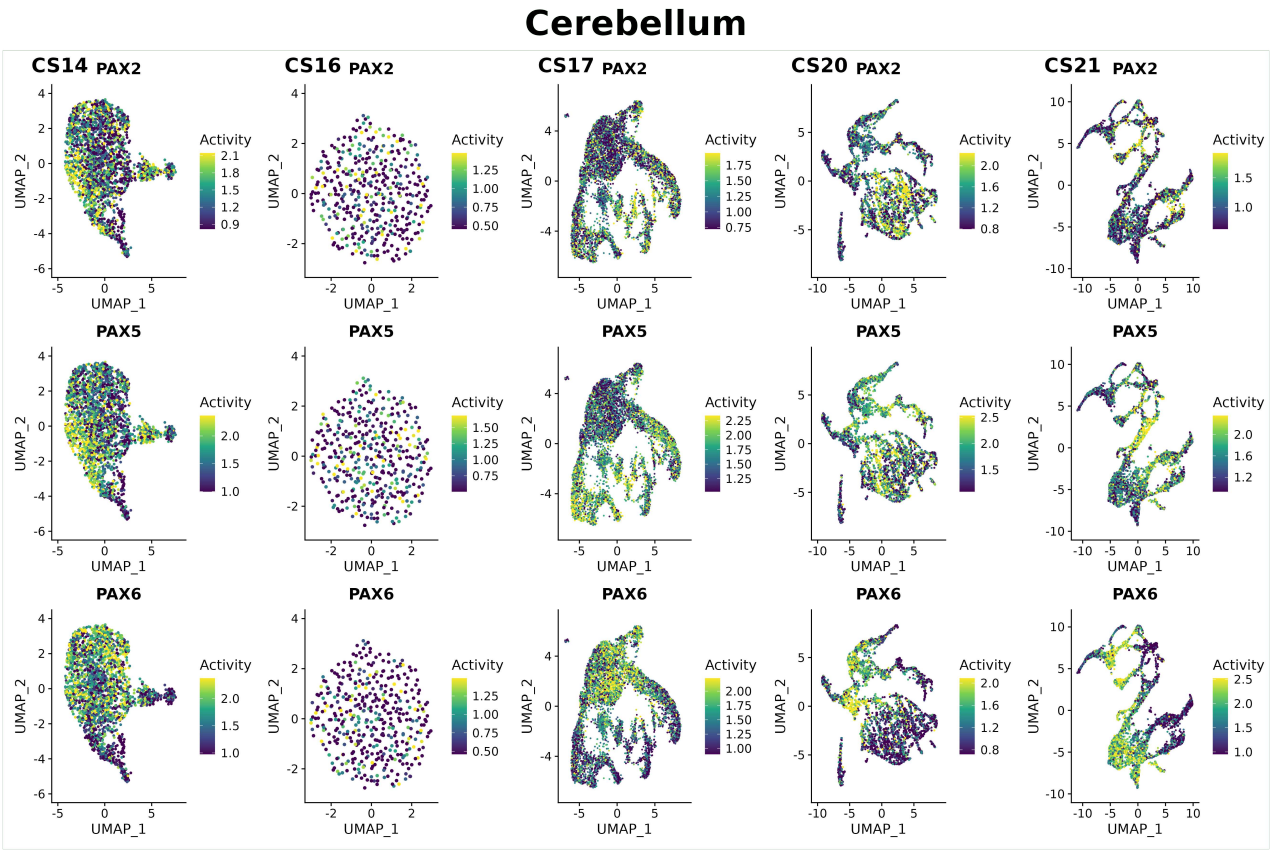


Figure 5.8: UMAPs of gene activity in the medulla for separate Carnegie Stages.

5.6.3 The Medulla

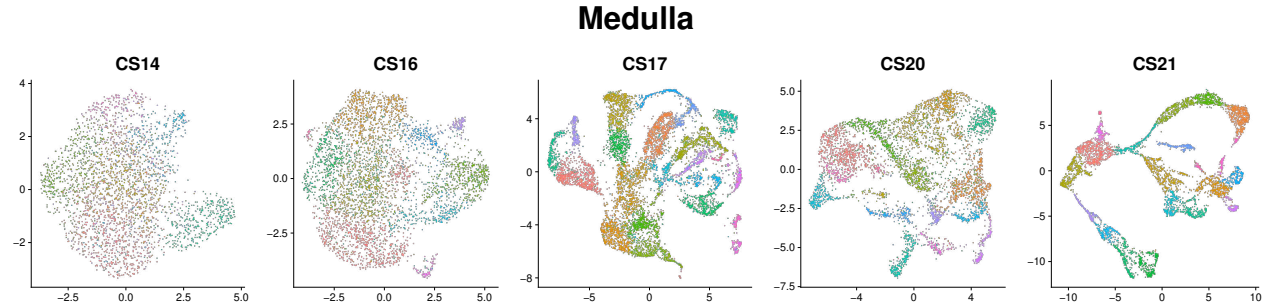


Figure 5.9: UMAPs of gene activity in the medulla for separate Carnegie Stages. For each Carnegie stage, the UMAP algorithm was run on dimensions 2-30 of the cellular LSI matrix of the medulla data. The first dimension is typically highly correlated with sequencing depth and is therefore excluded from the analysis. KNN clustering is then performed on these LSI dimensions. Gene activity is then visualsed on the first two UMAP dimensions, with identified clusters annotated by colour.

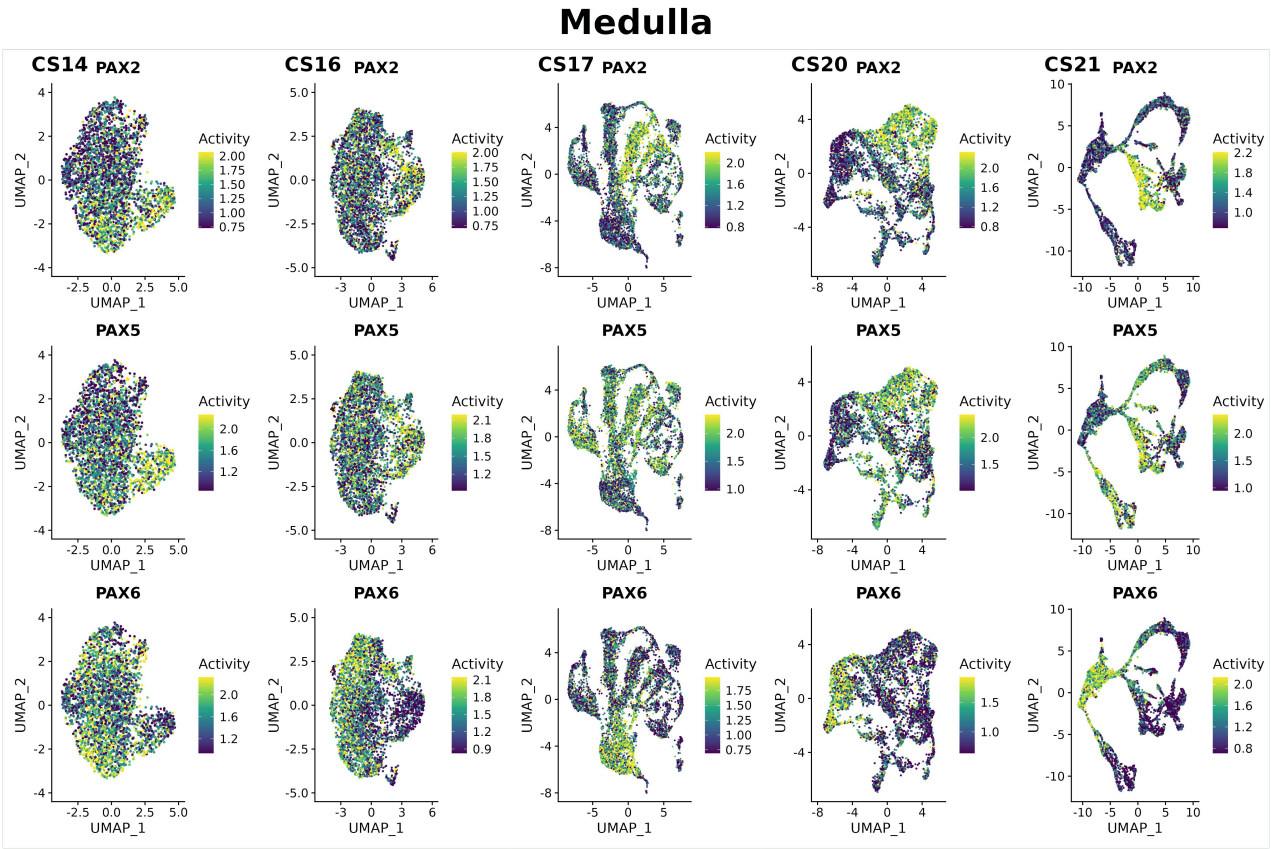


Figure 5.10: UMAPs of gene activity in the medulla for separate Carnegie Stages.

5.7 The Landscape Of Developmental Mutations

We now attempt to map the landscape of mutations that can be identified in early development. In this section, we contend with the difficulty posed by low-depth sequencing data which arose due to frozen tissue samples. First is the task of identifying the set of mutations from the data.

5.7.1 Shallow Depth In Frozen Samples

Figure 5.11 displays the limited depth of coverage visually, showing the average depth per cell at each position of the mitochondrial genome. The depth of coverage varies substantially both between samples and across the genome. In most single-cell datasets we have explored, variants to the reference genome are observed at almost every position of the genome in at least one cell on at least one read. Much of this is explained by sequencing or PCR error. In Chapter 3, we used robust, high-depth thresholds, at which an error was unlikely to occur so many reads at the same position by chance. In the regime of extremely limited depth, simple heuristic depth of coverage thresholds become less relevant as the required depth is seldom obtained.

Such limitations call for a different approach. The first approach is to aggregate data across cells, to gain understanding of mtDNA variation. We give a sketch of mutability across the mitochondrial genome at a pseudo bulk level.

5.7.2 Mutatbility Of Mitochondrial Genome

A simple approach to explore the landscape of the mitochondrial genome in early development is to aggregate data across all cells within a tissue to allow for a pseudo-bulk analysis of the data. We summed the total number of bases mismatching the reference, dividing by the total number of sequenced bases, for each gene to give the mutation rate for each gene. This was done separately for each sample, with the median mutation rate per base across samples displayed in Figure. 5.12. As sequencing error would be more or less uniform throughout the length of the genome, differences in the observed numbers of mutated reads reflects underlying biological sequence variation. The number of total bases sequenced for a particular locus ranged from

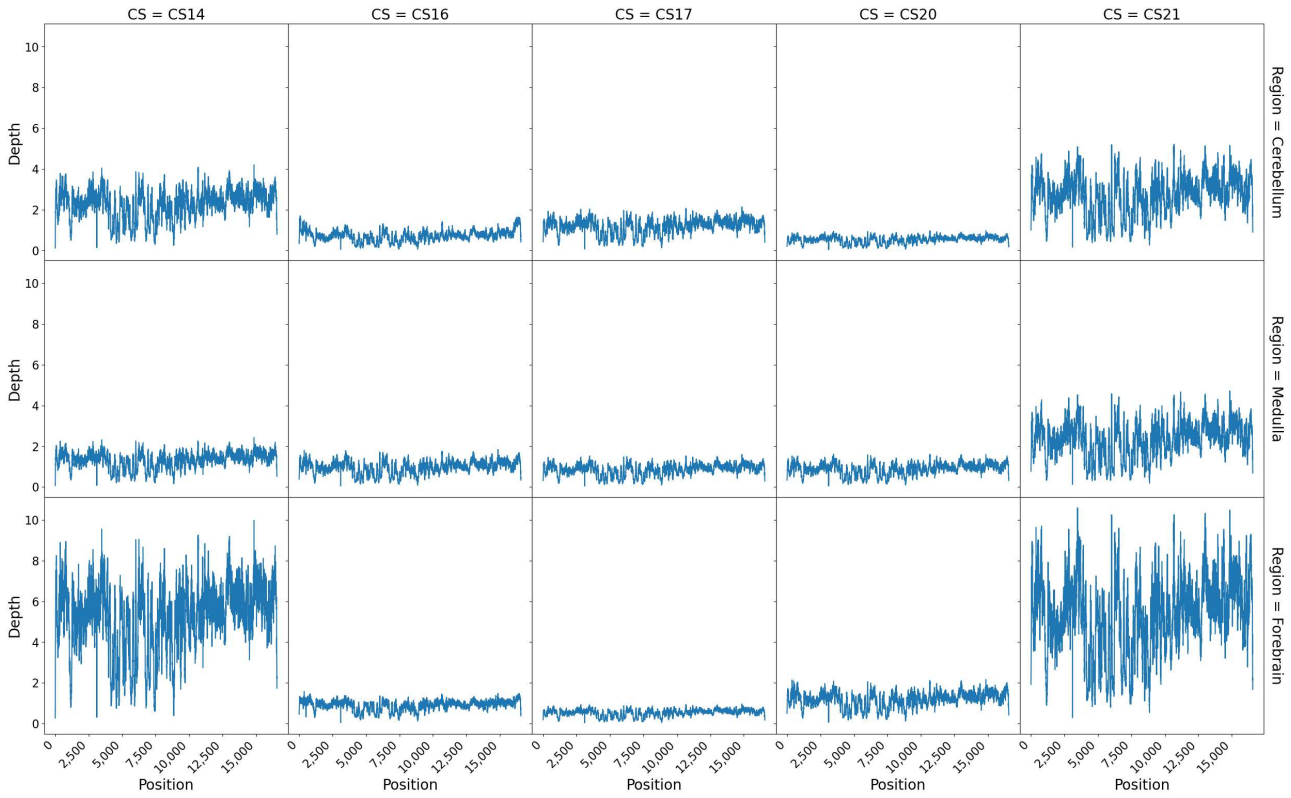


Figure 5.11: Mean depth of coverage per cell across the mitochondrial genome. The degree of variation in the depth of coverage both between datasets and across positions within a dataset is highly variable. Moreover, the typical depth observed is typically below ten reads highlighting the need for statistical methods which are capable of modeling this variation.

$39,496$ in MT-NC4 to $265,299,170$ in MT-ND5.

As expected, the D-Loop, a non-coding region, has the highest mutability, consonant with previous observations [211]. Across all of our datasets, we were able to find a consistent pattern mutations, with the most likely position for a mutation to occur being in the D-Loop.

We separated CDS mutations into missense and synonymous mutations. The median mutation rate of synonymous mutations was significantly greater than non-synonymous mutations with rates 0.0024 and 0.0014 respectively ($P = 0.016$ in two-sided Mann-Whitney U rank test). This is indicative of a selection pressure against potentially damaging non-synonymous mutations in the developing human brain, whether they be germline or early developmental somatic mutations.

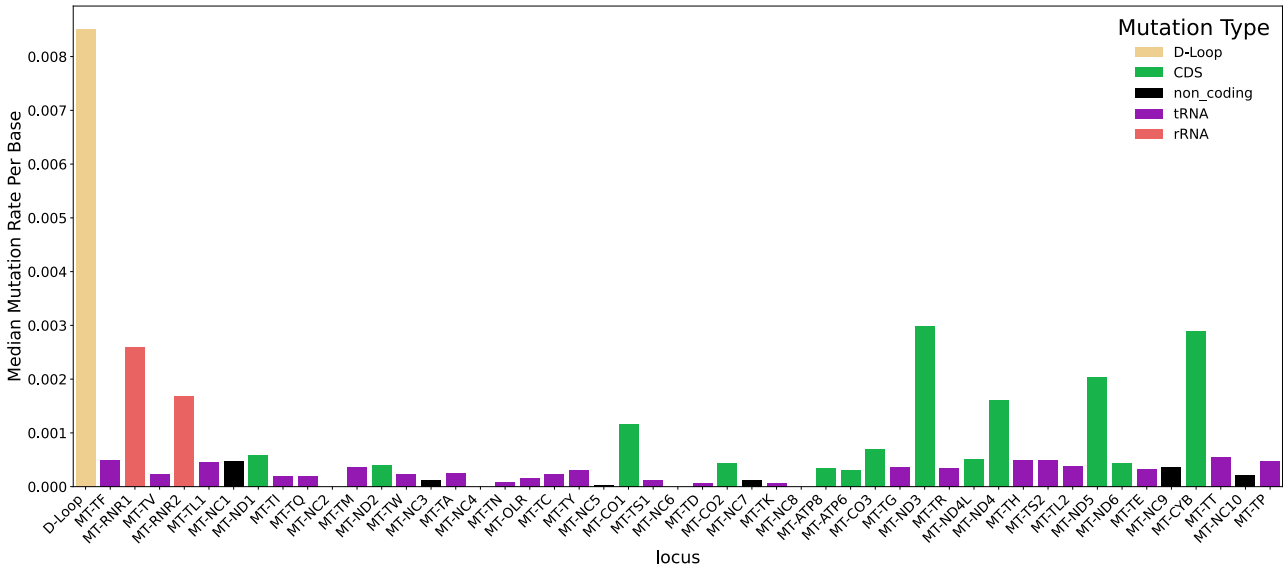


Figure 5.12: Mutability across different mtDNA genomic loci is highly variable when expressed as the proportion of all sequenced bases mismatching the reference genome, quantified as the *mutation rate per base*. Here, we display the median mutation rate for each locus across all sequenced samples. As expected, the non-coding D-Loop region is the most mutagenic region of the genome, in line with previous research [211]. This is followed by rRNA mutations, with CDS mutations also present in the data at a high rate.

5.7.3 Identifying Candidate Mutations For Further Analysis

Due to the severely limited depth available within these samples, identifying genuine mtDNA mutations from sequencing errors is not trivial. In most of the single-cell datasets we've investigated, mismatches to the reference genome are observed across the entire length of the genome. Sequencing up to $\approx$ 16kb across the mitochondrial genome, 1-100s times per cell, in 1,000s of cells inevitably raises multitudes of mismatches, even with low error rates. In previous chapters, we addressed this through the use of conservative depth thresholds, which cannot be applied here. Here, we seek to identify widespread mutations present in many cells by posing a null model of sequencing error, and control the false discovery rate of the mutations we accept. Typical sequencer usage, compared with idealized settings, is known to exhibit variation in the sequencing error rate, with some Illumina sequencers having been estimated to exceed 2%. With this in mind, we conservatively place the sequencing error rate at 2.5% [212]. Reducing the rate could potentially uncover more mutations at the expense of more errors entering the analysis.

We make the following definitions.

$P_{i,j}$:= Probability cell i, has at least one error at position j

$DF_{i,j}$:= Depth in cell i at Position j

$N :=$ Cells

$\epsilon :=$ error rate per sequenced base (approx 2.5%)

Putting this together, we see that the probability of having an error at site j in cell i is:

$$P_{i,j} = 1 - (1 - \epsilon)^{DF_{i,j}}$$

$$X_{i,j} := \begin{cases} 1 & \text{if cell } i \text{ has reference mismatch at position j} \\ 0 & \text{otherwise ,} \end{cases}$$

$X_j :=$ Total Number of cells with an error at position j

$$X_j = \sum_i X_{i,j}$$

By the central limit theorem for the summation of many iid random variables, X_i will be approximately normally distributed, with following moments:

$$
\begin{aligned}
E[X_j] &= \sum_i E[X_{i,j}] = \sum_i P_{i,j} \\
\text{Var}[X_j] &= E[X_j^2] - E[X_j]^2 \\
E[X_j^2] &= E[(\sum_i X_{i,j})^2] = E[\sum_i X_{i,j}^2] + E[(2\sum_{i<k} X_{i,j}X_{k,j})] = \sum_i P_{i,j} + 2\sum_{i<k} P_{i,j}P_{k,j} \\
E[X_j]^2 &= (\sum_i P_{i,j})^2 = \sum_i P_{i,j}^2 + 2\sum_{i<k} P_{i,j}P_{j,K} \\
\text{Var}[X_j] &= E[X_j^2] - E[X_j]^2 = \sum_i P_{i,j} + 2\sum_{i<k} P_{i,j}P_{k,j} - \sum_i P_{i,j}^2 + 2\sum_{i<k} P_{i,j}P_{j,K} \\
&= \sum_i P_{i,j} - P_{i,j}^2 \\
\text{Var}[X_j] &= \sum_i P_{i,j}(1 - P_{i,j})
\end{aligned}
$$

By matching the moments of X_j to those of the normal distribution, we obtain that:

$$X_j | DF_{i,j} \sim Normal(\sum_i^n P_{i,j}, \sum_i^n P_{i,j}(1 - P_{i,j}))$$

This can be considered from an alternative perspective. If we are in the large n, low error regime, all $P_{i,j}$ will have small values then we're in the *Poisson paradigm*. The Poisson paradigm applies when a random variable is the sum of the total number of rare Bernoulli events, each with a small probability of success. The resulting sum has a Poisson distribution in which the rate parameter is the sum of the individual Bernoulli successes. Poisson random variables have equal mean and variance. The correspondence between our normal approximation above and the Poisson paradigm can be seen in the limit of all $P_{i,j} \to 0$. For the normal result above, the variance term, $\sum_i^n P_{i,j}(1 - P_{i,j}) \to \sum_i^n P_{i,j}$, identical to the mean term of the normal distribution stated above. This shows the correspondence between the normal approximation and the Poisson approximation in this regime, as both distributions have identical mean and variance (see Ross, page 68 [164]).

We can state our schema for selecting likely mutations in the following manner. Given the read depth in all cells at a given position, the simple model stated above serves as a reasonable

null model of the distribution of the number of cells harboring a mutation, given that it was produced by sequencing error.

First, we can compute the number of cells we observe each mutation in. With the per-base error rate and the depth of coverage in each cell at a given position for a given dataset, we can compute the distribution of the number of cells a sequencing error would be present in. Using this as our null, we can compute a p-value as the quantile of this distribution for the observed mutation cell count. We can repeat this for every position on the mitochondrial genome, computing a p-value for each.

Then, we apply the standard Benjamini-Hochberg procedure to these computed p-values, ensuring the False discovery rate stays below 5% [26]. This results in a reduced list of mutations which produce a pattern of data that would be unlikely under the presumption of sequencing error.

5.7.4 Properties Of Identified Mutations

We performed this analysis separately for each brain region within each sample, giving a list of high-confidence mismatches from the genome that were deemed unlikely to be caused by sequencing errors. This analysis identified 93 different mutations across all 15 datasets. Due to the way in which high confidence mutations were identified, the 93 mutations are all highly prevalent in the sense of being present in numerous cells in the data. Mutations present in this many cells at such an early point in development will either have been inherited through the germline, or have occurred very early in the developmental process. The constrained depth of our data means we are unable to explore the landscape of cryptic mutations present in early development. Furthermore, simulations in Chapter 3 showed that any cryptic mutations present during this early stage of life have a negligible chance of reaching a pathological threshold and are not of concern in what follows.

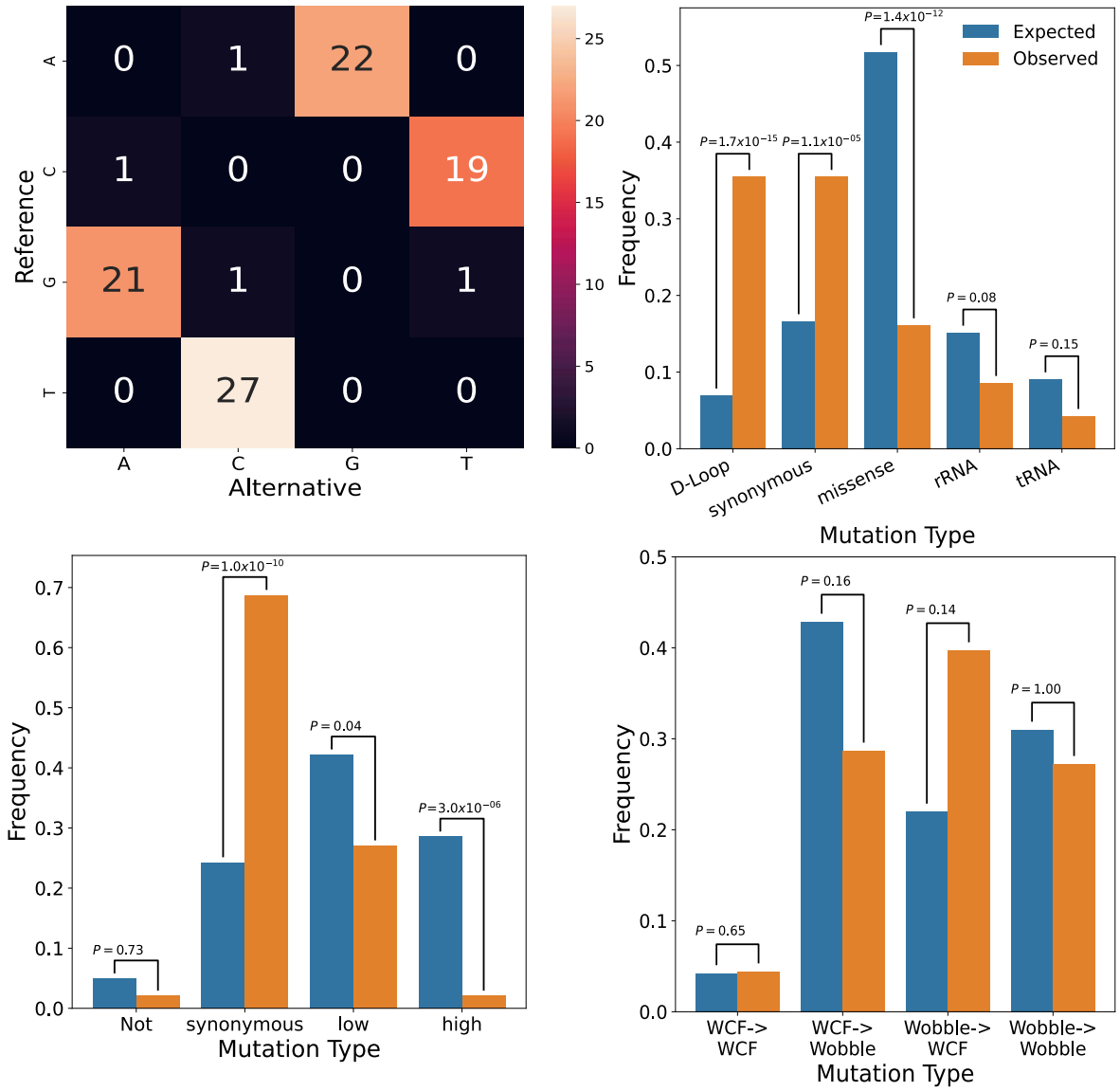


Figure 5.13: We display the profile of the 93 mutations that we were able to confidently identify from the reference genome. (Top left) The base mismatch from the reference genome to the observed alternative. This profile has the characteristic of *transversion* mutations, characteristic of mtDNA mutations previously observed in the literature [32]. (Top right) The frequency of mutations types had they occurred randomly across the mitochondrial genome, is compared to the observed set of mutations. The observed frequency of the 93 mutations shows "statistically significantly" more D-Loop and synonymous mutations were observed than expected using a two-sided binomial test statistic. Missense mutations, those that would be most likely to occur by chance, are "significantly" under-represented in our data. This indicates that synonymous mutations and those in the D-Loop are not selected against, consonant with expectations, whilst missense mutations are aggressively selected against in early development. (Bottom left) CDS mutations were annotated by MutPred to give an indication of whether they were low, high or not pathological based on predicted impact of protein function [213]. We can see that low pathogenicity mutations were under-represented and highly pathogenic mutations were strongly and "statistically significantly" under-represented (P-values computed using two sided binomial test). (Bottom right) Shows the observed vs expected frequency of sub-types of synonymous mtDNA mutations. Of the identified mutations, 33 were synonymous, limiting the statistical power to identify "significant" relationships using a two-sided binomial test statistic. Nevertheless, the observed pattern is in-line with expectations that WCF ->Wobble mutations would be under-represented, due to their diminished codon affinity in translation and Wobble ->WCF over-represented due to enhanced binding affinity, line with previous results [214].

A recurring observation within the mitochondrial literature is that most mitochondrial mutations arise from DNA copying errors. Such mutations have a characteristic profile in that they are overwhelmingly likely to be DNA transitions, that is, A<->G and C<->T mutations [32]. This is precisely the pattern observed within our own reduced list of mutations (see Figure. 5.13, top left). This lends credence to our ability to filter out otherwise random errors that occur due to technical artifacts by showing coherence with established literature about the nature of mtDNA mutations

We made comparisons between the frequency of observed characteristics of mtDNA mutations and those that would be expected by chance if they occurred randomly along the length of the mtDNA. We generated all possible mtDNA single nucleotide mutations by creating all three variants at every position in the rCRS reference genome. We then annotated all of these mutations using MitoMap to identify whether they occurring in the D-Loop, are synonymous, missense, rRNA or tRNA mutations [215]. With this we computed the normalised frequency of expected mutations that could occur across the genome. Figure 5.13 (top right) shows the results of comparisons using an exact binomial test, with the reference genome frequencies serving as a null distribution to compare against. We observe "statistically significant" over-representation of D-Loop and synonymous mutations, with strong "statistically significant" under-representation of missense mutations. This is in line with expectations that D-Loop and synonymous mutations have little effect on transcription products, whilst missense mutations can result in pathological effects on the proteins they code for.

We further explored the properties of CDS mutations observed in our data. We annotated missense mutations using `MutPred2`, which gives pathogenicity scores based on the impact of variants on protein structure [213]. For all possible mtDNA mutations we identified whether they were synonymous or "low", "high", or "not" pathological according to `MutPred2`, yielding the null frequency of subcategories of CDS mutations. Again we compared this to the observed distributions in our data, displayed in figure 5.13. Synonymous mutations were seen to be strongly and significantly represented with low and highly pathological mutations displaying a strong under-representation.

Finally, we further explored the properties of the synonymous mutations present in our data. Synonymous mutations occur due to the degeneracy of the genetic code, in which multiple triplet codons correspond to the same amino acid. For instance, GUU, GUC, GUA, and GUG are all RNA codons that code for Valine, with only the third position changing. The third position of the codon is the most common position of degeneracy for different amino acids. This pattern is described by the *wobble rules*, in which the first two codons of the RNA pair match precisely according to the standard AU/GC rules, whilst the third position does not need to form such precise hydrogen bonds, tolerating a degree of mismatch. This enables some tRNAs to match multiple codons, and some codons to match multiple tRNAs. For instance, the tRNA anticodon, 3–AAG–5, binds perfectly with 5–UUC–3, through canonical Watson-Crick-Franklin (WCF) binding, but is still capable of binding with 5–UUU–3 through wobble dependent translation [216].

This leads to four subcategories of synonymous mutations: WCF ->WCF; WCF ->Wobble; Wobble ->WCF; and Wobble ->Wobble. Though all synonymous mutations do not alter the amino acid content of the proteins *transcribed*, the shift in the *translation* mechanism in some synonymous mutations could still be detrimental. Previous results have indicated that the discrepancies in codon affinity in synonymous mtDNA mutations perturb standard cellular function [214]. We sought to understand whether the mutations we identified in the early developing human brain were consonant with previous findings. We identified all 8,248 synonymous mtDNA mutations and determined which of the four types they were, establishing the expected frequency of these mutations. Of the 93 high confidence variants we identified in our data, only 33 were synonymous, limiting the power to find "statistically significant" results. Nevertheless, when compared to expectation by chance, we saw an under-representation of WCF ->Wobble and an over-representation of Wobble ->WCF synonymous mutations. This is in line with expectations that diminished codon binding is selected against whilst enhanced codon binding is selected for. The data are suggestive that widespread synonymous mutations in the early brain are subject to selection pressures. Whether this selection occurs in the germline or early development itself is unclear.

We display the features of these putative mutations in figure 5.14. Each heatmap in the figure

follows an identical structure, with the row being a mutation, and columns signifying which Carnegie stage and brain region, as indicated by the coloured headers. Going from left to right, the first heatmap displays % cells the mutation was observed in that it could have been observed in. We performed hierarchical agglomerative clustering on the rows. All other heatmaps retain the same row order. The second heatmap shows the mean heteroplasmy of the observed cells, computed using the naive heteroplasmy estimate as $\frac{NM_i}{D_i}$. The third heatmap displays the % of cells observed to be heteroplasmic, meaning neither 0% nor %100 heteroplasmy.

Mutations were almost exclusively identified independently in all three brain regions within each embryo indicating these mutations are either inherited through the germline or occurred very early in development.

Based on these metrics, we further filtered the set of mutations carried forwards for further analysis, displayed in the final heatmap. These mutations had to have a mean heteroplasmy of $< 90\%$, be observed in $> 30\%$ of cells and have a mean depth > 1 in all three brain regions. obtaining list of 93 unique mutations. These metrics were chosen to select mutations that were present in enough cells to allow the negative binomial regression with heteroplasmy uncertainty from Chapter 4 to be fit with confidence. Additionally, we selected mutations that were likely to display enough variation in their heteroplasmy that they could plausibly be associated with differences in gene activity.

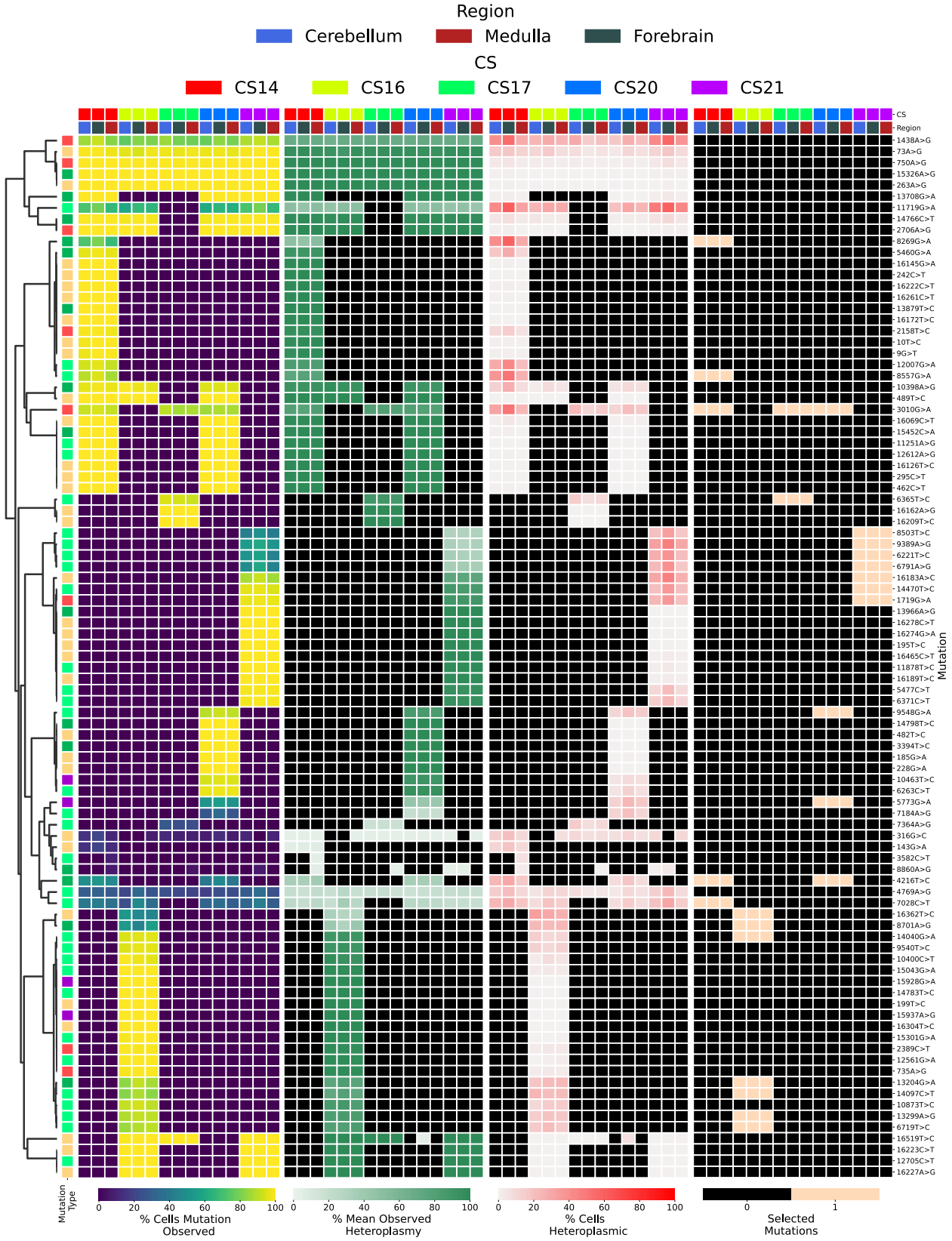

Region
Cerebellum
Medulla
Forebrain
CS
CS14
CS16
CS17
CS20
CS21
CS
Region
1438A>G
73A>G
750A>G
15326A>G
263A>G
13708G>A
11719G>A
14766C>T
2706A>G
8269G>A
5460G>A
16145G>A
242C>T
16222C>T
16261C>T
13879T>C
16172T>C
2158T>C
10T>C
9G>T
12007G>A
8557G>A
10398A>G
489T>C
3010G>A
16069C>T
15452C>A
11251A>G
12612A>G
16126T>C
295C>T
462C>T
6365T>C
16162A>G
16209T>C
8503T>C
9389A>G
6221T>C
6791A>G
16183A>C
14470T>C
1719G>A
13966A>G
16278C>T
16274G>A
195T>C
16465C>T
11878T>C
16189T>C
5477C>T
6371C>T
9548G>A
14798T>C
482T>C
3394T>C
185G>A
228G>A
10463T>C
6263C>T
5773G>A
7184A>G
7364A>G
316G>C
143G>A
3582C>T
8860A>G
4216T>C
4769A>G
7028C>T
16362T>C
8701A>G
14040G>A
9540T>C
10400C>T
15043G>A
15928G>A
14783T>C
199T>C
15937A>G
16304T>C
15301G>A
2389C>T
12561G>A
735A>G
13204G>A
14097C>T
10873T>C
13299A>G
6719T>C
16519T>C
16223C>T
12705C>T
16227A>G
Mutation
Mutation Type
0 20 40 60 80 100
% Cells Mutation Observed
0 20 40 60 80 100
% Mean Observed Heteroplasmy
0 20 40 60 80 100
% Cells Heteroplasmic
0 1
Selected Mutations

Figure 5.14: We compared the observed number of cells each mutations was observed in to the null error model for each data sample separately, controlling the FDR at 5%. we visualise the results in four heatmaps displaying different features. Each heatmap is structured such that the columns follow a hierarchical structure, with the first level being the Carnegie stage of the sample, and the second being the brain region. Each column therefore represents one of the 15 samples in our data. Going from left to right, the first heatmap displays % cells the mutation was observed in that it could have been observed it. We performed hierarchical agglomerative clustering on the rows. All other heatmaps retain the same row order. The first thing to note is that mutations were almost exclusively identified independently in all three brain regions within each embryo indicating these mutations are either inherited through the germline or occurred very early in development. The second heatmap shows the mean heteroplasmy of the observed cells, computed using the naive heteroplasmy estimate as $\frac{NM_i}{D_i}$. The third heatmap displays the % of cells observed to be heteroplasmic, meaning neither 0% nor %100 heteroplasmy. Based on these metrics, we further filtered the set of mutations carried forwards for further analysis, displayed in the final heatmap. These mutations had to have a mean heteroplasmy of $< 90\%$, be observed in $> 30\%$ of cells and have a mean depth > 1 in all three brain regions.

5.8 Association Between Heteroplasmy And Gene Activity

5.8.1 Negative Binomial Regression Including Heteroplasmy Uncertainty

To measure the functional impacts of heteroplasmy on gene expression, we used negative binomial regression with heteroplasmy as an uncertain predictor variable. This methodology is outlined in detail in Chapter 4, were the model was validated *in silico*. Before turning our attention to the results of this application, we explore the software engineering challenges inherent in applying such a model *millions* of times.

5.8.2 Bayesian Inference At A Massive Scale: Software Engineering Challanges

The model was implemented in `STAN`, with posterior samples generated using the Hamiltonian Monte Carlo. Markov Chain Monte Carlo methods are computationally expensive. Hamiltonian Monte Carlo yields sampling efficiency improvements by exploiting the structure of the logarithm of the posterior. Nevertheless, even HMC is a computationally demanding algorithm to run, with sampling efficiency depending on the geometry of the posterior. In particular, we used `CmdStan` (version 2.30.1) through the `CmdStanPy` (version 1.0.8) interface (neither of which should be confused with `StanPy`). The typical time to obtain 1,000 samples from 4 different chains takes around 80-90 seconds on one of our University's AMD EPYC 7742 64-Core Processors. This is using the default settings of `CmdStanPy` which runs each chain in a separate process, allocating each a CPU core, without any within or cross chain multithreading.

As a rough order of magnitude calculation, consider 20,000 genes $\times$ 90 seconds means each sample would take approximately 500 (hrs) of continuous compute time to apply the model to every mutation gene pair, within just a single sample. When considering reapplication to 3 samples, for 10s of mutations, total compute times are sensibly counted in years. Obviously, a naive serial approach to such computations is untenable if we are to make any progress in measuring the associations for multiple samples.

Exploiting *concurrency*, by use of multithreading, allowing multiple tasks to be progressed at once is one approach, though the virtues of this approach are most obvious for IO bound problems. The current problem however is CPU bound. *Parallelisation* is, therefore, the obvious solution, but the correct approach to such parallelisation is not obvious. Our university compute cluster offers many CPU cores, but running thousands of separate compute cores rapidly blocks wait queues on schedulers of shared compute resources, with more time spent waiting to run code rather than executing, another bottleneck. We therefore opted to request large jobs on our HPC server with 200 CPU cores per job.

By leveraging Python's built-in `multiprocessing` module, we could exploit process-based parallelism to run approximately 40 models in parallel, on a single large compute node. Each model would run in parallel, with each of the 4 HMC sampling chains running across cores, allowing up to 160 HMC chains to sampled in parallel. Such an approach is however extraordinarily memory intensive as separate processes do not have access to the same shared memory spaces, with each process having its own separate memory space. When a variable is passed to a subprocess, a new copy of the underlying object is created, ensuring that the two processes do not conflict. In our application, multiple models attempting to access the same data table representing heteroplasmy and gene activity would result in upto 40 duplicates living in memory. Moreover, each sampling chain running in parallel once again duplicates the memory used. This results in extraordinary memory-hungry jobs, requiring upto 300Gb in RAM to prevent compute jobs from being canceled on the high-performance computer. In spite of this, such an approach was able to get the approximately 3-weeks of continuous compute time required for a single set of regressions for a single mtDNA variant against the set of all genes, down to a 3-4 days. By drawing upon multiple compute nodes with 200 CPUs, we were further able to parallelise the process of fitting models. At peak usage, we were able to get around 10 clusters running at this level on our university HPC. With 40 models, each 4 chains on 10 compute nodes allows for up to 1,600 HMC chains to be running simultaneously. By using `snakemake`, we were able to coordinate all jobs to be run through the high-performance computer, with models for each mutation in each sample being run separately, in parallel. This parallelisation approach allowed years of total computing time for the entire set of selected mutations to take

place over weeks.

5.8.3 Posterior Predictive Checks

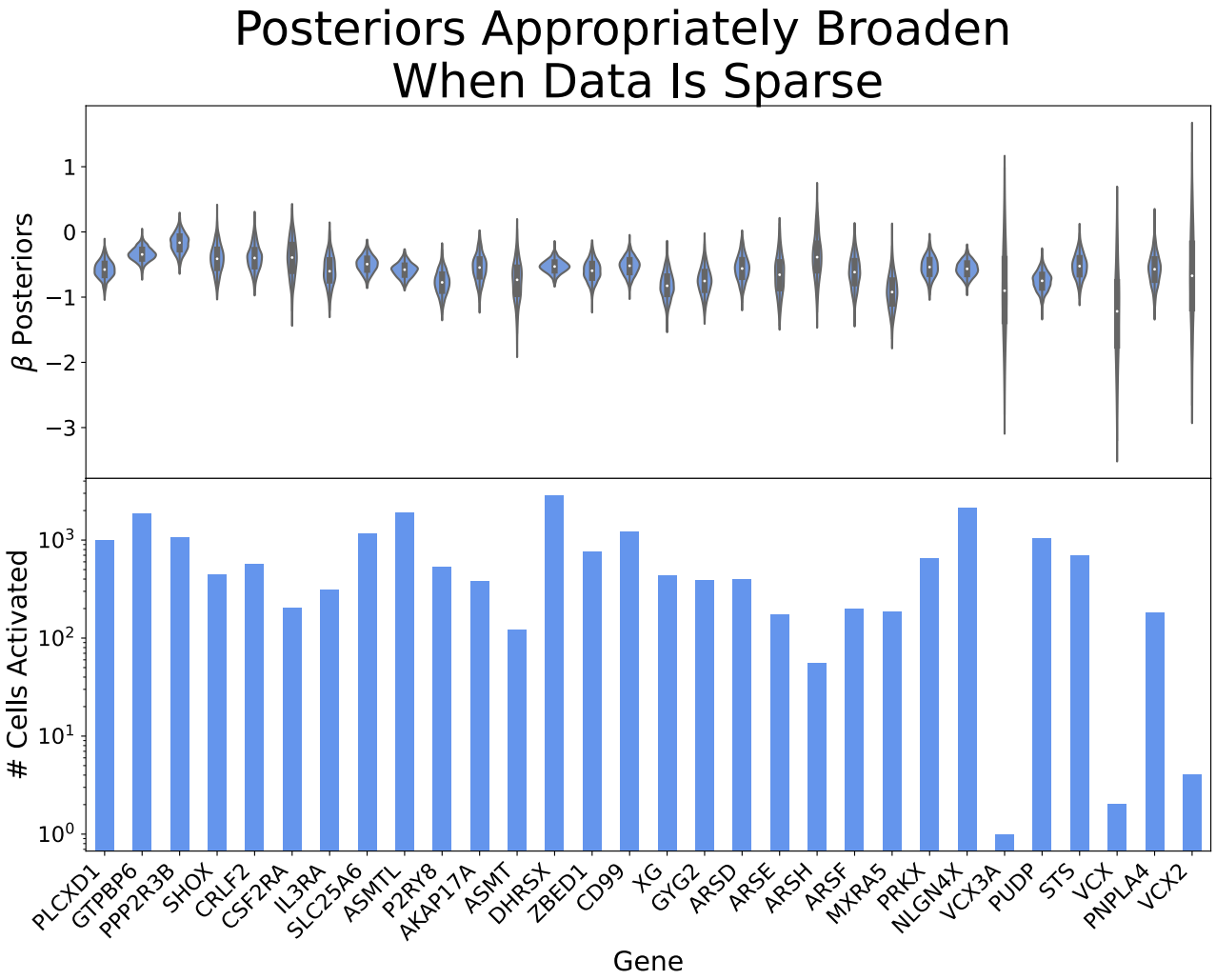


Figure 5.15: Posterior regression coefficients, β, displayed as violin plots for a sample of 30 randomly selected genes from the CS21 forebrain data. In particular, these are the regression coefficients between 9389A>G and each gene's activity. The posteriors exhibit a range of widths with many tightly centered and others displaying a huge amount of uncertainty.In the bottom panel, we display the number of cells in which a gene was active, defined as being the number of cells with 1 or more counts for that particular gene. The relationship between the number of active cells and posterior width displays a clear pattern across the range of genes. Regression coefficients for VCX3A, VCX, and VCX2 display abnormally large widths, which is explained by each of them having only 10 cells which are active. This is indicative that the model performs reasonably, being more uncertain with less data on which to base its inferences.

Before proceeding to give widespread analysis of posterior results across all samples, we give an concrete illustration of the model on a small subset of data. In particular we focus on regression performed between the heteroplasmy of 9389A->G in the forebrain data of the CS21

sample. This is to demonstrate the general characteristics and behaviour of the model, as well as demonstrating that the model is capable of fitting selected data appropriately.

In the top row of Figure. 5.15 we display the posterior for the β coefficient for 30 genes that we fit using HMC. In the bottom row of the figure we display the number of cells which had that gene activated at least once. Most genes have 100s or 1,000s of active cells. There is a clear relationship between the number of cells which are active in the data, and the uncertainty of the regression coefficient. There are three genes in which only a few cells appear to be active, namely, VCX3A, VCX, and VCX2. It can be seen that the posterior distributions for these regression coefficients are abnormally broad, reflecting the fact that these models are indeed uncertain due to the limited data. This exemplifies another upside of this approach as the model honestly reports cases where it is unsure about the true nature of the relationship between heteroplasmy and a gene's activity.

No divergent transitions were present for the HMC. We display the trace plots of MCMC chains of the parameters of interest in figure 5.16. All chains mix and appear to exhibit stationarity, implying the MCMC chains are not stuck in local optima and have converged. Correspondingly, all $\hat{R}$ values were very near 1. This is in line with the results in our simulated setting in the previous chapter.

An important component of checking the validity of Bayesian models is performing *posterior predictive checks* (PPCs). The concept of posterior predictive checking is that if the model fits, then data generated from posterior samples should resemble the observed data that the model was fitted to, or be plausible under the posterior predictive distribution [1]. Deviations from the observed data and the posterior predictive distribution would indicate potential limitations of the model that could be used in turn to develop the model further [153, 154].

Figure 5.17 displays the posterior predictive distributions for three of the genes in the figure above. We display the 50%, 90%, and 95% predictive posteriors quantiles. In order to compute this, one takes a single posterior parameter sample, for α, β, and ϕ. We then compute $\mu = exp(\alpha + H\beta)$ with a range of values $H \in [0, 1]$. Samples are then drawn from the negative binomial distribution for each H, using the values of μ_i and ϕ. This process is then repeated

[1] This is the distribution of new data points sample from likelihood, when parameter values are sampled from the posterior distribution, rather than the prior $P(y_{\text{new}}|y) = \int P(y_{\text{new}}|\theta)P(\theta|y)\,d\theta$

over multiple draws from the posterior — we used 500 for the visual. This gives a separate distribution for a range of values of H. Quantiles from this distribution are computed and displayed in figure 5.17.

Conventionally, in a posterior predictive check of a regression, one would plot the raw data points. Here, however, our predictors are themselves latent variables. For the purpose of displaying the data, we took 5 posterior samples of H for each cell and plotted the sample H against the observed gene activity. This overcomes the discretisation artifacts seen in the previous chapter and allows use to check that the model is self-consistent in its posteriors for H and and the relationship between H and gene activity.

For each of these draws of H, we can compute the proportion of samples for which y_i, the associated gene activity, is less that posterior predictive quantiles. Table 5.1 displays the coverage for the three different genes in the figure. We can see that the posterior predictive distribution reasonably captures different realisations of the data, with deviations of a few percent at most.

	90%	**95%**
DHRSX	88.06	93.42
PPP2R3B	84.19	92.56
NLGN4X	88.93	93.88

Table 5.1: The proportion of data samples less than or equal to different posterior quantiles. We can see that the proportion of data points at or less than different values is roughly in agreement, showing the model's ability to describe the data.

The final component of posterior predictive checking that we explore is the behavior of the heteroplasmy posteriors, and how the model interprets them. Figure 5.18 shows how the posterior mean of the heteroplasmy is displaced from the naive heteroplasmy point estimate. The obvious behavior is that cells with low depth do not place meaningful weight onto the posterior positions with their posteriors being broad and the location of the mean freely moving. Once depth exceeds 10 reads, the displacement in heteroplasmy point estimates rapidly attenuates to no more than a few percent. This behavior is both reassuring and expected in that when depth is high, we should expect our inferences of heteroplasmy to be driven primarily by the proportion of reads observed to carry a mutation. Indeed, it would be alarming if when 5 of 25 were observed to be mutated, the model confidently concluded the real heteroplasmy of the

cell was in fact closer to 90%. Situations akin to this are never observed. Furthermore, the stability of these high-depth heteroplasmies suggests the model takes them seriously, leaving them where they are and weights them heavily in the regression. Low-depth heteroplasmy cells on the other hand are flexible and, in some sense, are informed by the combination of the proportion of mutated reads and gene activity. By leveraging the regression coefficient, which is mainly informed by high depth cells, the model can in turn come to infer the heteroplasmy of low depth cells which is consistent with the gene activity observed, with posterior width varying accordingly.

For any particular regression, we get a complement of posterior distributions for the heteroplasmy from each cell. In general, there is no reason to expect the posteriors of individual cells to be the same from regression to regression. High-depth cells would however yield consistent posterior inferences, and be the most informative in each regression. In some sense, these differences can be considered a limitation of our current approach. Further developments could consider additional exploration of how to combine information from regressions across different genes in order to come up with well-informed estimates of heteroplasmy. As mentioned in the previous chapter, we explored this option but ran into computational issues with the model being unable to generate HMC samples when hierarchically pooling regression coefficients across multiple genes at once. Given these limitations, we use the model independently for each regression, focusing on the regression coefficients as measures of association between underlying heteroplasmy and gene activity.

5.8.4 Bayesian Inference At A Massive Scale: Validating Model Behaviour

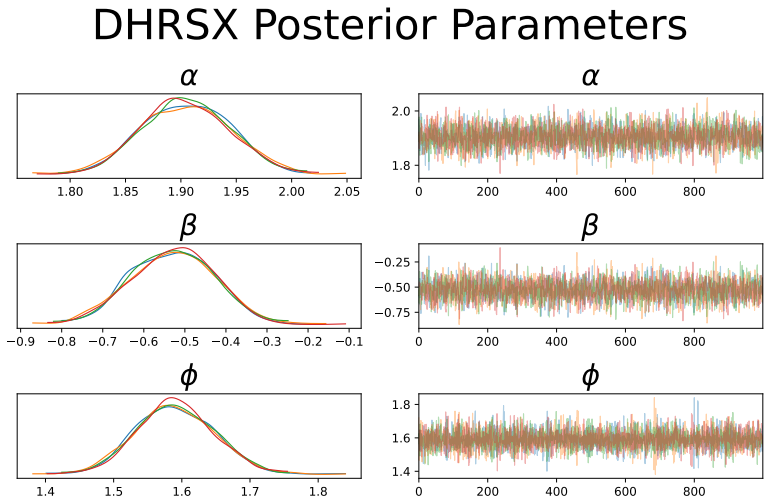
We performed negative binomial regression with uncertain heteroplasmy predictors more than 1.2 million times. This amounts to fitting more than 1.2 million separate Bayesian models, all with 1,000s of parameters, with each parameter having been sampled 1,000 times in 4 separate chains. If the additional thousand warmup steps for each chain is included, this leads to on the order of tens of *trillions* HMC samples being obtained, with half of these (the post warmup samples) being saved to disk. This combines to occupy close to 100Gb of disk space when saved

as pandas dataframes in the `.parquet` file format.

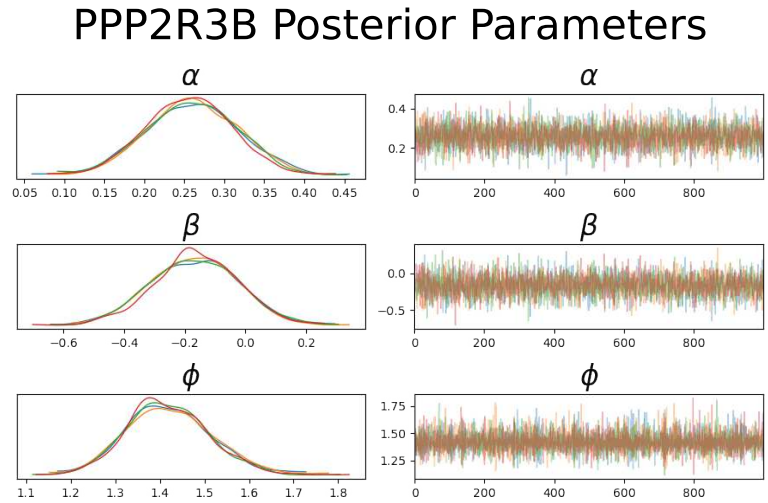
This poses the challenge of processing the outputs of MCMC performed at this massive scale and verifying that MCMC has accurately captured the posterior. The first thing to note is that manual inspection of MCMC chains to observe mixing has taken place is beyond the realm of practicality. We manually inspect these for a few samples as a sanity check as in Fig 5.16. However, we make use of the standard $\hat{R}$ to assess between chain mixing and monitor convergence for each model and further estimate the number of effectively independent simulation draws. Given the scale of the number of parameters and models, we focus on performance bounds rather than manual inspection of each parameter separately. For each model, we took the largest $\hat{R}$ value across all parameters to assess convergence. We placed a practical bound of 1.05 on $\hat{R}$, more stringent than used elsewhere [153]. For a given mutation, typically far fewer than 1% of all regression models possessed any parameters for which $\hat{R}$>1.05, reassuring us about the model's robust performance. We sought to further understand the few cases where the model did exhibit poor convergence in some parameters. Table 5.2 divides the regressions into those performed on *active* and *inactive* genes for all regressions performed in the CS21 sample, across all three brain regions. We define inactivity as having less than 200 cells in the dataset with at least one activity count. It can be seen that nearly 100% of regressions performed on *active* had all of their parameters with $\hat{R}$<1.05. Inactive genes were thus the primary source of models with $\hat{R}$>1.05, which might be due to the lack of data resulting in a posterior distribution that is difficult to explore. Even in the class of the inactive genes, however, we see that more than 99% of models possessed satisfactory convergence in all parameters. Based on this, we exclude genes inactive genes from the analysis.

	% Models with all params $\hat{R}$<1.05					
Region	**Cerebellum**		**Forebrain**		**Medulla**	
MutationID	**Inactive**	**Active**	**Inactive**	**Active**	**Inactive**	**Active**
1719G>A	99.52	100.0	99.62	100.0	99.62	100.00
6221T>C	99.38	100.0	99.47	100.0	99.42	100.00
6791A>G	99.38	100.0	99.47	100.0	99.47	100.00
8503T>C	99.47	100.0	99.71	100.0	99.62	100.00
9389A>G	99.28	100.0	99.52	100.0	99.18	100.00
14470T>C	99.28	100.0	99.57	100.0	99.38	99.99
16183A>C	99.71	100.0	99.76	100.0	99.23	99.99

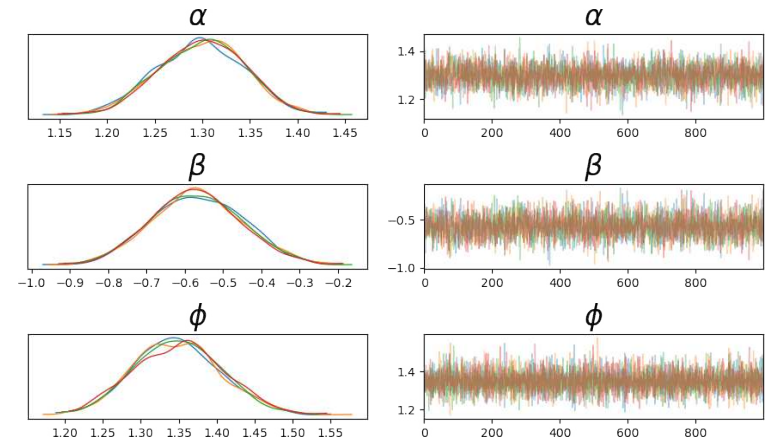
Table 5.2: To every model we measured the degree of convergence using the "potential scale reduction factor", $\hat{R}$, for each parameter. There a thousands of parameters for each model, so we focus on the worst-performing parameter each model had. Some, though very few models, had at least one parameter exceeding the 1.05 cutoff value. Upon further inspection, we discovered that the models that exhibited poor convergence were those regressions for genes that were largely *inactive*. We define inactivity as having less than 200 cells in the dataset with at least one activity count. It can be seen that approximately 100% of regressions performed on *active* had all of their parameters with $\hat{R}$<1.05. Inactive genes were thus the primary source of models with $\hat{R}$>1.05, where there might not have been enough data to produce a posterior that could be effectively sampled by HMC. Nevertheless, even among the set of inactive regressions, more than 99% of models performed as expected.



(a)



(b)



(c)

Figure 5.16: Traceplots for the α, β, and ϕ parameters of the Negative Binomial regression model with uncertain predictors for two sets of real data. It can be seen that all chains converge and mix well, indicative of good sampling. Lack of divergences indicates no obvious problems occurred with the model

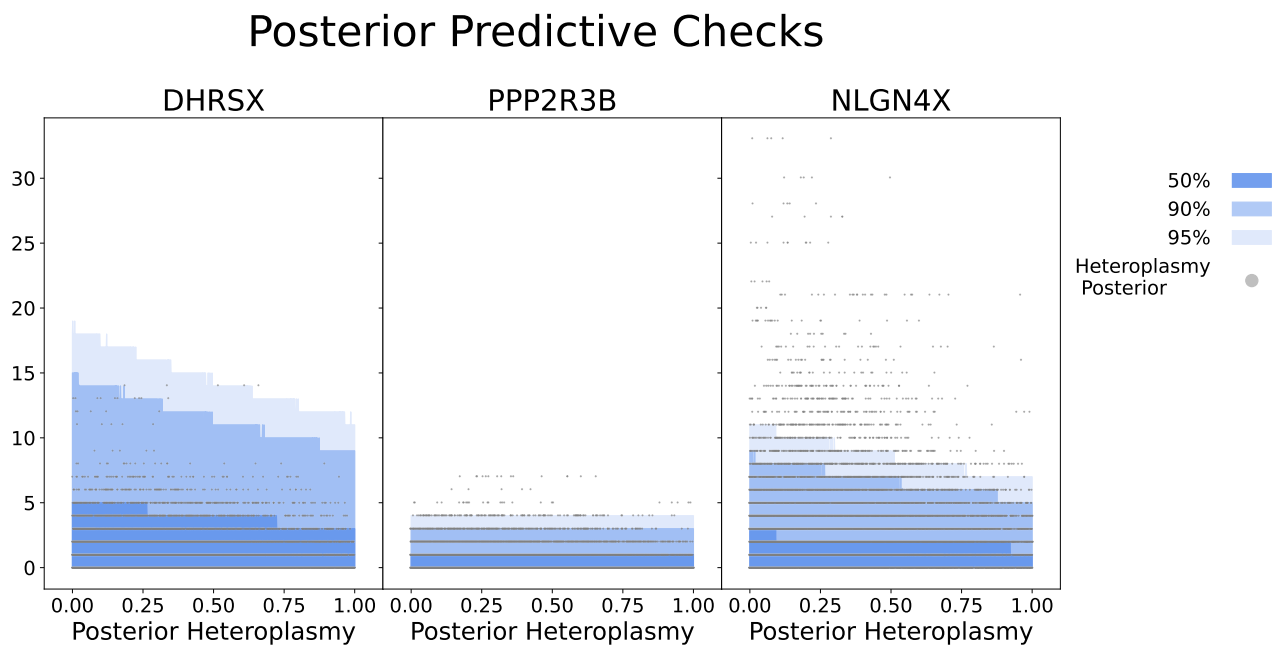


Figure 5.17: Illustrative posterior predictive plots for three genes. Blue shading indicates different quantiles of the posterior predictive distribution, which appropriately displays a progressive increase in the proportion of the data covered.

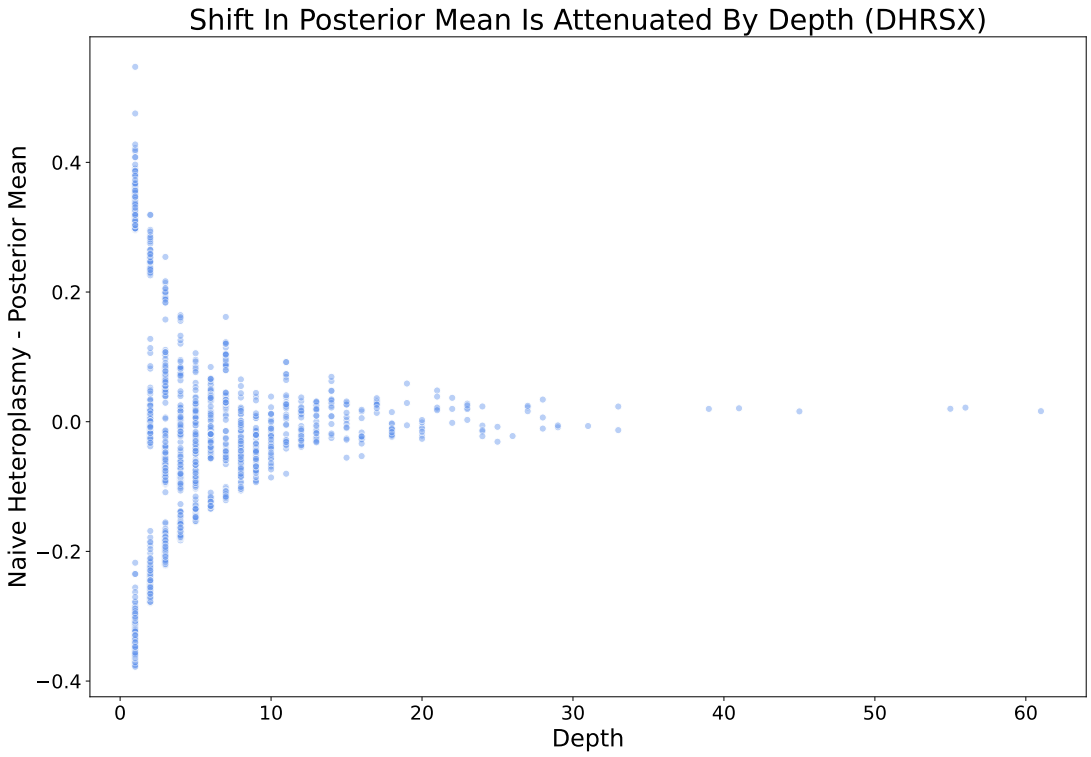


Figure 5.18: As depth increases, the degree to which posterior means are shifted from the naive point estimates is attenuated. Intuitively, this can be thought of as the model being reluctant to shift heteroplasmy posteriors on high-depth data, as is intuitively desirable.

5.9 Functional Relationships

Having obtained posterior samples from the model, we can now use it to investigate the associations observed between the heteroplasmy of mtDNA mutations and gene activity in the developing brain. We highlight that the following approach addresses the heteroplasmy of each mutation independently. Moreover, it treats the effect of a mutation's heteroplasmy on each gene independently, and does not account for cell type structure in the data. Though independence assumptions feature in other single-cell genomics algorithms, it is reasonable to question this assumption [21]. Moreover, this is a continued piece of analysis and should be judged accordingly. The following should be taken as indicative and the conclusions drawn as subject to further refinements and improvements. We discuss this further in the future work section at the end of this chapter.

Figure. 5.19 illustrates the progression from individual model parameters to a broader aggregated analysis. We selected 3010G>A for illustrative purposes. 3010G>A is situated in MT-RNR2, one of the two ribosomal genes in the mtDNA and has previously been identified *in silico* to alter the tertiary structure of the ribosomal RNA [217]. Previous evidence has suggested that mtDNA mutations causing structural changes to rRNA genes may play a role in disease [218]. It was also among the most common in our data allowing us to search for patterns that were recurrent across samples.

For each regression model, we obtain a posterior distribution on the regression coefficient, β, which captures the association between a mutation's heteroplasmy and the activity of a gene. Figure. 5.19 a) displays the posterior distributions of regression coefficients between 3010G>A and five genes in the CS14 medulla data. These posteriors display confidence in a range of effect sizes. The null parameter value is highlighted by the vertical dashed line. The two-tailed posterior quantile, denoted q, of the null is then computed as in Chapter 4. This is repeated for all genes analysed in the CS14 medulla data.

To visualise the range of effect sizes and quantiles we produce a *Bayesian volcano plot*, displayed in Figure. 5.19 b). We convert the regression coefficient into the implied $\log_2$ fold change in gene activity that is expected between cells with 0% and 100% heteroplasmy, and compute the posterior mean of this value. Plotting the posterior mean $\log_2$ fold change against the $-\log_{10}(q)$

gives a Bayesian analog of the conventional volcano plot. A range of effect sizes and posterior quantiles are observed. At more extreme posterior quantiles, we see horizontal lines emerging. These artifacts arise from the limited number of posterior samples drawn for each model. When parameter values are sufficiently extreme in either direction, only a small number of samples will be on the other side of $\beta = 0$ to the majority of the samples. In some instances, there is no overlap of the posterior with zero. We map the posterior quantiles of these values to 10^{-3} to make $-\log_{10}(q)$ a finite number. With the set of posterior quantiles obtained for this analysis, we can apply the Benjamini-Hochberg procedure to obtain a list of genes with high-confidence associations to the mutation's heteroplasmy.

Fig 5.19 c) shows the pairplot of posterior samples for parameter values used in the negative binomial regression. The α and β exhibit a negative correlation, typical of the behaviour observed in Bayesian regression models [153]. Inferences about over-dispersion, represented by ϕ, are less dependent on the other parameter values. Having looked at parameters for 3010G>A in CS14 medulla we expand the analysis to encompass all three brain regions. We computed the posterior mean estimate for every negative binomial regression parameter in the data. We then computed the correlation of posterior means for each parameter across all genes between each of the three different brain regions. Fig 5.19 d) displays these correlation matrices. The intercept and overdispersion for parameters for each gene are strongly correlated between brain regions. This is consonant with each gene having a typical degree of activity, embodied in the intercept term, α, and variability in activity, represented by the overdispersion control parameter, ϕ. The association between heteroplasmy and gene activity is seen to be far less stable across brain regions. This is plausible for at least two reasons. First, any associations measured within a brain region could plausibly vary due to the complex nature of the interactions present, each mutation with disparate effects in different brain regions. Secondly, many genes are presumably affected to a negligible degree, with effectively null associations. In these null instances, the precise location of the posterior mean would therefore be attributable to sampling variation, diminishing the correlation observed in the estimate of these values between brain regions.

Having performed the Benjamini-Hochberg procedure on each of the datasets containing 3010G>A separately, we obtained a list of genes. In order to look at systematically recur-

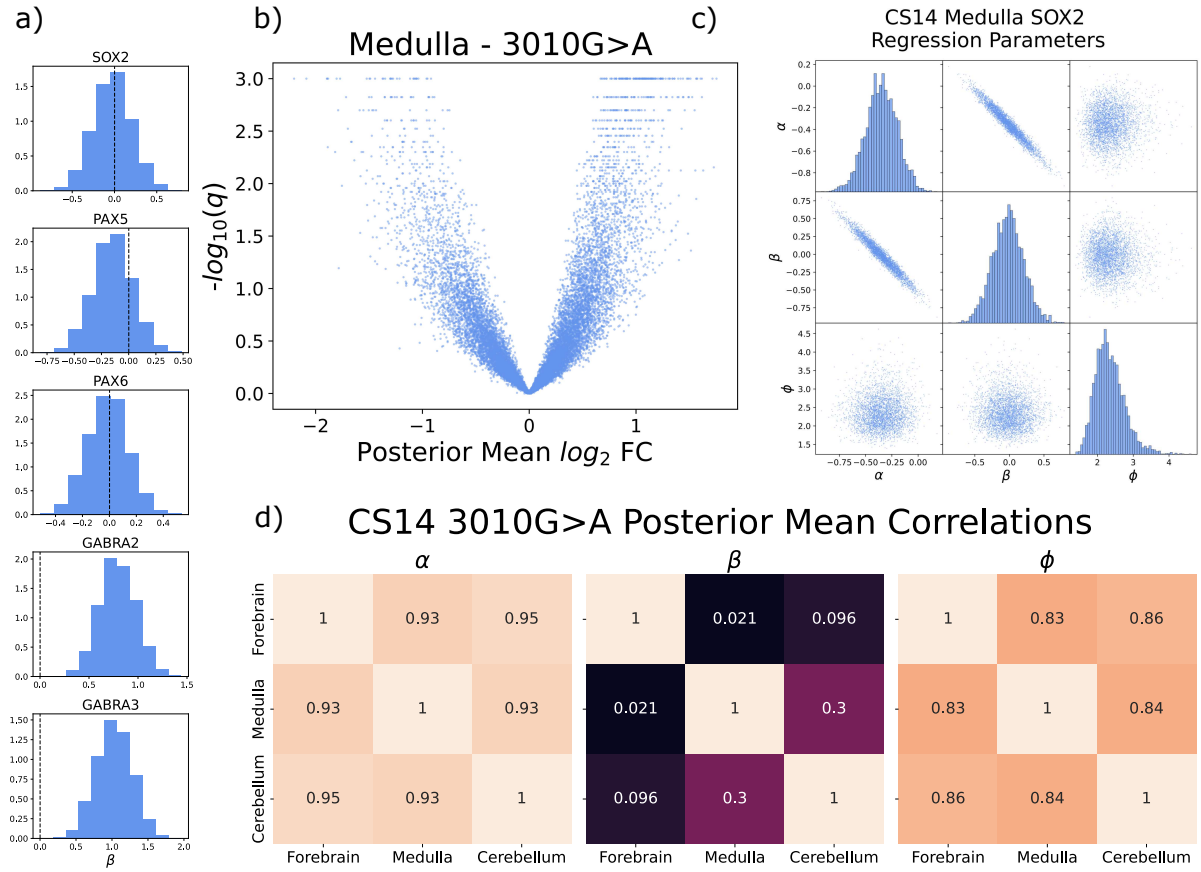


Figure 5.19: a) Illustrative posterior distribution of regression coefficients measuring the association between the heteroplasmy of 3010G>A and gene activity. The null value of $\beta = 0$ is represented by vertical dashed lines. b) The Bayesian volcano plot displays the posterior mean log_2 fold change against the posterior quantile of the $\beta = 0$ for all genes in the 3010G>A dataset. c) Pairplots displaying the posterior samples of the Negative Binomial model's regression coefficients for the regression of the heteroplasmy of 3010G>A against SOX2 activity within the CS14 Medulla data. We can see that the α and β posterior samples are negatively correlated, a common behavior in Bayesian regression models, whilst the over-dispersion parameter ϕ is less associated with the other parameter values. d) To get a broader view of model parameters across the data, we take the posterior mean of all model parameters for each gene. Then, for each parameter type, we compute the correlation between the posterior means across all three brain regions. The observed correlation for the intercept, α, and the observed correlation for the overdispersion, ϕ, is high across all three brain regions. This is consonant with each gene having a degree of activity and variability in expression that is stable across the brain. The association between the heteroplasmy of 3010G>A and gene activity is positive, though the strength of the correlation across brain regions is much weaker than for the other parameters. Two factors diminish the strength of this association: 1) The heteroplasmy of a mutation plausibly has a varying effect strength in different brain regions; 2) Many genes will possess no association with heteroplasmy, with the precise location of their posterior mean being attributable to sampling variation, attenuating the correlation between β means across brain regions.

ring effects, we looked at genes which were identified as "significant" in at least 5 datasets, producing a list of 403 genes. We looked to enriched GO terms in this set of genes using the `python` package `GSEAPy` 1.0.3 [219]. The results are displayed in table 5.3. These terms are linked to synapses and neurotransmitters [220]. These GO terms were highlighted by frequently occurring genes GABRB1, GABRA2 and GABRA3 three being present in the list of significant genes (see Figure. 5.19 a) for examples of posteriors of regression coefficients for these gene.) These genes are known to play are role in GABA receptors [221, 222]. Together, these observations suggest that the heteroplasmy of 3010G>A may modify the activity of genes that play a role in synapse function in the developing brain.

Gene_set	**Term**	**Adjusted P-value**
GO_Molecular_Function_2023	Transmitter-Gated Monoatomic Ion Channel Activity	0.004
	Ligand-Gated Monoatomic Anion Channel Activity	0.004
	Chloride Channel Activity	0.005
	Inhibitory Extracellular Ligand-Gated Monoatomic Ion Channel Activity	0.009
	Voltage-Gated Monoatomic Cation Channel Activity	0.01
	Voltage-Gated Monoatomic Ion Channel Activity	0.011
GO_Biological_Process_2023	Inorganic Anion Transmembrane Transport	0.027
	Calcium Ion Transmembrane Import Into Cytosol	0.027
	Chloride Transmembrane Transport	0.027
	Monoatomic Anion Transmembrane Transport	0.045

Table 5.3: Enriched GO terms for genes showing "significant" associations with the heteroplasmy of 3010G>A in at least 5 datasets.

5.9.1 Functional Effects Across All Mutations

Having focused on a specific mutation, we expand our analysis to explore all mutations together. Using the same procedure for computing posterior quantiles outlined above, we obtained a list of "significant" genes for every mutation in each Carnegie stage and brain region separately. We identified enriched GO terms for each of these lists. Table 5.4 displays the ten most recurrent GO terms from each of three gene sets: GO Molecular Function, GO Cellular Component, and GO Biological Process. From the the set regarding molecular function and biological processes, we see terms related to GTPAse, Phosphorylation, and the related term Guanyl-Nucleotide Exchange Factor Activity. GTPAse dynamins are known to play a role in mitochondrial fission and fusion dynamics in multiple species [223, 224, 225]. In particular DNM1L (typically referred to as DRP1) and OPA1, which were found to be significant in 17 and 18 of our

datasets respectively, play key roles. Previous research has shown that disruptions to normal GTPAse function are mediated by phosphorylation processes, modifying normal fission and fusion dynamics with disruptions to apoptosis [225]. Moreover, disruptions to DRP1 have been observed to disrupt cerebellar development in mice [226]. Additional research shows that modifications to DRP1 have been shown to alter cell fate decisions in neurogenesis [227, 228]. In terms of cellular components, we see multiple terms regarding the components of neural cells typically found in the brain, including: Dendrite, Axon and Neuron Projection. Previous work has shown that molecular manipulation of OPA1 and DRP1 reduce dendritic mitochondria content and lead to a loss of synapses and dendritic spines [229].

Gene Set	**GO Term**	**Counts**
GO_Molecular_Function_2023	GTPase Regulator Activity	23
	Protein Serine/Threonine Kinase Activity	22
	Guanyl-Nucleotide Exchange Factor Activity	21
	Cadherin Binding	20
	Tubulin Binding	19
	Microtubule Binding	19
	Adenyl Ribonucleotide Binding	18
	GTPase Binding	17
	Small GTPase Binding	17
	RNA Binding	15
GO_Cellular_Component_2023	Cytoskeleton	23
	Actin Cytoskeleton	23
	Axon	21
	Neuron Projection	20
	Cell-Substrate Junction	19
	Dendrite	19
	Intracellular Membrane-Bounded Organelle	18
	Focal Adhesion	18
	Basement Membrane	18
	Bounding Membrane Of Organelle	17
GO_Biological_Process_2023	Protein Phosphorylation	22
	Phosphorylation	22
	Protein Modification Process	21
	Regulation Of Intracellular Signal Transduction	21
	Peptidyl-Serine Modification	20
	Peptidyl-Serine Phosphorylation	19
	Regulation Of Small GTPase Mediated Signal Transduction	18
	Positive Regulation Of Transcription By RNA Polymerase II	18
	Vesicle-Mediated Transport	18
	Positive Regulation Of DNA-templated Transcription	18

Table 5.4: We obtained a list of "significant" genes for every mutation at each Carnegie stage and brain region separately. We identified enriched GO terms for the three set of gene sets: GO Molecular Function, GO Cellular Component, and GO Biological Process. We display the ten most recurrent GO terms in our data for each of these sets.

5.10 Conclusion

In this chapter, we explored the landscape of single-cell mitochondrial mutations in the developing brain. A comprehensive cartographic analysis was precluded by the shallow depth of coverage of the mitochondrial genome, due to the tissue samples being frozen prior to sequencing. Nevertheless, we were able to sketch the contours of the terrain at a coarse grained level.

The improved pipeline developed in Section 5.4 allowed for an end-to-end analysis, going from raw scATAC sequencing data to association analysis of widespread mtDNA mutation heteroplasmy and gene activity. The pipeline can seamlessly generalise its application across many samples and be run, in parallel, in a university high-performance-computing environment.

We performed a brief exploratory data analysis of the gene-activity, derived from the scATAC modality. The clustering analysis showed that, when aggregated, data most obviously clustered by the brain region. Samples from the same brain region were sequenced separately, precluding this clustering being due to batch effects and reflecting the underlying biology of similar cell types being present in a brain region through development. We then observed how the activity of genes associated with forebrain and hindbrain development clustered in a manner consonant with previous literature.

The challenges posed by the low-depth of the mtDNA in this data were then highlighted. To understand the sequence variability of the mitochondrial genome, we performed a pseudo-bulk analysis and measured the rate at which reads originating from different mitochondrial genomic loci deviated from the reference genome. In line with expectations, the D-Loop region was observed to be the most mutable, followed by CDS and rRNA mutations.

We then went on to explore mutations that were widespread, being present in multiple single-cells. The low-depth of the mtDNA posed a problem of distinguishing sequencing errors from genuine sequencing errors. In Section 5.7.3, we developed a conservative statistical test capable of identifying widespread single-cell mutations by comparing the number of cells carrying a sequence mismatch at a given position to the number that would be expected by sequencing errors. We performed this analysis separately on data from each brain region and Carnegie stage. The identified list of mutations has the classic *transversion* pattern of mtDNA mutations. We

also saw how this procedure reliably identified the same mutations in all three brain regions of each individual, across all Carnegie stages. We classified mutations according to whether they were in the D-Loop, synonymous, missense, rRNA, or tRNA mutations. We compared the observed frequency of mutation types to the expected frequency that would occur if mutation of each position of the mtDNA genome was equally likely. This revealed that D-Loop and synonymous mutations were "significantly" over-represented, with strong and "significant" under-representation of missense mutations. This pattern is consonant with the generally benign nature of the former and the deleterious functional effects the latter can have. This was further backed up by comparing the expected frequency to the observed of CDS mutations as a whole, with low and high-pathogenic missense mutations being strongly underrepresented. Recent studies have presented evidence that synonymous mtDNA mutations can have a functional effect according to the wobble pattern they display [214]. Our data produced a similar frequency pattern when comparing the expected and observed frequency of different wobble mutations, though, due to the small sample size lacked the power to achieve "significance".

The set of identified, high-confidence mutations was then reduced to those that exhibited characteristics that suggested they were not homoplasmic/haplogroup mutations and were present in enough cells with sufficient depth to measure the association using the methodology developed in Chapter 4. We then measure the association between the heteroplasmy of these mutations and the gene activity of genes in the data. This required Bayesian inference to be performed *millions* of times. Generally, Bayesian inference using an MCMC algorithm is far more computationally intensive than a gradient-based statistical model such as MLE. Software engineering challenges required to scale this Bayesian inference procedure to such a massive repeated use had to be overcome. This was primarily achieved through the use of parallelization at different scales. At the smallest, model implementations leveraged multiprocessing techniques to run 10s of models in parallel across 100s of CPU cores. At a larger scale, models being fit for different brain regions and Carnegie stages could be deployed on separate compute nodes, each with 100s of CPUs and 100s GBs of RAM at the same time. This approach reduced over a year of sequential computation to the time scale of weeks. The resulted in 100s GBs of parameter posterior samples for the negative binomial regression model being saved to disc.

As an illustration, standard visual checks of trace plots of some samples were performed, though this approach is unfeasible for the large number of models that we obtained. Standard MCMC quality statistics suggested sampling was widely good, and that mixing was generally satisfactory. Posterior predictive checks also suggested the predicted quantiles of gene-activity, as predicted by heteroplasmy, were in line with observation, suggesting adequate model fit.

Finally, we turned our attention to understanding the aggregated behaviour of thousands of model parameters. We illustrated how starting from the individual posteriors, we could compute the posterior quantile of the null in the data for the 3010G>A mutation in the CS14 medulla. We saw the range of effect sizes and posterior quantiles in the Bayesian volcano plot for all regressions performed in the 3010G>A mutation in the CS14 medulla. This analysis was then performed for each mutation-gene pair, in each sample of the data separately. Analysis of the posterior mean estimates of the regression parameters in all three brain regions was performed. We saw that the α and ϕ parameters were strongly associated for each gene in all three regions. This is line with α determining the general level of the activity of that gene and ϕ its over dispersion. These properties would generally be expected to be stable across brain regions, and correlations in posterior mean estimates cohere with this. Regression coefficients β between the heteroplasmy of 3010G>A and gene activity of each gene was only weakly correlated. This is expected. Firstly, the uncertainty in regression coefficients would be uncertain due to the low-depth data. Secondly, it is not expected that the heteroplasmy would display strong systematic effects across brain regions for many genes in the genome.

Exploration of enrichment GO terms showed evidence of a link between the heteroplasmy of early developmental mutations and the GTPAses activity. We repeatedly found DRP1 and OPA1 in the list of genes significantly associated with mitochondrial heteroplasmy. These GTPAses are key regulators of mitochondrial network structure with known impacts on neurogenesis and brain development, which has previously been linked to mitochondrial function [227, 228, 229]. Together, these observations suggest that the heteroplasmy of widespread developmental mtDNA mutations is associated with modifications to regulators of mitochondrial network structure and that this in turn is associated with alterations to structural components of neurones and their precursors in early development.

Taken together, the results of this chapter show a first glimpse at the landscape of widespread, single-cell, developmental, mitochondrial mutations and their functional impacts. This was made possible through the novel use of bioinformatics and statistics.

5.11 Future Directions

The opening epigraph of Chapter 4 emphasised how no models are true, and all models leave out details of the reality they represent. Negative binomial regression including heteroplasmy uncertainty is no exception. It is a substantive improvement over models that fail to include heteroplasmy uncertainty into their predictions. As such, the model is more applicable than alternatives, especially to low and variable depth data. Nevertheless, in so far as we have applied it here, the model relies upon numerous independence assumptions, which is common in single-cell data. An obvious source of confounding is cell type. The analysis could be developed by allowing assigned cell types to be incorporated into the model's structure. A step in this direction could be to allow varying for intercept and regression coefficient terms and modelling them hierarchically. With only a few cell types, the dimensionality of the problem need not be substantively increased. However, hierarchical models have complex geometries and ensuring fast, reliable sampling takes place is known to be challenging [161]. This is a problem when the model needs to be applied to thousands of genes without much manual tuning and validation, which MCMC often requires. In circumstances where cell counts are high enough, we could attempt to address cell type concerns through a separated analysis of each cell type. Further interesting questions surround how we can tease apart the impact of the heteroplasmy of one mutations against another, and how to address this.

Further refinements to the analysis performed in this chapter would seek to decompose the range of functional effects further. This would involve identifying systematic effects within mutation types, answering questions such as how do rRNA genes and missense mutations differ in their effects? What systematic impacts WCF>Wobble do mutation have?

Recent work has also proposed theoretical links between mitochondrial network dynamics and heteroplasmy [230]. Further work could involve interpreting the evidence presented in this chapter in light of these previous theoretical developments.

To end this chapter, we make some observations about how the computational frameworks deployed in this thesis might benefit from further improvements. All of the preceding work in this thesis made extensive use of computational pipelines that connected various software tools and statistical techniques. As bioinformatics pipelines continue to grow in complexity,

it is vital that additional emphasis is placed on assessing their reproducibility, validity, and integrity. The work in this thesis strove to abide by best practises as far as possible regarding reproducibility. This was done by using a powerful pipeline framework, `snakemake`, that allows code to be modular and structured in manner that can be shared with other researchers. Though progress in the area of reproducibility continues, the ability researchers to reproduce one another's work is still constrained by various factors. This includes the time resources required to understand another researcher's code, which can be of varying quality, as well as the time required to execute computationally intensive bioinformatics programs. These constraints on reproducibility are further compounded by the fact that many researchers lack the resources to store large amounts of bioinformatics data and lack the necessary compute to execute intensive bioinformatics programs.

With these constraints in mind, we point to further development that could be performed on the pipeline used in this work as well as adopted more broadly in bioinformatics practice: 1) Standardised unit testing; 2) Baselining module outputs with checksums; 3) Null testing a pipeline.

Unit testing is standard in industrial settings and involves testing individual components of a software application in isolation to ensure each part functions correctly. For example, a function that is defined within a code library is given a specified input and the returned value compared against a specified benchmark. Testing in this manner allows code in a code base to be modified and enhanced, whilst ensuring that functionality is not compromised. In terms of reproducibility, this allows researchers to ensure that software developed in another researcher's software ecosystem operates in their own compute environment as intended. In a manner adjacent to this, we believe that pipelines executed on data actually used for publication should produce "baseline values" which can be used for comparison by other groups. The simplest example of this would be the use of *checksums* at various steps in a pipeline. Checksums are usually deployed to ensure that large data files have been correctly downloaded. In it's simplest terms, a checksum is an alphanumeric string obtained by running a data set through a checksum algorithm. Good checksum algorithms substantially alter the value of the checksum for minor changes introduced to the data. By publishing checksums for data produced at

various steps of the pipeline, researchers attempting to reproduce an analysis can execute the pipeline on published data, produce the output, compute the checksums and compare their checksum against a published value. Upon seeing a match, they can be sure the input and output data are identical, and any deviations can be investigated more thoroughly. Although checksum compliance is available as a tool, it is not widely adopetd.

Finally, we believe that further validation of bioinformatics tools such as controlling false positives through the use of null testing would be beneficial to this work and in bioinformatics work more broadly. At a biological level, this could include the use of negative trial laboratory data sets to against which pipelines are benchmarked. For instance, multiple samples collected under identical conditions, from the same organism. Valid pipelines, data manipulations and statistical techniques should not return meaningful statistical differences in these settings. At the level of bioinformatics, even more generic scrambling of indices associated with tissues or cells could be used to ensure that pipelines do not return erroneous positive samples. In computational statistics, analogous techniques such as the bootstrap and permutation testing randomly resample data with different group labels attached. For example we might have random samples belonging to two groups, A and B. The null assumption is that data with different labels are exchangeable and that such labeling is in no way informative about the distribution of data. Repeated resampling is then used to empirically estimate the distribution of a test statistic under this null assumption. This empirical distribution is then used to assess the "statistical significance" of the test statistic for the real data [231]. Though intensive end-to-end re-running of a bioinformatics pipeline on resampled data is unfeasible due to computational constraints, components of a pipeline could be evaluated against subsets of data to place further controls on false positive elements of a pipeline.

Chapter 6

Addendum — Observing mtDNA Motion

6.1 Summary

In this short chapter, we move beyond exploring mitochondrial DNA sequencing data to the physical mtDNA that can be observed under a microscope. Super-resolution microscopy, in conjunction with the appropriate staining techniques, offers the ability to directly observe mtDNA nucleoids and track their movement across time. We develop a simple model that can guide interpretations of observations of mtDNA under the microscope. The model development treats mtDNA within the mitochondrial matrix as undergoing Brownian motion in a space constrained by the cristae. In particular, this model is useful for guiding microscopists in understanding how observations of the distances the mtDNA travel through time can be distorted by the boundaries of the cristae.

6.2 Tracking mtDNA

Modern microscopes are capable of tracking the positions in an x-y plane at a set frame rate. By taking the measured distance between sequential frames, one can measure the total distance traversed by an mtDNA over a time interval.

Previous research has shown that the movement of mtDNA is consistent with a random walk [232]. Cristae are distributed throughout mitochondria, and it is reasonable to suppose that these act as obstacles, capable of trapping mtDNA in the mitochondrial matrix in the inter-cristae space. In this setting, the motion of mtDNA is considered to be a Brownian motion confined to a box of finite size without the ability to exit. As lateral motion within the mitochondria is severely limited, this can be considered in 1-dimension for simplicity. This motion can be modelled through the use of reflecting boundary conditions. In the section 6.3 we demonstrate how to generate sample paths for a Brownian motion in a finite box. Section 6.4 follows by showing how failure to account for a limited observation framerate can distort distance measurements of nucleoids trap between cristae. In particular, we show how this can lead to the illusion that mtDNA trapped within larger spaces travel faster.

6.3 Simulating Brownian Motions In Finite Boxes

We can measure the coordinates of the nucleoid, $x_i(t_i)$, at a discrete sequence of times $t_0 \leq t_1 \leq ...t_i \leq t_{i+1}... \leq t_n$. Brownian motion is *Markovian*, meaning we can obtain a distribution over the coordinates at all time points as:

$$p(x_{i+1}, t_{i+1}; ...; x_{i+1}, t_{i+1}; x_i, t_i; ...; x_0, t_0) = p(x_0, t_0) \prod_{i=1}^{i=n} p(x_{i+1}, t_{i+1} | x_i, t_i) \tag{6.1}$$

Conditional transition probabilities are governed by the solution to the Fokker-Planck equation 6.2 [233]. Given an initial coordinate, x_0, at time $t = 0$, $p(x, t'|x_0, 0)$ describes the distribution over position, x, at a later time t'.

Diffusion on a finite domain is the solution to the diffusion equation:

$$\frac{\partial p(x, t|x_0, 0)}{\partial t} = D \frac{\partial^2 p(x, t|x_0, 0)}{\partial x^2} \tag{6.2}$$

We can represent a 1-dimensional Brownian motion in a finite box with left-hand boundary at $x = 0$ and right-hand boundary condition at $x = L$. Confinement to the box is accounted for with reflecting boundary conditions at $x = 0$ and $x = L$, being equivalent to $\frac{\partial p(0,t|x_0,0)}{\partial x} = \frac{\partial p(L,t|x_0,0)}{\partial x} = 0$. The initial condition will be $p(x, t = 0|x_0, 0) = \delta(x - x_0)$

The following is the solution to equation 6.2:

$$p(x, t|x_0, 0) = \frac{1}{L} + \frac{2}{L} \sum_{n=1}^{\infty} e^{-\left(\frac{n^2 \pi^2 D t}{L^2}\right)} \cos\left(\frac{n\pi x}{L}\right) \cos\left(\frac{n\pi x_0}{L}\right) \tag{6.3}$$

We can check that this satisfies the diffusion equation by performing the appropriate partial derivatives on the LHS and RHS to equation 6.2:

$$\frac{\partial p(x,t|x_0,0)}{\partial t} = -\frac{2\pi^2 D}{L^3}\sum_{n=1}^{\infty} n^2 e^{-\left(\frac{n^2\pi^2 Dt}{L^2}\right)} \cos\left(\frac{n\pi x}{L}\right)\cos\left(\frac{n\pi x_0}{L}\right) \tag{6.4}$$

$$D\frac{\partial^2 p(x,t|x_0,0)}{\partial x^2} = D\left(-\frac{2\pi^2}{L^3}\sum_{n=1}^{\infty} n^2 e^{-\left(\frac{n^2\pi^2 Dt}{L^2}\right)} \cos\left(\frac{n\pi x}{L}\right)\cos\left(\frac{n\pi x_0}{L}\right)\right) \tag{6.5}$$

$$\frac{\partial p(x,t|x_0,0)}{\partial t} = D\frac{\partial^2 p(x,t|x_0,0)}{\partial x^2}$$

We can further check that this satisfies the reflecting boundary conditions with all terms going to zero at the edge of the domain:

$$\frac{\partial p(x=0,t|x_0,0)}{\partial x} = -\frac{2\pi}{L^2}\sum_{n=1}^{\infty} n e^{-\left(\frac{n^2\pi^2 Dt}{L^2}\right)} \sin\left(\frac{n\pi x}{L}\right)\cos\left(\frac{n\pi x_0}{L}\right)\Bigg|_{x=0} = 0 \tag{6.6}$$

$$\frac{\partial p(x=L,t|x_0,0)}{\partial x} = -\frac{2\pi}{L^2}\sum_{n=1}^{\infty} n e^{-\left(\frac{n^2\pi^2 Dt}{L^2}\right)} \sin\left(\frac{n\pi x}{L}\right)\cos\left(\frac{n\pi x_0}{L}\right)\Bigg|_{x=L} = 0 \tag{6.7}$$

Using the equation 6.3, we could attempt to generate sample paths of the Markov chain sequentially. Given a state x_i, we can generate the next state at a later time t later from $p(x_{i+1}, t|x_i, 0)$. This can be done recursively to generate coordinates at sequential times. Obtaining samples from 6.3 is non-trivial, especially when the infinite sum is considered. To do so, we can use Hamiltonian Monte Carlo to obtain samples. The probabilistic programming language `stan`, though designed to obtain samples from posteriors distributions conditional on data, can be used to obtain samples from any arbitrary distribution. Appendix B contains the `stan` required to do so, using the sum truncated at 1,000 terms. Some examples of this are displayed along side the the analytic functional form of equation 6.3.

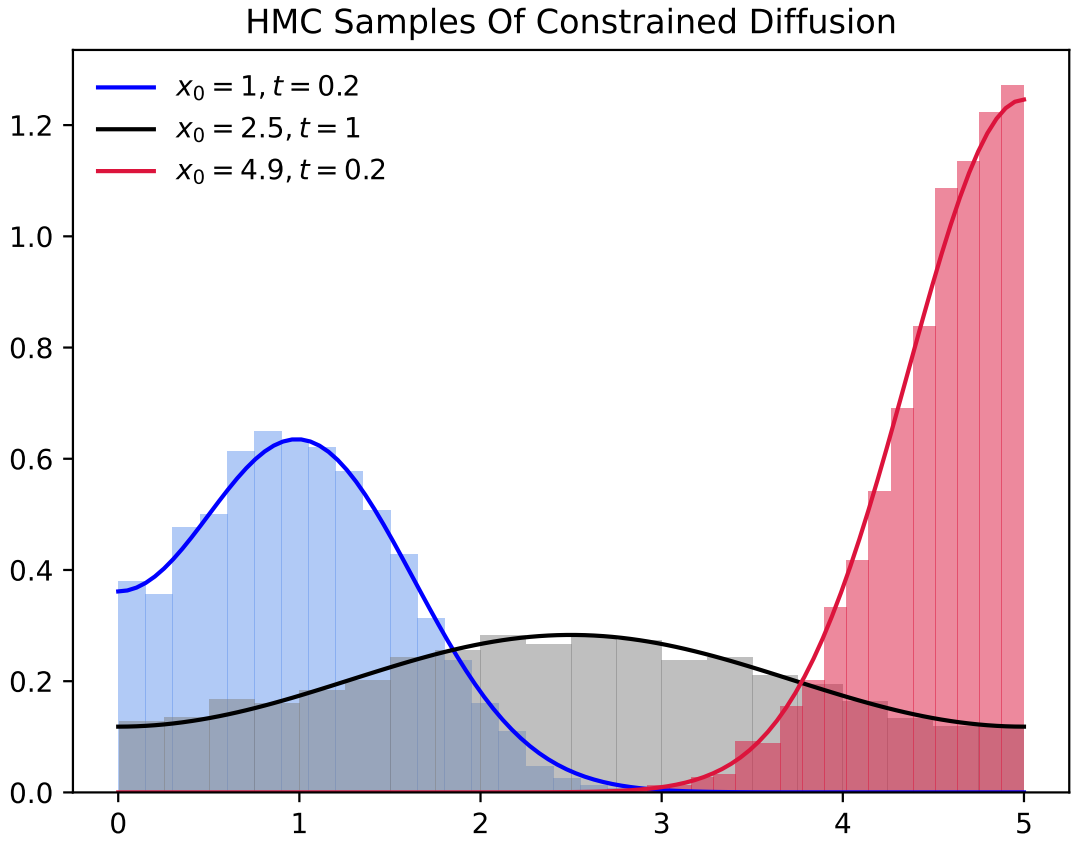


Figure 6.1: HMC samples from equation 6.3 in a box with $L = 5$. Three different values for initial start points, x_0, and the elapsed time intervals t are used. Normalised histograms show samples obtained from HMC are nicely in line functional form 6.3 (solid lines). We can see that $\frac{\partial p(x,t|x_0,0)}{\partial x} = 0$ at both boundaries as required.

Though it is possible to generate transitions in this manner it is computationally expensive to perform a full HMC for each time interval in the Markov Chain. Instead, we give an alternative, more direct solution which is equivalent to equation 6.3. Fokker-Plank equations have corresponding stochastic differential equations that can be used to generate sample paths [233]. This makes use of a Wiener process, W_t[1]. Conventionally, we can simulate sample paths of Brownian motion (without drift) on the infinite domain by using SDEs in the following manner:

[1]A Wiener process $W(t)$ is a continuous-time stochastic process that satisfies the following properties: 1. $W(0) = 0$ 2. $W(t)$ is almost surely continuous. 3. $W(t)$ has independent increments: for $0 \leq s < t$, the increment $W(t) - W(s)$ is independent of the past values $\{W(u) : 0 \leq u \leq s\}$. 4. $W(t)$ has Gaussian increments: $W(t) - W(s) \sim \mathcal{N}(0, t - s)$ for $0 \leq s < t$, where $\mathcal{N}(0, t - s)$ denotes the normal distribution with mean 0 and variance $t - s$.

$$dX_t = \sigma dW_t = \sqrt{2D} dW_t \tag{6.8}$$

Given some initial start point x_0, one adds increments for time intervals of length t using equation 6.8. This is the *Euler–Maruyama* approximation to the SDE. We now demonstrate that in the reflecting boundary conditions this is equivalent to a *reflected Gaussian distribution.* Now that we have a reflecting domain, we need to ensure that the particle stays within the specified range $x \in (0, L)$. If, at a certain time increment, the particle's proposed new position, x, jump were to exit beyond the bound L by an amount $(X - L)$, we would reflect it back into the domain to the position $X := L - (X - L) = 2L - X$. If this X were to extend in the negative direction beyond 0, we would reflect about 0, assigning X to $-X$. If this were in turn beyond L, we would repeat the first step, and so on, until the particle ended up back in the finite domain. A particle that would have traveled some total distance and landed outside the box, had the boundaries not existed, would still travel the same total distance between time intervals. Collisions with the boundaries would retain it inside the box. This procedure of reflecting the Weiner proposal back inside the boundary amounts to a reflected Gaussian distribution.

We now show that the reflected Gaussian distribution representation of this Brownian motion in a box is equivalent to 6.3. Given a start position at $x_0 \in (0, L)$, and a normal distribution of length L, we recursively reflect the probability mass out the boundaries back into the $(0, L)$ domain. This is similar but distinct from the *wrapped Gaussian* which puts mass back into the range by subtraction, rather than reflection [234]. If we set $L = \pi$, then the distribution of the reflected Gaussian for $x \in (0, \pi)$ is given by:

$$p(x, t|x_0, 0) = \sum_{k=\pm\infty} N(x + 2\pi k|x_0, \sigma^2) + N(x + 2\pi k + 2(\pi - x)|x_0, \sigma^2) \tag{6.9}$$

If temporarily we extend the range of x to $x \in (-\pi, \pi)$, we can later discard $x \in (-\pi, 0)$. By using this this temporary extension and re-writing the sum as:

$$p(x,t|x_0,0) = \sum_{k=\pm\infty} N(x+2\pi k|x_0,\sigma^2) + N(x+2\pi k|x_0-2(\pi-x),\sigma^2) \tag{6.10}$$

Which is equivalent to the sum of two wrapped Gaussians on $x \in (-\pi,\pi)$, one centered on $\mu_1 = x_0$ and another centered on $\mu_2 = x_0 - 2(\pi - x)$. The formula for a wrapped Gaussian with initial coordinates μ, σ^2 can be written as:

$$p_{WG}(\theta|\mu,\sigma^2) = \frac{1}{2\pi} + \frac{1}{\pi}\sum_{n=1}^{\infty} e^{-\left(\frac{n^2\sigma^2}{2}\right)} \cos\left(n(\theta-\mu)\right) \tag{6.11}$$

Subbing this formula into the sum above gives the transition probability of the finite domain diffusion as being:

$$p(x,t|x_0,0) = \frac{1}{\pi} + \frac{1}{\pi}\sum_{n=1}^{\infty} e^{-\left(\frac{n^2\sigma^2}{2}\right)} \left(\cos\left(n(x-\mu_1)\right) + \cos\left(n(x-\mu_2)\right)\right) \tag{6.12}$$

$$= \frac{1}{\pi} + \frac{2}{\pi}\sum_{n=1}^{\infty} e^{-\left(\frac{n^2\sigma^2}{2}\right)} \cos\left(\frac{n(x-\mu_1)+(x-\mu_2)}{2}\right)\cos\left(\frac{n(x-\mu_1)-(x-\mu_2)}{2}\right) \tag{6.13}$$

$$= \frac{1}{\pi} + \frac{2}{\pi}\sum_{n=1}^{\infty} e^{-\left(\frac{n^2\sigma^2}{2}\right)} \cos\left(nx - \frac{n}{2}(\mu_1+\mu_2)\right)\cos\left(\frac{n(\mu_2-\mu_1)}{2}\right) \tag{6.14}$$

$$\tag{6.15}$$

Subbing $\mu_1 + \mu_2 = 2(x_0 - \pi + x)$, $\mu_2 - \mu_1 = 2(\pi - x)$

$$p(x,t|x_0,0) = \frac{1}{\pi} + \frac{2}{\pi}\sum_{n=1}^{\infty} e^{-\left(\frac{n^2\sigma^2}{2}\right)} \cos\Big(nx - \frac{n}{2}2(x_0-\pi+x)\Big)\cos\Big(\frac{n}{2}2(\pi-x)\Big) \tag{6.16}$$

$$= \frac{1}{\pi} + \frac{2}{\pi}\sum_{n=1}^{\infty} e^{-\left(\frac{n^2\sigma^2}{2}\right)} \cos\left(n(x_0-\pi)\right)\cos\left(n(x-\pi)\right) \tag{6.17}$$

$$\tag{6.18}$$

Using $cos(A)cos(B) = cos(A + n\pi)cos(B + n\pi)$, we obtain the final result that:

$$p(x, t|x_0, 0) = \frac{1}{\pi} + \frac{2}{\pi}\sum_{n=1}^{\infty} e^{-\left(\frac{n^2\sigma^2}{2}\right)} \cos(nx_0)\cos(nx)) \tag{6.19}$$

$$\tag{6.20}$$

We simply need to restrict our attention to the initial domain $x \in (0, \pi)$ to retain the appropriate normalisation. We can perform a change of variables which rescales the length of system whilst keeping the physics the same by setting $x = x'\frac{\pi}{L}$ such that in the x' perspective the box has length L. Also rescaling $x_0 = x_0'\frac{\pi}{L}$, and $\sigma = \sigma'\frac{\pi}{L}$ as appropriate, we obtain:

$$p(x', t|x_0, 0) = \left|\frac{dx}{dx'}\right| p(x, t|x_0, 0) \tag{6.21}$$

$$= \frac{\pi}{L} p(x, t|x_0, 0) \tag{6.22}$$

$$= \frac{1}{L} + \frac{2}{L}\sum_{n=1}^{\infty} e^{-\left(\frac{n^2\sigma'^2}{2L^2}\right)} \cos\left(\frac{n\pi x_0}{L}\right)\cos\left(\frac{n\pi x}{L}\right) \tag{6.23}$$

This recovers

$$p(x, t|x_0, 0) = \frac{1}{L} + \frac{2}{L}\sum_{n=1}^{\infty} e^{-\left(\frac{n^2\pi^2 Dt}{L^2}\right)} \cos\left(\frac{n\pi x}{L}\right)\cos\left(\frac{n\pi x_0}{L}\right) \tag{6.24}$$

This shows the equivalence between the reflected Gaussian SDE approach and the solution to the Fokker-Plank equation with reflecting boundaries. This gives a method of simulating sample paths for the SDE runs, using the algorithm outlined above which recursively reflects mass back into the sample domain. Additionally, it yields two sums for approximating the likelihood depending on the scale of σ. If σ is large, then the cosine representation works well. Alternatively, if the σ is small, only a few terms of the reflected Gaussian sum will be required to get a reasonable approximation of the probability.

This simple procedure allows one to generate proposal states directly using a standard Normal

random number generator, available in numerous programming languages, rather than having to rely on the expensive computation of HMC. Appendix B shows the python code that can be used to obtain sample runs using the reflected Gaussian. Verifying this works, figure 6.2 shows that samples generated from the reflected Gaussian correspond line up with the functional form of equation 6.3.

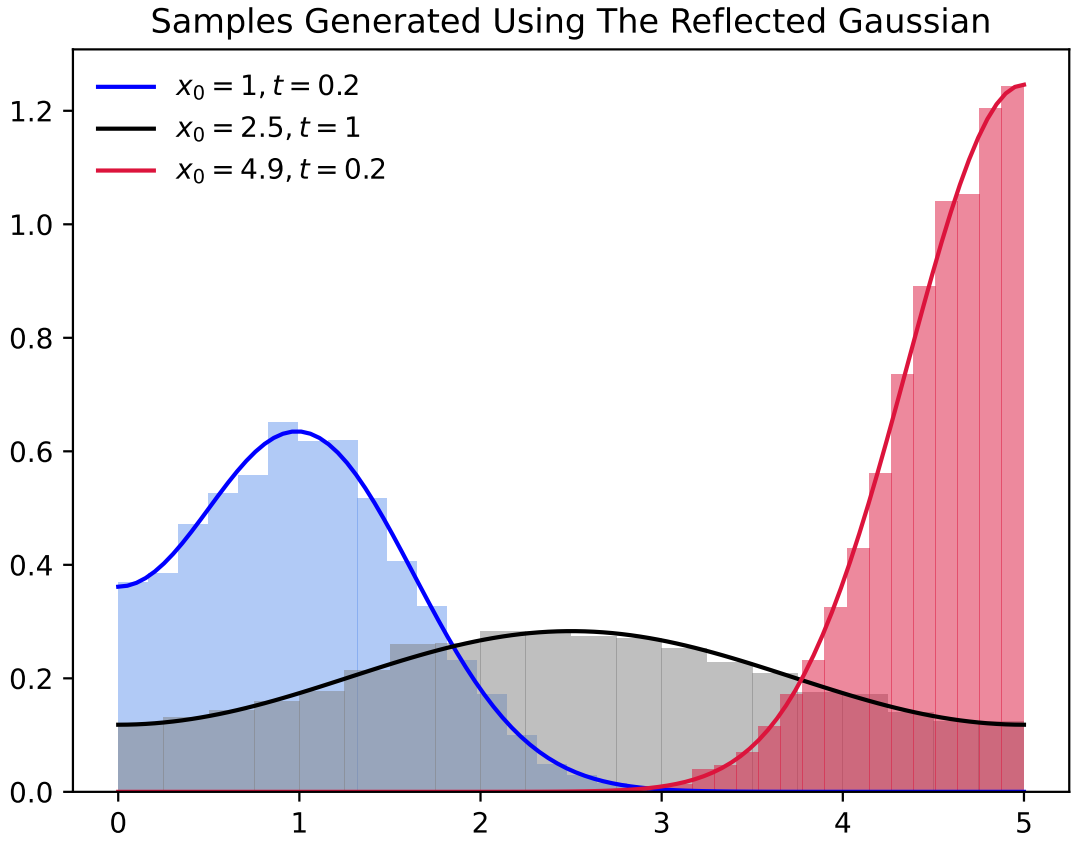


Figure 6.2: Samples from the reflected Gaussian generated using python code (see Appendix B). Three different values for initial start points, x_0, and the elapsed time intervals t are used. Normalised histograms show samples obtained from the reflected Gaussian are nicely in line functional form 6.3 (solid lines). We can again see that $\frac{\partial p(x,t|x_0,0)}{\partial x} = 0$ at both boundaries as required. The reflected Gaussian is far more advantageous in that one does not need to obtain a full set of samples, and then selected on in the manner HMC requires. Instead, we can use the Wiener process to generate a single sample position and reflect about the boxes edges as required, saving considerable compute.

6.4 Cristae Distort Distance Measurements

Section 6.3 showed how the reflected Gaussian can be used to simulate sample paths and for Brownian motion in a finite domain.

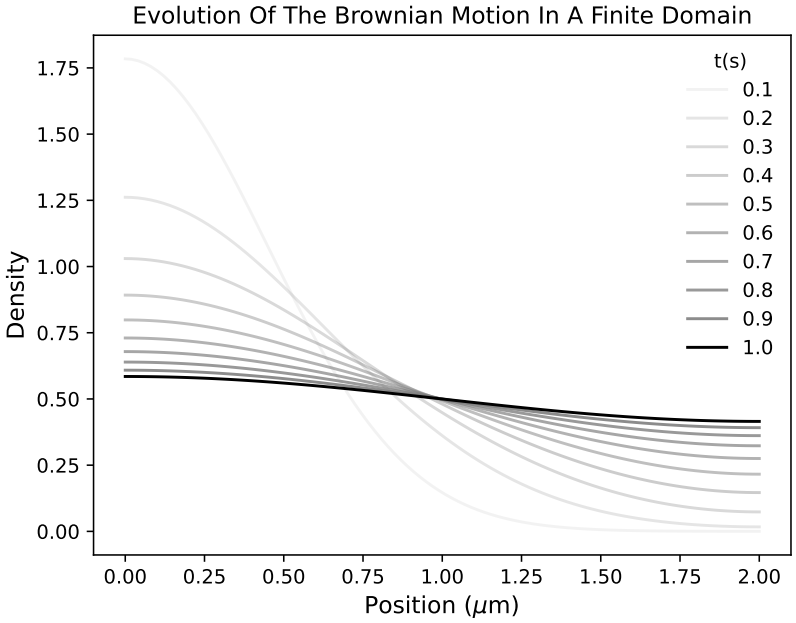
We can obtain indicative values of the diffusion constant for nucleoids by using the $v = 8.3 \times 10^{-22}\text{m}^3$, which assuming a spherical nucleoid gives $r = 5.8 \times 10^{-8}\text{m}$ [235]. The viscosity is around $\eta = 10\,\text{cP} = 0.01\,\text{kg} \cdot \text{m}^{-1} \cdot \text{s}^{-1}$ [236]. Using the Einstein-Smoluchowski equation, the diffusion constant is estimated to be:

$$\begin{aligned} D &= \frac{k_B T}{6\pi\eta r} \\ &= \frac{(1.38 \times 10^{-23}) \cdot 300}{6\pi \cdot (0.01) \cdot (5.8 \times 10^{-8})}\text{m}^2\text{s}^{-1} \\ &= 3.79 \times 10^{-13}\text{m}^2\text{s}^{-1} \\ &= 3.79 \times 10^{-1}\mu\text{m}^2\text{s}^{-1} \end{aligned}$$

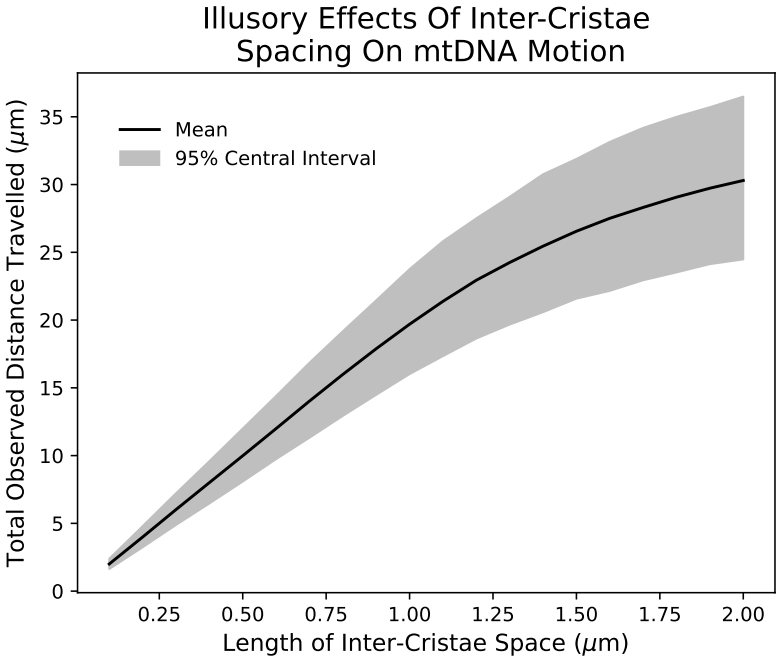
Using these parameter values, we can observe how equation 6.3 evolves through time. Figure 6.3 a) displays the distribution of position at 0.1s intervals in a space 2 μm dimension, consonant with typical observations of cristae spacing in mitochondria [237]. The system tends towards uniformity, the stationary state as $t \to \infty$, eventually reaching a uniform distribution. This poses a challenge for advanced microscopy that aims to track the motion of mtDNA within the inter-cristae spacing as framerates are typically on the 1s timescale. In even smaller inter-cristae spaces, the rate of relaxation is high enough that location of mtDNA in sequential frames can be considered to be iid uniform random variables, where the location of the mtDNA in one frame is uninformative about the location in the next frame.

This relaxation phenomenon can lead to erroneous conclusions about the effects of cristae spacing on mtDNA motion. Fig 6.3 b) shows how the observed distance travelled by a nucleoid recorded once a second for a minute is observed to grow linearly with intercristae space, before plateauing, implying an increase in the velocity of nucleoids. The underlying dynamics are

the same Brownian motion, regardless of inter-cristae length. As we get to larger inter-cristae spacings of around 2μm, we see this linear behaviour stops. Intuitively, when the inter-cristae space becomes large enough, the nucleoid would effectively be undergoing free Brownian motion unconstrained bound the bounding cristae.



(a)



(b)

Figure 6.3: a) The evolution of equation 6.3 through time. We can see as time proceeds, the system tends towards uniformity, the stationary state as $t \to \infty$. b) Illusory relationships between mtDNA motion and cristae spacing can be produced by finite sampling rates. We simulate 60 second long recordings, with frames captured once every second. Brownian motion sample paths were produced using the reflected Gaussian distribution. Between each frame, absolute changes in position were recorded. Summing over the course of a minute to give a naive measurement of distance traveled by a nucleoid over the time interval $|\Delta x|$. We produced 10,000 sample paths, for a range of different values of inter-cristae spacing. The mean and 95% central interval of outcome is observed to grow. In reality, all mtDNA traveled an equal distance over the duration of the minute. The long time gap between frames fails to account for the collisions between mtDNA in a small window. Without observing this effect, the my be thought to dampen the motion of nucleoids.

6.5 Conclusion

This brief addendum chapter exploring the motion of mtDNA between cristae could be modeled using a Brownian motion in a finite domain with reflecting boundary conditions. This simple model gives a qualitative account of how the finite frame rate of the tracked motion of mtDNA can give rise to the illusion that mtDNA nucleoids travel faster when inter-cristae spacing increases.

Chapter 7

Conclusion

7.1 Thesis Summary

This thesis has been an in-depth exploration of how nascent single-cell technologies can be used to explore the biology of mtDNA in novel ways. We have made several contributions to understanding the power of this technology to deliver novel insight into the role that mtDNA plays in health and disease.

Chapter 2 began with material from a review article we published, detailing the new range of sequencing tools available to research scientists. We described the new single-cell omics technologies that have come into existence and detailed how they work. In particular, we focused on scATAC and scRNA sequencing technologies. Sequencing approaches offered by 10x genomics in particular offer the ability to genotype thousands of separate single cells, within a single experiment. As mtDNA is the most abundant DNA in the cell by molecular count, single-cell technologies offer a rich source of information that can be used to develop novel insights. The article emphasized how single-cell techniques allow us to explore the cellular heterogeneity of mtDNA. We laid out terminology for the variations in *depth* and *breadth* of coverage of the mitochondrial genome which is essential to understanding the quality of single-cell data with regards to mtDNA genotyping.

We detailed some of the few publications that have explored the use of these technologies to study mtDNA in different contexts. Primarily, these pioneer studies focused on the use of mtDNA as cellular barcodes which could be used in lineage tracing. Though this is an

interesting approach, we noted how much of the mitochondrial genome in these studies was under-explored, focusing only on highly prevalent mutations without exploring other forms of mitochondrial heterogeneity at the cellular level. We also highlighted some of the limitations encountered in present studies and gave some comparisons between different modalities, in particular comparing 10x 3' single-cell RNA sequencing and other modalities. The former offered a consistent, reliable depth of coverage near the ends of genes, whilst other approaches such as ATAC sequencing of Smart-Seq2 tend to offer a better view of the length of the entire genome. 10x 3' sequencing is, however, abundant within the literature with numerous publicly available datasets on 1,000s of cells available to be explored.

Building on the context set in chapter 2, our own novel single-cell research was presented in 3. A simple model from population genetics, the *Moran model*, provides an understanding of how somatic mutations in post-mitotic cells would accumulate and increase in heteroplasmy throughout life. In this simple setting, the model is governed by two parameters — the mutation rate during replication, and the population size of mtDNA within a single cell. Remarkably, simulations using literature values for these parameters indicated that the heteroplasmy of somatic mutations, some of which will be deleterious and would begin to reach pathological levels on timescales consistent with modern human lifespans. Such mutations have the ability to contribute to cellular damage as we age.

As most of these mutations occur in single cells, they would be missed from the bulk tissue analyses that dominated the literature in previous years. As such, we sought to shed light on these *cryptic mutations* empirically by leveraging various single-cell genomics datasets that already existed in the literature, particularly single-cell RNA datasets which are abundant and offer the ability to link cryptic mutations with pathology. As there was no existing computational pipeline available to process this data in the manner we required, we developed one. Through substantial effort, we developed a pipeline capable of going from raw sequencing data right to heteroplasmy estimates of single cells.

With this pipeline to hand, we were able to process single-cell data from multiple tissues and numerous individuals in a straightforward fashion. We empirically demonstrated that the heteroplasmy of cryptic mtDNA mutations increased with age and did so in a manner

consonant with simple theory. We further gave indicative evidence that an established enhancer of longevity, caloric restriction, impeded this increase in heteroplasmy in rats. Beyond these initial findings, we were able to link the presence of non-synonymous cryptic mutations to many of the established hallmarks of ageing. Together, these findings highlight a previously missed and important component of human ageing: heterogeneity in the mitochondrial genome at the single-cell level has the potential to lead to a mosaic of cellular mitochondrial diseases.

Moving on from the role of cryptic mutations in ageing, and thus life's end, the thesis explored the role of widespread mitochondrial mutations in human development, at life's beginning. Rather than relying on previously available publicly available datasets, our collaborators at the Mitochondrial Biology Unit at the University of Cambridge presented us with novel single-cell sequencing data from embryonic human brains. This data presented the opportunity to get a first glimpse at the landscape of mtDNA mutations in the developing brain. Unfortunately, as the specimen was necessarily frozen, this severely limited the depth of coverage of the mtDNA that could be obtained. These constraints precluded us from being able to comprehensively map the landscape of development mtDNA mutations. Instead, we set out to sketch the contours of the terrain at a coarse level. To do even this required a radical new toolset.

Just as genomics has undergone a revolution in recent years, advances in Bayesian statistics and computation have taken great strides forward. Pre-existing probabilistic programming languages now make the implementation and fitting of bespoke Bayesian models using advanced samplers like Hamiltonian Monte Carlo, simpler than times when one would have to implement a Metropolis Hastings sampler from scratch. Such tools make Bayesian inference, a natural language for describing uncertainty, practical to solve real-world problems. With this in mind, we sought to incorporate the uncertainty of our heteroplasmy estimates in low-depth data when measuring the association between heteroplasmy and gene activity. Such an approach, we hoped, might allow us to glimpse the landscape of developmental mitochondrial mutations. Chapter 4 went through steps of methodological development, showing how we could retain the uncertainty in cellular heteroplasmy estimates imposed by limited depth, and still measure the association between heteroplasmy and gene activity. We developed a novel, negative binomial heteroplasmy regression model which retained predictor uncertainty. A simple demonstration

showed how a corresponding regression using naive point estimates of heteroplasmy would go woefully astray. Our Bayesian model, retaining uncertainty, was able to recover the true parameter values even in extremely low-depth regimes.

We went on to strain our model into very low-depth regimes, comparable to that observed in our later data. The model could still perform well. We also ensured that the model's implementation details were correct through the use of Simulation Based Calibration. Distinct from this, we showed how the model recovered the appropriate regression coefficients with enough cells, even with miss-specified priors on the regression coefficient compared to the data generating mechanism. Reassuringly, this behaviour was seen for the numbers of cells we typically possess in single-cell experiments, 100-1,000s. This dovetailed with the final section of the chapter, which considered an extreme form of model misspecification, in which the regression coefficients were firmly zero. This particular flavour of model misspecification carries a resemblance to conventional null hypothesis testing, for which multiple test procedures are appropriate. In general, there is no standard Bayesian approach for multiple hypothesis corrections. However, in some regimes, the posterior quantile of null value $\beta = 0$, becomes uniformly distributed in a manner akin to p-values. The Benjamini-Hochberg procedure for controlling the false discovery rate, relies on the uniform(0,1) distribution of p-values under the null. As the posterior quantiles of $\beta = 0$ for data generated from $\beta = 0$ are uniformly distributed for some Bayesian models, the Benjamini-Hochberg can be applied to these quantiles in an analogous manner. To apply this, we first had to show that the model's posterior quantiles of β were uniformly distributed when the data were generated from $\beta = 0$. Some Bayesian models possess uniform quantiles about the true parameter when data are generated from a single unique parameter value, as the number of data points increases. We demonstrated that the regression coefficients of our Bayesian model were among this set of models. Furthermore, through simulation, we showed that we could control the false discovery rate in line with the theoretical bound.

Chapter 5 then set about deploying this new machinery to real embryonic brain data, sequenced by our collaborators at Cambridge University's Mitochondrial Biology Unit, Nell Nie and Patrick Chinnery. We developed a statistical approach capable of confidently identifying significant mutations by comparing the observed number of cells the mutation was found into

the distribution that would be expected by sequencing error. Upon application of this method, the identified mutations were seen to be consistently present in all three brain regions for each individual analysed. Furthermore, the characteristics of identified mutations were in line with expectations based on previous literature. We then measured the associations between the heteroplasmy of these mutations and the activity of each gene. This required performing Bayesian inference *millions* of times. Applying MCMC this many times is a daunting computational task. This was overcome through engineering a `snakemake` bioinformatics pipeline that can scale repeated analyses seamlessly and run them in parallel on different compute nodes in an HPC environment. Additionally, the use of multiprocessing techniques allowed numerous MCMC runs to be done in parallel, reducing more than a year of sequential compute time to weeks in parallel.

This resulted in posterior distributions over the regression coefficients measuring the association between heteroplasmy and gene activity. Applying the Benjamini-Hochberg procedure to posterior quantiles, as justified in Chapter 4, resulted in lists of high-confidence genes for each mutation's effect, in each brain region. Analysis of the functional roles of these genes suggests that widespread developmental mutations could alter mitochondrial network dynamics, and impinge on normal neuronal functioning and development.

Chapter 6 turned our attention away from mtDNA sequencing to a model of mtDNA motion. This brief chapter described how observations of mtDNA motion with finite frame rates could lead to illusory conclusions about how the speed of mtDNA varies with inter-cristae spacing.

7.2 Closing Remarks

The work in this thesis has centered around mtDNA and its study in single-cells, differentiating it from previous bulk studies. Single-cell omics data is still nascent and much more work needs to be done to use this data to its full potential. The work reported in this thesis exploits this data in novel manner whilst being critical in the application of new methods. As new omics methods are developed, our ability as a research community to gather and analyse data will relentlessly continue into new territories. However, data is not knowledge or understanding. In order to unravel signal from noise and find meaningful, replicable, and reproducible conclusions from our data will require developments in the analytic tools to match the development of sequencing technologies. These developments will ultimately be grounded in *bioinformatics and statistics*.

Glossary

3'/5' Transcriptomics Approaches to transcriptomics that synthesise cDNA from the 3' or alternatively the 5' of a transcript. These approaches typically result in greater depth of coverage within 150 bp of the respective ends of the transcript.

BH (Benjamini-Hochberg Procedure) A statistical method used to control the false discovery rate (FDR) in multiple hypothesis testing.

Breadth of coverage The proportion of bases in a target genome that attained a certain depth of coverage threshold. The breadth of coverage is therefore depth threshold dependent .

CellRanger A software platform offered by 10x genomics to process data output from their broad range of protocols.

De-duplication Input cDNA from many single platforms undergoes PCR amplification to enable the small amount of RNA from individual cells to be read on next generation sequencing platforms. De-duplication is the process by which reads derived from the same starting cDNA molecule are grouped together, and a representative read is chosen. This quantifies how many molecules were present prior to PCR and counteracts PCR amplification bias.

Depth of coverage The number of reads from a sequencing experiment that align to a particular base in the genome. This can be in bulk or broken down per cell.

EB (Empirical Bayes) Empirical Bayes estimates the prior distribution from the data, rather than assuming it is known a priori.

FDR (False Discovery Rate) A statistical method used in multiple hypothesis testing to control or estimate the expected proportion of incorrectly rejected null hypotheses (false positives).

Full length transcriptomics Approaches to transcriptomics that obtain sequence data covering the entire length of a transcript. For instance, the Smart-Seq2 platform.

FWER (Family-Wise Error Rate) The probability of making one or more false discoveries, or Type I errors, when performing multiple hypotheses tests. It is a measure used to control or assess the expected proportion of incorrect rejections among all the hypotheses when multiple comparisons are made.

GLM (Generalised Linear Model) A flexible generalization of ordinary linear regression that allows for response variables to have error distribution models other than a normal distribution. .

HMC (Hamiltonian Monte Carlo) A type of MCMC algorithm that uses concepts from physics to produce efficient proposals for the Markov chain. It leverages the Hamiltonian dynamics to explore the target distribution more efficiently, which can lead to faster convergence compared to traditional MCMC methods.

Long read sequencing Sequencing on machines capable of obtaining sequences from reads in excess of 1,000 Kbp. This includes machines offered by Oxford Nanopore and Pacific Biosciences.

MCMC (Markov-Chain Monte-Carlo) A class of algorithms used to sample from probability distributions based on constructing a Markov chain that has the desired distribution as its equilibrium distribution. MCMC algorithms are used in Bayesian statistics to compute posterior distributions.

mgatk Mitochondrial Genome Analysis Toolkit. A computational tool developed to obtain high quality variants from mtscATAC-seq data for lineage tracing.

mtscATAC-seq mitochondrial single-cell assay for transposase-accessible chromatin with sequencing. A high throughput approach for the genotyping mitochondrial and obtaining chromatin accessability in single cells.

Multiplexing In high throughput single cell approaches, reads are barcoded with a nucleotide sequence to identify which cell they come from. This enables reads from all cells to be pooled together on the same flowcell during sequencing, thereby increasing throughput. This pooling is termed multiplexing.

NB2 (Negative Binomial version 2) A variant of the Negative Binomial distribution, often used in modeling count data where the variance exceeds the mean. It is particularly useful for overdispersed count outcome variables in regression models. NB2 introduces an additional parameter to the Negative Binomial distribution to model the overdispersion directly.

NUMT Nuclear DNA of mitochondrial origin. These homologous sequences can result in errors of heteroplasmy quantification when use short read sequences.

NUTS (No-U-Turn Sampler) An extension of Hamiltonian Monte Carlo designed to choose the optimal number of steps in the trajectory automatically. It improves efficiency by preventing the trajectory from retracing its steps, hence the name"No-U-Turn.".

PCR stutter A technical error occurring during PCR, typically near short tandem repeats, resulting in base pairs randomly being skipped during PCR. This results in results in reads derived from the same transcript having different starting alignment positions .

Probabilistic Programming Language (PPL) A specialized programming framework that facilitates the specification of probabilistic models and automates the process of inference. These languages typically offer built-in mechanisms for specifying prior distributions, likelihood functions, and latent variables, along with advanced algorithms for posterior inference such as Markov Chain Monte Carlo (MCMC), Variational Inference (VI), or Hamiltonian Monte Carlo (HMC)..

scATAC-seq Single cell Assay for Transposase Accessible Chromatin using sequencing. A single cell omics approach for obtaining chromatin accessible regions of the genome, an epigenomic cell state.

scRNA-seq Single cell RNA sequencing. A single cell omics approach for obtaining transcriptional state.

Short read sequencing Sequencing on machines that are typically capable of sequencing reads around 150-300 kbp in length. Illumina short read sequencers are a common platform used for this approach.

TCR T-cell Receptors are proteins used by T-cells for detecting antigens from foreign bodies. Many different TCRs are used by the immune system to detect a broad range of potential pathogens. Roughly 10^{15} possible TCR sequences exist, of which only 10^{11} will be found in an individual human, making it unlikely that two identical TCR sequences will emerge independently in the same person. TCR sequences can therefore be used to lineage trace T-Cells as any two cells sharing this sequence will likely be derived from a common ancestor.

V(D)J Recombination The process by which different TCR sequences are produced during Thymic maturation.

Variant Positions where sequenced reads differ from a reference genome. This is used to infer the presence of mutations.

Appendix A

Supplementary Material For Chapter 4

A.1 Stan Code

Listing A.1: Negative Binomial Regression Model With Heteroplasmy Uncertainty

```
data {
  int<lower=0> n_cells;
  int<lower=0> DF[n_cells];
  int<lower=0> x[n_cells];
  int<lower=0> y[n_cells];
}

parameters {
  real a;
  real b;
  real<lower = 0> phi;
  vector<lower=0, upper=1>[n_cells] H;
}

model {
  phi ~ exponential(2);
  a ~ normal(0,1);
  b ~ normal(0,1);
  x ~ binomial(DF, H);
  for (i in 1:n_cells){
    y[i] ~ neg_binomial_2(exp(a + H[i] * b), phi);
  }
}
```

Appendix B

Supplementary Material For Chapter 6

B.1 Stan Code

Listing B.1: Obtaining Samples For Diffusion In A Box

```
functions {
  real summand(real x, real x0, real t, real L, real D, real n) {
    // real arg1 = n*pi();
    return cos(n*pi()*x/L)*cos(n*pi()*x0/L)*exp(- square(n) * square(pi()) * D * t
   / (square(L)) );
  }
  real finite_diffusion_lpdf(real x, real x0, real t, real L, real D) {
    real pdf = 1;
    for (n in 1:1000) {
      pdf = pdf + 2 * summand(x, x0, t, L, D, n) ;
    }
    return log(pdf); // return the logarithm of the pdf
  }
}
data {
  real x0;
  real t;
  real L;
  real D;
}
parameters {
  real<lower=0, upper=L> x;
}
model {
  target += finite_diffusion_lpdf( x | x0, t, L, D);
}
```

Listing B.2: Python code to generate samples from the reflected gaussian distribution

```python
import numpy as np

def reflected_gaussian(mean, std_dev, lower_bound, upper_bound, num_samples):
    samples = np.random.normal(mean, std_dev, num_samples)

    for i in range(num_samples):
        while samples[i] < lower_bound or samples[i] > upper_bound:
            if samples[i] < lower_bound:
                samples[i] = lower_bound + (lower_bound - samples[i])
            elif samples[i] > upper_bound:
                samples[i] = upper_bound - (samples[i] - upper_bound)

    return samples
```

Bibliography

[1] P. Siekevitz, "Powerhouse of the Cell," *Scientific American*, vol. 197, no. 1, pp. 131–144, 1957.

[2] B. Alberts, *Molecular Biology of the Cell*. W.W. Norton, 2017.

[3] J. P. Abrahams, A. G. W. Leslie, R. Lutter *et al.*, "Structure at 2.8 Åresolution of F1-ATPase from bovine heart mitochondria," *Nature*, vol. 370, p. 621–628, 1994.

[4] Y. Hatefi, "The Mitochondrial Electron Transport And Oxidative Phosphorylation System," *Annual Review of Biochemistry*, vol. 54, no. 1, pp. 1015–1069, 1985.

[5] X. Li, P. Fang, J. Mai *et al.*, "Targeting mitochondrial reactive oxygen species as novel therapy for inflammatory diseases and cancers," *Journal of Hematology and Oncology*, vol. 6, 2013.

[6] H. M. McBride, M. Neuspiel, and S. Wasiak, "Mitochondria: More Than Just a Powerhouse," *Current Biology*, vol. 16, 2006.

[7] G. Hajnóczky, G. Csordás, S. Das *et al.*, "Mitochondrial calcium signalling and cell death: Approaches for assessing the role of mitochondrial Ca2+ uptake in apoptosis," *Cell Calcium*, vol. 40, pp. 553–560, 2006.

[8] C. M. Klinge, "Estrogenic control of mitochondrial function and biogenesis," *Journal of Cellular Biochemistry*, vol. 105, pp. 1342–1351, 12 2008.

[9] S. Victorelli, H. Salmonowicz, J. Chapman *et al.*, "Apoptotic stress causes mtDNA release during senescence and drives the SASP," *Nature*, vol. 622, pp. 627–636, 2023.

[10] L. Sagan, "On the origin of mitosing cells," *Journal of Theoretical Biology*, vol. 14, no. 3, pp. 225–IN6, 1967.

[11] W. F. Martin, S. Garg, and V. Zimorski, "Endosymbiotic theories for eukaryote origin," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 370, 2015.

[12] M. W. Gray, G. Burger, and B. F. Lang, "Mitochondrial Evolution," *Science*, vol. 283, no. 5407, pp. 1476–1481, 1999.

[13] W. Wei, K. R. Schon, G. Elgar *et al.*, "Nuclear-embedded mitochondrial DNA sequences in 66,083 human genomes," *Nature*, vol. 611, pp. 105–114, 2022.

[14] P. F. Chinnery, "Mitochondrial disease in adults: what's old and what's new?" *EMBO Molecular Medicine*, vol. 7, no. 12, pp. 1503–1512, 2015.

[15] P. Kaufmann, K. Engelstad, Y. Wei *et al.*, "Natural history of MELAS associated with mitochondrial DNA m.3243A>G genotype," *Neurology*, vol. 77, no. 22, pp. 1965–1971, 2011.

[16] P. Y. W. Man, P. G. Griffiths, D. T. Brown *et al.*, "The Epidemiology of Leber Hereditary Optic Neuropathy in the North East of England," *Am. J. Hum. Genet*, vol. 72, pp. 333–339, 2003.

[17] P. F. Chinnery, N. Howell, R. N. Lightowlers *et al.*, "Molecular pathology of MELAS and MERRF. The relationship between mutation load and clinical phenotypes." *Brain*, vol. 120, no. 10, pp. 1713–1721, 1997.

[18] P. F. Chinnery, R. M. Andrews, D. M. Turnbull *et al.*, "Leber hereditary optic neuropathy: Does heteroplasmy influence the inheritance and expression of the G11778A mitochondrial DNA mutation?" *American Journal of Medical Genetics*, vol. 98, no. 3, pp. 235–243, 2001.

[19] R. Rossignol, B. Faustin, C. Rocher *et al.*, "Mitochondrial threshold effects," *Biochemical Journal*, vol. 370, no. 3, pp. 751–762, 2003.

[20] A. Regev, S. A. Teichmann *et al.*, "The human cell atlas," *eLife*, vol. 6, pp. 1–30, 2017.

[21] "Best practices for single-cell analysis across modalities," *Nature Reviews Genetics*, vol. 24, pp. 550–572, 2023.

[22] M. D. Luecken and F. J. Theis, "Current best practices in single-cell RNA-seq analysis: a tutorial," *Molecular Systems Biology*, vol. 15, no. 6, 2019.

[23] B. Carpenter, A. Gelman, M. D. Hoffman *et al.*, "Stan: A Probabilistic Programming Language," *Journal of Statistical Software*, vol. 76, no. 1, p. 1–32, 2017.

[24] A. Gelman, A. Vehtari, D. Simpson *et al.*, "Bayesian Workflow," *arxiv*, 2020. [Online]. Available: https://doi.org/10.48550/arXiv.2011.01808

[25] Tabula Muris Consortium, "A single-cell transcriptomic atlas characterizes ageing tissues in the mouse." *Nature*, vol. 583, no. June 2019, 2020.

[26] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[27] A. S. Marshall and N. S. Jones, "Discovering Cellular Mitochondrial Heteroplasmy Heterogeneity with Single Cell RNA and ATAC Sequencing," *Biology*, vol. 10, no. 6, p. 503, Jun. 2021.

[28] G. S. Gorman, P. F. Chinnery, S. DiMauro *et al.*, "Mitochondrial diseases," *Nature Reviews Disease Primers*, vol. 2, pp. 1–23, 2016.

[29] B. A. Payne, I. J. Wilson, P. Yu-Wai-Man *et al.*, "Universal heteroplasmy of human mitochondrial DNA," *Human Molecular Genetics*, vol. 22, no. 2, pp. 384–390, 2013.

[30] S. R. Kennedy, J. J. Salk, M. W. Schmitt *et al.*, "Ultra-Sensitive Sequencing Reveals an Age-Related Increase in Somatic Mitochondrial Mutations That Are Inconsistent with Oxidative Damage," *PLoS Genetics*, vol. 9, no. 9, 2013.

[31] J. B. Stewart and P. F. Chinnery, "The dynamics of mitochondrial DNA heteroplasmy: Implications for human health and disease," *Nature Reviews Genetics*, vol. 16, no. 9, pp. 530–542, 2015.

[32] T. E. Kauppila, J. H. Kauppila, and N. G. Larsson, "Mammalian Mitochondria and Aging: An Update," *Cell Metabolism*, vol. 25, no. 1, pp. 57–71, 2017.

[33] J. Aryaman, I. G. Johnston, and N. S. Jones, "Mitochondrial heterogeneity," *Frontiers in Genetics*, vol. 10, pp. 1–16, 2019.

[34] J. B. Stewart and P. F. Chinnery, "Extreme heterogeneity of human mitochondrial DNA from organelles to populations," *Nature Reviews Genetics*, vol. 22, no. 2, pp. 106–118, 2021.

[35] H. R. Elliott, D. C. Samuels, J. A. Eden *et al.*, "Pathogenic Mitochondrial DNA Mutations Are Common in the General Population," *American Journal of Human Genetics*, vol. 83, no. 2, pp. 254–260, 2008.

[36] M. A. Walker, C. A. Lareau, L. S. Ludwig *et al.*, "Purifying Selection against Pathogenic Mitochondrial DNA in Human T Cells," *New England Journal of Medicine*, vol. 383, no. 16, pp. 1556–1563, 2020.

[37] N. Cloonan, A. R. Forrest, G. Kolle *et al.*, "Stem cell transcriptome profiling via massive-scale mRNA sequencing," *Nature Methods*, vol. 5, no. 7, pp. 613–619, 2008.

[38] K. Kretzschmar and F. M. Watt, "Lineage tracing," *Cell*, vol. 148, no. 1-2, pp. 33–45, 2012.

[39] E. Nekhaeva, N. D. Bodyak, Y. Kraytsberg *et al.*, "Clonally expanded mtDNA point mutations are abundant in individual cells of human tissues," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 99, no. 8, pp. 5521–5526, 2002.

[40] R. Stark, M. Grzelak, and J. Hadfield, "RNA sequencing: the teenage years," *Nature Reviews Genetics*, vol. 20, p. 631–656, 2019.

[41] B. Hwang, J. H. Lee, and D. Bang, "Single-cell RNA sequencing technologies and bioinformatics pipelines," *Experimental & Molecular Medicine*, vol. 50, pp. 1–14, 2018.

[42] L. Minnoye, G. K. Marinov *et al.*, "Chromatin accessibility profiling methods," *Nat. Rev. Methods Primers*, vol. 1, no. 1, 2021.

[43] S. Linnarsson and S. A. Teichmann, "Single-cell genomics: Coming of age," *Genome Biology*, vol. 17, no. 1, pp. 16–18, 2016.

[44] J. Lee, D. Y. Hyeon, and D. Hwang, "Single-cell multiomics: technologies and data analysis methods," *Experimental and Molecular Medicine*, vol. 52, no. 9, pp. 1428–1442, 2020.

[45] F. Tang, C. Barbacioru, Y. Wang *et al.*, "mRNA-Seq whole-transcriptome analysis of a single cell," *Nature Methods*, vol. 6, no. 5, pp. 377–382, 2009.

[46] G. X. Zheng, J. M. Terry, P. Belgrader *et al.*, "Massively parallel digital transcriptional profiling of single cells," *Nature Communications*, vol. 8, 2017.

[47] V. Svensson, R. Vento-Tormo, and S. A. Teichmann, "Exponential scaling of single-cell RNA-seq in the past decade," *Nature Protocols*, vol. 13, no. 4, pp. 599–604, 2018.

[48] A. T. Satpathy, J. M. Granja, K. E. Yost *et al.*, "Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion," *Nature Biotechnology*, vol. 37, 2019.

[49] N. Rajewsky *et al.*, "LifeTime and improving European healthcare through cell-based interceptive medicine," *Nature*, vol. 587, pp. 377–386, 2020.

[50] T. Gomes, S. A. Teichmann, and C. Talavera-López, "Immunology Driven by Large-Scale Single-Cell Sequencing," *Trends in Immunology*, vol. 40, no. 11, pp. 1011–1021, 2019.

[51] X. Ren, B. Kang, and Z. Zhang, "Understanding tumor ecosystems by single-cell sequencing : promises and limitations," *Genome Biol*, pp. 1–14, 2018.

[52] B. Lim, Y. Lin, and N. Navin, "Review Advancing Cancer Research and Medicine with Single-Cell Genomics," *Cancer Cell*, vol. 37, no. 4, pp. 456–470, 2020.

[53] M. A. Diroma, A. S. Varvara, M. Attimonelli *et al.*, "Investigating human mitochondrial genomes in single cells," *Genes*, vol. 11, no. 5, 2020.

[54] M. Safra, A. Sas-Chen, R. Nir *et al.*, "The m1A landscape on cytosolic and mitochondrial mRNA at single-base resolution," *Nature*, vol. 551, no. 7679, pp. 251–255, 2017.

[55] D. Sims, I. Sudbery, N. E. Ilott *et al.*, "Sequencing depth and coverage: Key considerations in genomic analyses," *Nature Reviews Genetics*, vol. 15, no. 2, pp. 121–132, 2014.

[56] A. P. Voigt, K. Mulfaul, N. K. Mullin *et al.*, "Single-cell transcriptomics of the human retinal pigment epithelium and choroid in health and macular degeneration," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 116, no. 48, pp. 24 100–24 107, 2019.

[57] M. Enge, H. E. Arda, M. Mignardi *et al.*, "Single-Cell Analysis of Human Pancreas Reveals Transcriptional Signatures of Aging and Somatic Mutation Patterns," *Cell*, vol. 171, no. 2, pp. 321–330.e14, 2017.

[58] S. Picelli, O. R. Faridani, Å. K. Björklund *et al.*, "Full-length RNA-seq from single cells using Smart-seq2," *Nature Protocols*, vol. 9, no. 1, pp. 171–181, 2014.

[59] M. D. Cesare, Y. Wang, P. Piazza *et al.*, "Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis," *F1000Res*, vol. 6, no. 100, 2017.

[60] A. M. Klein, L. Mazutis, I. Akartuna *et al.*, "Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells," *Cell*, vol. 161, no. 5, pp. 1187–1201, 2015.

[61] E. Z. Macosko, A. Basu, R. Satija *et al.*, "Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets," *Cell*, vol. 161, no. 5, pp. 1202–1214, 2015.

[62] S. Islam, A. Zeisel, S. Joost *et al.*, "Quantitative single-cell RNA-seq with unique molecular identifiers," *Nature Methods*, vol. 11, no. 2, pp. 163–166, 2014.

[63] A. Dobin, C. A. Davis, F. Schlesinger *et al.*, "STAR: Ultrafast universal RNA-seq aligner," *Bioinformatics*, vol. 29, no. 1, pp. 15–21, 2013.

[64] W. Chen, Y. Li, J. Easton *et al.*, "UMI-count modeling and differential expression analysis for single-cell RNA sequencing," *Genome Biology*, vol. 19, no. 1, pp. 1–17, 2018.

[65] T. Smith, A. Heger, and I. Sudbery, "UMI-tools: Modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy," *Genome Research*, vol. 27, no. 3, pp. 491–499, 2017.

[66] K. Lebrigand, V. Magnone, P. Barbry *et al.*, "High throughput error corrected Nanopore single cell transcriptome sequencing," *Nature Communications*, vol. 11, no. 1, pp. 1–8, 2020.

[67] M. Philpott, J. Watson, A. Thakurta *et al.*, "Highly accurate barcode and UMI error correction using dual nucleotide dimer blocks allows direct single-cell nanopore transcriptome sequencing," *bioRxiv*, 2021. [Online]. Available: https://doi.org/10.1101/2021.01.18.427145

[68] R. Volden and C. Vollmers, "Highly Multiplexed Single-Cell Full-Length cDNA Sequencing of human immune cells with 10X Genomics and R2C2," *bioRxiv*, 2020. [Online]. Available: https://doi.org/10.1101/2020.01.10.902361

[69] M. Hagemann-Jensen, C. Ziegenhain, P. Chen *et al.*, "Single-cell RNA counting at allele and isoform resolution using Smart-seq3," *Nature Biotechnology*, vol. 38, no. 6, pp. 708–714, 2020.

[70] J. Baran-Gale, T. Chandra, and K. Kirschner, "Experimental design for single-cell RNA sequencing," *Briefings in Functional Genomics*, vol. 17, no. 4, pp. 233–239, 2018.

[71] C. A. Lareau, F. M. Duarte, J. G. Chew *et al.*, "Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility," *Nature Biotechnology*, vol. 37, p. 916–924, 2019.

[72] R. Poplin, V. Ruano-Rubio, M. A. DePristo *et al.*, "Scaling accurate genetic variant discovery to tens of thousands of samples," *bioRxiv*, 2018. [Online]. Available: https://doi.org/10.1101/201178

[73] T. Zhu, K. Liao, R. Zhou *et al.*, "ATAC-seq with unique molecular identifiers improves quantification and footprinting," *Commun. Biol.*, vol. 3, no. 675, 2020.

[74] R. Desai, D. A. East, L. Hardy *et al.*, "Mitochondria form contact sites with the nucleus to couple prosurvival retrograde response," *Science Advances*, vol. 6, no. 51, pp. 1–16, 2020.

[75] R. W. Taylor, G. A. Taylor, S. E. Durham *et al.*, "The determination of complete human mitochondrial DNA sequences in single cells: Implications for the study of somatic mitochondrial DNA point mutations," *Nucleic Acids Research*, vol. 29, no. 15, pp. 1–8, 2001.

[76] E. Perales-Clemente, A. N. Cook, J. M. Evans *et al.*, "Natural underlying mt DNA heteroplasmy as a potential source of intra-person hi PSC variability," *The EMBO Journal*, vol. 35, no. 18, pp. 1979–1990, 2016.

[77] L. S. Ludwig, C. A. Lareau, J. C. Ulirsch *et al.*, "Lineage Tracing in Humans Enabled by Mitochondrial Mutations and Single-Cell Genomics," *Cell*, vol. 176, no. 6, pp. 1325–1339.e22, 2019.

[78] J. D. Buenrostro, B. Wu, H. Y. Chang *et al.*, "ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide," *Current Protocols in Molecular Biology*, vol. 109, 2015.

[79] C. Zheng, L. Zheng, J. K. Yoo *et al.*, "Landscape of Infiltrating T Cells in Liver Cancer Revealed by Single-Cell Sequencing," *Cell*, vol. 169, no. 7, 2017.

[80] X. Guo, Y. Zhang, L. Zheng *et al.*, "Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing," *Nature Medicine*, vol. 24, no. 7, pp. 978–985, 2018.

[81] A. Giustacchini, S. Thongjuea, N. Barkas *et al.*, "Single-cell transcriptomics uncovers distinct molecular signatures of stem cells in chronic myeloid leukemia," *Nature Medicine*, vol. 23, no. 6, pp. 692–702, 2017.

[82] J. Xu, K. Nuno, U. M. Litzenburger *et al.*, "Single-cell lineage tracing by endogenous mutations enriched in transposase accessible mitochondrial DNA," *bioRxiv*, 2019. [Online]. Available: https://doi.org/10.1101/480202

[83] M. R. Corces-Zimmerman, W. J. Hong, I. L. Weissman *et al.*, "Preleukemic mutations in human acute myeloid leukemia affect epigenetic regulators and persist in remission," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 111, no. 7, pp. 2548–2553, 2014.

[84] C. A. Lareau, L. S. Ludwig, C. Muus *et al.*, "Massively parallel single-cell mitochondrial DNA genotyping and chromatin profiling," *Nature Biotechnology*, vol. 39, no. 4, pp. 451–461, 2021.

[85] J. M. Shoffner, M. T. Lott, A. M. Lezza *et al.*, "Myoclonic epilepsy and ragged-red fiber disease (MERRF) is associated with a mitochondrial DNA tRNA (Lys) mutation," *Cell*, vol. 61, no. 6, pp. 931–937, 1990.

[86] P. F. Chinnery, N. Howell, R. N. Lightowlers *et al.*, "MELAS and MERRF. The relationship between maternal mutation load and the frequency of clinically affected offspring," *Brain*, vol. 121, no. 10, pp. 1889–1894, 1998.

[87] T. E. Miller, C. A. Lareau, J. A. Verga *et al.*, "Mitochondrial variant enrichment from high-throughput single-cell RNA-seq resolves clonal populations," *bioRxiv*, 2021. [Online]. Available: https://doi.org/10.1101/2021.03.08.434450

[88] H. Medini, T. Cohen, and D. Mishmar, "Mitochondrial gene expression in single cells shape pancreatic beta cells' sub-populations and explain variation in insulin pathway," *Scientific Reports*, vol. 11, 2021.

[89] J. A. Sena, G. Galotto, N. P. Devitt *et al.*, "Unique Molecular Identifiers reveal a novel sequencing artefact with implications for RNA-Seq based gene expression analysis," *Scientific Reports*, vol. 8, no. 1, pp. 1–13, 2018.

[90] S. Schwartz, R. Oren, and G. Ast, "Detection and removal of biases in the analysis of next-generation sequencing reads," *PLoS ONE*, vol. 6, no. 1, 2011.

[91] N. Konari, K. Nagaishi, S. Kikuchi *et al.*, "Mitochondria transfer from mesenchymal stem cells structurally and functionally repairs renal proximal tubular epithelial cells in diabetic nephropathy in vivo," *Scientific Reports*, vol. 9, no. 1, pp. 1–14, 2019.

[92] D. Liu, Y. Gao, J. Liu *et al.*, "Intercellular mitochondrial transfer as a means of tissue revitalization," *Signal Transduction and Targeted Therapy*, vol. 6, no. 1, 2021.

[93] J. R. Brestoff, C. B. Wilen, J. R. Moley *et al.*, "Intercellular Mitochondria Transfer to Macrophages Regulates White Adipose Tissue Homeostasis and Is Impaired in Obesity," *Cell Metabolism*, vol. 33, 2021.

[94] A. T. Satpathy, N. Saligrama, J. D. Buenrostro *et al.*, "Transcript-indexed ATAC-seq for precision immune profiling," *Nature Medicine*, vol. 24, no. 5, pp. 580–590, 2018.

[95] H. M. Kang, M. Subramaniam, S. Targ *et al.*, "Multiplexed droplet single-cell RNA-sequencing using natural genetic variation," *Nature Biotechnology*, vol. 36, no. 1, pp. 89–94, 2018.

[96] S. L. Wolock, R. Lopez, and A. M. Klein, "Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data," *Cell Systems*, vol. 8, no. 4, 2019.

[97] S. Yang, S. E. Corbett, Y. Koga *et al.*, "Decontamination of ambient RNA in single-cell RNA-seq with DecontX," *bioRxiv*, pp. 1–15, 2019. [Online]. Available: /doi.org/10.1101/704015

[98] M. D. Young and S. Behjati, "SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data," *GigaScience*, vol. 9, no. 12, pp. 1–10, 2020.

[99] "Scanpy - Clustering Tutorial," https://scanpy-tutorials.readthedocs.io/en/latest/pbmc3k.html, accessed: 29-04-2021.

[100] "Seurat - Guided Clustering Tutorial," https://satijalab.org/seurat/archive/v2.4/pbmc3k_tutorial.html, accessed: 29-04-2021.

[101] T. Ilicic, J. K. Kim, A. A. Kolodziejczyk *et al.*, "Classification of low quality cells from single-cell RNA-seq data," *Genome Biology*, vol. 17, no. 1, pp. 1–15, 2016.

[102] T. R. Mercer, S. Neph, M. E. Dinger *et al.*, "The human mitochondrial transcriptome," *Cell*, vol. 146, no. 4, pp. 645–658, 2011.

[103] D. Osorio and J. J. Cai, "Systematic determination of the mitochondrial proportion in human and mice tissues for single-cell RNA-sequencing data quality control," *Bioinformatics*, pp. 1–5, 2020.

[104] A. A. AlJanahi, M. Danielsen, and C. E. Dunbar, "An Introduction to the Analysis of Single-Cell RNA-Sequencing Data," *Molecular Therapy - Methods and Clinical Development*, vol. 10, pp. 189–196, 2018.

[105] G. C. Kujoth, C. Leeuwenburgh, and T. A. Prolla, "Mitochondrial DNA mutations and apoptosis in mammalian aging," *Cancer Research*, vol. 66, no. 15, pp. 7386–7389, 2006.

[106] S. Márquez-Jurado, J. Díaz-Colunga, R. P. Das Neves *et al.*, "Mitochondrial levels determine variability in cell death by modulating apoptotic gene expression," *Nature Communications*, vol. 9, no. 1, 2018.

[107] K. Auré, G. Fayet, J. P. Leroy *et al.*, "Apoptosis in mitochondrial myopathies is linked to mitochondrial proliferation," *Brain*, vol. 129, no. 5, pp. 1249–1259, 2006.

[108] H. Mei, S. Sun, Y. Bai *et al.*, "Reduced mtDNA copy number increases the sensitivity of tumor cells to chemotherapeutic drugs," *Cell Death and Disease*, vol. 6, no. 4, 2015.

[109] F. M. Calabrese, D. L. Balacco, R. Preste *et al.*, "NumtS colonization in mammalian genomes," *Scientific Reports*, vol. 7, no. 1, pp. 1–10, 2017.

[110] M. P. Cristiano, D. C. Cardoso, and T. M. Fernandes-Salomão, "Could pseudogenes be widespread in ants? Evidence of numts in the leafcutter ant Acromyrmex striatus (Roger, 1863) (Formicidae: Attini)," *Comptes Rendus - Biologies*, vol. 337, no. 2, pp. 78–85, 2014.

[111] H. Maude, M. Davidson, N. Charitakis *et al.*, "NUMT confounding biases mitochondrial heteroplasmy calls in favor of the reference allele," *Frontiers in Cell and Developmental Biology*, vol. 7, no. SEP, pp. 1–7, 2019.

[112] A. A. Petti, S. R. Williams, C. A. Miller *et al.*, "A general approach for detecting expressed mutations in AML cells using single cell RNA-sequencing," *Nature Communications*, vol. 10, no. 1, 2019.

[113] D. o. E. United Nations and P. D. Social Affairs, "World Population Prospects 2022: Summary of Results," United Nations, New York, NY, Tech. Rep., 2022.

[114] Sarah Harper, "Economic and social implications of aging societies," *Science*, vol. 346, no. 6209, pp. 583–587, 2014.

[115] C. López-Otín, M. A. Blasco, L. Partridge *et al.*, "The hallmarks of aging," *Cell*, vol. 153, no. 6, p. 1194, 2013.

[116] D. Harman, "The free radical theory of aging: Effect of age on serum copper levels," *Journal of Gerontology*, vol. 20, pp. 151–153, 1965.

[117] M. F. Alexeyev, "Is there more to aging than mitochondrial DNA and reactive oxygen species?" *FEBS Journal*, vol. 276, no. 20, pp. 5768–5787, 2009.

[118] K. N. Lewis, B. Andziak, T. Yang *et al.*, "The naked mole-rat response to oxidative stress: Just deal with it," *Antioxidants and Redox Signaling*, vol. 19, no. 12, pp. 1388–1399, 2013.

[119] X. G. Lei, J. H. Zhu, W. H. Cheng *et al.*, "Paradoxical roles of antioxidant enzymes: Basic mechanisms and health implications," *Physiological Reviews*, vol. 96, no. 1, pp. 307–364, 2015.

[120] A. Trifunovic, A. Wredenberg, M. Falkenberg *et al.*, "Premature ageing in mice expressing defective mitochondrial DNA polymerase," *Nature*, vol. 429, pp. 417–423, 2004.

[121] A. Trifunovic, A. Hansson, A. Wredenberg *et al.*, "Somatic mtDNA mutations cause aging phenotypes without affecting reactive oxygen species production," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, no. 50, pp. 17 993–17 998, 2005.

[122] B. Arbeithuber, J. Hester, M. A. Cremona *et al.*, "Age-related accumulation of de novo mitochondrial mutations in mammalian oocytes and somatic tissues," *PLoS Biology*, vol. 18, no. 7, pp. 1–37, 2020.

[123] L. C. Greaves, M. Nooteboom, J. L. Elson *et al.*, "Clonal expansion of early to mid-life mitochondrial DNA point mutations drives mitochondrial dysfunction during human ageing," *PLoS Genetics*, vol. 10, no. 9, pp. 1–10, 2014.

[124] M. Pinto and C. T. Moraes, "Mechanisms linking mtDNA damage and aging," *Free Radical Biology and Medicine*, vol. 85, pp. 250–258, 2015.

[125] P. A. P. Moran, "Random processes in genetics," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 54, no. 1, p. 60–71, 1958.

[126] T. Barrett, S. E. Wilhite, P. Ledoux *et al.*, "NCBI GEO: Archive for functional genomics data sets - Update," *Nucleic Acids Research*, vol. 41, pp. 991–995, 2013.

[127] R. Durrett, *Probability Models for DNA Sequence Evolution*. Springer, 2002.

[128] Z. Y. Tam, J. Gruber, B. Halliwell, and R. Gunawan, "Context-dependent Role of mitochondrial fusion-fission in clonal expansion of mtDNA mutations," *PLOS Computational Biology*, vol. 11, no. 5, pp. 1–22, 2015.

[129] A. M. D'Erchia, A. Atlante, G. Gadaleta *et al.*, "Tissue-specific mtDNA abundance from exome data and its correlation with mitochondrial transcription, mass and respiratory activity." *Mitochondrion*, vol. 20, pp. 13–21, 2015.

[130] W. Zheng, K. Khrapko, H. A. Coller *et al.*, "Origins of human mitochondrial point mutations as DNA polymerase — gamma mediated errors," vol. 599, pp. 11–20, 2006.

[131] M. J. Muraro, G. Dharmadhikari, D. Grün *et al.*, "A single-cell transcriptome atlas of the human pancreas," *Cell Systems*, vol. 3, no. 4, 2016.

[132] S. Ma, S. Sun, L. Geng *et al.*, "Caloric restriction reprograms the single-cell transcriptional landscape of rattus norvegicus aging," *Cell*, vol. 180, no. 5, 2020.

[133] C. Calabrese, D. Simone, M. A. Diroma *et al.*, "MToolBox: A highly automated pipeline for heteroplasmy annotation and prioritization analysis of human mitochondrial variants in high-throughput sequencing," *Bioinformatics*, vol. 30, no. 21, pp. 3115–3117, 2014.

[134] D. Clark, N. Pazdernik, and M. McGehee, *Molecular Biology*. Elsevier Science, 2018.

[135] P. J. Cock, C. J. Fields, N. Goto *et al.*, "The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants," *Nucleic Acids Research*, vol. 38, pp. 1767–1771, 2009.

[136] B. Kaminow, D. Yunusov, and A. Dobin, "STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data," *bioRxiv*, 2021. [Online]. Available: https://doi.org/10.1101/2021.05.05.442755

[137] P. Danecek, J. K. Bonfield, J. Liddle *et al.*, "Twelve years of SAMtools and BCFtools," *GigaScience*, vol. 10, no. 2, 2021.

[138] A. Green, "Mitochondrial DNA mutations in single-cell ageing and health," Ph.D. dissertation, Imperial College London, 2022.

[139] E. Tudurí, S. Soriano, L. Almagro *et al.*, "The pancreatic β -cell in ageing : Implications in age-related diabetes," *Ageing Research Reviews*, vol. 80, p. 101674, 2022.

[140] L. S. Lim, P. Mitchell, J. M. Seddon *et al.*, "Age-related macular degeneration," *The Lancet*, vol. 379, no. 9827, pp. 1728–1738, 2012.

[141] R. Arrojo, V. Lev-ram, S. Tyagi *et al.*, "Age mosaicism across multiple scales in adult tissues," *Cell Metabolism*, vol. 30, no. 2, pp. 343–351, 2019.

[142] L. Fontana and L. Partridge, "Promoting health and longevity through diet: From model organisms to humans," *Cell*, vol. 161, no. 1, pp. 106–118, 2015.

[143] K. Townsend and Y.-h. Tseng, "Brown adipose tissue," *Adipocyte*, vol. 1, no. 1, pp. 13–24, 2012.

[144] A. Picca, V. Pesce, F. Fracasso *et al.*, "A comparison among the tissue-specific effects of aging and calorie restriction on TFAM amount and TFAM-binding activity to mtDNA in rat," *Biochimica et Biophysica Acta - General Subjects*, vol. 1840, no. 7, pp. 2184–2191, 2014.

[145] A. P. Green, F. Klimm, A. S. Marshall *et al.*, "Cryptic mitochondrial ageing coincides with mid-late life and is pathophysiologically informative in single cells across tissues and species," *bioRxiv*, 2023. [Online]. Available: https://doi.org/10.1101/2023.07.04.547509

[146] D. S. Kerby, "The simple difference formula: an approach to teaching nonparametric correlation," *Comprehensive Psychology*, vol. 3, 2014.

[147] F. Reinecke, J. A. Smeitink, and F. H. Van Der Westhuizen, "OXPHOS gene expression and control in mitochondrial disorders," *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, vol. 1792, no. 12, pp. 1113–1121, 2009.

[148] H. Mi, D. Ebert, A. Muruganujan *et al.*, "PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API," *Nucleic Acids Research*, vol. 49, no. D1, 2021.

[149] T. Deuse, X. Hu, S. Agbor-Enoh *et al.*, "De novo mutations in mitochondrial DNA of iPSCs produce immunogenic neoepitopes in mice and humans," *Nature Biotechnology*, vol. 37, no. 10, pp. 1137–1144, 2019.

[150] M. Tigano, D. C. Vargas, S. Tremblay-Belzile *et al.*, "Nuclear sensing of breaks in mitochondrial DNA enhances immune surveillance," *Nature*, pp. 1–5, 2021.

[151] D. Matheoud, A. Sugiura, A. Bellemare-Pelletier *et al.*, "Parkinson's disease-related proteins PINK1 and Parkin repress mitochondrial antigen presentation," *Cell*, vol. 166, no. 2, pp. 314–327, 2016.

[152] G. Rasch, *Probabilistic Models for Some Intelligence and Attainment Tests*, ser. Studies in mathematical psychology. Danmarks Paedagogiske Institut, 1960.

[153] A. Gelman, J. Carlin, H. Stern *et al.*, *Bayesian Data Analysis, Third Edition*, ser. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2013.

[154] R. McElreath, *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*, ser. A Chapman & Hall book. CRC Press, 2020.

[155] D. J. Schad, M. Betancourt, and S. Vasishth, "Toward a principled Bayesian workflow in cognitive science," *arxiv*, 2020. [Online]. Available: https://doi.org/10.48550/arXiv.1904.12765

[156] S. Brooks, A. Gelman, G. Jones *et al.*, *Handbook of Markov Chain Monte Carlo*, ser. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, 2011.

[157] D. MacKay, D. Kay, and C. U. Press, *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.

[158] C. Bishop, *Pattern Recognition and Machine Learning*, ser. Information Science and Statistics. Springer, 2006.

[159] J. Blitzstein and J. Hwang, *Introduction to Probability, Second Edition*, ser. Chapman & Hall/CRC Texts in Statistical Science. CRC Press, 2019.

[160] M. Betancourt, "A Conceptual Introduction to Hamiltonian Monte Carlo," *arxiv*, 2017. [Online]. Available: https://doi.org/10.48550/arXiv.1701.02434

[161] M. J. Betancourt and M. Girolami, "Hamiltonian Monte Carlo for Hierarchical Models," *arxiv*, 2013.

[162] A. Gelman and D. B. Rubin, "Inference from iterative simulation using multiple sequences," *Statistical Science*, vol. 7, no. 4, pp. 457 – 472, 1992.

[163] T. Stuart, A. Srivastava, S. Madad *et al.*, "Single-cell chromatin state analysis with Signac," *Nature Methods*, vol. 18, pp. 1333–1341, 2021.

[164] S. Ross, *Introduction to Probability Models*, ser. Introduction to Probability Models. Elsevier Science, 2010.

[165] J. Hilbe, *Negative Binomial Regression*. Cambridge University Press, 2011.

[166] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biology*, vol. 15, 12 2014.

[167] D. Risso, F. Perraudeau, S. Gribkova *et al.*, "A general and flexible method for signal extraction from single-cell RNA-seq data," *Nature Communications*, vol. 9, 12 2018.

[168] W. Tang, F. Bertaux, P. Thomas *et al.*, "BayNorm: Bayesian gene expression recovery, imputation and normalization for single-cell RNA-sequencing data," *Bioinformatics*, vol. 36, pp. 1174–1181, 2020.

[169] S. R. Cook, A. Gelman, and D. B. Rubin, "Validation of software for Bayesian models using posterior quantiles," *Journal of Computational and Graphical Statistics*, vol. 15, no. 3, pp. 675–692, 2006.

[170] S. Talts, M. Betancourt, D. Simpson *et al.*, "Validating Bayesian inference algorithms with simulation-based calibration," *arxiv*, 2020. [Online]. Available: https://doi.org/10.48550/arXiv.1804.06788

[171] M. Modrák, A. H. Moon, S. Kim *et al.*, "Simulation-based calibration checking for Bayesian computation: The choice of test quantities shapes sensitivity," *Bayesian Analysis*, vol. 1, no. 1, Jan. 2023.

[172] J. H. Andrew Gelman and M. Yajima, "Why we (usually) don't have to worry about multiple comparisons," *Journal of Research on Educational Effectiveness*, vol. 5, no. 2, pp. 189–211, 2012.

[173] G. Casella and R. Berger, *Statistical Inference*. Cengage Learning, 2021.

[174] B. Efron and T. Hastie, *Computer Age Statistical Inference*, ser. Institute of Mathematical Statistics Monographs. Cambridge University Press, 2016.

[175] J. D. Storey, "The positive false discovery rate: a Bayesian interpretation and the q-value," *The Annals of Statistics*, vol. 31, no. 6, pp. 2013 – 2035, 2003.

[176] K. P. Murphy, *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023.

[177] Y. Yao, A. Vehtari, D. Simpson *et al.*, "Yes, but did It Work?: Evaluating variational inference," *arxiv*, 2018. [Online]. Available: https://doi.org/10.48550/arXiv.1802.02538

[178] J. Coxhead, M. Kurzawa-Akanbi, R. Hussain *et al.*, "Somatic mtDNA variation is an important component of Parkinson's disease," *Neurobiology of Aging*, vol. 38, 2015.

[179] A. M. J. T. Q. H. Fan Liu, Jianfei Lu, "Mitochondria in ischemic stroke: New insight and implications," *Aging and disease*, vol. 9, no. 5, pp. 924–937, 2018.

[180] P. Norat, S. Soldozy, J. D. Sokolowski *et al.*, "Mitochondrial dysfunction in neurological disorders: Exploring mitochondrial transplantation," *npj Regenerative Medicine*, vol. 5, 2020.

[181] Y. Nie, A. Murley, Z. Golder *et al.*, "Heteroplasmic mitochondrial DNA mutations in frontotemporal lobar degeneration," *Acta Neuropathologica*, vol. 143, pp. 687–695, 6 2022.

[182] A. Bender, K. J. Krishnan, C. M. Morris *et al.*, "High levels of mitochondrial DNA deletions in substantia nigra neurons in aging and Parkinson disease," *Nature Genetics*, vol. 38, pp. 515–517, 2006.

[183] F. Mölder, K. P. Jablonski, B. Letcher *et al.*, "Sustainable data analysis with Snakemake," *F1000Research*, vol. 10, p. 33, 2021.

[184] *Anaconda Software Distribution.* Anaconda Inc., 2020.

[185] M. Esposito, "Mutations and disorder in cellular systems," Ph.D. dissertation, Imperial College London, London, 20203.

[186] R. Satija, J. A. Farrell, D. Gennert, *et al.*

[187] Y. Zhang, T. Liu, C. A. Meyer *et al.*, "Model-based analysis of ChIP-Seq (MACS)," *Genome Biology*, vol. 9, 9 2008.

[188] S. R. Srivatsan, J. L. McFaline-Figueroa, V. Ramani *et al.*, "Massively multiplex chemical transcriptomics at single-cell resolution," *Science*, vol. 367, no. 6473, pp. 45–51, 2020.

[189] V. Buffalo, *Bioinformatics Data Skills: Reproducible and Robust Research with Open Source Tools.* O'Reilly Media, 2015.

[190] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," *arxiv*, 2020. [Online]. Available: https://doi.org/10.48550/arXiv.1802.03426

[191] T. Chari and L. Pachter, "The specious art of single-cell genomics," *PLOS Computational Biology*, vol. 19, no. 8, pp. 1–20, 08 2023.

[192] H. A. Pliner, J. S. Packer, J. L. McFaline-Figueroa *et al.*, "Cicero predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data," *Molecular Cell*, vol. 71, no. 5, pp. 858–871.e8, Sep. 2018.

[193] B. J. H. Butler, *An atlas for staging mammalian and chick embryos.* CRC Press, 1987.

[194] M. Götz and W. B. Huttner, "The cell biology of neurogenesis," *Nature Reviews Molecular Cell Biology*, vol. 6, no. 10, p. 777–788, Oct. 2005.

[195] U. C. Eze, A. Bhaduri, M. Haeussler *et al.*, "Single-cell atlas of early human brain development highlights heterogeneity of human neuroepithelial cells and early radial glia," *Nature Neuroscience*, vol. 24, no. 4, p. 584–594, Mar. 2021.

[196] A. Amador-Arjona, F. Cimadamore, C.-T. Huang *et al.*, "SOX2 primes the epigenetic landscape in neural precursors enabling proper gene activation during hippocampal neurogenesis," *Proceedings of the National Academy of Sciences*, vol. 112, no. 15, pp. E1936–E1945, 2015.

[197] P. Ellis, B. M. Fagan, S. T. Magness *et al.*, "SOX2, a persistent marker for multipotential neural stem cells derived from embryonic stem cells, the embryo or the adult," *Developmental Neuroscience*, vol. 26, no. 2-4, pp. 148–165, 2005.

[198] N. Heins, P. Malatesta, F. Cecconi *et al.*, "Glial cells generate neurons: the role of the transcription factor Pax6," *Nature Neuroscience*, vol. 5, no. 4, p. 308–315, Mar. 2002.

[199] R. Tao, K. N. Davis, C. Li *et al.*, "GAD1 alternative transcripts and DNA methylation in human prefrontal cortex and hippocampus in brain development, schizophrenia," *Molecular Psychiatry*, vol. 23, no. 6, p. 1496–1505, May 2017.

[200] T. Sagai, T. Amano, A. Maeno *et al.*, "SHH signaling mediated by a prechordal and brain enhancer controls forebrain organization," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 116, no. 47, pp. 23 636–23 642, 2019.

[201] I. Barravecchia, S. Mariotti, A. Pucci *et al.*, "MICAL2 is expressed in cancer associated neo-angiogenic capillary endothelia and it is required for endothelial cell viability, motility

and VEGF response," *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, vol. 1865, no. 9, pp. 2111–2124, 2019.

[202] J. Essen and S. Nayler and E. Becker and J. Jacob, "Deconstructing cerebellar development cell by cell," *PLOS Genetics*, vol. 16, no. 4, pp. 1–14, 04 2020.

[203] K. Leto, M. Arancillo, E. B. E. Becker *et al.*, "Consensus paper: Cerebellar development," *The Cerebellum*, vol. 15, no. 6, pp. 789–828, 2016.

[204] V. Y. Wang and H. Y. Zoghbi, "Genetic regulation of cerebellar development," *Nature Reviews Neuroscience*, vol. 2, no. 7, p. 484–491, Jul. 2001.

[205] S. M. Maricich and K. Herrup, "PAX-2 expression defines a subset of GABAergic interneurons and their precursors in the developing murine cerebellum," *Journal of Neurobiology*, vol. 41, no. 2, pp. 281–294, 1999.

[206] P. Urbánek, I. Fetka, M. H. Meisler *et al.*, "Cooperation of Pax2 and Pax5 in midbrain and cerebellum development," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 94, no. 11, pp. 5703–5708, 1997.

[207] J. Yeung, T. J. Ha, D. J. Swanson *et al.*, "A novel and multivalent role of Pax6 in cerebellar development," *Journal of Neuroscience*, vol. 36, no. 35, pp. 9057–9069, 2016.

[208] A. Pillai, A. Mansouri, R. Behringer *et al.*, "Lhx1 and Lhx5 maintain the inhibitory-neurotransmitter status of interneurons in the dorsal spinal cord," *Development*, vol. 134, pp. 357–366, 2007.

[209] N. Genethliou, E. Panayiotou, H. Panayi *et al.*, "Spatially distinct functions of PAX6 and NKX2.2 during gliogenesis in the ventral spinal cord," *Biochemical and Biophysical Research Communications*, vol. 382, no. 1, pp. 69–73, 2009.

[210] S. M. Glasgow, R. M. Henke, R. J. MacDonald *et al.*, "Ptf1a determines GABAergic over glutamatergic neuronal cell fate in the spinal cord dorsal horn," *Development*, vol. 132, no. 24, pp. 5461–5469, 2005.

[211] W. Wei, S. Tuna, M. J. Keogh *et al.*, "Germline selection shapes human mitochondrial DNA diversity," *Science*, vol. 364, no. 6442, 2019.

[212] N. Stoler and A. Nekrutenko, "Sequencing error profiles of Illumina sequencing instruments," *NAR Genomics and Bioinformatics*, vol. 3, 2021.

[213] V. Pejaver, J. Urresti, J. Lugo-Martinez *et al.*, "Inferring the molecular and phenotypic impact of amino acid variants with MutPred2," *Nature Communications*, vol. 11, 2020.

[214] C. A. Lareau, Y. Yin, J. C. Gutierrez *et al.*, "Codon affinity in mitochondrial DNA shapes evolutionary and somatic fitness," *bioarxiv*, 2023. [Online]. Available: https://doi.org/10.1101/2023.04.23.537997

[215] M. T. Lott, J. N. Leipzig, O. Derbeneva *et al.*, "mtDNA variation and analysis using Mitomap and Mitomaster," *Current Protocols in Bioinformatics*, 2013.

[216] R. Brooker, *Genetics: Analysis and Principles*. McGraw-Hill Higher Education, 2018.

[217] B. Rovcanin, J. Jancic, J. Samardzic *et al.*, "In silico model of mtDNA mutations effect on secondary and 3D structure of mitochondrial rRNA and tRNA in Leber's hereditary optic neuropathy," *Experimental Eye Research*, vol. 201, p. 108277, 2020.

[218] E. Ballana, E. Morales, R. Rabionet *et al.*, "Mitochondrial 12S rRNA gene mutations affect RNA secondary structure and lead to variable penetrance in hearing impairment," *Biochemical and Biophysical Research Communications*, vol. 341, no. 4, pp. 950–957, 2006.

[219] Z. Fang, X. Liu, and G. Peltz, "GSEApy: a comprehensive package for performing gene set enrichment analysis in Python," *Bioinformatics*, vol. 39, 2022.

[220] M. Galarreta and S. Hestrin, "Electrical synapses between Gaba-Releasing interneurons," *Nature Reviews Neuroscience*, vol. 2, p. 425–433, 2001.

[221] Q. Zhang, C. Forster-Gibson, E. Bercovici *et al.*, "Epilepsy plus blindness in microdeletion of GABRA1 and GABRG2 in mouse and human," *Experimental Neurology*, vol. 369, 2023.

[222] S. S. Jayakar, D. C. Chiara, X. Zhou *et al.*, "Photoaffinity labeling identifies an intersubunit steroid-binding site in heteromeric GABA type A ($GABA_A$) receptors," *Journal of Biological Chemistry*, vol. 295, pp. 11 495–11 512, 8 2020.

[223] N. Taguchi, N. Ishihara, A. Jofuku *et al.*, "Mitotic phosphorylation of dynamin-related GTPase Drp1 participates in mitochondrial fission," *Journal of Biological Chemistry*, vol. 282, p. 11521–11529, 2007.

[224] D. Otsuga, B. R. Keegan, E. Brisch *et al.*, "The Dynamin-related GTPase, Dnm1p, Controls Mitochondrial Morphology in Yeast ," *Journal of Cell Biology*, vol. 143, pp. 333–349, 1998.

[225] A. Santel and S. Frank, "Shaping mitochondria: The complex posttranslational regulation of the mitochondrial fission protein DRP1," *IUBMB Life*, vol. 60, no. 7, pp. 448–455, 2008.

[226] J. Wakabayashi, Z. Zhang, N. Wakabayashi *et al.*, "The dynamin-related GTPase Drp1 is required for embryonic and brain development in mice," *Journal of Cell Biology*, vol. 186, pp. 805–816, 2009.

[227] R. Iwata, P. Casimir, and P. Vanderhaeghen, "Mitochondrial dynamics in postmitotic cells regulate neurogenesis," *Science*, vol. 369, no. 6505, pp. 858–862, 2020.

[228] M. Cagalinec, D. Safiulina, M. Liiv *et al.*, "Principles of the mitochondrial fusion and fission cycle in neurons," *Journal of Cell Science*, 2013.

[229] Z. Li, K.-I. Okamoto, Y. Hayashi *et al.*, "The importance of dendritic mitochondria in the morphogenesis and plasticity of spines and synapses," *Cell*, vol. 119, p. 873–887, 2004.

[230] J. Aryaman, C. Bowles, N. S. Jones *et al.*, "Mitochondrial network state scales mtDNA genetic dynamics," *Genetics*, vol. 212, no. 4, pp. 1429–1443, 06 2019.

[231] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*, ser. Chapman & Hall/CRC Monographs on Statistics and Applied Probability. CRC Press, 1994. [Online]. Available: https://books.google.co.uk/books?id=MWC1DwAAQBAJ

[232] F. J. Iborra, H. Kimura, and P. R. Cook, "The functional organization of mitochondrial genomes in human cells," *BMC Biology*, vol. 2, 2004.

[233] C. Gardiner, *Handbook of Stochastic Methods for Physics, Chemistry, and the Natural Sciences*, ser. Springer complexity. Springer, 2004.

[234] G. Chirikjian, *Stochastic Models, Information Theory, and Lie Groups, Volume 1: Classical Results and Geometric Methods*, ser. Applied and Numerical Harmonic Analysis. Birkhäuser Boston, 2009.

[235] T. A. Brown, A. N. Tkachuk, G. Shtengel *et al.*, "Superresolution fluorescence imaging of mitochondrial nucleoids reveals their spatial range, limits, and membrane interaction," *Molecular and Cellular Biology*, vol. 31, pp. 4994–5010, 12 2011.

[236] C. E. J. Dieteren, S. C. A. M. Gielen, L. G. J. Nijtmans *et al.*, "Solute diffusion is hindered in the mitochondrial matrix," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 108, 2011.

[237] T. Stephan, A. Roesch, D. Riedel *et al.*, "Live-cell STED nanoscopy of mitochondrial cristae," *Scientific Reports*, vol. 9, 12 2019.