Bayesian Analysis of mtDNA Population Dynamics

Jordan Childs
(MMathStat PgDip)

Thesis submitted for the degree of
Doctor of Philosophy

Faculty of Medical Science
Newcastle University

Author declaration

This thesis is submitted for the degree of Doctor of Philosophy at Newcastle University. The research within was conducted in the Mitochondrial Research Group at Newcastle University under the supervision of Dr C Lawless, Dr A E Vincent, Prof A Golightly (Mathematics Department, University of Durham), Dr C Gillespie (Department of Mathematics, Statistics and Physics), Prof R Lightowlers, and Dr S Pickett.

I certify that none of the material presented within has previously been submitted for a degree or any other qualification at this or any other university by me or anyone else.

For Mum, with love

Abstract

Mitochondria are energy-producing organelles in eukaryotic cells with their own genome (mtDNA), which exists in multiple copies per cell. This allows for the coexistence of wild-type and pathogenic variant mtDNA, known as heteroplasmy. When the proportion of pathogenic mtDNA exceeds a critical threshold, mitochondrial function is impaired. Clonal expansion refers to the increase in pathogenic mtDNA within a cell, potentially leading to dysfunction, but its underlying mechanisms remain poorly understood.

Mathematical modelling has emerged as a powerful tool to investigate mtDNA dynamics, allowing simulation of long-term dynamics that cannot be measured experimentally. However, models rely on assumptions, simplifying the biological system, and require parameter inference from limited and often tissue-specific data, complicating their interpretation.

This thesis employs advanced statistical methods to improve our understanding of mtDNA population dynamics. First, a Bayesian model estimates the proportion of blood cells that have reached wild-type homoplasmy, demonstrating that T cell differentiation into memory cells selectively reduces pathogenic mtDNA. Second, a Bayesian classification model infers the proportion of skeletal muscle fibres with oxidative phosphorylation (OXPHOS) defects from OXPHOS protein abundance data, outperforming existing classification method. Finally, the practicalities of comparing theories of clonal expansion using OXPHOS deficient data and mathematical models is investigated using real and synthetic datasets.

Overall, this work underscores the potential of mathematical models in studying clonal expansion while highlighting the challenges posed by limited and variable biological data. It presents novel techniques for inferring mtDNA dynamics and emphasizes the need for more comprehensive experimental data to refine model accuracy and interpretability.

Acknowledgments

I must first thank my (numerous) supervisors, without their collective support and guidance, I would never have made it to this point.

Firstly, I would like to thank Conor Lawless for his (unfounded) belief in me and unrelenting, infectious enthusiasm for scientific research. Without you, this thesis would not have been possible. Secondly, I must thank Amy Vincent for her unwavering support and help throughout my project. Despite not always understanding what I was talking about, your constant presence and biological knowledge were invaluable, not to mention providing non-judgmental answers to my often embarrassingly simple questions about mitochondria. Thank you to Andy Golightly, your statistical guidance and friendly chats throughout my PhD and during my undergraduate degree were always a welcome relief. Without you, I certainly would not have a thesis. Thank you also to Colin Gillespie; discussions with you during the early part of my PhD were extremely helpful. Thank you to Bob Lightowlers, who was always available when called upon and, more importantly, was always free for a reassuring and uplifting talk. Lastly, I am eternally grateful to Dr Sarah Pickett. Who, for the first three years, always provided a friendly face and occasional guidance, but crucially agreed to become a part of the supervisory team during the thesis write-up stage. Without your feedback, my thesis would undoubtedly be significantly less readable and contain significantly less biological content. Our weekly meetings kept me almost on track and almost sane, thank you. A special thank you must be given to Imogen Franklin, Valeria Di Leo and Tiago Gomes, who willingly provided their datasets and expertise whenever asked.

Thank you to everyone based in room 4.052 of the Urban Sciences Building who made coming to the office almost enjoyable over the last four years, and thank you for persuading me to leave work early almost every Friday in support of our local pubs. Without you, my thesis would have been finished considerably earlier, but it would not have been nearly as fun. I must also thank my friends based in the Mitochondrial Research Group for always welcoming me with open arms and bottles of wine. Your steadfast encouragement over the last few years always made me feel better.

To all my friends who have patiently listened to my complaints and minor meltdowns on far too many occasions, thank you. Thank you for the afternoon coffees, snooker sessions, tennis matches, pub trips, pub crawls, nights out and bottomless'. Thank you for always being eager to take trains that were almost always delayed or cancelled to visit me in Newcastle, and thank you for letting me escape for a few days and visit you. Your collective ability to make me forget about work is unparalleled and will not be forgotten.

Finally, I must thank my family for their constant belief in me and their emotional support throughout not only my PhD but my entire university experience. Despite not knowing what a statistic or a mitochondrion is, you were always there. To Mum, thank you for your support, both financial and emotional. I truly could not have done this without you.

Publication list

Franklin, I. G., Milne, P., Childs, J., Boggan, R. M., Barrow, I., Lawless, C., Gorman, G. S., Ng, Y. S., Collin, M., Russell, O. M., & Pickett, S. J. (2023). T cell differentiation drives the negative selection of pathogenic mitochondrial DNA variants [Publisher: Life Science Alliance Section: Research Articles]. *Life Science Alliance*, *6*(11). https://doi.org/10.26508/lsa.202302271

Childs, J., Gomes, T. B., Vincent, A. E., Golightly, A., & Lawless, C. (2025). Bayesian classification of OXPHOS deficient skeletal myofibres [Publisher: Public Library of Science]. *PLOS Computational Biology*, *21*(2), e1012770. https://doi.org/10.1371/journal.pcbi.1012770

Contents

List of Figures

List of Tables

Abreviations

POLG Polymerase-γ.

TWNK Twinkle.

ABM agent-based model.

ATP adenosine triphosphate.

BER base excision repair.

BIC Bayesian information criterion.

COX cytochrome c oxidase.

CPEO chronic progressive external ophthalmoplegia.

dNTP deoxynucleotide.

ESS effeective sample size.

ETC electron transport chain.

ETP early T cell precursor.

FCD full-conditional distribution.

GMM Gaussian mixture model.

GPE Gaussian process emulator.

HDI high density interval.

HSC haematopoietic stem cell.

IF immunofluorescence.

IMC imaging mass cytometry.

IMM inner mitochondrial membrane.

IMS intermembrane space.

KSS Kearns-Sayre syndrome.

LHON Leber hereditary optic neuropathy.

MCMC Markov chain Monte Carlo.

MELAS mitochondrial encephalopathy, lactic acidosis and stroke-like episodes.

MHC major histocompatability complex.

MIDD maternally-inherited diabetes and deafness.

mt-rRNA mitochondrial ribosomal ribonucleic acid.

mt-tRNA mitochondrial transfer ribonucleic acid.

NCR non-coding region of mtDNA.

OMM outer mitochondrial membrane.

OXPHOS oxidative phosphorylation.

PDF probability density function.

PGC primordial germ cell.

PMF probability mass function.

PNN perinuclear niche.

QIF quardruple immunofluorescence.

RGD random genetic drift.

ROS reactive oxygen species.

SNV single-nucleotide variant.

SoS survival of the smallest.

SSD stochastic survival of the densest.

$\mathbf{T}_{\mathrm{CM}}$ central memory T cell.

$\mathbf{T}_{\mathrm{EMRA}}$ memory T cells re-expressing naïve marker CD45Ra.

$\mathbf{T}_{\mathrm{EM}}$ effector memory T cell.

TCR T cell receptor.

Probability distributions

Normal distribution

The probability density function of a normally distributed random variable is characteristic of a bell curve and defined by two parameters: mean μ and variance σ^2. However, it is often described in terms of precision, which is the inverse of variance. Both parameterisations are used throughout and highlighted where appropriate. A normally distributed random variable can take any value on the whole real line, this is known as its support. Let X be normally distributed with mean μ, variance σ^2 and let $\tau = 1/\sigma^2$. The following defines the support, notation and probability density function for a random variable X.

Support

$$X \in \mathbb{R}.$$

Notation

The distribution is denoted such that the second parameter always describes the variance, and, therefore, it is clear that τ is the precision as its inverse appears in the following notation

$$X \sim \mathrm{N}\left(\mu, \sigma^2\right),$$
$$X \sim \mathrm{N}\left(\mu, \tau^{-1}\right).$$

Definition

The definition of the density function is, for both the variance and precision parameterisations

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\left(x-\mu\right)^2\right)\right\},$$
$$f(x|\mu, \tau^{-1}) = \sqrt{\frac{\tau}{2\pi}} \exp\left\{-\frac{\tau}{2}\left(x-\mu\right)^2\right)\right\}.$$

Multivariate normal distribution

The univariate normal distribution can be extended to have multi-dimensional random vectors. The multivariate normal distribution still possesses the classical bell-shaped curve; however, it is within a multi-dimensional space. The distribution is defined by a mean vector and a covariance matrix. Let $\boldsymbol{X}$ be a p-dimensional random vector, i.e. $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)^{\mathrm{T}}$, where each element could be any real number. Notation throughout the thesis is kept as consistent as possible. Unless specified otherwise, a bold letter indicates a vector, such as $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)^{\mathrm{T}}$, and a non-bold letter indicates a scalar, such as x. The superscript T of a vector (or matrix) indicates its transpose. By convention, vectors are typically written as column vectors.

Support

$$\boldsymbol{X} \in \mathbb{R}^p.$$

Notation

The distribution is denoted similarly to the univariate case, and its multivariate nature is inferred from the parameters describing it, that is

$$\boldsymbol{X} \sim \mathrm{N}\left(\boldsymbol{\mu}, \Sigma\right).$$

Definition

$$f(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma) = (2\pi)^{-k/2}\det\left(\Sigma\right)^{-1/2}\exp\left\{-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)^{\mathrm{T}}\Sigma^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)\right\},$$

where det() is the determinant of the matrix. The PDF is properly defined only when Σ is a positive semi-definite matrix.

Exponential distribution

Exponentially distributed random variables have a decreasing probability density function whose mode is always located at 0.0. The function is defined with a single argument, λ. The rate definition of the density function is used throughout, not the scale (the inverse of the rate). For a random variabel X with support over the positive real line, $\mathbb{R}^+ = \{x \text{ such that } x \geq 0.0\}$.

Support

$$X \in \mathbb{R}^+.$$

Notation

$$X \sim \mathrm{Exp}\left(\lambda\right),$$

Definition

$$f(x|\lambda) = \lambda\exp\left\{-\lambda x\right\}.$$

Gamma distribution

A gamma random variable has support over positive real numbers, $\mathbb{R}^+$. The distribution is defined by two parameters: shape and rate. Although some chose to parameterise the distribution with the scale, the rate's inverse, this will not be done here.

Support

$$X \in \mathbb{R}^+$$

Notation

$$X \sim \mathrm{Ga}\left(\alpha, \beta\right),$$

Definition

$$f(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x},$$

where $\Gamma(\alpha)$ denotes the gamma function.

Beta distribution

The beta distribution is a continuous probability distribution constrained to be between 0.0 and 1.0, and it is therefore often used to model proportions or probabilities.

Support

$$X \in [0.0, 1.0]$$

Definition

$$f(x|\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}$$

where $\mathrm{B}(\alpha,\beta) = \frac{\Gamma(\alpha)+\Gamma(\beta)}{\Gamma(\alpha+\beta)}$.

Notation

$$X \sim \mathrm{Beta}\,(\alpha,\beta)$$

Binomial distribution

The binomial distribution models the number of successes for n independent events, where each event has a probability of success p and is therefore a discrete random variable, taking whole numbers between 0 and n. The Bernoulli distribution, not mentioned here, is a special case of the Binomial for a single trial, i.e. $n = 1$.

Support

$$X \in \{0, 1, 2, \ldots, n\}$$

Definition

$$f(x|n,p) = \binom{n}{x} p^x (1-p)^{n-x}$$

Notation

$$X \sim \mathrm{Binom}\,(n,p)$$

Truncated distributions

A truncated distribution limits the possible range of values for a random variable to a specific region and scales the distribution accordingly to account for this. Suppose that $f_X(x|\theta)$ is a PDF over the range $\mathcal{X}$ and that Y is a random variable with support $\mathcal{Y} \subseteq \mathcal{X}$. That is, the support of Y is a region within $\mathcal{X}$. The random variable Y can be modelled by $f_X(x)$ by truncating the distribution. Let the truncated distribution be denoted $f_Y(y|\theta)$.

Support

$$Y \in \mathcal{Y}$$

Definition

$$f_Y(y|\theta) = \frac{f_X(y|\theta)}{\Pr\left(X \in \mathcal{Y}\right)} \mathbb{I}(y \subseteq \mathcal{Y})$$

Where $\mathbb{I}(\cdot)$ is the indicator function and is equal to 1 if the condition inside the brackets is true and 0 otherwise. This ensures the density is zero outside the region $\mathcal{Y}$.

Notation

Throughout this thesis, a truncated distribution will be indicated by subscript notation within the distribution's notation. For example, a normal distribution truncated on the range $[a, b]$ will be indicated as

$$Y \sim \mathrm{N}_{[a,b]}\left(\mu, \sigma^2\right).$$

Where appropriate, truncation will also be highlighted within the text.

Chapter 1

Introduction

1.1 Mitochondrial Biology

1.1.1 Mitochondrial origins

Mitochondria are organelles in most eukaryotic cells and are most well-known for the production of adenosine triphosphate (ATP) via oxidative phosphorylation (OXPHOS). They also play key roles in other biochemical processes within the cell, such as iron-sulphur cluster formation and apoptosis (Duchen, 2004).

The α-probacteria endosymbiotic origins of mitochondria are well accepted (Esser et al., 2004; Sicheritz-Pontén et al., 1998); however, the origins of the eukaryotic cells are contested. Two main theories have been proposed in the literature: the Archezoan and the symbiogenesis hypotheses. The former suggests that a nucleated archezoa cell captured the endosymbiont (Cavalier-Smith, 1987; Roger et al., 2017; Yang et al., 1985), and the latter suggests that the endosymbiotic event occurred before the diversion of eukaryotes from prokaryotes (Martin & Müller, 1998).

1.1.2 Mitochondrial structure

Palade (1953) discovered the mitochondrial structure, proposing the "baffle" model in 1953. Palade *et al.* identified the inner mitochondrial membrane (IMM), which encloses the mitochondrial matrix, and the outer mitochondrial membrane (OMM), which surrounds the IMM, enclosing the intermembrane space (IMS) between the two membranes. Palade *et al.* identified cristae protruding from the IMM into the matrix; however, later evidence suggests that these are formed from invaginations of the membrane and are connected to the IMS by tubular cristae junctions (Perkins et al., 1997). The membrane structure is represented in Figure 1.1, which also highlights the respiratory chain complexes and mitochondrial DNA. The functions and structure of which are discussed in Chapters 1.1.4 and 1.1.6.

The OMM is relatively permeable, allowing many small and non-charged molecules to pass through without restriction. The IMM strictly controls molecule transportation, maintaining an electrochemical gradient between the mitochondrial matrix and IMM, as required for ATP production (Lemasters, 2007).

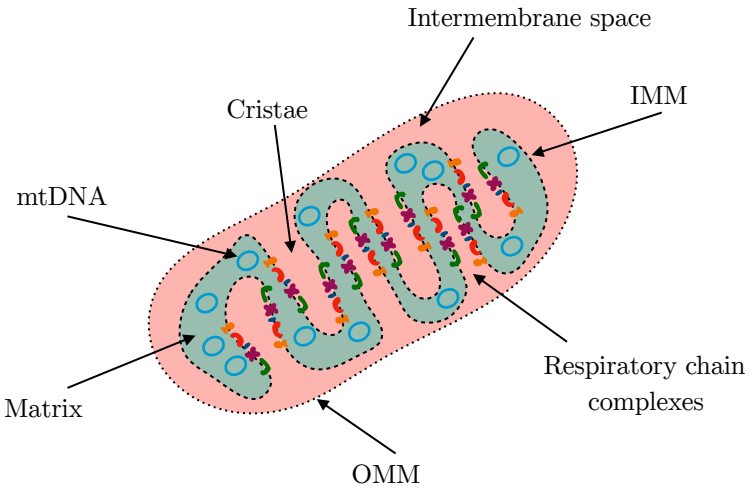


Figure 1.1: **Mitochondrial structure**. Example diagram of labelled components of the mitochondria. The two mitochondrial membranes are shown with black dotted lines, with the outer membrane having smaller dots. The IMS is shown in pink, and the matrix in blue/green. MtDNA is shown as blue circular structures within the matrix. The five complexes of the respiratory chain are found on the IMM, with each complex being shown in a different colour. Labels for each component are shown with arrows as appropriate.

1.1.3 Mitochondrial dynamics

Mitochondria are highly dynamic, continuously undergoing fission and fusion (Bereiter-Hahn & Vöth, 1994). The balance of these two processes plays a vital role in mitochondrial function and can change depending on cellular requirements (Kuznetsov et al., 2009). Fission and fusion rates vary between cell types; neurons have highly dynamic mitochondria that move along the neuron and constantly undergo fission and fusion. On the other hand, skeletal muscle fibres have a primarily static mitochondrial network (Kuznetsov et al., 2009). Mitochondrial dynamics enable the exchange of mitochondrial material, including mitochondrial DNA (mtDNA) and mitochondrial proteins, thereby reducing mitochondrial stress and increasing mtDNA stability (Chen et al., 2010). Disruptions to fission or fusion are associated with conditions such as Parkinson's disease (Van Laar & Berman, 2009) and Alzheimer's (Santos et al., 2010).

Fission

Fission is the process by which one mitochondrion splits into two distinct organelles. The exact mechanisms by which the IMM and OMM are spliced are unknown. However, the proteins involved are well documented. Early work identified the proteins, Dnm1 and Fis1, within yeast (Bleazard et al., 1999; Mozdy et al., 2000). Mammalian versions were later identified as dynamin-related protein 1 (DRP1) and mitochondrial fission 1 (FIS1), (James et al., 2003; Smirnova et al., 2001). DRP1 is recruited from outside the mitochondria and oligomerises, forming a helix. The molecule restricts the mitochondrion, which results in membrane scission and, ultimately, two distinct mitochondria (Legesse-Miller et al., 2003). Several proteins are thought to be involved in the recruitment of DRP1. FIS1, a protein found on the OMM, is one such protein (Mozdy et al., 2000). A diagram of mitochondrial fission can be seen in Figure 1.2. The removal or dysfunction of DRP1 or FIS1 results in hyper-fused and elongated mitochondria (Frank et al., 2001; Lee et al., 2016). Another protein, Septin 2, is localised to sites of mitochondrial fission,

and its depletion decreases DRP1's recruitment (Pagliuso et al., 2016). Fission also plays a key role in mitophagy; see Chapter 1.1.5.

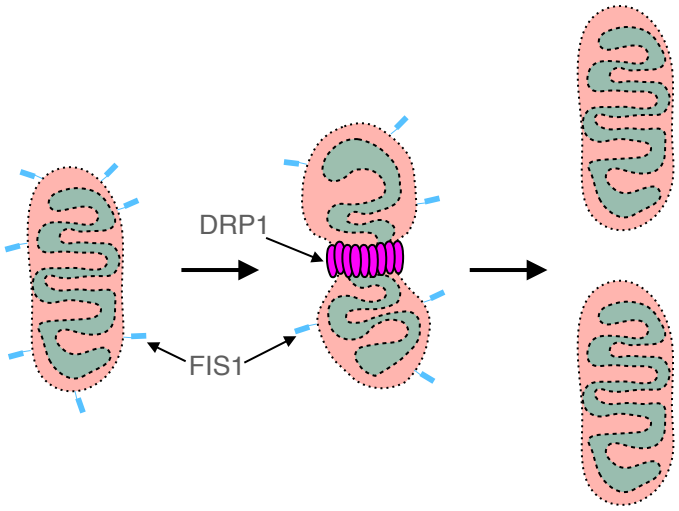


Figure 1.2: **Mitochondrial fission diagram**. Mammalian proteins FIS1 and DRP1 are represented as blue markers on the OMM and purple ovals, respectively. The DRP1 helix is shown as a chain of DRP1 molecules wrapping around the mitochondrion.

Fusion

Fusion is the process by which two distinct mitochondria become one, see Figure 1.3. The process is similarly complex to fission, requiring several proteins and complexes to successfully merge the two mitochondrial membranes. Two proteins, mitofusin 1 and 2 (MFN1 and MFN2), are crucial for mitochondrial fusion. Despite having similar roles, both are required for fusion. However, their significance is tissue-specific (Chen et al., 2003, 2005). MFN2 is additionally required for other processes, such as tethering mitochondria to the endoplasmic reticulum (de Brito & Scorrano, 2008) and normal glucose homeostasis (Sebastián et al., 2012). Mitochondrial fusion is believed to be mediated by OPA1 (Cipolat et al., 2004).

A disruption in fusion can result in a fragmented mitochondrial network, as one might expect, but can also lead to a reduction or complete loss of mtDNA within the cell (Chen et al., 2007; Hermann et al., 1998; Rapaport et al., 1998). A lack of mitochondrial fusion in mouse skeletal muscle fibres increases variant mtDNA within the fibre (Chen et al., 2010).

1.1.4 Mitochondrial genetics

Mitochondrial genome

Mitochondria possess their own circular, double-stranded genome (mtDNA) (Nass, 1966). The outer strand is rich in guanine, giving it a higher molecular mass compared to the inner strand. Therefore, they are referred to as the heavy (H) and light (L) strands, respectively. The human mitochondrial genome is small, relative to the nuclear genome, containing approximately 16.5kb and has a much lower proportion of non-coding regions. The genome contains 37 genes which code two mitochondrial ribosomal ribonucleic acid

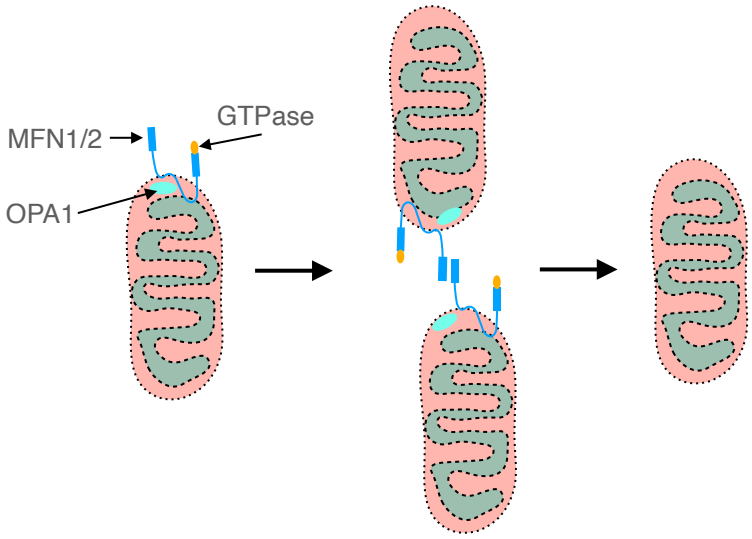


Figure 1.3: **Mitochondrial fusion diagram**. Key proteins related to mitochondrial fusion, the joining of two mitochondria into a single mitochondrion. MFN1/2 and OPA1 are shown in shades of blue, and the GTPase is shown in yellow.

(mt-rRNA), 22 mitochondrial transfer ribonucleic acid (mt-tRNA), and 13 OXPHOS subunits, the majority of which are encoded on the H-strand. MtDNA also contains the non-coding region of mtDNA (NCR) or D-loop, which plays a key role in mtDNA replication. Originally sequenced by Anderson et al. (1981) and later revised by Andrews et al. (1999), known as the Cambridge reference sequence. The genome is represented in Figure 1.4, highlighting the genes associated with each OXPHOS complex. The D-loop, depicted at the top of the molecule, contains O_H, the origin of replication for the heavy strand.

A single mitochondrion contains multiple copies of the genome, a state known as polyploidy. Each mtDNA molecule is packaged into a small nucleoprotein complex called a nucleoid (Bogenhagen, 2012; Satoh & Kuroiwa, 1991). However, the size of the cellular mtDNA population, copy number, varies depending on cell type; mature oocytes have a copy number of approximately 100,000 (Shoubridge & Wai, 2007), while skeletal muscle fibres have an estimated copy number of 3600 (Miller et al., 2003), and blood cells range in the hundreds (Kelly et al., 2012; Rausser et al., 2021).

MtDNA replication

MtDNA continuously undergoes replication, independent of the cell cycle (Bogenhagen & Clayton, 1977), requiring several proteins and complexes encoded in both the mitochondrial and nuclear genomes. Two key proteins are Polymerase-γ (*POLG*), responsible for replication of mtDNA, and Twinkle (*TWNK*), which unwinds the DNA double helix (Wanrooij & Falkenberg, 2010; Young & Copeland, 2016).

There are two prevailing theories for the mtDNA replication mechanism: synchronous, Figure 1.5, and asynchronous, Figure 1.6. Both replication models agree that replication of the H-strand begins at the H-strand origin, O_H, and replication of the L-strand begins at its origin, O_L. Robberson et al. (1972) suggests that replication of both strands begins simultaneously, being initiated at their respective strand origins and moving in opposing

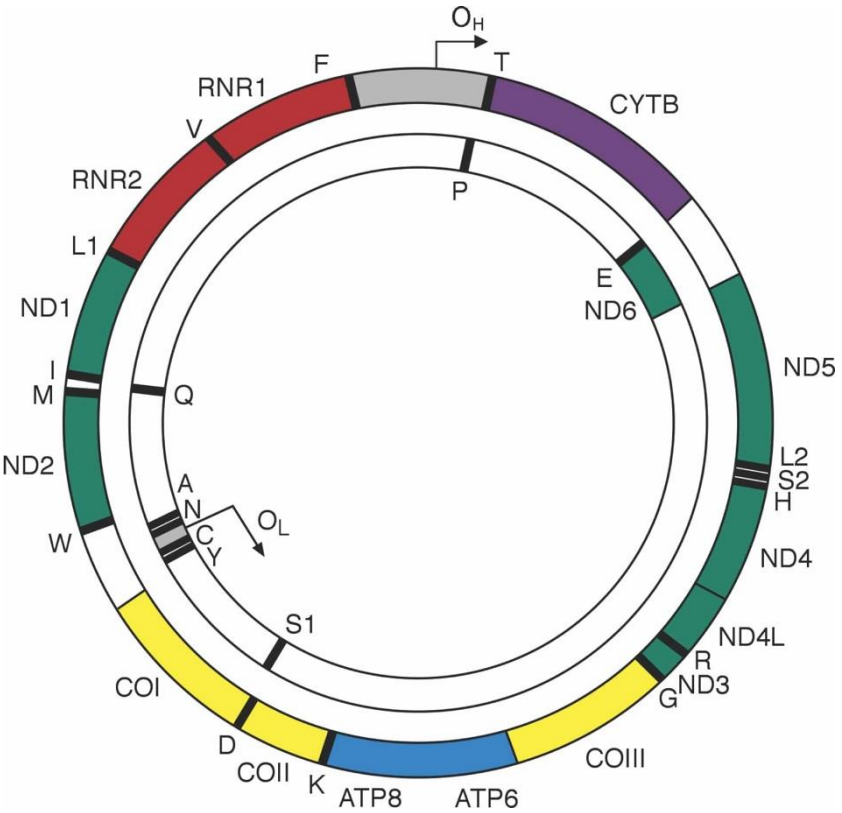


Figure 1.4: **Human mitochondria DNA organisation**. The larger (outer) circle maps the heavy strand, and the smaller circle maps the light strand. Genes encoding subunits of CI are green, and genes for CIII are purple. Catalytic subunits of CIV and CV are yellow and blue, respectively. Ribosomal RNA genes are red, and the black bars indicate transfer RNA genes and are labelled with their single-letter abbreviations. The heavy and light strand origins are denoted O_H and O_L. Taken from Greaves et al. (2012).

directions around the molecule. Asynchronous replication suggests that replication begins with the H-strand. After approximately two-thirds of the H-strand has been replicated, the replication reaches the O_L site, and then L-strand replication begins. From here, H-strand and L-strand replication continue simultaneously, in opposing directions, with L-strand replication 'lagging' behind H-strand (Brown et al., 2005; Clayton, 1982).

MtDNA repair

Clayton et al. (1974) proposed that mitochondria have no repair mechanisms for damaged mitochondrial DNA. However, later work identified multiple mechanisms. Base excision repair (BER) has been found to repair damage caused by reactive oxygen species (ROS) (Liu et al., 2008; Longley et al., 1998; Zheng et al., 2008) and single-stranded breaks utilise the same machinery (Kazak et al., 2012). The mechanisms for repairing double-

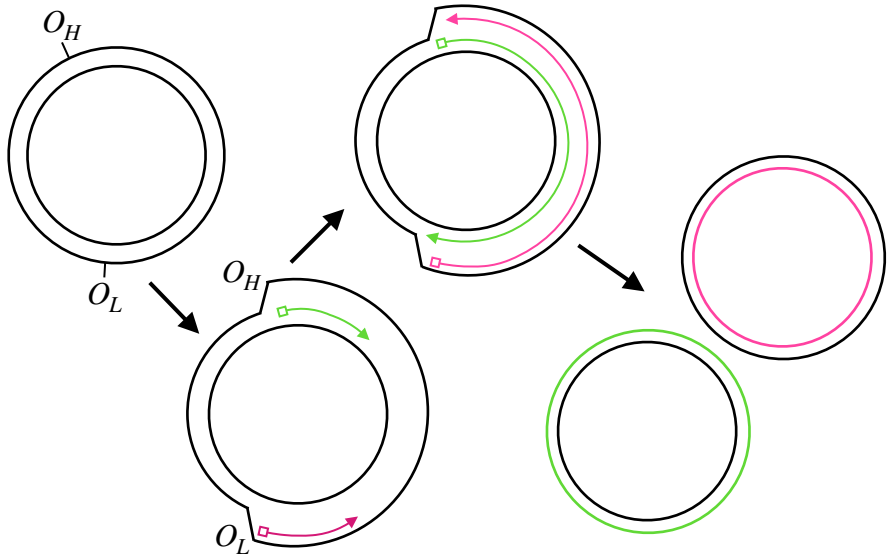


Figure 1.5: **Synchronous mtDNA replication**. A diagram depicting synchronous mtDNA replication, where both strands begin replication simultaneously at their respective origins, O_H and O_L. The pink and green lines show new strands being formed, and the arrows indicate the direction of movement for POLG and TWNK around the mtDNA.

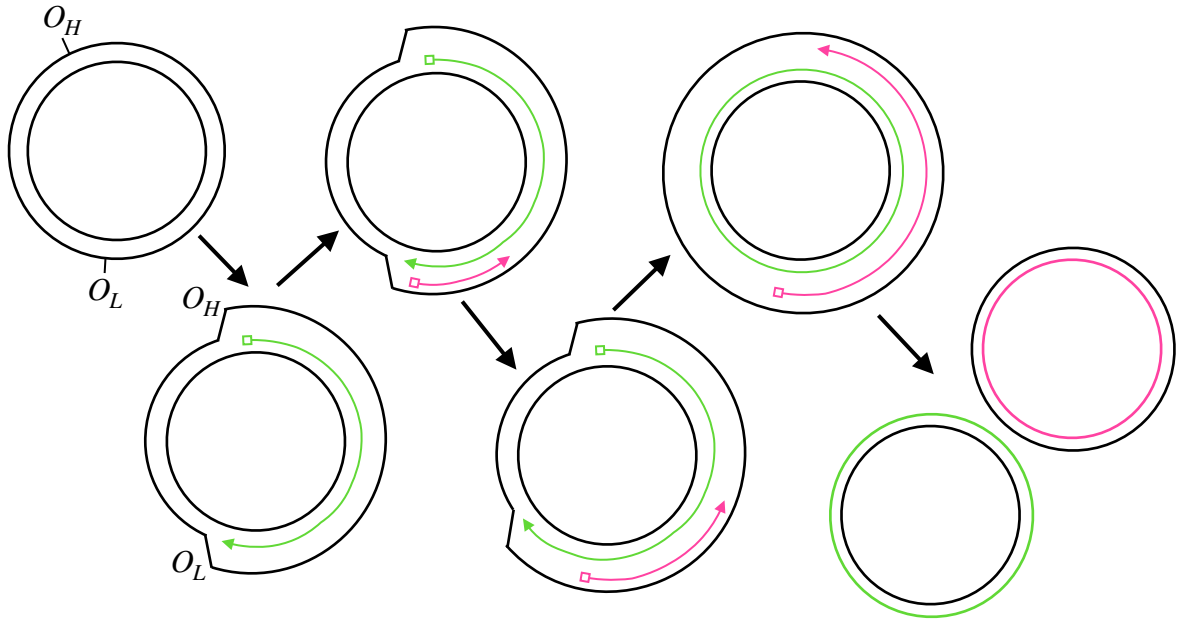


Figure 1.6: **Asynchronous mtDNA replication**. Diagram showing the steps of asynchronous mtDNA replication. The Heavy and light strand origins are denoted O_H and O_L respectively. The new heavy and light strands are shown in green and pink, respectively. Arrows show the direction of movement for POLG and TWNK.

strand breaks are not clear. However, some works have proposed recombination (Bacman et al., 2009; Damas et al., 2014; Fukui & Moraes, 2009).

1.1.5 Mitochondrial biogenesis and mitophagy

Mitochondrial turnover is independent of the cell and occurs at differing rates depending on cell type. Kim et al. (2012) used a rat model to compare the half-life of mitochondrial

proteins in liver and heart muscle, showing their half-life to be approximately four and 17 days, respectively.

Mitophagy is the process of degrading mitochondria and it's contents, including mtDNA. A reduction in membrane potential leads to the selective degradation of mitochondria (Kanki & Klionsky, 2008). The mitochondrion is segregated from the mitochondrial network by fission and is unable to re-fuse (Twig et al., 2008). Mitophagy removes the mitochondrion and any mtDNA within.

1.1.6 ATP production and oxidative phosphorylation

ATP is the primary energy source within a cell, and mitochondria are central to its production. The process of releasing energy from glucose comprises three sections: anaerobic glycolysis, tricarboxylic acid (TCA) cycle and OXPHOS. Glycolysis occurs outside of the mitochondria, in the cytosol, and produces two pyruvate molecules, two ATP molecules and reduces two NAD^+ to NADH molecules (Berg et al., 2015). The chemical reaction equation for anaerobic glycolysis can be seen in Eq. 1.1. Reduction of NAD^+ to NADH is key to ATP production, as NADH molecules are the utilised within the electron transport chain (ETC), described later.

$$\begin{aligned}\text{Glucose} + 2\text{NAD}^+ + 2\text{ADP} + 2\text{P}_\text{i} \rightarrow \\ 2\text{Pyruvate} + 2\text{NADH} + 2\text{H}^+ + 2\text{ATP} + 2\text{H}_2\text{O}\end{aligned} \tag{1.1}$$

The TCA cycle is a series of reactions in the mitochondrial matrix that begins with one pyruvate molecule converting to an acetyl-CoA molecule (Berg et al., 2015). The pyruvate to acetyl-CoA reaction also reduces one NAD^+ molecule, as described in Eq. 1.2.

$$\text{Pyruvate} + \text{CoA} + \text{NAD}^+ \rightarrow \text{Acetyl-CoA} + \text{NADH} + \text{H}^+ + \text{CO}_2 \tag{1.2}$$

The acetyl-CoA then undergoes a series of reactions, culminating in the formation of an oxaloacetate molecule (CoA). The reactions also produce one molecule of ATP and CO_2, two $FADH_2$ molecules, and three NADH molecules (Berg et al., 2015). Significantly, NADH and $FADH_2$ molecules act as electron carriers within the ETC. The net TCA reaction is described in Eq. 1.3.

$$\begin{aligned}\text{Acetyl-CoA} + 3\text{NAD}^+ + \text{FAD} + \text{ADP} + \text{P}_\text{i} + 2\text{H2O} \rightarrow \\ \text{CoA} + 3\text{NADH} + 2\text{H}^+ + \text{FADH}_2 + \text{ATP} + 2\text{CO2}\end{aligned} \tag{1.3}$$

OXPHOS is the final stage of ATP production and comprises of two parts: the ETC and ATP production. Respiratory chain complexes I-IV form the ETC and produce the energy required for ATP production. Through a series of reactions, electrons are moved to lower energy states, and the energy released is used to pump H^+ molecules out of the mitochondrial matrix into the IMS, forming an electrochemical gradient. The final respiratory chain complex, CV, utilises this potential energy to produce ATP (Berg et al., 2015). Each complex is briefly discussed in the next paragraph, and Figure 1.7 summarises the connections between complexes and key biochemical processes.

The first respiratory chain complex (CI) is the largest, consisting of approximately 45 subunits, seven of which are encoded by the mitochondrial genome (Zhu et al., 2016). The protein NDUFB8 is one such subunit and is often targeted when assaying CI levels via immunofluorescence or imaging mass cytometry (Grünewald A et al., 2014; Lehmann et al., 2019; Rocha et al., 2015), as will be seen in Chapter 4. Complex I obtains electrons by oxidation of NADH, producing two electrons to be passed through the ETC, and one NAD^+ molecule. Complex II (CII) comprises four subunits (Sun F et al., 2005). Importantly, CII is the only complex to be entirely nuclear-encoded, and can therefore be used to assess the mitochondrial genome's effect on OXPHOS protein levels. Although CII does not pump protons across the IMM, it plays key roles in OXPHOS. During the TCA cycle, CII is responsible for converting succinate to fumarate, producing an $NADH_2$ molecule, acting as the final entry point for electrons into the ETC. Complex III (CIII) consists of 11 subunits, of which only one, cytochrome *b*, is mitochondrially encoded. One role of CIII is to catalyse the transfer of electrons to cytochrome *c*, which transfers the electrons to complex IV, also known as cytochrome c oxidase (COX). Complex IV (CIV or COX) is the last complex in the ETC, and comprises 13 subunits, of which three are mitochondrially encoded; MTCOI, MTCOII, and MTCOIII (Tsukihara T et al., 1996). Similarly to NDUFB8, MTCOI is often targeted when assaying COX levels via immunofluorescence or imaging mass cytometry (Lehmann et al., 2019; Rocha et al., 2015). COX is responsible for reducing oxygen to water, being the final destination of electrons in the ETC.

Complex V (CV), also known as ATP synthase, is responsible for producing ATP. ATP production depends on the physical rotation of a CV subunit, which is driven by the electrochemical gradient and the movement of H^+ molecules from the IMS to the mitochondrial matrix (Noji et al., 1997). The rotation generates the energy required to phosphorylate ADP, releasing ATP.

1.1.7 DNA mutations

MtDNA mutations occur at a much higher frequency than in the nuclear genome, with Brown et al. (1979) reporting a 10-fold increase. The exact causes of this are unknown; however, several factors are likely to contribute. Continuous mtDNA turnover increases the probability of an error occurring during replication, which, if left uncorrected, continues to replicate, thereby increasing the variant mtDNA concentration. The proximity of mtDNA to the respiratory chain may lead to damage from ROS, a respiratory chain byproduct which can damage both proteins and DNA (Cui et al., 2012). In addition, Kunkel and Loeb (1981) demonstrated that polymerase-γ, the catalyst responsible for mtDNA replication, has a much lower fidelity rate than its nuclear counterpart.

MtDNA single-nucleotide variants

A single-nucleotide variant (SNV) is a change of nucleotide at a single base pair; they can be either inherited or appear throughout life and be benign or pathogenic. The majority of inherited SNVs affect mt-tRNA genes (Schaefer et al., 2008), with the most common SNV, m.3243A>G, being no exception.

M.3243A>G is located within the *MT-TL1* gene which encodes $tRNA^{Leu(UUR)}$. Therefore, the variant affects the translation of proteins encoded by the mitochondrial genome.

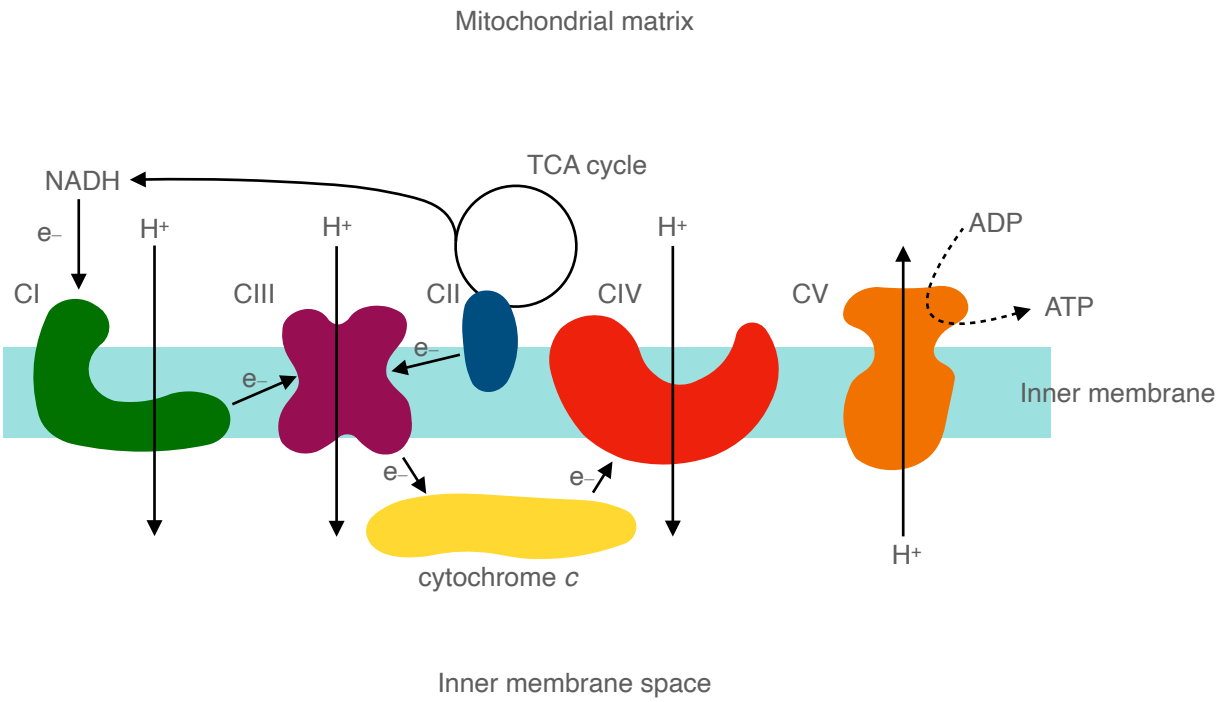


Figure 1.7: **Oxidative phosphorylation**. The five complexes involved in OXPHOS are embedded in the inner mitochondrial membrane. Complexes I-IV form the electron transport chain and pump protons across the inner membrane into the inter-membrane space. The resulting electrochemical gradient causes the physical rotation of a CV subunit, generating the energy required for ATP production. Each OXPHOS complex is labelled and shown in a different colour, as well as the electron carrier cytochrome *c*. The TCA cycle is highlighted, being the source of $FADH_2$ electron carrier for CII and NADH electron carrier for CI. The movement of protons and their direction through CI, CIII, CIV, and CV is indicated with arrows and H^+ labels at their original location.

Complex I has the largest number of mitochondrially encoded proteins, and its dysfunction, therefore, is commonly associated with m.3243A>G. However, deficiency in other complexes has also been detected. The resulting patient symptoms are varied (Pickett et al., 2018), but m.3243A>G is the leading cause of mitochondrial encephalopathy, lactic acidosis and stroke-like episodes (MELAS), with 80% of MELAS patients having m.3243A>G related mitochondrial disease (Urata et al., 2004). Other phenotypes include maternally-inherited diabetes and deafness (MIDD) and chronic progressive external ophthalmoplegia (CPEO) (Nesbitt et al., 2013).

The m.3243A>G variant has not been observed *in vivo* to have reached homoplasmic levels within a single cell, meaning patient cells harbour both m.3243A>G and wild-type mtDNA. A cell possessing more than one species of mtDNA is said to be heteroplasmic. Other mtDNA variants, however, can exist in a homoplasmic state and be the only mtDNA species within a cell. Leber hereditary optic neuropathy (LHON) disease is a form of mitochondrial disease, which can cause damage to the retinal ganglion, causing central blindness in patients. Three mtDNA variants cause over 95% of LHON cases: m.11778G>A, m.3460G>A, and m.14484T>C, all of which encode a subunit of CI of the respiratory chain and are found in a state of homoplasmy within most patients (Harding et al., 1995; Macmillan et al., 1998).

Similarly to m.3243A>G, all SNVs can have a large phenotypic variance. The cellular impact depends on the proportion of variant mtDNA, which is dynamic due to continuous mtDNA turnover. An inherited variant proportion can also drastically differ between neighbouring cells, affecting their function and the resulting cellular phenotype. These considerations are discussed further in Chapter 5.

MtDNA single large-scale deletions

A single, large-scale deletion is the loss of a continuous section of the mitochondrial genome. Deletions may arise from errors in replication or repair during embryo development, resulting in the same mtDNA variant being present in all cells (Krishnan et al., 2008; Shoffner et al., 1989). Although less common than SNVs, deletions account for approximately 16% of adult mtDNA variants and are responsible for 12% of mitochondrial diseases in adult patients (Gorman et al., 2015). Similarly to SNVs, patients present a wide variety of phenotypes, such as CPEO and Kearns-Sayre syndrome (KSS). KSS is a severe multi-system disease, often presenting with ptosis, Pigmentary retinopathy, and cardiac conduction abnormalities (Goldstein & Falk, 1993). Some patients with large-scale deletions also present non-syndromic symptoms, such as muscle weakness and ptosis (Mancuso et al., 2015).

OXPHOS deficiency is significantly correlated to variant load, the proportion of pathogenic variant mtDNA within a single cell, (Rocha et al., 2015), as well as deletion size and location (Rocha et al., 2018). The relationship between deletion size and patient phenotype is contested within the literature. Some studies have found a significant relationship between the two (Yamashita et al., 2008), whereas others have found none (López-Gallardo et al., 2009). Age of deletion formation has also been found to be significantly correlated to patient phenotype (Yamashita et al., 2008). Grady et al. (2014) found that deletion size and location, as well as variant load, are significantly related to disease progression and burden. In their study, a large cohort of 87 patients was collected and their data analysed using multiple regression mixture models. This highlighted the relationship between predictors and gave better context to conflicting previous reports, which used simpler statistical methods and looked at fewer explanatory variables.

One specific deletion, the 'common deletion', is far more frequent than the rest, accounting for approximately one-third of all deletions seen within single-deletion patients (Pitceathly et al., 2012). The common deletion is 4.977Kb in length, between bases 8470 and 13447, or approximately 30% of the mitochondrial genome (Schon et al., 1989). The deletion's impact on mitochondrially encoded proteins is significant, removing five genes encoding mt-tRNAs, seven genes associated with subunits of CI, one gene encoding a CIV subunit and two genes encoding CV subunits (Shoffner et al., 1989).

Nuclear DNA variants

The mitochondrial genome encodes 13 proteins required for mitochondrial function; the nuclear genome additionally encodes over 1,000 mitochondrial proteins, representing 99% of the mitochondrial proteome (Rath et al., 2021). Nuclear-encoded mitochondrial proteins have a wide range of functions, including OXPHOS complex subunits and assembly factors, mtDNA replication and maintenance, and the TCA cycle, to name a few. All

nuclear-encoded mitochondrial proteins must be targeted and imported into the mitochondria. If any part of their production, translation, or transportation is disrupted OXPHOS function can be impacted. Over 300 genes have been associated with mitochondrial disease, of which only 36 are mitochondrially encoded (Thompson et al., 2020). Therefore, a nuclear DNA variant can profoundly impact almost any aspect of mitochondria, including the mitochondrial genome itself.

MtDNA maintenance disorders are a group of mitochondrial diseases that affect mtDNA replication and maintenance. They are primarily associated with nuclear-encoded variants and can result in mtDNA sequence perturbations, such as single, large-scale deletions. The associated genes include: *TWNK* (Spelbrink et al., 2001), *POLG* (Van Goethem et al., 2001), (Rouzier et al., 2012) and OPA1 (Amati-Bonneau et al., 2008). Each gene has a large phenotypic variance depending on the variant present and other environmental and genetic factors. Variants of *POLG* are associated with progressive external ophthalmoplegia (PEO), male infertility, Alpers syndrome, and Parkinsonism, with PEO being the most common. A description of known phenotypes and associated *POLG* variant can be found in Longley et al. (2005).

1.1.8 Genetic bottleneck effect

Large inter-generational variation in the mitochondrial genome was first noted in Holstein cows and led to the proposal of a genetic 'bottleneck' (Ashley et al., 1989; Olivo et al., 1983); later work showed the same phenomena in many other species, including mice (Cao et al., 2007; Cree et al., 2008; Jenuth et al., 1996) and humans (Chinnery et al., 2000; Wilson et al., 2016). The genetic bottleneck creates a large amount of uncertainty in the proportion of variant mtDNA inherited from a mother. Asymptomatic mothers, with low pathogenic variant loads, are known to have offspring with variant loads exceeding 70% (Larsson et al., 1992; Pallotti et al., 2014). The bottleneck creates genetic variation between mature oocytes, resulting in variation in the mitochondrial genomes of offspring. Thus, the bottleneck directly impacts the inherited mtDNA level of heteroplasmic variants. Despite its significance, the genetic bottleneck and the biological mechanisms behind it remain poorly understood, with studies finding contradictory evidence.

One theory is that the bottleneck is caused by a drastic reduction in mtDNA copy number within primordial germ cells (PGCs), oocyte precursor cells, during the early stages of oogenesis (Floros et al., 2018). Using a mouse model, Cree et al. (2008) showed that mtDNA copy number fell from $\approx 4,000$ per cell before embryo implantation to ≈ 200 within PGCs 7.5 d.p.c. (days *post coitum*), an estimate in agreement with Jenuth et al. (1996). During PGC differentiation, Cree *et al.* found copy number to increase to $\approx 1,500$ in primary oocytes 14.5 d.p.c.. The variant-load variation explained by copy number reduction was estimated to be 70% via a mathematical model of fertilisation to oocyte development, and the remaining variation explained by the proliferation of mtDNA during PGC cell divisions (Cree et al., 2008). In contrast, some works find no evidence of a reduction in copy number during oogenesis (Cao et al., 2007, 2009) and suggest the inter-generational variation is caused by sub-populations of replicating and non-replicating mtDNA. Adding to the ambiguity, Wai *et al.* find that copy number significantly reduces in the early stages of oogenesis (Wai et al., 2008). However, they found no

significant difference in variant-load variation between any two stages of oogenesis. Wai *et al.* did, however, find a significant difference in genotypic variation between high copy number germ cells (mature ovulated oocytes, primary oocytes in secondary follicles), copy number $\geq 10,000$, and low copy number germ cells (PGCs, oogonia, primary oocytes, in primordial follicles), copy number $< 10,000$, concluding that the genotypic variation is the result of replicating (and non-replicating) mtDNA subpopulations. Whether the cause is a copy number depletion or replicating subpopulations, the reduction in the effective population size has a significant impact on the inherited variant mtDNA level.

1.1.9 Threshold effect

For pathogenic mtDNA variants, heteroplasmy is measured by the percentage of mtDNA molecules which differ from wild-type mtDNA, referred to as variant load. Continuous mtDNA turnover causes the variant load to be dynamic, and clonal expansion is the process by which the proportion of pathogenic variant mtDNA molecules changes over time to become the dominant species within a cell. The effect that mtDNA population dynamics have on cellular function is based on the 'threshold effect'. This principle states that a cell becomes OXPHOS deficient if the proportion of pathogenic mtDNA surpasses a biochemical threshold (Rossignol et al., 2003), see Figure 1.8.

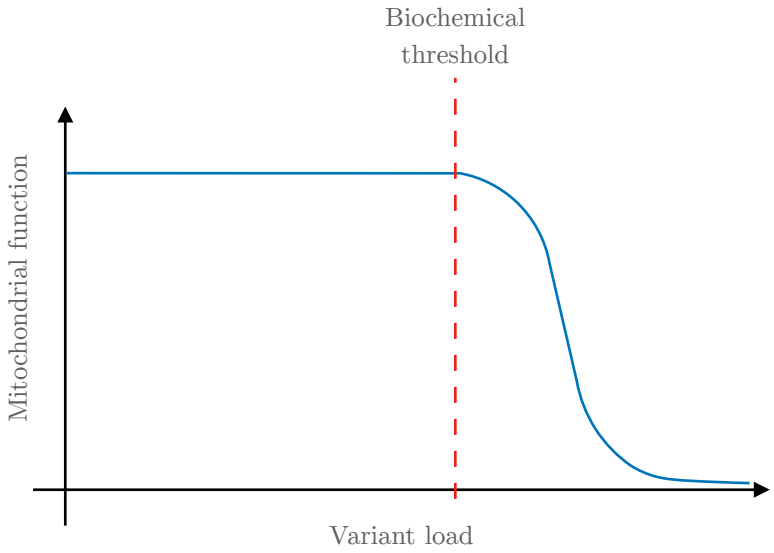


Figure 1.8: **Representation of the threshold effect**. How the threshold effect proposes mitochondrial function changes due to the mtDNA variant load within a single cell.

Moslemi et al. (1998) investigated the pathogenic threshold within six patients, across two families, with diagnosed m.8344A>G mtDNA single-point variant. Skeletal muscle fibre (myofibre) sections were taken, and COX deficient myofibres were identified through histochemical staining and manual classification. Polymerase chain reaction (PCR) analysis revealed the proportion of wild-type and variant mtDNA at a single myofibre level. A logistic regression model was fitted to the data, with the response variable being COX deficient status and a single explanatory variable of variant load; the pathogenic threshold was estimated by the variant load with a 50% probability of COX deficiency. For the six patients, the threshold estimates ranged from 95.3% to 97.7%, and no significant

difference in threshold was found between the two families.

Other work has shown the pathogenic threshold to be lower for single deletions. Hayashi et al. (1991) combined the mtDNA of patients with large single-deletions, and CPEO, with HeLa cells and isolated *trans*mitochondrial cybrid clones. When the mutation load passed 60%, the cybrid cells showed an inhibition of mt-tRNA translation and COX deficiency. Rocha et al. (2018) compared single-cell variant loads with OXPHOS protein abundances in six patients with characterised single large-scale deletions. The pathogenic thresholds were estimated to be between 56% to 82% in CI and 57% to 92% in CIV, and showed a significant relationship with deletion size and location.

Rossignol et al. (1999) used a rat model to compare the respiratory rates with inhibited levels of OXPHOS complexes; CI, CIII, and CIV. In muscle tissue, the thresholds (and S.E.) were estimated to be 74.5 (5.1), 85.2 (2.3), and 66.8 (4.4), respectively. Hernández-Ainsa et al. (2022) used a cybrid cell model to investigate the effect of mtDNA deletions in patients with Pearson's syndrome. CIV quantity and activity, and ATP levels were significantly reduced when the cellular variant load passed ≈60%. Mitochondrial respiration decreased in cybrids when their mutation load exceeded ≈70%.

1.1.10 Copy number regulation

Schultz et al. (1998) found that a decrease in mtDNA copy number leads to an increase in mtDNA replication; however, the mechanisms which control this replicative increase are unknown. Tang et al. (2000) suggested that copy number control is related to deoxynucleotide (dNTP) pools, essentially maintaining a relatively constant mtDNA mass within a cell. Clay Montier et al. (2009) suggest that the cell's dynamic ATP needs control the copy number. Increased mtDNA replication (and copy number) is signalled by increased cellular ATP requirements. However, evidence for this has yet to be found.

1.2 Clonal expansion of mtDNA

Variant mtDNA proportions can rise and fall over time, the dynamics of which are dependent on cell type and mtDNA variant. A number of studies have shown that the m.3243A>G mutation decreases in mitotic cells with time (de Laat et al., 2012; Grady et al., 2018; Sue et al., 1998). In contrast, the proportion of large-scale single deletions within skeletal muscle fibres increases with age (Bua et al., 2006). MtDNA dynamics have also been shown to be spatially dependent, with variant mtDNA spreading both cross-sectionally and longitudinally along a muscle fibre at differing rates (Bua et al., 2006; Vincent et al., 2018).

Clonal expansion is the term used to describe the changing population dynamics of mtDNA within a single cell and, consequently, how variant mtDNA becomes the dominant species. It is thought to be a factor in the progression of some forms of mitochondrial disease. However, the mechanisms governing clonal expansion are not known. Several theories have been proposed, but no single theory has been unilaterally agreed upon by

experts, nor has one been able to explain the wide range of experimentally observed data.

In vitro cell models of clonal expansion can be useful. For example, they allow the use of patient-specific human cells and the ability to monitor clonal expansion during cell development and replication. Such cell cultures have shown that SNV variants can arise, reach 25% of the mtDNA population, and go extinct within 21 weeks (Ludwig et al., 2019). However, they pose several problems. *In vivo* rates may differ from those *in vivo*, and it is difficult (practically impossible) to produce time scales similar to the human life span. In addition, many biological mechanisms act in a cell-type-specific manner (Herbers et al., 2019), and mitotic cells used within a culture do not reflect the mtDNA dynamics of post-mitotic cells, such as skeletal muscle fibres or neurons.

Animal models, often mice, allow the study of the tissue-specific aspects of mitochondrial disease by the collection of a range of tissues at varying stages of development and life, a difficult task when using human subjects (Burr & Chinnery, 2024). Specifically, animal models allow multiple measurements to be made within post-mitotic tissue (Dunn et al., 2012; Tyynismaa & Suomalainen, 2009). Despite their benefits, animal models suffer several limitations. Animal biology differs from human biology, which can impact mitochondrial dynamics. For instance, initially, human and mouse embryos have relatively similar mtDNA copy numbers, yet humans have many times the number of cells. The increased mtDNA proliferation and cellular division required during human development would lead to a higher inter-cellular variant-load variation in a stochastic system (Stewart, 2021). Additionally, the lower mtDNA copy number within mouse cells can increase the rate at which somatic mtDNA variants become dominant within a single cell (Elson et al., 2001). Biological differences in species and other considerations may result in less severe phenotypes in mice compared to human patients (Tyynismaa et al., 2005; van Riesen et al., 2006).

Computer models of mitochondrial DNA dynamics are gaining popularity within the literature (Chinnery & Samuels, 1999; Johnston & Jones, 2016; Kowald & Kirkwood, 2013; Lakshmanan et al., 2018). Mathematical models of clonal expansion will be discussed further in Chapter 5. These are simplified models of the biological system, focusing on the phenomena hypothesised to be the system's driving force. They offer additional benefits such as producing realistic time scales, knowing the system state at any point, and initialising the system with any state. Notably, an accurate mathematical model of clonal expansion may highlight biological mechanisms for further investigation and make patient-specific disease progression predictions. They do, however, have some disadvantages. In particular, assumptions about the biological system must be made, and model parameter values must be chosen from contested values in the literature or inferred. Slight changes to these values may have a drastic impact on the model output.

1.2.1 Clonal expansion theories

Random genetic drift

Random genetic drift (RGD) assumes no selective, replicative or any other advantage between mtDNA species and relies solely on the stochastic replication and degradation of mtDNA to explain the population dynamics (Chinnery & Samuels, 1999). Several other mathematical models and biological theories of clonal expansion incorporate no replicative

advantage between mtDNA species but impose other assumptions. For example, Capps et al. (2003) used a model of no replicative advantage to compare copy number control mechanisms.

Early mathematical modelling showed how variant load can reach pathogenic levels throughout the human life span via *de novo* mutation events (Chinnery & Samuels, 1999; Elson et al., 2001). Kowald and Kirkwood (2013) demonstrated that RGD cannot produce mutation loads seen in experimental data of short-lived species, such as rodents. Henderson et al. (2009) showed that the accumulation of single, large-scale deletions in neurons can be explained by RGD and Johnston et al. (2015) compared the mechanism governing the genetic bottleneck under the assumption of no replicative advantage. Both Henderson et al. (2009) and Johnston et al. (2015) inferred model parameters via Bayesian inference; however, often parameter values are fixed to values found in the literature (Capps et al., 2003; Elson et al., 2001; Insalata et al., 2022; Kowald & Kirkwood, 2013).

Parameter values within the literature, estimated from observed data, are scarce and come with a high degree of uncertainty. The uncertainty is significant as varying parameter values can drastically alter the resulting dynamics of a model. For example, experimental data of COX deficiency of neurons in the Substantia Nigra show a much higher deficiency level than predicted by Elson et al. (2001) (Itoh et al., 1996; Kraytsberg et al., 2006). However, a lower mtDNA half-life, higher mutation probability, or lower pathogenic threshold would result in a higher proportion of COX deficient cells within the same time frame, and random genetic drift could be capable of explaining the data. Parameter values within the literature are further discussed in Chapter 5.1.6.

Survival of the smallest

Survival of the smallest assumes that variant mtDNA, with large-scale deletions, replicate at a higher rate due to their decreased size (Wallace, 1989). The increased replication rate gives variant mtDNA a selective advantage, allowing them to clonally expand to high levels within a cell. Diaz et al. (2002) demonstrate that under relaxed replication, single-deletion mtDNA repopulates a cell faster than wild-type. Kowald et al. (2014) developed a mathematical model focused on the effect of single-deletions on replication time and showed that the replicative advantage is capable of explaining the accumulation of COX deficient cells in both long- and short-lived species. However, many have questioned the theory's explanation. Significantly, it does not explain the increase in SNVs with age (Greaves et al., 2014; Weber et al., 1997). In addition, Gitschlag et al. (2016) found that deletions of different sizes accumulate at the same rate. Hayashi et al. (1991) provided further evidence against the hypothesis by showing that the cell doubling time is higher for cells with mtDNA deletions. The varying evidence in the literature is thus inconclusive, indicating that this theory is not the key component driving clonal expansion. Nevertheless, the theory (or versions of it) is popular within the literature.

Perinuclear niche

Vincent et al. (2018) investigated subcellular distributions of COX deficiency in skeletal muscle fibres (myofibres). They found that COX deficiency and increased mtDNA copy

number localised to nuclei, which led to the proposed 'proliferative perinuclear niche' hypothesis. The hypothesis states that mtDNA replication rates are higher within the perinuclear niche, due to its dependence on nuclear-encoded proteins. Therefore, mtDNA variants which form in the niche affect OXPHOS locally before spreading throughout the myofibre. The theory is primarily focused on OXPHOS deficiency within myofibres, whose size allows sections of the myofibre to be OXPHOS deficient while other sections remain OXPHOS functional (Elson et al., 2002). Similarly to RGD, the theory assumes no replicative advantage between mtDNA species. Mathematical modelling of the perinuclear niche (PNN) has not been completed within the literature, likely due to its relatively recent proposal and the computational cost of a spatially dependent model.

Other theories

Insalata et al. (2022) proposed stochastic survival of the densest (SSD), a spatially dependent mathematical model with no replicative advantage between mtDNA species, to explain how variant mtDNA spreads along the length of a myofibre. Myofibres were modelled as a series of adjacent, independent compartments, which mtDNA can migrate between. Using parameter values from the literature, model predictions were compared to rhesus monkey data and showed a better fit than a replicative advantage model. However, the authors acknowledged a high amount of uncertainty in some parameter value choices.

Some theories proposed are based on the effect of free radicals. de Grey (1997) proposed 'survival of the slowest', a model of clonal expansion in which a negative feedback loop reduces the rate at which variant mtDNA is degraded. Their hypothesis states that IMM dysfunction is caused by free radicals, produced during ATP synthesis, and leads to the degradation of the mitochondrion (and its mtDNA). The reduced ATP synthesis of mitochondria harbouring variant mtDNA leads to reduced IMM damage and prolonged mitochondrial life. The theory, therefore, induces a selective pressure against the degradation of variant mtDNA, allowing the variant load to grow within a cell.

Before moving on to the analysis of datasets and the development of mathematical models to investigate the driving forces of clonal expansion, it is necessary to first discuss some of the statistical background required for this work. In the remainder of this chapter and Chapter 2, some key concepts in statistics are introduced, which underpin later work. Namely, Bayes' theorem, the key result responsible for Bayesian statistics, which is the method of statistical inference used throughout this thesis. Markov chains are also introduced, being a key concept required for Bayesian inference and models of mtDNA population dynamics. Direct applications and uses become clearer in later chapters, where these methods are applied, although some examples are given here.

1.3 Probability and Bayes' theorem

Before discussing Bayes' theorem, some key results in probability are introduced. Let A and B be two events, and their individual probabilities of occurring be denoted $\Pr(A)$ and $\Pr(B)$. The probability that both events occur is denoted $\Pr(A \cap B)$, and the conditional probability of event A occurring given that B has occurred is denoted $\Pr(A|B)$.

1.3.1 Conditional probability

The law of conditional probability states

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}, \tag{1.4}$$

note that $\Pr(A \cap B) = \Pr(B \cap A)$ implies that $\Pr(A \cap B)$ is, therefore, equal to both $\Pr(A|B)\Pr(B)$ and $\Pr(B|A)\Pr(A)$, and so

$$\Pr(A|B)\Pr(B) = \Pr(B|A)\Pr(A). \tag{1.5}$$

1.3.2 Total probability

Discrete

Let $\mathbf{B} = \{B_i : i = 1, \ldots, n\}$ be a finite, or countably infinite, set of mutually exclusive and collectively exhaustive events i.e. $\Pr(B_i \cap B_j) = 0$ for all $i \neq j$, and $\sum_{i=1}^{n} \Pr(B_i) = 1.0$. Then, the law of total probability states that the marginal probability of event A is equal to the sum of all unions,

$$\Pr(A) = \sum_i \Pr(A \cap B_i). \tag{1.6}$$

Substituting the law of conditional probability, this becomes

$$\Pr(A) = \sum_i \Pr(A|B_i)\Pr(B_i). \tag{1.7}$$

Continuous

Extending the law to continuous random variables is relatively easy. Let X be a random variable which can take any value in the set $\mathcal{X}$, referred to as the support, and its density be described by a probability density function (PDF) $f_X(x)$. The summation in the discrete version, Eq. 1.7, becomes an integral over all possible values of X:

$$\Pr(A) = \int_{\mathcal{X}} \Pr(A|X = x) f_X(x) dx. \tag{1.8}$$

1.4 Bayesian statistics

Unlike frequentist statisticians, Bayesian statisticians believe that there is no "true" value of the parameters in a statistical model. Instead, Bayesian methods summarise parameter beliefs after observing data (*a posteriori*) by a probability distribution, giving more weight to more likely values. Parameter beliefs before observing any data (*a priori*) are called prior beliefs, and are similarly summarised by a probability distribution. Prior beliefs can be as vague or informed as required to reflect the beliefs of relevant experts. The complete lack of prior knowledge can be reflected in extremely vague prior beliefs, with a high variance. In contrast, a large amount of prior knowledge from previous work may result in well-informed priors with high precision. Parameter beliefs are updated by combining prior beliefs and new evidence presented from a dataset. Bayesian methodology is fully probabilistic and considers model parameters, hidden states, as well as missing and observed data in the same vein. The coherent treatment of model parameters, data, and

hidden states has allowed Bayesian methods to be used in a wide range of inference problems. Some of which are discussed in the next paragraph, and more generally throughout this thesis.

Bayesian methods have gained popularity in all aspects of research. Traditionally, the limiting factor to conducting Bayesian analysis was computational expense. However, the rise of computing power is changing this. Applications within the biological and medical sciences are becoming increasingly common, enabling the use of complex models to explain complex biological systems with incomplete and messy datasets. Reviews by Wilkinson and Yau *et al.* demonstrate areas where Bayesian methods have been applied, including biological sequence analysis and microarray analysis (Wilkinson, 2007; Yau & Campbell, 2019). In other areas, Rickett et al. (2015) demonstrated that Bayesian methods can detect subtle differences in bacterial growth rates, and Russel-Buckland *et al.* used Bayesian methods within systems biology, inferring parameters of a mathematical model (Russell-Buckland et al., 2019). Mathematical models of biological systems commonly use Bayesian inference schemes, where uncertainty in unobserved data and hidden states can be coherently accounted for (Fisher et al., 2022; Georgoulas et al., 2016; Golightly & Wilkinson, 2011; Henderson et al., 2010).

1.4.1 Bayes' theorem

Bayes' theorem is the foundation of Bayesian statistics and provides the formula to find posterior beliefs. It follows immediately from the law of conditional probability, Eq. 1.4, by rearranging Eq. 1.5.

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}. \tag{1.9}$$

For practical use, this is viewed with A and B being parameter beliefs and data. Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ be a set of independent observations and $f_X(x|\theta)$ be a statistical model that is believed to explain the data. The model, $f_X(x|\theta)$, is dependent on a set of parameters, θ, which are unknown and must be inferred. Let Θ denote the support of the model parameters, θ. For example, in a simple linear model, $\theta = (m, c)^{\mathrm{T}}$ indicating the slope and intercept of the linear equation, and its support is the 2-dimensional real-plane, $\mathbb{R}^2$. The data-likelihood, $f(\boldsymbol{x}|\theta)$, is the probability of observing $\boldsymbol{x}$ given a set of parameter values, θ, and is the product of the model density evaluated at each observation given θ,

$$f(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} f_X(x_i|\theta). \tag{1.10}$$

The probability density summarising the prior beliefs is often denoted $p(\theta)$. Substituting these elements into Eq. 1.9 and using the law of total probability, Eq. 1.8, to write the denominator with known entities, we get Bayes' theorem;

$$p(\theta|\boldsymbol{x}) = \frac{p(\theta) f(\boldsymbol{x}|\theta)}{\int_\Theta f(\boldsymbol{x}|\theta) p(\theta) d\theta}. \tag{1.11}$$

Bayesian methods concern learning parameter values, θ. We, therefore, think of functions in Bayes' theorem not dependent on θ. The equation is often rewritten using the

parameter-likelihood function; its definition is identical to the data-likelihood function, but the data are considered known and, therefore, the given entity. Let $L(\theta|\boldsymbol{x})$ be the parameter-likelihood function and

$$L\left(\theta|\boldsymbol{x}\right) = f(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} f_X(x_i|\theta). \tag{1.12}$$

Bayes' theorem can then be rewritten using the parameter-likelihood

$$p\left(\theta|\boldsymbol{x}\right) = \frac{p(\theta)L\left(\theta|\boldsymbol{x}\right)}{\int_\Theta L(\theta|\boldsymbol{x})p(\theta)d\theta}. \tag{1.13}$$

The denominator, $\int_\Theta f(\theta|\boldsymbol{x})p(\theta)d\theta$, is referred to as the marginal likelihood and is a constant for a given dataset, $\boldsymbol{x}$. It is a scaling constant and ensures the posterior probability density integrates to 1.0. This can also be seen as the integrand, the function within the integral, is equal to the numerator in Eq. 1.13. As it is a constant, Bayes' theorem is often simplified to express the posterior up to proportionality,

$$p\left(\theta|\boldsymbol{x}\right) \propto p(\theta)L\left(\theta|\boldsymbol{x}\right). \tag{1.14}$$

The posterior distribution, up to proportionality, is therefore the prior density function multiplied by the likelihood function, the result is a function of θ. How Eq. 1.14 is used to find a posterior distribution is discussed next.

1.4.2 Conjugate analysis

In simple cases, the posterior distribution can be found analytically when the multiplication, $p(\theta) \times L(\theta|\boldsymbol{x})$, evaluates to be a known probability density up to proportionality. This is called conjugate analysis, an example is shown below. Specific models and prior beliefs are known to be conjugate. Although restricting, a variety of models and prior beliefs can be formed this way. Conjugacy is helpful because there is (almost) no computational cost to calculating the posterior distribution, only a human cost of doing the maths. However, conjugate models tend to be simple and so are not applicable in many situations. In Chapters 1.4.4 and 2.1 non-conjugate methods are discussed in more detail.

1.4.3 Conjugate analysis example

Here, an example of conjugate analysis is presented, implementing the theory presented in Chapter 1.4. Suppose $\boldsymbol{x}$ is a vector of n independent observations, and we wish to model each observation as coming from a normal distribution. Further, assume the data's population mean, μ, is known but the model precision, τ, must be inferred. For $i = 1, \ldots, n$, the model is

$$x_i|\mu, \tau \sim \mathrm{N}\left(\mu, \tau^{-1}\right). \tag{1.15}$$

The likelihood function is the product of the model densities evaluated at each observation. The model here is a normal density, and so the likelihood is the product of n normal

densities, with mean μ and precision τ, evaluated at x_i for $i = 1, 2, \ldots, n$. Continuing the notation from before, that is

$$L(\tau|\boldsymbol{x}) = \prod_{i=1}^{n} \sqrt{\frac{\tau}{2\pi}} \exp\left[-\frac{1}{2\tau}(x_i - \mu)^2\right]. \tag{1.16}$$

After consulting the literature and experts in the field, it is agreed that prior beliefs for the precision, τ, are summarised by a gamma distribution with shape and rate a and b,

$$\tau \sim \text{Ga}\,(a, b)\,. \tag{1.17}$$

The prior density function is, therefore, the gamma density function with shape and rate: a and b,

$$p(\tau) = \frac{b^a}{\Gamma(a)} \tau^{a-1} \exp\left(-b\tau\right). \tag{1.18}$$

The posterior distribution $p(\tau|\boldsymbol{x})$ can then be found up to proportionality by multiplying the prior density and likelihood functions.

$$\begin{aligned} p(\tau|\boldsymbol{x}) &\propto \frac{b^a}{\Gamma(a)} \tau^{a-1} \exp\left(-b\tau\right) \times \prod_{i=1}^{n} \sqrt{\frac{\tau}{2\pi}} \exp\left[-\frac{1}{2\tau}(x_i - \mu)^2\right] \\ &\propto \tau^{a+n/2-1} \exp\left\{-\left[b + \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right]\tau\right\} \end{aligned} \tag{1.19}$$

The form of $p(\tau|\boldsymbol{x})$, up to proportionality, has the same form as a gamma distribution and, therefore, τ must follow a gamma distribution *a posteriori*. Hence, the posterior belief for τ is summarised by a gamma distribution with shape $a + n/2 - 1$ and rate $\left[b + \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right]$, and the posterior distribution can be written as

$$\tau|\boldsymbol{x} \sim \text{Ga}\left(a + n/2 - 1, b + \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right). \tag{1.20}$$

In this example, the posterior distribution can be found analytically and is a known distribution, and, therefore, the model is conjugate.

1.4.4 Non-conjugate analysis

Although useful, there are relatively few models which are fully conjugate. However, parameter values can still be inferred if this is not the case. For non-conjugate analysis, the posterior cannot be found analytically, although it can be summarised by a large number of random draws. Suppose, we are tasked with finding the mean and variance of a normal distribution whose parameter values are unknown, however it is possible to randomly sample from the distribution. If a large number of samples are drawn, these can be inspected to gain an understanding of the mean and variance. Similarly, in non-conjugate analysis, the posterior distribution is repeatedly sampled from and those posterior draws are analysed. How the draws are generated varies depending on the model in question; more complex models often require more complex algorithms and more computationally expensive methods to sample from the posterior. In the next chapter, Chapter 1.5, some of the concepts underpinning the algorithms designed to sample from such posterior distributions are introduced. Specific sampling methods are discussed in Chapter 2.1.

1.5 Markov chains

A Markov chain is a sequence of random variables whose current state depends only on the previous state and is independent of all states before that. This condition is known as the memoryless or Markov property. Their applications span many areas of statistics, including parameter inference and statistical modelling; both areas will be demonstrated in Chapters 2.1 and 5 of this thesis.

The memoryless property implies that the current state is a random variable whose possible values are described by a probability density dependent on the previous state. Being sequential, Markov chains are usually thought to progress with time, with updates occurring at discrete or continuous intervals. The state space of a Markov chain, the possible values the random variables, can be either multi- or univariate and be discrete or continuous. In the remainder of this chapter, discrete-time Markov chains are introduced to highlight some of the mathematical theories behind Markov chains.

1.5.1 Discrete state-space Markov chain

Let $\{X_t : t = 0, 1, 2, \dots\}$ be a sequence of random variables which form a discrete-time, discrete state-space, stochastic process. That is, a Markov chain which updates at discrete and fixed time increments and whose value can be a finite number of possibilities. Let $\mathcal{S} = \{S_1, S_2, \dots, S_m\}$ denote the set of m possible values that X_t can be i.e. $X_t \in \mathcal{S}$ for all $t = 0, 1, 2, \dots$. Assume that the probability of moving between any two states, $S_i, S_j \in \mathcal{S} i \neq j$, is known and independent of time and be denoted p_{ij} such that $\Pr(X_t = S_j | X_{t-1} = S_i) = p_{ij}$, for all ≥ 0 and for all $1 \leq i, j \leq m$.

Given that at time index t the previous state is $X_{t-1} = S_i$, the transition probabilities $\{p_{i1}, p_{i2}, \dots, p_{im}\}$ form a probability mass function (PMF) describing the distribution of X_t. The system can, therefore, be forward-simulated with a known starting condition by recursively sampling from the appropriate PMF. The next section discusses a discrete state-space example to solidify some of their key concepts.

Discrete state-space example

Suppose a system can be in four distinct states, $\{S_1, S_2, S_3, S_4\}$, and the probabilities of moving between states are known. The directed graph in Figure 1.9 summarises the system. A directed graph is a convenient way to describe a Markov chain when the state space is relatively small. For example, by inspection of the graph, it can be seen that when the system is in S_1, the next state can be either S_2 or S_3, with equal probability.

Intuitively, within the system described in Figure 1.9, the probability distribution of the current system state depends only on the previous state and not any state before that. It can also be seen in the directed graph that if the system reaches S_4, it will remain in this state indefinitely. This is an example of a stationary distribution of the Markov chain. Given an infinite amount of time, the system will reach S_4 with probability 1, and once it does, it can not escape. In this example, the stationary distribution has all weight associated with S_4 and none to the other states. Although in more complex systems, this may not be the case. For example, the system depicted in Figure 1.10, shows a stationary

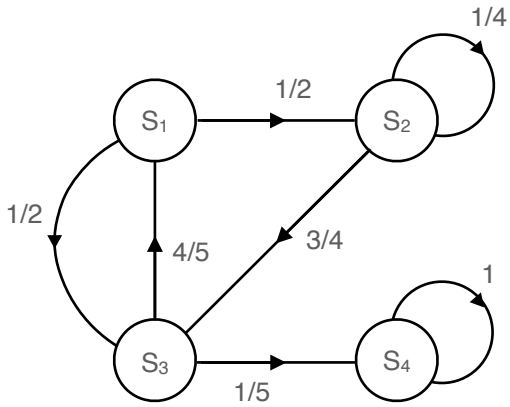


Figure 1.9: **Discrete state-space Markov chain**. A directed graph showing the possible transitions and their probabilities of the Markov chain. The directed edges between nodes indicate a possible transition over a discrete time interval. For example, moving from S_2 to S_3, over one time increment, is possible, but not from S_3 to S_2. Each transition is associated with probability, seen adjacent to the edge connecting to nodes.

distribution split across two states; S_4 and S_5. By inspection of the directed graph, it can be seen that once the system enters S_4, it will oscillate between S_4 and S_5 indefinitely.

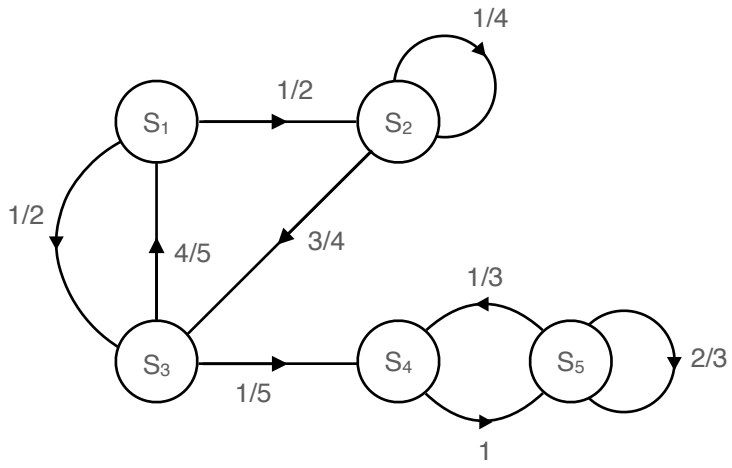


Figure 1.10: **Discrete state-space Markov chain with non-singular stationary distribution**. A directed graph showing the possible transitions and their probabilities of the Markov chain. The connections between nodes indicate that it is possible to transition between the two states over a discrete time interval. The arrow of these connections indicates in which direction it is possible to move. The stationary distribution of this Markov chain has equal weight associated with states S_4 and S_5 but zero weight associated with the remaining nodes.

A stationary distribution represents the time-limiting behaviour of the system and the proportion of time the chain spends in each state as the system time tends to infinity. If one exists, the system's stationary distribution can be calculated analytically for discrete-time, discrete-state-space Markov chains, see Chapter 1.5.2.

As the size of the state space increases, summarising the possible transitions and their probabilities within a directed graph becomes inconvenient. Therefore, this information is often summarised in a transition matrix, whose (i, j)-th element is the probability of transitioning from state i to j, p_{ij}. For the system described in Figure 1.10, the transition

matrix, P, is

$$P = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 \\ 0 & 1/4 & 3/4 & 0 & 0 \\ 4/5 & 0 & 0 & 1/5 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1/3 & 2/3 \end{pmatrix}.$$

The i-th row of the transition matrix is the PMF of the current state if the system was previously in the i-th state. For example, the 3rd row, $(4/5, 0, 0, 1/5, 0)$ gives probabilities $4/5$ and $1/5$ of being in the 1st and 4th states, respectively, and all other states having a zero probability. This is reflected in Figure 1.10 where the only directed edges leaving S_3 are to S_1 and S_4, with the appropriate weights.

Transition matrices

Let $\{X_t : t = 0, 1, 2, \dots\}$ be a discrete state-space stochastic process, such that $X_t \in \{S_1, S_2, \dots, S_m\}$ for all ≥ 0, and P be it's transition matrix. The (i, j)-th element of P is, therefore, the probability of transition from S_i-th to the S_j-th state,

$$(P)_{i,j} = \Pr\left(X_t = S_j | X_{t-1} = S_i\right) = p_{ij}.$$

If the previous system state, $X_{t-1} = S_i$, is written as a length m, unit row-vector $\boldsymbol{\pi}_{t-1}$, whose i-th element is 1.0 and all other elements 0.0, i.e. a PMF with all weight associated with a single value. Then, the distribution of the current state, $\boldsymbol{\pi}_t$, can be found by multiplying the unit vector and the transition matrix. Which returns the i-th row of the transition matrix, as would be expected,

$$\begin{aligned} \boldsymbol{\pi}_t &= \boldsymbol{\pi}_{t-1} P, \\ &= (0, \dots, 0, 1, 0, \dots, 0)\, P, \\ &= (p_{i1}, p_{i2}, \dots, p_{im}). \end{aligned}$$

Similarly, if uncertainty exists in the previous state's value and is described by a non-singular PMF, $\boldsymbol{\pi}_{t-1}$, then the distribution of the current system state, $\boldsymbol{\pi}_t$ is found through matrix multiplication,

$$\boldsymbol{\pi}_t = \boldsymbol{\pi}_{t-1} P.$$

The distribution of the future state, $\boldsymbol{\pi}_{t+1}$, dependent on X_{t-1}, can also be calculated by substitution,

$$\boldsymbol{\pi}_{t+1} = \boldsymbol{\pi}_t P = \left[\boldsymbol{\pi}_{t-1} P\right] P = \boldsymbol{\pi}_{t-1} P^2.$$

Recursive substitution allows the distribution of system state to be calculated for any time step into the future,

$$\boldsymbol{\pi}_{t+k} = \boldsymbol{\pi}_{t-1} P^{k+1}.$$

Here, it is shown that not only is it possible to forward simulate the system by randomly drawing the next state from the appropriate distribution, but it is also possible to calculate the PMF of any future state, given an initial state.

1.5.2 Stationary distribution

As mentioned, stationary distributions are an important aspect of Bayesian inference. When models are not fully conjugate, posterior distributions are sampled by constructing Markov chains whose stationary distributions are the desired posterior density, see Chapter 2.1. Here, stationary distributions for discrete state-space Markov chains are discussed.

Once a Markov chain reaches its stationary distribution, $\boldsymbol{\pi}$, it will remain in this distribution indefinitely. Therefore, after each update of the Markov chain, the value of the random variable, X_t, follows the same distribution, and the stationary distribution must satisfy the equation

$$\boldsymbol{\pi} = \boldsymbol{\pi} P, \tag{1.21}$$

where P is the transition matrix and $\boldsymbol{\pi}$ a row vector. The stationary distribution can be found by solving Eq. 1.21. The existence and uniqueness of such a distribution require certain criteria of the Markov chain, which will not be discussed here; a detailed discussion can be found in Norris (1997). Importantly, Markov chains can be constructed with such requirements to ensure a unique stationary distribution exists.

Discrete state-space example continued

Recall the discrete time, discrete state-space Markov chain depicted in Figure 1.10, the system has five possible states, $\{S_1, S_2, S_3, S_4, S_5\}$, and a transition matrix,

$$P = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 \\ 0 & 1/4 & 3/4 & 0 & 0 \\ 4/5 & 0 & 0 & 1/5 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1/3 & 2/3 \end{pmatrix}.$$

By inspecting the directed graph, Figure 1.10, it can intuitively be seen that once the system reaches S_4, it remains in S_4 or S_5 indefinitely. However, the distribution of time spent in each state within the limit is unclear. The stationary distribution can be found by solving the system of equations from Eq. 1.21. The system equations can be rearranged such that,

$$\begin{aligned} \boldsymbol{\pi} &= \boldsymbol{\pi} P \\ \boldsymbol{\pi}\left(I_5 - P\right) &= \mathbf{0} \end{aligned}$$

where I_5 is the 5×5 identity matrix, which has 1's on the diagonal and 0's everywhere else, and $\mathbf{0}$ is a 5-dimensional row vector with all zero elements. Both sides of the equation are transposed before substituting the transition matrix and calculating the matrix multiplication. This only makes the vectors easier to read on the page and does not affect

the solution. The stationary distribution can then be calculated as follows

$$(I_5 - P)^{\mathrm{T}} \boldsymbol{\pi}^{\mathrm{T}} = \mathbf{0}^{\mathrm{T}},$$

$$\begin{pmatrix} 1-0 & 0 & -4/5 & 0 & 0 \\ -1/2 & 1-1/4 & 0 & 0 & 0 \\ -1/2 & -3/4 & 1-0 & 0 & 0 \\ 0 & 0 & 0 & 1-0 & -1/3 \\ 0 & 0 & 0 & -1 & 1-2/3 \end{pmatrix} \begin{pmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \\ \pi_4 \\ \pi_5 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix},$$

$$\begin{pmatrix} \pi_1 - \frac{4}{5}\pi_3 \\ -\frac{1}{2}\pi_1 + \frac{3}{4}\pi_2 \\ -\frac{1}{2}\pi_1 - \frac{3}{4}\pi_2 + \pi_3 \\ -\frac{1}{5}\pi_3 + \pi_4 - \frac{1}{3}\pi_5 \\ -\pi_4 + \frac{1}{3}\pi_5 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Solving the system of equations, we find that $\pi_1 = \pi_2 = \pi_3 = 0$ and $\pi_4 = \frac{1}{3}\pi_5$. Recall that $\boldsymbol{\pi}$ is a PMF i.e. $\pi_i \geq 0.0$ for all i and $\sum_i \pi_i = 1.0$. Using these constraints values for π_4 and π_5 can be found;

$$\begin{aligned} \sum_{i=1}^{5} \pi_i &= 1.0, \\ \pi_4 + \pi_5 &= 1.0, \\ \frac{4}{3}\pi_5 &= 1.0, \\ \pi_5 &= \frac{3}{4}. \end{aligned}$$

This implies that $\pi_4 = \frac{1}{3}\left[\frac{3}{4}\right] = 1/4$, and so the stationary distribution of system is $\boldsymbol{\pi} = \left(0, 0, 0, \frac{1}{4}, \frac{3}{4}\right)$. Therefore, the Markov chain described by the directed graph in Figure 1.10 will spend 75% of its time in the state S_5 and the remainder in S_4, as the time limit tends to infinity.

Although this was a simple example, stationary distributions are essential in Markov chains and Bayesian inference. Specific algorithms which target the posterior density are discussed in Chapter 2.1. Stationary distributions are not discussed in further detail; a detailed review of theory and practice in Bayesian inference can be found in Gamerman and Lopes (2006). The following chapter introduces a real-world Markov chain example to highlight their use in mathematical modelling.

1.5.3 Cell culture population example

Consider the population of cells within a culture. A cell can divide, resulting in two daughter cells, or die, removing the cell from the population. Assuming that the events are independent of the current population and occur randomly, the population after the next event depends only on the current population and not any previous state. Thus, the culture's population could be modelled as a Markov chain. The state space of the chain is the set of whole numbers greater than or equal to zero, commonly denoted $\mathbb{N}$.

Let X be the current population of the system, the cell culture, and Y be the population after the next event. Also, let p_x be a known replication probability when $X = x$.

The transition probabilities of moving from the current state to any other in the state space can then be defined. Assume the current population $X = x > 0$, then

$$\begin{aligned}\Pr(Y = x + 1|X = x) &= p_x,\\ \Pr(Y = x - 1|X = x) &= (1.0 - p_x),\\ \Pr(|Y - x| > 1|X = x) &= 0.0.\end{aligned}$$

Note that if $X = 0$, there are no cells to divide or die, and the Markov chain cannot escape this state. As such, the transition probabilities require two additional probabilities to be comprehensive;

$$\begin{aligned}\Pr(Y = 0|X = 0) &= 1.0,\\ \Pr(Y > 0|X = 0) &= 0.0.\end{aligned}$$

The example hopefully highlights how Markov chains may appear in the real world. The probability of replication was purposefully kept vague in the example, as, in reality, this is likely to depend on resources available within the culture. The example models the population level of a cell culture stochastically, recreating the random fluctuations which would be seen in a real dataset. Further examples of such models are discussed in Chapter 2.6.5 and 2.6.6. Chapter 5 considers mtDNA populations within a single cell as stochastically evolving processes and develops mathematical models based on the principles of Markov chains to model their dynamics. Details of the mathematical theory underpinning the models used are discussed in Chapter 2.6.

1.5.4 Continuous state-space Markov chain

As mentioned, Markov chains are key to sampling from posterior distributions during Bayesian inference. However, only discrete state spaces have been discussed so far, and many models require inference of continuous-valued parameters. Here, continuous state-space Markov chains are briefly introduced. A detailed discussion of Markov chains, discrete and continuous state-space, can be found in Norris (1997).

Similarly to the discrete state-space version, the continuous state-space Markov chain possesses the memoryless property. However, due to the continuous state space, the transition probabilities can no longer be expressed exactly and must be considered using transition densities. Much like the difference between discrete and continuous random variables.

Let $\{X_t : t \geq 0\}$ be discrete-time stochastic process, and $X_t \in \mathcal{S}$ be a continuous state-space, such as $\mathbb{R}$. The memoryless property implies that the distribution of the current state, X_t, depends only on the previous, X_{t-1}, and is independent of $X_{t-2}, X_{t-3}, \ldots, X_0$. For a transition density f that is

$$f(X_t|X_{t-1}, X_{t-2}, \ldots, X_0) = f(X_t|X_{t-1}). \tag{1.22}$$

Continuous state-space Markov chains possess many of the same properties as their discrete counterparts. However, their transition probabilities cannot be succinctly summarised in a matrix. Importantly, stationary distributions of continuous state-space Markov chains can still exist, subject to some requirements (Norris, 1997), and they can be forward simulated. An example of a continuous state-space Markov chain is introduced here.

Continuous state-space example

Let $\{X_t : t = 0, 1, 2, \dots\}$ be real-valued random variables that are sequentially updated by adding normally distributed independent random noise. That is, given an initial value $X_0 = x_0$, for $t = 1, 2, \dots$

$$\begin{aligned} X_t &= X_{t-1} + \varepsilon_t, \\ \varepsilon_t &\sim \mathrm{N}\left(0, \sigma^2\right), \qquad \text{indep.} \end{aligned}$$

The random variables $\{X_t : t = 0, 1, 2, \dots\}$ form a continuous state-space, discrete-time Markov chain, referred to as a normal random walk. Whose transition density is a normal distribution centred around the previous state, with variance σ^2. Figure 1.11 shows a number of realisations from this model, with varying standard deviations of the additive random noise.

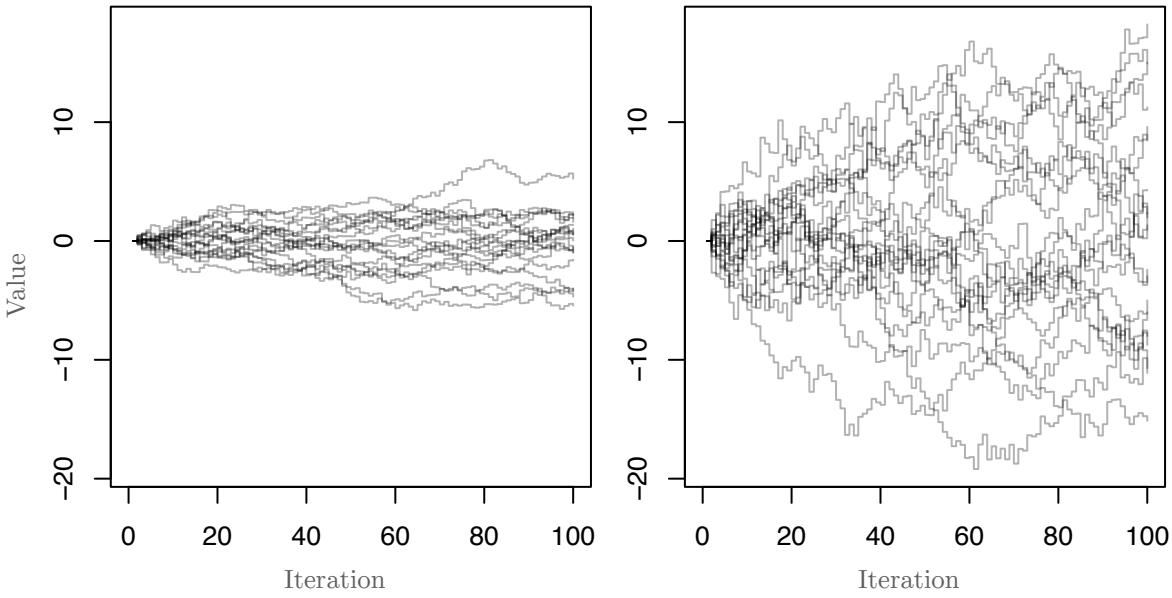


Figure 1.11: **Example simulations of normal random walk**. Twenty realisations from a normal random walk, with initial value $X_0 = 0$ (both) and standard deviations $\sigma = 0.25$ (left) and $\sigma = 1.0$ (right).

As presented here, the normal random walk is univariate; however, this need not be the case, and extending the Markov chain to be multivariate is relatively straightforward. Multivariate (and univariate) normal random walks are commonly used within Bayesian inference schemes; see Chapter 2.1.2.

1.6 Project aims

This thesis aims to use advanced statistical techniques to infer mtDNA dynamics and investigate whether biological theories of clonal expansion can be compared using Bayesian methods. To this end, mathematical models of clonal expansion are developed that reflect biological theories proposed in the literature. It is hoped that a comprehensive statistical analysis would allow for the robust comparison of the mathematical models, giving weight

to the biological theories they reflect.

Chapter 2

Methods

In this chapter, statistical models and inference schemes used in Chapters 3, 4, and 5 are discussed. The details of the models and parameter inference schemes are specific to each analysis and are therefore outlined in their respective chapters. However, the statistical theory and methods which are applicable throughout the thesis are discussed here. As mentioned, this thesis focuses on the use of Bayesian methods; therefore, this section begins with a discussion of algorithms for sampling from posterior distributions in non-conjugate analysis.

2.1 Markov chain Monte Carlo

Markov chain Monte Carlo (MCMC) is a common method used to sample from a posterior distribution when the analysis is not fully conjugate. The premise of the method is to construct a Markov chain whose stationary distribution is the posterior density. Once the Markov chain has converged to the posterior, any sample generated by the chain will be a realisation from the posterior distribution. The period before the chain reaches the stationary distribution is referred to as the 'burn-in' period, and these samples are removed before analysing the posteriors. Often, multiple chains are executed to ensure they converge to the same distribution, and checks are implemented to ensure this. Methods to assess convergence are discussed in Chapter 2.1.4.

In this chapter a generic statistical model, $f(x|\boldsymbol{\theta})$, is referred to, which depends on a set of m parameters, $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)^{\mathrm{T}}$. The support of $\boldsymbol{\theta}$ is denoted Θ. The parameters are assumed to be unknown and inferred for an observed dataset, $\boldsymbol{x}$.

2.1.1 Gibbs sampler

The Gibbs sampler, Algorithm 1, was introduced by Geman and Geman (1984), and has since become one of the most common methods to construct MCMC schemes. It is generally less computationally expensive than the Metropolis-Hastings algorithm, Algorithm 2. However, it imposes some restrictions on the models which can be used. Before describing the idea and algorithm, full-conditional distributions (FCDs) must be introduced. An FCD is the distribution of a single parameter, θ_i (the i-th element of $\boldsymbol{\theta}$), conditional on every other entity (parameters and data). Such a distribution is denoted similarly to any

other conditional distribution by stating its dependents,

$$p(\theta_i|\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_m, \boldsymbol{x}).$$

Suppose we wish to infer a set of unknown parameters in a model, whose posterior, $p(\boldsymbol{\theta}|\boldsymbol{x})$, cannot be found analytically, nor can it be sampled from. However, the FCDs for each element of $\boldsymbol{\theta}$ can be found analytically and sampled from. Then, a Markov chain can be constructed by sequential sampling of the FCDs of each element of $\boldsymbol{\theta}$, and its stationary distribution is the joint-posterior, $p(\boldsymbol{\theta}|\boldsymbol{x})$ (Gamerman & Lopes, 2006). The Gibbs sampling algorithm is described in Algorithm 1.

Algorithm 1 Gibbs sampler

1. Initialise the state of the chain $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \ldots, \theta_m^{(0)})^{\mathrm{T}}$ and set $j = 1$.
2. Generate $\boldsymbol{\theta}^{(j)}$ by sequential realisations from full-conditionals

$$\begin{aligned}
\theta_1^{(j)} &\sim p(\theta_1|\theta_2^{(j-1)}, \ldots, \theta_m^{(j-1)}, \boldsymbol{x}) \\
\theta_2^{(j)} &\sim p(\theta_2|\theta_1^{(j)}, \theta_3^{(j-1)} \ldots, \theta_m^{(j-1)}, \boldsymbol{x}) \\
&\vdots \\
\theta_m^{(j)} &\sim p(\theta_m|\theta_1^{(j)}, \ldots, \theta_{m-1}^{(j)}, \boldsymbol{x})
\end{aligned}$$

3. Increment j by 1 and return to step 2.

The Gibbs sampler produces an m-dimensional Markov chain which will eventually, after a burn-in period, sample from the joint-posterior density. However, the time taken to reach the stationary distribution is not known, and the quality of the posterior draw produced is not guaranteed. How to assess both aspects of inference is discussed in Chapter 2.1.4.

2.1.2 Metropolis-Hastings

The Metropolis-Hastings algorithm was initially proposed by Metropolis et al. (1953) and later developed by Hastings (1970), and it can be used when sampling from full-conditionals is not available, i.e. the FCDs can not be sampled from directly. Unlike the Gibbs sampler, the method requires new $\boldsymbol{\theta}$ values to be actively proposed before being accepted or rejected. The acceptance probability combines information from the proposal distribution and posterior density. The proposal distribution, denoted $q(\cdot|\cdot)$, often depends on the previously accepted value of $\boldsymbol{\theta}$. Following this notation, the algorithm is described in Algorithm 2.

Note that the acceptance probability contains a ratio of posterior densities, the ratio cancels out the normalising constant, and thus is equal to a ratio of prior multiplied by likelihood. The form of the acceptance probability in Algorithm 2 ensures that the Markov chain's stationary distribution is the posterior density of interest, with proposed values in areas of higher posterior density having a higher acceptance probability (Gamerman & Lopes, 2006).

Algorithm 2 Metropolis-Hastings algorithm

1. Initialise the state of the chain $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \ldots, \theta_m^{(0)})^{\mathrm{T}}$ and set $j = 1$.
2. Propose a new parameter value

$$\boldsymbol{\theta}^* \sim q(\cdot|\boldsymbol{\theta}^{(j-1)}).$$

3. Calculate the acceptance probability

$$\begin{aligned}\alpha(\boldsymbol{\theta}^{(j-1)}, \boldsymbol{\theta}^*) &= \min\left\{1, \frac{p(\boldsymbol{\theta}^*|\boldsymbol{x})}{p(\boldsymbol{\theta}^{(j-1)}|\boldsymbol{x})}\frac{q(\boldsymbol{\theta}^{(j-1)}|\boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(j-1)})}\right\} \\ &= \min\left\{1, \frac{p(\boldsymbol{\theta}^*)L(\boldsymbol{\theta}^*|\boldsymbol{x})}{p(\boldsymbol{\theta}^{(j-1)})L(\boldsymbol{\theta}^{(j-1)}|\boldsymbol{x})}\frac{q(\boldsymbol{\theta}^{(j-1)}|\boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(j-1)})}\right\}.\end{aligned}$$

4. Set $\boldsymbol{\theta}^{(j)} = \boldsymbol{\theta}^*$ with probability $\alpha(\boldsymbol{\theta}^{(j-1)}, \boldsymbol{\theta}^*)$, otherwise set $\boldsymbol{\theta}^{(j)} = \boldsymbol{\theta}^{(j-1)}$.
3. Increment j by 1 and return to step 2.

The proposal distribution, $q(\cdot|\boldsymbol{\theta})$, is of particular importance, as it determines the algorithm's efficiency. Ideally, the proposal distribution closely resembles the stationary distribution (posterior distribution). If this is not the case, the chain may have inefficient exploration of the parameter space, which can lead to long burn-in periods and increased computational cost. Common proposal distributions are discussed in the following sections. Similarly to the Gibbs sampler, no guarantees are made about the quality of posterior draws or burn-in length.

Independent proposal

An independent proposal is one that does not rely on the current state of the Markov chain. The proposal density can then be written $q(\boldsymbol{\theta}^*)$, without the $\boldsymbol{\theta}^{(j-1)}$ dependence, and the acceptance probability becomes

$$\begin{aligned}\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*) &= \min\left\{1, \frac{p(\boldsymbol{\theta}^*|\boldsymbol{x})}{p(\boldsymbol{\theta}|\boldsymbol{x})}\frac{q(\boldsymbol{\theta})}{q(\boldsymbol{\theta}^*)}\right\}, \\ &= \min\left\{1, \frac{p(\boldsymbol{\theta}^*)L(\boldsymbol{\theta}^*|\boldsymbol{x})}{p(\boldsymbol{\theta})L(\boldsymbol{\theta}|\boldsymbol{x})}\frac{q(\boldsymbol{\theta})}{q(\boldsymbol{\theta}^*)}\right\}.\end{aligned}$$

Note that the superscript of the previously accepted value of $\boldsymbol{\theta}^{(j-1)}$ has been dropped for brevity. A special case of the independent proposal is when the prior distribution is used to propose values. In this case, the acceptance ratio simplifies to become the ratio of the likelihoods,

$$\begin{aligned}\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*) &= \min\left\{1, \frac{p(\boldsymbol{\theta}^*)L(\boldsymbol{\theta}^*|\boldsymbol{x})}{p(\boldsymbol{\theta})L(\boldsymbol{\theta}|\boldsymbol{x})}\frac{p(\boldsymbol{\theta})}{p(\boldsymbol{\theta}^*)}\right\}, \\ &= \min\left\{1, \frac{L(\boldsymbol{\theta}^*|\boldsymbol{x})}{L(\boldsymbol{\theta}|\boldsymbol{x})}\right\}.\end{aligned}$$

A prior proposal distribution is efficient if the prior closely resembles the posterior. Often, this is not the case, and the algorithm leads to a low acceptance probability and

slow exploration of the parameter space.

Symmetric proposal

A proposal is considered symmetric if $q(\boldsymbol{\theta}|\boldsymbol{\theta}^*) = q(\boldsymbol{\theta}^*|\boldsymbol{\theta})$ for all $\boldsymbol{\theta}, \boldsymbol{\theta}^* \in \Theta$. When this is true, the acceptance probability simplifies to a ratio of the stationary distributions, as such

$$\begin{aligned}\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*) &= \min\left\{1,\ \frac{p(\boldsymbol{\theta}^*|\boldsymbol{x})}{p(\boldsymbol{\theta}|\boldsymbol{x})}\right\} \\ &= \min\left\{1,\ \frac{p(\boldsymbol{\theta}^*)L(\boldsymbol{\theta}^*|\boldsymbol{x})}{p(\boldsymbol{\theta})L(\boldsymbol{\theta}|\boldsymbol{x})}\right\}.\end{aligned}$$

Symmetric proposals are common when constructing MCMC schemes, the most common of which is the random walk proposal.

Random walk proposal

A special case of the symmetric distribution is a random walk, which adds independent and identically distributed random noise to the previously accepted parameter values. That is,

$$\boldsymbol{\theta}^* = \boldsymbol{\theta}^{(j-1)} + \boldsymbol{\varepsilon}_j, \tag{2.1}$$

where $\boldsymbol{\varepsilon}_j$ are independent and identically distributed random variables. Usually, the noise is normally distributed and centred around $\mathbf{0}$ (m-dimensional zero vector). This is the multivariate extension of the normal random walk example in Section 1.5.4. Let

$$\boldsymbol{\varepsilon}_j \sim \mathrm{N}\left(\mathbf{0}, \Sigma\right),$$

for all j, then the a proposed value of $\boldsymbol{\theta}$ is drawn from

$$\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(j-1)} \sim \mathrm{N}\left(\boldsymbol{\theta}^{(j-1)}, \Sigma\right).$$

It remains to choose a covariance matrix, the choice of which is important for the efficiency of the inference scheme. Similarly to before, a quick exploration of parameter space is preferred. To achieve this, a covariance matrix is needed that shows similar correlations to the posterior distribution and marginal variances which are not too big nor too small. Small marginal variances lead to slow exploration of the parameter space with many proposed values being accepted. Large marginal variances will lead to too few proposed values being accepted and slow exploration.

In practice, Roberts and Rosenthal (2001) proposed that the covariance matrix should depend on the posterior covariance and the number of parameters being inferred, suggesting that

$$\mathrm{Var}(\boldsymbol{\varepsilon}_j) = \frac{2.38^2}{m}\widehat{\mathrm{Var}}(\boldsymbol{\theta}|\boldsymbol{x}). \tag{2.2}$$

Here, $\widehat{\mathrm{Var}}(\boldsymbol{\theta}|\boldsymbol{x})$ denotes a sample variance taken from an initial run of the inference. Initial runs of inference would add additional computational cost to the inference scheme.

However, for complex models with complex posterior distributions, the benefits of a more efficient sampling algorithm would generally outweigh the additional cost. The normal random walk proposal mechanism is implemented in Chapter 5.

2.1.3 Metropolis-within-Gibbs

Hybrid MCMC schemes are capable of implementing Metropolis-Hastings updates to a subset of parameters whose FCDs are not tractable (solvable) while allowing component-wise Gibbs updates for those that are.

The Metropolis-within-Gibbs algorithm is a combination of Algorithms 1 and 2. Prior to inference, a proposal distribution for the Metropolis-Hastings update must be chosen. The chain is initialised before new values are drawn for the parameters with tractable FCDs. The Metropolis-Hastings update follows Algorithm 2; new values are proposed, an acceptance probability is calculated, and the chain is updated accordingly. A proposal distribution is only required for parameters whose FCDs can not be sampled from. Note that the updates can occur in any order, and the Metropolis-Hastings update can be separated into several distinct updates, each with its own proposal distribution and acceptance probability. Such schemes are beneficial as high dimensional parameter spaces are often more difficult to explore, with increased difficulty in finding appropriate proposal distributions. See Gamerman and Lopes (2006) for a detailed description.

2.1.4 Convergence and autocorrelation

Convergence

Once a suitable inference scheme is constructed, a key question remains to be answered; has the chain reached the target distribution? Despite the guarantees that in the limit, as the number of iterations approaches infinity, the chain will have reached its stationary distribution, providing it exists, it is not practical, or even possible, to run an MCMC scheme for an infinite number of iterations (Roberts & Rosenthal, 2004). Therefore, practical applications of how to answer the question must be considered. Not providing a suitable answer may have dramatic consequences on the results of analysis, skewing posterior distributions, predictions and any conclusions drawn from them.

An initial convergence check is usually done by eye. A converged chain should sample from a steady distribution. Conversely, a chain that does not maintain some equilibrium shows signs of non-convergence and may still be in its burn-in period. Whether or not the chain is moving centred upon some steady state should be visible in a trace plot, a time series line plot of the chain's value at each iteration. Also informal, although more convincing, is the inspection of multiple chains, initialised at a range of values. All chains should reach the same stationary distribution, and overlayed trace plots are an easy method to highlight when this is or is not the case. When the chains do not overlap, this could indicate a lack of convergence. Multiple chains have the additional benefit that the posterior draws can be combined, provided checks have been made that they come from the same distribution, essentially parallelising inference.

Several formal convergence diagnostics have been proposed. Geweke (1991) suggested that the averages of sequential realisations from the Markov chain be compared. One set of realisations from the beginning of the chain (after burn-in) and one set of the most recent realisations. Geweke showed that as the number of realisations gets larger, the normalised difference between the two averages tends to a standard normal distribution. Therefore, large standardised differences show signs of non-convergence. A more popular method is a statistic which compares the within and between chain variation of multiple chains. The statistic, $\hat{\mathrm{R}}$, tends to 1.0 as the number of iterations increases and was proposed by Gelman and Rubin (1992). Many other convergence diagnostics have been suggested, however, they will not be discussed in detail here, see Mengersen et al. (1999) for a review.

Autocorrelation

Autocorrelation refers to the correlation between realisations of the Markov chain at varying lags, the number of iterations between two realisations, and consequently draws from the posterior of interest. Ideally, all posterior draws would be independent but this is generally not possible. The autocorrelation can be reduced by only saving every k-th realisation from the chain, and deleting the rest, referred to as thinning. However, this ignores the information that could be gained from the deleted realisation. A better method is to increase the number of iterations to gain a better understanding of the posterior distribution and reduce the uncertainty in posterior statistics, e.g. mean and variance. Thinning the output can be reserved for when an extremely large number of iterations are needed and the associated storage costs become too high.

The amount of autocorrelation can observed informally by plotting the autocorrelation at a range of lags. More formally, the effective sample size (ESS) can be calculated (Plummer et al., 2006). This is the number of independent realisations which have the same estimation power as the correlated realisations from the Markov chain. When autocorrelation is low, ESS and the number of realisations will be very close in value. When autocorrelation is high the ESS can be drastically smaller. The ESS gives a good indication of the accuracy of moments calculated from the posterior draws. Single and multi-variate ESS calculations exist and both can used together.

2.2 Statistical software

Constructing an appropriate inference can be complicated, and finding appropriate proposal distributions may be difficult for high-dimensional problems. To overcome this hurdle, software exists that can implement a Bayesian inference scheme (semi-)automatically. The two main pieces of software are STAN (Carpenter et al., 2017) and JAGS (Plummer, 2003). Both are available within R and Python, as well as in many other places, and they both have advantages and disadvantages. The largest disadvantage for both is the initial time spent needing to learn how to write a model in their respective probabilistic programming languages.

STAN, available in `R` through the '`rstan`' package (Stan Development Team, 2020), implements another Bayesian inference algorithm called Hamiltonian Monte Carlo (HMC), which generally offers better exploration of the parameter space (Hoffman & Gelman,

2014). STAN quickly finds the main areas of the target support, and once there, the samples generated are almost uncorrelated. However, due to the specific algorithm STAN uses, categorical variables cannot directly be a part of the model. For many models, including the one used in Chapter 4, this is not an issue and a 'workaround' can be implemented.

JAGS (Plummer, 2003) is available in `R` through the '`rjags`' package (Plummer et al., 2024). JAGS is not as efficient as STAN in its exploration of the parameter space and its posteriors often produce more correlated samples. However, JAGS is easier to use and allows categorical variables.

Both STAN and JAGS are likely faster than hard-coded inference schemes, both using optimised C code. However, it is not always possible to use STAN or JAGS, as will seen in Chapter 5, necessitating hard-coding an inference scheme.

2.3 Bayesian hypothesis testing

The t-test is one of the most common tests used in statistics. It compares the means of two normally distributed datasets by calculating a test statistic and comparing it to the t-distribution. Variations of the test exist, including constant and differing variance between groups, paired and unpaired, and non-normality in the data. Here, Bayesian hypothesis testing is briefly discussed, giving an indication of how a two-sample t-test would be conducted within a Bayesian context. Extensions to other situations are available. However, the following example illustrates the main concept.

Let $\{\boldsymbol{X}_1, \boldsymbol{X}_2\}$ be two independent and normally distributed datasets, with means μ_1 and μ_2. Assume the variance of the two groups is known and equal, although this need not be the case. In a Bayesian context, the posterior distributions $p(\mu_1|\boldsymbol{X}_1)$ and $p(\mu_2|\boldsymbol{X}_2)$ are inferred, in the usual manner by constructing a prior distribution and combining with the data likelihood. Let the difference of the group means define a random variable, $\Delta\mu = \mu_1 - \mu_2$. The posterior distribution $p(\Delta\mu|\boldsymbol{X})$ where $\boldsymbol{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2\}$, is of interest here and can be found either analytically or by MCMC, by either method, the difference in the group means can be inspected by analysing $p(\Delta\mu|\boldsymbol{X})$. For example, to answer the question; Is the mean of group 1 larger than group 2? A Bayesian statistician would calculate the posterior probability $\Pr(\mu_1 > \mu_2|\boldsymbol{X})$, which is equivalent to $\Pr(\Delta\mu > 0|\boldsymbol{X})$. If $p(\Delta\mu|\boldsymbol{X})$ is found via MCMC, the probability is approximated

$$\Pr(\Delta\mu > 0|\boldsymbol{X}) \approx \frac{1}{N_{\text{MCMC}}} \sum_i \mathbb{I}(\mu_1^{(i)} > \mu_2^{(i)}) \tag{2.3}$$

where N_{MCMC} is the number of posterior realisations and $\mu_1^{(i)}$ and $\mu_2^{(i)}$ are the i-th realisations from the posterion. the indicator function, $\mathbb{I}(\cdot)$, equals 1 if the expression is true and 0.0 otherwise. Although a Bayesian statistician would not use the word *significant*, one may say that if the probability is low, less than 0.05, there is *substantial* evidence to suggest that the mean of group 1 is larger than the mean of group 2.

Other questions regarding the two group means can be proposed, and their answer formulated in a similar manner by inspection of the appropriate posterior distribution.

For a review and discussion of how Bayesian statistical tests are implemented and a comparison to frequentist methods, see Kruschke and Liddell (2018).

An important aspect of hypothesis testing in frequentist statistics is the confidence interval. In the Bayesian context, an analogue of the confidence interval is often used, the high density interval (HDI). Their uses are very similar to the confidence interval, but their theory is based on the parameter beliefs rather than asymptotic assumptions via the central limit theorem.

2.3.1 High density intervals

For a univariate random variable, a $100(1-\alpha)\%$ HDI defines a region of the support that $100(1-\alpha)\%$ of the distribution and the probability density at all points in that region is higher than for any point outside. The latter condition ensures that the region is unique, assuming that the density is not flat at any point within the interval.

For example, consider the distribution in Figure 2.1. The symmetric distribution in Figure 2.1(a) has the same HDI as an equi-tailed confidence interval. When the distribution is skewed or bimodal, such as Figures 2.1(b) and 2.1(c), this is not the case. The equi-tailed confidence interval for a skewed distribution is shifted compared to the HDI. The HDI of the bimodal distribution in Figure 2.1(c) contains the two modes but does not contain the mean, zero. For bimodal and strongly skewed distributions, the use of confidence interval (CI) does not make sense, and HDIs are preferred, but for symmetric distributions, there is little difference.

Throughout this thesis, HDIs are used to assess whether a random variable is substantially different from a particular value, usually zero. Similarly to a CI, if zero lies outside of a random variable HDI, it can be concluded that it is substantially different from zero.

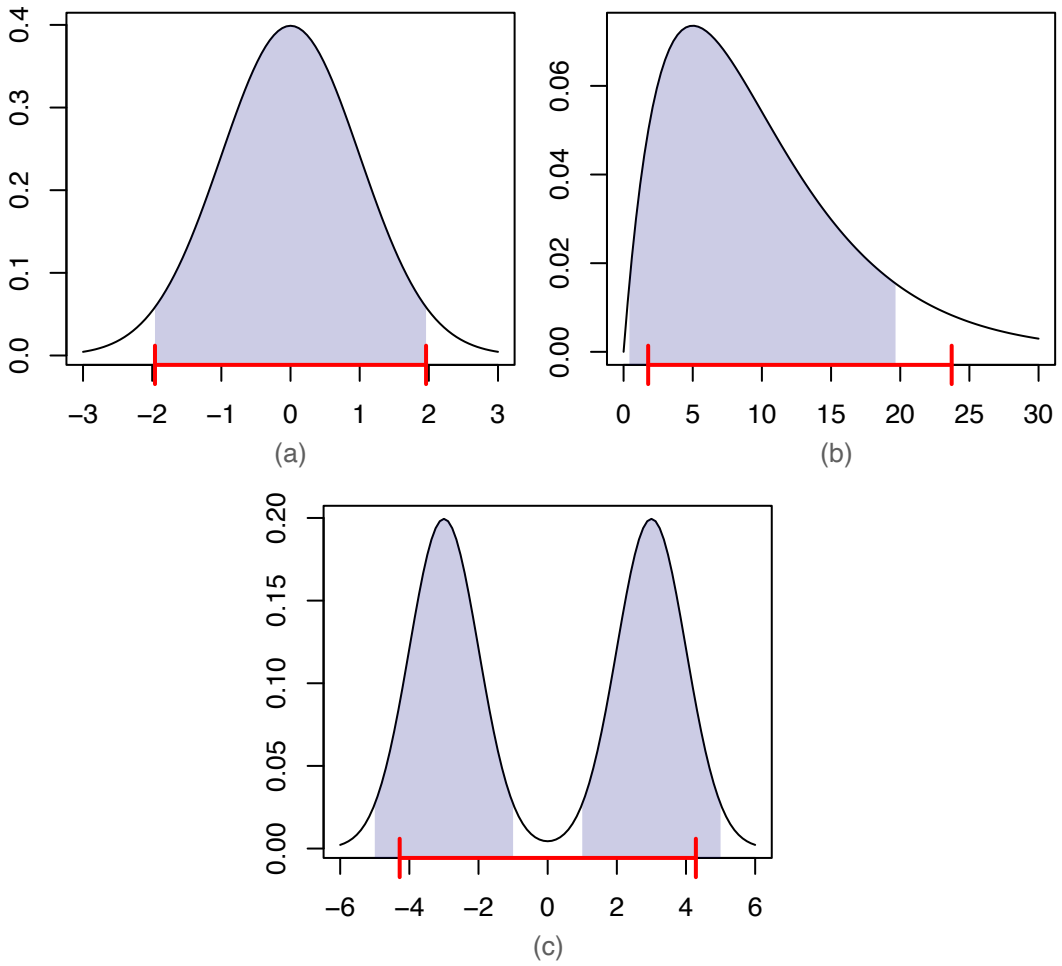


Figure 2.1: **High density intervals compared to equi-tailed confidence intervals**. High-density intervals for (a) unimodal, (b) skewed, and (c) bimodal distributions are illustrated by the shaded blue area below the density curve. Equivalent equi-tailed confidence intervals are shown by red bars along the x-axis.

2.4 Mixture models

In this chapter, a special type of statistical model is introduced and discussed: the mixture model. Mixture models are a diverse set of statistical models which can take an endless number of forms. Here, the discussion is focused on finite mixture models, as these are used within this thesis. The most common form of the mixture model is likely to be the Gaussian mixture model (GMM) (Reynolds, 2009), which is used for a variety of applications, including clustering, an unsupervised classification method, (Shervegar & Bhat, 2018) and traditional modelling (Costa et al., 2012). Mixture models are used in Chapters 3 and 4. Before discussing mixture models, mixture distributions are introduced. Although not explicitly discussed here, the chapter continues the focus on Bayesian methods. Discussions of latent variables, classification and label switching are all specific to the Bayesian paradigm.

2.4.1 Mixture distributions

A mixture distribution is defined as the weighted sum of multiple probability density functions. Let $f_1(x), f_2(x), \ldots, f_k(x)$ be probability density functions, and $w_1, w_2, \ldots, w_k$ a set of weights, such that $w_i \geq 0$ for all i and $\sum_i w_i = 1$. Then the mixture distribution, $f(x)$, is defined as

$$f(x) = \sum_{i=1}^{k} w_i f_i(x). \tag{2.4}$$

Note that if the component densities are well defined, then the mixture density integrates to 1, and all values are non-negative, as required for a PDF. The component densities, $f_1(x), f_2(x), \ldots, f_k(x)$, should all be appropriately discrete or continuous and have the same dimension for the random variable. However, it is not required that they have the same density with different parameters, and mixtures can combine distributions of various types.

A mixture distribution can be useful when dealing with multimodal or skewed data that is not representative of a standard distribution. In Bayesian statistics, mixture distributions provide greater flexibility in the choice of prior beliefs.

Normal mixture distribution example

To illustrate the flexibility of mixture distributions, consider a three-component normal mixture distribution. Let X be a random variable described by the mixture distribution, and π_i, μ_i and σ_i be the component weight, mean and standard deviation of the i-th component. The distribution can be written as

$$X \sim \pi_1 \mathrm{N}\left(\mu_1, {\sigma_1}^2\right) + \pi_2 \mathrm{N}\left(\mu_2, {\sigma_2}^2\right) + \pi_3 \mathrm{N}\left(\mu_3, {\sigma_3}^2\right). \tag{2.5}$$

Figure 2.2 shows the PDF with a range of parameter values; it is clear from this that a wide variety of densities are possible from just the combination of normal distributions.

2.4.2 Mixture modelling

A mixture model combines appropriately weighted component models to form a single model, representing the natural extension of mixture distributions. A mixture model is formed by the appropriately weighted sum of multiple models, much like a mixture distribution. The weights associated with each component can be interpreted in two ways: the proportion of the population belonging to the component or the probability that a new observation belongs to that component.

Mixtures can be beneficial when data shows sub-populations, groups within the dataset, exhibit inter-group variation. A major benefit of this model type is that group labels are not needed. In fact, a physical interpretation of the groups is not required, and identification of 'like' groups can lead to further investigation. Consequently, mixture models can be used to classify datasets into subpopulations, often referred to as clustering. Unlike other clustering algorithms, the number of sub-populations must be chosen before modelling begins. However, this does not stop the fitting of multiple models with varying

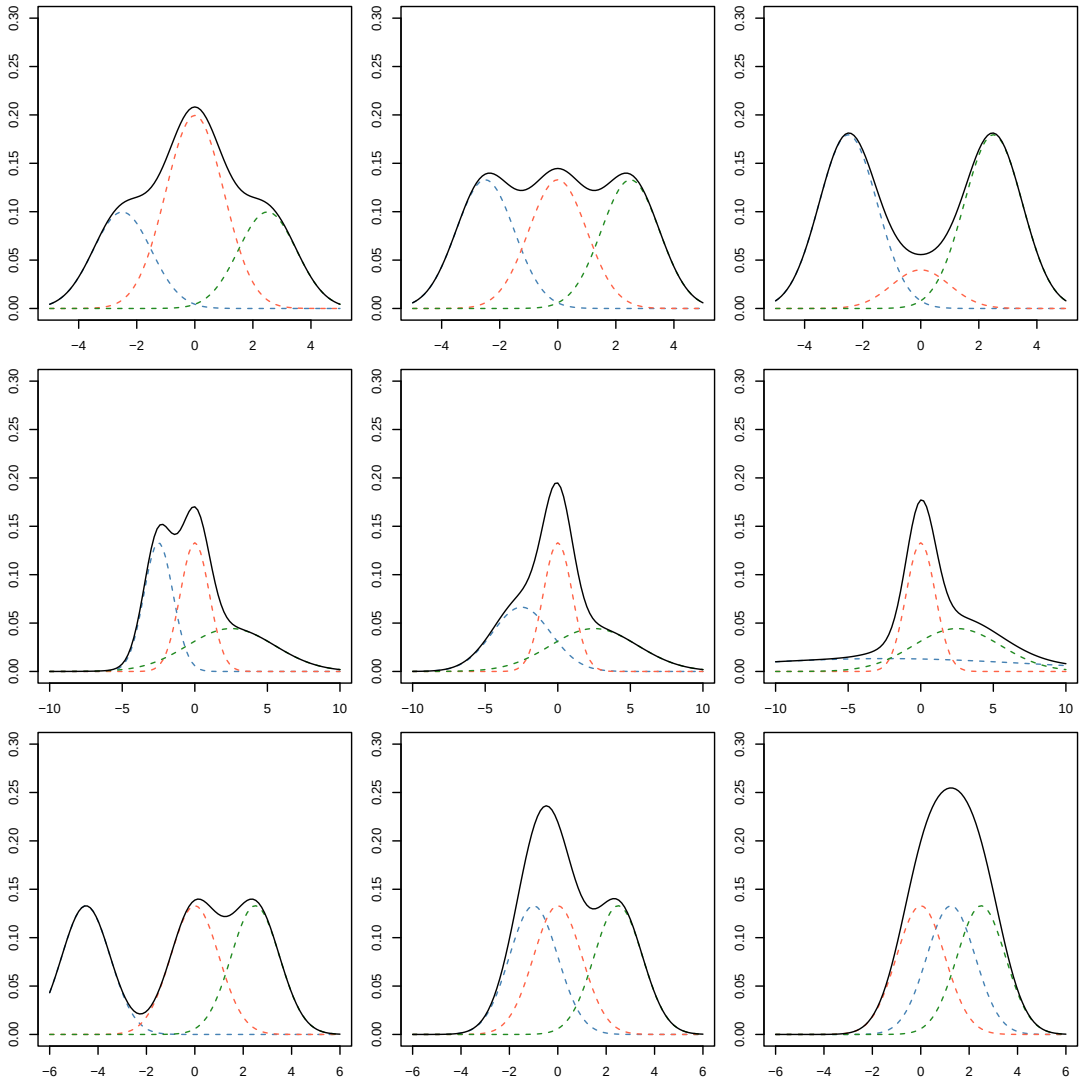


Figure 2.2: **Variety of three-component normal mixture distribution**. Each row alters only a single parameter. The top row of distributions alters the proportion of the red component distribution, decreasing from left to right and splitting the proportion evenly between the other components. The middle row alters the variance of the blue component, increasing from left to right while keeping all other parameters the same. The bottom row only changes the mean of the blue component and all other parameters are kept the same.

numbers of components and the comparison of model fit.

Let $\{\boldsymbol{x}, \boldsymbol{y}\}$ denote a dataset of n independent observations on explanatory variable x and response variable y, such that $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^{\mathrm{T}}$ and $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^{\mathrm{T}}$, and suppose that it is believed that the data split naturally into two groups. The first group is modelled by $f_1(y|\theta_1, x)$ and the second by $f_2(y|\theta_2, x)$, with respective weights and model parameters w_1, w_2, θ_1 and θ_2. The mixture model, $f(y|\theta_1, \theta_2, x)$, is defined

$$f(y|\theta_1, \theta_2, x) = w_1 f_1(y|\theta_1, x) + w_2 f_2(y|\theta_2, x). \tag{2.6}$$

More generally, a J-component mixture model is written as

$$f(y|\boldsymbol{\theta}, x) = \sum_{j=1}^{J} w_j f_j(y|\theta_j, x), \tag{2.7}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_J)$, the set of component parameters. The general likelihood function, for observations $\boldsymbol{y}$, is

$$L(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y}) = \prod_{i=1}^{n} \left\{ \sum_{j=1}^{J} w_j f_j \left(y_i|\theta_j, x_i\right) \right\}. \tag{2.8}$$

By expanding the likelihood, it can be seen that this becomes increasingly complex when components or data are added. In particular, the likelihood is an n-degree polynomial in the component weights.

2.4.3 Latent states

Classification by mixture model is the task of learning the unobserved, latent states, which identify each observation as belonging to a model component or sub-population. The latent states are, as such, categorical variables, taking values in the range $\{1, 2, \ldots, J\}$, where J is the number of components.

Suppose the group allocation is known, i.e. the data is labelled. The model and likelihood function becomes

$$L\left(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{w}, \boldsymbol{Z}\right) = \prod_{i=1}^{n} w_{Z_i} f_{Z_i}(y_i|\theta_{Z_i}, x_i). \tag{2.9}$$

Simplifying the model density removes the n-degree polynomial in the likelihood function, drastically decreasing its computational complexity. Evaluation of the likelihood function is critical to many parameter inference schemes in both frequentist and Bayesian settings. As such, including the latent variables could reduce the computational costs associated with inference and facilitate data classification in the process.

2.4.4 Classification and latent state inference

Conditioning the likelihood function on latent states can reduce its complexity. However, latent states are not necessarily known; when this is the case, the latent states must be inferred. In Bayesian methodology, all unknown quantities are treated in the same manner, and so we aim to find the posterior distribution of the latent states.

Let Z_i be the latent state of the i-th observation, which is distributed according to the component weights *a priori*,

$$\Pr\left(Z_i = j|\boldsymbol{w}\right) = w_j. \tag{2.10}$$

Inference for the latent states can progress by considering their FCDs. The law of conditional probability implies that the FCD is proportional to the joint density, $p\left(\boldsymbol{y}, \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{w}\right)$. The probability can be calculated up to proportionality, and so any func-

tion not dependent on Z_i can be removed

$$\begin{aligned}
\Pr\left(Z_i = j|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{w}\right) &\propto p\left(\boldsymbol{y}, \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{w}\right), \\
&= p\left(\boldsymbol{y}, \boldsymbol{Z}|\boldsymbol{\theta}, \boldsymbol{w}\right) p(\boldsymbol{\theta}, \boldsymbol{w}), \\
&\propto p\left(\boldsymbol{y}|\boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{w}\right) p\left(\boldsymbol{Z}|\boldsymbol{\theta}, \boldsymbol{w}\right), \\
&= \prod_{l=1}^{n} p\left(y_l, |Z_l, \boldsymbol{\theta}\right) p\left(Z_l|\boldsymbol{w}\right), \\
&= p\left(y_i, |Z_i, \boldsymbol{\theta}\right) p\left(Z_i|\boldsymbol{w}\right), \\
&\propto w_j f_j\left(y_i|\theta_j\right).
\end{aligned} \tag{2.11}$$

The joint density, $p(\boldsymbol{y}, \boldsymbol{Z}|\boldsymbol{\theta}, \boldsymbol{w})$, is split into the observed data likelihood, $p(\boldsymbol{y}|\boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{w})$ and the latent state density $p(\boldsymbol{Z}|\boldsymbol{\theta}, \boldsymbol{w})$, and simplified. The probability that the i-th observation is classified as belonging to the j-th component is found to be proportional to the weighted density of the j-th component, evaluated at y_i. After calculating the normalising constant, the posterior probability can be written exactly,

$$\Pr\left(Z_i = j|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{w}\right) = \frac{w_j f_j\left(y_i|\theta_j\right)}{\sum_{k=1}^{J} w_k f_k\left(y_i|\theta_k\right)}. \tag{2.12}$$

This classification naturally arises from the Bayesian mixture model; as such, it is called the Bayes classifier. Inference for other model parameters can proceed using an appropriate inference scheme, such as Gibbs or Metropolis within Gibbs, conditioning on latent states. At each iteration of the inference scheme, the latent states for each observation are then randomly assigned using the above PMF.

2.4.5 Label switching

Label switching is an issue seen during the inference of mixture model parameters. Consider the two-component mixture model,

$$f(y|\theta_1, \theta_2, \pi, X) = \pi f_1(y|\theta_1, x) + (1 - \pi) f_2(y|\theta_2, x),$$

where f_1 and f_2 are the same model family, i.e. both normal distributions. The problem arises due to the exchangeability of parameters, that is, the model, $f(y|\theta_1, \theta_2, \pi, x)$ is indistinguishable from $f(y|\theta_2, \theta_1, 1 - \pi, x)$. The result of this exchangeability is that the Markov chains targeting the posterior $p(\theta_1|\boldsymbol{y})$ and $p(\theta_2|\boldsymbol{y})$ can jump between the two states (Jasra et al., 2005). This can lead to errors when inspecting posterior moments of the distributions and errors when inspecting predictions drawn from the model. If label switching has occurred, it is often evident from inspecting trace plots and convergence diagnostics. Autocorrelation would appear to be extremely high, and the chains will have a large within-chain variation leading to a high $\hat{\mathrm{R}}$.

Theoretically, it is possible to prevent label switching through a careful selection of prior beliefs; however, this is not guaranteed to work, and such beliefs may not accurately represent the true beliefs. A more practical approach is to impose a constraint on the parameters θ_1 and θ_2. Enforcing $\theta_1 < \theta_2$ within the inference scheme can prevent the two chains from jumping between the posterior distributions. The ordering constraint does

impose some restrictions on posterior analysis. Suppose we wish to know the probability that $\theta_1 > \theta_2$. In the Bayesian paradigm, the probability is approximated by the proportion of posterior draws such that $\theta_1 > \theta_2$, as in Chapter 2.3. This method will fail as the constraint is enforced. Additionally, imposing this constraint becomes more difficult when the number of unknown parameters is high.

Another approach is to initialise the Markov chain at realistic initial values, close to the posterior modes. This does not guarantee that the chains will not transition between modes, and selecting appropriate initial values may prove challenging.

2.5 Hierarchical modelling

In this chapter, another type of model is introduced, one that is synonymous with Bayesian methods, the hierarchical model (Gelman & Hill, 2006). Hierarchical models are popular as they allow simple models to be combined to form a complex structure. Often, hierarchical models are constructed as a form of mixture model, and therefore, the previous discussion about latent variables and classification is also relevant here.

The ability of hierarchical models to reflect complex systems means they have been applied to a variety of modelling situations. Heydari et al. (2016) proposed a hierarchical model, which can reflect the experimental design, to model yeast population growth rates and their genetic interactions. Their model was able to detect more subtle interactions between genotypes than an existing non-hierarchical method. Berry et al. (2013) demonstrated that a hierarchical model can reduce the number of patients required in clinical trials by an average of 4-7 patients compared to existing methods. Within this thesis, hierarchical models are used within Chapters 3 and 4.

2.5.1 Bayesian hierarchical model

A Bayesian hierarchical model imposes dependencies between unknown parameters. In practice, prior parameters are considered unknown and to be inferred, reflecting uncertainty in the prior specification itself. To this end, hyper-priors, defined by hyper-parameters, are chosen to summarise beliefs on the prior parameters. The dependent structure enables the construction of complex models from simple components.

Hierarchical models are often used when the dataset can be split into non-overlapping groups, which may impact the response variable. The hierarchy allows each group to be modelled with a different parameter while learning the population-level parameters. The top level of the hierarchy describes beliefs about the population-level parameters, which concern the whole dataset. After that, each level will describe a smaller subset of the data.

One benefit of using a Bayesian hierarchical model is that information is shared between groups through the learning of top-level parameters, referred to as borrowing strength between groups. Whole-population parameters beliefs are passed on to sub-population parameters, minimising over-fitting risks for groups with few data points.

Mixture models and hierarchical models are often combined, as they both concern data sub-populations. When this is the case, inference can proceed as usual, but care must be taken when calculating FCDs due to the model's more complex structure. An example of such a model is shown in Chapter 4.

2.5.2 Bayesian hierarchical model example

Suppose we wish to model the proportion of students achieving a pass in at least five GCSEs within schools across the UK. It is reasonable to assume that the pass rate will vary between schools. As such, aggregating the UK-wide data would not be informative of the whole picture. Alternatively, each school could be modelled independently, but this would ignore the information within the dataset from other schools. The hierarchical model addresses both issues, allowing each school to have a unique pass rate while borrowing strength across schools.

Suppose the dataset contains information for N uniquely identified schools across the UK. Let n_i be the number of students in the i-th school, and Y_i be the number of students achieving the five GCSEs in that school. Table 2.1 shows a snippet of what such a dataset may look like.

School	n	Y
1	120	95
2	99	65
3	108	71
⋮	⋮	⋮
N	79	70

Table 2.1: **Example dataset of UK school students achieving five GCSEs**.

Assume Y_i follows a binomial distribution, with probability of success θ_i, and the number of children in each school is known, then

$$Y_i|n_i, \theta_i \sim \text{Bin}\left(n_i, \theta_i\right). \tag{2.13}$$

The parameters $\theta_1, \theta_2, \ldots, \theta_N$ are unknown and to be inferred. The pass rates are proportions, so it is reasonable to model them using a Beta distribution,

$$\theta_i|\alpha, \beta \sim \text{Beta}\left(\alpha, \beta\right) \tag{2.14}$$

The hierarchy is implemented by considering the prior parameters α and β unknown and, therefore, also to be inferred. Both are continuous and positive; independent gamma distributions are one choice for the hyper-priors.

$$\begin{aligned} \alpha &\sim \text{Ga}\left(a_1, b_1\right) \\ \beta &\sim \text{Ga}\left(a_2, b_2\right) \end{aligned} \tag{2.15}$$

The hyper-parameters, $\{a_1, b_1, a_2, b_2\}$, are known and chosen by us to summarise prior beliefs. By inferring α and β, the distribution of a nationwide success rate can be learnt from the whole dataset. Which, in turn, affects the distribution of school-level success

rates. The relationship can be seen by calculating the parameter FCDs up to proportionality.

First, consider the FCD of the prior parameters α

$$\begin{aligned} p(\alpha|\beta,\boldsymbol{\theta},\boldsymbol{y}) &\propto p\left(\alpha,\beta,\boldsymbol{\theta},\boldsymbol{y}\right), \\ &\propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\alpha,\beta)p(\alpha)p(\beta), \\ &\propto p(\boldsymbol{\theta}|\alpha,\beta)p(\alpha), \\ &\propto \prod_{i=1}^{n}\left\{{\theta_i}^{\alpha-1}\left(1-\theta_i\right)^{\beta-1}\right\}\times\alpha^{a_1-1}e^{-b_1\alpha}, \\ &\propto \alpha^{a_1-1}e^{-b_1\alpha}\prod_{i=1}^{n}{\theta_i}^{\alpha-1}. \end{aligned} \tag{2.16}$$

Unfortunately, this is not the form of a known probability distribution; as such, the analysis is not semi-conjugate. Despite the analysis not being semi-conjugate, inference can still go ahead with the implementation of Metropolis-Hastings or hybrid algorithms, as previously discussed. Nevertheless, it can be seen that given a_1, b_1 and $\boldsymbol{\theta}$, $p(\alpha|\beta,\boldsymbol{\theta},\boldsymbol{y})$ is independent of β and $\boldsymbol{y}$. Similarly, calculating the FCD of β shows that given a_2, b_2 and $\boldsymbol{\theta}$, it is independent of the observations $\boldsymbol{y}$ and α.

$$p(\beta|\alpha,\boldsymbol{\theta},\boldsymbol{y}) \propto \alpha^{a_\beta-1}e^{-b_\beta\beta}\prod_{i=1}^{n}\left(1-\theta_i\right)^{\beta-1} \tag{2.17}$$

The final parameters to consider are the individual school success rates

$$\begin{aligned} p(\theta_i|\alpha,\beta,\boldsymbol{\theta}_{-i},\boldsymbol{y}) &\propto p\left(\alpha,\beta,\boldsymbol{\theta},\boldsymbol{y}\right), \\ &\propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\alpha,\beta), \\ &\propto \prod_{i=1}^{n}\left\{{\theta_i}^{y_i}\left(1-\theta_i\right)^{n_i-y_i}\right\}\times\prod_{i=1}^{n}\left\{{\theta_i}^{\alpha-1}\left(1-\theta_i\right)^{\beta-1}\right)\right\}, \\ &\propto {\theta_i}^{y_i+\alpha-1}\left(1-\theta_i\right)^{n_i-y_i+\beta-1}. \end{aligned} \tag{2.18}$$

This is the form of a Beta distribution with parameters $y_i+\alpha$ and $n_i-y_i+\beta$, and, therefore,

$$\theta_i|\alpha,\beta,y_i \sim \text{Beta}\left(y_i+\alpha, n_i-y_i+\beta\right). \tag{2.19}$$

The FCD of the individual school success rates depends on the observed data of only that school, y_i, and the population-level parameters, α and β.

By learning high-level parameters, α and β, information about the global population is passed to lower-level parameters, θ_i, which are specific to a sub-population. Therefore, information is transferred to sub-population-specific parameters from the whole dataset, borrowing strength across groups. When there is little data in sub-populations, sharing information across groups is particularly useful; alternatively, treating each sub-population independently may lead to over-fitting in either a Bayesian or frequentist setting.

Hierarchical models can become more complex when additional layers to the hierarchy are added. For example, suppose the schools are grouped into geographical regions. Another layer could be added which learns region-specific success rates.

2.6 Stochastic kinetic models

Stochastic kinetic models were developed to model chemical reactions within a system (Lecca, 2013). However, they have since become popular methods to model complex biological systems (Wilkinson, 2018). Stochastic kinetic models often simplify much larger and more complex systems, as computationally it would be impossible to account for and describe all the possible events and molecules within a system. The models, therefore, focus on the key reactions considered to be driving system dynamics. The stochastic nature of the models allows the seemingly random fluctuations caused by unknown and unaccounted-for elements within the system to be simulated in a way that a deterministic model would not be able to do. Random fluctuations are often most significant when populations are small and a single event can make a substantial difference in the relative populations of the system (Elowitz et al., 2002). This is why their use in biological datasets is increasing in popularity.

Stochastic kinetic models have been used to model mtDNA population dynamics (Henderson et al., 2009; Hoitzing et al., 2017; Johnston et al., 2015), as well as a wide variety of other complex biological systems, including: how the spatial distribution of cancer cells affects bulk and single-cell sequencing data (Chkhaidze et al., 2019), the coordination of plant growth through interacting hormones and genes (Jackson et al., 2020), and the effects of cortisol levels on neuronal activity (McAuley et al., 2009). In this thesis, the population dynamics of mtDNA is considered and referred to throughout. Stochastic kinetic models of these dynamics are considered in Chapter 5. The remainder of this chapter introduces the theory behind stochastic kinetic models, as well as their simulation methods. The chapter concludes with two examples of stochastic kinetic models and realisations from their simulation.

2.6.1 Chemical reactions

Before describing stochastic kinetics models, associated theory, and simulation methods, let us consider a system of biochemical reactions. Suppose the system consists of n species, denoted $X_1, X_2, \ldots, X_n$, and set of m possible chemical reactions, denoted $R_1, R_2, \ldots, R_m$. The set of reactions describes all possible interactions that can occur between the species within the system. A reaction changes the number (or concentration) of one or more species in the system. For example, the species X_1 may split into two molecules of type X_2 and X_3. This would decrease the amount of X_1 by one unit and increase the amount of X_2 and X_3 by one unit. The reaction would be denoted by a pseudo-equation,

$$X_1 \xrightarrow{k} X_2 + X_3. \tag{2.20}$$

The species present before the reaction are called the reactants, and the species present after are called the products. Under a set of assumptions, discussed in Chapter 2.6.2, the reaction will occur at a given rate, k, often noted above the reaction's direction arrow.

A general reaction network for a system can be written as follows,

$$\begin{aligned}
R_1: &\quad a_{11}X_1 + a_{12}X_2 + \cdots + a_{1n}X_n \xrightarrow{k_1} b_{11}X_1 + b_{12}X_2 + \cdots + b_{1n}X_n, \\
R_2: &\quad a_{21}X_1 + a_{22}X_2 + \cdots + a_{2n}X_n \xrightarrow{k_2} b_{21}X_1 + b_{22}X_2 + \cdots + b_{2n}X_n, \\
&\qquad\qquad \vdots \qquad\qquad\qquad\qquad \vdots \\
R_m: &\quad a_{m1}X_1 + a_{m2}X_2 + \cdots + a_{mn}X_n \xrightarrow{k_m} b_{m1}X_1 + b_{m2}X_2 + \cdots + b_{mn}X_n,
\end{aligned} \tag{2.21}$$

where $\boldsymbol{k} = (k_1, k_2, \ldots, k_m)^{\mathrm{T}}$ are the reaction rates of the system.

The coefficients of the reactants and products are non-negative real numbers denoting the quantity of each species required for that reaction. In Eq. 2.20, the only non-zero reactant coefficient was for X_1, and it was equal to 1.0; the only non-zero product coefficients were for X_2 and X_3, both equal to 1.0. The coefficients are referred to as stoichiometric coefficients. It is useful to summarise the coefficients in matrix form. Let A denote the pre-reaction stoichiometric matrix, whose (i, j)-th element is equal to a_{ij}, and B denote the post-reaction stoichiometric matrix, whose (i, j)-th element is $b_{i,j}$.

The net change in X_j species level after the i-th reaction is therefore $b_{ij} - a_{ij}$. The change in species due to each reaction is stored in the net effects matrix, $Q = B - A$, whose (i, j)-th element is equal to $b_{ij} - a_{ij}$. The stoichiometry matrix, S, is the transpose of the net effect matrix,

$$S = Q^{\mathrm{T}}, \tag{2.22}$$

which is often used in stochastic kinetic models.

2.6.2 Markov jump process

Assuming the molecules are within a well-mixed container of fixed volume at thermal equilibrium, it can be shown that their movement is governed by Brownian motion. Brownian motion describes the movement of a molecule over a time step, δt, as normal random noise, whose variance is proportional to δt. The rate of reactions is constant over a small time step (Gillespie, 1992), and so the reactions must have associated stochastic rate constants, previously denoted $\boldsymbol{k}$. The assumption of mass action kinetics states that the reaction rate is proportional to the product of the reactant levels required for the reaction. Under these assumptions, the hazard associated with any reaction can be defined as the stochastic rate constant multiplied by the product of the reactant levels required.

Consider the first-order chemical reaction,

$$X_1 \xrightarrow{k} X_2. \tag{2.23}$$

Only one molecule of X_1 is required for the reaction to take place. The reaction hazard is

$$h(\boldsymbol{x}, k) = kx_1, \tag{2.24}$$

where x_1 is the level of species X_1. The second-order reaction

$$X_1 + X_2 \xrightarrow{k} X_3, \tag{2.25}$$

requires two molecules to interact, one molecule of X_1, the one of X_2. If the current levels of the species are x_1 and x_2, respectively, then there are $x_1 \times x_2$ possible combinations of molecules to interact, and the resulting hazard is

$$h(\boldsymbol{x}, k) = k x_1 x_2. \tag{2.26}$$

Consider the reaction,

$$2X_1 \xrightarrow{k} X_2, \tag{2.27}$$

which is also a second-order reaction as two X_1 molecules are required. If the current X_1 level is x_1, then the number of combinations for two of them to interact is $x_1(x_1 - 1)/2$. Therefore, the reaction hazard is

$$h(\boldsymbol{x}, k) = k \frac{x_1(x_1 - 1)}{2}. \tag{2.28}$$

Continuing the logic and considering the combinations of interacting molecules, we find that the hazard is proportional to the product of the binomial coefficients. Let k_i be the stochastic rate constant associated with a reaction, R_i, and $h_i(\boldsymbol{x}, k_i)$ be the reaction hazard. The generalised form of the hazard function is

$$h_i(\boldsymbol{x}, k_i) = k_i \prod_{j=1}^{n} \binom{x_j}{a_{ij}}. \tag{2.29}$$

The total hazard, describing the rate of any reaction occurring, is the sum of the individual reaction hazards,

$$h_0(\boldsymbol{x}, \boldsymbol{k}) = \sum_{i=1}^{m} \left\{ k_i \prod_{j=1}^{n} \binom{x_j}{a_{ij}} \right\}. \tag{2.30}$$

A reaction event changes the state of the system based upon the species involved. If reaction i occurred at time t_n, the system state, $\boldsymbol{x}$ is updated to be

$$\boldsymbol{x}(t_n) = \boldsymbol{x}(t_m) + S^i. \tag{2.31}$$

Where $\boldsymbol{x}(t_m)$ is the system state after the last reaction, and S^i is the i-th column of the stoichiometry matrix. The evolution of $\boldsymbol{x}(t)$ is then described by a discrete state-space, continuous-time Markov process.

The state of the system, $\boldsymbol{x}(t)$, at time t can be written as

$$\boldsymbol{x}(t) = \boldsymbol{x}_0 + \sum_i S^i R_{i,t}, \tag{2.32}$$

where x_0 is the initial state of the system and $R_{i,t}$ denotes the number of reactions of type i in the interval $(0, t]$. $R_{i,t}$ is a counting process with intensity equivalent to the reaction hazard, $h_i(\boldsymbol{x}(t), k_i)$, which can be expressed as

$$R_{i,t} = Y_i \left(\int_0^t h_i(\boldsymbol{x}(u), k_i) du \right). \tag{2.33}$$

Where Y_i for $i = 1, 2, \ldots, m$, are m independent unit rate Poisson processes. A detailed discussion on this representation of Markov jump processes can be found in Kurtz (1972)

and more recently in Wilkinson (2018).

A Markov jump process can be forward-simulated by a variety of simulation algorithms. However, the computational expense of simulation can be high, particularly when the number of species and reactions is high, which generally increases the reactivity of the system. A highly reactive system is often more computationally expensive to simulate as the number of reaction simulations increases. As a result, approximation algorithms are often used, which reduce the computational cost. Two methods to simulate a system under these assumptions are described below. The first being an exact simulation algorithm and the second a popular approximation. However, there are many other approximation algorithms (Golightly & Gillespie, 2013). The two methods presented here are used in Chapter 5.

2.6.3 Gillespie's direct method

Gillespie's direct method (Gillespie, 1977) is an exact simulation algorithm and the most common method to exactly simulate the Markov jump process representation of the reaction network. The method describes the inter-event time as an exponential random variable, whose rate is the total system hazard, $h_0(\boldsymbol{x}, \boldsymbol{k})$, and the probability of reaction R_i occurring is proportional to the reaction hazard, $h_i(\boldsymbol{x}, \boldsymbol{k})$. Let $\boldsymbol{x}_0$ be the initial system state, and assume the stochastic rate constants, $\boldsymbol{k}$, are known. To simulate the system up to a maximum time, T_{max}, the algorithm is described in Algorithm 3.

Algorithm 3 Gillespie's direct simulation method

1. Set $t = 0$. Initialise the system state $\boldsymbol{x} = \boldsymbol{x}_0$.
2. Calculate reaction hazards, for $i = 1, 2, \ldots, m$

$$h_i(\boldsymbol{x}, k_i) = k_i \prod_{j=1}^{n} \binom{x_j}{a_{ij}}.$$

3. Calculate the system hazard

$$h_0(\boldsymbol{x}, \boldsymbol{k}) = \sum_{i=1}^{m} h_i(\boldsymbol{x}, k_i).$$

4. Simulate inter-event time, $\delta t \sim \text{Exp}\left(h_0(\boldsymbol{x}, \boldsymbol{k})\right)$, and put $t := t + \delta t$.
5. Simulate the reaction index, j, from a discrete distribution with probabilities. For $j = 1, 2, \ldots, m$,

$$\Pr(\text{reaction index} = j) = \frac{h_j(\boldsymbol{x}, k_j)}{h_0(\boldsymbol{x}, \boldsymbol{k})}.$$

6. Update system state, $\boldsymbol{x}$, according the reaction R_j.
7. If $t < T_{max}$, return to step 2.

The exact simulations of the direct method can lead to high computational expense, particularly for complex or highly reactive systems. An exact history of the system can also be computationally expensive to store in a computer's memory. In practice, the

system state is saved a set of predefined system times, as the exact history is not practical and often not of interest.

2.6.4 Poisson and tau-leap algorithms

The Poisson leap algorithm is an approximate simulation algorithm that approximates the number of reactions of each type over a time step, Δt. The approach is based on a Poisson process, which simulates a Markov jump process with a constant hazard rate and, consequently, assumes that the system's hazard rate is approximately constant over the time step, Δt. Intuitively, this makes sense, as a constant hazard implies no change in the system's reactivity, and a system whose reactivity changes significantly over a time step cannot be accurately approximated. To reduce computational expense Δt should be much larger than the typical inter-event time calculated by Gillespie's direct method. For a known set of stochastic rate constants, $\boldsymbol{k}$, initial condition, $\boldsymbol{x}_0$, and time step, Δt, the Poisson algorithm is described in Algorithm 4.

Algorithm 4 Poisson leap approximate simulation algorithm

1. Set $t = 0$, and initialise the system state $\boldsymbol{x} = \boldsymbol{x}_0$.
2. Calculate reaction hazards, for $i = 1, 2, \ldots, m$

$$h_i(\boldsymbol{x}, k_i) = k_i \prod_{j=1}^{n} \binom{x_j}{a_{ij}}.$$

3. Calculate the system hazard

$$h_0(\boldsymbol{x}, \boldsymbol{k}) = \sum_{i=1}^{m} h_i(\boldsymbol{x}, k_i).$$

4. Simulate the m-dimensional reaction vector, $\boldsymbol{r}$, such that

$$r_i \sim \text{Po}(h_i(\boldsymbol{x}, k_i)\Delta t).$$

5. Update $\boldsymbol{x} := \boldsymbol{x} + S\boldsymbol{r}$ and put $t := t + \Delta t$.
6. If $t < T_{max}$, return to step 2.

Clearly, the choice of Δt is important; too small and computational savings are little to none; too big, and the approximation accuracy falls. Additionally, the appropriate size of Δt may change throughout the simulation as the system becomes more or less reactive. The Poisson leap algorithm requires a constant choice of Δt. However, Gillespie proposed a τ-leap algorithm, whose time step, τ, is variable and changes depending on the system's current state (Gillespie, 2001). The choice of the next time step, τ, is still difficult. The ideal option is a trade-off between maximising computational savings (making τ as large as possible) and maintaining the assumption of an approximately constant system hazard. Gillespie suggested the value of τ should result in a relatively small change in the reaction hazards, $h_i(\boldsymbol{x}, k_i)$, thus maintaining a relatively constant system hazard. The algorithm follows similar steps to the Poisson leap algorithm, with the addition of a τ calculation before simulating the number of reactions of each type in Step 4.

2.6.5 Example: birth-death model

To give context to the stochastic kinetic theory discussed in this chapter, we discuss some examples. The first, discussed here, is the birth-death model. It is a fairly simple model of population dynamics which only considers one species and two possible reactions: birth and death. The model is a continuous-time extension of the cell population example described in Chapter 1.5.3.

Let X denote the species and x be its population. If a birth occurs, the species' population increases by one; alternatively, if a death occurs, the population decreases by one. Using the chemical reaction network introduced earlier, the system can be described by the equations

$$\begin{aligned} X &\overset{k_{\text{birth}}}{\rightarrow} 2X, \\ X &\overset{k_{\text{death}}}{\rightarrow} \phi. \end{aligned} \tag{2.34}$$

The symbol ϕ indicates that no elements are present on that side of the equation. The stochastic rate constants of the two reactions are denoted k_b and k_d for birth and death, and the resulting reaction hazards are $k_b x$ and $k_d x$, respectively. The stoichiometry matrix is a row vector with $S = (1, -1)$.

The dynamics of the population depend on the values of the birth and death rates. Equal birth and death rates imply that the population is expected to remain fairly stable. However, due to the stochastic nature of the system, fluctuations in the population are natural, as shown in Figure 2.3(a). A higher birth rate, $k_{\text{birth}} > k_{\text{death}}$, results in an increasing population size that grows without bounds. A higher death rate, $k_{\text{birth}} < k_{\text{death}}$, means the population is destined for extinction, a state which cannot be escaped. Both scenarios are depicted in Figures 2.3(b) and 2.3(c), respectively. The stochasticity of the system means that no particular behaviour is guaranteed, not considering steady states. However, the overall behaviour summarised by many simulations of the system can be described.

2.6.6 Example: mtDNA population dynamics

General model

In this chapter, stochastic kinetic models of single-cell mtDNA population dynamics are introduced. As is common in mathematical models of mtDNA dynamics, we consider two species of mtDNA: wild-type, W, and variant-type, V, (Ainsworth, 2014; Elson et al., 2001; Henderson et al., 2009). Eq. 2.35 shows the four first-order chemical reactions describing mtDNA replication, degradation.

$$\begin{aligned} W &\overset{k_1}{\rightarrow} 2W \\ V &\overset{k_2}{\rightarrow} 2V \\ W &\overset{k_3}{\rightarrow} \phi \\ V &\overset{k_4}{\rightarrow} \phi \end{aligned} \tag{2.35}$$

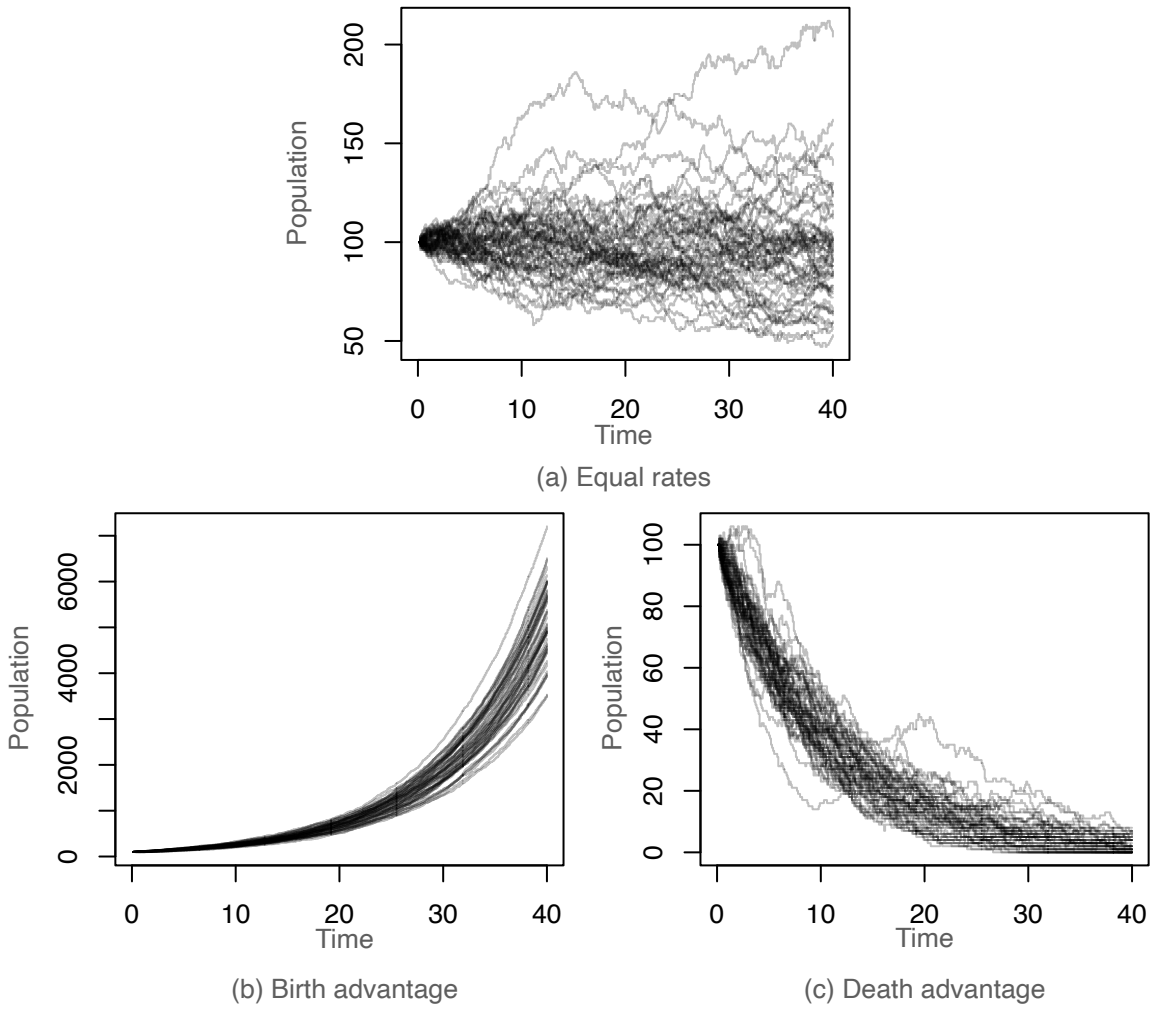


Figure 2.3: **Realisations from the birth-death model**. Fifty realisations from the birth-death model with a varying death rate. (a) Shows equal birth and death rates, $k_{\text{birth}} = k_{\text{death}} = 1.0 \times 10^{-1}$. (b) Shows a birth-rate advantage, $k_{\text{birth}} = 2.0e - 1$ and $k_{\text{death}} = 1.0e - 01$. (c) Shows a death-rate advantage $k_{\text{birth}} = 1.0 \times 10^{-1}$ and $k_{\text{death}} = 2.0 \times 10^{-2}$. All simulations were executed using Gillespie's direct method and initialised with the same initial population of 100.

The resulting net effect matrix is

$$S^{\mathrm{T}} = \begin{pmatrix} -1 & 0 \\ 0 & -1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}, \tag{2.36}$$

and the reaction hazard functions are

$$\begin{aligned} h_1(\boldsymbol{x}, k_1) &= k_1 W, \\ h_2(\boldsymbol{x}, k_2) &= k_2 V, \\ h_3(\boldsymbol{x}, k_3) &= k_3 W, \\ h_4(\boldsymbol{x}, k_4) &= k_4 V. \end{aligned} \tag{2.37}$$

Theories of clonal expansion can be imparted into the model by the constraints on the stochastic rate constants, $\boldsymbol{k}$. For example, random genetic drift would assume $k_1 = k_2$, giving no replicative advantage between the species, and survival of the smallest would

constrain $k_2 > k_1$, imposing a variant-type replicative advantage. Homoplasmy, of either species, is a variant load stationary distribution, which the system cannot escape. The raw output of the simulations are the population levels of the two species. However, it is more informative to view the mtDNA copy number and variant load. The two statistics capture all dynamics within the system, and no information is lost by the transformation. For the remainder of this chapter and thesis, simulation output of mtDNA populations is presented in this form. The inclusion of mtDNA mutation events is discussed in Chapter 5.

MtDNA copy number

A key component to modelling mtDNA population dynamics is copy number control. Copy number remains fairly constant in healthy cells, and without a mechanism controlling it, a mathematical model would allow mtDNA to go extinct or the population to explode to unrealistic values. The effect of an uncontrolled population can be seen in the birth-death example, Figure 2.3. When the birth rate is less than the death rate, the population falls towards extinction; conversely, when the birth rate is higher than the death rate, the population grows without bounds. The same would be seen in a model of mtDNA populations without a copy number control mechanism. For the remainder of this chapter, a copy number controller has been implemented to maintain a relatively stable population size, but the specifics are not discussed here. Control mechanisms are discussed in Chapter 5. Figure 2.4 demonstrates the effect of the implemented copy number control mechanism over a period of 10 years. The system considers the target copy number to be 200, and maintains this by dynamically altering the replication rate of mtDNA. The result maintains a copy number centred around 200 with a range of approximately 50.

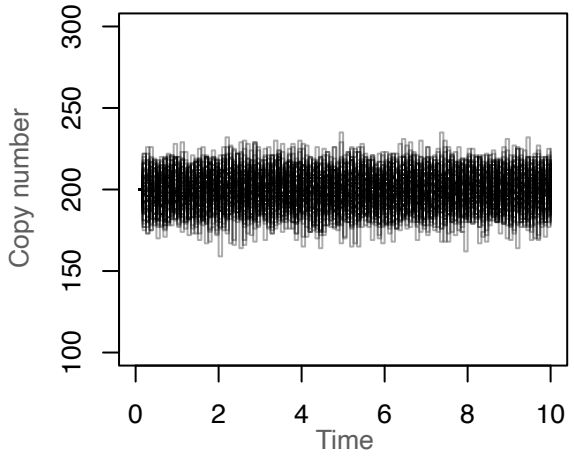


Figure 2.4: **Copy number control**. Copy number from 100 independent simulations of mtDNA dynamics described in Equation 2.35, with an implemented copy number control mechanism targeting a copy number of 200.

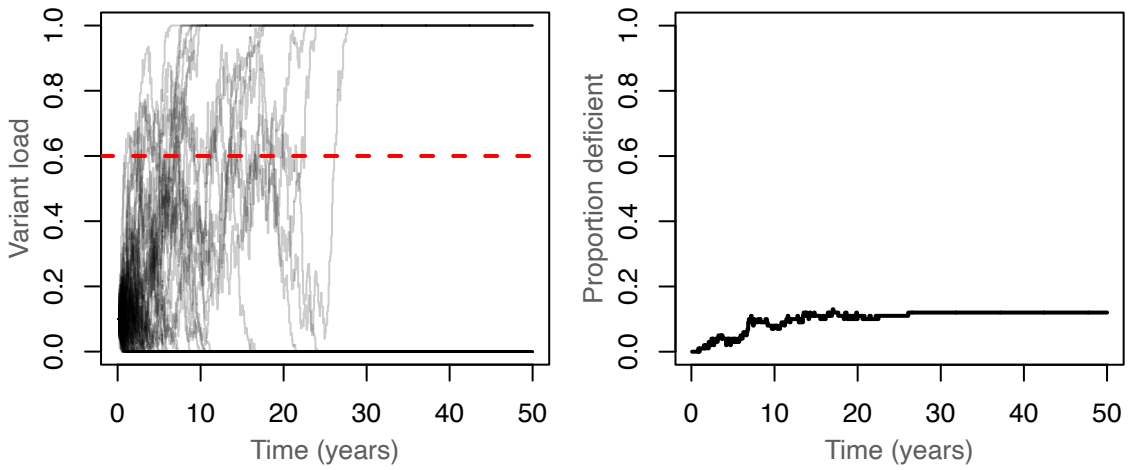
Simulations

The system defined in Equation 2.35 can yield a wide variety of dynamics depending on the reaction rates and initial conditions. Here, a number of simulations are presented with alterations to their reaction rates or initial conditions. The copy number associated

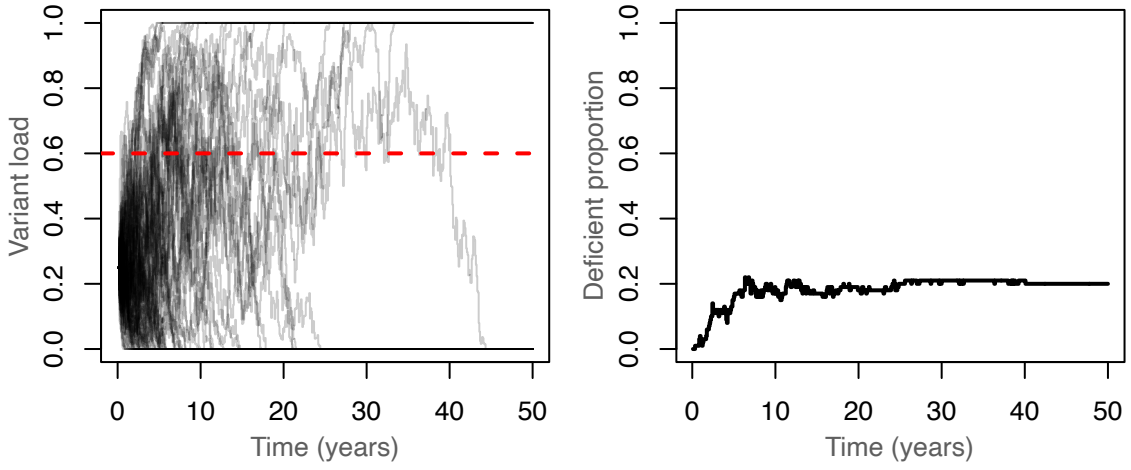
with each simulation is not of particular interest, as they are tightly controlled; see Figure 2.4 for an example, as such they not shown. The degradation rate was taken from the literature (Gross et al., 1969). All simulations possessed the same copy number control mechanism and target copy number, 200. Also included within the simulations is a pathogenic threshold; if the simulated cell's variant load is above this, the cell is considered to be OXPHOS deficient. The pathogenic threshold is set at 60% (Hayashi et al., 1991; Hernández-Ainsa et al., 2022). A more detailed discussion about these parameter values is given in Chapter 5.

Figure 2.5 demonstrates how mtDNA dynamics are affected by the initial variant load, assuming zero probability of *de novo* mutations and no species advantage. Unsurprisingly, an increase in initial variant load results in increased OXPHOS deficiency. Perhaps more surprising is that the proportion of deficiency reaches a plateau. This is primarily caused by simulations reaching mtDNA homoplasy, a state which they cannot escape. The deficiency proportion at the plateau appears to be directly correlated with the initial variant load, with higher initial variant load having higher variant load plateaus.

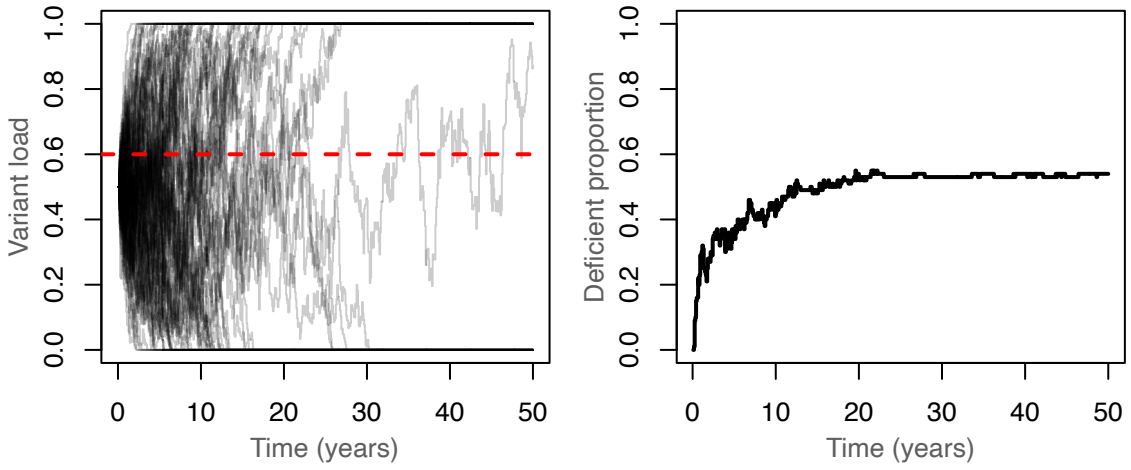
As mentioned, there is a fairly high degree of uncertainty in the rate of mtDNA replication, Chapter 1.2. Figure 2.6 demonstrates how an increase in degradation (and replication) rate(s) accelerates the system dynamics, increasing the rate at which simulations reach stationary distributions (mtDNA homoplasmy), for a model of no species advantage and no *de novo* mutations. The values of degradation rate used capture the range of reported half-lives in the literature, 1d and 100d (Insalata et al., 2022). The difference between the two simulations is clear. When mtDNA has a very low half-life, mtDNA turnover is very fast, accelerating dynamics. All simulations with low mtDNA half-life reach a homoplasmic state within 3 years. In contrast, a high half-life slows dynamics, drastically increasing the time taken for the system to reach a homoplasmic state.



(a) Initial variant load: 10%

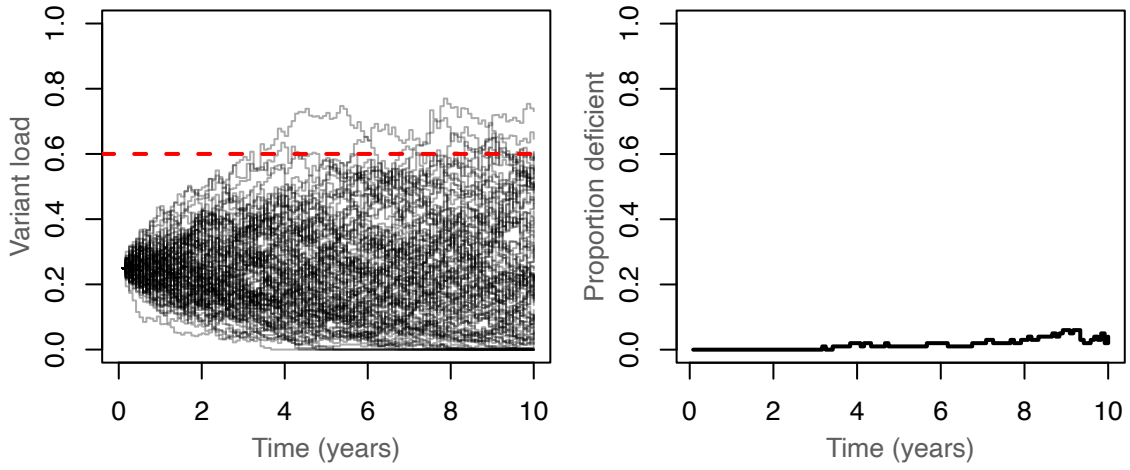


(b) Initial variant load: 25%

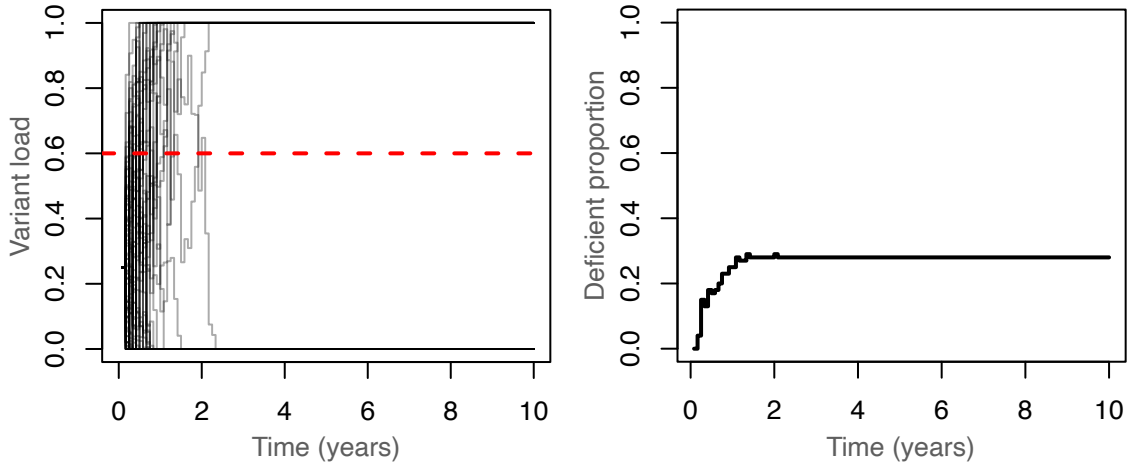


(c) Initial variant load: 50%

Figure 2.5: **Initial variant load impact on OXPHOS deficiency**. One hundred simulations of single-cell mtDNA population dynamics. All simulations used the same copy number control mechanism and mtDNA degradation rates, equivalent to a half-life of 17.7d (Gross et al., 1969). The pathogenic threshold is assumed to be 60% (Hayashi et al., 1991; Hernández-Ainsa et al., 2022). The simulations were executed using Gillespie's direct method, Algorithm 3 (Gillespie, 1977).



(a) Low replication rate



(b) High replication rate

Figure 2.6: **Impact of mitochondrial turnover on mtDNA dynamics**. One hundred simulations of single-cell mtDNA population dynamics. All simulations used the same copy number control mechanism, targeting 200 mtDNA molecules, and initialised with a 25% variant load. MtDNA degradation rates were equivalent to mtDNA half-lives of 100d and 2d for plots (a) and (b), respectively (Diaz & Moraes, 2008; Gross et al., 1968; Johnston & Jones, 2016). The pathogenic threshold is assumed to be 60% (Hayashi et al., 1991; Hernández-Ainsa et al., 2022). The simulations were executed using Gillespie's direct method, Algorithm 3 (Gillespie, 1977).

Chapter 3

Blood Cell Analysis

3.1 Introduction

Direct measurements of mtDNA population dynamics are difficult to collect. Ethical and financial issues often limit data collection to a relatively small number of patients, and data are rarely gathered from multiple time points within the same patient. Some of these problems may be eased by collecting measurements from blood cells, which is comparatively easy compared to other tissues, such as muscle. Hence, multiple studies have investigated mitochondrial disease in the blood.

Investigating clonal expansion within blood cells presents a new level of biological complexity compared to post-mitotic cells. Some blood cells are highly dynamic, with constant cellular replication and degradation increasing variant-load stochasticity through the random segregation of mtDNA during cell division and periods of high mtDNA proliferation. Additionally, blood cells have a complex lineage of maturation and development. Nevertheless, if these problems can be overcome, blood is a convenient tissue for collecting data for longitudinal studies of clonal expansion, which would be much more difficult in other tissues.

In this chapter, statistical techniques are used to analyse single-cell variant load data from various blood cell types. The development of a mathematical model of blood cell development is discussed in the context of an investigation into clonal expansion. Before this, the function and development of relevant blood cells are discussed.

3.1.1 Blood cell biology

Blood cells are synonymous with the immune system, with most cells involved in the immune response being found within the blood. Although a large variety of blood cells exists, here, the discussion is limited to cells within the lymphoid lineage, as these are the subject of the dataset used in this chapter. The cells within the lymphoid lineage can be broadly categorised into two main types: B cells and T cells.

T cells

T cells are a part of the adaptive immune response and can generally be separated by their ability to respond to foreign antigens. Antigens are presented by the extracellular major histocompatability complexs (MHCs), of which there are two types. MHCs of Type I are intracellular and present proteins foreign to the cell, while MHCs of Type II are exogenous and present peptides from broken down, endocytosed bacteria (Sun et al., 2023). CD4+ T cells, also known as helper T cells, recognise MHC Type II, while CD8+ T cells, or cytotoxic T cells, recognise MHC Type I molecules (Sun et al., 2023). Each T cell possesses an antigen-specific T cell receptor (TCR), which binds to an MHC. During cell development, precursor naïve T cells undergo TCR recombination and mass proliferation, ensuring that the naïve T cell population possesses a wide variety of antigen-specific cells. Self-reactive or under-reactive precursor naïve T cells are removed from the population by apoptosis.

Once an MHC molecule is recognised, a T cell becomes active and undergoes proliferation; before this, the cell is in a naïve state. Activated T cells undergo asymmetric cellular division to produce one cell primed to be a memory T cell and one primed to be an effector T cell (Pollizzi et al., 2016; Verbist et al., 2016), which go on to differentiate into their non-naïve counterparts. CD8+ T cells proliferate and develop into effector or memory CD8+ T cells. Effector CD8+ T cells kill infected cells and induce phagocytosis of bacterial pathogens via the release of cytokines, while memory cells are long-lived and remain in the body for a quick and efficient response to future reinfection. An activated CD4+ T cell also proliferates into effector and memory cells. Effector CD4+ T cells are involved in coordinating the adaptive and innate immune response. Like CD8+ memory T cells, CD4+ memory T cells are long-lived, remaining in the blood for a rapid response to reinfection. It has been shown that the activation of effector cells is OXPHOS dependent, and deficiencies in CI and CIV lead to activation defects and proliferation inhibition (Tarasenko et al., 2017; Vardhana et al., 2020; Yi et al., 2006). However, the mechanisms controlling this are unclear, and a switch to a glycolytic state facilitates mass proliferation.

Memory T cells, for both CD8+ and CD4+ cells, can be further sub-categorised based on their function. Central memory T cell (T_{CM}) migrate to the secondary lymphoid organs, where they remain until reinfection. Effector memory T cell (T_{EM}) remain in the blood (Sallusto et al., 1999) and are the first response to reinfection. T_{EM} cells are short-lived cells compared to T_{CM} cells, with a lower proliferative ability; their role in the immune system response is to contain the infection, allowing T_{CM} cells to proliferate to high numbers. It has been shown that gene methylation is lost at a higher rate in T_{EM} compared to T_{CM} and in T_{CM} compared to naïve T cells, indicating an increase in cell maturity from naïve T cells to T_{CM} to T_{EM} cells (Abdelsamed et al., 2017; Durek et al., 2016). Another memory T cell sub-category is the memory T cells re-expressing naïve marker CD45Ra (T_{EMRA}) cell, which re-expresses naïve cell markers. The maturity of T_{EMRA} cells is debated in the literature. Some work suggests that they are the most terminally differentiated of the T cell compartment, showing lower proliferative ability and shorter telomeres, compared to T_{CM} and T_{EM} cells (Geginat et al., 2003; Verma et al., 2017). However, Rufer et al. (2003) suggest that T_{EMRA} cells are intermediary cells between naïve and effector cells, citing their similarities in cell markers to naïve cells. Later work by Verma et al. (2017) found that T_{EMRA} cells could be divided into two populations: *young* and *old*, based on the cell marker CD57. CD57- T_{EMRA} cells show higher

proliferative ability, longer telomeres, and terminal placisity, while CD57+ T_{EMRA} cells showed terminal differentiation. Each sub-category of memory T cell can differentiate into effector T cells, indicating that the maturity of these cells is non-linear.

B cells

B cells drive the humoral immune response. Their development occurs in the bone marrow before release into the periphery as immature B cells. Similar to T cells, B cells have high fidelity and cells with self-reactive clones or dysfunctional B cell receptors (BCRs), an extracellular antigen-specific protein binding sites, are removed (Kelsoe, 1996).

The activation of naïve B cells is a two-signal process. The first signal begins with an antigen binding to the BCR before the antigen is endocytosed and expressed by an MHC type II molecule. The majority of B-cell activation is T-cell dependent. Once the antigen is presented by the MHC, an activated effector CD4+ T cell can bind to it and signal activation; without both signals, the B cell undergoes apoptosis (Akkaya et al., 2018; Chesnut & Grey, 1981).

An activated B cell starts to proliferate and develop into effector and memory B cells. Mass proliferation is accompanied by an increase in glycolysis (Jellusova et al., 2017), somatic mutations to the BCR, and class switch recombination (Victora & Nussenzweig, 2022). These processes result in a smaller proportion of cells undergoing apoptosis compared to T cells and increase BCR binding affinity. Memory B cells are similar to memory T cells in that they remain in the blood for long periods in preparation for a second infection. Effector B cells, also known as plasma cells, proliferate to high numbers and begin to produce antigen-specific antibodies. Plasma cells can be further segregated into two types: long- and short-lived. Short-lived plasma cells are responsible for producing antibodies for the immediate immune response. Long-lived plasma cells move to the bone marrow, where they continue to produce and secrete antibodies for many years (Cyster & Allen, 2019).

Cell lineage

The origin of blood cells is the multipotent haematopoietic stem cell (HSC), found in the bone marrow. HSCs differentiate to become progenitor cells, a range of cell types which go on to further differentiate into specialised cells, including early T cell precursors (ETPs) and B cell and natural killer (B/NK) cell progenitors, which differentiate into T and B cells, respectively. Figure 3.1 shows a lineage diagram that highlights the relationships between B and T cells and how cell differentiation progresses between types.

3.1.2 M.3243A>G and blood cells

The dataset analysed here consists of patients with a diagnosed m.3243A>G single-point variant. Early studies of the mtDNA variant noted a significant difference between its level in blood cells and post-mitotic muscle cells, showing a significantly lower, or undetectable, variant load compared to muscle tissue (Rahman et al., 2001; Sue et al., 1998).

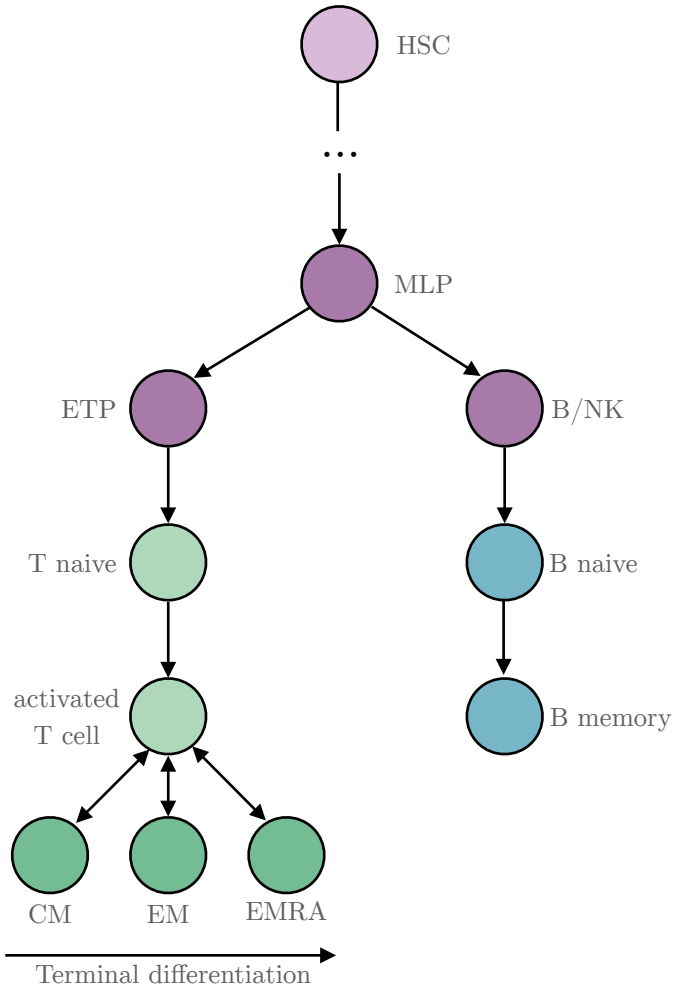


Figure 3.1: **Blood cell lineage diagram**. Blood cells originate from HSC, the root cell in the diagram. A number of cell types exist between HSCs and multilymphoid progenitors (MLP). However, these are not represented in the dataset. The T cell compartment is green, originating from early T cell precursors (ETPs). The B cell compartment is blue and originates from B and natural killer cell progenitors (B/NKs). The diagram is not specific to either CD4+ or CD8+ cells and is applicable to both. CD34+ progenitor cells are shown in purple. The maturation of memory T cell sub-types is shown in the terminal differentiation, although the maturation of T_{EMRA} is not clear.

Since then, longitudinal studies have shown that m.3243A>G variant loads decline over time (de Laat et al., 2012; Langdahl et al., 2018; Veitia, 2018). Mean variant-load in post-mitotic muscle tissue has been shown to remain fairly constant throughout the human lifespan. Additionally, the age-adjusted blood variant load is consistent with variant loads measured in muscle (Grady et al., 2018).

The decline in m.3243A>G load is hypothesised to occur during HSC replication. Rajasimha et al. (2008) proposed a mathematical model of HSCs, which randomly developed into progenitor cells during replication. The mtDNA populations of each cell were modelled stochastically, and any cell with a variant load above a pathogenic threshold (several values were considered) was removed from the population. The pathogenic threshold mechanism removes the variant homoplasmy steady state of mtDNA dynamics. Therefore, given enough time, all cells in the model would reach wild-type homoplasmy,

resulting in a decline in variant load throughout the simulations. The HSC population was simulated for approximately one lifespan, 100 years, and the decline in variant mtDNA was consistent with those observed in patients. Rajasimha *et al.* also simulated symmetric cellular division for short-lived cells, ≈25d (days), and showed that these maintain a steady mean variant level. The shorter lifespan of the cells is not enough time for the cells to reach wild-type homoplasmy or for high variant-load cells to be degraded, resulting in no detectable reduction in mean variant load. The results of the model are supported by observations that the variant level within foetal tissues is consistent (Monnot et al., 2011), implying that a negative selection against variant mtDNA is causing the decline rather than inheritance. In addition, a significant reduction in mtDNA variant level has been noted in other mitotic cells, indicating that cellular replication is driving the selection (Frederiksen et al., 2006; Su et al., 2018).

3.1.3 General aim

The decline in m.3243A>G levels within blood cells makes them uniquely interesting in the investigation of clonal expansion. In this chapter, the feasibility of using blood cell data to investigate the mechanisms of clonal expansion is investigated.

3.2 Data

To investigate the enhanced decline of m.3243A>G within blood cells, investigators collected aggregate and single-cell variant load data from a cohort of patients, all with m.3243A>G characterised mitochondrial disease, and a range of cell types (Franklin et al., 2023). Patients were recruited through the NHS Highly Specialised Service for Rare Mitochondrial Disorders in Newcastle upon Tyne, UK, and must not have been showing symptoms of an infection, confirmed via lymphocyte and myeloid cell counts. Prior to participation, informed, written consent was gathered from all patients. The Newcastle and North Tyneside Research Ethics Committee provided ethical approval (REC:19/LO/0117).

3.2.1 Aggregate variant load findings

The variant load (aggregate) was measured for a number of cell types, including CD34+ progenitors, B cells, as well as CD4+ and CD8+ T cells, within the patient samples. Across cell types, a significant negative correlation was found between variant load and age. All T cell subsets showed a significant reduction in the m.3243A>G level when compared to the CD34+ progenitors. Memory T cells showed a significant reduction when compared to the naïve T cells, indicating an enhanced reduction during naïve T cell differentiation into memory cells. The maturity-related decline in variant load was not seen in the B cell compartment.

Previous work shows that the decline in m.3243A>G level with age is not found in other point mutations (Altmann et al., 2016; Shoffner et al., 1990; Yoneda et al., 1990). Analysis of the m.8344A>G patient samples collected here showed results consistent with

the decline in m.3243A>G levels. Memory CD4+ T cells, memory CD8+ T cells, and memory B cells all showed a reduction in m.8433A>G levels compared to CD34+ progenitor cells.

The enhanced reduction in m.3243A>G level within the T cell compartment led the investigators to query mtDNA dynamics within these cells. Therefore, it was decided to gather single-cell variant-level data for a subset of patients.

3.2.2 Single-cell m.3243A>G level data

Six patients from the cohort were selected for single-cell analysis. This subset is the dataset which is under consideration in the remainder of this chapter and is summarised in Table 3.1. The chosen patients' ages range between 18 and 46, with an NMDAS scaled score range of 17.6 to 45.2. Single-cell m.3243A>G levels were collected on a number of cell types within the T-cell compartment. CD34+ progenitor cells are directly dervived from HSC compartment and are the most naive cells investigated, and are therefore used a reference to compare the accelerated loss of m.3243A>G within the T cell compartment. Data were collected from specific memory T cell types: T_{CM}, T_{EM}, and T_{EMRA}. Additionally, variant load measurements were collected for B naïve/memory cells and monocytes. Monocytes, also originally derived from HSCs, are distantly related to B and T cells. Their lineage breaks away from the T and B cells prior to the multilymphoid progenitors. These are another cell type which can be used to investigate whether the enhanced reduction is specific to the T cell compartment.

Patient ID	Age	Sex	Scaled Total MDAS	Variant load (%)	
				Blood (age)	Urine
P10	24	M	17.6	59 (22)	92
P15	45	M	45.2	23 (45)	85
P17	46	F	17.6	9 (46)	64
P18	24	F	14.5	36 (24)	80
P19	20	F	24.9	59 (20)	70
P22	18	F	24.5	60 (18)	96

Table 3.1: **Patient summary for single-cell analysis**. The subset of patients from the cohort which were chosen for single-cell analysis. Aggregate variant loads are given in the table.

The observed variant loads spanned (almost) the entire support, with CD34+ progenitors and monocytes generally exhibiting the widest ranges, a trend which is particularly highlighted in younger patients. All patients show a spike of near-zero variant loads within the T cell compartments, which becomes more pronounced with the patient's age, and older patients show the same spike in CD34+ progenitors, monocytes and B cells. See Figure 3.2 for a subset of the data.

To reveal some of the underlying mtDNA dynamics that are driving the eradication, the proportion of cells that have completely removed the variant from their population, and thus reached wild-type homoplasmy, must be quantified. Unfortunately, observational error is believed to be in the data, and all cells whose variant load is within the near-zero spike are believed to have reached wild-type homoplasmy. For the remainder of

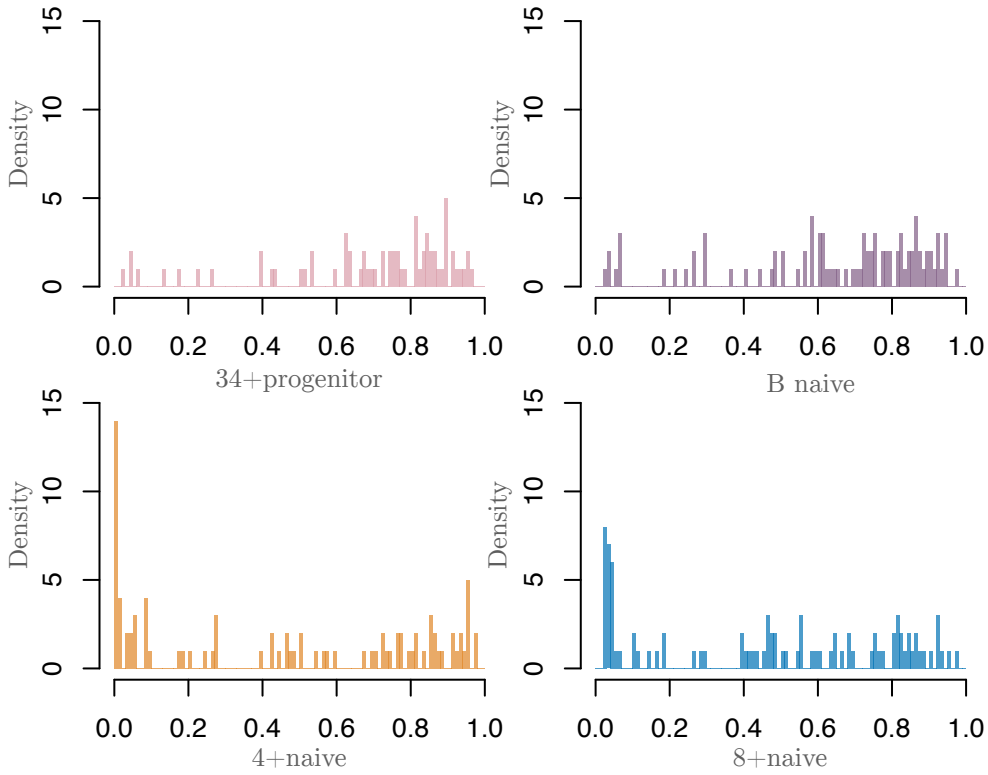


Figure 3.2: **Single-cell variant load shows a near-zero spike within T cell compartment**. Histograms of variant-load within specific cell types collected from patient P22, aged 18. Approximately 100 single-cell observations were collected per cell type, per patient, leading to a non-smooth density.

this chapter, all cells within the near-zero variant load spike are considered to have eradicated m.3243A>G from their mtDNA population and be wild-type homoplasmic, and, therefore, references to the proportion of wild-type homoplasmic cells are in fact referring to the proportion of cells belonging to the near-zero spike.

3.2.3 Data analysis aim

Robustly estimate the proportion of blood cells which have reached wild-type homoplasmy for each cell type within the six patients with single-cell observations.

3.3 Methods

3.3.1 Mixture model

A mixture model naturally fits the modelling task, incorporating component proportions and allowing different density functions to describe each aspect of the data. To allow the wild-type homoplasmy to be a natural parameter in the model, one component of the mixture model must model this aspect of the data. Proportion data is constrained in the range $[0.0, 1.0]$, which may suggest the use of a Beta distribution. However, Beta distributions are somewhat restrictive in their shape and location. Instead, a truncated

normal distribution is proposed, which allows more freedom in the location and variance of the near-zero spike. Data outside the near-zero spike is relatively flat, with no observable or physical interpretation of other sub-populations. Therefore, we propose the second component of the mixture to be a uniform distribution, whose density is non-zero in the range $[0.0, 1.0]$. Although the second component's parameters (range) are fixed, the density of the component is still variable due to its associated proportion. Let $Y_{i,j,k}$ be the variant load of the k-th observation form the j-th cell-type in the i-th patient, the model is

$$Y_{i,j,k}|\pi_{i,j}, \mu_{i,j}, \sigma_{i,j} \sim \pi_{i,j} N_{[0,1]}\left(\mu_{i,j}, {\sigma_{i,j}}^2\right) + (1 - \pi_{i,j})U(0, 1). \tag{3.1}$$

The parameters to be inferred are $\pi_{i,j}, \mu_{i,j}$, and $\sigma_{i,j}$ for all i and j, denoting the wild-type homoplamsy proportions, spike centres and spike standard deviations, respectively.

The model assumes no relationship between wild-type homoplamsy proportions between cell types or patients, as the spike is believed to be a consequence of noise due to sequencing errors. A hierarchy could be added to the model, which infers collection errors for each batch. However, batch effects are not of interest here. In addition, if the model were to be used again for different datasets where this experimental structure did not exist, it would be beneficial to know if the model was still able to infer the wild-type homoplamsy proportion. Therefore, each patient and cell-type combination is treated as an independent dataset, and the model notation can be simplified to reflect this by removing subscripts. Let Y_k be the k-th observation of the current dataset (patient and cell type). The model is

$$Y_k|\pi, \mu, \sigma \sim \pi N_{[0,1]}\left(\mu, \sigma^2\right) + (1 - \pi)U(0, 1). \tag{3.2}$$

3.3.2 Prior beliefs

The prior beliefs for π, the wild-type homoplasmy proportion, vary depending on cell type and patient. To allow the model to be applied to other datasets, a general prior belief is proposed, which can be used for all patients and cell types. Another consideration is that the spike is not seen in all patients and cell types. However, it is not known which cell types and patients the spike will appear in *a priori*. We, therefore, wish to construct a prior belief that reflects the uncertainty in the spike's existence.

Let π_0 be the unknown probability of spike existence. The probability naturally leads to a Beta prior, whose support is $[0, 1]$, and so

$$\pi_0 \sim \text{Beta}(a_0, b_0). \tag{3.3}$$

A priori its value is very uncertain; an uninformed, flat prior can be achieved with values $a_0 = b_0 = 1$.

The probability of spike existence is used as part of a two-component mixture distribution for the prior beliefs of the proportion of observations belonging to the spike. The first component, weighted with π_0, is associated with no spike existing. If no spike exists, no observations can belong to it, and so the component must have a proportion of 0.0 with probability 1.0. The second component, weighted by $1 - \pi_0$, is the prior belief of the proportion, given the spike exists and should have non-zero weight over the support.

Let δ_0 be a density function, such that if $X \sim \delta_0$, $\Pr(X = 0) = 1.0$ i.e. it's only possible value is 0.0. The two-component mixture prior distribution of π is then

$$\pi|\pi_0 \sim \pi_0\delta_0 + (1 - \pi_0)\text{Beta}(a, b). \tag{3.4}$$

The parameters a and b must be chosen. Experts have a high degree of uncertainty in spike proportion given its existence, so $a = b = 1$ are chosen to give equal probability to all areas of support. The resulting prior density for π has a point mass at $\pi = 0$ and is flat on the range $(0, 1]$, see Figure 3.3.

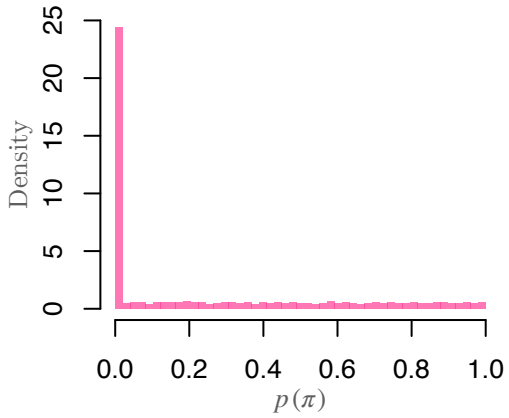


Figure 3.3: **Spike proportion prior beliefs**. A histogram of 100,000 random draws from the prior beliefs of the near-zero spike proportion, $p(\pi)$.

Prior beliefs must be chosen for the remaining two parameters: the spike location and variance. If a near-zero spike is present, its location must be, by definition, near zero. A flat prior on the interval $[0, 0.2]$ is chosen to reflect this. Outside of this range, the prior density is zero, giving no weight to that area of the support, and the prior is $\mu \sim \text{U}(0, 0.2)$. If a spike exists, it must have a relatively small variance to be a visible spike in the data. Therefore, prior beliefs for the spike's standard deviation are $\sigma \sim \text{Exp}(5)$. The resulting expected spike variance is $\text{E}[\sigma] = 0.2$. The model structure can be seen in Figure 3.4.

3.3.3 Computational methods

The model is neither conjugate nor semi-conjugate, and therefore, the posterior cannot be found analytically or sampled using a Gibbs sampler; it must be sampled from using some other MCMC scheme. Constructing a Metropolis-Hastings scheme and selecting an appropriate proposal distribution can be challenging. Instead, parameter inference is conducted by the statistical software `JAGS` (Plummer, 2003). Initial runs of the MCMC scheme showed that the posterior draws possess a high degree of autocorrelation, so the output was thinned with a lag of 100, resulting in almost uncorrelated posterior draws. For each cell type and patient data, three chains were executed with 1,000,000 iterations (not including lag), of which the first half was considered to be the burn-in period and was removed. The remaining posterior draws were thinned, resulting in 5,000 (almost) uncorrelated draws from the posterior. No chains showed signs of non-convergence by inspection of trace plots and $\hat{\text{R}}$ values.

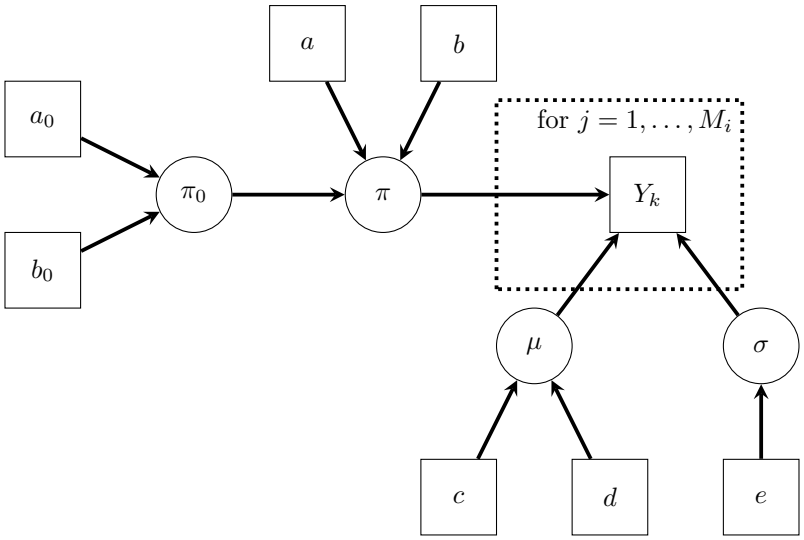
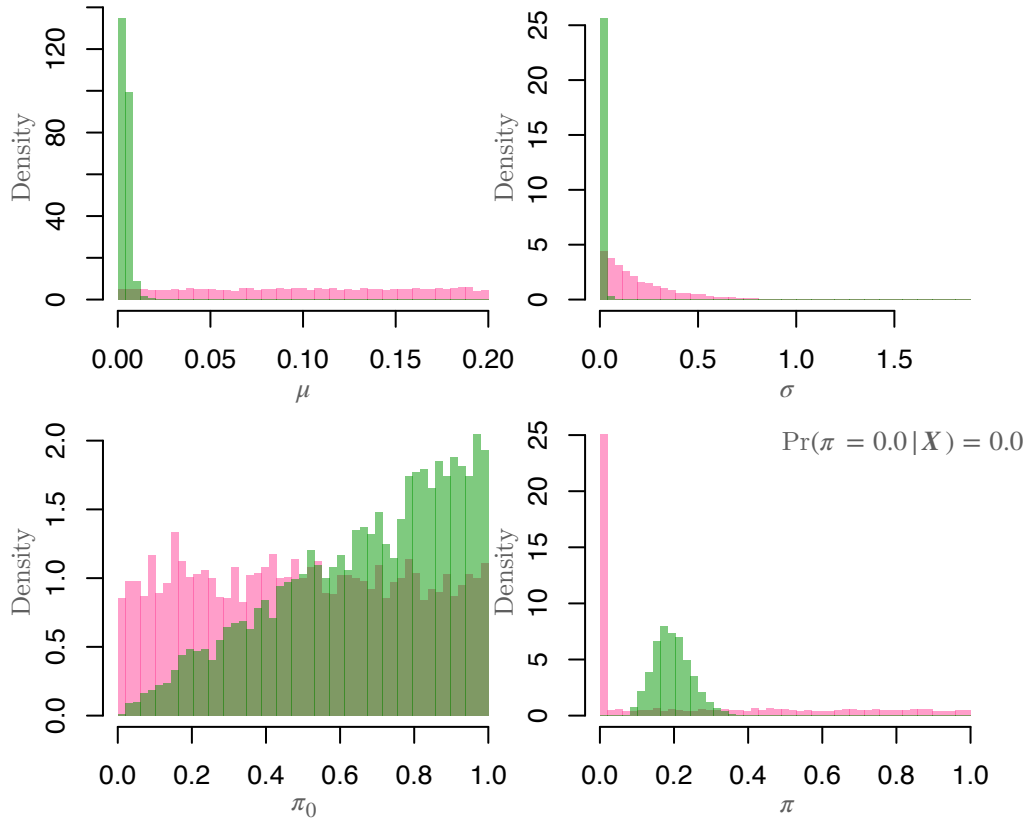


Figure 3.4: **Blood cell varient level model directed acyclic graph**. Bayesian mixture model of variant loads in blood cell data. The observed data of variant load proportions is denoted Y_k, for the k-th single-cell observation of a specific patient and cell type. The mean and standard deviation of the truncated normal fitted to the spike data are μ and σ. The spike proportion is denoted by π and the probability of spike existence, π_0. Prior parameters are denoted by lowercase roman letters in square nodes, and unknown parameters, to be inferred, are Greek letters within circular nodes.

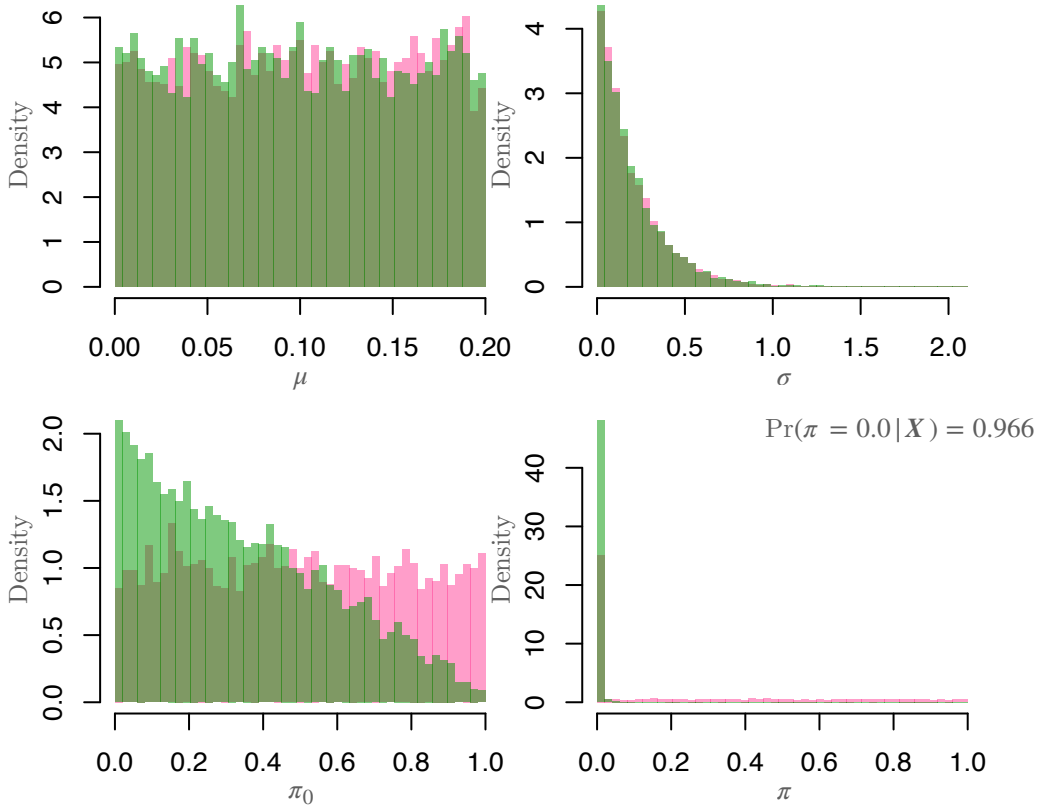
3.4 Results

3.4.1 Model output

First, we inspect the parameter posteriors to check if the parameter beliefs have been updated *a posteriori*. Figures 3.5 show example posterior distributions resulting from fitting the model to variant load data for CD34+ Progenitor cells and CD4+ T memory cells, both patient P22. First, consider the CD34+ progenitor output, the posterior beliefs for π, the wild-type homoplasmy proportion, is weighted mostly at 0.0; in fact, the posterior probability of the proportion being equal to 0.0 is 0.966. This means that the data is overwhelmingly modelled by the second component of the mixture, which is reflected in the posterior beliefs about μ and σ, whose posterior distributions resemble their respective priors. As the data is almost completely modelled by the second component, the model cannot infer the parameters of the first component. The probability of spike existence, π_0, is skewed toward 0.0. Although the beliefs about π_0 have been updated, there is a lot of posterior uncertainty. This is likely due to the hierarchical aspect of the model. The skewed posterior with high uncertainty is typical of $p(\pi_0|Y)$ when fitting the model to this data.



(a) P22 CD34+progenitor



(b) P22 T_{CM}

Figure 3.5: **Posterior proportion of wild-type homoplasmy updates in spike absence**. The prior (pink) and posterior (green) beliefs for the unknown model parameters. After fitting the proposed model to single-cell CD34+ Progenitor variant load data from patient P22.

Next, consider the posterior beliefs after fitting the model to the variant load data for CD4+ naive T cells from patient P22; this dataset shows a distinct population of wild-type homoplasmic cells (see Figure 3.2). In contrast to the posterior beliefs of the CD34+ Progenitor cell data, the beliefs about μ and σ have been updated, greatly reducing the parameter uncertainty. The posterior beliefs about the spike proportion have also been updated. Here, the posterior beliefs about the spike probability, π_0, are skewed towards 1.0 but have high uncertainty, mirroring the posterior of the CD34+ Progenitor cells.

3.4.2 Model fit

Model fit can be visually assessed by inspecting the posterior predictive distributions compared to the observed data. If the model shows a poor model fit, few similarities will be seen between the predicted and observed values. Figure 3.6 shows the posterior predictive distributions for a subset of cell types for patient P22, after integrating out parameter uncertainty. The posterior predictive distributions appear to match the dataset well, accurately identifying the existence and proportion of spikes within the datasets.

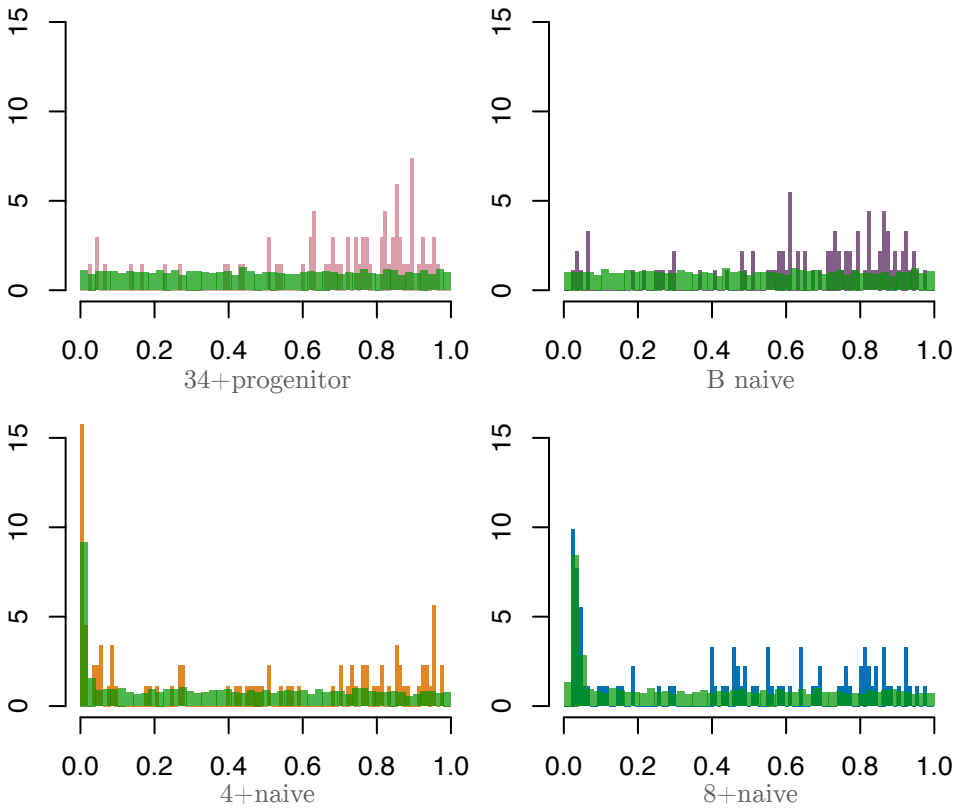


Figure 3.6: **Posterior predictive distribution shows a good resemblance to observed data**. Posterior predictive distribution of variant load (green) compared against the appropriate observed variant load data (not green). The data shown is a subset of cell types for patient P22.

3.4.3 Proportion of cells reaching wild-type homoplasmy

The analysis aimed to find the cell-type-specific proportion of cells which have reached wild-type homoplasmy. For the model used here, this is described by the parameter π.

Therefore, we can inspect its posteriors directly. Figure 3.7 shows the posterior of π for each cell type where data was collected for patient P22. It can be seen that almost no cells have eradicated variant mtDNA in the CD34+ Progenitor cells or monocytes. CD4+ and CD8+ naive T cells show similar levels of wild-type homoplasmy, with expected posterior values of ≈ 0.2. Indeed, there is no substantial difference in their values when inspecting the 95% HDI of their posterior difference. In contrast, naive B cells show a substantially lower wild-type homoplasmy proportion and a posterior expected value of 0.001.

The posterior beliefs show substantial differences between naïve T cells and their more mature counterparts in patient P22. The CD4+ T_{CM} and T_{EM} cells have substantially higher proportions than the CD4+ naive T cells when comparing their posterior differences using 95% HDI. This suggests a negative selection against variant mtDNA during T cell maturation. Within the CD8+ T cells, a substantial difference is found when comparing the T_{EM} and T_{EMRA} cells against the naive cells. No substantial posterior difference is found between CD8+ naive cells and CD8+ T_{CM} cells, as indicated by the posterior 95% HDI. However, the expected proportion of cells reaching wild-type homoplasmy is higher within CD8+ T_{CM} cells.

The increased proportion of wild-type homoplasmy within the T cell compartment indicates a negative selection against the m.3243A>G variant mtDNA during T cell differentiation. The same result is observed for all patients within the dataset, as shown in Figure 3.7. Substantial differences in the proportion of wild-type homoplasmy appear to be less common as patients age. This is likely due to the posteriors being constrained in the range [0.0, 1.0], as they are proportions, leading to skewed distributions with similar expectations (close to 1.0).

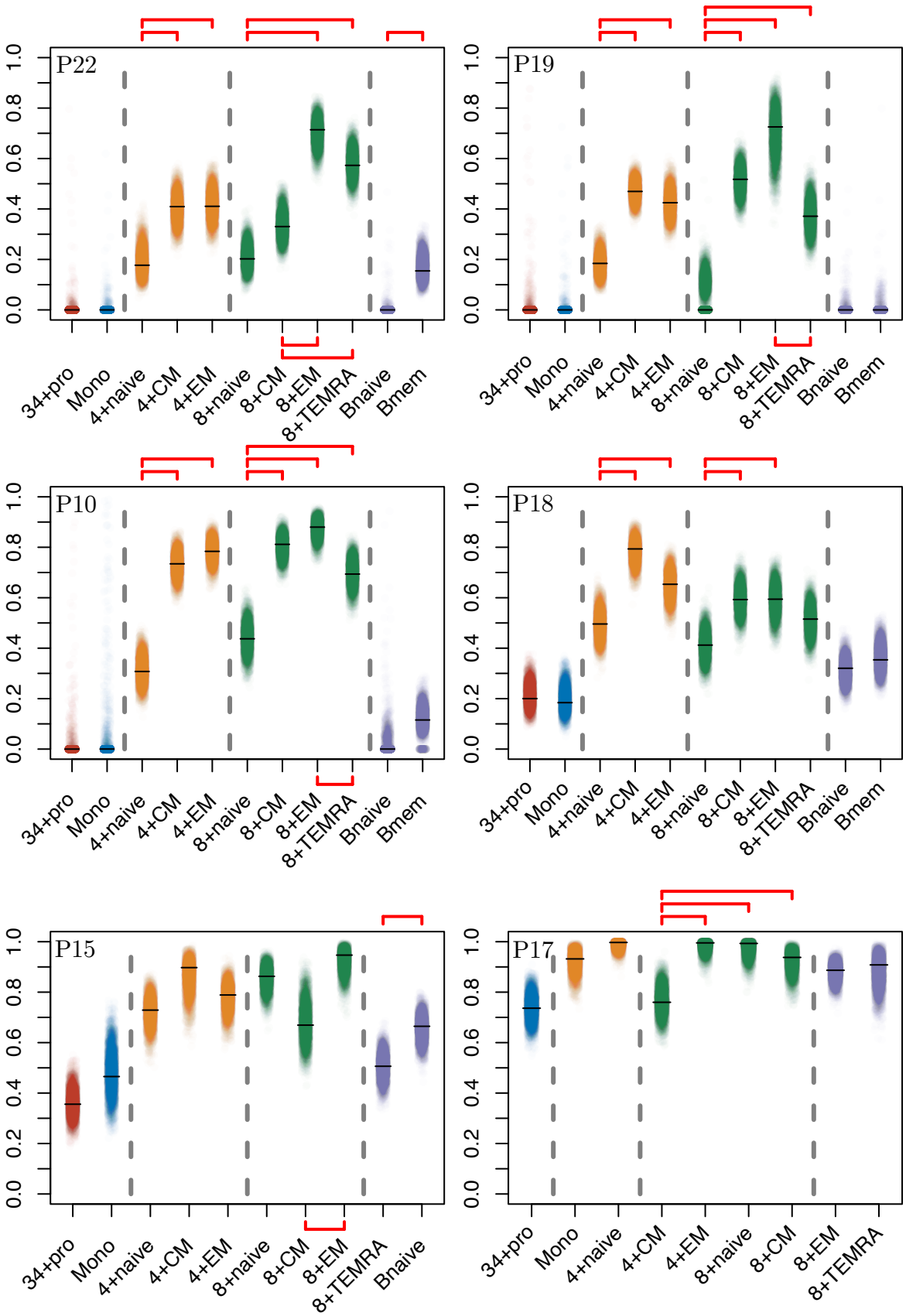


Figure 3.7: **Wild-type homoplasmy is increased within the memory T cells across patients**. Posterior beliefs of the proportion of cells which have reached wild-type homoplasmy (y-axis) across cell-types (x-axis) for all patients. Posterior beliefs are represented by 5,000 (almost) uncorrelated draws from the marginal posterior distribution $p(\pi|\boldsymbol{Y})$. Each cell compartment is uniquely coloured and separated by dashed black lines. Posterior modes are indicated with a solid black line. Red lines indicate a substantial difference between posterior proportions for cells within the same compartment, evaluated using a 95% HDI.

3.5 Discussion

3.5.1 Negative selection against m.3243A>G

The developed model robustly estimates the proportion of blood cells reaching wild-type homoplasmy, coherently accounting for parameter uncertainty and noise in the data. Inspection of the posterior beliefs for wild-type homoplasmy proportion confirms previous findings showing that m.3243A>G reduces over time (de Laat et al., 2012; Langdahl et al., 2018; Rahman et al., 2001; Veitia, 2018; Walker et al., 2020). Previous work suggested that this reduction is driven by the asymmetric cellular division of HSCs in the bone marrow (Rajasimha et al., 2008). Importantly, this analysis shows that negative selection against m.3243A>G is enhanced during T-cell differentiation and maturation.

The generally accepted linear model of T cell differentiation, T_{CM} to T_{EM} to T_{EMRA} (Geginat et al., 2003; Verma et al., 2017), would suggest that T_{EMRA} would have the highest levels of wild-type homoplasmy. However, this is not what is found in this dataset. T_{EMRA} cells are found to have substantially lower m.3243A>G clearance than T_{EM} cells in two patients (P10 and P19) and only one patient, P15, shows a substantially higher clearance in T_{EMRA} compared to T_{EM} cells. The reduction in m.3243A>G clearance among T_{EMRA} cells suggests that these cells may be less mature than previously thought, in agreement with findings from (Rufer et al., 2003). However, findings by Verma et al. (2017) show that two sub-populations of T_{EMRA} cells exist, meaning that the data here is aggregated variant load from *young* and *old* T_{EMRA} cells. The aggregation of cell-type data could be causing the mixed results of T_{EMRA} maturation. In addition, only one patient, P22, shows a substantial increase in m.3243A>G clearance between T_{CM} and T_{EM} cells in their CD8+ T cell population. These findings also suggest that T cell differentiation is non-linear. The non-linear differentiation of memory T cells is supported by T_{CM} and T_{EM} being able to differentiate into effector cells (Youngblood et al., 2017).

In this study, only patients harbouring the m.3243A>G point mutation were selected for single-cell analysis. However, the larger study also showed that the proportions of m.8433A>G variant decrease with time. Unfortunately, single-cell data were not collected from these patients. Therefore, it can not be confirmed whether the enhanced selection against m.3243A>G during memory T cell development is seen in other pathogenic variants. To investigate whether the same enhanced negative selection is seen in other pathogenic variant mtDNA, the model and resulting analysis can be applied to other datasets, with single-cell observations for other pathogenic mtDNA variants.

3.5.2 Mathematical model of B and T cell development

The dataset used in this chapter provides single-cell measurements of mtDNA variant load, providing direct observations of mtDNA population dynamics. The relative ease at which blood can be collected could allow for longitudinal study, with multiple observations within the same patient, albeit not within the same cells. This would allow for an incredibly rich dataset which could be used to develop and fit mathematical models of clonal expansion and test their underlying theories. However, there are a number of obstacles which need to be addressed to accomplish this.

There is a large degree of uncertainty in the parameter values governing a mathematical model. As discussed, mitochondrial function is cell-specific, so any prior parameter inference in other tissues and cell types are not directly comparable. Without extensive prior knowledge, parameter values should be inferred to fairly compare the models (and theories) under consideration.

Building a mathematical model of a complex biological system such as this is not easy, and a number of model aspects must first be considered. The first of which is the biological system. Blood cells have a complex lineage of development and differentiation, which is not fully understood, and the development of these cells would have to be considered. In this dataset, observations are made on a number of blood cells derived from HSCs but mostly within the B and T cell compartments. Given the available data, building a mathematical model of the entire HSC lineage is not practical and likely unnecessary for an investigation into clonal expansion. A model of the B and T cell compartments is more realistic, starting with naïve cells and modelling their development into long-lived memory cells (the primary source of data in this project). In the rest of this chapter, key components of the mathematical model of B and T cell development are discussed.

Infection

It is important to note that an infection initiates B and T cell development, and thus, an infection mechanism is a crucial part of the mathematical model. Infection could be considered to occur at discrete or continuous time intervals and be fixed or stochastic. Both mechanisms would require assumptions about the frequency with which the immune response is triggered.

In response to an infection, the appropriate antigen-specific naïve cell begins a period of high proliferation. In the mathematical model, how this cell is selected must be decided. Unless a more appropriate mechanism is proposed, randomly sampling from the naïve cell population seems reasonable. The possibility of reinfection should also be considered. Reinfection causes the long-lived memory and effector cells (not naïve) to begin proliferation in response to the infection and thus will further drive the negative selection against m.3243A>G seen in long-lived memory cells. The long-lived memory cells should, therefore, be labelled, allowing them to be appropriately sampled in the event of reinfection.

Naïve cell population

As discussed in Chapter 3.1, variant load in blood is negatively correlated with time. A mathematical model of cell development from naïve cell to long-lived memory cells would require special consideration of naïve cell variant load. Simulating the mtDNA population dynamics of naïve cells would not be appropriate as the cells arise from a series of cellular developments within the HSC lineage, and, as discussed, modelling the whole lineage is not practicable. Therefore, a model of naïve cell variant load must be chosen. This could be deterministic or stochastic, but would have to reflect the decrease in time seen in patients. Grady et al. (2018) developed a formula to estimate the age-adjusted (inherited) variant load for m.3243A>G patients, although they did not have access to

cell-type-specific data and used blood-cell aggregate variant load data to develop their model. Nevertheless, the method could be used to estimate not only the inherited variant load but also the variant load throughout a patient's life.

Cell development

Upon activation, a naïve T cell undergoes asymmetric cellular division, producing one cell primed to be an effector T cell and one primed to be a memory T cell (Chang et al., 2007; Pollizzi et al., 2016; Verbist et al., 2016). From here, the two primed cells undergo rapid proliferation and become a heterogenous population of activated memory and effector cells. As mentioned, T cell activation is dependent on OXPHOS, assuming the threshold effect, the cellular variant loads must be considered during this proliferative period and cells whose variant load is above a threshold must be degraded.

Rajasimha et al. (2008) developed two mathematical models, independently modelling stem cells (HSCs) and high proliferative cells (T and B cells) and showed that division of stem cells with a variant-load threshold, above which the cell dies, drives the removal of m.3243A>G. However, they did not consider the development and differentiation of B and T cells, instead regarding them as short-lived and highly proliferative. In their model, HSCs differentiated with a defined probability, allowing for both symmetric and asymmetric division. Additionally, Moeller et al. (2024) developed a mathematical model of the HSC population throughout periods of proliferation and population stability. Their model allowed HSCs to differentiate both symmetrically and asymmetrically, with defined rates. Borsa et al. (2019) later showed that memory T cells are also able to asymmetrically divide, possibly explaining the enhanced negative selection seen in the T cell compartment. Although not a confirmed mechanism, random differentiation during cell division would appear to be an appropriate starting point for the mathematical mechanism of memory T cell differentiation. Indeed, if the model accurately predicts the observed data, then this would provide further evidence that random differentiation is the biological mechanism.

Assuming a model in which memory T cell differentiation occurs through asymmetric cellular division, the possible daughter cells for each type must be considered. The non-linearity of T cell maturation shown in the dataset analysed in this chapter would suggest that a differentiation pattern of T_{CM}-T_{EM}-T_{EMRA} is not appropriate. If no appropriate mechanism is suggested, then the probability of asymmetric division between each cell type can be inferred. The increased uncertainty in the model and additional model parameters are likely to add to the difficulties of parameter inference. However, if the probabilities were estimated, the results could be very interesting and have significant implications for the biological mechanisms of T cell maturation.

MtDNA dynamics

In addition to modelling the cell-level differentiation of naïve to long-lived memory cells, the mathematical model would also have to consider the mtDNA dynamics within each individual cell. MtDNA population dynamics can be modelled using the methods discussed in Chapter 2.6. The highly proliferative nature of B and T cells during infection provides two problems: copy number regulation and mtDNA segregation during cell division. Both aspects can have a large impact on the cellular variant loads. However,

their biological mechanisms are not clear, and mathematical modelling of mtDNA population dynamics has largely ignored dividing cells, likely due to the additional complexities associated with them. However, Johnston et al. (2015) investigated mechanisms associated with the genetic bottleneck through mathematical models of oocyte development. They concluded that, given their available data, binomial partitioning of mtDNA between daughter cells is the most likely segregation mechanism when compared to clustered partitioning and deterministic partitioning. As part of their model, cells were considered to be in a number of phases throughout their development, each with its own replication and degradation rates. Rajasimha et al. (2008) assumed cellular division randomly segregates mtDNA between two daughter cells, approximately halving copy number relative to a target value and allowed copy number to rise post-division. Following these works, it would be sensible for mtDNA segregation to be binomial and a cell to be in two possible states: *recently divided* and *maintenance*. Where recently divided cells have an increased replication rate, allowing copy number to be replenished, although this mechanism has not been experimentally confirmed. Indeed, it may be possible to infer the mechanism within a mathematical model, given an appropriate dataset of cellular variant load during this period of high cellular proliferation. Although it is unlikely to be inferred from the current data of long-lived cells, where mtDNA dynamics during cellular division are likely obscured by the long period of mtDNA dynamics post-division.

The mathematical model must implement a copy number control mechanism to prevent unrealistic copy numbers or mtDNA extinction. Copy number also affects mtDNA dynamics, with increased copy numbers slowing the clonal expansion of variant mtDNA (Elson et al., 2001). As part of the larger study, copy number measurements were taken within multiple cell types within the blood cell lineage from both patients and control subjects (Franklin et al., 2023). Copy number was found to have a weak positive correlation with aggregate m.3243A>G variant load (from multiple cell types) in patients and have a stronger negative correlation with age. No difference was found between the copy numbers of varying cell types in patients or control subjects. Mathematical models which reflect the increased copy number have been proposed (Capps et al., 2003; Hoitzing et al., 2019; Insalata et al., 2022). Capps et al. (2003) suggested a series of nuclear-controlled mtDNA replication models, which allowed copy number to increase with the increase in variant load. Hoitzing et al. (2019) investigated the energetic requirements of cells and how they can most efficiently use their resources. Although, their model is likely too complex for our purposes, requiring information on cellular energy and resources. Insalata et al. (2022) focused on modelling the spread of variant mtDNA variant load in skeletal muscle fibres. However, their model would not be appropriate for blood cells, given the very different structure and dynamics of blood cells compared to skeletal muscle. Another popular mechanism, although one that does not naturally give rise to a positive correlation between copy number and variant load, is the linear feedback mechanism, which decreases the mtDNA replication rate by an amount proportional to the difference between the current copy number and a target copy number (Aryaman et al., 2019; Hoitzing et al., 2017; Insalata et al., 2022). The target copy number is often fixed, but given the relationship between copy number and age found by Franklin et al. (2023), and the lack of appropriate biological or mathematical mechanisms which recreate this, the target copy number could be a function of age, following the linear model in their work.

Model summary

Here, the ideas discussed so far are pulled together. Figure 3.8 shows the simulation of a single infection, starting with naïve cells, drawing their initial variant load and copy number from appropriate distributions, and their development into the differentiating memory T cells. The diagram ignores the simulation of mtDNA populations within single cells, as these are described in detail in Chapters 2.6.2 and 5, and would change depending on the theory of clonal expansion in place. A general simulation algorithm for a single patient for a length of time, $T_{\max}$ is described in Algorithm 5. Similarly to Figure 3.8, and for the same reason, the algorithm ignores the simulation of mtDNA population dynamics within a single cell. In the algorithm, the time taken for the immune response to finish is denoted $t_{\text{infection}}$, and is assumed constant. The algorithm has been further simplified by not allowing reinfections.

Algorithm 5 Model of B and T cell development (no reinfections)

1. Set the system time $t = 0$.
2. Simulate the time until the next infection

$$t' \sim [\text{inter-infection-time distribution}].$$

3. Update the population of long-lived memory cells. Simulate each long-lived memory cell between $(t, t + t' + t_{\text{infection}}]$, including their mtDNA populations and possible differentiations to other cell types
4. Simulate infection
   - i Calculate the variant load and copy number distributions for B and T naïve cells variant load, at time $t + t'$
   - ii Sample the antigen-specific naïve B and T cells, with the appropriate variant load and copy number
   - iii Simulate the development of the chosen naïve cells and their differentiation into long-lived memory B cells, T_{CM}, T_{EM}, and T_{EMRA} cells over the time interval $(t + t', t + t' + t_{\text{infection}}]$
   - iv Add newly differentiated long-lived memory cells to their respective populations
5. Put $t := t + t' + t_{\text{infection}}$
6. If $t < T_{\max}$, return to Step 2.

Parameter inference

The mathematical model of blood cells, as described, could be created if given some further consideration. However, due to the complexities of the biological system, the model has become considerably computationally expensive. The cost of simulating the mtDNA population dynamics of a single cell is relatively low, but with the large number of cells which are part of the system, combined with the additional complexities of cellular divisions, both symmetric and asymmetric, the model is likely to be very computationally expensive. Parameter inference would require many thousands or millions of simulations of the system and, unfortunately, is key to comparing theories of clonal expansion. Previous work implemented Bayesian inference to infer parameters for mathematical models of

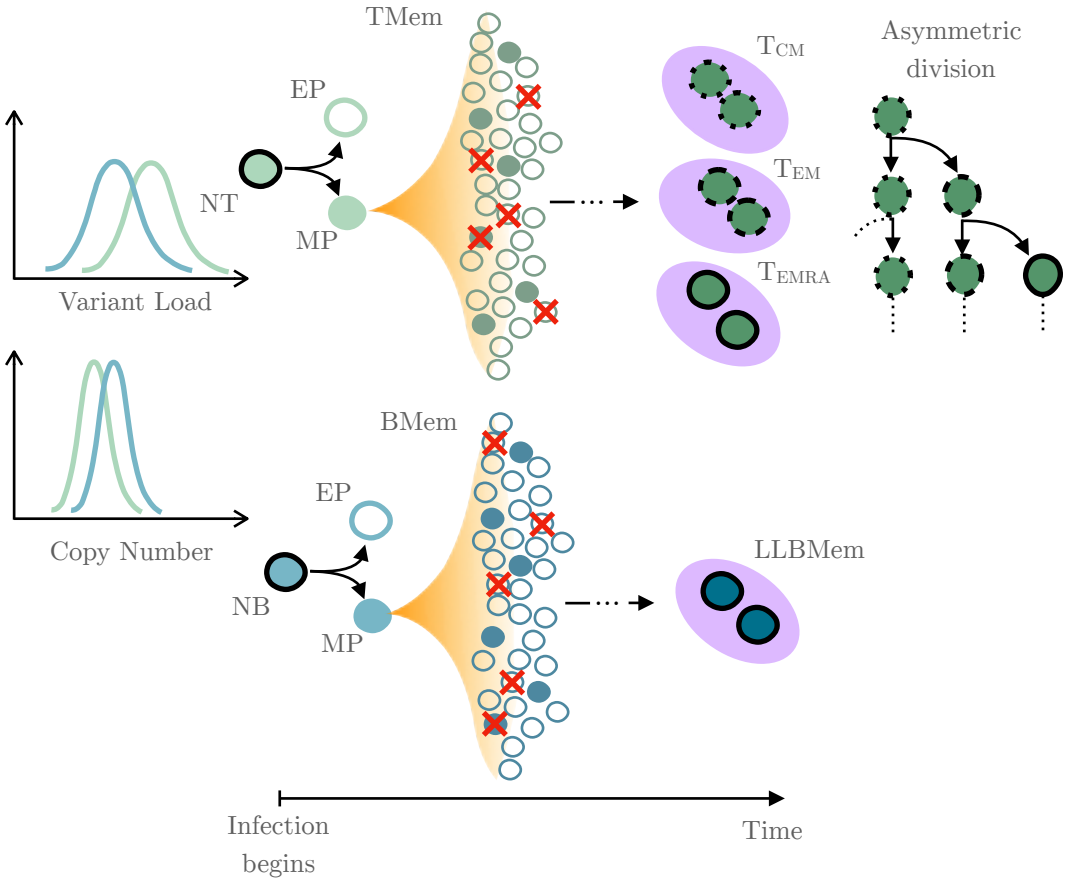


Figure 3.8: **Mathematical model of B and T cell development**. A diagram showing a mathematical model of the B and T cell compartments based on the single-cell data used in Chapter 3. The initial variant load and mtDNA copy number of the B (BN) and T (TN) naïve cell are taken from a calculated distribution based on the age at the time of the simulated infection. The naïve cells divide asymmetrically into effector (EP) and memory (MP) progenitors before proliferating and becoming active, differentiating into memory B (BMem) and T (TMem) cells. During activation, cells which show signs of OXPHOS deficiency are removed from the population, indicated by a red cross. Lastly, a subset of the active BMem and TMem differentiates into long-lived cells. Active B memory cells differentiate into long-lived B memory cells (LLBMem), and active memory T cells differentiate into three long-lived memory cells: T_{CM}, T_{EM}, and T_{EMRA}. Assuming linear differentiation between cell types with T_{CM} being the least mature and T_{EMRA} being the most, the cells undergo random asymmetric cell division. Cells with black outlines indicate cell types with observed data in the current dataset. Long-lived cells of the same type are also highlighted by purple groupings.

mtDNA population dynamics and used model emulators to reduce the computational cost of inference (Ainsworth, 2014; Henderson et al., 2009). However, emulation still requires a large number of simulations to build the emulator (Henderson et al., 2009). Furthermore, parallelising single-cell simulations is not possible due to the time-dependent nature of population dynamics and cell differentiation, making it challenging to reduce the computational cost.

3.5.3 Concluding remarks

The dataset used here provides a lot of information regarding the population dynamics of mtDNA and could provide an excellent basis for comparing theories of clonal expansion. However, as mentioned, a significant number of hurdles must be overcome. The majority of these are the choice of mechanisms to reflect biological behaviour. However, if experts could agree upon reasonable assumptions about these, then the development of the mathematical model would be feasible, and the model could be used to investigate clonal expansion. Bayesian inference schemes could be used to infer model parameters and compare theories to the observed dataset.

Conversely, although not the aim of this thesis, mathematical models could be used to reflect theories of the biological mechanism of memory T cell differentiation. By constructing a number of mathematical mechanisms which reflect different biological mechanism theories, and comparing their simulated output to the observed data. Again, this could be achieved by Bayesian inference, following work by Johnston *et al.*, who used Bayesian inference to compare theories of the genetic bottleneck (Johnston et al., 2015). In theory, this could be combined with a clonal expansion investigation into one large inference problem. However, given the large amounts of uncertainty in all areas, this would likely be a very difficult task.

Unfortunately, the work required to both develop the mathematical model and infer parameter values, is likely too high in a reasonable time frame of this thesis given the number of uncertainties in; the biological mechanisms, the mathematical modelling mechanisms to reflect these, and the lack of prior knowledge in parameter values, and the computational expense.

Chapter 4

Classification of myofibre OXPHOS status

The difficulties associated with developing a mathematical model of clonal expansion in blood cells were largely associated with the complexity of the biological system. Therefore, a simpler system is sought. Non-dividing and terminally differentiated cells would provide a significant reduction in biological (and modelling) complexity, entirely removing uncertainties in cellular development and replication, which are part of the blood cell system. Skeletal muscle fibres (myofibres) are one such example. Throughout this chapter, the term myofibre is used as the data under consideration was collected from myofibre samples. However, the work presented here is not restricted to myofibres and can be applied to many other cell types.

4.1 Introduction

In this chapter, Bayesian methods are used to infer measurements of mtDNA population dynamics from OXPHOS protein abundance data in myofibres. Mitochondrial dysfunction can be identified by inspecting the level of OXPHOS proteins within a single myofibre. Proteins required for ATP production are tightly regulated within a functioning cell, maintaining an appropriate level for myofibre size and ATP needs; a discrepancy from this indicates that the mitochondria cannot produce the required level of ATP. Mitochondrial dysfunction usually presents within myofibres by under-representing OXPHOS proteins (Ahmed et al., 2017; Hellebrekers et al., 2019), although over-representation can also occur (Warren et al., 2020). Assuming the biochemical threshold theory (Rossignol et al., 2003), mitochondrial function is disrupted when the proportion of pathogenic mtDNA passes the pathogenic threshold. The identification of myofibres which show a reduction in OXPHOS proteins is, therefore, equivalent to cells whose variant load is greater than the pathogenic threshold (Bua et al., 2006; Campbell et al., 2014).

Previously, two methods have been used in the literature to classify the OXPHOS status of single myofibres using their protein abundances. Both methods inspected single-myofibre OXPHOS abundances in relation to mitochondrial mass. OXPHOS protein abundances were inspected by their relation to mitochondrial mass, to standardise the protein abundances within the myofibres. However, direct measurements of mitochondrial mass are difficult to make, and, therefore, it was assumed that mitochondrial mass is pro-

portional to the abundance of specific proteins found on the OMM. Therefore, reference to mitochondrial mass within this chapter refers to the abundance of one such protein, VDAC1, which is used within both observed datasets in this chapter.

4.1.1 Previous work

Frequentist linear model

Rocha et al. (2015) used immunofluorescence (IF) to measure NDUFB8 (CI) and MTCO1 (CIV) abundances in myofibres and considered how a patient's abundance differed from a healthy control. They noted that protein abundances increased linearly with mitochondrial mass on the log-scale within control subjects, and considered patient myofibres which show the same relationship to be healthy. A classification pipeline was constructed by fitting a frequentist linear regression model to the logged OXPHOS protein abundance from healthy control subjects, using a single explanatory variable, the logged measure of mitochondrial mass. Patient myofibres were classified as having mitochondrial dysfunction if their (logged) protein abundance lies outside of the 95% predictive interval of the linear model. The model and classification are described in Equations 4.1 and 4.2, where logged OXPHOS protein abundance and logged mitochondrial mass of the j-th control myofibre be denoted $\{X_j^{\mathrm{C}}, Y_j^{\mathrm{C}}\}$,

$$\begin{aligned} Y_j^{\mathrm{C}} &= mX_j^{\mathrm{C}} + c + \varepsilon_j, \\ \varepsilon_j &\sim \mathrm{N}\left(0, \sigma^2\right). \end{aligned} \tag{4.1}$$

Where m and c denote the slope and intercept of the linear regression, and σ denotes the model error.

Rocha *et al.* go further, classifying patient myofibres into four levels of deficiency based upon a patient myofibre's Z-score, the vertical distance between the observed protein abundance and the expected protein abundance from the linear model fitted to the control subject data. Later, Warren et al. (2020) showed that most patient protein abundances are naturally divided into two groups: not-like-control and like-control. The binary classification is also consistent with the biochemical threshold theory of clonal expansions (Rossignol et al., 2003), where a mitochondrial dysfunction occurs only after the variant load has passed the pathogenic threshold. In addition, Warren *et al.* noted that some not-like-control patient myofibres showed an increased OXPHOS protein abundance compared to the control data and, therefore, did not implement the Z-score classification of Rocha *et al.*. Instead, Warren *et al.* used a ternary classification, by first fitting the linear model to the control subject's abundances and classifying patient myofibres lying outside the 95% predictive interval as not-like-control, before sub-categorising them as under- or over-expressed.

Following the binary classification implemented by Warren *et al.*, and combining under- and over-expressed myofibres into a single not-like-control group. Let $\{X_j^{\mathrm{P}}, X_j^{\mathrm{P}}\}$ be the logged measure of mitochondrial mass and OXPHOS protein abundance of the j-th patient myofibre, it is classified as like-control if it lies within the 95% predictive interval of the frequentist linear mode fitted to the aggregated control data. Otherwise, it is classified as not-like-control. Let Z_j be the classification variable, then

$$Z_j = \begin{cases} 1 & Y_j^{\mathrm{P}} \notin [L_j, U_j], \\ 0 & \text{otherwise} \end{cases}. \tag{4.2}$$

The bounds of the 95% predictive interval at X_j^{P} are denoted L_j and U_j, and the j-th myofibre is considered not-like-control if $Z_j = 1$.

The classification pipeline proposed by Rocha *et al.* has become widespread in the literature, (Baty et al., 2021; Hellebrekers et al., 2019; Ng et al., 2020; Sithamparanathan et al., 2018). However, the model's rigidity may misclassify a large number of myofibres within a patient sample. To be classified as like-control, a patient's protein abundance and mitochondrial mass must resemble the relationship seen in the control data. The natural variation between subjects means this is often not the case.

Gaussian mixture model

Another classification method was proposed by Vincent et al. (2024), who classified patient myofibres as being OXPHOS deficient using a two-component GMM, independent of the control subject's data, inferring model parameters in a frequentist setting. The resulting clusters were then compared to the control data to decide which cluster was like-control and which was not. The model was able to correctly identify deficiency in two of the three OXPHOS proteins within the dataset: NDUFB8 and MTCO1. However, it failed to correctly classify the CYB (a subunit of CIII) status of all myofibres due to some myofibres over-expressing the protein and, thus, distorting the not-like-control group. To the author's knowledge, this method has only been used within this work and has not been replicated elsewhere. Due to its limited use and varied results, it will not be replicated for comparison within this chapter and, instead, the focus is given to the more widely used frequentist linear model described previously.

4.1.2 General aim

The identification of not-like-control patient myofibres is critical to establishing disease severity and estimating disease progression. In this chapter, we show that the frequentist linear model, proposed by Rocha et al. (2015), is not able to accurately classify OXPHOS status of patients whose OXPHOS abundance differs from controls. The aim of this chapter is, therefore, to develop a model which can better cope with the natural inter-subject variability.

4.2 Data

Within this chapter, two observed datasets are used to test the proposed model's classification. Another synthetic dataset is also used, which is discussed in Chapter 4.3.4. In this chapter, two observed datasets are introduced and analysed using the frequentist linear model.

4.2.1 Vincent dataset

Data collection

Vincent et al. (2024) collected skeletal muscle samples from 12 patients, each having nuclear-encoded mitochondrial disease affecting mtDNA replication and maintenance. The disease causes mtDNA mutation events to occur at a much higher frequency, and consequently, variant mtDNA clonally expands at a faster rate compared to healthy subjects. Skeletal muscle samples were taken from the hamstrings of four (healthy) control subjects while undergoing anterior cruciate ligament surgery. The tissue was cut into 6µm thick cross-sections and assessed by imaging mass cytometry (imaging mass cytometry (IMC)), generating a pseudo-image of myofibres where colour intensities indicate protein abundances of specific proteins. The image is segmented to find single-myofibre protein abundances, using the cell membrane marker DMD and the Mitocyto segmentation tool (Warren et al., 2020). Single-myofibre protein abundances were calculated as the average protein abundance within a segmented single-myofibre of the pseudo image. In this dataset, VDAC, a protein found on the OMM, is assumed to be proportional to mitochondrial mass.

Tissue collection and data analysis were conducted in accordance with protocols approved by the Ethical Committee of Martin Luther University Halle-Wittenberg. Subjects in the study provided written informed consent prior to their participation. Two subjects, P07 and P08, were investigated with informed consent by the Newcastle Tyneside Local Research Ethics committees (REC ref. 2002/205). Control tissue samples were also collected with informed consent from patients undergoing cruciate ligament surgery, approved by Newcastle and North Tyneside Local Research Ethics Committees (RED ref. 12/NE/0395).

Protein abundance data was collected on various proteins in the experiment. However, this work's analysis is restricted to the three OXPHOS proteins: CYB, NDUFB8, and MTCO1. The latter two proteins, NDUFB8 and MTCO1, are the most commonly used OXPHOS subunits used to assess mitochondrial function within the literature (Ahmed et al., 2017; Lehmann et al., 2019; Rocha et al., 2015; Rocha et al., 2018). A detailed description of the data collection process and methods can be found in Vincent et al. (2024). The IMC process failed for P08 in the dataset, so the patient was removed from all analyses, leaving a total of 11 patients within the dataset to be analysed.

For this dataset, manual classification of the OXPHOS status of patient myofibres were made to use a benchmark, against which model performance can be evaluated. Three experts inspected 2Dmito plots, visualisations of patient and control protein abundances defined below, see Figures 4.2 and 4.3, and selected myofibres believed to be OXPHOS deficient. Unfortunately, there can be disagreement between experts on the overlap between myofibre populations. Here, only myofibres which had total agreement from the panel were considered OXPHOS deficient and classified as not-like-control.

Data exploration

Rocha et al. (2015) log-transformed their data so that the control subjects' protein abundances followed the assumptions of a linear model. The untransformed protein abundances

by Vincent et al. already show these assumptions, and they are not significantly improved or worsened after log transformation. Nevertheless, the Box-Cox power transformation is used on the protein abundances from each control subject independently. No transformation is suggested for C02 or C03 for NDUFB8 and CYB abundances and a transformation of approximately $x :\to x^{1.5}$ is suggested for MTCO1. In contrast, no transformation is suggested for MTCO1 in control subject C04, whereas a transformation of $x :\to x^{0.75}$ is suggested for NDUFB8 and CYB. Given the little difference made to this dataset and the reduction in non-normality in other datasets (Rocha et al., 2015; Warren et al., 2020), the protein abundances (including VDAC) are logged prior to analysis.

OXPHOS protein abundance data is commonly presented in the form of a 2Dmito plot. This is a scatter plot of protein abundances, with mitochondrial mass on the x-axis and OXPHOS protein abundance on the y-axis. Within this thesis, a 2Dmito plot is largely constrained to contain data from a single patient and a number of control subjects. However, 2Dmito plots containing only control subject abundances are used during initial data exploration, see Figure 4.1. By inspection of Figure 4.1, a strong linear relationship is seen in the control subjects.

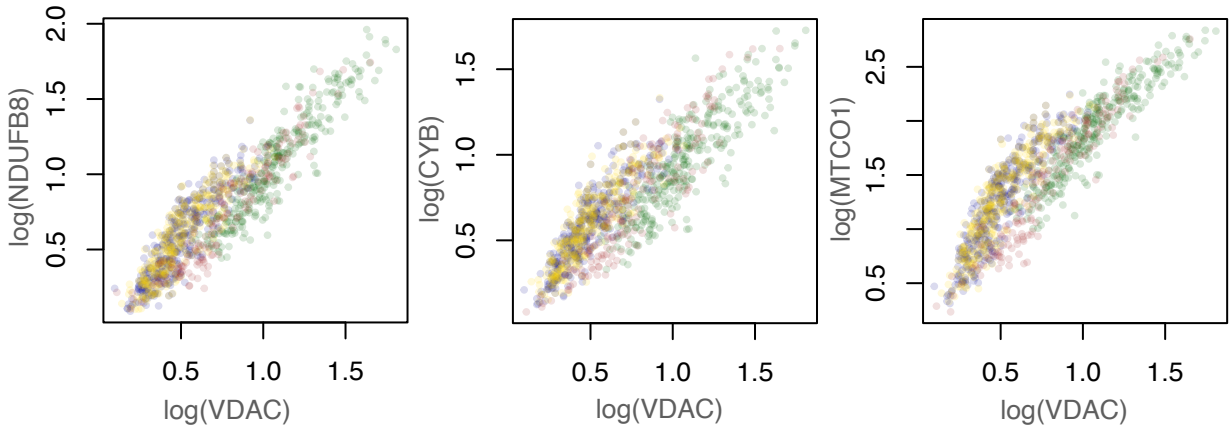


Figure 4.1: **Control subject protein abundances, Vincent *et al.***. Single-myofibre OXPHOS protein abundances collected by IMC from four healthy control subjects: C01 (green), C02 (blue), C03 (yellow) and C04 (red), with 300, 363, 338, and 154 single-myofibre observations, respectively.

Figure 4.2 shows example 2Dmito plots for the three OXPHOS proteins in patient P09, chosen due to the variety of abundance profiles seen across the proteins. P09 shows a population of myofibres whose OXPHOS protein abundances and mitochondrial mass have a similar relationship to those seen in the control subjects. Experts believe that the population of patient myofibres showing a linear relationship, lying close to the control data, is healthy. The remaining myofibres, which lie below the control data and show little to no relationship, are assumed to be OXPHOS deficient. Following Warren *et al.*, the two populations are referred to as like-control and not-like-control, respectively. The protein abundances in P09 highlight that patients do not need the same relationship between proteins to have distinct populations of like-control and not-like-control myofibres.

A clear linear relationship exists between (log) OXPHOS protein abundance and (log) VDAC abundance in control subjects and like-control patient myofibres. Therefore, it

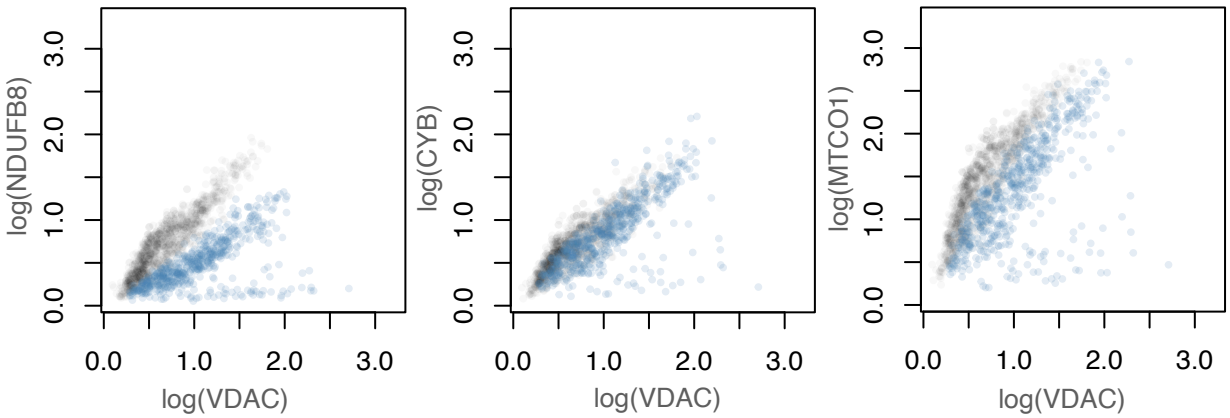
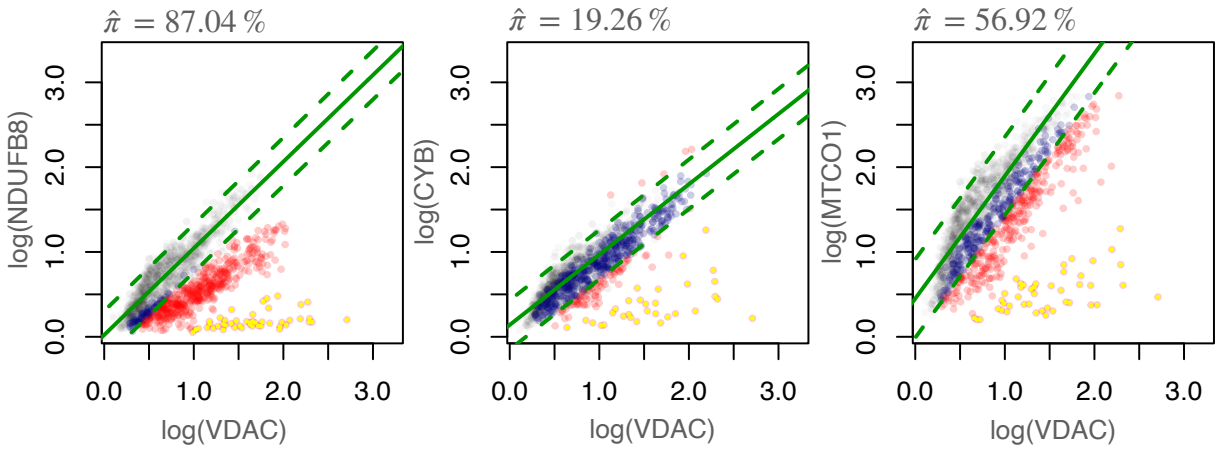


Figure 4.2: **Patients single-myofibre OXPHOS protein abundances split into two populations, Vincent dataset**. Single-fibre protein abundances were collected by Imaging Mass Cytometry (IMC) from skeletal muscle myofibres from four healthy control subjects (grey, 1,155 myofibres) and one patient, P09, (blue, 571 myofibres).

is sensible first to fit the frequentist linear model classification pipeline (Rocha et al., 2015; Warren et al., 2020). Figure 4.3 shows the 2Dmito plots, with classifications by the frequentist model, for all OXPHOS proteins in patient P09. Using the expert manual classifications as a benchmark, also seen in Figure 4.3, the classifications are varied in quality. The classified CYB status of most myofibres is consistent with the manual classification, but a few patient myofibres, lying adjacent to the controls, have been misclassified. Classification of myofibres in both MTCO1 and NDUFB8 status has failed. For MTCO1, the population of like-control patient myofibres has been split in two, resulting in 281 misclassified myofibres. The NDUFB8 status of the majority of like-control patient myofibres has been incorrectly classified. The like-control patient myofibres have shifted from the controls. As a result, they have been almost entirely missed by the linear model's predictive interval. Compared to the manual classification, 14% - 79% of patient myofibres were misclassified across the three OXPHOS proteins, as highlighted in the confusion matrices, also in Figure 4.3. All misclassifications were false positives, resulting in an overestimation of the proportion of deficient myofibres.



	NDUFB8	Manual	
		0	1
Freq	0	74	0
	1	451	46

	CYB	Manual	
		0	1
	0	461	0
	1	79	31

	MTCO1	Manual	
		0	1
	0	246	0
	1	281	44

Figure 4.3: **Frequentist classification arbitrarily splits healthy myofibre populations**. Frequentist model's 95% predictive interval and classifications for all three OXPHOS proteins in P09 with 571 myofibres (coloured points). Control myofibres are shown in grey (1,155 from four healthy subjects). Patient myofibres are blue or red, depending on whether the model classified them as like-control or not-like-control, respectively. The 95% predictive interval and fitted values for the model are shown in green. The manually classified not-like-control myofibres are shown with a small yellow dot within the myofibre's data point. The confusion matrices for the classification compare the frequentist linear model and manual classifications, following Eq. 4.1 and 4.2 where a myofibre classification of 1 indicates not-like-control.

4.2.2 Gomes dataset

Data collection

Gomes et al. (2025) also collected OXPHOS abundance data in skeletal muscle, their study aimed to investigate the within-patient variability of OXPHOS protein abundance and OXPHOS deficiency. Multiple post-mortem tissue samples were collected from three patients, with m.3243A>G related mitochondrial disease, within two tissue types: quadricep (QD) and Tibialis Anterior (TA). The tissue samples were collected under Newcastle Brain Tissue Resource ethics (IRAS 255808) by application 2021031. Three control tissues were collected for each patient, provided by the Newcastle Mitochondrial Research Biobank, and used under their ethics approval (REC Ref: 16/NA/0267) by application MRBOC ID043.

Tissue samples were analysed by quardruple immunofluorescence (QIF) (Rocha et al., 2015), and, similarly to the Vincent *et al.* dataset, single-myofibre abundances are considered to be the mean of the protein abundance within a single segmented myofibre. Myofibre segmentation was conducted by Quadruple Immuno Analyser and manually corrected (Rocha et al., 2015). Protein abundance data was collected for NDUFB8, MTCO1, VDAC and cell membrane marker LAMA1, used to myofibre segmentation in the pseudo image. Tissue samples (blocks) were cut into three adjacent sections, assumed to contain

the same myofibres and, therefore, the same proportion of OXPHOS deficiency. A single myofibre cross-section is labelled as follows: PX-XX-BX-SX, indicating the patient, tissue type, block identifier and slice, respectively.

Data exploration

Similarly to the Vincent *et al.* dataset, the protein abundances of healthy myofibres in control subjects show a fairly strong linear relationship on the log scale, Figure 4.4. The Box-Cox power transformation test suggests that the control data be log-transformed or remain at their raw values. For the reasons described previously, the protein abundances are log-transformed.

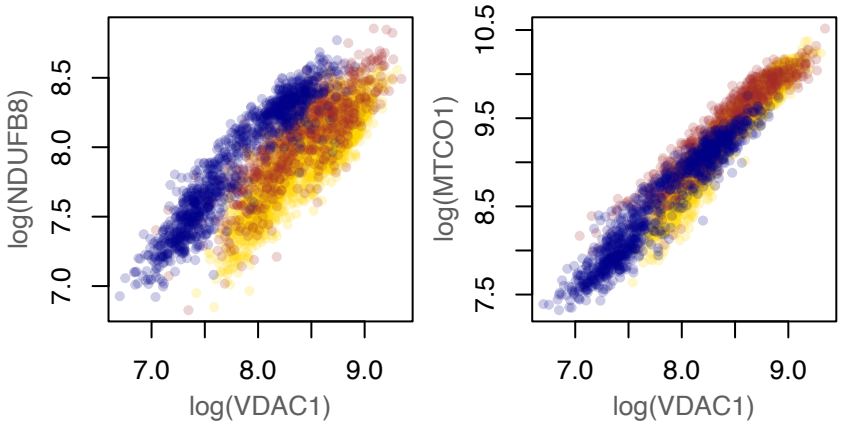


Figure 4.4: **Control subject protein abundances show linear relationship, Gomes dataset**. Single-myofibre protein abundances were collected by QIF from three healthy control subjects: C01 (yellow), C02 (red), and C03 (blue), the samples have 2537, 719, and 1031 single-myofibre observations, respectively.

Figure 4.5(a) shows the 2Dmito plots of three serial sections of TA tissue from patient P01 in the Gomes dataset. Similarly to the patient abundances shown in Figure 4.2, a population of myofibres with protein abundances resembling the controls can be seen, as well as a population of myofibres below these. Interestingly, here, the like-control patient NDUFB8 abundances of section S3 are split into two disjoint sub-populations, both showing a similar relationship to the control data. The same phenomenon is also observed in protein abundances in section S1, albeit to a lesser degree.

Patient P03 in the Gomes dataset possesses OXPHOS abundances which appear different from those seen in P01, P02, or the Vincent dataset, where the distinction between patient and control abundances is less pronounced, Figure 4.5(b). This could be due to a higher proportion of not-like-control myofibers, which increases the presence of this population and diminishes the separation between myofibre populations. In contrast to Figures 4.2 and 4.5(a), the like-control myofibre population is considerably less clear, and it is not obvious, from visual inspection, which myofibres are like-control and which are not. Interestingly, patient P03 had the longest time from death until tissue collection, which may have affected tissue quality.

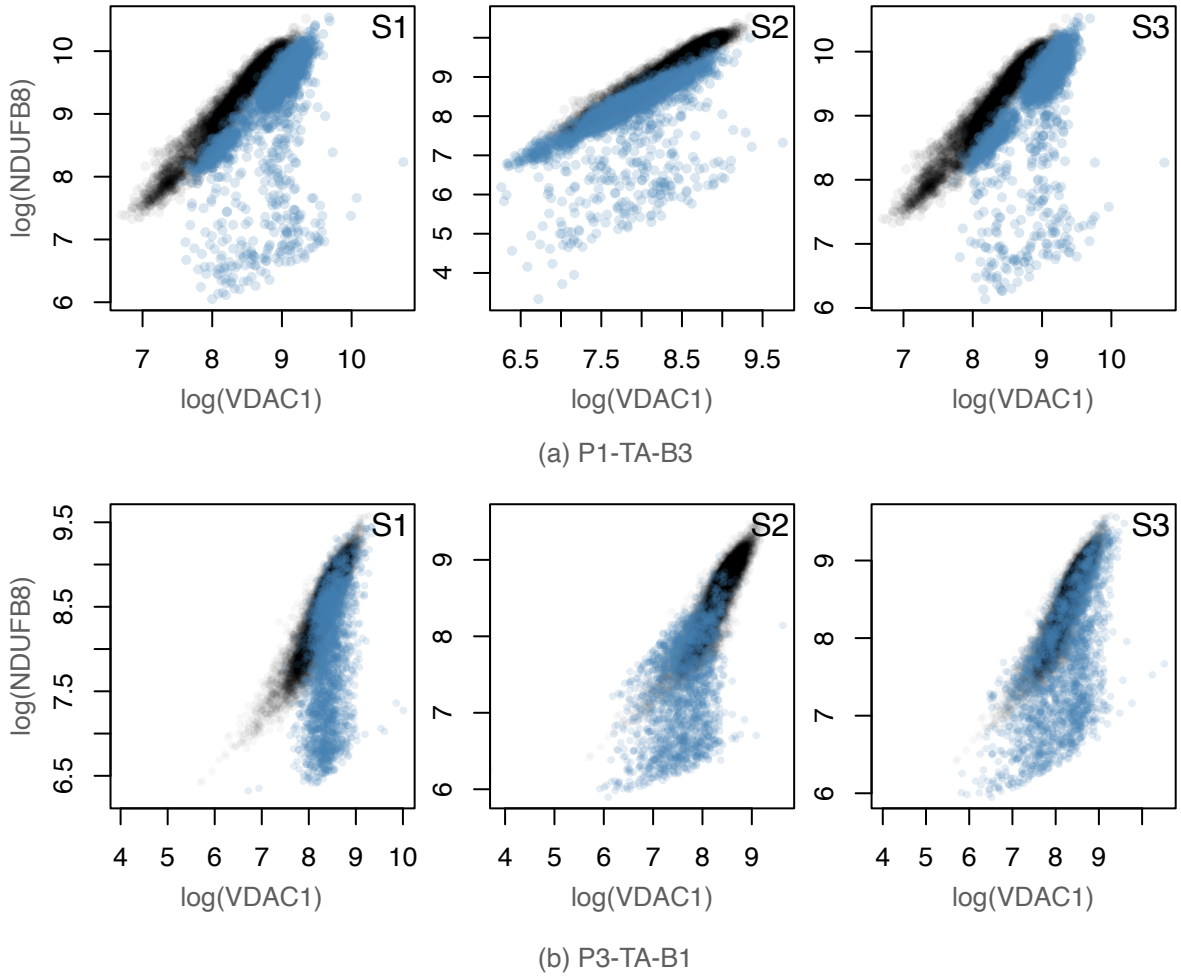


Figure 4.5: **OXPHOS protein abundance profiles show high variance in the Gomes dataset**. NDUFB8 abundance data collected from serial sections of TA muscle from patients P01 and P03 in the Gomes dataset. Protein abundances from three control subjects (black), totalling 4,287 single myofibre observations (a) and 9,118 observations (b). Protein abundances in patient myofibres (blue) total 1930 (P1-TA-B3-S1), 1895 (P1-TA-B3-S2), and 1736 (P1-TA-B3-S3) single-myofibre observations, and 2553 (P3-TA-B1-S1), 2048 (P3-TA-B1-S2), and 2437 (P3-TA-B1-S3).

Manual classifications are not available for the Gomes dataset; however, there is an alternative method for comparing model performance. As mentioned, multiple tissue samples were collected from within each patient, and it is expected that the proportion of OXPHOS deficiency be approximately consistent between adjacent sections, tissue blocks and between tissues. Therefore, the consistency of the estimated proportion of deficiency can be inspected to compare model performances. The frequentist classification pipeline for three serial sections of tissue within patient P01 can be seen in Figure 4.6(a). The model splits the like-control myofibre population for all three sections and yields very different estimates for the proportion of NDUFB8-deficient myofibres. It is not believable that the proportion of deficiency varies so drastically in adjacent sections of tissue, which are assumed to contain the same myofibres. A large variation in the proportion of deficiency estimates is seen throughout the tissue samples in P01, Figure 4.6. However, the variation between estimates is less pronounced in P02 and P03, seen in Figures 4.15 and 4.16.

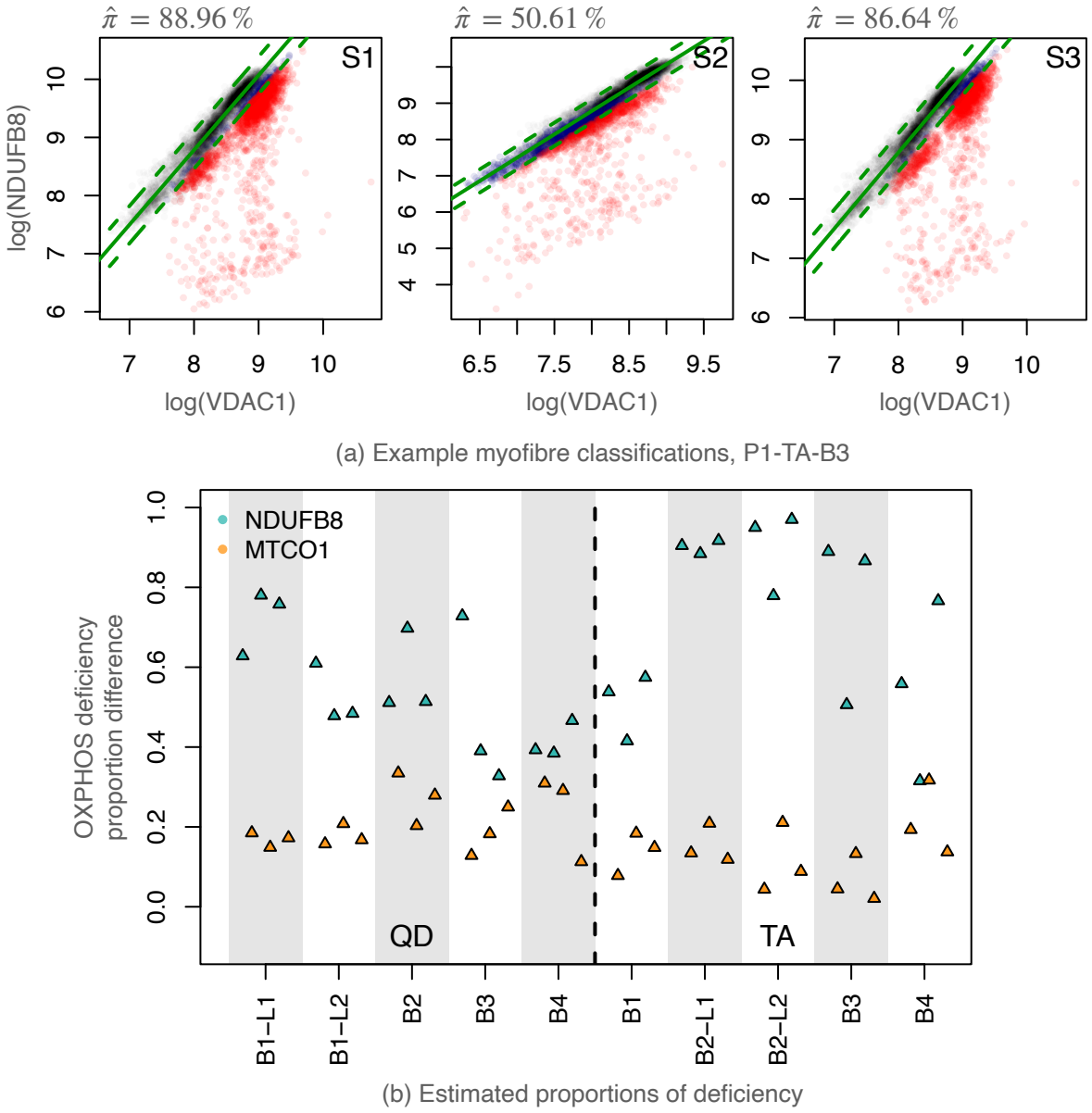


Figure 4.6: **Frequentist pipeline shows high variation within patient P01, Gomes data.** (a) OXPHOS protein abundance data from three adjacent sections of TA from patient P01 in the Gomes data. Patients' single-myofibre abundances are classified according to the frequentist pipeline, coloured blue if classified as like-control and red otherwise. The expectation and 95% predictive interval of a linear model fitted to the control data (black data-points) are shown in green. The proportion of NDUFB8-deficient myofibres, $\hat{\pi}$, is noted above each 2Dmito plot. (b) The estimated proportion of OXPHOS-deficient myofibres for each tissue sample collected from patient P01 in the Gomes dataset, estimates were calculated as the proportion of not-like-control myofibres in the sample. Tissue sections within the same block are grouped by alternating grey and white bands. The thick dashed line separates tissues collected from different muscle types, left QD and right TA.

4.2.3 Data analysis aim

The assumptions of the frequentist linear model classifier are likely too strong. Problems arising from its rigidity are highlighted here when fitting to datasets which contain more variability in OXPHOS protein abundances. Variability is a natural phenomenon caused

by both genetic and environmental factors. Further, ethical and financial considerations often only allow for a small number of control subjects, typically three or four. As we will demonstrate, this is insufficient to capture the variability in the population.

This project aims to improve the existing method of classifying the OXPHOS status of single myofibres within a tissue sample. Accurate and robust classifications, in turn, allow more accurate and robust beliefs about the proportion of not-like-control myofibres and, consequently, the proportion of myofibres whose variant load has passed the pathogenic threshold. A Bayesian hierarchical model is proposed to infer the proportion of deficiency in a patient tissue sample. A hierarchical model naturally lends itself to account for inter-subject variability, seen within the data, and Bayesian methods allow for coherent propagation of parameter uncertainty throughout the hierarchical structure.

4.3 Methods

4.3.1 Bayesian hierarchical model

The linear relationship between log protein abundance and mitochondrial mass is well documented in healthy control myofibres (Rocha et al., 2015; Vincent et al., 2024; Warren et al., 2020). The same relationship exists in healthy myofibres in patients within both observed datasets, and thus a linear model is likely a good fit for the population of healthy myofibres in both control subjects and patients. Given the variety in size, structure and proteins of interest in OXPHOS protein datasets, a general model is sought, one that does not depend on the specifics of these datasets. Therefore, a model is proposed that fits to the data represented in a single 2Dmito plot, the protein abundance of one OXPHOS protein for one patient and a number of controls. The independence between patients allows the model to not rely on a large number of patient samples or OXPHOS proteins, and it is hoped this makes the model more robust for use on other datasets. The model could include additional layers, e.g. patient-level hierarchies. However, there is likely not enough patient-level data to include this, and it makes the model less generalisable. A Bayesian hierarchical model can naturally account for the inter- and intra-subject variability while borrowing strength between subjects and a mixture model gives rise to the myofibre classification. Incorporation of prior knowledge through Bayesian methods gives weight to areas of the parameter space which are known to be more likely, reducing the risk of misclassification.

The model described below is fitted to the data represented in a single 2Dmito plot, and, therefore, all parameters described in the model are specific to that particular patient and OXPHOS protein. Consequently, for the Vincent *et al.* dataset, the model will be fitted independently 33 times, 11 patients each with three OXPHOS proteins. Similarly to the frequentist linear model, the single-myofibre protein abundances from $k-1$ control subjects are utilised in the model to identify like-control protein abundances. However, in the hierarchical approach, each control subject and the patient are considered separate experimental units in the hierarchy, totalling k groups. All experimental units have many single-myofibre OXPHOS protein abundance observations and corresponding measures of mitochondrial mass.

The proposed model is a two-component mixture model, allowing for the classification

of patient myofibres. The first component is intended to model the like-control myofibre populations in both control subjects and the patient. All control subject myofibres are assumed healthy and belong to the first component with probability 1.0. Not-like-control patient myofibres are intended to be modelled by the second component of the mixture model.

Like-control myofibres show a linear relationship. However, not-like-control myofibres can show a range of abundance profiles, which a single model cannot characterise. Consequently, choosing a specific model for the second component would likely lead to over-fitting and may result in misclassification. The second component is chosen to be a linear model with the same slope and intercept as the first, but with a constant and arbitrarily large variance. The component is intended to be dispersed enough to model all myofibres whose protein abundances are inconsistent with the first component. The model is not symmetric in the component parameters, and the model parameters are identifiable. This removes the risk of label switching commonly associated with mixture models.

The model hierarchy is placed on the slope and intercept, and so each experimental unit (subject) is modelled with a different slope and intercept, accounting for the inter-subject variability. Many previous studies have demonstrated a correlation between OXPHOS protein abundance and VDAC abundance in healthy control subjects (Rocha et al., 2015; Vincent et al., 2024; Warren et al., 2020). Furthermore, the works have elucidated a similar relationship in the sub-population of patient myofibres and separate sub-populations without this relationship, meaning that a single patient sample consists of myofibres that are a mix of like-control and not-like-control. Accordingly, the model error is assumed to be equal across all experimental units in the hierarchy.

The model, written below, describes the distribution of the j-th myofibre in the i-th subject for the data represented in a 2Dmito plot. Let $\{X_{ij}, Y_{ij}\}$ be the log-transformed OXPHOS protein and mitochondrial mass marker abundances. Discussion of the prior specification is left to the next section. The constant precision of the second component is denoted γ; all other parameters described in this section are considered unknown and to be inferred. The choice of γ is discussed in Chapter 4.4.4.

Let M_i be the number of observations for the i-th subject, and $i = 1, \ldots, k-1$ be the control subject indexes. Then, for $i = 1, \ldots, k-1$ and $j = 1, 2, \ldots, M_i$

$$Y_{ij}|m_i, c_i, \tau \sim \mathrm{N}\left(m_i X_{ij} + c_i, \tau^{-1}\right), \tag{4.3}$$

where m_i and c_i are subject-specific slopes and intercepts, and the model precision, denoted τ, is for all subjects. Myofibre protein abundance for the patient, $i = k$, is modelled using a two-component mixture model,

$$Y_{kj}|m_k, c_k, \tau, \gamma \sim \ (1-\pi)\,\mathrm{N}(m_k X_{kj} + c_k, \tau^{-1}) + \pi\,\mathrm{N}(m_k X_{kj} + c_k, \gamma^{-1}). \tag{4.4}$$

Let the latent variable, Z_{ij}, be the classification of the j-th myofibre from the i-th subject. If i is a control subject its value is fixed, for $i = 1, \ldots, k-1$ and $j = 1, \ldots, M_i$,

$$Z_{ij} = 1. \tag{4.5}$$

For $i = k$, it is a random variable, and for $j = 1, \ldots, M_k$

$$Z_{kj}|\pi \sim \text{Bern}(\pi). \tag{4.6}$$

The hierarchy appears in the prior distribution for the slope and intercept. For the priors are $i = 1, \ldots, k$

$$\begin{aligned} m_i|\mu_m, \tau_m &\sim \text{N}_{[0.1,\infty)}(\mu_m, \tau_m^{-1}), \\ c_i|\mu_c, \tau_c &\sim \text{N}(\mu_c, \tau_c^{-1}). \end{aligned} \tag{4.7}$$

Where μ_m, μ_c, τ_m, and τ_c are to be inferred, again, their prior distributions are discussed in the coming section. Note the distribution of m_i's is a truncated normal distribution. The truncation prevents the slope from becoming negative or being too close to 0.0, imparting an assumption of a positive correlation between mitochondrial mass and OXPHOS protein abundance (Rocha et al., 2015; Vincent et al., 2024; Warren et al., 2020).

4.3.2 Prior specification

Protein abundance data differs in shape and scale between OXPHOS proteins and collection methods. For example, compared to IMC, datasets collected by IF have larger scales due to the higher resolution and bit depth of IF. To keep the model generalisable and prevent the elicitation of a new prior for each dataset, we propose informing prior beliefs from control data collected with the same experimental conditions. An understanding of the relationship between (log) OXPHOS protein abundance and mitochondrial mass marker abundances can be gained by fitting a linear model to each control subject independently, which yields a set of estimates for the slopes, intercepts, and precision. The initial parameter estimates can be used to inform the expected parameters *a priori*, by setting the prior expectation to the mean of the appropriate estimates.

The aim of the analysis is to classify individual myofibres as not-like-control and to infer the proportion of not-like-control myofibres in a patient sample. The use of control subject data for the construction of prior beliefs allows more information to be imparted from the control subjects before formal inference is made. If there were more control subject samples collected in the experiment this may not be necessary, as the model should be able to learn more about global parameter values through the borrowing of strength in the hierarchy. The analysis still allows parameter values to move away from these values, and account for natural inter-subject variability. Prior uncertainty in model parameter values is not informed by the control data and was chosen to reflect our beliefs; the effect of prior uncertainty on the model outcome is discussed in Chapter 4.4.4.

Normal distributions are used to summarise prior beliefs about expected slope and intercept, μ_m and μ_c, $\mu_m \sim N(a_m, b_m)$ and $\mu_c \sim N(a_c, b_c)$. Their prior expected values, a_m and a_c, are informed from the control data and are the means of the appropriate parameter estimates after fitting linear models to the control subject data. Consequently, the uncertainty in their values is fairly low, and $b_m = b_c = 0.25^2$ is chosen to reflect this.

A gamma distribution is used to summarise the model precision prior beliefs, $\tau \sim \text{Ga}(g, h)$. The shape and rate are chosen such that the mode *a priori* is equal to the mean of the model precision estimates from the linear models independently fitted to the

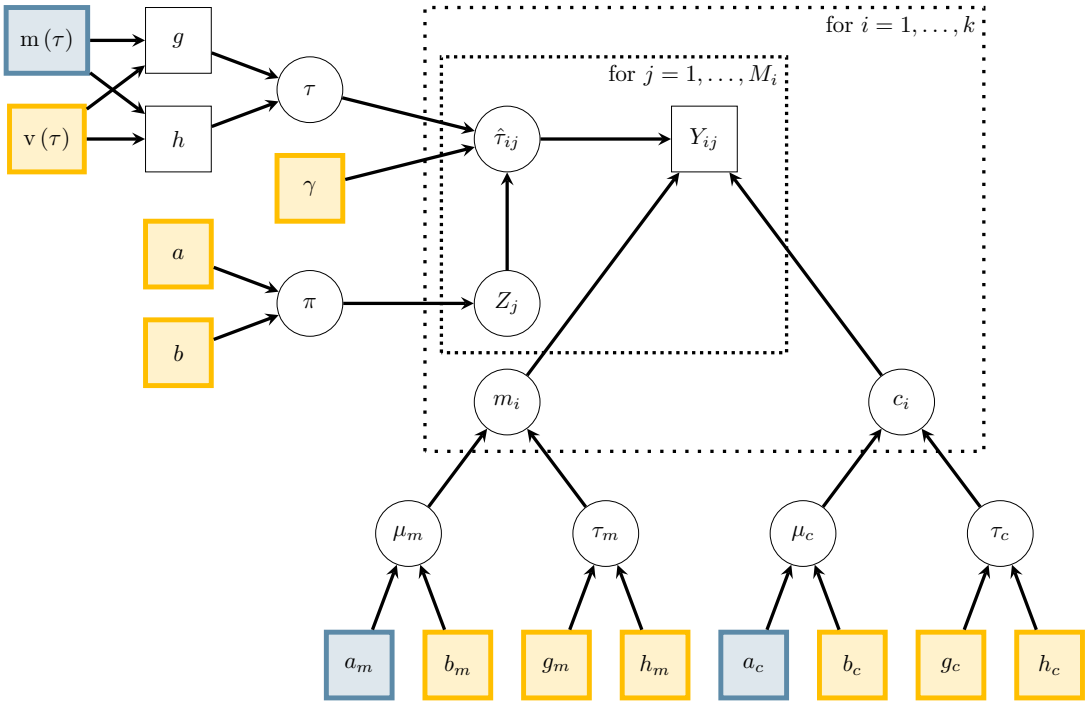


Figure 4.7: **DAG of Bayesian hierarchical linear mixture model to classify myofibre's OXPHOS status**. Parameters to be inferred are shown in a circular node, and known (hyper-) parameters are in a square node. Prior parameters informed through initial inspection of the control data are highlighted in blue, and prior parameters we chose to reflect uncertainty are in yellow. Let k be the number of subjects in the 2Dmito plot and M_i be the number of myofibres in the i-th subject sample. The subject-specific slope and intercept are labelled m_i and c_i, with expectations and precisions: μ_m and τ_m, and μ_c and τ_c, respectively. The hyper-prior parameters are labelled with appropriate subscripts. The prior parameters of the inferred model precision, τ, are informed by previous estimates, where $\mathrm{m}(\cdot)$ and $\mathrm{v}(\cdot)$ define the mean and variance of a set of estimates.

control subjects' data. The variance of τ is chosen to equal 10.0, reflecting a relatively high amount of uncertainty.

The slope and interceptes prior uncertainties, τ_m and τ_c, are believed to be independent *a priori* and summarised by gamma distributions, $\tau_m \sim \mathrm{Ga}\,(g_m, h_m)$ and $\tau_c \sim \mathrm{Ga}\,(g_c, h_c)$. There is a fair degree of uncertainty in the variation between subjects and, consequently, the variation between models fitted to the data from those subjects. Therefore, it is reasonable to expect the precisions τ_m and τ_c to be in the range $[1.0, 100]$, and prior parameters are chosen to reflect this. We chose a prior mode and variance of 50.0, which is equivalent to the shape and rate parameters being $g_m = g_c = 1.020$ and $h_m = h_c = 51.981$.

It is possible to elicit a prior for the proportion of not-like-control myofibres in a patient sample, however the beliefs are specific to a single patient given their age, mutation type, and OXPHOS protein in question. For the Vincent *et al.* dataset, no patient is believed to have a not-like-control proportion greater than 50% in any OXPHOS protein (Lehmann et al., 2019). Maintaining some generality, we choose to model each dataset represented in a 2Dmito plot with the same prior beliefs. After consultation with experts, the amount of uncertainty for a general prior was deemed so high that a flat prior on the range [0.0.5] was appropriate; therefore, $\pi \sim \mathrm{U}(0, 0.5)$.

Lastly, the constant precision of the second component, γ, is chosen to be four or five orders of magnitude smaller than the precision of the first component, τ, and fixed at $\gamma = 0.0001$. The impact of this choice is assessed in Section 4.4.4.

Figure 4.7 shows a DAG of the Bayesian hierarchical model described in this section. Parameter dependencies and prior parameters informed by control subjects are highlighted. An additional parameter is introduced in the DAG for convenience. Let

$$\hat{\tau}_{ij} = \begin{cases} \tau, & \text{if } i \text{ is a control subject or } Z_{ij} = 0 \\ \gamma & \text{otherwise} \end{cases}, \tag{4.8}$$

then $\hat{\tau}_{ij}$ is the linear model precision, conditioned on the latent state classification of the (i, j)-th myofibre.

4.3.3 Computational methods

The Bayesian model is not analytically tractable, however, its posterior can be sampled via MCMC methods. The model is, in fact, semi-conjugate, meaning the posterior can be sampled via a Gibbs sampler.

Gibbs sampler

The calculation of the FCDs can be found in Appendix A.1. Continuing the notation introduced in for DAG, $\hat{\tau}_{ij} = \tau$ if j-th myofibre from subject i is classified as like-control, for control subjects ($i = 1, \ldots, k-1$), this is always true. The Gibbs sampler targeting the joint posterior distribution of the parameters in the Bayesian model is described in Algorithm 6.

The complex nature of the joint posterior led to the Gibbs sampler suffering from very high autocorrelation and slow exploration of the parameter space. The MCMC chains were also prone to becoming stuck in local maxima and could not escape to reach global maxima within a practical timescale. Executing many chains helped the local maxima issue, but running the same inference scheme a large number of times is impractical. Instead, model inference was moved to STAN (Stan Development Team, 2020). STAN was chosen as it generally offers an efficient exploration of the parameter space and little autocorrelation compared to JAGS (Plummer, 2003).

STAN

The parameters were inferred using STAN (Carpenter et al., 2017), via the R package `rstan` (Stan Development Team, 2020). Three chains were executed and checked for convergence by multi- and univariate effective sample size, $\hat{\mathrm{R}}$ values, and inspection of trace plots (Vats et al., 2019; Vehtari et al., 2021). Any chain showing signs of non-convergence was ignored. The chain with the highest minimum univariate ESS was used as the basis for further analysis to maintain a consistent number of posterior draws between different models.

Computation cost

The computational cost of the model is greater than the frequentist model. However, for a single dataset, the computational expense is relatively low. To indicate the time and computing power required to fit the model. The Vincent *et al.* dataset contains 1,155 control subject myofibres and between 152 - 1,199 patient myofibres, resulting in 1,307 to 2,354 data points per 2Dmito plot and per inference scheme. The wall clock time for a single execution of the inference scheme (one chain for the data represented in a single 2Dmito plot) ranged between 44 and 155 seconds. The Gomes *et al.* dataset contained larger tissue samples, with more myofibres. For patient P04, the total number of myofibres per inference scheme ranged from 10,880 to 12,609, including both control and patient myofibres. The wall clock execution time for one chain of the inference scheme ranged from 4,174 to 35,114 seconds, a significant increase in inference time. Each scheme was executed for 22,000 iterations, of which the first 20,000 were considered part of the burn-in period, an amount found satisfactory for convergence for the majority of the dataset. Inference was executed on a 2023 MacBook Pro with an M2 Pro chip and 16GB of RAM.

Algorithm 6 Hierarchical linear mixture model Gibbs sampler

1. Initialise the state of the chain

$$\boldsymbol{\theta}^{(0)} = ({\tau_m}^{(0)}, {\tau_c}^{(0)}, {\mu_m}^{(0)}, {\mu_c}^{(0)}, {m_1}^{(0)}, {m_2}^{(0)}, \ldots, {m_k}^{(0)}, {c_1}^{(0)}, {c_2}^{(0)}, \ldots, {c_k}^{(0)}, \tau^{(0)})^{\mathrm{T}}$$

and set $t = 1$.

2. Sample the latent states $\boldsymbol{Z}^{(t)}$. For $i = 1, \ldots, k-1$, ${Z_{ij}}^{(t)} = 0$ for all j, and for $i = k$ and $j = 1, 2, \ldots, M_k$

$$Z_{kj}^{(t)} | \boldsymbol{\theta}^{(t-1)} \sim \text{Bern}\,(p_{kj})$$

$$p_{kj} = \frac{\sqrt{\gamma}\exp\left[\frac{-\gamma}{2}(Y_{kj} - \tilde{\mu}_{kj})^2\right]}{\sqrt{\tau}\exp\left[\frac{-\tau}{2}(Y_{kj} - \tilde{\mu}_{kj})^2\right] + \sqrt{\gamma}\exp\left[\frac{-\gamma}{2}(Y_{kj} - \tilde{\mu}_{kj})^2\right]}$$

$$\tilde{\mu}_{kj} = m_i X_{kj} + c_i$$

3. Generate $\boldsymbol{\theta}^{(j)}$ by sequential random draws from the parameters' FCDs. Let $\tilde{\mu}_{ij} = m_i X_{ij} + c_i$ and $\hat{\tau}_{ij}^{(t)} = \tau^{(t-1)}\mathbb{I}({Z_{ij}}^{(t)} = 0)) + \gamma\mathbb{I}({Z_{ij}}^{(t)} = 1)$

i ${\tau_m}^{(t)} | \boldsymbol{Y}, \boldsymbol{Z}^{(t)}, \boldsymbol{\theta}^{(t)}_{-\tau_m} \sim \text{Ga}\left(g_m + N/2,\ \ h_m + \frac{1}{2}\sum_{i=1}^{k}({\mu_m}^{(t-1)} - {m_i}^{(t-1)})^2\right)$

ii ${\tau_c}^{(t)} | \boldsymbol{Y}, \boldsymbol{Z}^{(t)}, \boldsymbol{\theta}^{(t)}_{-\tau_c} \sim \text{Ga}\left(g_c + \frac{N}{2},\ \ h_c + \frac{1}{2}\sum_{i=1}^{k}({\mu_c}^{(t-1)} - {c_i}^{(t-1)})^2\right)$

iii ${\mu_m}^{(t)} | \boldsymbol{Y}, \boldsymbol{Z}^{(t)}, \boldsymbol{\theta}^{(t)}_{-\mu_m} \sim \text{N}\left(\frac{a_m b_m + {\tau_m}^{(t)}\sum_{i=1}^{k} {m_i}^{(t-1)}}{b_m + {\tau_m}^{(t)} N},\ \ \left(b_m + N{\tau_m}^{(t)}\right)\right)$

iv ${\mu_c}^{(t)} | \boldsymbol{Y}, \boldsymbol{Z}^{(t)}, \boldsymbol{\theta}^{(t)}_{-\mu_c} \sim \text{N}\left(\frac{a_c b_c + {\tau_c}^{(t)}\sum_{i=1}^{k} {c_i}^{(t-1)}}{b_c + {\tau_c}^{(t)} n},\ \ \left(b_c + N{\tau_c}^{(t)}\right)\right)$

v For $i = 1, 2, \ldots, k$,

$${m_i}^{(t)} | \boldsymbol{Y}, \boldsymbol{Z}^{(t)}, \boldsymbol{\theta}^{(t)}_{-m_i} \sim \text{N}\left(\frac{{\mu_m}^{(t)}{\tau_m}^{(t)} + \tilde{\mu}_m\tilde{\tau}_m}{{\tau_m}^{(t)} + \tilde{\tau}_m},\ \ ({\tau_m}^{(t)} + \tilde{\tau}_m)^{-1}\right)$$

$$\tilde{\tau}_m = \sum_{j=1}^{M_i} \hat{\tau}_{ij}^{(t-1)} {X_{ij}}^2$$

$$\tilde{\mu}_m = \frac{1}{\tilde{\tau}_m^{(t)}} \sum_{j=1}^{M_i} \hat{\tau}_{ij}^{(t-1)} \left(X_{ij}Y_{ij} - {c_i}^{(t-1)}X_{ij}\right)$$

vi For $i = 1, 2, \ldots, k$,

$${c_i}^{(t)} | \boldsymbol{Y}, \boldsymbol{Z}^{(t)}, \boldsymbol{\theta}^{(t)}_{-c_i} \sim \text{N}\left(\frac{{\mu_c}^{(t)}{\tau_c}^{(t)} + \sum_{j=1}^{M_i} \hat{\tau}_{ij}^{(t-1)} \left(Y_{ij} - {m_i}^{(t)}X_{ij}\right)}{\tau_c + \sum_{j=1}^{M_i} \hat{\tau}_{ij}^{(t-1)}},\ \ (\tau_c + \sum_{j=1}^{M_i} \hat{\tau}_{ij}^{(t-1)})^{-1}\right)$$

vii $\tau^{(t)} | \boldsymbol{Y}, \boldsymbol{Z}^{(t)}, \boldsymbol{\theta}^{(t)}_{-\tau} \sim \text{Ga}\left(\tilde{g}_\tau, \tilde{h}_\tau\right)$

$$\tilde{g}_\tau = g + \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{M_i} \mathbb{I}({Z_{ij}}^{(t)} = 0), \quad \tilde{h}_\tau = h + \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{M_i} (\tilde{\mu}_{ij} - Y_{ij})^2 \mathbb{I}({Z_{ij}}^{(t)} = 0)$$

4. Increment t by 1 and return to step 2.

R package

The STAN files and inference functions were stored in an `R` package, allowing the Bayesian hierarchical model to be implemented by others. The package is available for download through GitHub, (https://github.com/jordanbchilds/analysis2Dmito). In addition to functions for performing inference, the package includes several plotting functions to help visualise output, as well as function documentation and a detailed README file on the GitHub page. Which describes data format requirements and an example inference.

4.3.4 Generating synthetic data

The model assumptions are investigated through two synthetic datasets. One dataset has OXPHOS protein abundances, which resemble IMC data in scale, and the other dataset has larger protein abundances reflecting the higher resolution data collected by IF, referred to as DO1 and DO2, respectively. A larger inter-subject variability is imposed in D02 to demonstrate the model's flexibility.

D01

To generate the synthetic dataset D01, ground truth parameter values are set by randomly sampling from the marginal posterior distributions after fitting the Bayesian hierarchical model to the Vincent *et al.* dataset. Therefore, the synthetic data contains four control subjects, 11 patients and protein abundances for CYB, MTCO1, and NDUFB8. Control data is repeated between 2Dmito plots, and so the Bayesian model is independently fit to the control data for each patient and OXPHOS protein. The resulting control subject posterior beliefs are, therefore, not consistent across posteriors. Consequently, a different set of control subject ground-truth parameters and synthetic protein abundances are generated for each patient and OXPHOS protein.

To generate ground-truth parameter values low-level parameters $(\mu_m, \mu_c, \tau_m, \tau_c)$ are set at their posterior expectations and high-level parameters are randomly sampled conditioning on their value. Let an asterisk, *, indicate a ground-truth parameter. For a given patient and OXPHOS protein, the ground-truth slope and intercept for each experimental unit is sampled

$$\begin{aligned} m_i^* | \mu_m^*, \tau_m^* &\sim \mathrm{N}\left(\mu_m^*, \tau_m^{*\,-1}\right) \\ c_i^* | \mu_c^*, \tau_c^* &\sim \mathrm{N}\left(\mu_c^*, \tau_c^{*-1}\right). \end{aligned} \tag{4.9}$$

The ground-truth proportion of deficiency is randomly sampled from the appropriate posterior distribution, $\pi^* \sim p(\pi|\boldsymbol{Y})$. The ground-truth OXPHOS status of each patients' myofibres are independently sampled, dependent on the ground-truth proportion of deficiency, from a Bernoulli distribution. Ground truth OXPHOS status for control subject myofibres was set to like-control, i.e. $Z_{ij} = 0$ for $i = 1, \ldots, k-1$ and $\forall j$, patient ground-truth OXPHOS status is randomly sampled. For $j = 1, 2, \ldots, M_k$

$$Z_{kj}^* \sim \mathrm{Bern}(\pi^*). \tag{4.10}$$

Synthetic OXPHOS protein abundances for control and like-control patient myofibres were sampled from a linear model, with slopes and intercepts being the ground-truth

parameters as appropriate, dependent on the observed (log) VDAC abundances. For $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, M_i$,

$$Y_{ij}|Z^*_{ij} = 0 \sim \mathrm{N}\left(m^*_i X_{ij} + c^*_i, \tau^{*-1}\right). \tag{4.11}$$

Synthetic OXPHOS protein abundances for not-like-control myofibres were sampled from a linear model with a decreased expectation and higher variance than the linear model that sampled the patient's like-control protein abundances. For $j = 1, 2, \ldots, M_k$

$$Y_{kj}|Z^*_{kj} = 1 \sim \mathrm{N}\left(m^*_k X_{kj} + c^*_k - \frac{11}{\sqrt{\tau^*}}, \frac{10}{\tau^*}\right). \tag{4.12}$$

D02

The synthetic dataset D02 is created with a larger scale and has increased inter-subject variability. The generation of the D02 closely follows D01; however, two extra steps are implemented. Firstly, after the sampling from the posterior, the model precisions, τ_m, τ_c and τ, are decreased by a factor of 4, increasing between-subject variability. Secondly, the observed (log) VDAC abundances are linearly transformed, increasing the scale of the data. For $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, M_i$

$$\begin{aligned} Y_{ij}|Z^*_{ij} = 0 &\sim \mathrm{N}\left(m^*_i(3X_{ij} + 8) + c^*_i, \tau^{*-1}\right) \\ Y_{ij}|Z^*_{ij} = 1 &\sim \mathrm{N}\left(m^*_i(3X_{ij} + 8) + c^*_i - \frac{11}{\sqrt{\tau^*}}, \frac{10}{\tau^*}\right). \end{aligned} \tag{4.13}$$

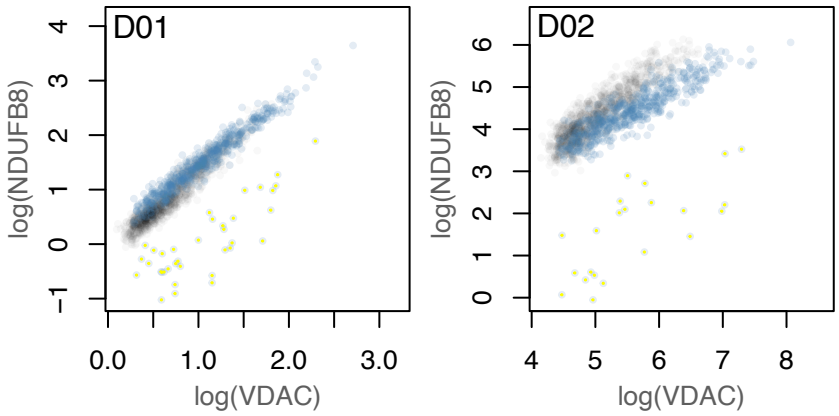


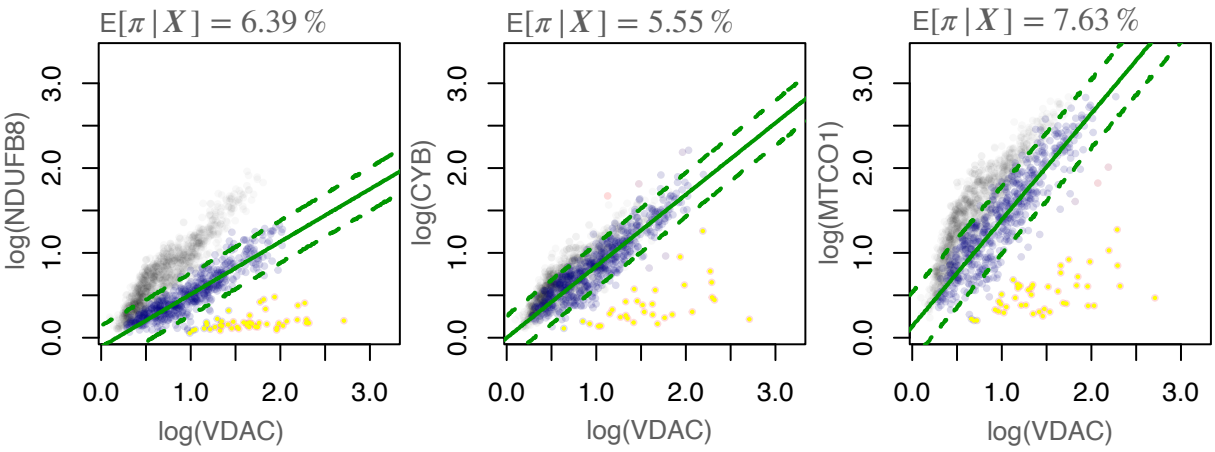
Figure 4.8: **Example 2Dmito plots from synthetic datasets D01 and D02**. Synthetic control subject abundances are shown in black and synthetic patient data is blue. Single-myofibre abundances whose ground-truth OXPHOS status is deficient are highlighted with a single yellow dot. The data were generated using the methods described in Chapter 4.3.4. Using the logged VDAC abundances and posterior distributions associated with fitting the model to logged NDUFB8 abundances for P09 in the Vincent *et al.* dataset.

4.4 Results

4.4.1 Vincent dataset

The Bayesian model fit is shown for an example patient, P09, in Figure 4.9. The 95% posterior predictive intervals match the like-control patient myofibres well and have moved

away from the control myofibres where needed. Initial visual inspection suggests that the Bayesian model provided a better classification than the frequentist linear model, see Figure 4.3. Unlike the frequentist model, the Bayesian linear model has not bisected the like-control patient myofibre population. The confusion matrices, Figure 4.9, confirm that the Bayesian model has classifications much more in line with the manual classifications, with misclassification rate ranging from 0.5% to 1.8% in patient P09.



NDUFB8

Bayes \ Manual	0	1
0	525	10
1	0	36

CYB

Bayes \ Manual	0	1
0	525	3
1	2	41

MTCO1

Bayes \ Manual	0	1
0	539	2
1	1	29

Figure 4.9: **Bayesian model correctly identifies the majority of like-control patient myofibres**. Model posterior and classifications for three OXPHOS proteins for P09 with 571 myofibres (coloured points). Control myofibres are shown in grey (1,155 myofibres from four healthy subjects). Patient myofibres are shown on a scale of blue to red, depending on their probability of being not-like-control. The 95% posterior predictive interval and fitted values for the healthy patient (log) OXPHOS abundance are shown in green. The tables show confusion matrices when classifying P09 patient myofibres, comparing the Bayesian classification method to the manual classifications. A single myofibre was characterised as like-control if the expected marginal posterior probability of the myofibre being not-like-control was less than 50%.

Marginal posterior parameter beliefs for an example model fit are shown in Figure 4.10. Prior and posterior beliefs are very similar for τ_m and τ_c, indicating little has been learnt about their value after observing the data. This is not unique to the data represented in this 2Dmito plot and can be seen across the Vincent *et al.* dataset. However, the beliefs about the expected slope and intercept, μ_m and μ_c, have been consistently updated across the dataset. Their posterior expectation has not changed drastically, but this is to be expected given the high degree of certainty in their values *a priori*. A reduction in uncertainty can be seen in the population slope and intercept, primarily due to the reduction in uncertainty in their expectations, μ_m and μ_c. A large decrease in parameter uncertainty can also be seen in the not-like-control proportion and the model precision.

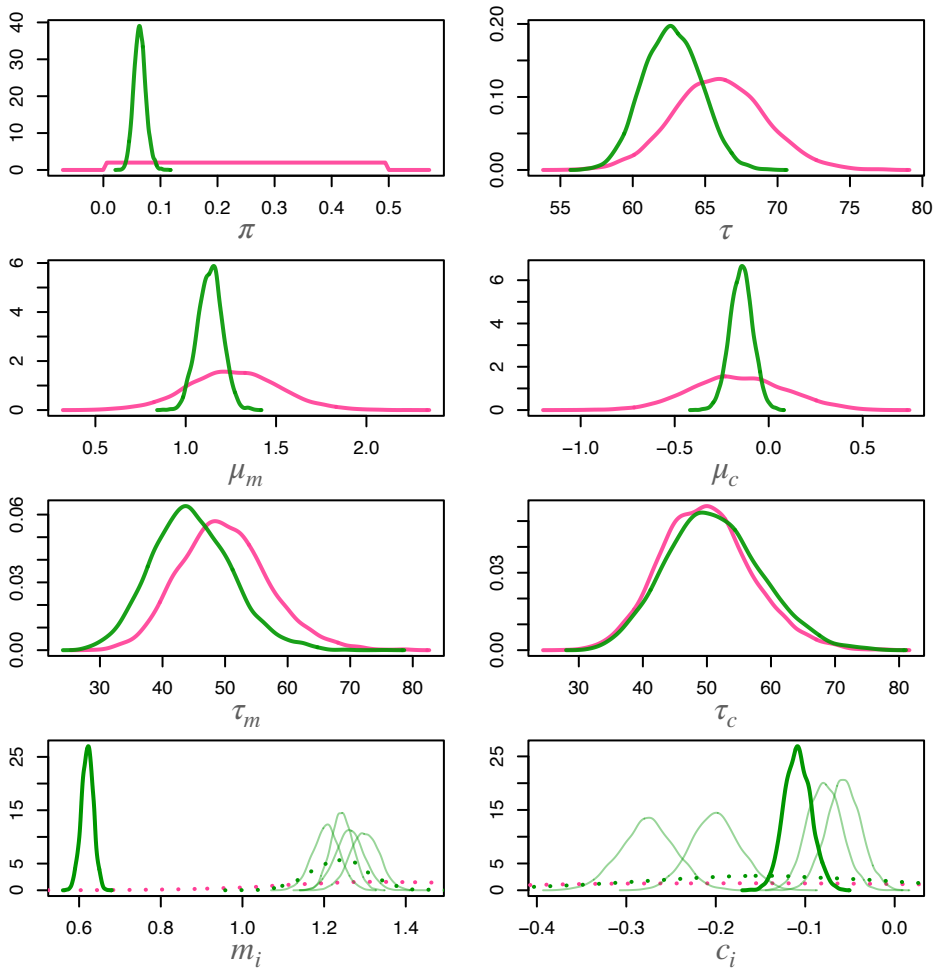


Figure 4.10: **Marginal prior and posterior densities for all parameters after classifying myofibres from P09 by NDUFB8**. Kernel density estimates of 20,000 draws from prior (pink) and posterior (green) distributions. Posterior densities of patient slope and intercept are thick, solid green. The control posteriors are shown as transparent green. Dotted lines indicate the population marginal densities of the population-level distributions of the slope and intercept, $\mathrm{N}\left(\mu_m, \tau_m^{-1}\right)$ and $\mathrm{N}\left(\mu_c, \tau_c^{-1}\right)$.

The inter-subject variability is highlighted in the marginal posterior distributions for the subject-specific slope and intercepts, Figure 4.10. For this patient, the linear model's slope, m_i, shows a substantial difference from the control subject parameters, with no overlap in the bulk of the posterior densities. Similar differences in the subject-specific posterior beliefs can be seen throughout the dataset.

The differences in the output from the two models are highlighted in the inferred/estimated not-like-control proportions. The difference between the frequentist model's estimates and the manual classification estimate ranges from 7% to 85%, with a mean difference of 32%. The expected difference in the Bayesian approach ranges from -3% to 6%, with a mean expected difference of -0.1%. The posterior expectation was calculated by the mean of the posterior draws. The not-like-control proportions from the two models are also substantially different, the probability of observing the frequentist model's estimates, given the appropriate Bayesian posterior, range is $(\leq 2.0 \times 10^{-4}, 6.0 \times 10^{-4}]$.

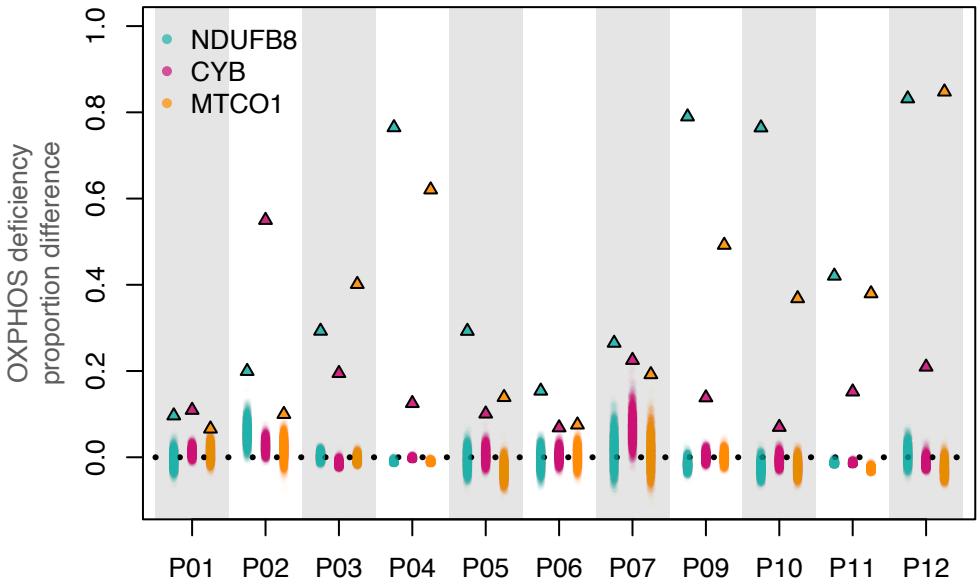
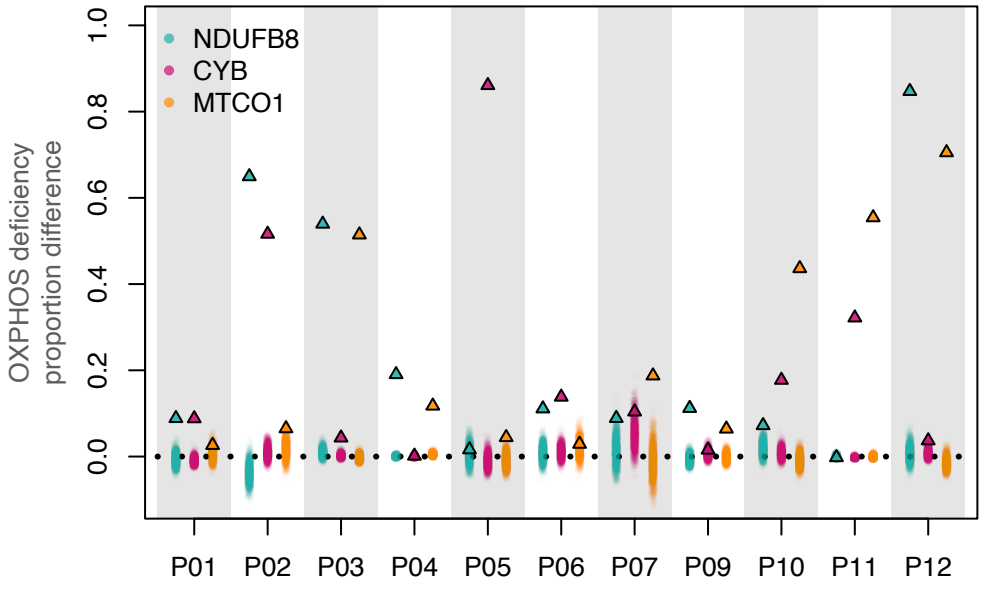


Figure 4.11: **The difference between the frequentist and manual estimates of the proportion of not-like-control is larger than that of the Bayesian and manual estimates**. The differences between the manual and frequentist classifications are represented by point estimates, shown as triangles. The difference between the Bayesian and manual classifications is distributions summarised by 5,000 posterior draws. Each posterior sample of the difference distribution is shown as a small transparent circle. The dashed line is zero. Therefore, the distance between the dashed line and the points is the difference in the estimated proportion of not-like control myofibres from the two methods models and the manual classification.

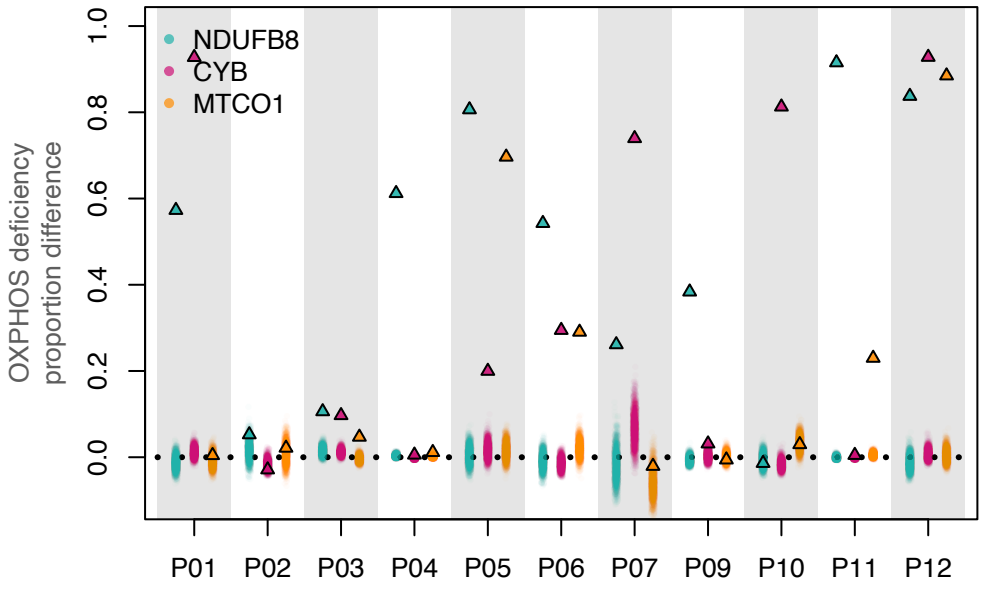
4.4.2 Synthetic data

The synthetic datasets were classified by both the Bayesian and frequentist classification pipelines. The performance of the two models is evaluated by comparing the inferred/estimated not-like-control proportion and the inferred/estimated OXPHOS status of individual myofibres compared to the ground-truth values.

For convenience, a single myofibre is characterised as not-like-control by the Bayesian model if the expectation of its marginal posterior probability of belonging to the second component is above 0.5. After fitting the Bayesian classifier to the synthetic dataset D01, six patient myofibres, of a possible 19,553, were characterised in opposition to their ground truths. Five of the misclassifications are false-negatives, incorrectly classifying a single myofibre as like-control. By chance, the myofibres misclassified as like-control showed abnormally high protein abundances, which closely resembled those of like-control myofibres. The single false-positive misclassification showed a particularly high protein abundance compared to the rest of the data. A small number of synthetic myofibres with (log) OXPHOS protein abundances outside of the like-control range is expected when generating a synthetic dataset due to the random nature of its generation. Therefore, some misclassifications are to be expected.



(a) D01



(b) D02

Figure 4.12: **Bayesian model accurately estimates ground-truth not-like-control proportion, D01 and D02**. The differences between the ground-truth and frequentist classifications are point estimates and are shown as triangles. The difference between the inferred proportion of deficiency and ground-truth distributions is summarised by 5,000 posterior draws. Each posterior sample of the difference distribution is shown as a small transparent circle. The dashed line is zero. Therefore, the distance between the dashed line and the points is the difference in the estimated proportion of not-like-control myofibres from the two models and the ground-truth value used to generate the synthetic data.

Upon inspection of the 99% posterior HDIs, the posterior proportion of not-like-control myofibres shows no difference to the ground-truth values. The expected posterior difference between the inferred proportion and ground-truth is $[-0.07, 0.03]$. When fitting the frequentist model, the difference in the ground-truth and estimated proportion of deficiency range is $[-0.006, 0.868]$. Further, a substantial difference is found in the not-

like-control proportion estimates of the two models for the majority of synthetic datasets. The posterior probability of observing the frequentist estimate, or more extreme, was less than 1% for 24 out of 33 samples. The differences in the not-like-control proportion between the two models and the ground-truth can be seen in Figure 4.12.

The Bayesian model also shows more accurate estimates of the not-like-control proportion in D02 than the frequentist model. The expected posterior difference between the Bayesian model and the ground-truth is $[-0.058, 0.066]$, while the frequentist difference range is $[-0.029, 0.928]$. Similar to D01, the ground-truth values lie within the appropriate 99% HDIs, indicating no substantial difference.

The Bayesian model's susceptibility to over-fitting was assessed by randomly splitting D01 into training and validation subsets. Approximately 80% of each subject's myofibres were used for inferring the parameters of the Bayesian model, and the remaining 20% of patient myofibres were used for validating model classifications at a single myofibre level. The marginal posterior probabilities of unseen myofibres being not-like-control were compared to the marginal posterior probabilities when the whole dataset was seen during inference. No evidence of a difference between the probabilities was found for any myofibre by checking that 0.0 lay within the 99% HDIs of the posterior differences.

4.4.3 Gomes dataset

P01

Within patient P01 the model successfully identified all like-control myofibre populations in both OXPHOS proteins, even when the population was split. The inferred proportion of OXPHOS deficiency shows a higher degree of consistency between adjacent serial sections and tissue blocks than the frequentist method for all samples in patient P01; see Figure 4.13. In all but one case the frequentist model estimated the deficient proportion to be substantially higher than the Bayesian analysis when examining the 99% HDI of the posterior proportions, consistent with results from the Vincent *et al.* and synthetic datasets. It is clear from Figure 4.13 that the within-block frequentist estimates have a much higher variance than the Bayesian model, indicating a much higher degree of consistency.

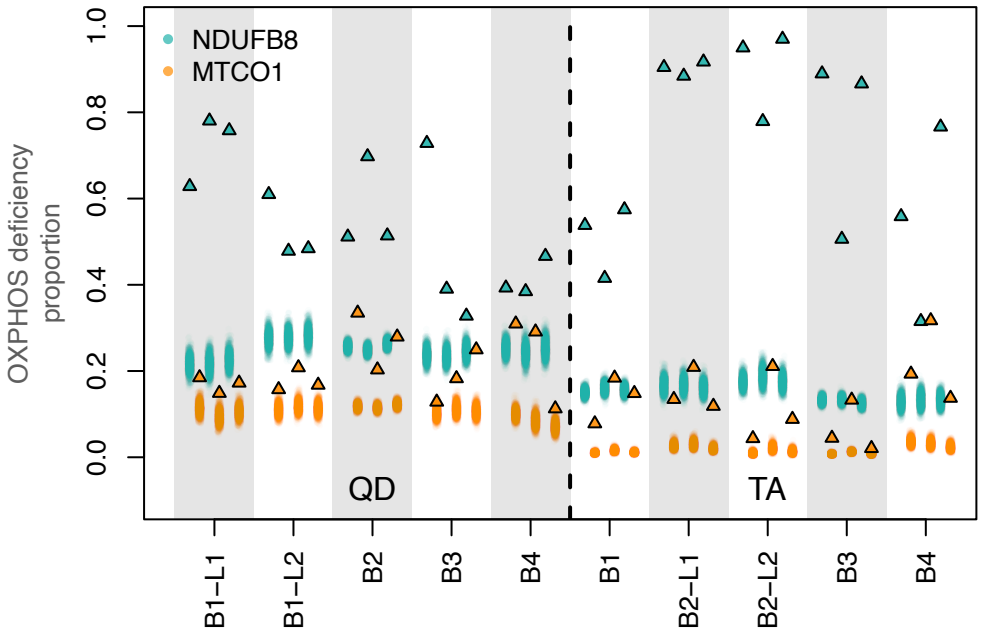


Figure 4.13: **Bayesian model consistently infers OXPHOS deficiency proportion, P01 Gomes dataset**. Posterior beliefs of the proportion of OXPHOS deficiency for all tissue samples in patient P01, from the Gomes dataset. Tissue serial sections are grouped by grey vertical bands and samples from different tissues (QD and TA) are separated by the dashed black line, and labelled appropriately. Bayesian posterior distribution is summarised by 2,000 draws from the posterior distribution inferred via STAN. Single-point frequentist estimates of the proportion are shown as triangles outlined in black.

P02

The Bayesian model successfully identified the MTCO1 status of all samples and 19 (out of 21) samples for NDUFB8, identifying the population of like-control myofibres and inferring a reasonable value of the proportion of deficiency. Unfortunately, the inference chains for the remaining two schemes failed to converge to one posterior distribution; in both cases, three chains were executed, and one of them was considered a successful classification. The number of chains was then increased to 20, and they were similarly divided between two distributions: one with successful classification and one classifying all patient myofibres as like-control. Unlike other examples, the chains could not be distinguished by signs of non-convergence. Therefore, the inference scheme was rerun with a larger burn-in, 100,000 iterations, compared to the 20,000 iterations previously used. The increased burn-in period aided inference, and two of the three executed chains correctly identified the like-control population of myofibres. However, the chains were still indistinguishable in their univariate and multivariate effeective sample sizes (ESSs).

A further increase to the burn-in period would, in theory, allow all chains to converge to the same posterior distribution. However, increasing the number of burn-in iterations to 200,000 was computationally prohibitive and exhausted the memory (RAM) of a 16GB machine. Inference was then attempted using a Gibbs sampler, Algorithm 6, which was able to better deal with the memory requirements of the inference scheme but its slow exploration of the parameter space and high autocorrelation meant it was also impractical, and after a 2,000,000 iteration burn-in period the chains had not converged.

The chains of the inference scheme, executed in STAN, were re-inspected and drastic differences were seen in their posterior likelihoods. Inference schemes which correctly classified like-control myofibres showed substantially higher posterior likelihoods than those that didn't, see Figure 4.14. The higher posterior likelihood indicates that this distribution is in a region of higher posterior density, and given sufficient time, all chains would converge to this distribution, assuming this is the posterior maximum. The chains with a lower posterior likelihood are likely stuck in a local maximum, which is difficult to escape.

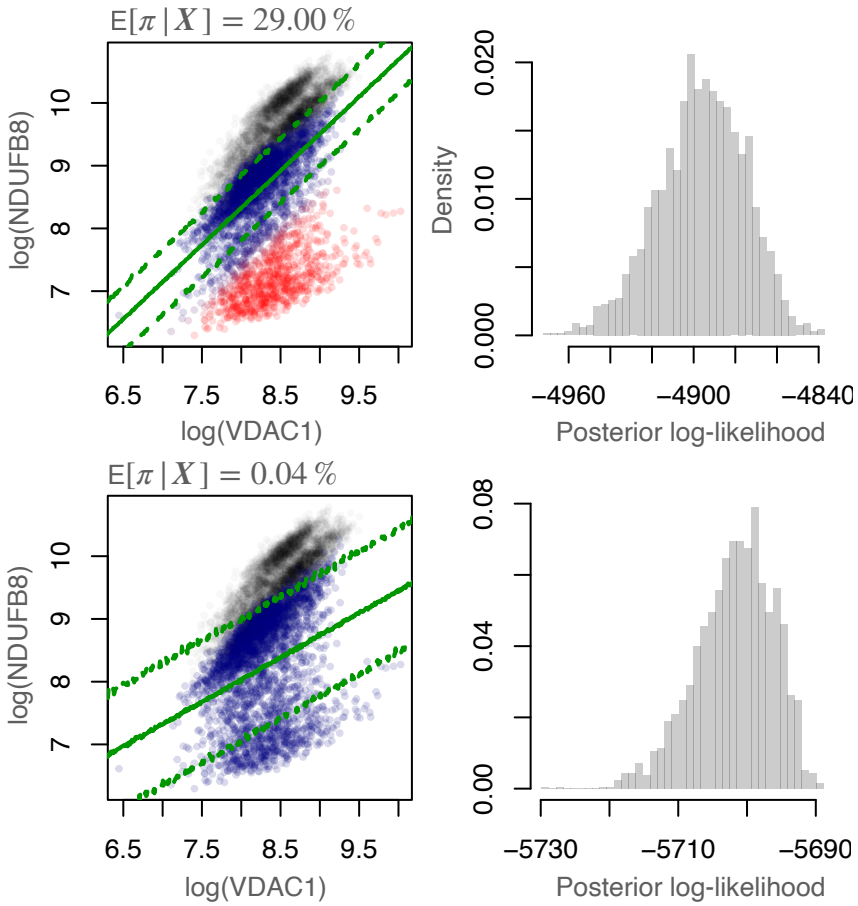


Figure 4.14: **Inference chains with high posterior likelihood correctly classify patient myofibres**. The two posterior distributions found when fitting the Bayesian classifier to a patient sample, P2-TA-B1-S2, from the Gomes dataset, as well as the posterior likelihoods. Control data are coloured black, and patient data are blue or red depending on their posterior classification. Single myofibres classified as being like-control are blue, and not-like-control myofibres are red. The 95% posterior predictive for the linear model fitted to the like-control patient data is shown in green. The expected proportion of deficient myofibres, $\mathrm{E}[\pi|\boldsymbol{X}]$, is shown for both inference schemes.

Other than these two samples, the Bayesian model performed well, inferring the proportion of deficiency more consistently than the frequentist method, which, similarly to before, produced substantially higher proportion estimates, as shown in Figure 4.15.

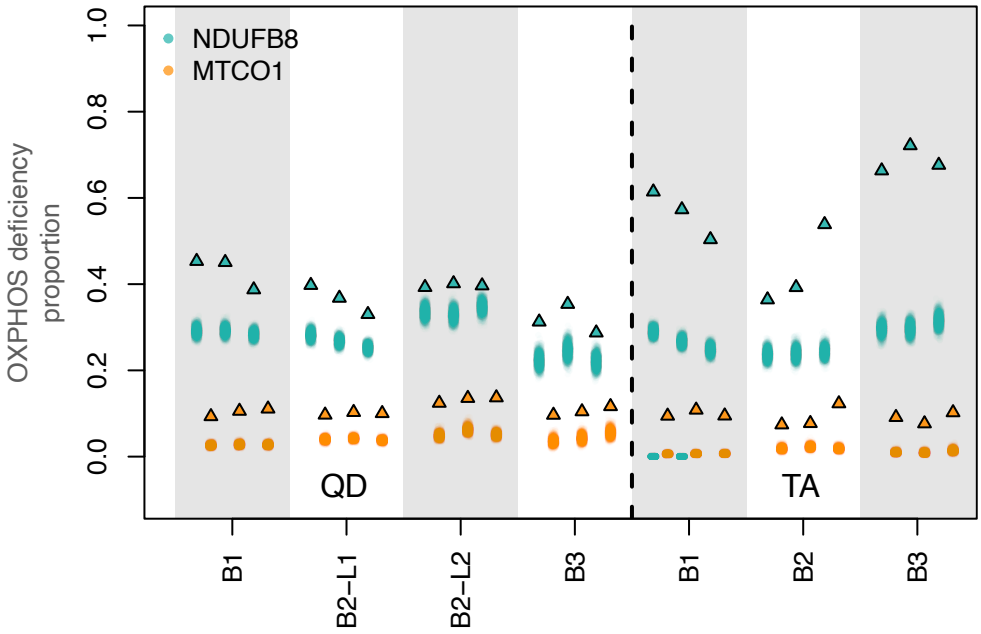


Figure 4.15: **Bayesian model consistently infers deficiency proportion in P02, Gomes data**. Posterior beliefs of the proportion of OXPHOS deficiency for all tissue samples in patient P02, from the Gomes dataset. Grey vertical bands group tissue serial sections, and samples from different tissues (QD and TA) are separated by the dashed black line, and labelled appropriately. Bayesian posterior distribution is summarised by 2,000 draws from the posterior distribution inferred via STAN. Single-point frequentist estimates of the proportion are shown as black triangles. For the two schemes which did not converge, TA-B1-S1 and TA-B1-S2, the inferred proportion from both posterior modes are shown.

P03

Recall that P03 of this dataset showed different OXPHOS protein abundances compared to patients P01 and P02 and the patients within the Vincent *et al.* dataset, lacking a clear distinction between like-control and not-like-control myofibres, shown in Figure 4.5. Unfortunately, the Bayesian model largely failed to classify the myofibres within this patient. Nine tissue samples were analysed by QIF, the estimated proportion of NDUFB8 deficiency for six of the samples was approximately zero, in contradiction to expert prior beliefs. All samples showed very low levels of MTC01 deficiencies, by inspection of their 2Dmito plots, and this was reflected in Bayesian posteriors, Figure 4.16. Apart from one chain, which was removed, the inference chains of the misclassified output converged to the same posteriors and showed no signs of non-convergence. Therefore, there is no evidence to suggest that this misclassification is a convergence issue, such as that seen for P02, and the posteriors should be taken as they are.

The misclassifications, although disappointing, are not surprising. The model assumes that the like-control myofibres show a clear linear relationship between OXPHOS and VDAC abundances on the log-scale. However, there is almost no separation between like-control and not-like-control patient protein abundances for the samples in this patient, see Figure 4.5.

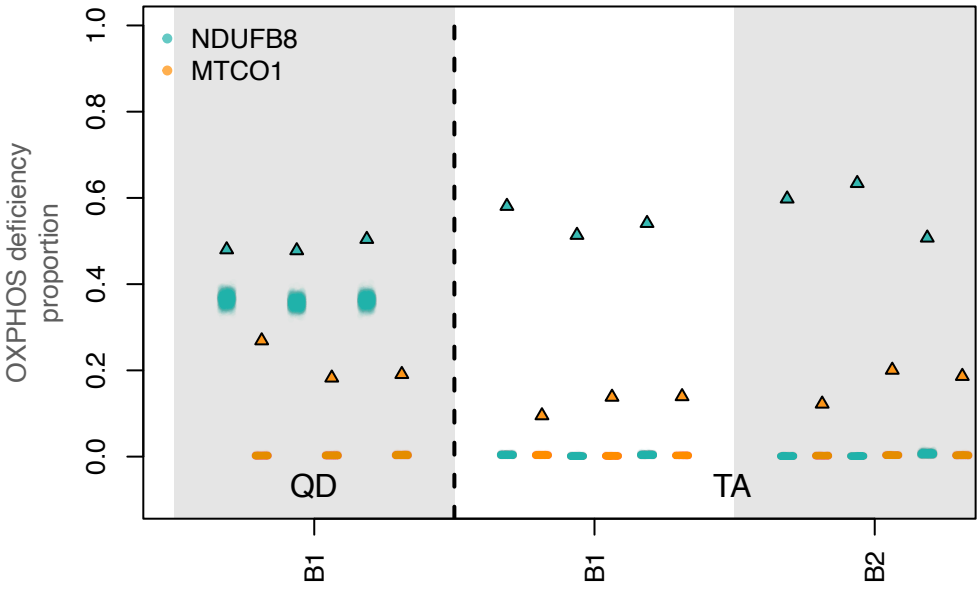


Figure 4.16: **Bayesian model fails to consistently classify myofibres within non-distinct healthy OXPHOS abundances**. Posterior beliefs of the proportion of OXPHOS deficiency for all tissue samples in patient P02, from the Gomes dataset. Tissue serial sections are grouped by grey vertical bands and samples from different tissues (QD and TA) are separated by the dashed black line, and labelled appropriately. Bayesian posterior distribution is summarised by 2,000 draws from the posterior distribution inferred via STAN. Single-point frequentist estimates of the proportion are shown as triangles outlined in black.

4.4.4 Sensitivity to prior specification

The Bayesian model requires the prior specification of several parameters, the value of some has been informed by the control subject's data, while some have not. Here, we assess the impact of prior parameters, which were chosen to reflect our prior beliefs. To this end, the inference for Vincent *et al.* was rerun with a varying set of prior parameters. First, we consider the parameter γ, the fixed precision of the second component of the mixture model. Since the parameter is not inferred, it is reasonable to assume that its choice can significantly impact the model output.

The Bayesian model is fit to Vincent *et al.* dataset with γ fixed at varying values between $[0.0000001, 1.0]$. The resulting model fits are compared by inspecting the mean absolute difference (MAD) between the inferred not-like-control proportion and the estimate from manual classification. Figure 4.17 shows that the value which minimises this criterion is $\gamma = 0.0001$. Values of γ in the range $[0.00001, 0.01]$ show similarly low MAD values, indicating that these could also be appropriate.

To assess the impact of prior uncertainty on the model output, the model parameters were inferred using two prior parameter sets: one with increased prior uncertainty and one with decreased prior uncertainty. An assessment of the prior uncertainty in π was not considered, as its prior is flat, and its choice was discussed in Section 4.3.2. Prior uncertainty for the rest of the parameters; μ_m, τ_m, μ_c, τ_c and τ, was considered. The prior variance for each parameter was increased and decreased by a factor of five. The resulting models were inspected through the not-like-control proportion. There was no evidence of a substantial difference between the proportions throughout the dataset when

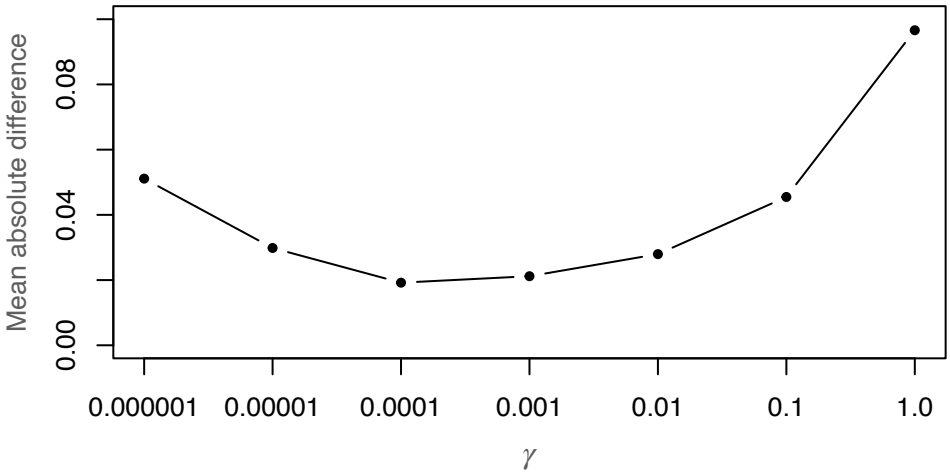


Figure 4.17: **Difference between Bayesian model and manual classification is minimised at $\gamma = 0.0001$**. The mean absolute difference (MAD) between the Bayesian proportion of difference and the manual classification calculated across all samples and proteins within the dataset was calculated for varying values of γ.

checking 0.0 lay within the 95% HDI of their posterior difference.

4.5 Discussion

The proportion of patient myofibres which are not-like-control (assumed to have some OXPHOS defect) is an important tool for measuring the pathological progression of mitochondrial disease or assessing the effect of treatments over time. In addition, robustly classifying individual myofibres as like-control allows their direct comparison for further investigation to learn other differences associated with OXPHOS dysfunction.

The Bayesian model, proposed here, can be fitted by the `R` package, `analysis2Dmito`, available on GitHub (https://github.com/jordanbchilds/analysis2Dmito). The package contains functions to infer model parameters and visually asses model output. Prior construction is automated using control subject data provided by the user and the pipeline described in Section 4.3.2. However, all (hyper-)prior parameters can be selected manually if desired.

4.5.1 Key findings

A new model was proposed to identify single myofibres with OXPHOS deficiency using protein abundances was proposed. A hierarchical mixture model is used to account for the natural inter-subject variability seen within patients and to infer the latent state of myofibre classification. The Bayesian paradigm is well suited to classification tasks and complex hierarchical modelling, inferring uncertainty in the posterior classification probabilities.

The Bayesian and frequentist models have been tested on the same dataset and compared against expert manual classification of the myofibres, and the Bayesian model was

shown to identify manually classified like-control myofibre populations, which the frequentist model missed. Additionally, the frequentist model was shown to produce many false-positive classifications, resulting in overestimation of the proportion of not-like-control myofibres within a patient sample. The Bayesian model's posterior beliefs for the proportion of not-like-control myofibres show substantially higher agreement with the manual classifications, achieving a 2% error rate across the Vincent *et al.* dataset.

The models' performances were further assessed on two synthetic datasets, where the Bayesian model showed a higher agreement with the ground-truth OXPHOS status of single-myofibres, as well as the ground-truth not-like-control proportions. The model's susceptibility to over-fitting was tested by splitting a synthetic dataset into training and validation subsets. The posterior myofibre classifications showed no substantial difference when the model parameters were inferred using the whole dataset, underscoring the predictive power of the model.

Prior beliefs are often difficult to elicit from experts when model parameters are abstract. Here, a prior construction pipeline which allows prior beliefs to be informed from previous experiments or control subject data was proposed and implemented on all datasets. The model's sensitivity to the chosen prior parameters was investigated by inflating and deflating prior uncertainty and the model proved to be fairly insensitive to the prior specification. The prior construction pipeline was applied to the synthetic datasets, where the model successfully retrieved the ground-truth parameter values.

Although mtDNA dynamics are not directly observed in OXPHOS abundance data, information can be gained about their dynamics indirectly. Robust and accurate estimation of the proportion of myofibres with an OXPHOS defect also yields robust and accurate estimates for the proportion of myofibres whose variant load has passed the pathogenic threshold, assuming the biochemical threshold theory. Hence, it would be possible to use such datasets in the investigation of clonal expansion.

4.5.2 Limitations

The Bayesian model requires healthy patient myofibres to show a strong linear relationship in (log) OXPHOS protein and VDAC abundances, similar to control subject data, and be distinct from the OXPHOS deficient patient myofibres. When this is not the case, the model classifications are not reliable, as seen in patient P03 of the Gomes dataset. Additionally, the Gomes dataset highlighted the complex nature of the posterior distribution and the resulting difficulties in sampling from it. Some inference chains become stuck at local maxima, which are difficult to escape from. In theory, longer runs of the inference scheme would solve this issue. However, in practice, this could incur a large computational cost and require the analysis of the chain output. Both of which may be challenging to overcome when non-statistical or computational researchers implement the model.

The Vincent *et al.* dataset contained no patient whose proportion of not-like-control myofibres was greater than 50%. A majority of not-like-control patient myofibres would likely result in the linear model fitting to the not-like-control population. Therefore, the

classification is not reliable for patients with a majority of OXPHOS deficient myofibres. In a diagnostic context, the proportion of not-like-control myofibres would not be known *a priori* but would be visible during exploratory data analysis. Nevertheless, for patients with a majority like-control myofibre population, the approach seems to closely match the expert manual classifications.

Using OXPHOS abundance data to investigate clonal expansion poses some significant challenges. Clonal expansion is a temporal process whose starting conditions drastically affect the resulting dynamics. Estimating the inherited variant load from a patient is challenging due to the bottleneck effect, and ethical considerations make it difficult to collect tissue samples from a single patient at multiple time points. If this were to be done, samples would likely be collected from different tissues, which have tissue-specific effects on mtDNA dynamics. In addition, uncertainty in model parameters such as mtDNA replication and mutation rates, and the biochemical threshold, means that parameter fixing is unsatisfactory, and parameter inference is difficult.

4.5.3 Future work

To address the issue of convergence, seen in samples P2-TA-B2-S1 and P2-TA-B2-S2 of the Gomes dataset, different inference schemes should be proposed and tested to ensure a fast exploration of the parameter space and an ability to escape local maxima of the posterior within a reasonable time. Another significant issue persists: patients must show a specific type of OXPHOS abundance. Namely, like-control myofibres show a linear relationship distinct from the not-like-control myofibre abundances, and the proportion of OXPHOS deficient myofibres be less than 50%. Although this data type is common, seen in all patients of the Vincent *et al.* dataset, P01 and P02 of the Gomes dataset, and the dataset presented in Warren et al. (2020), other OXPHOS abundances prolifes exist, such as that in P03 of the Gomes data. When this is the case, the Bayesian model's results are unreliable. To robustly classify such data, it is likely that a new model will be needed, one that views the data in a different manner, i.e. not a 2Dmito plot. The problem is exacerbated by the small number of control subjects within each experiment; however, ethical and financial constraints make it unlikely that this will improve. What may help the classification of OXPHOS deficient myofibres is aggregating datasets together. However, this brings its own problems; collection methods alter the scale, and the formatting of data is inconsistent. Not to mention problems such as data privacy, an unwillingness to share, or data being lost. Nevertheless, if a large aggregate dataset were to be constructed, a Bayesian hierarchical model, not necessarily the one presented here, is well-suited to account for inter-experimental variation.

The inferred proportion of OXPHOS deficiency could be used in an investigation into clonal expansion, where the proportion serves as a data point used to fit a mathematical model. The model would have to simulate the mtDNA population dynamics of individual myofibres and estimate the deficiency proportion by a number of simulations, comparing variant loads to a pathogenic threshold. Parameter inference for both the mathematical model and the Bayesian hierarchical model could be inferred via a single, large inference scheme. However, this would require considerable work and significant computational expense. A more pragmatic approach would be to use the posterior expectations or manual

classifications, if available.

Chapter 5

Modelling Clonal Expansion

Following Chapter 4, here OXPHOS deficiency data is used to compare models of clonal expansion. The steps required to develop a new agent-based, spatially dependent model and the practicalities of inferring model parameters with real and synthetic datasets are considered.

5.1 Introduction

As discussed in Chapter 1, the mechanisms governing many of the cellular processes controlling mtDNA dynamics are unknown, adding uncertainty to the driving factors of clonal expansion. Therefore, developing therapies and treatments to prevent its progression is extremely difficult. Several studies have proposed and investigated different clonal expansion theories using experimental results and mathematical models. However, the high degree of uncertainty within all aspects of the biological mechanisms leads to high uncertainty in the mathematical models and the parameters which govern them. Some studies have gone further to compare predictions from different models, but, to the author's knowledge, none have inferred model parameter values as well; instead, they favour the use of experimentally-derived estimates found in the literature. As will be shown here, a range of parameter values are capable of replicating the dynamics within a single dataset, and, therefore, parameter fixing is undesirable. To overcome this, a comprehensive investigation that accounts for the parameter uncertainty and formally compares clonal expansion theories is needed. Bayesian inference is well-suited to such tasks due to its ability to handle complex model structures, intractable likelihoods, and the propagation of parameter uncertainty. Bayesian methods also work well with model selection methods, such as Bayesian information criterion (BIC), which can account for parameter uncertainty and provide quantitative justifications for specific models.

Unfortunately, such an investigation could not be done due to this project's time and resource limitations. Instead, the practicalities of using OXPHOS deficiency proportion data, like that seen in Chapter 4, are investigated to compare mathematical models of clonal expansion. Chapter 5.3 discusses the mathematical and statistical details of the models and comparison techniques. The remainder of this section discusses existing work using mathematical modelling and related parameter estimates found in the literature.

5.1.1 Random genetic drift

RGD assumes no replicative or degradative advantage between wild-type and variant mtDNA and, therefore, it assumes equal reaction rates or probabilities between the species (Chinnery & Samuels, 1999). Since its introduction, the model has become very popular within the literature, often being used as a neutral model to compare against or as a basis for investigating other mechanisms, such as copy number control (Capps et al., 2003; Hoitzing, 2017) or the genetic bottleneck (Johnston et al., 2015).

The earliest mathematical models of clonal expansion reflected the assumptions of RGD. Chinnery and Samuels (1999) investigated the probability of reaching variant and wild-type homoplasmy within post-mitotic cells for patients with inherited variant loads and mtDNA maintenance disorders. They developed a discrete-time stochastic model to simulate mtDNA dynamics and estimate the proportion of cells reaching homoplasmy for various starting conditions and model parameters. Their model's replication rate was deterministic, with the number of mtDNA replications within a single time-step calculated depending on the system's current state. However, the degradation of molecules was stochastic, following a Poisson process (Chinnery & Samuels, 1999). Elson et al. (2001) also executed discrete-time stochastic simulations to show that post-mitotic cells in healthy patients could reach COX deficient status throughout the human life-span. In their work, Elson *et al.* showed that the rate of clonal expansion is dependent on copy number, with high copy numbers slowing mtDNA dynamics. Coller et al. (2001) modelled mtDNA dynamics in mitotic cells, while investigating the mtDNA mutations in the colonic crypt. Their model differed slightly, focusing on replenishing mtDNA between cellular divisions and simulating the number of variant mtDNA that arise before the next division from a Poisson distribution, rather than randomly simulating each mtDNA replication.

Few works have focused on inferring the mathematical model parameters, and instead have used values from the literature. Henderson et al. (2009) developed a continuous-time stochastic model reflecting RGD for mtDNA within neurons and implemented Bayesian inference to infer model parameters (Henderson et al., 2009). However, they found the computational cost of simulating from the model via the Gillespie algorithm (or approximation algorithms) too great for inference, and therefore implemented an emulator that inferred the mathematical model output based on parameter value inputs. The emulator is far less computationally expensive to sample from and so, once built, can massively reduce inference time. Later work by Henderson et al. (2010) combined stochastic simulation and a statistical model, which can use multiple data sources, again inferring parameter values by Bayesian inference and an emulator. In both cases, the inferred parameters showed decreased uncertainty *a posteriori* and the predictive distributions showed a strong resemblance to the data.

Some works show contradictory evidence, suggesting that RGD is not an appropriate model for mtDNA dynamics. Kowald and Kirkwood (2013) developed a discrete-time mathematical model of mtDNA dynamics reflecting RGD to simulate the proportion COX deficient cells. They showed that to obtain the observed proportion of COX deficiency for short-lived species, the mutation probability must be increased by a factor of 200, compared to the probability when recreating data for long-lived species. However, in their investigation, all other parameters (mtDNA half-life and copy number) remained

unchanged, which may not be appropriate given the species-specific nature of their work.

Aryaman et al. (2019) used a neutral model of mtDNA dynamics to assess the impact of mitochondrial network fragmentation on variant load mean and heteroplasmy. They developed a continuous-time, discrete state-space model, simulated by the Gillespie algorithm (Gillespie, 1977), with parameter values found in the literature. Johnston et al. (2015) used a neutral model of mtDNA population dynamics as part of a larger model to investigate the mechanisms of the genetic bottleneck. Varying mechanisms of mtDNA segregation during cellular division were considered, although none distinguished between mtDNA species. Single-cell mtDNA population dynamics were simulated by the Gillespie algorithm (Gillespie, 1977). Using a dataset collected from mouse tissue, parameter values were inferred using approximate Bayesian computation (Sunnåker et al., 2013). Hoitzing et al. (2019) investigated the impact of mtDNA populations on the cell's energy requirements, using the Gillespie algorithm to simulate from model of random drift.

5.1.2 Survival of the smallest

As discussed in Chapter 1.2, the survival of the smallest hypothesis assumes that the smaller size of mtDNA with single, large-scale deletions yields a shorter replication time and thus, a replicative advantage. However, this does not explain the increase in single-point variants observed (Greaves et al., 2014; Weber et al., 1997). Nevertheless, a variant mtDNA replicative advantage theory persists in the literature to explain clonal expansion, although its biological reasoning remains unclear (Insalata et al., 2022; Johnston et al., 2015).

Kowald et al. (2014) developed an agent-based, discrete-time mathematical model to investigate the impact of a reduced replication time in variant mtDNA. Their work considered mtDNA to be in two states: free, non-replicating, or busy, replicating. Replication times were fixed, but not equal, for the two species, with variant mtDNA molecules replicating at a specified fraction of wild-type's. Mutation events occurred during replication of wild-type molecules with a defined probability, producing one variant and one wild-type mtDNA molecule. Using parameters from the literature, Kowald *et al.* showed that mtDNA half-life (and turnover) is a key factor in mitochondrial dynamics, where increased mitochondrial turnover increases the rate of clonal expansion. From their simulations, Kowald *et al.* concluded that a reduction in replication time is unlikely to be the biological phenomenon driving clonal expansion, due to the much longer time required to reach observed levels of COX deficiency (Kowald et al., 2014). However, the acknowledged uncertainty in their modelling parameters does not conclusively rule out the model. Later work by Kowald and Kirkwood (2014) combined a variant mtDNA replicative advantage and replication feedback mechanism based on ATP levels. The model is further discussed in Chapter 5.1.4. Other studies have included a model of replicative advantage as a comparison, such as (Insalata et al., 2022), also discussed in Chapter 5.1.4.

5.1.3 Perinuclear niche hypothesis

Vincent et al. (2018) investigated subcellular areas of OXPHOS deficiency within myofibres, finding that localised OXPHOS deficiency was limited to the perinuclear niche and hypothesised that mtDNA replication is increased within the region due to its resource abundance. The increased replication increases the rate of variant accumulation due to mtDNA replication error (Kowald et al., 2014), resulting in localised OXPHOS deficiency. Vincent *et al.* also showed that partial OXPHOS deficiency in the cross-sectional direction is less common than partial OXPHOS deficiency in the longitudinal direction, and therefore, summised that mtDNA spread is slower in the longitudinal direction. Later work showed that the mitochondrial network is more highly connected in the myofibre cross section (Vincent et al., 2019), allowing for the faster spread of mtDNA. The mitochondrial network of myofibres is critical to the proposed theory. Other cell types do not have the same network structure, and smaller sizes do not allow for significant spread of variant mtDNA and OXPHOS deficiency. Therefore, the perinuclear niche hypothesis is primarily considered within myofibres.

To the author's knowledge, a mathematical model of the perinuclear niche theory has yet to be developed. This is likely because the theory was proposed much more recently than the other theories discussed in this chapter, published in 2018 (Vincent et al., 2018). Another reason which may have prevented its development is the spatially dependent nature of the hypothesis. Spatial dependence brings added complexity and computational cost, which may hinder its practical application. An agent-based mathematical model reflecting the assumptions of the perinuclear niche theory is developed in Chapter 5.1.3.

5.1.4 Other notable models

Free radical models

Mao *et al.* proposed a model of mitochondrial dynamics (not mtDNA), which included the effect of free radicals (ROS) on mtDNA mutation rate, mitochondrial membrane potential, and mtDNA replication rates (Mao et al., 2006). The model was expressed as a series of ordered differential equations (ODEs). ODEs are fully deterministic and ignore random fluctuations, which are significant for small populations such as those seen when a variant mtDNA population first arises. Nevertheless, the model was compared to data observed in mice and showed a good fit for variant mtDNA accumulation. Where appropriate, parameters for the model were taken from the literature, and an estimate of the mutation rate was found by optimisation.

ATP models

Kowald *et al.* proposed a mathematical model which included ATP production and use, and a replicative advantage for variant mtDNA (Kowald & Kirkwood, 2014). Discrete-time stochastic simulations of the model were compared to those from Kowald's models of RGD (Kowald & Kirkwood, 2013) and survival of the smallest (SoS) (Kowald et al., 2014). Model parameters were taken from the literature or derived as appropriate. Although it appears that the replication rate within this model was taken from the mtDNA

replication time and did not consider the time between replications. The increased mitochondrial turnover would increase the model's clonal expansion rate. However, a fair comparison would be achieved if the other models were executed with the same parameters. Their model was able to predict realistic levels of COX deficiency for short- and long-lived species, whereas their models of RGD and SoS could not. Parameters were kept consistent between simulations. However, similarly to before, the investigation did not consider the differences in parameter values between species. The probability of mutation for each model was chosen so that the model simulations resembled realistic output, 10% COX deficiency at various ages. The parameters relating to ATP generation and use were given no justification from the literature, and only those values were considered. Nevertheless, the ATP dependent model showed a better fit to observed results, such as the number of distinct variant mtDNA species and lower mutation probabilities.

Spatially dependent models

Insalata et al. (2022) proposed the stochastic-survival-of-the-densest model of clonal expansion. The model focuses on the spatial dynamics of mtDNA and how mitochondrial dysfunction can spread throughout a myofibre (Elson et al., 2002; Vincent et al., 2018). Much like the perinuclear niche hypothesis, the model is primarily concerned with myofibres, where the spread of mitochondrial dysfunction can be seen due to their large size and post-mitotic nature. The authors employ a spatial Gillespie model, which models the myofibre as a series of compartments, each controlled by a single nucleus, which mtDNA can migrate between. No species-specific replicative advantage was assumed; therefore, the model was essentially a spatial extension of RGD. Parameter inference was not done, and parameter values were taken from the literature. Their proposed model more closely estimated the rate of variant mtDNA spread than a model of variant mtDNA replicative advantage.

5.1.5 Copy number control

Healthy cells without pathogenic variant mtDNA, whose energy requirements are approximately constant, should have a stable copy number as the cell maintains an equilibrium state (Clayton, 1996). The biological mechanisms governing copy number control remain largely unknown. As such, various mathematical models of copy number control have been used throughout the literature. Implementing such controls in a mathematical model is important to prevent the system copy number from reaching unrealistic values, and, in some cases, reflect observed change in copy number seen in OXPHOS deficient myofibres (Grady et al., 2018).

Strict control

The simplest copy number control mechanism strictly controls the population to remain exactly at a target value. The copy number is maintained by ensuring that an mtDNA degradation event is immediately followed by replication (Elson et al., 2001; Kowald & Kirkwood, 2013). Although not biologically realistic, the method is convenient in the

absence of other information.

Feedback control

Mechanisms which allow copy number to vary are more common and have been used since the earliest mathematical models of mtDNA dynamics. Maintaining an approximately constant copy number can be achieved by altering mtDNA replication or degradation rates based on the current state of the myofibre. By decreasing the replication rate relative to the degradation rate, the copy number naturally decreases over time and vice versa. It is usually assumed that when the copy number equals its target value, the replication and degradation rates are equal; otherwise, the copy number would tend to increase or decrease. This will be the case for all mechanisms discussed here unless otherwise indicated.

The most commonly used mechanism considers the degradation rate constant and alters the replication rate as needed. A base replication rate, equal to degradation, is linearly altered by an amount proportional to the difference between the current and target copy numbers. Let $\tilde{k}(W, V)$ be the mtDNA replication rate and W and V be the current wild-type and variant population sizes. The mechanism can be written as

$$\tilde{k}(W, V) = k_0 + k_1 \left[C_0 - (W + V)\right], \tag{5.1}$$

where k_0 is the base replication rate, equal to the degradation rate, k_1 is a parameter determining the strength of the controller, and C_0 is the target copy number. A slight variation on this mechanism includes a variant-mtDNA detection parameter, $0 \leq \delta \leq 1$. In this case, the replication rate $\tilde{k}(W, V)$ can be written as

$$\tilde{k}(W, V) = k_0 + k_1 \left[C_0 - (W + \delta V)\right]. \tag{5.2}$$

The variant detection parameter allows the copy number to increase with the rise of variant mtDNA. At variant homoplasy, the equilibrium copy number becomes C_0/δ, recall $\delta \leq 1.0$. Although it is referred to by different names throughout the literature, it is commonly used (Aryaman et al., 2017; Capps et al., 2003; Hoitzing et al., 2019; Insalata et al., 2022).

Chinnery and Samuels (1999) set the mtDNA replication rate to be a function of wild-type population size, assuming that variant mtDNA did not affect the control mechanism. Unlike the linear feedback mechanism in Equation 5.2, their method altered a base replication rate multiplicatively. The Chinnery *et al.* mechanism allowed the replication rate to reach a maximum value when wild-type mtDNA became extinct and depended on the target copy number, mtDNA half-life and a control parameter. Capps et al. (2003) extended this control mechanism to include the variant mtDNA detection parameter and stop mtDNA replication when the copy number reached a defined maximum capacity. The Capps replication rate function is described in Equation 5.3, where $C_{\max}$ is the maximum copy number in the system. The wild-type dependent control mechanism of Chinnery *et al.* can be achieved by setting $\delta = 0$ in Equation 5.3 (Chinnery & Samuels, 1999). Thus,

$$\tilde{k}(W, V) = \begin{cases} k_0 \left[k_1 - (k_1 - 1)\frac{W + \delta V}{C_0}\right], & W + \delta V \leq C_{\max} \\ 0, & W + \delta V > C_{\max} \end{cases} \tag{5.3}$$

Kowald and Kirkwood (2014) assumed that mtDNA copy number is dependent on the cellular ATP level, a concept previously introduced by Clay Montier et al. (2009). By altering the copy number in response to the ATP levels, the copy number dynamically reacts to the cellular requirements. Unfortunately, the model was not compared to ATP-related data. Kowald *et al.* induced copy number control by making the stochastic replication rate inversely proportional to copy number, giving

$$\begin{aligned} \tilde{k}_W(W,V) &= \frac{k_W}{1+ATP/k_1}, \\ \tilde{k}_V(W,V) &= \frac{k_V}{1+ATP/k_1}. \end{aligned} \tag{5.4}$$

Where k_W and k_V are the maximum replication rates of each mtDNA species, achieved when no ATP is present within the myofibre, the parameter k_1 controls the ATP's effect on replication rate. In their model, ATP was produced at a constant rate proportional to the wild-type copy number. Within the system, ATP was degraded proportional to the weighted sum of the cellular copy number and ATP population, where the weights are model parameters.

5.1.6 Parameter values

MtDNA half-life

Although most mathematical models are not parameterised to include half-life, favouring reaction rates or probabilities for a given time-step, half-life is the most common measurement in biological experiments for mtDNA turnover. Fortunately, the transformation between the two is trivial, and works often quote half-life estimates to justify reaction rates. Therefore, discussion of mtDNA degradation rate and half-life is interchangeable and will be considered as such throughout this thesis. Early works simulating mathematical models of clonal expansion largely informed parameter values used in models decades later. Chinnery and Samuels (1999) used a variety of mtDNA half-lives, ranging from $1-10$d, which was followed by Elson et al. (2001) who used an mtDNA half-life of 10d. This value has since become ubiquitous within the literature with many studies using it (Capps et al., 2003; Hoitzing et al., 2019; Insalata et al., 2022; Johnston & Jones, 2016; Kowald & Kirkwood, 2013; Kowald et al., 2014). This half-life estimate originates from tissue-specific, cell-type aggregate measurements made using a rat model (Gross et al., 1968), see Table 5.1. It is clear from the work of Gross *et al.* that mtDNA half-life is tissue-specific and should, therefore, be considered as such. Significantly for this chapter, none of these estimates are made within skeletal muscle. To the author's knowledge, only two estimates of mtDNA half-life in skeletal muscle exist. Both estimates were made using an animal model and produced starkly different half-life values. Korr et al. (1998) estimated half-life in mouse skeletal muscle to be 17.7d, and later Collins et al. (2003) produced an estimate of 700d using a rat model. Both studies include estimates in other tissues, and Korr et al. (1998) produced several cell-type-specific estimates, as shown in Table 5.1.

Tissue	Half-life (days)		
	Gross *et al.*	Korr *et al.*	Collins *et al.*
Kidney	10.4 ± 1.2	$[13.1, 17.7]$	-
Liver	9.4 ± 0.7	8.4	-
Brain	31.0 ± 16.0	$[11.6, 23.2]$	-
Heart	6.7 ± 1.0	-	350
Skeletal muscle	-	17.7	700

Table 5.1: **MtDNA half-life estimates**. MtDNA half-life estimates in a number of tissues from three different studies. Gross *et al.* and Collins *et al.* used a rat model and made estimates from aggregated tissue samples (Collins et al., 2003; Gross et al., 1969). Korr *et al.* used a mouse model and made tissue and cell-type-specific estimates in a number of tissues (Korr et al., 1998). The range of estimates is given where half-life was estimated in multiple cell types per tissue.

Henderson et al. (2009) inferred the model parameters of a mathematical model of random drift, variant load data collected within brain tissue of patients with Parkinson's disease. Their posterior expected value for the rate of mtDNA degradation and base replication was 1.27×10^{-7} s^{-1}, with the 95% credible interval being $(7.13 \times 10^{-8}, 5.83 \times 10^{-4})$. Unfortunately, the full posteriors are not given in the paper; as such, the posterior half-life distribution cannot be found. Instead, the posterior expectation and interval are transformed, which are 63.17d and $(0.0138, 112.518)$. The posterior expectation is much higher than experimental estimates found by Gross et al. (1969) and Korr et al. (1998), but the posterior uncertainty is fairly large, and all experimental values are within the approximate 95% posterior probability.

Copy number

Elson et al. (2001) employed a range of copy numbers, $100 - 10,000$, to observe their effect on the rate of clonal expansion, and used a copy number of 1,000 when inspecting the effect of other parameters. A wild-type equilibrium copy number of 1,000 is now common in the literature, for both strict and feedback control mechanisms when simulating post-mitotic cells (Capps et al., 2003; Chinnery & Samuels, 1999; Henderson et al., 2009; Kowald & Kirkwood, 2013, 2014; Kowald et al., 2014; Mao et al., 2006).

Miller et al. (2003) estimated mtDNA copy number within myofibres of healthy subjects to be 3650 ± 620 per diploid nuclear genome, and showed that this is significantly lower than the copy number per diploid nuclear genome in the myocardium. This value was later used to inform the copy number used by Insalata et al. (2022). Bruusgaard et al. (2003) investigated the number and distribution of nuclei within myofibres, the inter-nuclei-distance in the extensor digitorum longus muscle was found to be ≈ 31µm. Using these values, the equilibrium wild-type copy number for a section or whole myofibre can be estimated.

Pathogenic threshold

As discussed in Chapter 1.1.9, several studies have investigated and estimated the value of the pathogenic threshold, and it is clear from their work that its value is cell-type and

variant-specific. Within mathematical modelling, work which directly incorporates the pathogenic threshold to simulate OXPHOS deficiency has generally used a single value, 60% (Hayashi et al., 1991), (Elson et al., 2001; Kowald & Kirkwood, 2013, 2014; Kowald et al., 2014; Lakshmanan et al., 2018).

Mutation probability

Elson et al. (2001) tested a range of mutation probabilities in their work, 1.0×10^{-6} to 1.0×10^{-4} per wild-type mtDNA replication event, and showed that a probability of 1.0×10^{-6} allows variant mtDNA to clonally expand to pathogenic levels throughout a human life-span. Similar probabilities were used by Kowald and Kirkwood (2013), and later work by Kowald *et al.* varied the mutation probability to reflect observed data in short- and long-lived animals (Kowald & Kirkwood, 2014). Mathematical models of random drift and replicative advantage required mutation probabilities of 1×10^{-2} to 1×10^{-5} to produce realistic levels of OXPHOS deficiency. The mathematical model including ATP production, proposed in Kowald and Kirkwood (2014), required mutation probabilities of approximately 1×10^{-5} to 1×10^{-7}.

Henderson et al. (2009) fitted a mathematical model, with the assumptions of RGD, to variant load data from neurons of Parkinson's disease patients. Unfortunately, mutation probability was not a parameter in their study. However, their model included the mutation event rate and a parameter which resembles the base replication rate. A crude posterior estimate of the mutation probability can be calculated as the expected posterior mutation rate divided by the sum of the expected posterior mutation and replication rates. This finds the approximate expected posterior mutation probability to be 0.00344. Mao et al. (2006) combined experimental data from healthy mice and a mathematical model to estimate the mutation rate of mtDNA. Tissue samples were collected in the brain and liver; their approximate mutation probabilities were 8.18×10^{-7} and 3.88×10^{-8}, respectively. Shenkar et al. (1996) developed a statistical technique to estimate the mutation probability of a genome and applied this to a mtDNA deletion mutation. Data collected from fibroblast cybrid cells was enucleated with cytochalasin B and fused to the mtDNA-less celline $\rho^0$143B.206 (King & Attardi, 1989). Their method estimates the probability of mutation as 5.95×10^{-8} with a conservative confidence interval of $(4.60 \times 10^{-9}, 1.07 \times 10^{-7})$.

Variant mtDNA replicative advantage

Kowald et al. (2014) assumed that the time reduction of the variant replication was proportional to the size of the single deletion, and simulated deletion sizes from 50% to 70% of the molecule, and wild-type mtDNA replication was assumed to take 2 hours (Berk & Calyton, 1974). Holt and Davies (2022) also chose to reduce variant mtDNA replication time proportionally to the deletion size. In the context of stochastic models, the increased replication rates for wild-type and variant mtDNA can be estimated. Assuming an mtDNA half-life of 10d and that mtDNA replication and degradation are in equilibrium, maintaining a stable copy number, the expected time between replications of a single molecule is approximately 14d. A 50% reduction in variant mtDNA size increases the replication rate by approximately 0.3%. Increasing half-life to 17.7d, the estimate for

skeletal muscle, the variant mtDNA replication rate is increased approximately 0.1%.

Insalata et al. (2022) compared their proposed model to a mathematical model of replicative advantage. An increased replication rate for variant mtDNA was achieved by an additive factor of 2.4d^{-1}. This value was calculated to ensure a desired increase in mtDNA copy number within the cell. However, the factor is 34 times larger than the base wild-type replication rate, when the system is in equilibrium, 0.07d^{-1}.

5.1.7 Aims of investigation

This thesis investigates whether meaningful inference can be made about the parameters governing a mathematical model of clonal expansion using single observations of the proportion of deficient myofibres within patients with mtDNA maintenance disorders. Secondly, whether a mathematical model can be distinguished from similar data using model comparison techniques and Bayesian inference is considered.

To this end, three mathematical models of mtDNA dynamics are proposed: RGD, SoS, and PNN, and a Bayesian inference scheme is constructed to infer parameter values given a dataset of OXPHOS deficiency proportions. Due to time and resource restrictions, only the parameters of the RGD model are inferred. However, they are inferred given several datasets. Firstly, the parameters are inferred given an observed dataset, to ascertain posterior uncertainty for small (realistically sized) single datasets. Secondly, the parameters are independently inferred given three synthetic datasets, generated from the three mathematical models. This is to test whether a model of random genetic drift can suitably fit OXPHOS deficient data whose true underlying mtDNA dynamics are known.

5.2 Data

The dataset under consideration in this chapter is the OXPHOS deficiency data collected by Vincent et al. (2024), and previously introduced in Chapter 4. Recall that the dataset contains skeletal muscle fibre samples from 11 patients with mtDNA maintenance-related mitochondrial disease. Unfortunately, the age at which the tissue sample was taken was not available for patients P11 and P12, and so these patients were removed from the dataset. The proportion of OXPHOS deficient myofibres for each patient was taken to be the estimate from the manual classification of the samples' 2Dmito plot, discussed in Chapter 4.2.1, each patient, therefore, has one observation per OXPHOS protein.

This dataset was chosen over the Gomes dataset, also introduced in Chapter 4, or other datasets, as the patient phenotype (mtDNA maintenance disorders) allows the assumption that at birth each patient's myofibres have a zero variant load. Although it is possible that variant mtDNA arises during fetal development, it is assumed that there is not sufficient time for variant mtDNA to clonally expand before birth. The assumption is key to being able to infer parameter values because it assumes an initial state of wild-type homoplasmy for the mathematical models. Without this assumption, it is likely that longitudinal data would be required due to the variety in population dynamics, which are

possible with varying parameter rates, within a single mathematical model.

In developing the mathematical model of the perinuclear niche theory, a dataset was used to inform the modelling space. The data used is a single myofibre, segmented by Mitocyto (Warren et al., 2020). The sample was collected from a healthy male control subject, aged 20, following ethical approval through the Newcastle biobank under the application number 042 (17/NE/0361).

5.3 Methods

In this section, the details of mathematical models and their simulation methods for each theory of clonal expansion are described, as well as the statistical model connecting the mathematical models to the data. Three synthetic datasets are generated as part of the investigation; their generation and ground-truth parameters are also discussed. Lastly, the section comments on the computational cost and practical implications of the Bayesian inference for model inference with this data type, OXPHOS deficiency proportions.

The discussion of the perinuclear niche model is separate to the discussion of the RGD and SoS models, as these are relatively simple to simulate. Having no spatial dependence, both models can be implemented using stochastic kinetic models and simulated via the Gillespie algorithm, introduced in Chapter 2.6. The spatial dependence of the perinuclear nuche theory means its model development is more involved, and each aspect of model development is discussed in a later section.

5.3.1 Stochastic kinetic models

Stochastic kinetic models can simulate both RGD and SoS theories of clonal expansion, as neither theory assumes spatial dependence on mtDNA behaviour. The mitochondrial network is assumed to be well-connected enough that freely moving mtDNA molecules are an appropriate approximation. Therefore, the two theories can be simulated using Gillespie's direct method or one of its approximations, discussed in Chapter 2.6.

Random genetic drift

A mathematical model of random genetic drift was introduced in Chapter 2.6.6, although the model did not include the possibility of *de novo* mutations. Allowing *de novo* mutations here is crucial because the dataset consists of patients with mtDNA maintenance disorders. Therefore, an additional reaction is added to the system, which allows *de novo* mutations to arise. Variant mtDNA is assumed to arise from a slippage event during wild-type mtDNA replication (Guo et al., 2010; Shoffner et al., 1989). To reduce model complexity, all mtDNA variants are considered to be the same species within the model. The model considers two species: wild-type, W, and pathogenic variant, V, with a total of five reactions. Both species can replicate and degrade, and a mutation event during wild-type replication is assumed to lead to one molecule of each species. The pseudo chemical reaction equations describing the system are

$$
\begin{aligned}
W &\xrightarrow{k_1} 2W, \\
V &\xrightarrow{k_2} 2V, \\
W &\xrightarrow{k_3} \phi, \\
V &\xrightarrow{k_4} \phi, \\
W &\xrightarrow{k_5} W + V.
\end{aligned} \tag{5.5}
$$

RGD assumes no replicative advantage between species, so replication and degradation rates must be equal when the system is in equilibrium. Due to the assumption that a mutation occurs during wild-type replication, the reaction rates of the wild-type replication and mutation must be considered together, and their sum must equal the rate of variant replication. Otherwise, an implicit replicative advantage would be present in the model. It may be more natural to parameterise the reaction rates using a general replication rate for both species, k_{rep}, and a mutation probability, ρ. Not including a copy number control mechanism, the stochastic rate constant for a successful wild-type replication is then $k_1 = (1-\rho)k_{\text{rep}}$, and the rate constant for a mutation is $k_5 = \rho k_{\text{rep}}$. Under the assumptions discussed in Chapter 2.6.1, this parameterisation implies that a wild-type and variant replication occurs with equal rates and that a wild-type molecule is successfully replicated with probability $(1-\rho)$, otherwise a mutation event occurs.

Copy number control and OXPHOS deficiency are discussed in the coming sections, as the methods are relevant to both RGD and SoS models. Let the copy number controller additively alter the general mtDNA replication rate, k_{rep}, of both species based on the system's current state. The controller is defined by a function, $\tilde{k}(W, V)$. The reactions and their rates become:

$$
\begin{aligned}
W \to 2W \quad & (1-\rho)\left[k_{\text{rep}} + \tilde{k}(W,V)\right], \\
V \to 2V \quad & k_{\text{rep}} + \tilde{k}(W,V), \\
W \to \phi \quad & k_{\text{deg}}, \\
V \to \phi \quad & k_{\text{deg}} \\
W \to W + V \quad & \rho\left[k_{\text{rep}} + \tilde{k}(W,V)\right].
\end{aligned}
$$

Not including parameters of the copy number control mechanism, the mathematical model of random genetic drift, therefore, contains three parameters: general mtDNA replication rate, k_{rep}, mtDNA degradation rate, k_{deg}, and mutation probability, ρ. It is assumed that the cell is in a state of equilibrium, such that when the current copy number is equal to the target copy number, the general replication and degradation rates are equal. Therefore, it can also be assumed that $k_{\text{rep}} = k_{\text{deg}} = k_0$. The parameter k_0 is referred to as the base reaction rate throughout the rest of the thesis. The parameter controls mtDNA turnover and overall reactiveness of the system. A high base reaction rate increases the number of reactions within the system, and the reaction rates can be rewritten, such that

$$
\begin{aligned}
W \to 2W \quad & (1-\rho)\left[k_0 + \tilde{k}(W,V)\right], \\
V \to 2V \quad & k_0 + \tilde{k}(W,V), \\
W \to \phi \quad & k_0, \\
V \to \phi \quad & k_0, \\
W \to W+V \quad & \rho\left[k_0 + \tilde{k}(W,V)\right].
\end{aligned}
$$

Survival of the smallest

SoS assumes a variant mtDNA replicative advantage, due to the physically smaller size of a single, large-scale deletion mtDNA molecule. Previous work modelled replication directly, considering mtDNA to be in replicative or free state (Kowald et al., 2014), here, however, a simpler model is used that is akin to the model of RGD previously described. A replicative advantage is achieved through an increased replication rate of variant mtDNA by a specified factor, γ. The model is defined by five equations, with the same parameterisation of a base reaction rate and probability of mutation as previously discussed. To impose a variant replicative advantage, γ is constrained to be greater than 1.0, otherwise a wild-type advantage would be imposed. The reaction network is described by the same set of equations as the RGD model, Equations 5.5, but the rate for variant mtDNA is altered,

$$
\begin{aligned}
W \to 2W \quad & (1-\rho)\left[k_0 + \tilde{k}(W,V)\right], \\
V \to 2V \quad & \gamma k_0 + \tilde{k}(W,V). \\
W \to \phi \quad & k_0, \\
V \to \phi \quad & k_0, \\
W \to W+V \quad & \rho\left[k_0 + \tilde{k}(W,V)\right].
\end{aligned}
$$

Copy number control

As discussed, a number of copy number control mechanisms have been proposed, and no controller has a consensus of agreement. As the data under consideration does not contain copy number information and copy number mechanisms are not investigated within this chapter, a linear feedback controller is chosen, Equation 5.6, follwing previous work in the literature (Aryaman et al., 2017; Capps et al., 2003; Hoitzing et al., 2019; Insalata et al., 2022). The controller alters the mtDNA replication rates proportional to the difference between the current copy number and a target copy number denoted C_0,

$$
\tilde{k}(C_0, W, V) = c_1 \left[C_0 - (W+V)\right]. \tag{5.6}
$$

The control mechanism adds two new parameters to the model: C_0, the target copy number, and c_1, which determines how tightly the copy number is controlled. Due to the lack of copy number data within the dataset, a constant, fixed C_0 is assumed, and a value of $C_0 = 1,000$ is chosen, following the literature (Capps et al., 2003; Chinnery & Samuels, 1999; Henderson et al., 2009; Kowald & Kirkwood, 2013, 2014; Kowald et al., 2014; Mao et al., 2006).

The controller parameter is assumed to depend on the replication rate, k_0, and is set to $k_1 = k_0/100$. This allows the copy number to be controlled relatively to the base reaction rate. Assuming independence between k_0 and k_1 could lead to negative reaction rates during inference, if an unusually low reaction rate and an unusually high controller parameter are sampled.

Simulation methods

The stochastic models of RGD and SoS are continuous-time, discrete state-space models and follow the assumptions of mass action kinetics, allowing them to be simulated by the Gillespie algorithm, Chapter 2.6. The computational expense of a single simulation is low, but the massive number of simulations required for inference means that runtime should be reduced as much as possible. Therefore, the τ-leap simulation method is used, Chapter 2.6.4. As discussed, the size of the time step is critical to the accuracy of the approximation. For a tightly controlled copy number, the overall reactivity of the system is relatively consistent, meaning a large step size should closely approximate an exact simulation method and considerably reduce computational cost. An adaptive step size is chosen, which leaps forward by approximately the expected time for one hundred reactions to occur within the system. Recall that the inter-reaction time is exponentially distributed, with rate equal to the total system hazard, $h_0(\boldsymbol{x}, \boldsymbol{k})$. Therefore the time step, Δt, is defined as

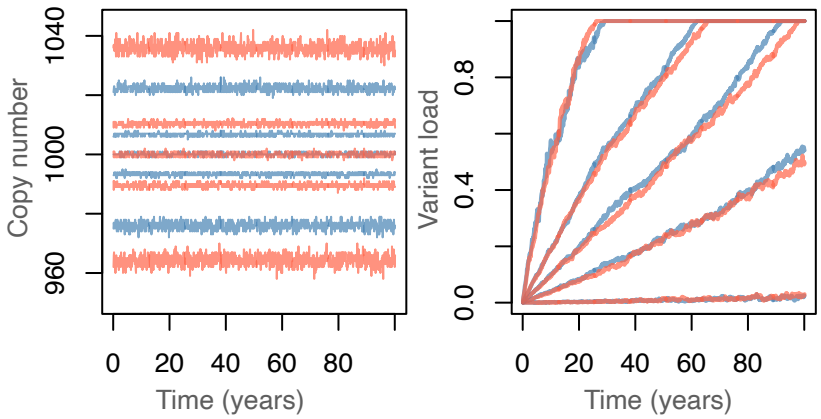
$$\Delta t = \frac{100}{h_0(\boldsymbol{x}, \boldsymbol{k})}. \tag{5.7}$$

Figure 5.1(a) compares the mtDNA dynamics of system exactly and approximately simulated. The τ-leap approximation shows a close approximation of the model's variant load. However, the use of the Poisson leap results in higher variance in the system's copy number. As copy number is not under consideration here, the approximation seems appropriate for our uses.

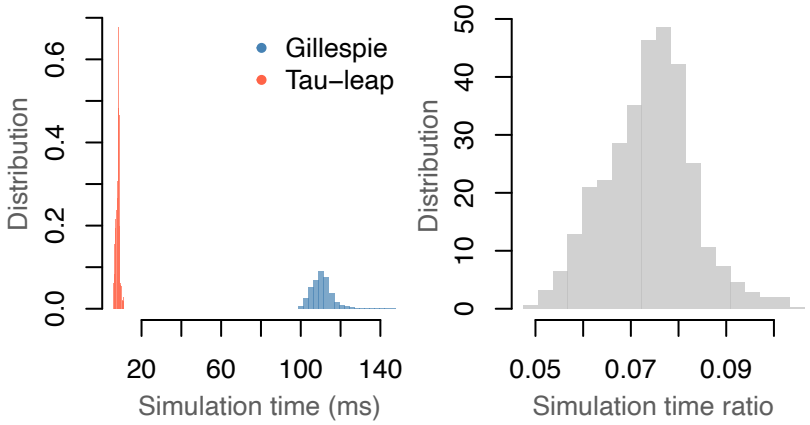
The τ-leap algorithm with the adaptive time-step reduces the computation time more than 10 fold, when compared to simulation via Gillespie's direct method, as seen in Figure 5.1(b), giving a considerable reduction in the computational expense.

OXPHOS deficiency

Following the biochemical threshold theory, a myofibre is considered to be OXPHOS deficient if the variant load passes a pathogenic threshold. Previous work has shown that the threshold depends on the complex in question, Chapter 1.1.9, and, therefore, a different threshold is used for each OXPHOS protein in the dataset, τ_1, τ_2, τ_3. To simulate the three OXPHOS deficiency measurements for one patient within the observed dataset, the mtDNA dynamics of a single myofibre is simulated once, and the myofibre's variant load is compared to the three pathogenic thresholds, giving three values of OXPHOS status. This process is independently repeated a number of times, N_{sim}, storing the OXPHOS status of each repetition, from which the proportion of OXPHOS deficiency, p_1, p_2, p_3, is calculated. Let $f(\boldsymbol{\theta}, x)$ denote the simulator of the mtDNA dynamics for a single myofibre, given a set of model parameters $\boldsymbol{\theta}$ and time points x, and let ϕ_ℓ be the



(a) Simulation output



(b) Simulation time output

Figure 5.1: **Exact vs. approximate simulation of mtDNA dynamics**.(a) The 1%, 25%, 50%, 75%, and 99% quantiles from 1,000 simulations of a system using the exact Gillespie algorithm in red (Gillespie, 1977) and the τ-leap algorithm in blue (Gillespie, 2001), with an adaptive step-size which leaps forward by the expected time for 100 reactions given the system current reactivty. (b) Wall clock simulation times (left panel) and time ratios of *tau*-leap-simulation-time:Gillespie-simulation-time (right panel), from 1,000 independent simulations. All simuatlions were executed using the model of RGD described in this chapter with parameters: $k_0 = 4.53 \times 10^{-7}$, $\rho = 1 \times 10^{-3}$, $k_1 = k_0/100$, $C_0 = 1,000$, and an initial variant load of 0%, on a 2023 Macbook Pro with an M2 Pro chip and 16GB of RAM.

simulated variant load for the ℓ-th repeat. The proportion estimates are then calculated by

$$\begin{aligned} &\text{for } i = 1, 2, \ldots, N_{\text{sim}} \\ &\qquad \phi_\ell \sim f(\boldsymbol{\theta}, x), \quad \text{indep.} \end{aligned}$$

$$\begin{aligned} &\text{for } j = 1, 2, \ldots, N_{\text{OX}} \\ &\qquad p_j = \frac{1}{N_{\text{sim}}} \sum_{\ell=1}^{N_{\text{sim}}} \mathbb{I}\left(\phi_\ell > \tau_j\right). \end{aligned}$$

Where N_{OX} is the number of OXPHOS proteins, here $N_{\text{OX}} = 3$, and $\mathbb{I}(\cdot)$ is the indicator function, returning 1 if the statement within it is true and 0 otherwise.

5.3.2 Agent-based model

An agent-based model (ABM), sometimes called an individual-based model, simulates each agent within a system individually, tracking their movements and reactions. They more naturally allow for spatial dependencies compared to a stochastic kinetic model, whose spatial extensions model independent well-mixed systems between which molecules can migrate at a specified rate (Wilkinson, 2009). A spatial model is required for the final theory of clonal expansion under consideration, the perinuclear niche theory, as mtDNA replication is spatially dependent.

In this section, the development of a mathematical model of the perinuclear niche theory is discussed. The aim is to develop a model which reflects the spatial elements of a myofibre and the behaviour of mtDNA. The model proposed follows the assumptions of mass action kinetics, where possible, but models the system at a molecular level rather than a population level. To construct an ABM of the perinuclear niche theory several things must be considered, first, the model space is discussed.

System space

The system space must reflect the mitochondrial network within a myofibre. The mitochondrial network is mostly static (Huang et al., 2013) and forms planes of well-connected mitochondria which are perpendicular to the longitudinal direction of the myofibre (Dubowitz & Sewry, 2006). The well-connected mitochondrial bands, z-bands, run the length of the myofibre and are separated by mitochondrially sparse areas with fewer connections (Vincent et al., 2018, 2019), see Figure 5.2. In the mathematical model, the z-bands are simplified to be 2-dimensional planes that are assumed well connected enough that mtDNA are able to move freely within them. The myofibre space is constructed by a series of parallel z-bands between which mtDNA can migrate. Migrations are considered on an individual mtDNA basis and are independent of the molecules' position within the z-band. MtDNA movement within and between bands is discussed in detail in the coming section.

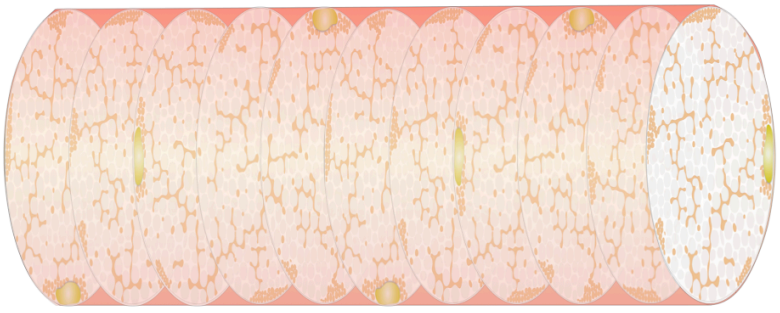


Figure 5.2: **Theoretical diagram of the z-band structure of the mitochondrial network in myofibres**. Mitochondrial z-bands are depicted as flat, 2-dimensional planes with highly connected mitochondria (light brown), between which lie sparsely connected spaces. Nuclei are depicted on the periphery of the myofibre, along the cell membrane, in gold. The within z-band mitochondrial network is shown spreading throughout the z-band and as clusters along the membrane. Figure adapted from Vincent et al. (2018).

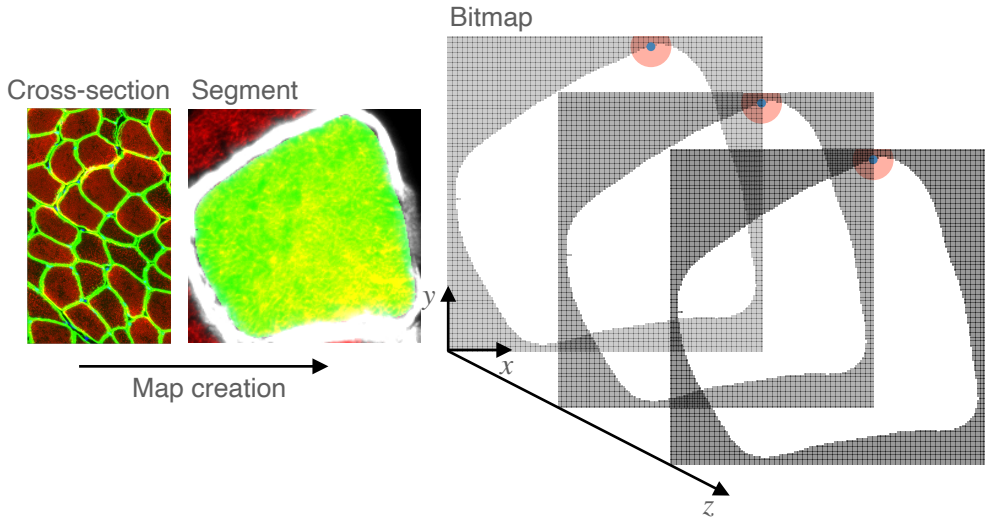


Figure 5.3: **A schematic diagram of the myofibre space within the mathematical model**. A diagram to show the construction of the myofibre modelling space for the perinuclear niche theory. The process begins with an IMC image of tissue cross-section, collecting data for a set of OXPHOS protein abundances. The tissue cross-section is segmented using the mitocyto tool (Warren et al., 2020), which segments and isolates individual myofibers. From the segmented myofibres, bitmaps are created that outline and locate the nuclei, forming a single 2-dimensional plane. The perinuclear region(s) are then added, defined to be a specified radius centre on the middle of each nuclei within the image. The IMC image (left) shows the protein abundances of dystrophin (green), a protein found on the cell membrane, and VDAC1 (red). The segmented image (middle) shows the segmented outline of a single myofibre from the whole tissue image, which is isolated to form the bitmap. The cross-sectional space of the myofibre is shown in green, and the myofibre membrane is now shown in white; the VDAC1 level is shown in red. The bitmaps (right) combine the myofibre cross-section space and the nuclei location to form the z-band space. The perinuclear region is defined to be within a specified radius of the nucleus centre (blue dot). Multiple z-bands are used to make a 3-dimensional modelling space within which mtDNA can move.

The z-band space must be decided upon, as well as the location of the nuclei and the perinuclear region. Both the outline of a myofibre and nuclei location can be inferred from observed datasets. Protein abundance data collected by IMC or QIF, which has been segmented to collect single-myofibre observations, are used to create the z-band space. Segmented myofibres can be isolated and converted into a bitmap, a matrix with binary elements indicating whether each pixel is part of the myofibre or not. The bitmap defines the 2-dimensional plane that reflects the z-band, as shown in Figure 5.3. Nuclei locations for a specific myofibre can be found by the complete absence of mitochondrial protein markers. A single nucleus is defined as a cluster of pixels, absent of mitochondrial protein abundances, which are connected in any direction, and the perinuclear niche is defined as the region within a specified radius of the nuclei centres. It is assumed that mtDNA molecules can not pass through a nucleus and, therefore, the nuclei identified in each myofibre are removed from the z-band bitmap. The pixel size of the pseudo-image is known, and so the map and perinuclear niche can be drawn to scale. Unfortunately, the dataset does not contain serial images of the same tissue section, and even if it did, it would be difficult and time-consuming to match myofibres between serial sections. Therefore, the 3-dimensional model is created by repeating a single z-band map a desired number of times. For this project, a single myofibre space was created, using a single, segmented

myofibre from a pseudo-image collected by IMC. Assuming a 2µm gap between z-bands (Dubowitz & Sewry, 2006; Vincent et al., 2018), the model space had five z-bands, reflecting a tissue sample thickness of 10µm.

Molecule movement

Following the assumptions of a Gillespie-type model, mtDNA movement within a z-band is assumed to be driven by Brownian motion. Therefore, over a time-step, Δt, the within-band x and y coordinates of an mtDNA molecule, $\mathcal{M}_i$, are updated according to a bivariate normal distribution with mean zero and marginal variances proportional to Δt. Let $x_i(t), y_i(t)$ be the x and y within-z-band coordinates of $\mathcal{M}_i$ at time t, the distribution of its position after a time-step, Δt, is

$$\begin{pmatrix} x_i(t+\Delta t) \\ y_i(t+\Delta t) \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} x_i(t) \\ y_i(t) \end{pmatrix}, \Delta t \begin{pmatrix} {\sigma_x}^2 & 0 \\ 0 & {\sigma_y}^2 \end{pmatrix} \right), \tag{5.8}$$

where σ_x and σ_y are the diffusion rates in the x and y direction, within the z-band. It is assumed that $\sigma_x = \sigma_y$, as the mitochondrial network is equally connected within all directions of the z-band (Vincent et al., 2019), the within z-band diffusion rate is denoted $\sigma_x = \sigma_y = \sigma_\mathrm{d}$, and the movement update is simplified,

$$\begin{pmatrix} x_i(t+\Delta t) \\ y_i(t+\Delta t) \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} x_i(t) \\ y_i(t) \end{pmatrix}, {\sigma_\mathrm{d}}^2 \, \Delta t \, I_2 \right). \tag{5.9}$$

For an individual mtDNA molecule, the probability of migrating to an adjacent band over a time-step Δt is assumed to be

$$\Pr\left(\mathcal{M}_i \text{ migration}\right) = 1 - \exp\left(-k_\mathrm{mig} \Delta t\right). \tag{5.10}$$

Given that a migration event occurs, equal weight is given to both adjacent bands. The migration rate is assumed to be small enough that the probability of moving more than one z-band within a time-step is zero. The parameter k_mig controls the migration rate and the spread of mtDNA in the longitudinal direction. A small value of k_mig will result in few migrations and slow longitudinal spread of mtDNA. Conversely, a large value will result in many migration events and a quick spread of mtDNA.

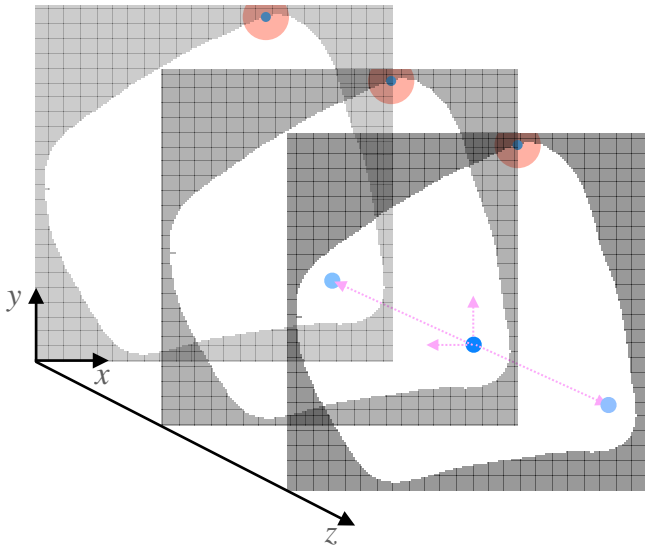


Figure 5.4: **Diagram of mtDNA movement over a time step**. Movement within the z-band is described by Brownian motion whose marginal variances in the x and y directions are proportional to the size of the time-step. Movement between z-bands is described by a probability of migration, described in Equation 5.10.

Simulating reactions

Each molecule can undergo a pre-defined set of reactions (replication, degradation, etc.) with associated reaction rates, which depend on the molecule's location. Given that a molecule experiences an event (any reaction), the specific reaction which occurs must be simulated, and the probability of each reaction is proportional to the specific reaction's hazard. The perinuclear niche theory assumes no advantage between species. However, a spatial dependence must be imposed, constraining which molecules can replicate based on their location within the myofibre. A strict constraint is assumed, which allows only the mtDNA within the perinuclear niche to replicate. As mutation events are assumed to be failed replications, these can also only occur within the niche. It is assumed that any mtDNA molecule can degrade, as this mechanism is related to mitochondrial degradation and independent of mtDNA replication.

The possible reactions are the same as the previous models. The reaction rates indicate that outside of the perinuclear niche region, $\mathcal{R}$, the replication and mutation reactions have a zero probability of occurring. We have that

$$
\begin{aligned}
W \to 2W \quad & \begin{cases} k_1 & W \in \mathcal{R} \\ 0 & W \notin \mathcal{R} \end{cases} , \\
V \to 2V \quad & \begin{cases} k_2 & V \in \mathcal{R} \\ 0 & V \notin \mathcal{R} \end{cases} , \\
W \to \phi \quad & k_3 \, , \\
V \to \phi \quad & k_4 \, , \\
W \to W + V \quad & \begin{cases} k_5 & W \in \mathcal{R} \\ 0 & W \notin \mathcal{R} \end{cases} .
\end{aligned}
$$

The reaction rates are left generic, allowing results to arise naturally when discussing copy number control mechanisms in the next section. This allows for the simulation of specific reactions of molecules; however, the decision on how molecules are chosen to react must be made.

The sum of a molecule's reaction rates is referred to as the molecular hazard and is proportional to the probability that that molecule undergoes a reaction, given that a reaction occurs. For a given molecule, $\mathcal{M}_i$, the molecular hazard is, therefore,

$$
\begin{aligned}
h\left(\mathcal{M}_i \mid \text{wild-type}\right) &= (k_1 + k_3)\mathbb{I}(\mathcal{M}_i \in \Omega) + k_5, \\
h\left(\mathcal{M}_i \mid \text{mutant}\right) &= k_2\mathbb{I}(\mathcal{M}_i \in \Omega) + k_4,
\end{aligned}
$$

where Ω denotes the set of molecules within the perinuclear niche. The system hazard is the sum of all molecular hazards and indicates the overall reactivity of the system,

$$
h_0 = \sum_{\forall i} h(\mathcal{M}_i),
$$

and, similarly to Gillespie's exact method, the system hazard is used to simulate the inter-event time of the system,

$$
\Delta t \sim \text{Exp}\left(h_0\right).
$$

Therefore, for each system update, the inter-event, reacting molecule, and specific reaction event can be simulated, depending only on the reaction hazards of each molecule. A full description of the simulating algorithm is given at the end of this section, Algorithm 7.

Copy number control

Previously discussed copy number control mechanisms alter the replication rates given the system's current state. Assuming a system in equilibrium, the probability of a population increase or decrease is equal when the current and target copy numbers are equal. To achieve this, the base replication and degradation rates are equal. Unfortunately, this method is not appropriate here, as the number of molecules able to replicate is dynamic. A constant replication rate will result in replicative (or degradative) advantage, leading to an unstable copy number. Therefore, the base replication rate must also be dynamic

to reflect a changing system.

A new copy number controller is proposed, one that alters the replication rate to produce a pre-defined probability of copy number increase. This way, it can be assured that the probability of population increase is 50% when the current and target copy numbers are equal, regardless of the number of mtDNA within the niche. Following the linear controller previously implemented, we propose to alter the probability linearly, proportional to $C_0 - (W + V)$ where W and V are the current wild-type and variant populations. The maximum deviation from the target copy number is chosen to be 200 molecules; outside of this range, the function imposes a probability of a population increase to be 0.0 or 1.0, as appropriate. Let $\epsilon = C_0 - (W + V)$ and $\alpha(\epsilon)$ be the function of population increase, then

$$\alpha(\epsilon) = \begin{cases} 0.0, & \epsilon < -200 \\ \frac{\epsilon}{200} + 0.5, & |\epsilon| \leq 200 \\ 1.0, & \epsilon > 200 \end{cases} . \tag{5.11}$$

The replication rate can now be calculated, depending on $\alpha(\epsilon)$ and the degradation rates, to ensure a stable copy number. Let $\mathcal{W}$ and $\mathcal{V}$ denote the set of wild-type and variant mtDNA molecules, respectively, and Ω denote the set of mtDNA molecules currently within the perinuclear niche. Continuing the notation of W and V being wild-type and variant population sizes, and introduce W_Ω and V_Ω to be the population sizes within the perinuclear niche.

The added complexity of modelling individual mtDNA necessitates the use of notation defining a set of mtDNA molecules, let $\mathcal{W}$ and $\mathcal{V}$ denote the set of wild-type and variant mtDNA molecules within a system. Given that a reaction occurs at a specific time, the probability that a specific mtDNA molecule, $\mathcal{M}_i$, reacts is proportional to its molecular hazard. That is

$$\begin{aligned} \Pr\left(\mathcal{M}_i \text{ reacts } | \mathcal{M}_i \in \mathcal{W}\right) &= \frac{k_3 + (k_1 + k_5)\mathbb{I}(\mathcal{M}_i \in \Omega)}{h_0}, \\ \Pr\left(\mathcal{M}_i \text{ reacts } | \mathcal{M}_i \in \mathcal{V}\right) &= \frac{k_2 + k_4\mathbb{I}(\mathcal{M}_i \in \Omega)}{h_0}, \end{aligned}$$

where h_0 denotes the total hazard of the system,

$$\begin{aligned} h_0 &= \sum_{\mathcal{M}_i \in \mathcal{W}} \left[k_1\mathbb{I}(\mathcal{M}_i \in \Omega) + k_3 + k_5\mathbb{I}(\mathcal{M}_i \in \Omega)\right] + \sum_{\mathcal{M}_i \in \mathcal{V}} \left[k_2\mathbb{I}(\mathcal{M}_i \in \Omega) + k_4\right], \\ &= k_3 W + (k_1 + k_5)W_\Omega + k_2 V_\Omega + k_4 V. \end{aligned} \tag{5.12}$$

Given that molecule $\mathcal{M}_i$ reacts and its species, the probabilities of reactions which result in an increased copy number are

$$\begin{aligned} \Pr\left(\mathcal{M}_i \text{ replicates } | \mathcal{M}_i \in \mathcal{W}\right) &= \frac{k_3}{k_1 + k_3 + k_5}\mathbb{I}\left(\mathcal{M}_i \in \Omega\right), \\ \Pr\left(\mathcal{M}_i \text{ mutates } | \mathcal{M}_i \in \mathcal{W}\right) &= \frac{k_5}{k_1 + k_3 + k_5}\mathbb{I}\left(\mathcal{M}_i \in \Omega\right), \\ \Pr\left(\mathcal{M}_i \text{ replicates } | \mathcal{M}_i \in \mathcal{V}\right) &= \frac{k_2}{k_2 + k_4}\mathbb{I}\left(\mathcal{M}_i \in \Omega\right). \end{aligned}$$

Next, the probability of a population increase, as a function of the reaction rates, must be found. The probability can be split by conditioning on the type of molecule to react,

$$\begin{aligned}\Pr(\text{population increase}) = \Pr(&\text{population increase} \mid \text{wild-type reaction}) \\ &+ \Pr(\text{population increase} \mid \text{variant reaction}).\end{aligned}$$

The conditional probabilities are considered in turn, the probability of a population increase given that a wild-type molecule reacts is

$$\begin{aligned}\Pr(\text{pop. incr.}|\text{Wild-type}) &= \Pr(\text{rep. }|\text{Wild-type}) + \Pr(\text{mut.}|\text{Wild-type}), \\ &= \sum_{\mathcal{M}_i \in \mathcal{W}} \Pr(\text{rep. } |\mathcal{M}_i \text{ reacts}) \times \Pr(\mathcal{M}_i \text{ reacts}) \\ &\quad + \sum_{\mathcal{M}_i \in \mathcal{W}} \Pr(\text{mut. } |\mathcal{M}_i \text{ reacts}) \times \Pr(\mathcal{M}_i \text{ reacts}), \\ &= \sum_{\mathcal{M}_i \in \mathcal{W}} [\Pr(\text{rep. } |\mathcal{M}_i \text{ reacts}) + \Pr(\text{mut. } |\mathcal{M}_i \text{ reacts})] \times \Pr(\mathcal{M}_i \text{ reacts}), \\ &= \sum_{\mathcal{M}_i \in \mathcal{W}} \left[\frac{k_1 \mathbb{I}\left(\mathcal{M}_i \in \Omega\right)}{k_1 + k_3 + k_5} + \frac{k_5 \mathbb{I}(\mathcal{M}_i \in \Omega)}{k_1 + k_3 + k_5}\right] \times \frac{k_1 + (k_3 + k_5)\mathbb{I}(\mathcal{M}_i \in \Omega)}{h_0}, \\ &= \frac{(k_1 + k_5)W_\Omega}{h_0}.\end{aligned}$$

Similarly, the probability of a population increase given a variant mtDNA reaction is

$$\begin{aligned}\Pr(\text{pop. incr. }|\text{Variant}) &= \Pr(\text{rep. }|\text{Variant}), \\ &= \sum_{\mathcal{M}_i \in \mathcal{V}} \Pr(\text{rep. } |\mathcal{M}_i \text{ reacts}) \times \Pr(\mathcal{M}_i \text{ reacts}), \\ &= \frac{k_2 V_\Omega}{h_0}.\end{aligned}$$

Combining these, the marginal probability of a population increase can be expressed as

$$\begin{aligned}\Pr(\text{population increase}) &= \frac{(k_1 + k_5)W_\Omega + k_2 V_\Omega}{h_0}, \\ &= \frac{(k_1 + k_5)W_\Omega + k_2 V_\Omega}{(k_1 + k_5)W_\Omega + k_2 V_\Omega + k_3 W + k_4 V}.\end{aligned} \tag{5.13}$$

This is the general form of the probability and can be simplified for the perinuclear niche model by imposing assumptions about model parameters, namely no replicative advantage and mtDNA mutations ocurring during wild-type replication. These imply that $k_1 + k_5 = k_2 = k_{\text{rep}}$ and $k_3 = k_4 = k_{\text{deg}}$. The above probability equals $\alpha(\epsilon)$, a pre-defined function, and setting the expression equal to $\alpha(\epsilon)$ and re-arranging the form k_{rep} is found

$$\begin{aligned}\alpha(\epsilon) &= \frac{k_{\text{rep}} W_\Omega + k_{\text{rep}} V_\Omega}{k_{\text{rep}} V_\Omega + k_{\text{rep}} V_\Omega + k_{\text{deg}}(W + V)}, \\ k_{\text{rep}} &= \frac{\alpha(\epsilon)[W + V]}{(1 - \alpha(\epsilon))[W_\Omega + V_\Omega]} k_{\text{deg}}.\end{aligned} \tag{5.14}$$

Therefore, the general replication rate of mtDNA, for both wild-type and variant mtDNA, can be defined to achieve a desired probability of population increase, thus controlling

mtDNA copy number.

The dynamic replication rate alters depending on the number of mtDNA molecules within the perinuclear niche, $W_\Omega + V_\Omega$, and the overall copy number. How its definition causes the system hazard to change with time should be inspected. One of the modelling assumptions is that the system hazard does not change in-between reactions; otherwise, the simulation method would not be exact. Therefore, the hazard function for this form of k_{rep} is calculated as

$$\begin{aligned}
h_0(t) &= \sum_{\mathcal{M}_i \in \mathcal{W}} \left[(1-\rho)k_{\text{rep}}\mathbb{I}(\mathcal{M}_i \in \Omega) + k_{\text{deg}} + \rho k_{\text{rep}}\mathbb{I}(\mathcal{M}_i \in \Omega)\right] + \sum_{\mathcal{M}_i \in \mathcal{V}} \left[k_{\text{rep}}\mathbb{I}(\mathcal{M}_i \in \Omega) + k_{\text{deg}}\right], \\
&= \sum_{\mathcal{M}_i \in \mathcal{W}} \left[k_{\text{rep}}\mathbb{I}(\mathcal{M}_i \in \Omega) + k_{\text{deg}}\right] + \sum_{\mathcal{M}_i \in \mathcal{V}} \left[k_{\text{rep}}\mathbb{I}(\mathcal{M}_i \in \Omega) + k_{\text{deg}}\right], \\
&= k_{\text{rep}}(W_\Omega + V_\Omega) + k_{\text{deg}}(W+V), \\
&= \frac{\alpha(\epsilon)k_{\text{deg}}(W+V)}{(1-\alpha(\epsilon))(W_\Omega + V_\Omega)}(W_\Omega + V_\Omega) + k_{\text{deg}}(W+V) \\
&= \frac{k_{\text{deg}}(W+V)}{1-\alpha(\epsilon)}.
\end{aligned}$$

This shows that the total system hazard, h_0, is dependent on the copy number, $W+V$, the degradation rate, k_{deg}, and the defined probability of population increase, $\alpha(\epsilon)$. Under the assumption of a constant degradation rate and $\alpha(\epsilon)$, the hazard is constant between reactions and, therefore, the system can be exactly simulated.

OXPHOS deficiency

A synthetic, simulated myofibre is considered OXPHOS deficient if its variant load passes the protein-specific pathogenic threshold, similarly to models of RGD and SoS previously described. Variant load is calculated by all the mtDNA within the system, not by specific z-bands. The proportion of OXPHOS deficiency is simulated following the same steps as before, see Chapter 5.3.1.

Simulating perinuclear niche model

Algorithm 7 outlines the general method for simulating from the continuous-time, agent-based mathematical model of the perinuclear niche hypothesis.

Computational cost

The requirement of simulating the movement of each molecule adds considerable computational expense to the simulations, and the linearity of the simulation algorithm means it is difficult to reduce this. Therefore, the mathematical model was written in C++, a high-performance programming language, and steps to reduce the computational cost were taken, including compilation optimisation and memory allocation considerations. However, the model is still very expensive to simulate compared to the RGD or SoS models. The expected simulation time of the perinuclear niche model is approximately 1400

Algorithm 7 Mathematical model of the perinuclear niche theory

1. Make the simulation space.
   i Select a single myofibre cross-section to be the z-band space and create a bit-map of its cross-section from a segmented pseudo-image
   ii Decide the size of the simulation space and the number of z-bands required for this.
2. Initialise the model. Define the initial mtDNA populations of wild-type, $\mathcal{W}$, and variant, $\mathcal{V}$, and their location within the modelling space. Set the system time $t = 0$.
3. Update the system
   i Calculate the system hazard, $h_0 = \sum_{\forall \mathcal{M}_i} h(\mathcal{M}_i)$
   ii Sample the time to the next event, $\Delta t \sim \mathrm{Exp}\,(h_0)$
   iii Simulate mtDNA movement. Let (x_i, y_i, z_i) be the x and y coordinates within the z-band and the z-band identifier for the molecule $\mathcal{M}_i$. For $\mathcal{M}_i \in \mathcal{W} \cup \mathcal{V}$, simulate the within z-band movement

$$\begin{pmatrix} x_i(t+\Delta t) \\ y_i(t+\Delta t) \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} x_i(t) \\ y_i(t) \end{pmatrix}, \begin{pmatrix} {\sigma_\mathrm{d}}^2 & 0 \\ 0 & {\sigma_\mathrm{d}}^2 \end{pmatrix} \Delta t \right),$$

   and the between z band movement, let $p = 1 - \exp\left(-k_\mathrm{mig} \Delta t\right)$

$$z_i(t+\Delta t) = \begin{cases} z_i(t) + 1, & p/2 \\ z_i(t), & 1-p \\ z_i(t) - 1, & p/2 \end{cases}$$

   and update the $\mathcal{W}_\Omega$ and $\mathcal{V}_\Omega$ accordingly
   iv Calculate replication rate

$$k_\mathrm{rep} = \frac{\alpha(\epsilon)[W+V]}{(1-\alpha(\epsilon))[W_\Omega + V_\Omega]} k_\mathrm{deg}$$

   and update the reaction rates of each molecule
   v Randomly sample the molecule to react

$$\Pr\left(\mathcal{M}_i \text{ reacts}\right) = h_i / h_0$$

   vi Randomly sample mtDNA reaction

$$\Pr\left(\mathcal{M}_i \text{ reaction } j\right) = k_j / h_i$$

   and update mtDNA molecule sets accordingly $\mathcal{W}$ and $\mathcal{V}$
4. If $t < T_\mathrm{max}$, put $t := t + \Delta t$ and return to Step 3.

times larger than that of the RGD, simulated by the Gillespie algorithm.

5.3.3 Statistical model

It is assumed that there is inherent noise in the observations from the collected data; therefore, we must model this observational error. As the data are proportions, the error is assumed to be on the logit scale. The logit is a function maps $[0, 1]$ to the whole real line,

$$g(x) = \log \frac{x}{1-x}$$

for $x \in [0, 1]$. The logit-transformation does not handle raw values of $x = 0, 1$ well, resulting in non-finite values. Deficiency proportions of 100% are not expected to be observed within the data. However, deficiency proportions of 0% could be seen in some OXPHOS proteins within younger patients, and both can appear during inference of simulated output. Therefore, an alteration is made to the transformation. Let x be a proportion and y its transformed value, so that

$$y = \log \frac{x+\delta}{1-x+\delta} = \text{elogit}(x),$$

where δ is a small number. This transformation maps a proportion to a finite range, dependent on the choice of δ.

Assuming normal random noise on an approximately logit scale, the transformed observed data Y_i would be normally distributed and centred around the 'true' deficiency proportion, with a precision ψ. However, the 'true' proportion of deficiency is unknown and is therefore a latent variable. The value of the proportion is estimated via a mathematical model. Let $\hat{p}_i$ be the estimated deficiency proportion, then the statistical model is

$$Y_i | \hat{p}_i, \psi \sim \text{N}\left(\text{elogit } \hat{p}_i, \psi^{-1}\right).$$

The dataset in question possesses a single OXPHOS deficiency proportion observation per patient, per OXPHOS protein. Let Y_j^i and $\hat{p}_j^i$ be the transformed observed and estimated proportion of deficiency for the i-th patient and j-th OXPHOS protein. Including the number of proteins per patient, the model becomes: for $i = 1, 2, \ldots, N_{\text{pat}}$ and $j = 1, 2, N_{\text{OX}}$

$$Y_j^i | \hat{p}_j^i, \psi \sim \text{N}\left(\text{elogit } \hat{p}_j^i, \psi^{-1}\right), \tag{5.15}$$

where N_{pat} is the number of patients in the dataset and N_{OX} is the number of OXPHOS proteins.

The proportion of deficiency for one patient is estimated using N_{sim} independent simulations of the mathematical model. For data from a 2Dmito plot, the number of myofibres is known and can be used as the number of independent simulations; here, the number ranges from 153 to 1,199. However, the computational cost of simulation is significant, so $N_{\text{sim}} = 100$ is chosen to reduce computational expense and is found to be an appropriate approximation. The estimate, $\hat{p}_j^i$ is then found by the proportion of simulations whose variant load is greater than the pathogenic threshold.

Let $\boldsymbol{\tau}$ be the set of pathogenic thresholds for each OXPHOS protein and ϕ_ℓ^i be the ℓ-th simulated variant load of the i-th patient, the estimated proportion of deficiency for that patient is

$$\hat{p}_j^i = \frac{1}{N_{\text{sim}}} \sum_{\ell=1}^{N_{\text{sim}}} \mathbb{I}\left(\phi_\ell^i > \tau_j\right). \tag{5.16}$$

5.3.4 Prior beliefs

Base reaction rate

Estimates of mtDNA half-life range between a few days to a few hundred days (Burgstaller et al., 2014; Collins et al., 2003; Gross et al., 1968; Menzies & Gold, 1971). Within this range, experts do not believe some values are less or more likely than others, due to a lack of consensus in studies. Investigations which have inferred replication/degradation rates impart vague or uniform prior beliefs to reflect this uncertainty (Henderson et al., 2009; Insalata et al., 2022; Johnston et al., 2015). The prior belief is constructed on the log scale, where it is easier to reflect uncertainties about the order of the reaction rate. A vague prior is appropriate given the variety and limited number of existing estimates. A normal distribution is chosen to summarise prior beliefs of the base reaction rate on the log-scale, $\theta_1 \sim \mathrm{N}\left(-14.036, 2.0\right)$ for $\theta_1 = \log k_0$.

Mutation probability

No estimates were found for the mtDNA mutation probability for a patient with an mtDNA maintenance disorder, and after consulting experts, they agreed that the level of uncertainty *a priori* is high. However, some estimates of the mutation probability exist within healthy controls and other patient phenotypes. These estimates, within other patient phenotypes, range from approximately 10^{-8} to 10^{-3} (Henderson et al., 2009; Mao et al., 2006; Shenkar et al., 1996). Experts agreed that the mutation probability for patients with mtDNA maintenance disorders is higher than those. The prior is placed on the log-scale, which can better handle small numbers and removes possible complications when proposing values close to the limit of the support. It is not expected that the probability of mutation reaches 1.0, and so the upper limit is not of concern. Given the limited prior beliefs, a vague prior is used, let $\theta_2 \sim \mathrm{N}_{(-\infty,0]}\left(-4.605, 5.0\right)$, where $\theta_2 = \log \rho$.

Pathogenic threshold

Several estimates of the pathogenic threshold exist, but their tissue- and mutation-specific nature means that many estimates are not appropriate, and the ones that are available come with caveats, Chapter 1.1.9. However, no estimate places the pathogenic threshold below 40%, so the prior support is truncated to $[0.4, 1.0]$. Again, the prior is placed on the log-scale, due to difficulties of sampling close to the support limits. The three pathogenic thresholds on the log-scale, $\theta_3 = \log \tau_1, \theta_4 = \log \tau_1, \theta_5 = \log \tau_1$, are all assumed to have the same prior and be independent *a priori*, maintaining the generalisability of the method. The prior beliefs are also summarised by a normal distribution, for $i = 3, 4, 5$ $\theta_i \sim \mathrm{N}_{[\log 0.4, 0.0]}\left(-0.511, 0.2\right)$.

Model error

Inter-patient variability is inherent within mitochondrial disease; as such, there is expected to be a relatively high degree of noise in this dataset. Recall that the noise is on the logit scale and the range of transformed deficiency proportions is approximately $[-6, 6]$. Due to the high variability in the data, it is believed that the model precision, ψ, should not be too large, with the bulk of the prior being below 2.0. This prior belief is summarised by a gamma distribution, with shape $\alpha = 1$ and rate $\beta = 0.1$, that is $\psi \sim \text{Ga}(\alpha, \beta)$.

5.3.5 Inference

A hybrid algorithm can achieve Bayesian inference for the stochastic kinetic model and the model precision. Due to the choice of a gamma prior, the model precision can be updated using a Gibbs update. In contrast, the mathematical model parameters must be updated via a Metropolis-Hastings step. Let $\{\boldsymbol{y}, \boldsymbol{x}\}$ be the set of observed transformed deficiency proportions and age of patients at each observation, and $\boldsymbol{z}$ be the set of the transformed latent variables of the 'true' proportion of deficiency. The model error is described by precision ψ and is assumed constant in time and for all OXPHOS proteins. Lastly, the mathematical model parameters are denoted $\boldsymbol{\theta} = (\log k_0, \log \rho, \log \tau_1, \log \tau_2, \log \tau_3)^T$, then the joint posterior density is given by

$$p(\boldsymbol{\theta}, \psi, \boldsymbol{z}|\boldsymbol{y}, \boldsymbol{x}) \propto p(\psi)p(\boldsymbol{\theta})p(\boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{x})p(\boldsymbol{y}|\boldsymbol{z}, \psi). \tag{5.17}$$

The two-step update Metropolis-within-Gibbs, Algorithm 8, directly samples from the joint posterior density. The details of sampling the mathematical model of variant load dynamics is omitted, and denoted $\phi \sim f(\boldsymbol{\theta}, x)$, corresponding to a simulation from a mathematical model, $f(\cdot)$, given model parameters $\boldsymbol{\theta}$, between birth and x years.

The acceptance probability in Step 3.iv. of Algorithm 8 has already been simplified. The full form includes terms involving the density of the latent variables' density. However, these cancel as they are on both the numerator and denominator of the acceptance ratio. Therefore, it is possible to directly sample from the joint posterior density without evaluating the latent data density.

The proprosal density, $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)})$, corresponds to a normal random walk i.e.

$$\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)} \sim \text{N}\left(\boldsymbol{\theta}^{(t-1)}, \Sigma_\Theta\right), \tag{5.18}$$

where Σ_Θ is an appropriately defined covariance matrix. For the inference schemes executed in this chapter, the proposal covariance matrix was estimated using an initial run of 10,000 iterations of the scheme, and the covariance was calculated by the methods proposed in Roberts and Rosenthal (2001). The acceptance probability in Step 3.iv. of Algorithm 8 has been simplified, using the symmetry of the proposal distribution.

5.3.6 Synthetic datasets

This section describes the methods used to generate the synthetic datasets. Three synthetic datasets are created, one for each of the mathematical models described in the

Algorithm 8 Hybrid MCMC scheme to infer parameters of a mathematical model of clonal expansion

1. Initialise the system
   i Set $t = 1$.
   ii Set
   $$\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \theta_3^{(0)}, \theta_4^{(0)}, \theta_5^{(0)})^T$$
   to be appropriate values from the support.
   iii Independently sample the initial variant loads from the mathematical model, for $i = 1, 2, \ldots, N_{\text{pat}}$ and $\ell = 1, 2, \ldots, N_{\text{sim}}$
   $$\phi_\ell^{i\,(0)} \sim f(\boldsymbol{\theta}^{(0)}, x_i).$$
   iv Let $\boldsymbol{\theta}_\tau{}^{(0)} = (\theta_3, \theta_4, \theta_5)^T$ and calculate the initial latent state. For $i = 1, 2, \ldots, N_{\text{pat}}$ and $j = 1, \ldots, N_{\text{OX}}$
   $$\hat{p}_j^{i\,(0)} = \frac{1}{N_{\text{sim}}} \sum_{\ell=1}^{N_{\text{sim}}} \mathbb{I}\left(\log \phi_\ell^{i\,(0)} > \theta_{\tau j}{}^{(0)}\right),$$
   $$z_j^{i\,(0)} = \text{elogit}\, \hat{p}_j^{i\,(0)}.$$
2. Update model error
   $$\psi^{(t)} | \boldsymbol{\theta}^{(t-1)} \sim \text{Ga}\left(\tilde{\alpha}, \tilde{\beta}\right)$$
   $$\tilde{\alpha} = \tfrac{1}{2}\alpha + N_{\text{pat}} N_{\text{OX}}, \quad \tilde{\beta} = \textstyle\sum_{i=1}^{N_{\text{pat}}} \sum_{j=1}^{N_{\text{OX}}} \left(y_j^i - z_j^{i\,(t-1)}\right)^2.$$
3. Update mathematical model parameters
   i Propose new parameters
   $$\boldsymbol{\theta}^* \sim q(\boldsymbol{\theta}^* | \boldsymbol{\theta}^{(t-1)})$$
   ii Generate OXPHOS deficiency proprtion latent variables. For $i = 1, 2, \ldots, N_{\text{pat}}$ and $\ell = 1, 2, \ldots, N_{\text{sim}}$
   $$\phi_\ell^{i\,*} \sim f(\boldsymbol{\theta}^*, x_i),$$
   iii Calculate the proportions of deficiency for each OXPHOS protein. For $i = 1, 2, \ldots, N_{\text{pat}}$ and $\ell = 1, 2, \ldots, N_{\text{sim}}$
   $$\hat{p}_j^{i\,*} = \frac{1}{N_{\text{sim}}} \sum_{\ell=1}^{N_{\text{sim}}} \mathbb{I}\left(\log \phi_\ell^{i\,*} > \theta_{\tau j}^*\right),$$
   $$z_j^{i\,*} = \text{elogit}\, \hat{p}_j^{i\,*}.$$
   iv Calculate the probability of acceptance
   $$\alpha(\boldsymbol{\theta}^{(t-1)}, \boldsymbol{\theta}^*) = \frac{p(\boldsymbol{\theta}^*) p(\boldsymbol{y} | \boldsymbol{z}^*, \psi^{(t)})}{p(\boldsymbol{\theta}^{(t-1)}) p(\boldsymbol{y} | \boldsymbol{z}^{(t-1)}, \psi^{(t)})}.$$
   v With probability $\alpha(\boldsymbol{\theta}^{(t-1)}, \boldsymbol{\theta}^*)$, set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^*$ and $\boldsymbol{z}^{(t)} = \boldsymbol{z}^*$. Otherwise set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$ and $\boldsymbol{z}^{(t)} = \boldsymbol{z}^{(t-1)}$.
4. Put $t := t + 1$ and return to Step 2.

previous sections; these are denoted: $\boldsymbol{X}_{\text{RGD}}$, $\boldsymbol{X}_{\text{SoS}}$, and $\boldsymbol{X}_{\text{PNN}}$, generated via the mathematical models of RGD, SoS and PNN, respectively. First, the ground-truth parameter values used for their generation and their justification are discussed.

Fixing ground-truth parameters

The mtDNA half-life estimated by Korr et al. (1998), 17.7d, was chosen as the ground-truth value. This estimate was preffered to that found by Collins et al. (2003) as it is more consistent to other half-life estimates. Therefore, the base replication rate is set to its equivalent value, $k_0 = \frac{\log 2}{17.7\times60\times60\times24} = 4.33 \times 10^{-7}\text{s}^{-1}$.

Little information about the mutation probability for patients with an mtDNA maintenance disorder is available. Some studies estimate the mutation rate in healthy subjects and patients with other phenotypes, as discussed previously in Chapter 5.1.6. These can be used as a guide, noting that the probability of mutation for mtDNA maintenance disorder patients will be higher. A mutation probability an order of magnitude larger than previous estimates is chosen, and $\rho = 0.001$.

Following the work of Rossignol et al. (1999), the pathogenic thresholds are set to be 70% for CI deficiency, 85% for CIII, and 66.8% for CIV deficiency. Although these are not the exact estimates for complexes CI and CIII, they are within the estimates standard errors. A smaller threshold for CI was chosen, 70% compared to 74.5%, to vary the increments between the thresholds, allowing the simulation study to better investigate whether substantial differences between thresholds can be found when their 'true' values are closer together. As the synthetic data is not concerned with specific OXPHOS proteins, the synthetic proteins are referred to by OX1, OX2, and OX3, and their associated ground-truth threshold are $\tau_1 = 0.668$, $\tau_2 = 0.700$, and $\tau_3 = 0.850$.

Iborra et al. (2004) estimated the diffusion rate of mtDNA along the length of a myofibre to be $1.1 \times 10^{-3}\mu\text{m}^{-1}\,\text{s}^{-1}$. This value was then used by Insalata et al. (2022) o estimate the migration rate of mtDNA between adjacent sections of a myofibre to be $\approx 0.1\text{d}^{-1}$, or $1.116 \times 10^{-6}\text{s}^{-1}$. Assuming that the mtDNA movement along a myofibre is random, using this estimate as the basis of the probability of moving between z-bands in the mathematical model of PNN is reasonable. Mitochondrial connectivity within z-bands is approximately four times greater than connectivity between z-bands (Vincent et al., 2019). Therefore, an appropriate estimate for within-band diffusion would be four times greater than the longitudinal diffusion rate, $4.4 \times 10^{-3}\mu\text{m}^{-1}\,\text{s}^{-1}$. However, implementing these values of mtDNA diffusion results in mtDNA clustered around the perinuclear niche and not evenly spread throughout the z-band. Therefore, the within z-band diffusion rate is increased to give a more realistic spatial distribution of mtDNA, and set $\sigma_{\text{d}} = 2.31 \times 10^{-5}$, and the same factor similarly inflates the migration rate to maintain a four fold increase, $k_{\text{mig}} = 5.79 \times 10^{-6}$.

The shape of the perinuclear niche is simplified to be a circle around the centre of the nuclei. The radius of the niche is set to be 4.0µm, upon consultation with experts and in conjunction with observations by Vincent et al. (2018).

The variant advantage parameter is thought to arise from the physically smaller mtDNA molecule. Assuming an mtDNA half-life of 17.7d and that copy number in equilibrium (Korr et al., 1998), the expected time between replications of a single molecule is approximately 25h. The replication time for an mtDNA molecule is between one and two hours (Clayton, 1982). Using these values, the replicative advantage of the replication rate can be calculated. The common deletion removes approximately 30% of the mtDNA molecule and, assuming the replication time is proportional to the length of the molecule, the variant mtDNA replication rate is approximately 0.1% larger than wild-type and, therefore, it is chosen that $\gamma = 1.001$.

Parameter	Value	Model
k_0	$4.53 \times 10^{-7}\text{s}^{-1}$	RGD, SoS, PNN
ρ	0.001	RGD, SoS, PNN
τ_1	0.668	RGD, SoS, PNN
τ_2	0.70	RGD, SoS, PNN
τ_3	0.850	RGD, SoS, PNN
γ	1.001	SoS
σ_d	$2.31 \times 10^{-5}\mu\text{m s}^{-1}$	PNN
k_mig	$5.79 \times 10^{-6}\text{s}^{-1}$	PNN
ψ	1.0	RGD, SoS, PNN

Table 5.2: **Ground-truth parameter values mathematical models of clonal expansion**. Ground-truth values are used to generate synthetic datasets.

Generating synthetic data

The synthetic datasets are constructed to resemble the structure of the observed dataset. Each patient has three measurements of OXPHOS deficiency from a single tissue sample, and each measurement is from a different OXPHOS protein: OX1, OX2, and OX3. However, for the synthetic dataset, the number of patients, N_pat, is increased from nine to twenty. Patient age, $\boldsymbol{x}$, at the time of observation, was sampled independently from a uniform distribution between 40 and 90 years old, and then rounded to the nearest whole number. The same set of patient ages was used for all synthetic datasets.
For all three datasets, synthetic proportions of deficiency were generated for each patient by independently sampling from the appropriate model $N_\text{sim} = 100$ times and calculating the proportion of simulated variant loads above the ground-truth pathogenic thresholds, $\boldsymbol{\tau}$. This yielded a dataset with 60 observations, 20 patients having one observed proportion of deficiency for each OXPHOS protein.

The ground-truth proportion data, $\boldsymbol{z}$, were then transformed to the elogit-scale to add random noise and generate the synthetic observations, $\boldsymbol{y}$. After adding random noise, the synthetic data was transformed back to the proportion scale. Figure 5.5 shows the resulting synthetic dataset $\boldsymbol{X}_\text{RGD}$, the remaining synthetic datasets can be seen in Figures 5.9, 5.10 and B.1.

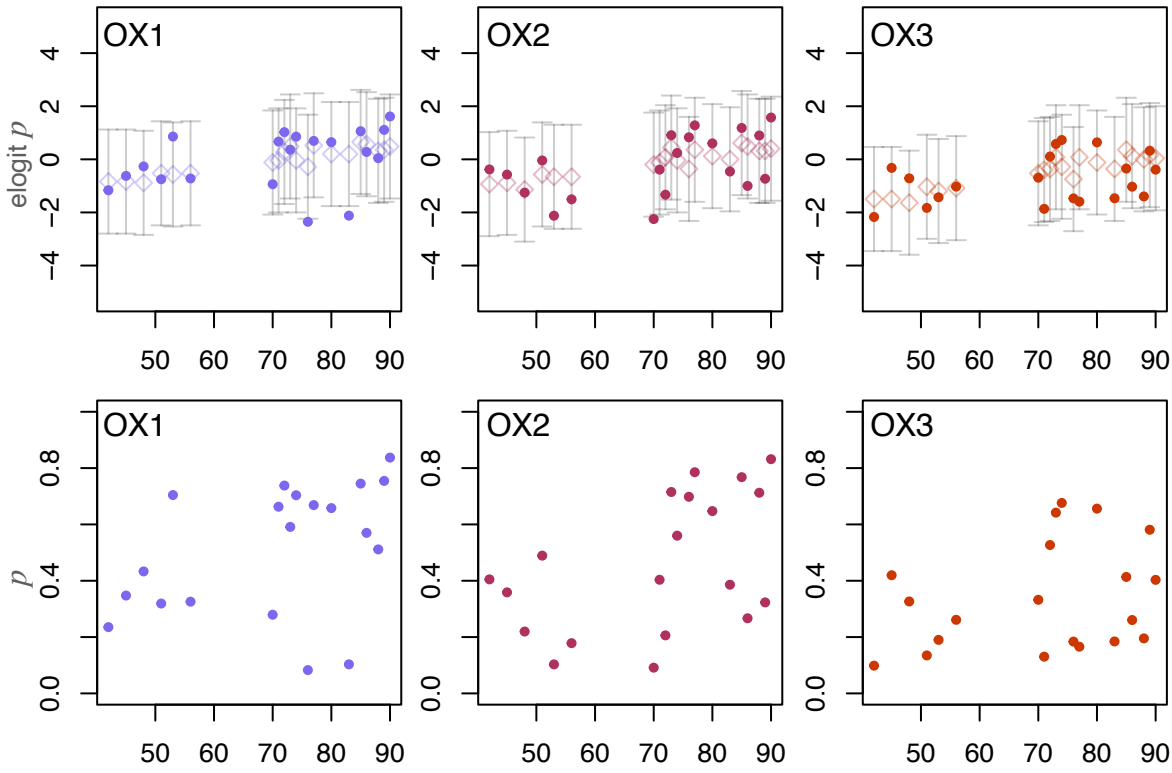


Figure 5.5: **Synthetic RGD OXPHOS deficiency proporiton dataset**. Synthetic datasets generated by the mathematical model of the random genetic drift. The top row depicts the latent, ground-truth deficiency proportions (y-axis) for the three OXPHOS proteins (diamonds) and the synthetic observed data, after the addition of random noise on the transformed scale (solid circle) against time in years (x-axis). The error bars show the 95% interval of the observed value, given its ground-truth value. The bottom row shows the synthetic observed data, after adding random noise, on the natural scale (y-axis) against time in years.

5.3.7 Computational cost of inference and implementation

As mentioned, the computational cost of inference is significant due to the many simulations required to generate the latent variable, $\hat{p}_j^i$, and is highly correlated to the base reaction rate for all mathematical models. Although not inferred here, mtDNA copy number is also correlated to the computational expense, again, because a high copy number leads to a more reactive system. The cost is further increased for larger datasets, with each new patient an additional N_{sim} simulations must be executed per iteration of the inference scheme.

The inference scheme was implemented in C++ to reduce the computational cost. However, it is still significant and highly variable. The time required to produce a single simulated deficiency proportion can range between 130ms and 94, 000ms. The times were calculated by taking the mean simulation time required from 100 repetitions, using the smallest and largest posterior values of the base reaction rate after fitting the mathematical model, $\min p(\theta_1|\boldsymbol{X}_{\text{RGD}})$ and $\max p(\theta_1|\boldsymbol{X}_{\text{RGD}})$, and simulating 100 years. The timings were executed on a 2023 MacBook Pro with an M2 Pro chip and 16GB of RAM. For each iteration of the inference, simulations must be executed for each patient, resulting in a total of nine repetitions for the observed dataset and 20 repetitions for the synthetic datasets. As a result, it was only possible to execute one chain per dataset. Each chain

was executed with 100,000 iterations (including burn-in), and the first 20,000 iterations were removed as burn-in. All chains suffered from extremely high autocorrelation and low ESSs.

5.4 Results

5.4.1 Observed data

Posterior inference for the observed dataset was relatively successful. Although, due to computational restraints, only a single chain was executed, which showed no signs of non-convergence. However, the posterior draws have high autocorrelation, and the resulting ESS is small. For 80,000 posterior draws, the single-variate ESS range was $[178, 462]$ for the parameters of the mathematical model and a multivariate ESS of 230. The ESS for the model precision was comparatively much higher, being over $2,000$.

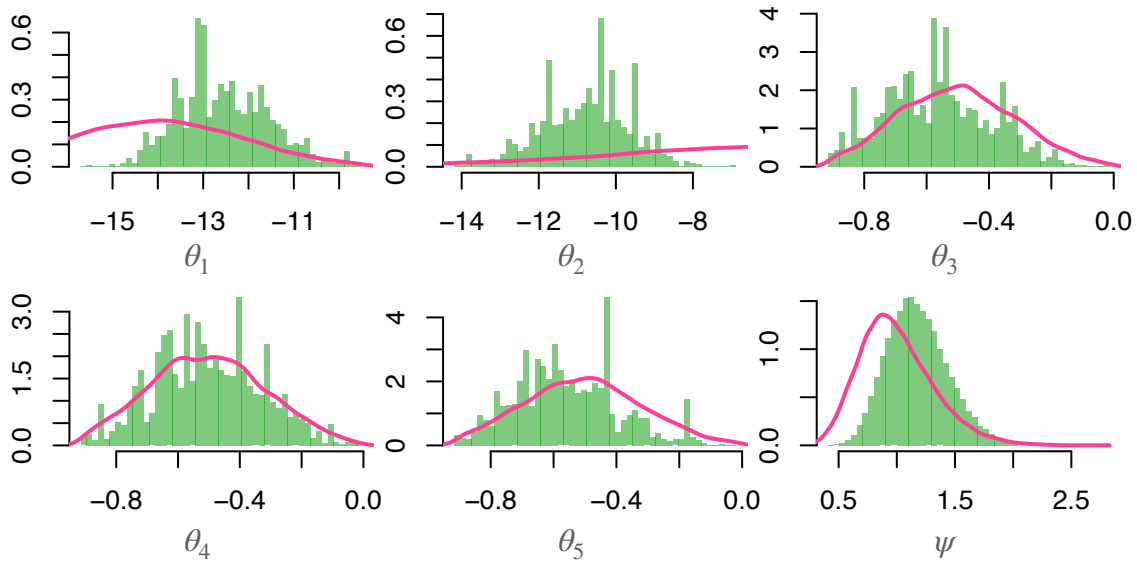
Posterior beliefs

Marginal posterior densities after fitting the model to the observed dataset, $\boldsymbol{X}_{\text{Obs}}$, can be seen in Figure 5.6. There is a fair degree of posterior uncertainty in model parameters, although this is not surprising given the relatively small amount of data. The marginal posteriors, however, do not show the relationships between model parameters. The parameters θ_1 and θ_2, the base reaction rate and mutation probability on the log-scale, are strongly correlated and $\text{cor}(\theta_1, \theta_2 | \boldsymbol{X}_{\text{Obs}}) = -0.9982$. This is to be expected, as the rate OXPHOS deficiency increases depends heavily on both parameters. Increased mitochondrial turnover increases the rate of the variant load increase, as it allows more opportunities for *de novo* mutation events. Similarly, an increased mutation probability increases the rate at which variant mtDNA clonally expands. Posterior beliefs about the pathogenic thresholds strongly resemble their prior beliefs, indicating insufficient information in the data to infer their value, and no evidence of a substantial difference between them is found *a posteriori*. The posterior beliefs for ψ, the model precision, have been updated, although its value has a fair degree of uncertainty. Posterior expectations and 95% HDIs are given in Table 5.3.

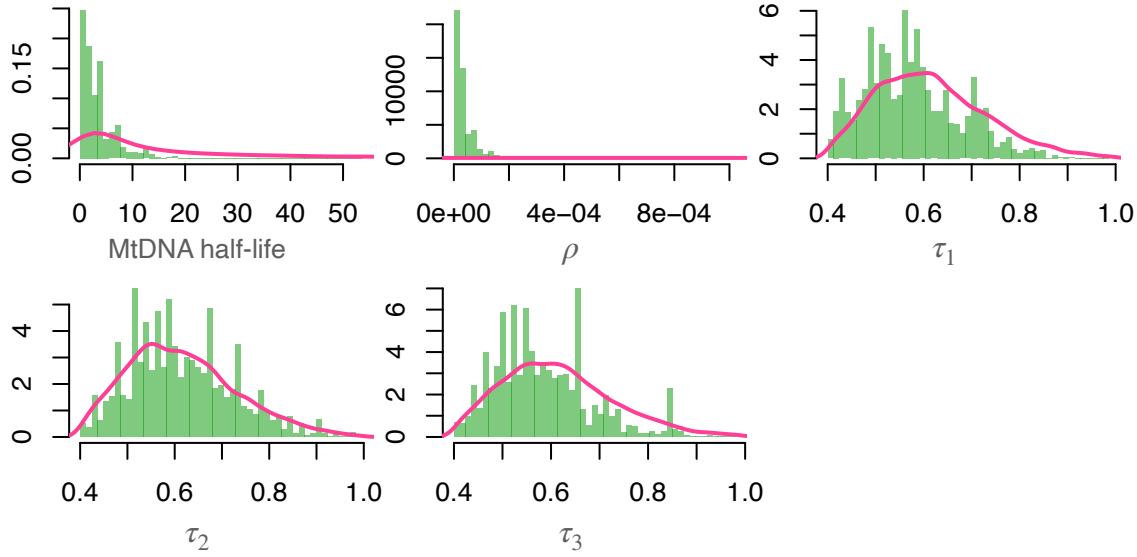
Posterior beliefs about the base replication rate, k_0, can be converted to the more interpretable half-life, and the resulting 95% posterior HDI is $(0.107, 10.530)$ days. This is a reduction in uncertainty from prior beliefs and contains the commonly used value in mathematical modelling of 10d, however, does not contain the experimental values 17.7d or 700d (Collins et al., 2003; Korr et al., 1998). The expected probability of mutation is within the range of previous estimates, although not for the same patient phenotype. The posterior beliefs of the pathogenic threshold are slightly lower than those found experimentally. Rocha et al. (2018) estimated the threshold for single, large-scale deletion patients to be between 56% to 82% in CI and 57% to 92% in CIV, corresponding to θ_3 and θ_5 respectively, which have 95% posterior HDIs of $[0.404, 0.759]$ and $[0.400, 0.773]$. Posterior beliefs on the biologically interpretable scales are shown for all parameters in Figure 5.6(b).

	$\mathrm{E}[\cdot\|\boldsymbol{X}_{\mathrm{Obs}}]$ (95% HDI)	$\mathrm{SD}(\cdot\|\boldsymbol{X}_{\mathrm{Obs}})$	$\mathrm{E}[e^{\theta}\|\boldsymbol{X}_{\mathrm{Obs}}]$ (95% HDI)	Parameter description
θ_1	$-12.642\,45$ $(-14.3373, -10.6555)$	0.999	3.232×10^{-6} $(1.773 \times 10^{-7}, 2.0483 \times 10^{-5})$	Base reaction (degradation) rate
θ_2	-10.6244 $(-12.866, -8.674)$	1.1315	2.432×10^{-5} $(7.607 \times 10^{-7}, 1.342 \times 10^{-4})$	Mutation probability
θ_3	-0.570 $(-0.881, -0.260)$	0.170	0.565 $(0.404, 0.759)$	Pathogenic thresh--old (NDUFB8)
θ_4	-0.510 $(-0.848, -0.192)$	0.173	0.600 $(0.428, 0.825)$	Pathogenci thresh--old (CYB)
θ_5	-0.574 $(-0.825, -0.170)$	0.165	0.563 $(0.400, 0.773)$	Pathogenic thresh--old (MTCO1)
ψ	1.160365 $(0.680, 1.712)$	0.268	- -	Model precision

Table 5.3: **Posterior expectations and 95% HDIs when fitting the mathematical model to the observed Vincent *et al.* dataset**. Inference for the model precision, ψ, is done on its natural scale, so values for its exponential transformation are not of interest. The posterior expectations, calculated by the median of the posterior draws, and 95% posterior HDI, calculated by the `R` package `HDIneterval`, are given on both scales (where appropriate) and indicated as such, after inferring model parameters, using the observed Vincent *et al.* dataset.



(a) Prior and posterior beliefs on scale of inference



(b) Prior and posterior beliefs on natural/interpretable scales

Figure 5.6: **Base reaction rate and mutation probability showed reduced uncertainty *a posteriori*.** (a) Prior (pink) and posterior (green) parameter beliefs on their inference scales for mathematical model parameters, $\theta_1, \ldots, \theta_5$, and model error, ψ. (b) Transformed parameter beliefs on biological interpretable scales. The base reaction rate has been converted to mtDNA half-life, and the remaining mathematical model parameters are shown on their natural scales. Model error is omitted here as it is shown on its natural scale in (a). Posterior beliefs are summarised by 100,000 draws from the posterior distribution. Prior beliefs are shown as kernel density estimates of the 100,000 samples from the prior distribution.

Despite the posterior uncertainty in the pathogenic threshold values, the posterior predictive distributions for NDUFB8 and MTC01 deficiency strongly resemble the data, matching the inter-patient variation well, Figure 5.7. However, the posterior predictive distribution for CYB does not resemble the data as strongly, with the 95% predictive interval being much wider than the observed proportions of deficiency. The modelling assumption that the variation in transformed OXPHOS deficiency proportions is constant

across proteins appears too strong. On the natural data-scale, the variation in OXPHOS deficiency proportion is notably different in CYB compared to the others.

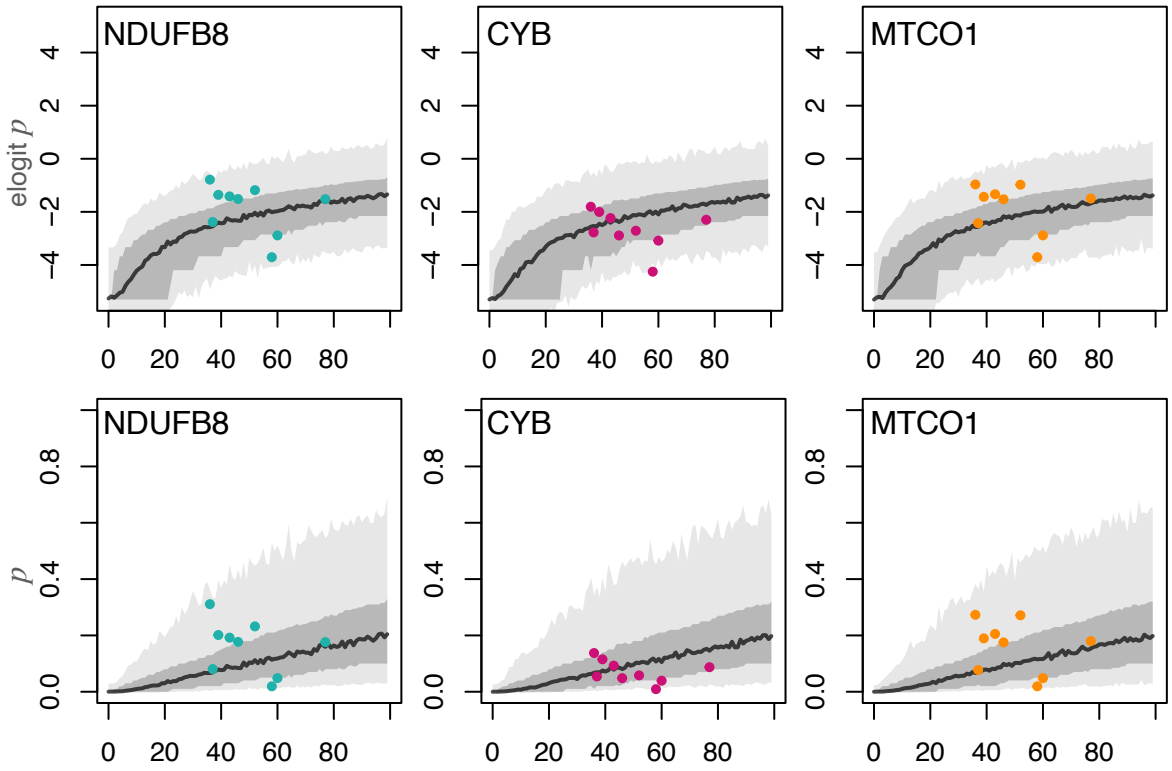


Figure 5.7: **Posterior predictive distributions match observed data well for NDUFB8 and MTCO1 proteins**. Posterior predictive intervals for all OXPHOS proteins (y-axis) throughout time in years (x-axis) in the observed dataset. Equitailed 95% posterior predictive probability intervals of the observed are represented as a light grey band, the posterior expected value is shown as a solid black line. The darker grey band shows the 95% posterior probability interval of the unobserved proportion of deficiency. Data for each protein is shown as coloured points. Predictive quantiles were made from 500 independent realisations of the mathematical model, each with a different set of parameter values, chosen to be equidistant intervals from the posterior draws.

5.4.2 Synthetic data

It is hoped that the increased dataset size of the synthetic datasets decreases the posterior uncertainty in parameter values compared to the observed dataset. Unfortunately, due to the computational expense of inference, it was not possible to investigate the dataset size and its effect on parameter uncertainty robustly. Nevertheless, it can be informally gleaned here.

In general, inference for the synthetic datasets produced less correlated output than the observed dataset, with multivariate ESSs ranging from 611 to 1,168 for the three datasets in question. The computational cost of inference was somewhat varied. When fitting the model to $\boldsymbol{X}_{\text{RGD}}$ and $\boldsymbol{X}_{\text{SoS}}$ inference time was approximately equivalent to when fitting the model to the observed dataset. However, the inference time was much shorter when fitting the model to $\boldsymbol{X}_{\text{PNN}}$. The reduction was largely due to the area of the parameter space the schemes were exploring. When fitting the model to $\boldsymbol{X}_{\text{PNN}}$, the

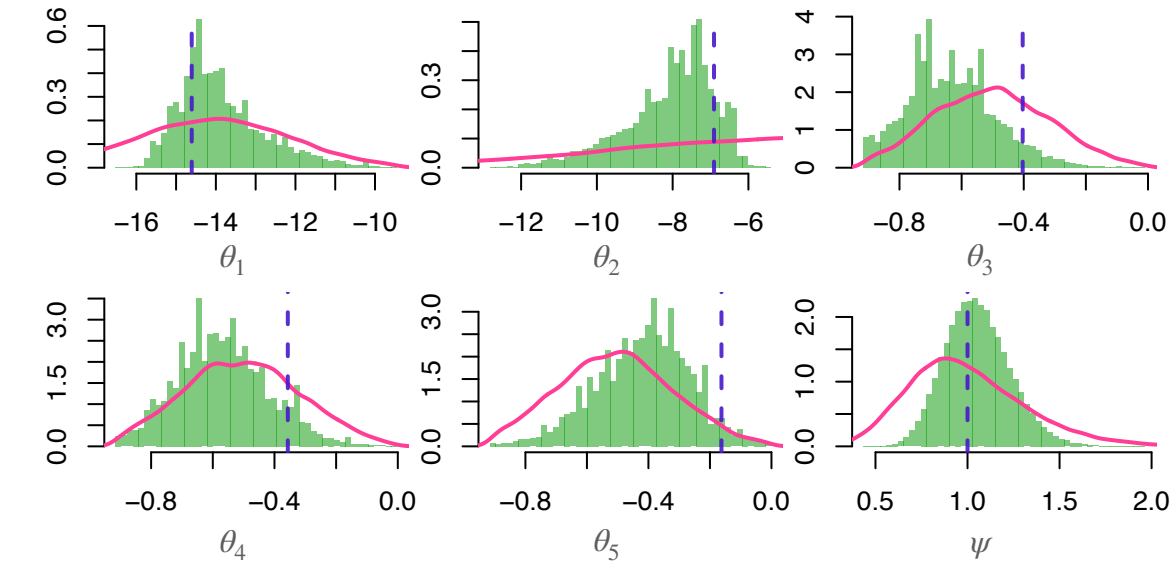
posterior beliefs for base reaction rate, k_0, were lower than the others, resulting in faster simulations and quicker inference. Inference for this dataset was still restricted to a single chain for consistency. No chains showed signs of non-convergence and were executed for 100,000 iterations, including a 20,000 iteration burn-in period.

RGD

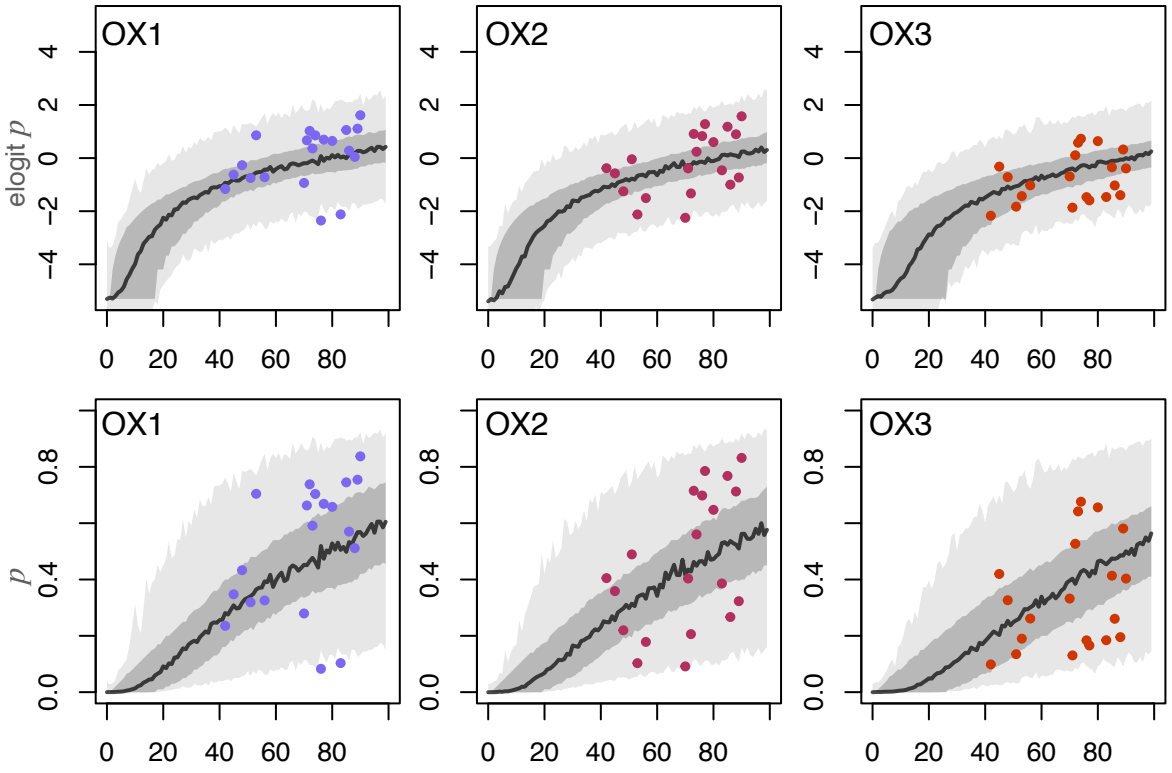
The inference successfully retrieves ground-truth parameter values used to generate the synthetic dataset, as all 95% posterior HDIs contained their ground-truth value, see Table 5.4. The posterior uncertainty in the pathogenic threshold values is slightly reduced compared to the inference based on the observed dataset. However, the posterior uncertainty in the other parameter values remains approximately the same. Despite the reduction in uncertainty, no substantial difference is found between the pathogenic threshold values *a posteriori*. However, the probability of θ_5 being larger than θ_3 and θ_4 is 0.892 and 0.814, respectively, mirroring its larger ground-truth value. Similarly, $\Pr(\theta_4 > \theta_3 | \boldsymbol{X}_{\text{RGD}}) = 0.694$. A large negative correlation is also found between θ_1 and θ_2 *a posteriori*, akin to that found in the posterior for $\boldsymbol{X}_{\text{Obs}}$. The posterior predictive distributions on the elogit scale closely resemble the data for proteins OX1 and OX2, but the resemblance is less profound in OX3, as shown in Figure 5.8. As noted, the synthetic data shows similar progressions for OX1 and OX2, due to their close ground-truths of the pathogenic threshold. The synthetic protein OX3 possesses a higher ground-truth pathogenic threshold, as reflected in the data, which shows lower OXPHOS deficiency proportions and with less variance.

	θ^*	$\mathrm{E}[\cdot\|\boldsymbol{X}_{\text{RGD}}]$ (95% HDI)	$\mathrm{SD}(\cdot\|\boldsymbol{X}_{\text{RGD}})$	e^{θ^*}	$\mathrm{E}[e^{\theta}\|\boldsymbol{X}_{\text{RGD}}]$ (95% HDI)
θ_1	-14.61	-14.081 $(-15.538, -11.496)$	1.080	4.53×10^{-7}	7.671×10^{-7} $(7.633 \times 10^{-8}, 8.466 \times 10^{-6})$
θ_2	-6.91	-7.822 $(-10.661, -6.303)$	1.184	1×10^{-3}	4.009×10^{-4} $(2.684 \times 10^{-6}, 1.410 \times 10^{-3})$
θ_3	-0.40	-0.641 $(-0.910, -0.380)$	0.141	0.668	0.527 $(0.402, 0.684)$
θ_4	-0.36	-0.572 $(-0.839, -0.268)$	0.147	0.70	0.564 $(0.414, 0.739)$
θ_5	-0.16	-0.405 $(-0.740, -0.145)$	0.151	0.85	0.667 $(0.472, 0.857)$
ψ	1.00	1.043 $(0.720, 1.404)$	0.176	-	- -

Table 5.4: **Posterior expectations and 95% HDIs, RGD dataset**. The ground-truth parameter values on the log- and natural scales, θ^* and e^{θ^*}, are given for $\theta_1, \theta_2, \theta_3, \theta_4$, and θ_5, representing the base reaction rate, mutation probability and pathogenic thresholds, respectively. Inference for the model precision, ψ, is done on its natural scale, so values for its exponential transformation are not required. The posterior expectations, calculated by the median of the posterior draws, and 95% posterior HDI, calculated by the `R` package `HDIneterval`, are given on both scales (where appropriate) and indicated as such, after inferring model parameters, using the synthetic RGD dataset.



(a) Prior and posterior beliefs



(b) Posterior predictive distributions

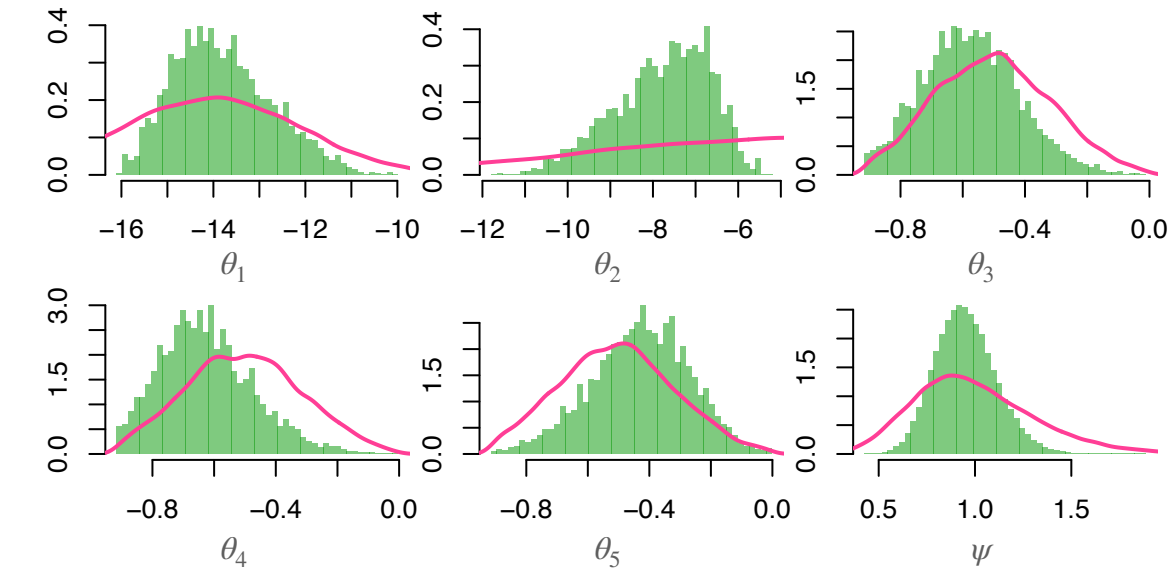
Figure 5.8: **Bayesian inference recovers ground-truth parameter values and posterior predictives show close resemblance to data**. (a) Prior densities (pink) and posterior beliefs (green) for all parameters in the model. Ground-truth parameter values are indicated with a dashed purple line. (b) Equitailed 95% posterior predictive probability intervals for the proportion of OXPHOS deficient myofibres (y-axis) throughout time in years (x-axis), on the elogit (top) and natural (bottom) scales, are represented as a light grey band. The posterior expected value is shown as a solid black line. The darker grey band shows the 95% posterior probability interval of the latent proportion of deficiency. Posterior predictions were calculated using independent simulations for 500 sets of parameter values, chosen to be equally spaced throughout the posterior draws. The synthetic proportion of deficiency data is shown as coloured points.

SoS

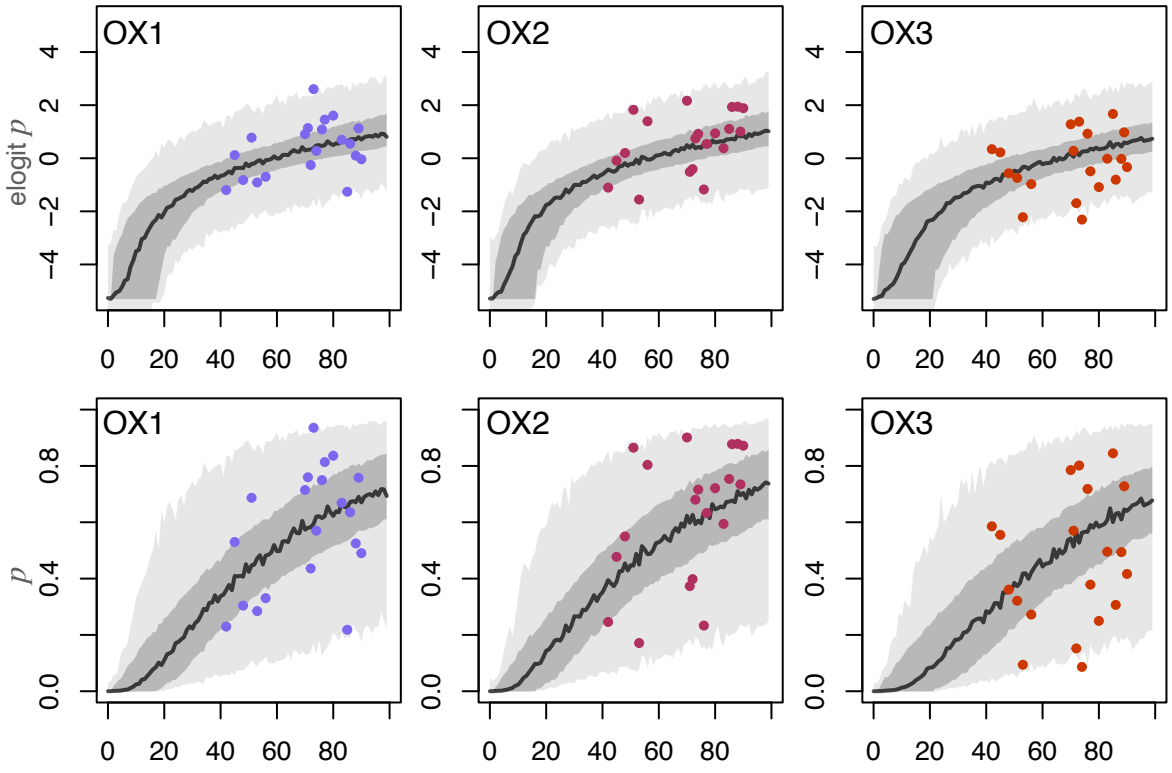
It is not appropriate to compare the posterior beliefs to ground-truth parameter values here because a different model was used to generate the data; therefore, they are not comparable. However, the posteriors can still be inspected for uncertainty and model fit. Posterior beliefs show a similar degree of uncertainty to the posteriors when fitting the model to the synthetic RGD dataset, $\boldsymbol{X}_{\text{RGD}}$, yet beliefs on all paramters have been updated, see Table 5.5 and Figure 5.9. The posterior beliefs follow the same story as $p(\boldsymbol{\theta}|\boldsymbol{X}_{\text{RGD}})$; a strong negative corrlation between θ_1 and θ_2, and no substantial differences between θ_3, θ_4, and θ_5. The pathogenic thresholds similarly showed relatively large posterior probabilities of the θ_5 being larger than θ_3 and θ_4, 0.791 and 0.833, respectively. The predictive distributions also mirror the results of the previous inference, showing a better fit to OX1 and OX2 when compared to OX3. This is perhaps expected given the similarities in the two synthetic datasets.

	$\text{E}[\cdot\|\boldsymbol{X}_{\text{SoS}}]$ (95% HDI)	$\text{SD}(\cdot\|\boldsymbol{X}_{\text{SoS}})$	$\text{E}[e^{\theta}\|\boldsymbol{X}_{\text{SoS}}]$ (95% HDI)
θ_1	-13.972 $(-15.691, -11.663)$	1.075	8.551×10^{-7} $(1.059 \times 10^{-7}, 7.190 \times 10^{-6})$
θ_2	-7.571 $(-10.038, -5.864)$	1.130	5.152×10^{-4} $(2.004 \times 10^{-5}, 2.100 \times 10^{-3})$
θ_3	-0.641 $(-0.880, -0.287)$	0.154	0.558 $(0.406, 0.736)$
θ_4	-0.64373 $(-0.907, -0.346)$	0.1496	0.525 $(0.403, 0.706)$
θ_5	-0.423 $(-0.752, -0.142)$	0.156	0.655 $(0.466, 0.861)$
ψ	0.944 $(0.651, 1.264)$	0.158	- -

Table 5.5: **Posterior expectations and 95% HDIs, SoS dataset**. The ground-truth parameter values on the log- and natural scales, θ^* and e^{θ^*}, are given for $\theta_1, \theta_2, \theta_3, \theta_4$, and θ_5, representing the base reaction rate, mutation probability and pathogenic thresholds, respectively. Inference for the model precision, ψ, is done on its natural scale, so values for its exponential transformation are not required. The posterior expectations, calculated by the median of the posterior draws, and 95% posterior HDI, calculated by the `R` package `HDIneterval` are given on both scales (where appropriate) and indicated as such, after inferring model parameters, using the synthetic RGD dataset.



(a) Prior and posterior beliefs



(b) Posterior predictive distributions

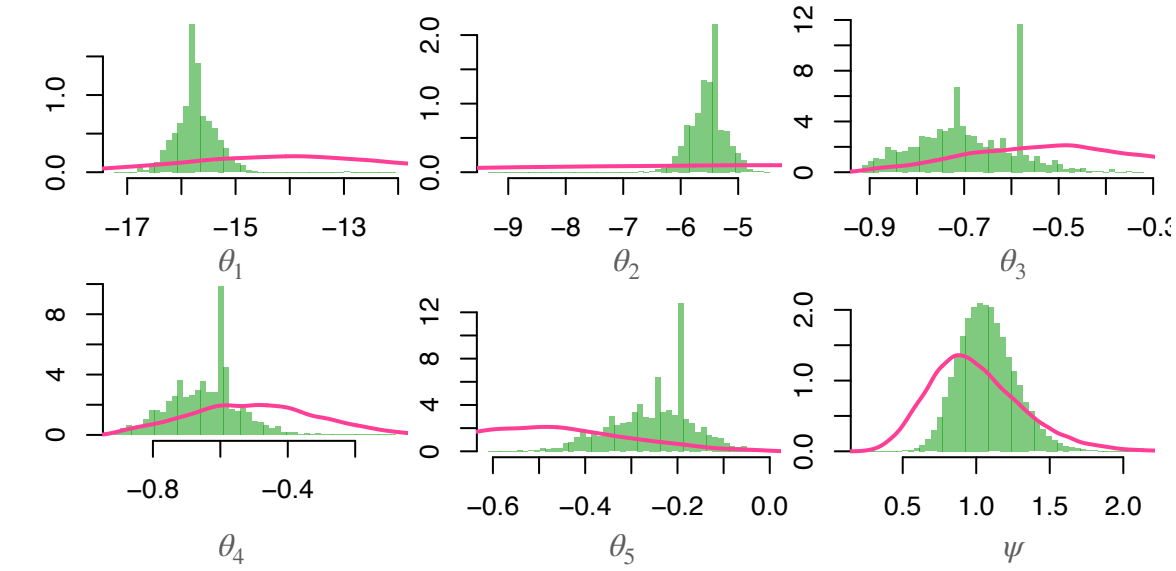
Figure 5.9: **Posterior parameter beliefs and predictive distribution when fitting model of RGD to synthetic SoS dataset**. (a) Prior densities (pink) and posterior beliefs (green) for all parameters in the model. (b) Equitailed 95% posterior predictive probability intervals for the proportion of OXPHOS deficient myofibres (y-axis) throughout time in years (x-axis), on the elogit (top) and natural (bottom) scales, are represented as a light grey band. The posterior expected value is shown as a solid black line. The darker grey band shows the 95% posterior probability interval of the latent proportion of deficiency. Posterior predictions were calculated using independent simulations for 500 sets of parameter values, chosen to be equally spaced throughout the posterior draws.

PNN

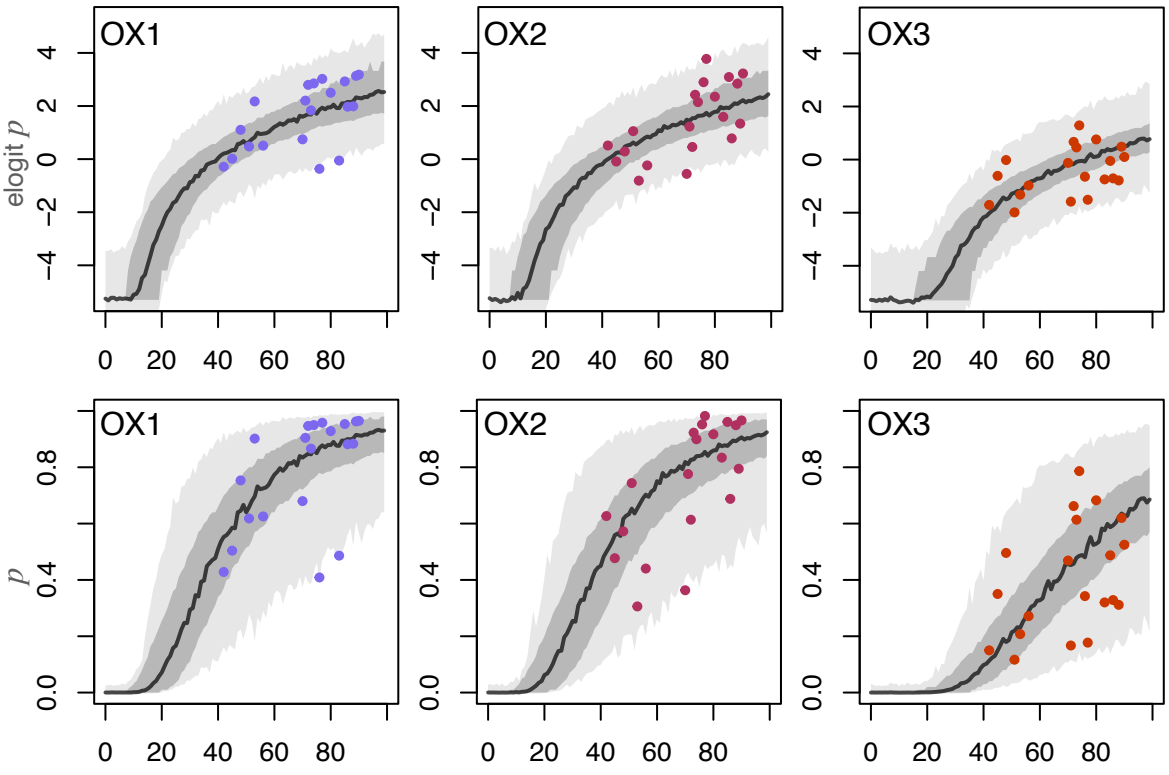
When fitting the model to the synthetic perinuclear niche data, $\boldsymbol{X}_{\text{PNN}}$, the posteriors appear markedly different when compared to other synthetic datasets. Notably, there is distinctly less uncertainty *a posteriori*, Table 5.6. Additionally, posterior beliefs of the base reaction rate and mutation probability differ substantially from the ground-truth parameters used to generate the data. The model was able to distinguish between the pathogenic thresholds, finding a substantial difference between θ_3 and θ_5, and between θ_4 and θ_5, when examining the 99% posterior HDIs of their difference distributions. No substantial difference was found θ_3 and θ_4, although it was found that $\Pr(\theta_4 > \theta_3 | \boldsymbol{X}_{\text{PNN}}) = 0.638$. The decrease in posterior uncertainty has also led to noticeable differences in the predictive distributions between proteins, Figure 5.10. The predictive distributions for all three proteins match the data well, and, significantly, the predictive distribution for OX3 is noticeably distinct from the others, predicting a slower accumulation of deficient myofibres as would be expected with a higher pathogenic threshold.

	$\text{E}[\cdot \mid \boldsymbol{X}_{\text{PNN}}]$ (95% HDI)	$\text{SD}(\cdot \mid \boldsymbol{X}_{\text{PNN}})$	$\text{E}[e^{\theta} \mid \boldsymbol{X}_{\text{PNN}}]$ (95% HDI)
θ_1	-15.7662 $(-16.447, -15.037)$	0.352	1.422×10^{-7} $(6.291 \times 10^{-8}, 2.704 \times 10^{-7})$
θ_2	-5.46514 $(-6.077, -4.897)$	0.298	4.232×10^{-3} $(2.061 \times 10^{-3}, 6.914 \times 10^{-3})$
θ_3	-0.691 $(-0.885, -0.513)$	0.106	0.501 $(0.402, 0.5856)$
θ_4	-0.637 $(-0.835, -0.428)$	0.102	0.529 $(0.410, 0.620)$
θ_5	-0.240 $(-0.442, -0.102)$	0.0913	0.786 $(0.643, 0.903)$
ψ	1.060 $(0.717, 1.461)$	0.193	- -

Table 5.6: **Posterior expectations and 95% HDIs, PNN dataset**. The ground-truth parameter values on the log- and natural scales, θ^* and e^{θ^*}, are given for $\theta_1, \theta_2, \theta_3, \theta_4$, and θ_5, representing the base reaction rate, mutation probability and pathogenic thresholds, respectively. Inference for the model precision, ψ, is done on its natural scale, and so values for its exponential transformation are not required. The posterior expectations, calculated by the median of the posterior draws, and 95% posterior HDI, calculated by the `R` package `HDIneterval`, are given on both scales (where appropriate) and indicated as such, after inferring model parameters, using the synthetic RGD dataset.



(a) Prior and posterior beliefs



(b) Posterior predictive distributions

Figure 5.10: **Posterior parameter beliefs and predictive distribution when fitting model of RGD to synthetic PNN dataset**. (a) Prior densities (pink) and posterior beliefs (green) for all parameters in the model. (b) Equitailed 95% posterior predictive probability intervals for the proportion of OXPHOS deficient myofibres (y-axis) throughout time (x-axis), on the elogit (top) and natural (bottom) scales, are represented as a light grey band. The posterior expected value is shown as a solid black line. The darker grey band shows the 95% posterior probability interval of the latent proportion of deficiency. Posterior predictions were calculated using independent simulations for 500 sets of parameter values, chosen to be equally spaced throughout the posterior draws.

Predictions of mtDNA dynamics

Synthetic data allows the model predictions to be compared to the unobserved mtDNA dynamics driving the OXPHOS deficiency. Figure 5.11 shows the posterior predictive quantiles of mtDNA variant load, for each synthetic dataset: $\boldsymbol{X}_{\text{RGD}}$, $\boldsymbol{X}_{\text{SoS}}$, and $\boldsymbol{X}_{\text{PNN}}$, compared to the true underlying dynamics used to generate the data. The posterior predictive quantiles show a fairly close match for $\boldsymbol{X}_{\text{RGD}}$ and $\boldsymbol{X}_{\text{SoS}}$, but with a slightly higher variance. The posterior expected variant load lies within the true 50% quantile for both datasets, although interestingly, both have consistently lower posterior expected values. The posterior predictions compared to the perinuclear niche hypothesis show the least agreement among the datasets. This is likely due to the difference between the underlying and fitted models. The true dynamics exhibit a stepped pattern as individual z-bands reach variant homoplasmy, before spreading or expanding into other bands, which is not reflected in the predictions made by the mathematical model. The differences in the predicted and true mtDNA dynamics of the perinuclear niche data, $\boldsymbol{X}_{\text{PNN}}$, highlight the versatility of a model random genetic drift. Despite a clear difference in the predicted mtDNA dynamics, the model shows a good fit to the OXPHOS deficiency data.

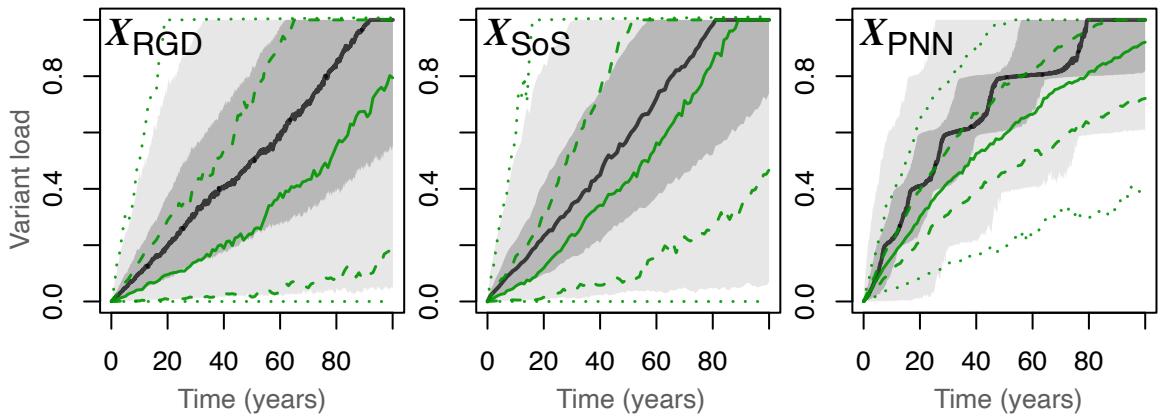


Figure 5.11: **Posterior predictions of mtDNA dynamcis**. The posterior predictive quantiles of the variant load given the three synthetic datasets: $\boldsymbol{X}_{\text{RGD}}$, $\boldsymbol{X}_{\text{SoS}}$, and $\boldsymbol{X}_{\text{PNN}}$. The posterior predictions are summarised by their 2.5% (dotted), 25% (dashed), 50% (solid), 75% (dashed), 97.5% (dotted) quantiles, giving the 95%, interquartile range and posterior expected values, shown in green. The true underlying mtDNA dynamics used to generate the synthetic datasets are summarised by the same quantiles, represented as grey bands and the expected value as a solid black line. Quantiles of the true system dynamics were calculated from 1,000 independent simulations of the system with the ground-truth parameters described in Table 5.2. Posterior quantiles were calculated from 500 independent simulations using equidistant parameter values from the posterior.

5.4.3 Model comparison

The mathematical model of RGD used in this chapter showed a good fit to the proportion of deficiencies in the synthetic datasets; however, clear differences are seen when comparing the unobserved mtDNA dynamics, indicating that the RGD theory of clonal expansion can describe OXPHOS deficiency data derived from other theories and biological mechanisms. An inference scheme that infers the most likely mathematical model of clonal expansion based upon some criteria could be constructed, although it was computationally infeasible to do so here. Instead, the posterior probabilities of observing the data given the model are calculated. This is the basis of the model comparison criteria BIC (Schwarz, 1978), which gives weight to models with higher likelihoods and penalises complex models with more parameters.

Figure 5.12 shows the posterior probabilities of observing the datasets given the mathematical model, $p(\boldsymbol{X}|M)$. This is equivalent to the posterior data likelihood after integrating out parameter uncertainty. Although similar, the probability of observing $\boldsymbol{X}_{\text{RGD}}$ is larger than observing $\boldsymbol{X}_{\text{SoS}}$ or $\boldsymbol{X}_{\text{PNN}}$, with probabilities 0.876 and 0.686 respectively. The larger posterior likelihood indicates that the mathematical model provides the best fit for $\boldsymbol{X}_{\text{RGD}}$. This may not be surprising, given that the mathematical model is a model of RGD, however, posterior beliefs and predictive distributions showed less posterior uncertainty when fitting the model to $\boldsymbol{X}_{\text{PNN}}$ and the posterior predictives of deficiency proportion showed a slightly better fit for OX3 within this dataset as well. Additionally, $\boldsymbol{X}_{\text{SoS}}$ and $\boldsymbol{X}_{\text{RGD}}$ exhibit a strong resemblance, yet a noticeable difference is seen within the data likelihoods.

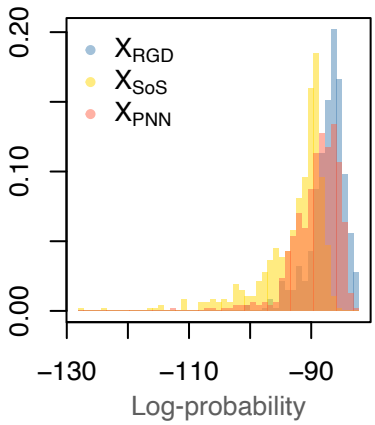


Figure 5.12: **Posterior probabilty of observing $\boldsymbol{X}_{\text{RGD}}$ is highest amongst synthetic datasets**. The posterior log-probability density when fitting the mathematical model of RGD to the synthetic datasets: $\boldsymbol{X}_{\text{RGD}}, \boldsymbol{X}_{\text{SoS}}$, and $\boldsymbol{X}_{\text{PNN}}$.

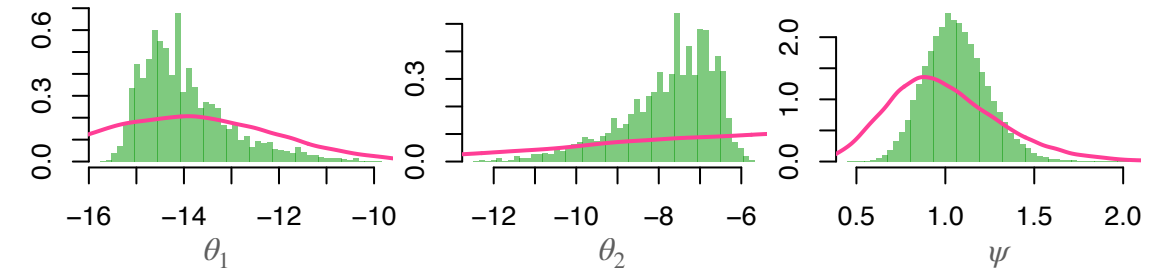
5.4.4 Fixing parameters

It may be possible to fix some of the parameters of a mathematical model using appropriate estimates from experimental datasets. Estimates of the pathogenic threshold are more prevalent within the literature, so they were chosen to be fixed at their ground-truth values. The remaining parameters were inferred using the RGD dataset, $\boldsymbol{X}_{\text{RGD}}$. Due to constraints, only a single chain of the model inference was executed with 60,000 iterations, including a 10,000 burn-in period. The resulting posterior showed improved

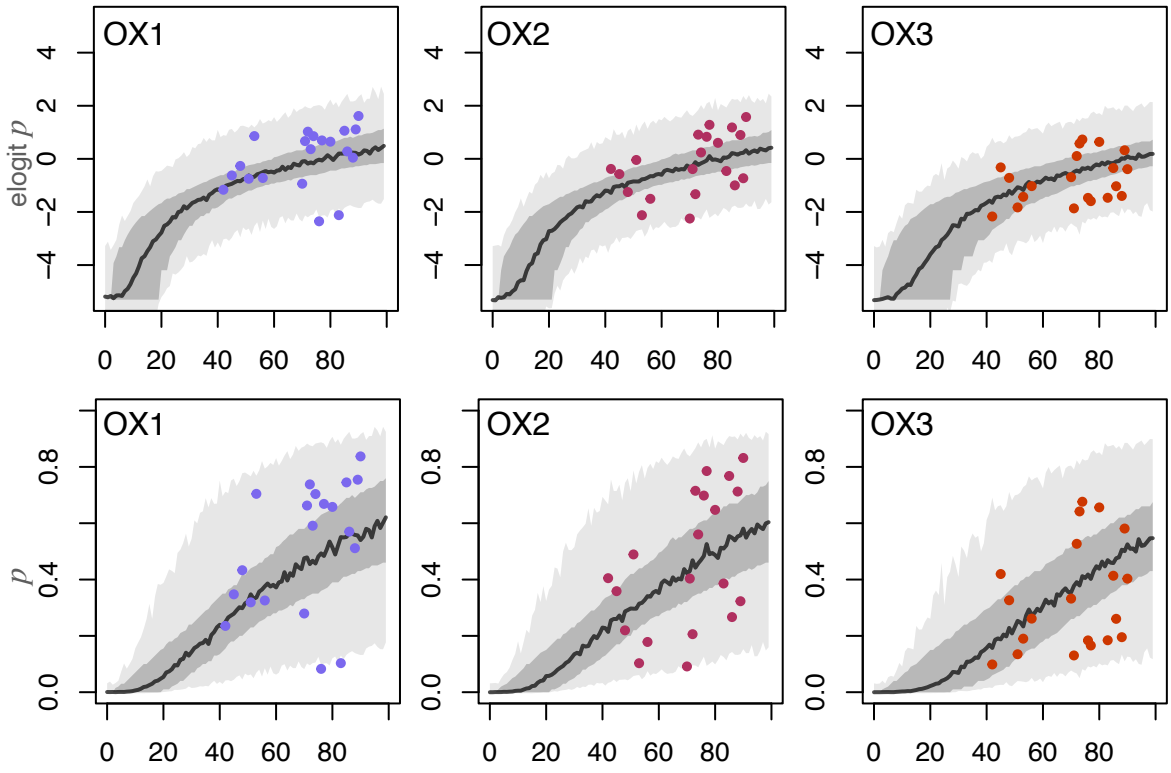
sampling compared to previous chains, with a much higher multivariate ESS of 1,485. However, the posterior uncertainty is (approximately) unchanged in parameter values and model predictive, Table 5.7 and Figure 5.13. The posterior likelihood is also approximately unchanged, and, unfortunately, fixing the pathogenic thresholds has not improved model fit. However, it did aid inference by drastically increasing multi- and single-variate ESSs.

	$\mathrm{E}[\theta \vert \boldsymbol{X}_{\mathrm{RGD}}]$ (95% HDI)	$\mathrm{SD}(\theta \vert \boldsymbol{X}_{\mathrm{RGD}})$	$\mathrm{E}[e^{\theta} \vert \boldsymbol{X}_{\mathrm{RGD}}]$ (95% HDI)
θ_1	-14.1834 $(-15.418, -11.946)$	0.951	6.922×10^{-7} $(1.6025 \times 10^{-7}, 6.104 \times 10^{-6})$
θ_2	-7.551 $(-10.140, -6.033)$	1.142	5.257×10^{-4} $(3.760 \times 10^{-6}, 16.837 \times 10^{-3})$
ψ	1.051575 $(0.731, 1.420)$	0.177	- -

Table 5.7: **Posterior expections and 95% HDIs**. The ground-truth parameter values on the log- and natural scales, θ^* and e^{θ^*}, for θ_1 and θ_2, representing the base reaction rate and the mutation load, respectively. Inference for the model precision, ψ, is done on its natural scale, so values for its exponential transformation are not required. The posterior expectations, calculated by the median of the posterior draws, and 95% posterior HDI, calculated by the `R` package `HDIneterval` are given on both scales (where appropriate) and indicated as such, after inferring model parameters, using the synthetic RGD dataset.



(a) Prior and posterior beliefs



(b) Posterior predictive distributions

Figure 5.13: **Fixing pathogenic thresholds maintains model**. (a) Posterior and prior beliefs for remaining parameters, when fixing pathogenic thresholds to their ground-truth values, and fitting the mathematical model to RGD dataset, $\boldsymbol{X}_{\text{RGD}}$. (b) Equitailed 95% posterior predictive probability intervals for the proportion of OXPHOS deficient myofibres (y-axis) throughout time in years w(x-axis), on the elogit (top) and natural (bottom) scales, are represented as a light grey band. The posterior expected value is shown as a solid black line. The darker grey band shows the 95% posterior probability interval of the latent proportion of deficiency. Posterior predictions were calculated using independent simulations for 500 sets of parameter values, chosen to be equally spaced throughout the posterior draws.

5.5 Discussion

5.5.1 Key findings

In this chapter, it was shown that a mathematical model of RGD can well describe synthetic OXPHOS deficiency data generated from a range of models. The unobserved mtDNA dynamics of each dataset were also compared to the posterior predictions of the

mathematical model. The model showed a fairly strong resemblance to two of the three synthetic datasets, $\boldsymbol{X}_{\text{RGD}}$ and $\boldsymbol{X}_{\text{SoS}}$. Despite the close fit of the OXPHOS deficiency data generated by the perinuclear niche model, $\boldsymbol{X}_{\text{PNN}}$, the predicted mtDNA dynamics showed substantial differences from the ground-truth. However, it was found that the synthetic dataset with the highest posterior likelihood was generated by a mathematical model of RGD. This might seem unsurprising; however, $p(\boldsymbol{\theta}|\boldsymbol{X}_{\text{PNN}})$ showed lower posterior uncertainty in parameter values, and the predictive distributions showed a closer resemblance to the data. The higher likelihood of the $\boldsymbol{X}_{\text{RGD}}$ indicates that it may be possible to distinguish between models of clonal expansion using OXPHOS deficiency data and a model selection criterion during Bayesian inference.

For the first time, a mathematical model of perinuclear niche theory was developed. Although the model described is specific to the PNN assumptions, it lays the foundation of other spatially dependent models and could be adapted to suit a variety of modelling assumptions.

The investigation highlights the difficulties when inferring the pathogenic threshold, compared to mtDNA half-life and mutation probability. The threshold posterior beliefs were scarcely updated from their priors in both the observed and synthetic datasets. A noticeable reduction in the posterior uncertainty of the pathogenic thresholds was seen between the observed dataset and all synthetic datasets, and it is hypothesised that the increased amount of data in the synthetic datasets could be driving this reduction.

5.5.2 Limitations

Computational expense

The investigation was hindered by the computational expense of inference, which is primarily driven by the cost of simulating the mathematical models. This is particularly true for the mathematical model of the perinuclear niche theory. The spatial dependence of the model requires agent-based modelling, which drastically increases the number of computations per simulation update. Even so, the computational cost of simulating the non-spatial models is still considerable when implementing Bayesian inference. Therefore, it was not possible, within an appropriate time frame, to construct and implement a large Bayesian inference scheme which included model selection and comparison. The expense was so great that it was not possible to execute multiple inference chains to ensure MCMC convergence to a single posterior distribution. The simulation cost during inference could be greatly reduced by using an emulator model (Henderson et al., 2009). Unfortunately, the project's time limitations did not allow for this. The cost of inference further impacted the simulation study. The reduction in posterior uncertainty of the pathogenic thresholds is attributed to the larger size of the synthetic data compared to the smaller observed dataset. However, this should be confirmed by a simulation study of how dataset size impacts posterior uncertainty and model fit.

Available data

Finding appropriate data for an investigation into clonal expansion is difficult. Here, the choice was made to use patients with mtDNA maintenance disorders, and it was assumed that the mtDNA mutation events did not occur during cell development. This assumption gave the simulations a known initial condition of zero variant load. Without this assumption, or prior information about the patients' inherited variant load, it would be almost impossible to distinguish between a system with deficiency proportion caused by inherited variant load or by mtDNA dynamics, based on a single time point. However, it is possible variant mtDNA arose post-conception, during cell development. Including cellular development in the model is possible, but would significantly increase model complexity and computational cost. Without knowledge of the inherited variant load, meaningful inference could be achieved by a longitudinal study. However, ethical considerations often prevent longitudinal studies that collect multiple tissue samples throughout a patient's life. Furthermore, it is not known how many tissue samples or time points are needed, nor what the optimal inter-collection time would be for inference. Another implication of using patients with an mtDNA maintenance disorder is that they are likely to have different pathogenic mtDNA variants, which could lead to greater inter-patient variability in their OXPHOS deficient levels due to the variant-specific nature of the pathogenic threshold and resulting OXPHOS deficiency.

Additionally, the observed dataset used here contained no information regarding mtDNA copy number and the decision was made to use values found in the literature. However, these may not be biologically realistic, and inappropriate copy number values could affect the inferred parameter values, as they impact the rate of clonal expansion (Elson et al., 2001).

Observed dataset

The observed dataset used in this chapter was relatively small, containing nine patients, each with three OXPHOS deficiency measurements collected from a single tissue sample. A larger dataset, with more patients, would likely be able to reduce some of the posterior uncertainty seen here. Although not a perfect comparison, the synthetic datasets contained 20 patients, and a noticeable reduction in posterior variance of the pathogenic thresholds was observed, although the same effect was not seen in other parameters. The computational cost of inference prevented conducting a further simulation study with a range of dataset sizes to properly investigate its impact on posterior uncertainty.

Model fit

The mathematical models proposed here assumed that the variation in deficiency proportions was due to natural variation between patients and consistency across proteins. However, this may not be the case. After fitting the model to the observed dataset, the posterior predictive intervals highlight that the variation in CYB deficiency proportion is much lower than that of NDUFB8 and MTCO1. This indicates that the previous assumption may not be appropriate. The statistical model fitted to the transformed proportion data, Equation 5.15, could be updated to allow the model precision to vary between

OXPHOS proteins. Unfortunately, due to computational expense, it was not possible to update the model and re-run the inference.

5.5.3 Future work

Dataset

The observed dataset used in this chapter was relatively small. A larger dataset is likely required to investigate the mechanisms of clonal expansion fully. An increase in data points can occur in several ways. Continuing to work with mtDNA maintenance disorder patients, multiple observations can be made at a single time point from a patient using post-mortem tissue. Ideally, all samples would be collected from the same tissue; otherwise, inter-tissue variability must be taken into account. Post-mortem tissue would increase the number of observations without additional computational expense, as more observations from a single patient would not increase the number of simulations required during inference. Another option would be to conduct a longitudinal study, collecting multiple tissue samples throughout a patient's life. Ethical considerations would likely limit this to one sample per time point and tissue. The benefits of a longitudinal study are that it does not necessarily require patients with mtDNA maintenance disorders, and other patient genotypes can be considered. However, if two or more patient phenotypes are to be considered in a single dataset, the probability of a mutation event and the pathogenic threshold for each phenotype must be considered. A longitudinal study may allow for the inherited variant load to be inferred for patients where this is unknown and cannot be assumed. A robust simulation study would elucidate whether this is possible.

OXPHOS deficiency data from separate sources can be combined to form a single large dataset. It is likely that when combining datasets, not all patients have measurements on the same OXPHOS proteins. However, the pool of proteins in which data is collected is relatively small, so a large amount of overlap is expected. Indeed, unobserved OXPHOS deficiency proportions could be considered latent variables and be inferred during inference. When combining datasets, it must be decided whether it is reasonable to expect the same model precision for all datasets, and if not, this consideration must be taken into account.

Single-myofibre copy number and control mechanisms were not considered in this investigation, as they were assumed to be known, using literature standard values. Copy number affects the rate of clonal expansion (Elson et al., 2001), and so its value is important when investigating clonal expansion. Incorporating copy number data would enable the inference of control mechanism parameters and potentially facilitate the selection of specific mechanisms. Ideally, single-myofibre copy numbers would be collected from the same patients as the OXPHOS deficiency measurements. This may be achieved by taking a tissue block, which can be split into separate samples and analysed by different means. One tissue slice would be used for OXPHOS protein abundances and another to find single-myofibre mtDNA copy number. Using the same patients for copy number and OXPHOS deficiency would reduce complexities regarding phenotype and inter-patient variability, which would otherwise have to be considered.

Improving statistical model

The statistical model could be updated to allow varying model precision between OXPHOS proteins, which would hopefully provide a better fit to the data *a posteriori*. Depending on the size of the available dataset, it may be appropriate to implement a hierarchical aspect to the model, which infers both population-level and protein-specific model precision. With the addition of the precision hierarchy, the statistical model would be of the form: for $i = 1, \ldots, N_{\text{pat}}$ and $j = 1, \ldots, N_{\text{OX}}$

$$\begin{aligned} Y_j^i | \hat{p}_j^i, \psi_j &\sim \text{N}\left(\text{elogit } \hat{p}_j^i, {\psi_j}^{-1}\right), \\ \psi_j | \alpha_\psi, \beta_\psi &\sim \text{Ga}\left(\alpha_\psi, \beta_\psi\right), \\ \alpha_\psi &\sim \text{Ga}\left(g_\alpha, h_\alpha\right), \\ \beta_\psi &\sim \text{Ga}\left(g_\beta, h_\beta\right). \end{aligned} \tag{5.19}$$

Decreasing computational cost

As discussed, a major obstacle to this investigation is the cost of inference, primarily driven by the need to simulate the mathematical model. It is, however, possible to reduce this by implementing a model emulator (Henderson et al., 2009). Emulation is a statistical method which infers the output of a mathematical model given a set of input parameters. This reduces the computational cost of inference by removing the need for simulation and instead drawing from the model's emulator. The reduction in computational cost of inference may allow for the formal comparison of varying mathematical models of clonal expansion, by criteria such as BIC. This would allow the model that provides the best fit to the data, based on the criterion, to be robustly identified, and consequently give weight to that theory of clonal expansion.

Simulation study

Reducing the computational cost would enable an in-depth simulation study, which could inform the data required to make meaningful inferences with a real dataset. A simulation study could also demonstrate the potential gains in posterior certainty and predictive power that an increase in the number of patients or patient samples could yield. Additionally, a simulation study can inform the optimal time between sample collections in longitudinal studies.

5.5.4 Final remarks

Many questions remain unanswered regarding the mechanisms behind clonal expansion, and mathematical models are likely to play a crucial role in elucidating these mechanisms. Unfortunately, several hurdles are currently hindering their use. Mainly, the computational cost of simulating such models for inference and the lack of appropriate and agreed-upon biological estimates of model parameters. Here, we showed that OXPHOS deficiency data can be used to infer model parameters, and, significantly, showed that the underlying system generating the data can be distinguished when using Bayesian inference to infer model parameters. Consequently, it was also demonstrated that a mathematical model of RGD can explain data generated from different underlying systems by

varying model parameters, indicating that conclusions based on standard literature values regarding model fit should be taken with caution. A reduction in simulation costs, achieved through model emulation, would enable the formal comparison of models and the development of more complex models to investigate copy number control mechanisms.

Chapter 6

Discussion

Mitochondrial disease is a rare genetic disorder which is caused by variants in both the nuclear and mitochondrial genomes. The wide variety of possible pathogenic DNA variants and the varied roles mitochondria play in cellular processes result in a diverse phenotypic spectrum both between and within pathogenic DNA variants. The stochastic and continuous turnover of mtDNA also gives rise to high cellular phenotypic variance, adding to the overall complexity of the disease. Despite their significance, the biological processes which govern mtDNA dynamics and mitochondrial dysfunction are poorly understood, making the development of therapies and treatments difficult. Several theories linking mtDNA dynamics and mitochondrial function have been proposed, although none have been conclusively agreed upon to date. Mathematical models of mtDNA dynamics have become a popular method for the proposal and comparison of biological theories. Unfortunately, the uncertainty in biological mechanisms mirrors uncertainty in the parameters governing mathematical models. Parameter inference is challenging due to the limited direct observations of mtDNA dynamics and the complexity of the mathematical models, rendering the likelihood intractable and making standard statistical techniques inapplicable. Investigations, therefore, use parameter estimates found in the literature. However, the species- and tissue-specific nature of mitochondrial function, as well as the variant-specific nature of the disease, makes this undesirable.

Robust investigations comparing mathematical models of clonal expansion could shed light on the governing processes of mitochondrial dynamics, making it easier to predict disease progression and develop therapies. Within the literature, relatively few investigations have directly compared theories of clonal expansion and those that have used parameter values found in the literature.

6.1 Key findings

Due to the combined difficulties of data availability and model comparison, Chapters 3 and 4 of this thesis examine two data types for their use in comparing mathematical models of clonal expansion. The practicalities of inferring parameter values and comparing models of clonal expansion using Bayesian inference and non-direct measurements of mtDNA dynamics are considered in Chapter 5.

6.1.1 Blood cell analysis

In Chapter 3, a Bayesian mixture model was used to infer the proportion of cells which had reached wild-type homoplasmy across several cell types within six patients harbouring the pathogenic m.3243A>G mtDNA variant. The data, believed to contain noise due to sequencing errors, consisted of single-cell variant loads. The analysis confirmed previous findings that m.3243A>G is negatively selected against over time within blood cells (Walker et al., 2020). Additionally, it was found that negative selection is enhanced during T cell development, with more mature cells having substantially higher levels of wild-type homoplamsy. The increased levels of homoplasmy, however, are not consistent with the linear differentiation model of T cell development (Geginat et al., 2003; Verma et al., 2017), with T_{EMRA} cells showing reduced homoplasmy levels compared to T_{EM} cells across patients. Instead, suggesting a fluid differentiation model.

6.1.2 OXPHOS status classification

In Chapter 4, the OXPHOS deficiency status of single myofibres was inferred using a Bayesian hierarchical model. The model showed improved classifications compared to the existing standard classification method (Rocha et al., 2015) when evaluated against expert manual classifications for an observed dataset, collected by Vincent et al. (2024). When comparing the two models to synthetic datasets, the Bayesian model showed higher agreement with the ground-truth values, exhibited high predictive power, and robustness to prior specification. However, when comparing the classifications to a second observed dataset, collected by Gomes et al. (2025), a problem was highlighted. The Bayesian model assumes that healthy myofibres' OXPHOS abundance shows a linear relationship to mitochondrial mass, which is distinct from unhealthy myofibres' abundances. One patient, out of three, in the Gomes dataset did not satisfy this assumption, resulting in misclassifications. This dataset also highlighted issues related to computational cost and convergence. The increased computational cost due to larger datasets may become prohibitive due to a complex joint-posterior distribution. Inference chains were shown to become stuck in local maxima, and an increased burn-in period exceeded the capabilities of a standard machine with 16GB RAM. This issue arose twice out of a combined of 219 independent executions of the model on data represented in a 2Dmito plot across the datasets.

6.1.3 Modelling clonal expansion

In Chapter 5, the practicalities of inferring the parameter values for a mathematical model of mtDNA dynamics were considered when using the proportion of OXPHOS deficient myofibre data. In total, four datasets were considered: one observed (Vincent et al., 2024) and three synthetically generated based on assumptions of RGD, SoS, and PNN theories of clonal expansion. For each dataset, a mathematical model of RGD was fitted, inferring model error and the mathematical model parameters. For the observed dataset, although the posterior distributions showed a fair degree of uncertainty, the posterior predictive distributions showed a good fit to the data for two of the three OXPHOS proteins. The remaining OXPHOS protein showed lower levels of inter-patient variation in the proportion of OXPHOS deficient myofibres, which was not reflected in the posterior predictive distributions. Unfortunately, time constraints and computational expense prohibited model

development and re-fitting. When fitting the model to the synthetic datasets, it was found that a model of random genetic drift could successfully replicate the data, proportions of OXPHOS protein abundance, generated by different underlying mechanisms. However, when comparing posterior marginal data likelihoods, the dataset generated under the assumptions of RGD showed the highest likelihood, implying the 'true' model could be inferred using model comparison techniques such as BIC. Model parameters showed reduced posterior uncertainty when being inferred using the synthetic datasets, compared to the observed; it is believed this is due to the increase in the number of data points within the synthetic datasets.

6.2 Future work

Recent advances in biological techniques could form part of a larger study into clonal expansion, collecting data which is used to infer the parameters of a mathematical model. For example, Lareau et al. (2023) recently developed a large-scale single-cell DNA sequencing technique that is capable of reading thousands of single cells' DNA sequences at once. They investigated mtDNA sequences of blood cells within six patients, similarly to the single-cell dataset seen in Chapter 3. However, Lareau et al. (2023) collected single-cell measurements on over 200,000 cells compared to approximately 6,000 cells in the dataset collected by Franklin et al. (2023). The large number of cells analysed also enabled Lareau *et al.* to collect data on rare cell types, allowing a more complete picture of cellular differentiation. Their method was applied to blood cells, raising questions about its applicability to skeletal muscle due to the latter's much larger size. More recently, Bury et al. (2024) developed a method to collect subcellular tissue samples and measure mtDNA variant load within the subcellular region of myofibres. Subcellular data, such as this, has not been available before and could provide interesting data of how variant mtDNA spreads throughout the myofibre.

Additionally, a number of statistical methodologies could be utilised to aid the development of statistical/mathematical models relating to clonal expansion. Henderson et al. (2010) used a mixed effects statistical model to combine two independent datasets measuring different aspects of mtDNA clonal expansion. The first dataset consisted of multiple variant load measurements made from aggregated neurons within a cohort of 15 patients with Parkinson's disease. The second dataset contained neuronal survival count data from a cohort of healthy controls. Bayesian inference was conducted to infer the parameters governing a mathematical model of mtDNA dynamics and showed considerable less uncertainty than when using a single dataset of just the variant load measurements (Henderson et al., 2010). Their work highlights the potential of combining independent datasets to provide a more comprehensive picture of the underlying dynamics of mtDNA. This is significant given the scarcity of data availability in rare diseases, and shows the potential for combining variant load measurements and OXPHOS deficiency. Independent datasets concerning different measurements of mtDNA dynamics, possibly collected from different patient phenotypes, could be combined within a single statistical model, allowing for a richer, fuller dataset to compare models of clonal expansion.

As mentioned in Chapter 5.5, emulation is a powerful tool for reducing the computational expense of simulating a stochastic mathematical model, see Baker et al. (2022) for a

review of emulator methods. In particular, Gaussian process emulators (GPEs) are widely used and have been applied to a range of mathematical models (Conti et al., 2009; Fisher et al., 2022; Vernon et al., 2014), including both Gillespie-type (Fisher et al., 2022; Henderson et al., 2010) and agent-based models (Oyebamiji et al., 2017). Building a model emulator, however, requires a large number of simulations from the mathematical model, which can still be expensive. One method to reduce the computation burden is improved hardware. Graphcore have developed a high-performance computer processor that can massively parallelise computation. A significant reduction in computing time and cost per simulation could be made by employing this hardware, as shown in Appendix C. Another method, commonly used in making an emulator, is referred to as history matching. An emulator is built by simulating model behaviour given a set of inputs which spans the parameter support, history matching removes unlikely areas of the parameter space, and focuses on regions of non-negligible posterior density (Andrianakis et al., 2017; Vernon et al., 2014; Williamson et al., 2013).

Combining recent advances in biological technology with advanced statistical methodology would enable an in-depth investigation into the mechanisms governing clonal expansion. If focus continues to be with clonal expansion within myofibres, single-myofibre variant load data could be used as a second independent dataset to OXPHOS deficiency proportion. The spatial dynamics of mtDNA could be inferred by incorporating subcellular data, by comparing model predictions to subcellular regions of the variant load and copy number. Otherwise, large-scale single-cell sequencing could be used to generate a comprehensive dataset of blood cells within the T and B cell compartments. As discussed in Chapter 3.5, a mathematical model of blood cell development is associated with several challenges but a large and rich dataset of single-cell variant loads and copy numbers, collected following the methods of Lareau et al. (2023), could be enough to overcome these. The ease of data collection in blood tissue would more easily allow for longitudinal studies, where multiple data points are collected for each patient. The use of model emulators and the advanced processing units could alleviate many of the computational problems that have been discussed in Chapter 5.5 and within this chapter.

6.3 Closing remarks

In conclusion, the work detailed in this thesis has implemented Bayesian analysis to infer the aspects of mtDNA dynamics and assess the use of OXPHOS deficiency data to infer the parameters of a mathematical model. The work also highlights the possibilities for future research comparing theories of clonal expansion using model emulators and varied data sources. This will hopefully contribute to ongoing research aimed at understanding the mechanisms governing mtDNA dynamics and clonal expansion, improving predictive capabilities of disease progression and severity.

Appendix A

Classification of myofibre OXPHOS status - Appendix

A.1 Full-conditional distribution calculations of Bayesian hierarchical linear mixture model

For k subjects ($k-1$ controls and one patient) represented in a 2Dmito plot, each subject possesses M_i $i = 1, 2, \ldots, k$ single myofibre protein abundance measurements. The full-conditional distribution is proportional to the full-joint distribution of all parameters and data, $p(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\pi})$. Let $f(y|\mu, \tau)$ represent the PDF of normal distribution with mean μ and variance τ. The joint density is,

$$\begin{aligned} p(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\pi}) &= p(\boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{\theta}, \boldsymbol{\pi})\, p(\boldsymbol{\theta}, \boldsymbol{\pi}), \\ p(\boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{\theta}, \boldsymbol{\pi}) &= \prod_{i=1}^{N} \prod_{j=1}^{M_i} \left\{ \pi f\left(Y_{ij}|m_i X_{ij} + c_i, \tau\right) + (1-\pi) f\left(Y_{ij}|m_i X_{ij} + c_i, \gamma\right) \right\}, \\ p(\boldsymbol{\theta}, \boldsymbol{\pi}) &= p(\boldsymbol{\pi}) p(\tau) p(\tau_m) p(\mu_m) p(\tau_c) p(\mu_c) \prod_i p(m_i|\mu_m, \tau_m) p(c_i|\mu_c, \tau_c). \end{aligned} \tag{A.1}$$

The FCD for each parameter can then be calculated by simplifying the full-joint distribution up to proportionality. We start with the expected values of the slope and intercept parameters. The FCD of the expected slope is

$$\begin{aligned} p(\mu_m|\boldsymbol{Y}, \boldsymbol{Z}, \ldots) &\propto p(\mu_m) \prod_{i=1}^{k} p\left(m_i|\mu_m, \tau_m\right), \\ &\propto \exp\left[-\frac{m}{2}(a_m - \mu_m)^2\right] \prod_{i=1}^{k} \exp\left[-\frac{\tau_m}{2}(\mu_m - m_i)^2\right], \\ \mu_m|\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\theta}_{-\mu_m} &\sim \mathrm{N}\left(\frac{a_m b_m + \tau_m \sum_i m_i}{b_m + \tau_m k}, (b_m + \tau_m k)^{-1}\right). \end{aligned} \tag{A.2}$$

Similarly, the FCD of the expected intercept is

$$\begin{aligned} p(\mu_c|\boldsymbol{Y}, \boldsymbol{Z}, \ldots) &\propto p(\mu_c) \prod_{i=1}^{k} p\left(m_i|\mu_c, \tau_c\right), \\ \mu_c|\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\theta}_{-\mu_c} &\sim \mathrm{N}\left(\frac{a_c b_c + \tau_c \sum_i c_i}{b_c + \tau_c k}, (b_c + \tau_c k)^{-1}\right). \end{aligned} \tag{A.3}$$

The FCDs of the precision of the slope population level is

$$
\begin{aligned}
p\left(\tau_m|\boldsymbol{Y},\boldsymbol{Z},\dots\right) &\propto p(\tau_m)\prod_{i=1}^{k} p\left(m_i|\mu_m,\tau_m\right), \\
&\propto \tau_m^{g_m-1}\exp(-h_m\tau_m)\ \prod_{i=1}^{k}\sqrt{\tau_m}\exp\left[-\frac{\tau_m}{2}(\mu_m-m_i)^2\right], \qquad \text{(A.4)}\\
\tau_m|\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{\theta}_{-\tau_m} &\sim \mathrm{Ga}\left(g_m+k/2, h_m+\frac{1}{2}\sum_{i=1}^{k}(\mu_m-m_i)^2\right),
\end{aligned}
$$

Similarly, the FCD of the precision of the intercept population level is

$$
\begin{aligned}
p\left(\tau_c|\boldsymbol{Y},\boldsymbol{Z},\dots\right) &\propto p(\tau_c)\prod_{i=1}^{k} p\left(m_i|\mu_c,\tau_c\right), \\
\tau_c|\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{\theta}_{-\tau_c} &\sim \mathrm{Ga}\left(g_c+\frac{k}{2}, h_c+\frac{1}{2}\sum_{i=1}^{k}(\mu_c-c_i)^2\right).
\end{aligned} \qquad \text{(A.5)}
$$

For the i-th subject, the FCD of the slope is

$$
\begin{aligned}
p\left(m_i|\boldsymbol{Y},\boldsymbol{Z},\dots\right) &\propto p(m_i|\mu_m,\tau_m) \\
&\quad\times\prod_{j=1}^{M_i}\Big\{f\left(Y_{ij}|m_iX_{ij}+c_i,\tau\right)\mathbb{I}(Z_{ij}=0)+f\left(Y_{ij}|m_iX_{ij}+c_i,\gamma\right)\mathbb{I}(Z_{ij}=1)\Big\}, \\
&\propto \exp\left[-\frac{\tau_m}{2}(\mu_m-m_i)^2\right] \\
&\quad\times\prod_{j=1}^{M_i}\Big\{\exp\left[-\frac{\tau}{2}(Y_{ij}-m_iX_{ij}-c_i)^2\right]\mathbb{I}(Z_{ij}=0) \\
&\qquad\qquad +\exp\left[-\frac{\gamma}{2}(Y_{ij}-m_iX_{ij}-c_i)^2\right]\mathbb{I}(Z_{ij}=1)\Big\}, \\
m_i|\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{\theta}_{-m_i} &\sim \mathrm{N}\left(\frac{\mu_m\tau_m+\hat{\mu}_m\hat{\tau}_m}{\tau_m+\hat{\tau}_m}, (\tau_m+\hat{\tau}_m)^{-1}\right), \\
\hat{\tau}_{ij} &= \tau\mathbb{I}(Z_{ij}=0)+\gamma\mathbb{I}(Z_{ij}=1), \\
\hat{\tau}_m &= \sum_{j=1}^{M_i}\hat{\tau}_{ij}{X_{ij}}^2, \\
\hat{\mu}_m &= \hat{\tau}_m^{-1}\left\{\sum_{j=1}^{M_i}\left(X_{ij}Y_{ij}-c_iX_{ij}\right)\hat{\tau}_{ij}\right\}.
\end{aligned} \qquad \text{(A.6)}
$$

For the i-th subject, the FCD of the intercept is

$$
\begin{aligned}
p\left(c_i|\boldsymbol{Y},\boldsymbol{Z},\dots\right) &\propto p(c_i|\mu_c,\tau_c) \\
&\quad\times\prod_{j=1}^{M_i}\left\{f\left(Y_{ij}|m_iX_{ij}+c_i,\tau\right)\mathbb{I}(Z_{ij}=0)+f\left(Y_{ij}|m_iX_{ij}+c_i,\gamma\right)\mathbb{I}(Z_{ij}=1)\right\}, \\
&\propto \exp\left[-\frac{\tau_c}{2}(\mu_c-c_i)^2\right] \\
&\quad\times\prod_{j=1}^{M_i}\Bigg\{\exp\left[-\frac{\tau}{2}(Y_{ij}-m_iX_{ij}-c_i)^2\right]\mathbb{I}(Z_{ij}=0) \\
&\qquad\qquad+\exp\left[-\frac{\gamma}{2}(Y_{ij}-m_iX_{ij}-c_i)^2\right]\mathbb{I}(Z_{ij}=1)\Bigg\}, \\
c_i|\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{\theta}_{-c_i} &\sim \mathrm{N}\left(\frac{\mu_c\tau_c+\sum_j\hat{\tau}_{ij}\left(Y_{ij}-m_iX_{ij}\right)}{\tau_c+\sum_j\hat{\tau}_{ij}},\left(\tau_c+\sum_{j=1}^{M_i}\hat{\tau}_{ij}\right)^{-1}\right), \\
\hat{\tau}_{ij} &= \tau\mathbb{I}(Z_{ij}=0)+\gamma\mathbb{I}(Z_{ij}=1)
\end{aligned}
\tag{A.7}
$$

The FCD of the model's inferred precision, τ, is

$$
\begin{aligned}
p\left(\tau|\boldsymbol{Y},\boldsymbol{Z},\dots\right) &\propto p(\tau)\prod_{i=1}^{k}\prod_{j=1}^{M_i}\left\{f\left(Y_{ij}|m_iX_{ij}+c_i,\tau\right)\mathbb{I}(Z_{ij}=0)\right\}, \\
&\propto \tau^{g-1}\exp\left(-h\tau\right)\prod_{i=1}^{k}\prod_{j=1}^{M_i}\sqrt{\tau}\exp\left[-\frac{\tau}{2}(Y_{ij}-m_iX_{ij}-c_i)^2\right]\mathbb{I}\left(Z_{ij}=0\right), \\
\tau|\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{\theta}_{-\tau} &\sim \mathrm{Ga}\left(g+\frac{1}{2}\sum_{j=1}^{k}\sum_{i=1}^{M_i}\mathbb{I}(Z_{ij}=0),h+\frac{1}{2}\sum_{j=1}^{k}\sum_{i=1}^{M_i}(\hat{\mu}_{ij}-Y_{ij})^2\mathbb{I}(Z_{ij}=0)\right), \\
\hat{\mu}_{ij} &= m_iX_{ij}+c_i.
\end{aligned}
\tag{A.8}
$$

Lastly, the FCD for the proportion of not-like-control patient myofibres is

$$
\begin{aligned}
p\left(\pi|\boldsymbol{Y},\boldsymbol{Z},\dots\right) &\propto p(\pi)p\left(\boldsymbol{Z}|\pi\right), \\
&\propto \pi^{\alpha-1}(1-\pi)^{\beta-1}\prod_{j=1}^{M_k}\left\{\pi\mathbb{I}(Z_{kj}=0)+(1-\pi)\mathbb{I}(Z_{kj}=1)\right\}, \\
\pi|\boldsymbol{Z} &\sim \mathrm{Beta}\left(\alpha+\sum_{j=1}^{M_k}\mathbb{I}(Z_{kj}=0),\beta+\sum_{j=1}^{M_k}\mathbb{I}(Z_{kj}=1)\right).
\end{aligned}
\tag{A.9}
$$

A.2 Bayesian model output

The Bayesian model was independently fitted to the data represented in 219 2Dmito plots, and, therefore, reporting the MCMC output and model fit for each 2Dmito plot would not be feasible. Instead, a representative sample of output is shown across the datasets.

A.2.1 Vincent *et al.* dataset

The MCMC output for a single scheme is shown in Figures A.1-A.3, the output is representative of all other schemes unless otherwise stated, and for the sake of brevity, no other output will be shown.

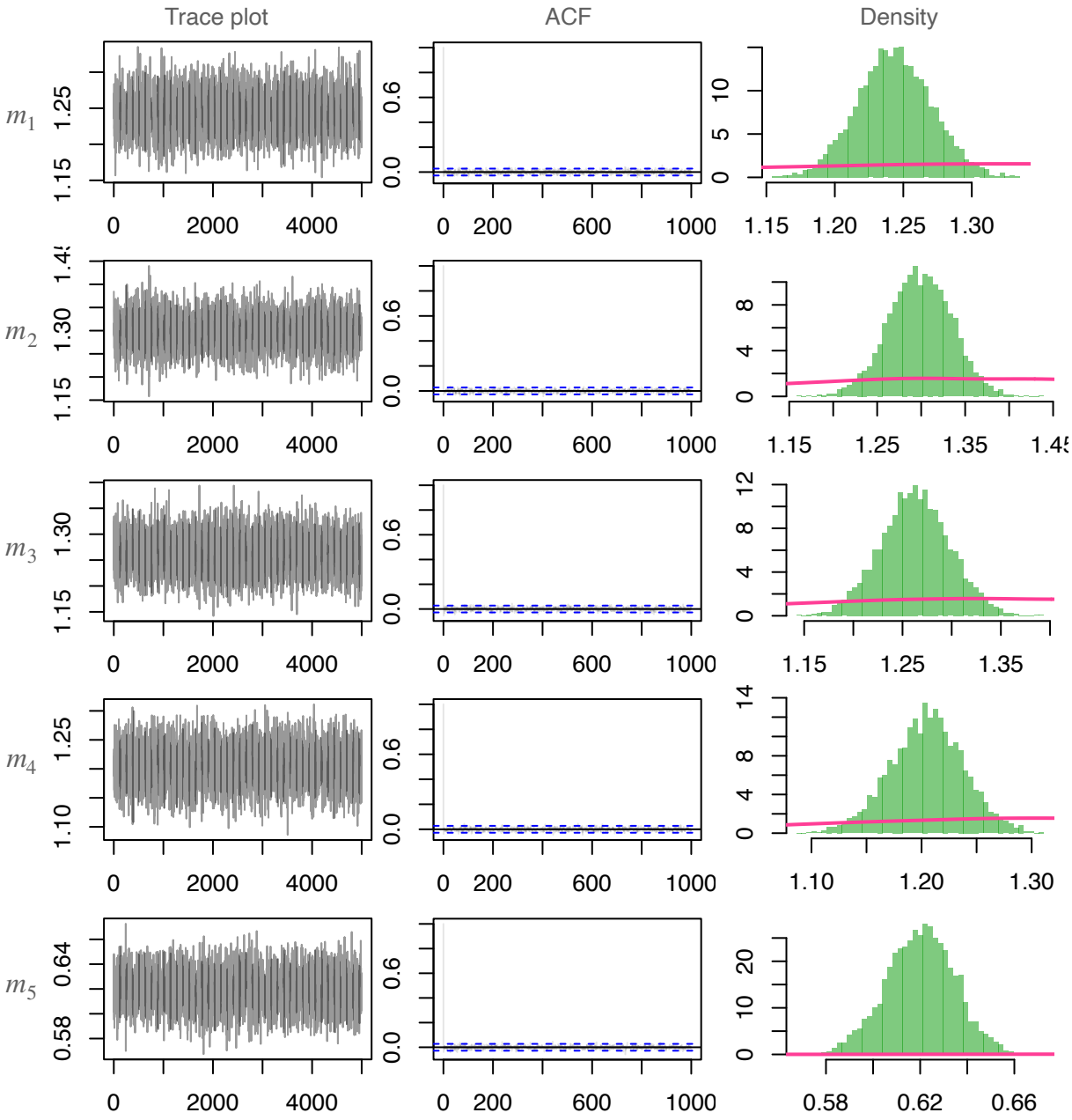


Figure A.1: **MCMC output for subject-specific slope parameters after fitting the Bayesian model to NDUFB8 abundance for patient P09 in the Vincent *et al.* dataset**. The posterior distribution is summarised by 5,000 almost independent draws from the posterior, generated by STAN, after removing a burn-in of 20,000 iterations. Histograms show the posterior densities (green) and the prior beliefs are indicated by their density (pink).

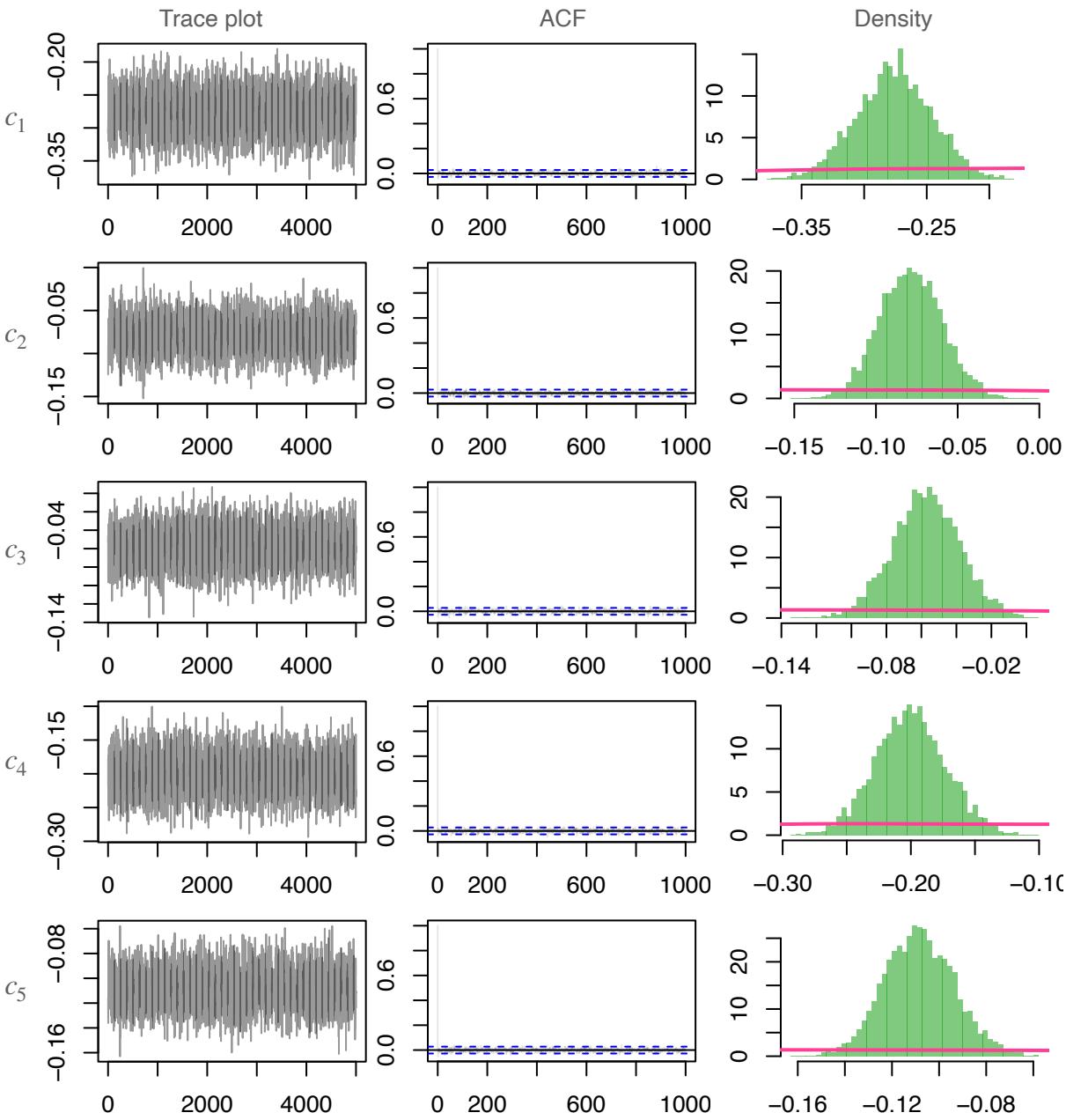


Figure A.2: **MCMC output for subject-specific intercept parameters after fitting the Bayesian model to NDUFB8 abundance for patient P09 in the Vincent *et al.* dataset**. The posterior distribution is summarised by 5,000 almost independent draws from the posterior, generated by STAN, after removing a burn-in of 20,000 iterations. Histograms show the posterior densities (green) and the prior beliefs are indicated by their density (pink).

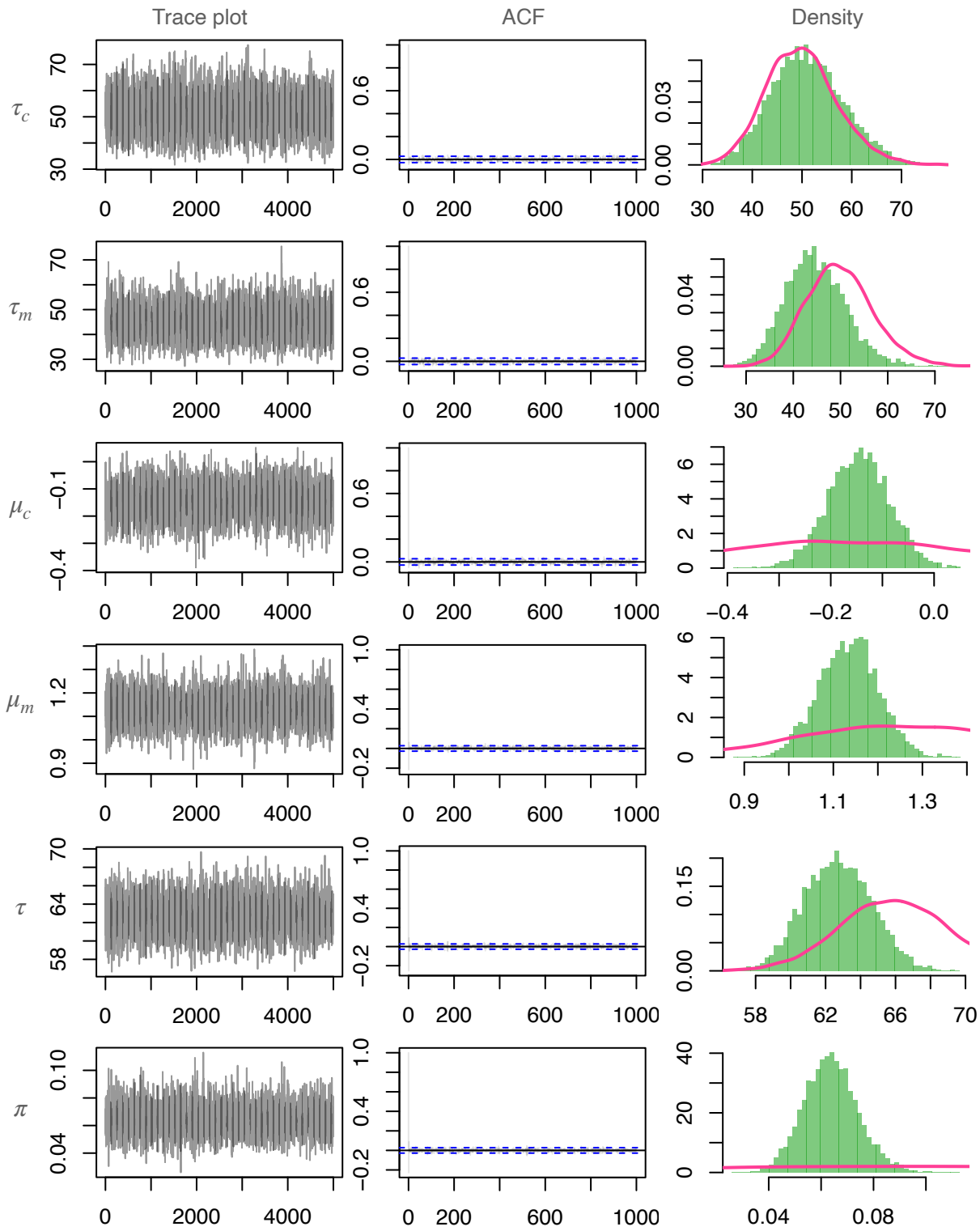


Figure A.3: **MCMC output for remaining parameters after fitting the Bayesian model to NDUFB8 abundance for patient P09 in the Vincent *et al.* dataset**. The posterior distribution is summarised by 5,000 almost independent draws from the posterior, generated by STAN, after removing a burn-in of 20,000 iterations. Histograms show the posterior densities (green) and the prior beliefs are indicated by their density (pink).

Example posterior classifications are shown in Figure A.4. The patients shown here were chosen due to their range in myofibre counts and abundance profiles in addition to those shown in Figure 5.7.

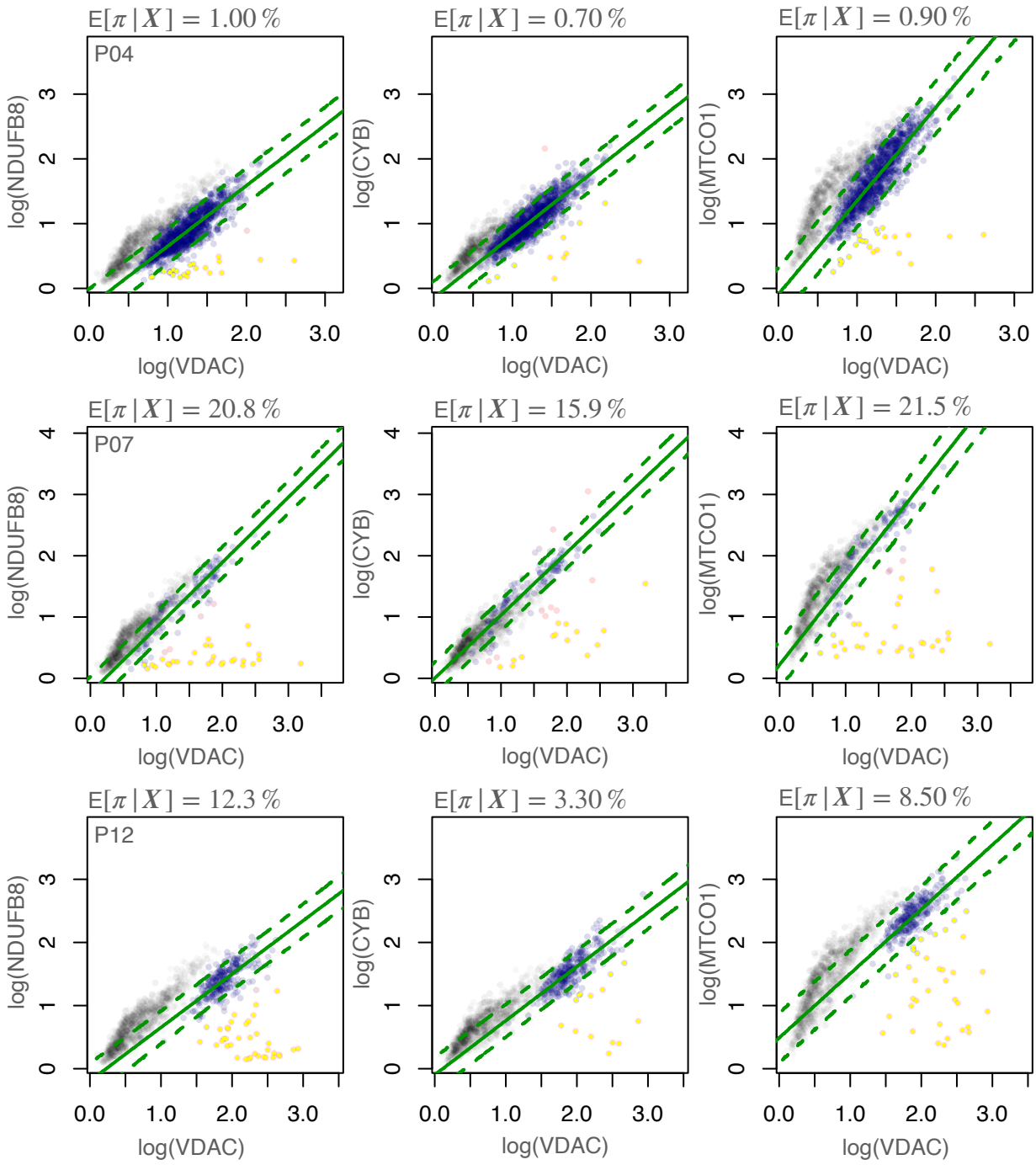


Figure A.4: **Example 2Dmito plots from the Vincent *et al.* dataset with Bayesian classification**. 2Dmito plots for three patient: P04, P07, and P12. Control subject data is shown (black) is aggregated and the model fit for them is not shown. The 95% posterior predictive distribution for like-control patient myofibres is shown (green dashed line) as well as the posterior expected value (solid green line). Posterior patient myofibre classifications are shown by colour; blue showing a like-control classification and red not-like-control. Manually classified not-like-control myofibres are indicated with a small yellow dot inside the data point. The posterior expected proportion of not-like-control myofibres is indicated above each 2Dmito plot.

A.2.2 Synthetic datasets

Example posterior classifications are from synthetically generated 2Dmito plots are shown in Figure A.5.

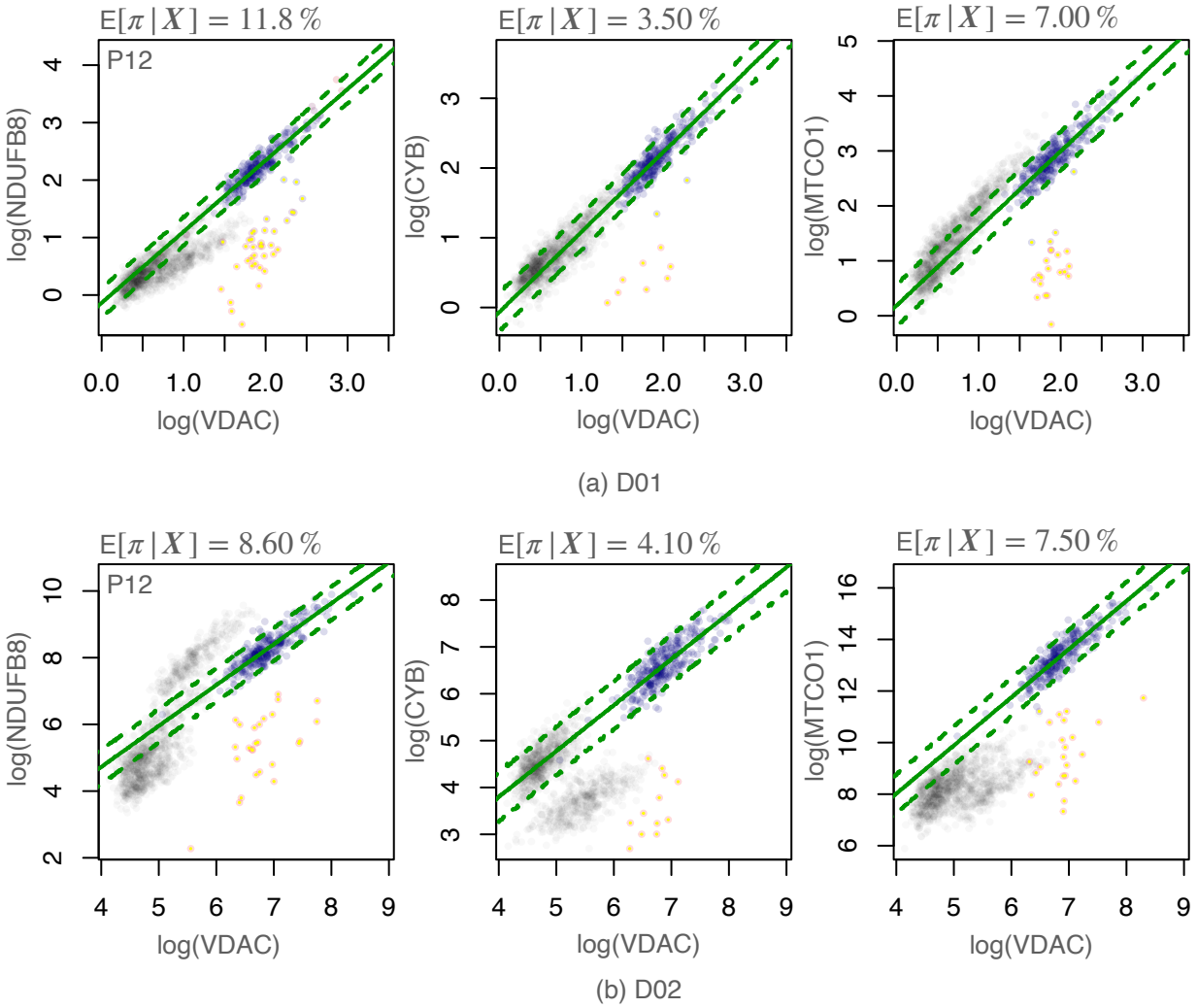


Figure A.5: **Example 2Dmito plots from the synthetic datasets with Bayesian classification**. 2Dmito plots for one patient, P12. Control subject data (black) is aggregated, and the model fit for them is not shown. The 95% posterior predictive distribution for like-control patient myofibres is shown (green dashed line) as well as the posterior expected value (solid green line). Posterior patient myofibre classifications are shown by colour; blue showing a like-control classification and red not-like-control. Ground-truth OXPHOS status of not-like-control myofibres are indicated with a small yellow dot inside the data point. The posterior expected proportion of not-like-control myofibres is indicated above each 2Dmito plot.

A.2.3 Gomes *et al.* dataset

The output for all samples in patient P03 of the Gomes *et al.* dataset are shown in Figures A.6-A.8. These 2Dmito were selected to be shown due to their varied success, and abundance profiles which are distinct from other patients and datasets.

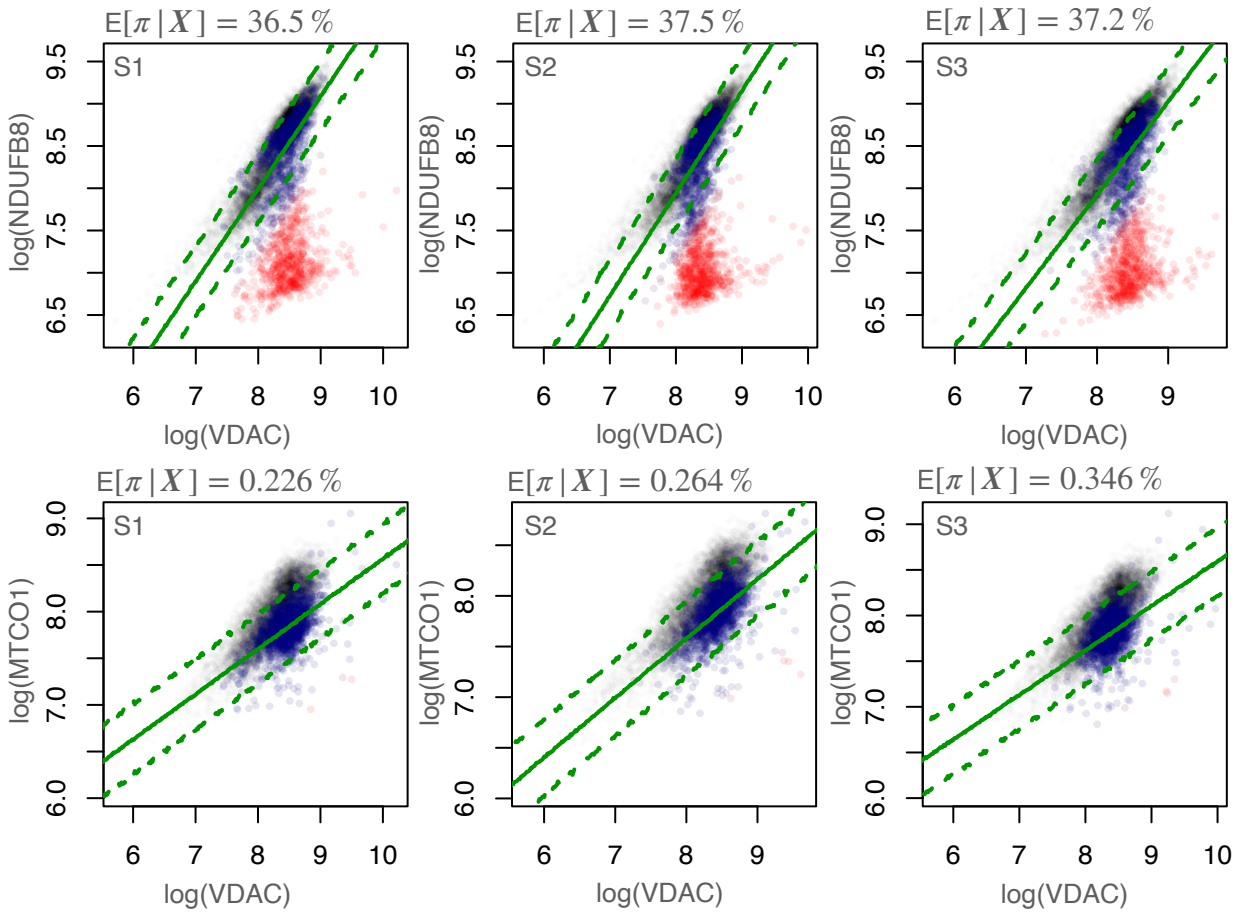


Figure A.6: **2Dmito plots for QD tissue from patient P03 in the Gomes *et al.* dataset with Bayesian classification**. 2Dmito plots for all QD samples in patient P03 of the Gomes dataset. Control subject data is shown (black) is aggregated and the model fit for them is not shown. The 95% posterior predictive distribution for like-control patient myofibres is shown (green dashed line) as well as the posterior expected value (solid green line). Posterior patient myofibre classifications are shown by colour; blue showing a like-control classification and red not-like-control. The posterior expected proportion of not-like-control myofibres is indicated above each 2Dmito plot.

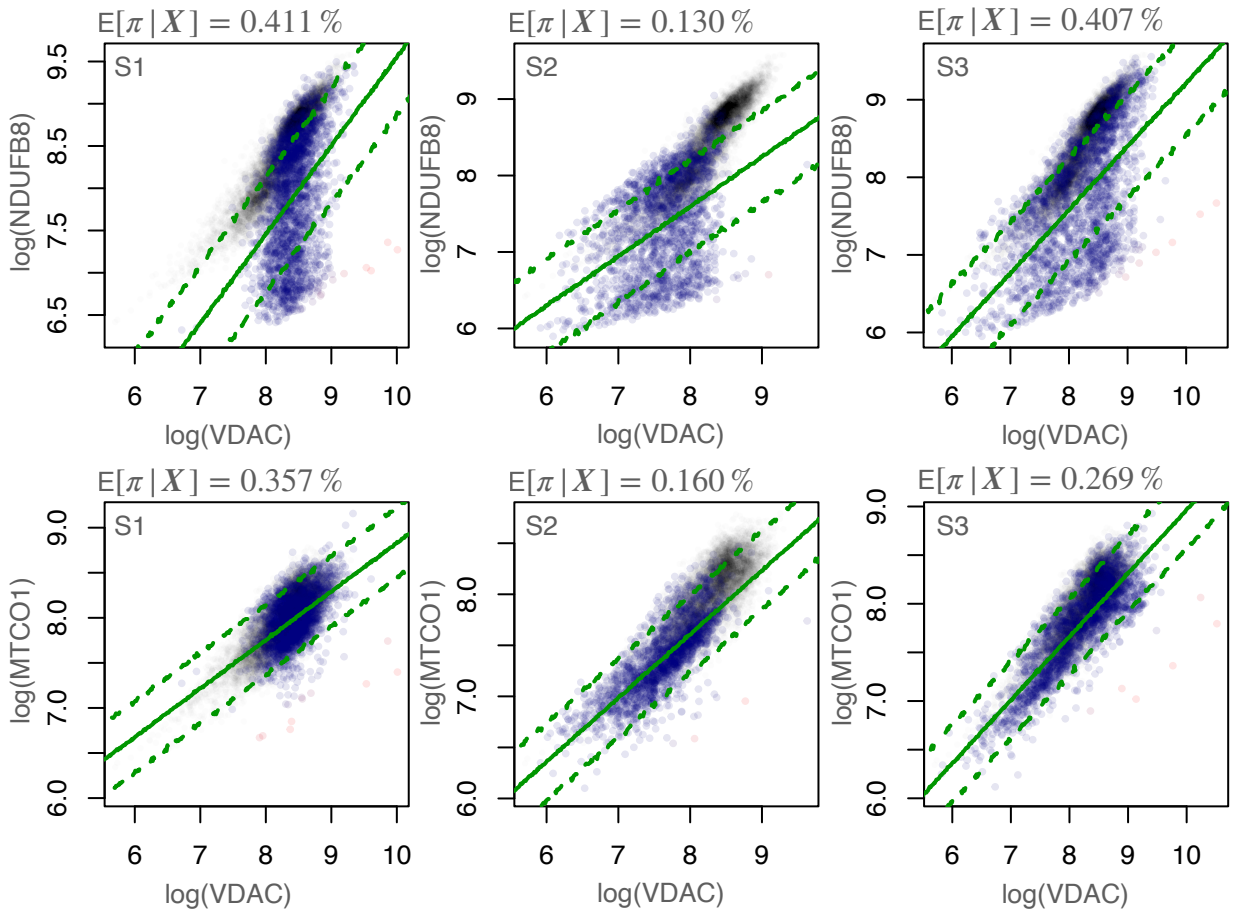


Figure A.7: **2Dmito plots for TA tissue (block 1) from patient P03 in the Gomes *et al.* dataset with Bayesian classification**. 2Dmito plots for TA-B1 samples in patient P03 of the Gomes dataset. Control subject data is shown (black) is aggregated and the model fit for them is not shown. The 95% posterior predictive distribution for like-control patient myofibres is shown (green dashed line) as well as the posterior expected value (solid green line). Posterior patient myofibre classifications are shown by colour; blue showing a like-control classification and red not-like-control. The posterior expected proportion of not-like-control myofibres is indicated above each 2Dmito plot.

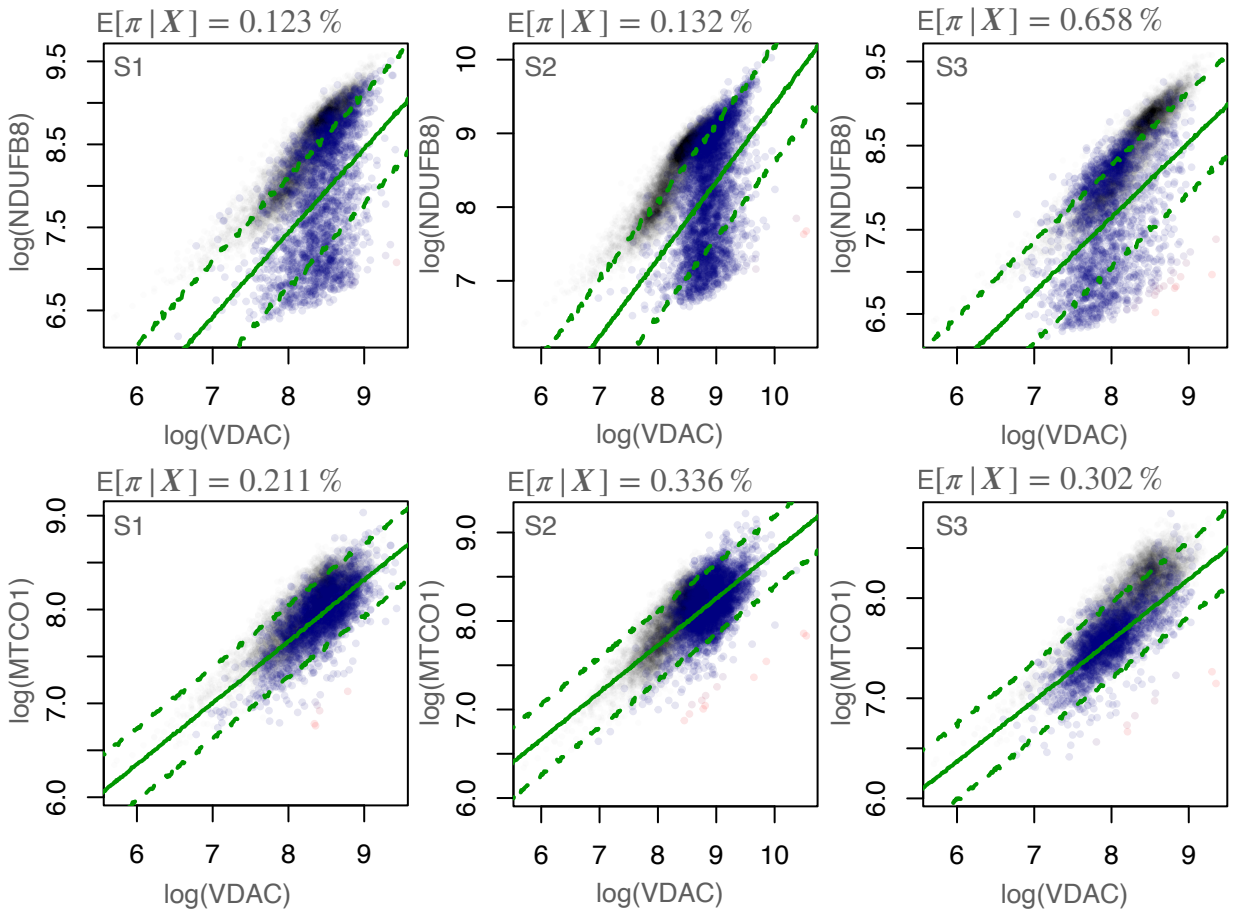


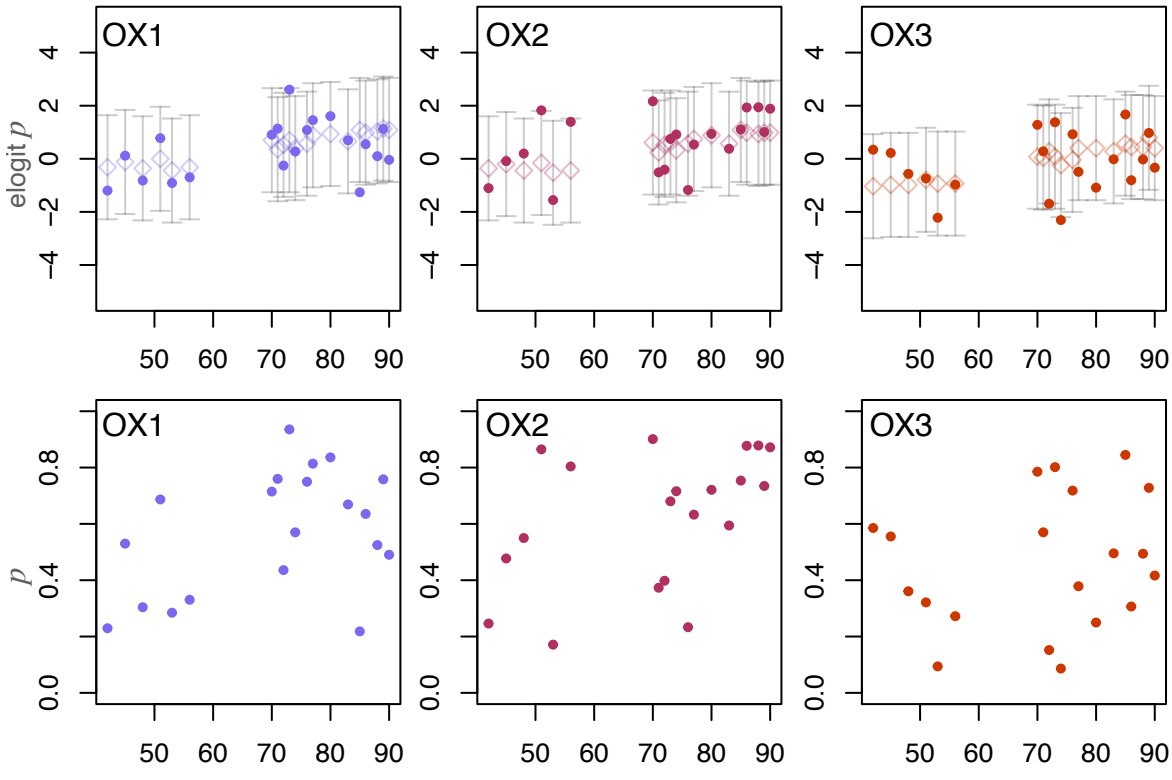
Figure A.8: **2Dmito plots for TA tissue (block 2) from patient P03 in the Gomes *et al.* dataset with Bayesian classification**. 2Dmito plots for TA-B2 samples in patient P03 of the Gomes dataset. Control subject data is shown (black) is aggregated and the model fit for them is not shown. The 95% posterior predictive distribution for like-control patient myofibres is shown (green dashed line) as well as the posterior expected value (solid green line). Posterior patient myofibre classifications are shown by colour; blue showing a like-control classification and red not-like-control. The posterior expected proportion of not-like-control myofibres is indicated above each 2Dmito plot.

Appendix B

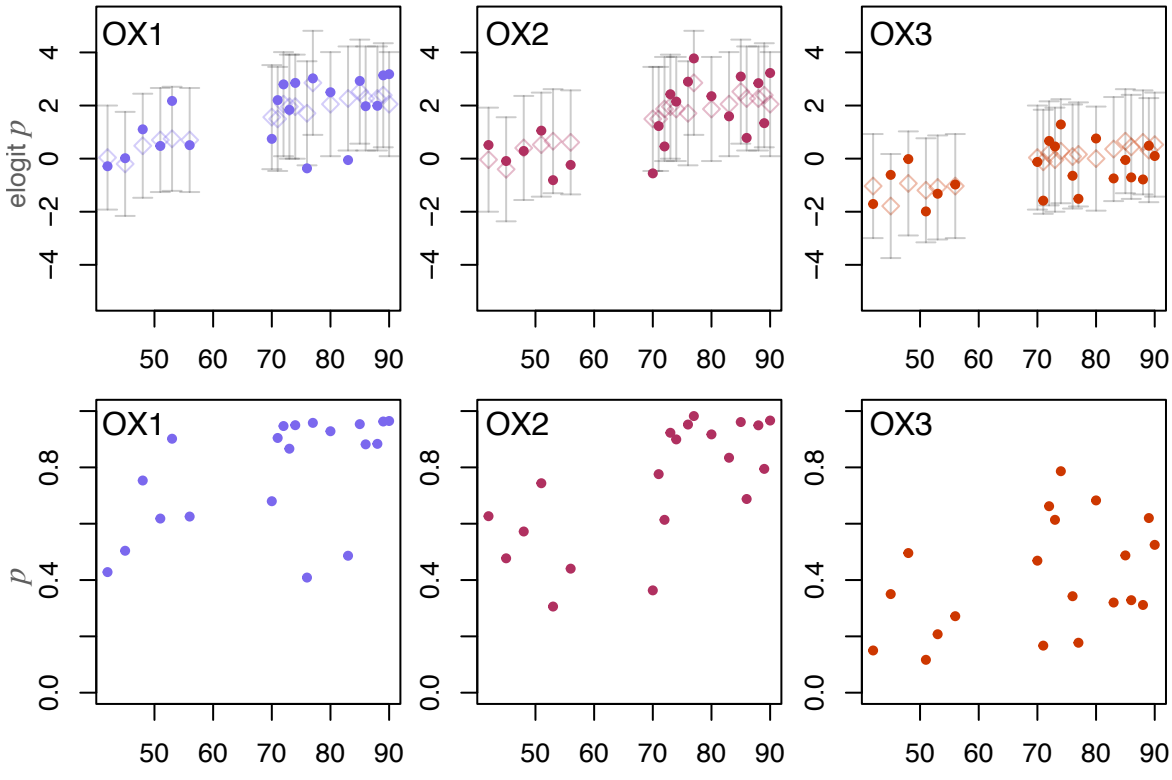
Modelling clonal expansion - Appendix

The generated synthetic datasets created from mathematical models of the SoS and PNN theories of clonal expansion are shown in Figure B.1.

B.1 Synthetic data



(a) X_{SoS}



(b) X_{PNN}

Figure B.1: **Synthetic OXPHOS deficiency proporiton dataset**. Synthetic datasets generated by the mathematical model of (a) survival of the smallest and (b) the perinuclear theory of clonal expansion. For each dataset, the top depicts the latent, ground-truth deficiency proportions for the three proteins (diamonds) and the synthetic observed data, after the addition of random noise on the transformed scale (solid circle). The error bars show the 95% interval of the observed value, given its ground-truth. The bottom row shows the synthetic observed data, after adding random noise, on the natural scale. The x axes for all plots depict time in years, i.e. the age of the tissue collection for the synthetic patients.

B.2 Inference output

The MCMC output for all executions of the inference scheme for the various datasets is shown in Figures B.2 - B.6.

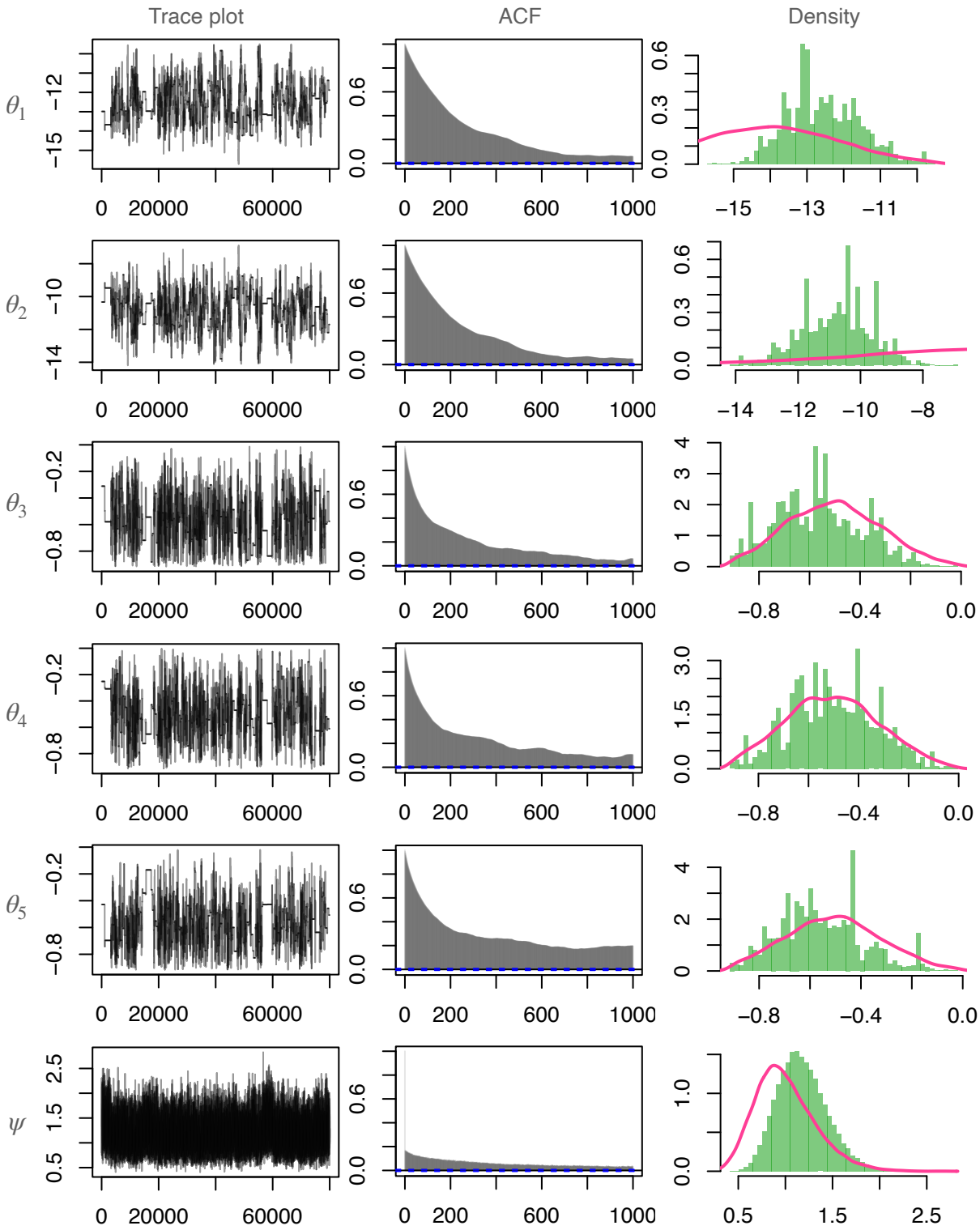


Figure B.2: **Results of parameter inference when fitting the mathematical model of clonal expansion to $\boldsymbol{X}_{\mathbf{Obs}}$.** The posterior distributions (green) are summarised by a histogram created with 100,000, not including burning draws from the joint posterior distribution. Kernel density estimates of the prior beliefs are shown in pink, as generated by 100,000 draws from the prior.

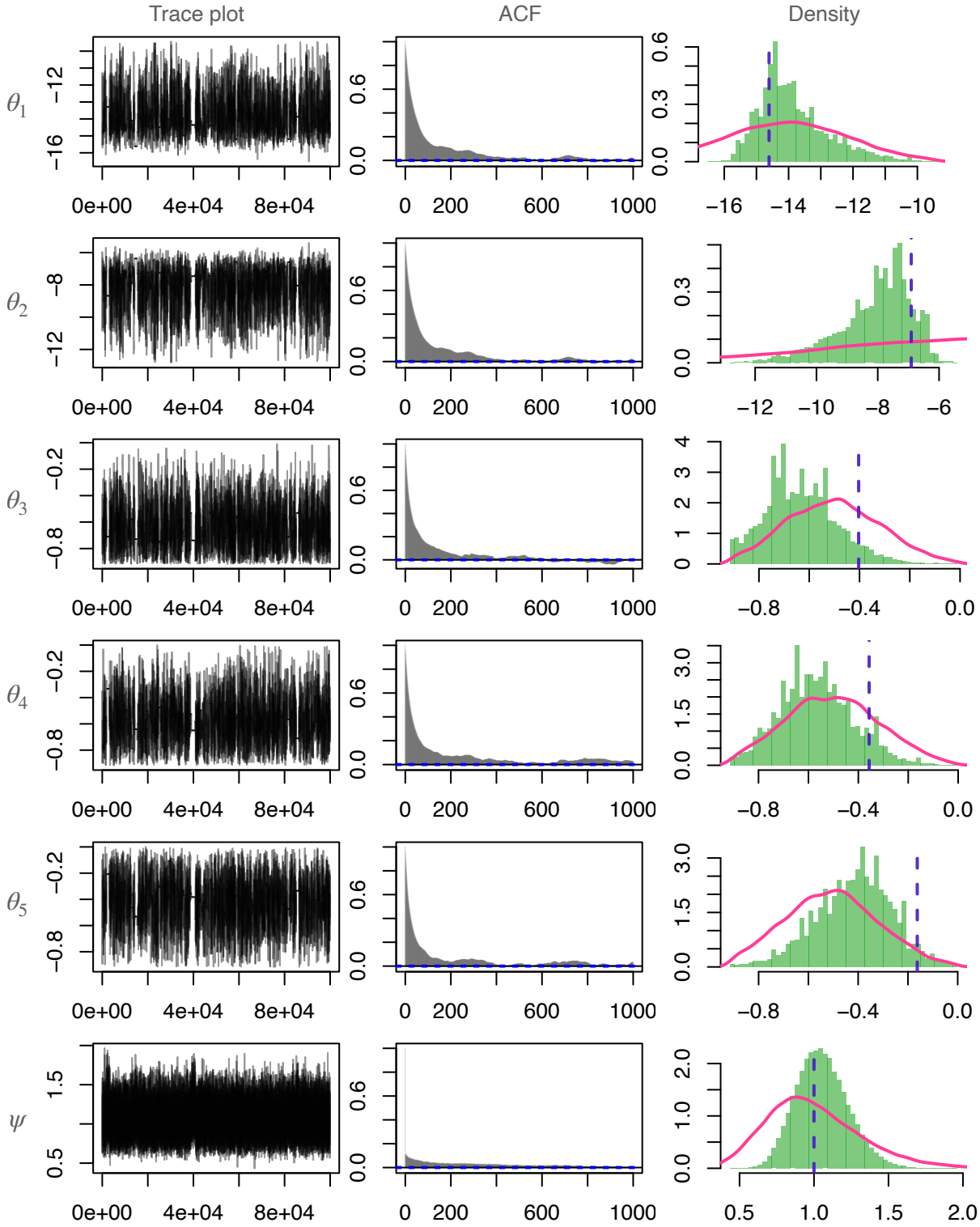


Figure B.3: **Results of parameter inference when fitting the mathematical model of clonal expansion to the synthetic RGD dataset**. The posterior distributions (green) are summarised by a histogram created with 100,000, not including burning draws from the joint posterior distribution. Kernel density estimates of the prior beliefs are shown in pink, as generated by 100,000 draws from the prior. Ground-truth parameter values, used to generate the synthetic data is shown as dashed blue line.

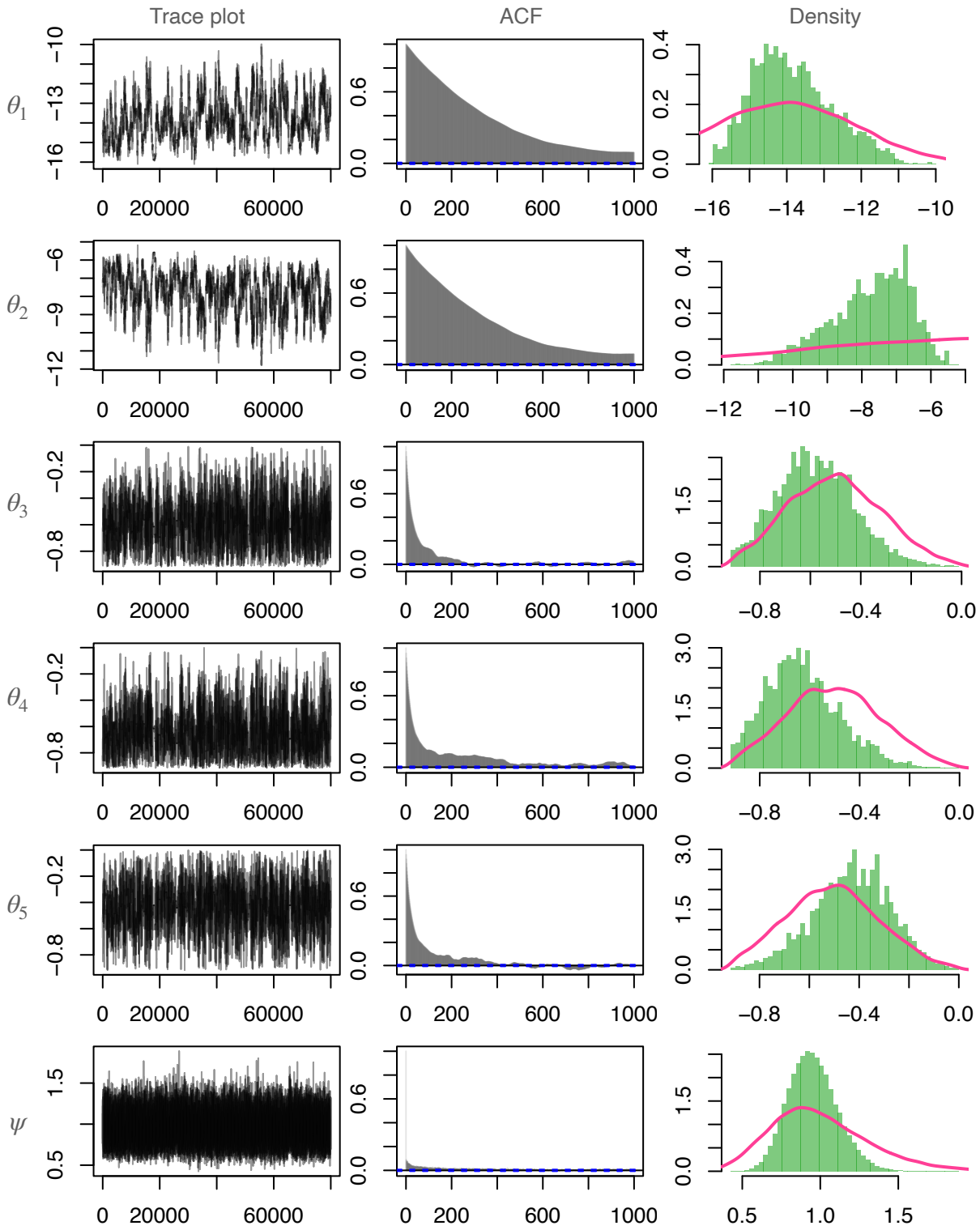


Figure B.4: **Results of parameter inference when fitting the mathematical model of clonal expansion to the synthetic SoS dataset**. The posterior distributions (green) are summarised by a histogram created with 100,000, not including burning draws from the joint posterior distribution. Kernel density estimates of the prior beliefs are shown in pink, as generated by 100,000 draws from the prior.

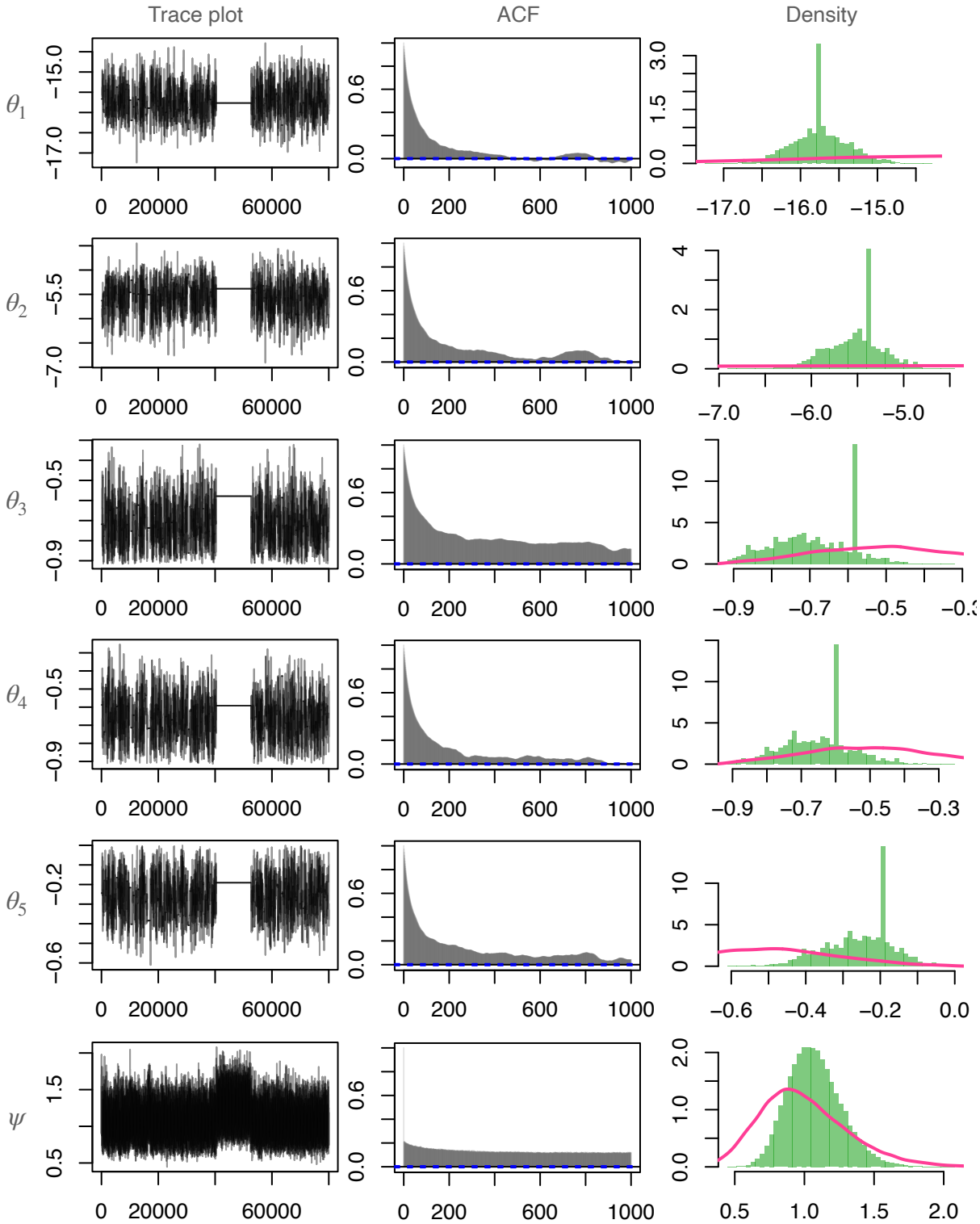


Figure B.5: **Results of parameter inference when fitting the mathematical model of clonal expansion to the synthetic PNN dataset**. The posterior distributions (green) are summarised by a histogram created with 100,000, not including burning draws from the joint posterior distribution. Kernel density estimates of the prior beliefs are shown in pink, as generated by 100,000 draws from the prior.

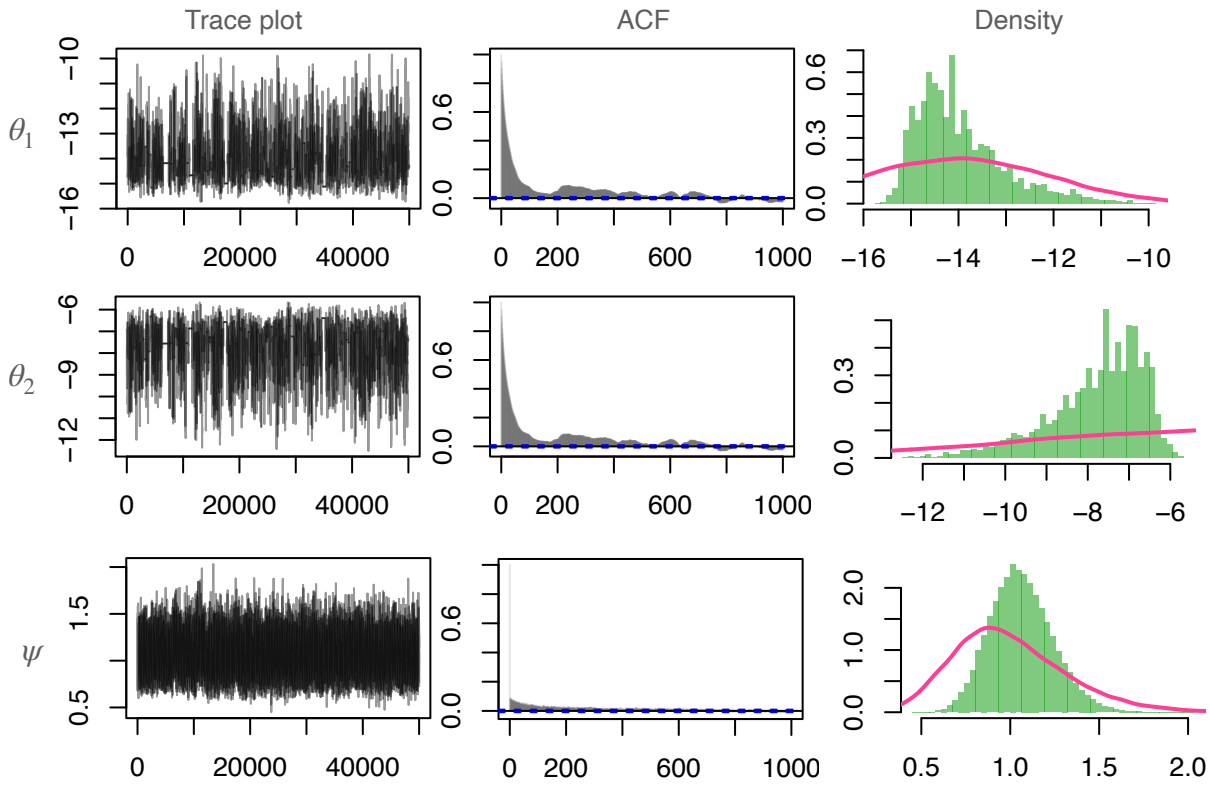


Figure B.6: **Results of parameter inference when fitting the mathematical model of clonal expansion to the synthetic RGD dataset**. The posterior distributions (green) are summarised by a histogram created with 100,000, not including burning draws from the joint posterior distribution. Kernel density estimates of the prior beliefs are shown in pink, as generated by 100,000 draws from the prior.

Appendix C

Hardware Accelerated Simulation

C.1 Introduction

Graphcore, a company based in Bristol, UK, has developed a new kind of processor with the aim of massively parallelising computation. The chip is called an Intelligence Processing Unit (IPU) and is now on its second generation, the Colossus™ MK2 IPU processor - the GC200. One GC200 contains 1,472 IPU-Tiles, one IPU-Tile contains one IPU-Core, which is capable of executing six independent threads in parallel; that is, one GC200 can execute 8,832 tasks in parallel. For example, if we were to run a single simulation, which is known to take 30 seconds on an IPU-Core, it would also take 30 seconds to run 8,832 simulations by utilising every thread on the GC200.

Originally designed to increase performance in machine learning applications dramatically, the IPUs architecture can also be used to provide big performance increases in many other areas. Here, we will show that the IPU's ability for massive parallelisation can be used to decrease the computational time for repeated simulation of stochastic models.

C.2 Method

The individual IPU-Cores are not very powerful when compared to most CPUs, but by using all available IPU-Tiles the full power of the processor can be utilised. Therefore, it would not be appropriate to only consider the execution time for a single simulation, as this effectively utilises 1/8,832 of the processor's power. Instead, we will consider 'batches' of simulations and compare the wall clock times for the GC200 processor and a CPU to execute 8,832 simulations (one batch). The CPU used as a comparison is the Intel® Xeon® Gold 6246R Processor, with an all-core turbo frequency of 4.0GHz and eight cores. It is of the same family, but a newer model, of the CPU used as a baseline comparison by Kulkarni et al. (2022). A single GC200 cannot be purchased; instead, they can only be purchased within an IPU-POD, which have a range of four to 256 processors. For the timings presented here, an IPU-POD4 was used, which contains four GC200 processors and can therefore execute 35,328 tasks in parallel. To show the scalability of the IPU, we will also show the time taken for the IPU-POD4 to execute 35,328 simulations, utilising every thread available within the IPU-POD4.

Lastly, we will compare the processor's computing power relative to the cost associated with them. There are several companies which rent time on an IPU-POD, much like any other cloud computing service. The IPUs used here were rented using Gcore's cloud infrastructure, and their cost will be used as a comparison. We have also used Gcore to rent time on the Intel Xeon, as Gcore allows the rental of bare metal processors, giving a more stable performance compared to virtual machines.

To compare the performance of the two chips, a model of mtDNA dynamics is used. Assuming random genetic drift and a zero probability of *de novo* mutations, the model considers the dynamics within patients with an inherited variant load. Similarly to Chapter 5, a linear copy number control is implemented.

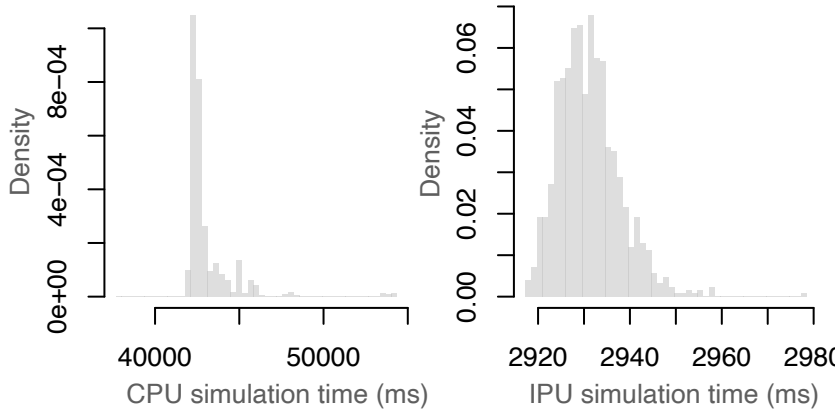
C.3 Results

As this is a stochastic process, the simulation time is not deterministic. Therefore, the simulations are executed a large number of times to gain an understanding of the runtime distribution. The results of running a single simulation are omitted; however, to demonstrate the relative power of a single IPU-Core compared to the Intel® Xeon® Processor, the mean time from 1,000 simulations is shown in Table C.1.

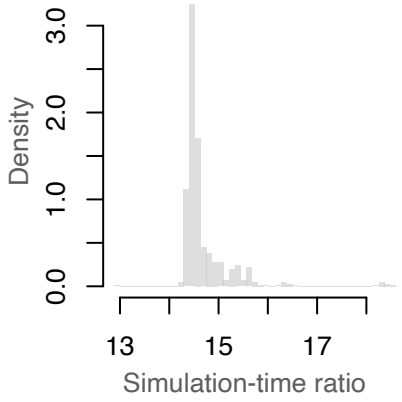
	Intel Xeon	GC200
Time (ms)	32.149	2822.955

Table C.1: **CPU is significantly faster when simulating a single realisation**. Mean wallclock execution time to simulate a single realisation from Gillespie algorithm, using a model of mtDNA dynamics with no *de novo* mutation events. The mean was calculated from 1,000 independent simulations.

It is clear that a one IPU-Core is much less powerful than a one core of the Intel Xeon. However, the massive number of IPU-Cores within a single processor, 1,472, combined with their ability to run up to six independent tasks without loss of performance, results in a powerful processor. Figure C.1(a) shows the distribution of 1,000 executions of simulation batches of size 8,832. These simulations were done in parallel. Each IPU-Core executed six simulations independently, resulting in 8,832 simulations being executed in parallel. The time to complete these simulations is essentially the time taken for the longest of the 8,832 simulations to finish, as the GC200 thread allocation and data streaming add little time. The parallelisation for the Intel® Xeon® processor was achieved by a Bash script executing N_{core} (the number of available cores) C++ scripts, each sequentially executing M Gillespie simulations, such that $M \times N_{core} = 8,832$. This approach was taken because the C++ standard library `std::thread` and the parallelisation software OpenMP added considerable computational overhead due to thread creation and task allocation, resulting in the parallel simulations taking significantly longer than if they were executed on a single thread. Although not ideal, the static scheduling used was not a massive disadvantage. The system maintains a relatively constant hazard due to copy number control and the assumption of no species advantage, resulting in the execution time of a single simulation having low variance. This is confirmed by trials with `std::thread` and OpenMP, where both static and dynamic scheduling schemes were used and the difference between them was negligible (results not shown).



(a) Time for batch simulations



(b) Time for batch simulations

Figure C.1: **IPUs massively reduce batch simulation time**. (a) Distribution of simulation times for a model of mtDNA dynamics, when executing the simulations on (left) a CPU, Intel® Xeon® , and (right) a GC200. (b) The ratio distribution of simulation times (CPU:IPU). The distributions were found by 1,000 timed simulations of *batches* of 8,832 simulations on each processor.

Figure C.1(a) shows that, compared to a single simulation, the time taken for the Intel® Xeon® to execute the batches has increased considerably (see Table C.1). However, the GC200 has only increased slightly, from $\approx$ 2800ms to $\approx$ 2930ms. Therefore, when comparing batch simulations, the GC200 processor was approximately 14 times faster than the Intel® Xeon® , see Figure C.1(b).

The batch simulations were conducted on a single GC200 processor. However, the IPU-POD4 contains four GC200 processors, and their scalability allows all cores and threads to be utilised while incurring little additional computational cost. The mean time to execute 35,328 simulations, one for each thread on IPU-POD4, was 2937.766ms, which is approximately equal to the time to execute 8,832 simulations on a single GC200. Figure C.2 shows how the simulation time distribution changes depending on the number of simulations executed in parallel on an IPU-POD4 machine.

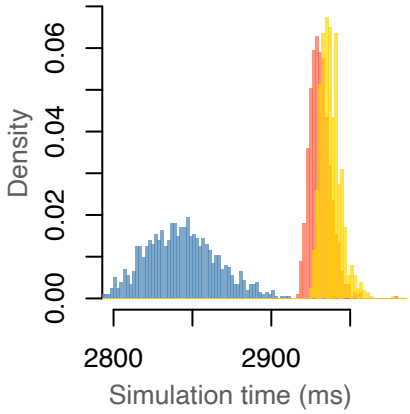


Figure C.2: **IPU-PODs allow seamless scalability**. The distribution of execution times for a single simulation (blue), 8,832 simulations (red), and 35,328 simulations (yellow) executed on an IPU-POD4. Each simulation was an independent Gillespie simulation given the same initial conditions, simulating the mtDNA dynamics of a single cell.

C.4 Discussion

As shown, when executing a large number of independent simulations, the GC200 can offer a considerable reduction in execution time compared to the Intel® Xeon® . This, however, does not take into account the cost of the processors. To compare the relative computing power per cost, seamless parallelisation between one CPU and four is assumed, such that four CPUs could execute four times the simulations of a single CPU without incurring any additional computational cost. For easy comparison, the median of the distributions is compared and divided by the cost of renting each machine from Gcore, giving the approximate cost of a single simulation on each machine. The results are presented in Table C.2.

	Intel Xeon	IPU-POD4
cost per hour	€2.5032 *	€6.12
median sim. time	42.5455s **	2.9428s
expected sim. cost	€0.024	€0.005

Table C.2: **IPU-POD4 provides a lower cost per simulation than CPU**. The expected cost to run one batch or 35,328 simulations using an Intel Xeon Processor and an IPU-POD4.

(*) The cost per hour for a single Intel® Xeon® Gold 6246R Processor is €0.5133.
(**) This is the expected simulation time assuming seamless scalability of CPUs.

Despite the assumptions made about the seamless scalability of the CPUs, the IPUs still offer a significant reduction in cost per simulation by a factor of almost five. Reducing the computational cost of simulation can decrease the cost of inference for mathematical models of biological systems, such as those presented in Chapter 5. Whether inference is being done by MCMC methods, ABC methods (Sunnåker et al., 2013), or by model emulation, each requires many thousands or millions of simulations to be made. However, it is usually not necessary to perform batch simulations of such large numbers during inference; the IPU enables each simulation to be executed with different parameters with

minimal effort. This can also be done on a CPU; however, it may require some thought and consideration due to the bash script method of parallelisation.

Despite the clear time advantage of the IPUs, there is a steep learning curve to master their use, even for users already familiar with C++. From personal experience, taking over a week to go from nothing to executing a Gillespie simulation. The CPU parallelisation method used here, however, is not without its faults. The Bash script method of parallelisation would require any desired output to be saved to a text file and then aggregated when all simulations are done. This would also add development time and a small amount of computational time.

The agent-based model of clonal expansion, described in Chapter 5, was not tested on the IPUs, as the model was developed after this investigation and time and resource limitations did not allow for this to be revisited. Due to the limited power of a single GC200 tile, it is unclear whether significant computational savings could be made when simulating such a computationally expensive model. However, demonstrating a reduction in the simulation time of the Gillespie algorithm is significant, given its widespread use.

Another problem that may arise when using the agent-based model to simulate clonal expansion on the IPUs, although it can be overcome, is retrieving model output. Attached to each IPU-POD is a CPU which sends commands to each GC200 processor; however, what can be sent and retrieved is somewhat limited. Only numerical vectors of a known length can currently be passed to and from a GC200 processor. This may seem restrictive, but with some consideration, the problem could be overcome by implementing a post-simulation pipeline that converts numerical vectors into the desired output.

Bibliography

Abdelsamed, H. A., Moustaki, A., Fan, Y., Dogra, P., Ghoneim, H. E., Zebley, C. C., Triplett, B. M., Sekaly, R.-P., & Youngblood, B. (2017). Human memory CD8 t cell effector potential is epigenetically preserved during in vivo homeostasis. *The Journal of Experimental Medicine*, *214*(6), 1593–1606. https://doi.org/10.1084/jem.20161760

Ahmed, S. T., Alston, C. L., Hopton, S., He, I. P., Langping andHargreaves, Falkous, G., Oláhová, M., McFarland, R., Turnbull, D. M., Rocha, M. C., & Taylor, R. W. (2017). Using a quantitative quadruple immunofluorescent assay to diagnose isolated mitochondrial complex i deficiency. *Scientific Reports*, *7*(1), 156676. https://doi.org/10.1038/s41598-017-14623-2

Ainsworth, H. F. (2014). *Bayesian inference for stochastic kinetic models using data on proportions of cell death* [Thesis]. Newcastle University [Accepted: 2015-02-09T14:01:11Z]. Retrieved February 19, 2025, from http://theses.ncl.ac.uk/jspui/handle/10443/2499

Akkaya, M., Traba, J., Roesler, A. S., Miozzo, P., Akkaya, B., Theall, B. P., Sohn, H., Pena, M., Smelkinson, M., Kabat, J., Dahlstrom, E., Dorward, D. W., Skinner, J., Sack, M. N., & Pierce, S. K. (2018). Second signals rescue b cells from activation-induced mitochondrial dysfunction and death [Publisher: Nature Publishing Group]. *Nature Immunology*, *19*(8), 871–884. https://doi.org/10.1038/s41590-018-0156-5

Altmann, J., Büchner, B., Nadaj-Pakleza, A., Schäfer, J., Jackson, S., Lehmann, D., Deschauer, M., Kopajtich, R., Lautenschläger, R., Kuhn, K. A., Karle, K., Schöls, L., Schulz, J. B., Weis, J., Prokisch, H., Kornblum, C., Claeys, K. G., & Klopstock, T. (2016). Expanded phenotypic spectrum of the m.8344a>g "MERRF" mutation: Data from the german mitoNET registry. *Journal of Neurology*, *263*(5), 961–972. https://doi.org/10.1007/s00415-016-8086-3

Amati-Bonneau, P., Valentino, M. L., Reynier, P., Gallardo, M. E., Bornstein, B., Boissière, A., Campos, Y., Rivera, H., de la Aleja, J. G., Carroccia, R., Iommarini, L., Labauge, P., Figarella-Branger, D., Marcorelles, P., Furby, A., Beauvais, K., Letournel, F., Liguori, R., La Morgia, C., … Carelli, V. (2008). OPA1 mutations induce mitochondrial DNA instability and optic atrophy 'plus' phenotypes. *Brain: A Journal of Neurology*, *131*, 338–351. https://doi.org/10.1093/brain/awm298

Anderson, S., Bankier, A. T., Barrell, B. G., de Bruijn, M. H. L., Coulson, A. R., Drouin, J., Eperon, I. C., Nierlich, D. P., Roe, B. A., Sanger, F., Schreier, P. H., Smith, A. J. H., Staden, R., & Young, I. G. (1981). Sequence and organization of the human mitochondrial genome [Publisher: Nature Publishing Group]. *Nature*, *290*(5806), 457–465. https://doi.org/10.1038/290457a0

Andrews, R. M., Kubacka, I., Chinnery, P. F., Lightowlers, R. N., Turnbull, D. M., & Howell, N. (1999). Reanalysis and revision of the cambridge reference sequence for

human mitochondrial DNA. *Nature Genetics*, *23*(2), 147. https://doi.org/10.1038/13779

Andrianakis, I., Vernon, I., McCreesh, N., McKinley, T. J., Oakley, J. E., Nsubuga, R. N., Goldstein, M., & White, R. G. (2017). History matching of a complex epidemiological model of human immunodeficiency virus transmission by using variance emulation. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *66*(4), 717–740. https://doi.org/10.1111/rssc.12198

Aryaman, J., Bowles, C., Jones, N. S., & Johnston, I. G. (2019). Mitochondrial network state scales mtDNA genetic dynamics. *Genetics*, *212*(4), 1429–1443. https://doi.org/10.1534/genetics.119.302423

Aryaman, J., Johnston, I. G., & Jones, N. S. (2017). Mitochondrial DNA density homeostasis accounts for a threshold effect in a cybrid model of a human mitochondrial disease. *Biochemical Journal*, *474*(23), 4019–4034. https://doi.org/10.1042/BCJ20170651

Ashley, M. V., Laipis, P. J., & Hauswirth, W. W. (1989). Rapid segregation of heteroplasmic bovine mitochondria. *Nucleic Acids Research*, *17*(18), 7325–7331. Retrieved December 18, 2024, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC334812/

Bacman, S. R., Williams, S. L., & Moraes, C. T. (2009). Intra- and inter-molecular recombination of mitochondrial DNA after in vivo induction of multiple double-strand breaks. *Nucleic Acids Research*, *37*(13), 4218–4226. https://doi.org/10.1093/nar/gkp348

Baker, E., Barbillon, P., Fadikar, A., Gramacy, R. B., Herbei, R., Higdon, D., Huang, J., Johnson, L. R., Ma, P., Mondal, A., Pires, B., Sacks, J., & Sokolov, V. (2022). Analyzing stochastic computer models: A review with opportunities [Publisher: Institute of Mathematical Statistics]. *Statistical Science*, *37*(1), 64–89. https://doi.org/10.1214/21-STS822

Baty, K., Farrugia, M. E., Hopton, S., Falkous, G., Schaefer, A. M., Stewart, W., Willison, H. J., Reilly, M. M., Blakely, E. L., Taylor, R. W., & Ng, Y. S. (2021). A novel MT-CO2 variant causing cerebellar ataxia and neuropathy: The role of muscle biopsy in diagnosis and defining pathogenicity. *Neuromuscular Disorders*, *31*(11), 1186–1193. https://doi.org/10.1016/j.nmd.2021.05.014

Bereiter-Hahn, J., & Vöth, M. (1994). Dynamics of mitochondria in living cells: Shape changes, dislocations, fusion, and fission of mitochondria. *Microscopy Research and Technique*, *27*(3), 198–219. https://doi.org/10.1002/jemt.1070270303

Berg, J., Tymoczko, J., Gatto Jr., G., & Stryer, L. (2015). *Biochemistry. glycolysis and gluconeogenesis* (8th edition). W.H. Freeman; Company.

Berk, A. J., & Calyton, D. A. (1974, March 29). *Mechanism of mitochondrial DNA replication in mouse l-cells : Asynchronous replication of strands, segregation of circular daughter molecules, aspects of topology and turnover of an initiation sequence.*

Berry, S. M., Broglio, K. R., Groshen, S., & Berry, D. A. (2013). Bayesian hierarchical modeling of patient subpopulations: Efficient designs of phase II oncology clinical trials [Publisher: SAGE Publications]. *Clinical Trials*, *10*(5), 720–734. https://doi.org/10.1177/1740774513497539

Bleazard, W., McCaffery, J. M., King, E. J., Bale, S., Mozdy, A., Tieu, Q., Nunnari, J., & Shaw, J. M. (1999). The dynamin-related GTPase dnm1 regulates mitochondrial fission in yeast. *Nature Cell Biology*, *1*(5), 298–304. https://doi.org/10.1038/13014

Bogenhagen, D., & Clayton, D. A. (1977). Mouse l cell mitochondrial DNA molecules are selected randomly for replication throughout the cell cycle [Publisher: Elsevier]. *Cell*, *11*(4), 719–727. https://doi.org/10.1016/0092-8674(77)90286-0

Bogenhagen, D. F. (2012). Mitochondrial DNA nucleoid structure. *Biochimica Et Biophysica Acta*, *1819*(9), 914–920. https://doi.org/10.1016/j.bbagrm.2011.11.005

Borsa, M., Barnstorf, I., Baumann, N. S., Pallmer, K., Yermanos, A., Gräbnitz, F., Barandun, N., Hausmann, A., Sandu, I., Barral, Y., & Oxenius, A. (2019). Modulation of asymmetric cell division as a mechanism to boost CD8+ t cell memory [Publisher: American Association for the Advancement of Science]. *Science Immunology*, *4*(34), eaav1730. https://doi.org/10.1126/sciimmunol.aav1730

Brown, T. A., Cecconi, C., Tkachuk, A. N., Bustamante, C., & Clayton, D. A. (2005). Replication of mitochondrial DNA occurs by strand displacement with alternative light-strand origins, not via a strand-coupled mechanism. *Genes & Development*, *19*(20), 2466–2476. https://doi.org/10.1101/gad.1352105

Brown, W. M., George, M., & Wilson, A. C. (1979). Rapid evolution of animal mitochondrial DNA. *Proceedings of the National Academy of Sciences of the United States of America*, *76*(4), 1967–1971. https://doi.org/10.1073/pnas.76.4.1967

Bruusgaard, J. C., Liestøl, K., Ekmark, M., Kollstad, K., & Gundersen, K. (2003). Number and spatial distribution of nuclei in the muscle fibres of normal mice studied in vivo. *The Journal of Physiology*, *551*, 467–478. https://doi.org/10.1113/jphysiol.2003.045328

Bua, E., Johnson, J., Herbst, A., Delong, B., McKenzie, D., Salamat, S., & Aiken, J. M. (2006). Mitochondrial DNA-deletion mutations accumulate intracellularly to detrimental levels in aged human skeletal muscle fibers. *American Journal of Human Genetics*, *79*(3), 469–480. https://doi.org/10.1086/507132

Burgstaller, J. P., Johnston, I. G., Jones, N. S., Albrechtová, J., Kolbe, T., Vogl, C., Futschik, A., Mayrhofer, C., Klein, D., Sabitzer, S., Blattner, M., Gülly, C., Poulton, J., Rülicke, T., Piálek, J., Steinborn, R., & Brem, G. (2014). mtDNA segregation in heteroplasmic tissues is common in vivo and modulated by haplotype differences and developmental stage [Publisher: Elsevier]. *Cell Reports*, *7*(6), 2031–2041. https://doi.org/10.1016/j.celrep.2014.05.020

Burr, S. P., & Chinnery, P. F. (2024). Origins of tissue and cell-type specificity in mitochondrial DNA (mtDNA) disease. *Human Molecular Genetics*, *33*, R3–R11. https://doi.org/10.1093/hmg/ddae059

Bury, A., Pyle, A., Vincent, A. E., Actis, P., & Hudson, G. (2024). Nanobiopsy investigation of the subcellular mtDNA heteroplasmy in human tissues [Publisher: Nature Publishing Group]. *Scientific Reports*, *14*(1), 13789. https://doi.org/10.1038/s41598-024-64455-0

Campbell, G., Krishnan, K. J., Deschauer, M., Taylor, R. W., & Turnbull, D. M. (2014). Dissecting the mechanisms underlying the accumulation of mitochondrial DNA deletions in human skeletal muscle. *Human Molecular Genetics*, *23*(17), 4612–4620. https://doi.org/10.1093/hmg/ddu176

Cao, L., Shitara, H., Horii, T., Nagao, Y., Imai, H., Abe, K., Hara, T., Hayashi, J.-I., & Yonekawa, H. (2007). The mitochondrial bottleneck occurs without reduction of mtDNA content in female mouse germ cells. *Nature Genetics*, *39*(3), 386–390. https://doi.org/10.1038/ng1970

Cao, L., Shitara, H., Sugimoto, M., Hayashi, J.-I., Abe, K., & Yonekawa, H. (2009). New evidence confirms that the mitochondrial bottleneck is generated without

reduction of mitochondrial DNA content in early primordial germ cells of mice [Publisher: Public Library of Science]. *PLOS Genetics*, *5*(12), e1000756. https://doi.org/10.1371/journal.pgen.1000756

Capps, G., Samuels, D., & Chinnery, P. (2003). A model of the nuclear control of mitochondrial DNA replication. *Journal of Theoretical Biology*, *221*(4), 565–583. https://doi.org/10.1006/jtbi.2003.3207

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*, 1. https://doi.org/10.18637/jss.v076.i01

Cavalier-Smith, T. (1987). The origin of eukaryotic and archaebacterial cells. *Annals of the New York Academy of Sciences*, *503*, 17–54. https://doi.org/10.1111/j.1749-6632.1987.tb40596.x

Chang, J. T., Palanivel, V. R., Kinjyo, I., Schambach, F., Intlekofer, A. M., Banerjee, A., Longworth, S. A., Vinup, K. E., Mrass, P., Oliaro, J., Killeen, N., Orange, J. S., Russell, S. M., Weninger, W., & Reiner, S. L. (2007). Asymmetric t lymphocyte division in the initiation of adaptive immune responses. *Science (New York, N.Y.)*, *315*(5819), 1687–1691. https://doi.org/10.1126/science.1139393

Chen, H., Chomyn, A., & Chan, D. C. (2005). Disruption of fusion results in mitochondrial heterogeneity and dysfunction. *The Journal of Biological Chemistry*, *280*(28), 26185–26192. https://doi.org/10.1074/jbc.M503062200

Chen, H., Detmer, S. A., Ewald, A. J., Griffin, E. E., Fraser, S. E., & Chan, D. C. (2003). Mitofusins mfn1 and mfn2 coordinately regulate mitochondrial fusion and are essential for embryonic development. *Journal of Cell Biology*, *160*(2), 189–200. https://doi.org/10.1083/jcb.200211046

Chen, H., McCaffery, J. M., & Chan, D. C. (2007). Mitochondrial fusion protects against neurodegeneration in the cerebellum. *Cell*, *130*(3), 548–562. https://doi.org/10.1016/j.cell.2007.06.026

Chen, H., Vermulst, M., Wang, Y. E., Chomyn, A., Prolla, T. A., McCaffery, J. M., & Chan, D. C. (2010). Mitochondrial fusion is required for mtDNA stability in skeletal muscle and tolerance of mtDNA mutations. *Cell*, *141*(2), 280–289. https://doi.org/10.1016/j.cell.2010.02.026

Chesnut, R. W., & Grey, H. M. (1981). Studies on the capacity of b cells to serve as antigen-presenting cells. *Journal of Immunology (Baltimore, Md.: 1950)*, *126*(3), 1075–1079.

Childs, J., Gomes, T. B., Vincent, A. E., Golightly, A., & Lawless, C. (2025). Bayesian classification of OXPHOS deficient skeletal myofibres [Publisher: Public Library of Science]. *PLOS Computational Biology*, *21*(2), e1012770. https://doi.org/10.1371/journal.pcbi.1012770

Chinnery, P. F., & Samuels, D. C. (1999). Relaxed replication of mtDNA: A model with implications for the expression of disease. *American Journal of Human Genetics*, *64*(4), 1158–1165. https://doi.org/10.1086/302311

Chinnery, P. F., Thorburn, D. R., Samuels, D. C., White, S. L., Dahl, H. M., Turnbull, D. M., Lightowlers, R. N., & Howell, N. (2000). The inheritance of mitochondrial DNA heteroplasmy: Random drift, selection or both? *Trends in genetics: TIG*, *16*(11), 500–505. https://doi.org/10.1016/s0168-9525(00)02120-x

Chkhaidze, K., Heide, T., Werner, B., Williams, M. J., Huang, W., Caravagna, G., Graham, T. A., & Sottoriva, A. (2019). Spatially constrained tumour growth affects

the patterns of clonal selection and neutral drift in cancer genomic data [Publisher: Public Library of Science]. *PLOS Computational Biology*, *15*(7), e1007243. https://doi.org/10.1371/journal.pcbi.1007243

Cipolat, S., Martins de Brito, O., Dal Zilio, B., & Scorrano, L. (2004). OPA1 requires mitofusin 1 to promote mitochondrial fusion. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(45), 15927–15932. https://doi.org/10.1073/pnas.0407043101

Clay Montier, L. L., Deng, J., & Bai, Y. (2009). Number matters: Control of mammalian mitochondrial DNA copy number. *Journal of genetics and genomics = Yi chuan xue bao*, *36*(3), 125–131. https://doi.org/10.1016/S1673-8527(08)60099-5

Clayton, D. A. (1982). Replication of animal mitochondrial DNA. *Cell*, *28*(4), 693–705. https://doi.org/10.1016/0092-8674(82)90049-6

Clayton, D. A., Doda, J. N., & Friedberg, E. C. (1974). The absence of a pyrimidine dimer repair mechanism in mammalian mitochondria. *Proceedings of the National Academy of Sciences of the United States of America*, *71*(7), 2777–2781. https://doi.org/10.1073/pnas.71.7.2777

Clayton, D. A. (1996). Mitochondrial DNA gets the drift [Publisher: Nature Publishing Group]. *Nature Genetics*, *14*(2), 123–125. https://doi.org/10.1038/ng1096-123

Coller, H. A., Khrapko, K., Bodyak, N. D., Nekhaeva, E., Herrero-Jimenez, P., & Thilly, W. G. (2001). High frequency of homoplasmic mitochondrial DNA mutations in human tumors can be explained without selection [Publisher: Nature Publishing Group]. *Nature Genetics*, *28*(2), 147–150. https://doi.org/10.1038/88859

Collins, M. L., Eng, S., Hoh, R., & Hellerstein, M. K. (2003). Measurement of mitochondrial DNA synthesis in vivo using a stable isotope-mass spectrometric technique [Publisher: American Physiological Society]. *Journal of Applied Physiology*, *94*(6), 2203–2211. https://doi.org/10.1152/japplphysiol.00691.2002

Conti, S., Gosling, J. P., Oakley, J. E., & O'Hagan, A. (2009). Gaussian process emulation of dynamic computer codes. *Biometrika*, *96*(3), 663–676. https://doi.org/10.1093/biomet/asp028

Costa, T., Boccignone, G., & Ferraro, M. (2012). Gaussian mixture model of heart rate variability. *PloS One*, *7*(5), e37731. https://doi.org/10.1371/journal.pone.0037731

Cree, L. M., Samuels, D. C., de Sousa Lopes, S. C., Rajasimha, H. K., Wonnapinij, P., Mann, J. R., Dahl, H.-H. M., & Chinnery, P. F. (2008). A reduction of mitochondrial DNA molecules during embryogenesis explains the rapid segregation of genotypes. *Nature Genetics*, *40*(2), 249–254. https://doi.org/10.1038/ng.2007.63

Cui, H., Kong, Y., & Zhang, H. (2012). Oxidative stress, mitochondrial dysfunction, and aging. *Journal of Signal Transduction*, *2012*, 646354. https://doi.org/10.1155/2012/646354

Cyster, J. G., & Allen, C. D. C. (2019). B cell responses: Cell interaction dynamics and decisions. *Cell*, *177*(3), 524–540. https://doi.org/10.1016/j.cell.2019.03.016

Damas, J., Carneiro, J., Amorim, A., & Pereira, F. (2014). MitoBreak: The mitochondrial DNA breakpoints database. *Nucleic Acids Research*, *42*, D1261–1268. https://doi.org/10.1093/nar/gkt982

de Brito, O. M., & Scorrano, L. (2008). Mitofusin 2 tethers endoplasmic reticulum to mitochondria [Publisher: Nature Publishing Group]. *Nature*, *456*(7222), 605–610. https://doi.org/10.1038/nature07534

de Grey, A. D. (1997). A proposed refinement of the mitochondrial free radical theory of aging. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, *19*(2), 161–166. https://doi.org/10.1002/bies.950190211

de Laat, P., Koene, S., van den Heuvel, L. P. W. J., Rodenburg, R. J. T., Janssen, M. C. H., & Smeitink, J. A. M. (2012). Clinical features and heteroplasmy in blood, urine and saliva in 34 dutch families carrying the m.3243a > g mutation. *Journal of Inherited Metabolic Disease*, *35*(6), 1059–1069. https://doi.org/10.1007/s10545-012-9465-2

Diaz, F., Bayona-Bafaluy, M. P., Rana, M., Mora, M., Hao, H., & Moraes, C. T. (2002). Human mitochondrial DNA with large deletions repopulates organelles faster than full-length genomes under relaxed copy number control. *Nucleic Acids Research*, *30*(21), 4626–4633. https://doi.org/10.1093/nar/gkf602

Diaz, F., & Moraes, C. T. (2008). Mitochondrial biogenesis and turnover. *Cell calcium*, *44*(1), 24–35. https://doi.org/10.1016/j.ceca.2007.12.004

Dubowitz, V., & Sewry, C. A. (2006, October 6). *Muscle biopsy: A practical approach*. Retrieved May 16, 2025, from https://shop.elsevier.com/books/muscle-biopsy-a-practical-approach/dubowitz/978-1-4160-2593-1

Duchen, M. R. (2004). Roles of mitochondria in health and disease. *Diabetes*, *53 Suppl 1*, S96–102. https://doi.org/10.2337/diabetes.53.2007.s96

Dunn, D. A., Cannon, M. V., Irwin, M. H., & Pinkert, C. A. (2012). Animal models of human mitochondrial DNA mutations. *Biochimica et Biophysica Acta (BBA) - General Subjects*, *1820*(5), 601–607. https://doi.org/10.1016/j.bbagen.2011.08.005

Durek, P., Nordström, K., Gasparoni, G., Salhab, A., Kressler, C., de Almeida, M., Bassler, K., Ulas, T., Schmidt, F., Xiong, J., Glažar, P., Klironomos, F., Sinha, A., Kinkley, S., Yang, X., Arrigoni, L., Amirabad, A. D., Ardakani, F. B., Feuerbach, L., … Polansky, J. K. (2016). Epigenomic profiling of human CD4+ t cells supports a linear differentiation model and highlights molecular regulators of memory development. *Immunity*, *45*(5), 1148–1161. https://doi.org/10.1016/j.immuni.2016.10.022

Elowitz, M. B., Levine, A. J., Siggia, E. D., & Swain, P. S. (2002). Stochastic gene expression in a single cell. *Science (New York, N.Y.)*, *297*(5584), 1183–1186. https://doi.org/10.1126/science.1070919

Elson, J. L., Samuels, D. C., Turnbull, D. M., & Chinnery, P. F. (2001). Random intracellular drift explains the clonal expansion of mitochondrial DNA mutations with age. *American Journal of Human Genetics*, *68*(3), 802–806. https://doi.org/10.1086/318801

Elson, J. L., Samuels, D. C., Johnson, M. A., Turnbull, D. M., & Chinnery, P. F. (2002). The length of cytochrome *c* oxidase-negative segments in muscle fibres in patients with mtDNA myopathy. *Neuromuscular Disorders*, *12*(9), 858–864. https://doi.org/10.1016/S0960-8966(02)00047-0

Esser, C., Ahmadinejad, N., Wiegand, C., Rotte, C., Sebastiani, F., Gelius-Dietrich, G., Henze, K., Kretschmann, E., Richly, E., Leister, D., Bryant, D., Steel, M. A., Lockhart, P. J., Penny, D., & Martin, W. (2004). A genome phylogeny for mitochondria among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. *Molecular Biology and Evolution*, *21*(9), 1643–1660. https://doi.org/10.1093/molbev/msh160

Fisher, H. F., Boys, R. J., Gillespie, C. S., Proctor, C. J., & Golightly, A. (2022). Parameter inference for a stochastic kinetic model of expanded polyglutamine proteins. *Biometrics*, *78*(3), 1195–1208. https://doi.org/10.1111/biom.13467

Floros, V. I., Pyle, A., Dietmann, S., Wei, W., Tang, W. C. W., Irie, N., Payne, B., Capalbo, A., Noli, L., Coxhead, J., Hudson, G., Crosier, M., Strahl, H., Khalaf, Y., Saitou, M., Ilic, D., Surani, M. A., & Chinnery, P. F. (2018). Segregation of mitochondrial DNA heteroplasmy through a developmental genetic bottleneck in human embryos [Publisher: Nature Publishing Group]. *Nature Cell Biology*, *20*(2), 144–151. https://doi.org/10.1038/s41556-017-0017-8

Frank, S., Gaume, B., Bergmann-Leitner, E. S., Leitner, W. W., Robert, E. G., Catez, F., Smith, C. L., & Youle, R. J. (2001). The role of dynamin-related protein 1, a mediator of mitochondrial fission, in apoptosis [Publisher: Elsevier]. *Developmental Cell*, *1*(4), 515–525. https://doi.org/10.1016/S1534-5807(01)00055-7

Franklin, I. G., Milne, P., Childs, J., Boggan, R. M., Barrow, I., Lawless, C., Gorman, G. S., Ng, Y. S., Collin, M., Russell, O. M., & Pickett, S. J. (2023). T cell differentiation drives the negative selection of pathogenic mitochondrial DNA variants [Publisher: Life Science Alliance Section: Research Articles]. *Life Science Alliance*, *6*(11). https://doi.org/10.26508/lsa.202302271

Frederiksen, A. L., Andersen, P. H., Kyvik, K. O., Jeppesen, T. D., Vissing, J., & Schwartz, M. (2006). Tissue specific distribution of the 3243a->g mtDNA mutation. *Journal of Medical Genetics*, *43*(8), 671–677. https://doi.org/10.1136/jmg.2005.039339

Fukui, H., & Moraes, C. T. (2009). Mechanisms of formation and accumulation of mitochondrial DNA deletions in aging neurons. *Human Molecular Genetics*, *18*(6), 1028–1036. https://doi.org/10.1093/hmg/ddn437

Gamerman, D., & Lopes, H. F. (2006, May 10). *Markov chain monte carlo: Stochastic simulation for bayesian inference, second edition* (2nd ed.). Chapman; Hall/CRC. https://doi.org/10.1201/9781482296426

Geginat, J., Lanzavecchia, A., & Sallusto, F. (2003). Proliferation and differentiation potential of human CD8+ memory t-cell subsets in response to antigen or homeostatic cytokines. *Blood*, *101*(11), 4260–4266. https://doi.org/10.1182/blood-2002-11-3577

Gelman, A., & Hill, J. (2006, December 18). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press. https://doi.org/10.1017/CBO9780511790942

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences [Publisher: Institute of Mathematical Statistics]. *Statistical Science*, *7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

Geman, S., & Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images [Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-6*(6), 721–741. https://doi.org/10.1109/TPAMI.1984.4767596

Georgoulas, A., Hillston, J., & Sanguinetti, G. (2016). Unbiased bayesian inference for population markov jump processes via random truncations. *Statistics and Computing*, *27*(4), 991. https://doi.org/10.1007/s11222-016-9667-9

Geweke, J. (1991). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. *Federal Reserve Bank of Minneapolis*. https://doi.org/https://doi.org/10.21034/sr.148

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, *81*(25), 2340–2361. https://doi.org/10.1021/j100540a008

Gillespie, D. T. (2001). Approximate accelerated stochastic simulation of chemically reacting systems. *The Journal of Chemical Physics*, *115*(4), 1716–1733. https://doi.org/10.1063/1.1378322

Gillespie, T. D. (1992, February). *Fundamentals of vehicle dynamics*. SAE International. https://doi.org/10.4271/R-114

Gitschlag, B. L., Kirby, C. S., Samuels, D. C., Gangula, R. D., Mallal, S. A., & Patel, M. R. (2016). Homeostatic responses regulate selfish mitochondrial genome dynamics in c. elegans. *Cell Metabolism*, *24*(1), 91–103. https://doi.org/10.1016/j.cmet.2016.06.008

Goldstein, A., & Falk, M. J. (1993). Single large-scale mitochondrial DNA deletion syndromes. In M. P. Adam, J. Feldman, G. M. Mirzaa, R. A. Pagon, S. E. Wallace, & A. Amemiya (Eds.), *GeneReviews®*. University of Washington, Seattle. Retrieved March 5, 2025, from http://www.ncbi.nlm.nih.gov/books/NBK1203/

Golightly, A., & Gillespie, C. S. (2013). Simulation of stochastic kinetic models. In M. V. Schneider (Ed.), *In silico systems biology* (pp. 169–187). Humana Press. https://doi.org/10.1007/978-1-62703-450-0_9

Golightly, A., & Wilkinson, D. J. (2011). Bayesian parameter inference for stochastic biochemical network models using particle markov chain monte carlo [Publisher: Royal Society]. *Interface Focus*, *1*(6), 807–820. https://doi.org/10.1098/rsfs.2011.0047

Gomes, T. B., Childs, J., Lawless, C., & Vincent, A. (2025). Quantification of OXPHOS deficiency variation in skeletal muscle. *Unpublished*.

Gorman, G. S., Schaefer, A. M., Ng, Y., Gomez, N., Blakely, E. L., Alston, C. L., Feeney, C., Horvath, R., Yu-Wai-Man, P., Chinnery, P. F., Taylor, R. W., Turnbull, D. M., & McFarland, R. (2015). Prevalence of nuclear and mitochondrial DNA mutations related to adult mitochondrial disease. *Annals of Neurology*, *77*(5), 753–759. https://doi.org/10.1002/ana.24362

Grady, J. P., Campbell, G., Ratnaike, T., Blakely, E. L., Falkous, G., Nesbitt, V., Schaefer, A. M., McNally, R. J., Gorman, G. S., Taylor, R. W., Turnbull, D. M., & McFarland, R. (2014). Disease progression in patients with single, large-scale mitochondrial DNA deletions. *Brain: A Journal of Neurology*, *137*, 323–334. https://doi.org/10.1093/brain/awt321

Grady, J. P., Pickett, S. J., Ng, Y. S., Alston, C. L., Blakely, E. L., Hardy, S. A., Feeney, C. L., Bright, A. A., Schaefer, A. M., Gorman, G. S., McNally, R. J., Taylor, R. W., Turnbull, D. M., & McFarland, R. (2018). mtDNA heteroplasmy level and copy number indicate disease burden in m.3243a>g mitochondrial disease. *EMBO molecular medicine*, *10*(6), e8262. https://doi.org/10.15252/emmm.201708262

Greaves, L. C., Nooteboom, M., Elson, J. L., Tuppen, H. A. L., Taylor, G. A., Commane, D. M., Arasaradnam, R. P., Khrapko, K., Taylor, R. W., Kirkwood, T. B. L., Mathers, J. C., & Turnbull, D. M. (2014). Clonal expansion of early to mid-life mitochondrial DNA point mutations drives mitochondrial dysfunction during human ageing. *PLoS genetics*, *10*(9), e1004620. https://doi.org/10.1371/journal.pgen.1004620

Greaves, L. C., Reeve, A. K., Taylor, R. W., & Turnbull, D. M. (2012). Mitochondrial DNA and disease [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/path.3028]. *The Journal of Pathology*, *226*(2), 274–286. https://doi.org/10.1002/path.3028

Gross, N. J., Getz, G. S., & Rabinowitz, M. (1968, August 26). *Apparent turnover of mitochondrial deoxyribonucleic acid and mitochondrial phospholipids in the tissues*

of the rat. Retrieved August 22, 2024, from https://www.jbc.org/article/S0021-9258(18)91795-3/pdf

Gross, N. J., Getz, G. S., & Rabinowitz, M. (1969). Apparent turnover of mitochondrial deoxyribonucleic acid and mitochondrial phospholipids in the tissues of the rat. *The Journal of Biological Chemistry*, *244*(6), 1552–1562.

Grünewald A, Lax Nz, Rocha Mc, Reeve Ak, Hepplewhite Pd, Rygiel Ka, Taylor Rw, & Turnbull Dm. (2014). Quantitative quadruple-label immunofluorescence of mitochondrial and cytoplasmic proteins in single neurons from human midbrain tissue [Publisher: J Neurosci Methods]. *Journal of neuroscience methods*, *232*(100). https://doi.org/10.1016/j.jneumeth.2014.05.026

Guo, X., Popadin, K. Y., Markuzon, N., Orlov, Y. L., Kraytsberg, Y., Krishnan, K. J., Zsurka, G., Turnbull, D. M., Kunz, W. S., & Khrapko, K. (2010). Repeats, longevity and the sources of mtDNA deletions: Evidence from 'deletional spectra'. *Trends in genetics: TIG*, *26*(8), 340–343. https://doi.org/10.1016/j.tig.2010.05.006

Harding, A. E., Sweeney, M. G., Govan, G. G., & Riordan-Eva, P. (1995). Pedigree analysis in leber hereditary optic neuropathy families with a pathogenic mtDNA mutation. *American Journal of Human Genetics*, *57*(1), 77–86. Retrieved March 4, 2025, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1801226/

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, *57*(1), 97–109. https://doi.org/10.1093/biomet/57.1.97

Hayashi, J., Ohta, S., Kikuchi, A., Takemitsu, M., Goto, Y., & Nonaka, I. (1991). Introduction of disease-related mitochondrial DNA deletions into HeLa cells lacking mitochondrial DNA results in mitochondrial dysfunction. *Proceedings of the National Academy of Sciences of the United States of America*, *88*(23), 10614–10618. Retrieved August 29, 2024, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC52980/

Hellebrekers, D. M. E. I., Blakely, E. L., Hendrickx, A. T. M., Hardy, S. A., Hopton, S., Falkous, G., de Coo, I. F. M., Smeets, H. J. M., van der Beek, N. M. E., & Taylor, R. W. (2019). A novel mitochondrial m.4414t>MT-TM gene variant causing progressive external ophthalmoplegia and myopathy. *Neuromuscular Disorders*, *29*(9), 693–697. https://doi.org/10.1016/j.nmd.2019.08.005

Henderson, D. A., Boys, R. J., & Wilkinson, D. J. (2010). Bayesian calibration of a stochastic kinetic computer model using multiple data sources. *Biometrics*, *66*(1), 249–256. https://doi.org/10.1111/j.1541-0420.2009.01245.x

Henderson, D. A., Boys, R. J., Krishnan, K. J., Lawless, C., & Wilkinson, D. J. (2009). Bayesian emulation and calibration of a stochastic computer model of mitochondrial DNA deletions in substantia nigra neurons [Publisher: ASA Website _eprint: https://doi.org/10.1198/jasa.2009.0005]. *Journal of the American Statistical Association*, *104*(485), 76–87. https://doi.org/10.1198/jasa.2009.0005

Herbers, E., Kekäläinen, N. J., Hangas, A., Pohjoismäki, J. L., & Goffart, S. (2019). Tissue specific differences in mitochondrial DNA maintenance and expression. *Mitochondrion*, *44*, 85–92. https://doi.org/10.1016/j.mito.2018.01.004

Hermann, G., Thatcher, J., Mills, J., Hale, K., Fuller, M., Nunnari, J., & Shaw, J. (1998). Mitochondrial fusion in yeast requires the transmembrane GTPase fzo1p [Publisher: J Cell Biol]. *The Journal of cell biology*, *143*(2). https://doi.org/10.1083/jcb.143.2.359

Hernández-Ainsa, C., López-Gallardo, E., García-Jiménez, M. C., Climent-Alcalá, F. J., Rodríguez-Vigil, C., García Fernández de Villalta, M., Artuch, R., Montoya, J.,

Ruiz-Pesini, E., & Emperador, S. (2022). Development and characterization of cell models harbouring mtDNA deletions for in vitro study of pearson syndrome. *Disease Models & Mechanisms*, *15*(3), dmm049083. https://doi.org/10.1242/dmm.049083

Heydari, J., Lawless, C., Lydall, D. A., & Wilkinson, D. J. (2016). Bayesian hierarchical modelling for inferring genetic interactions in yeast. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *65*(3), 367–393. https://doi.org/10.1111/rssc.12126

Hoffman, M. D., & Gelman, A. (2014). The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, *15*(1), 1593–1623.

Hoitzing, H. (2017, September). *Controlling mitochondrial dynamics: Population genetics and networks* [Doctoral dissertation]. Retrieved March 25, 2025, from http://hdl.handle.net/10044/1/58020

Hoitzing, H., Gammage, P. A., Haute, L. V., Minczuk, M., Johnston, I. G., & Jones, N. S. (2019). Energetic costs of cellular and therapeutic control of stochastic mitochondrial DNA populations. *PLoS computational biology*, *15*(6), e1007023. https://doi.org/10.1371/journal.pcbi.1007023

Hoitzing, H., Johnston, I. G., & Jones, N. S. (2017). Stochastic models for evolving cellular populations of mitochondria: Disease, development, and ageing. In D. Holcman (Ed.), *Stochastic processes, multiscale modeling, and numerical methods for computational cellular biology* (pp. 287–314). Springer International Publishing. https://doi.org/10.1007/978-3-319-62627-7_13

Holt, A. G., & Davies, A. M. (2022). A comparison of mtDNA deletion mutant proliferation mechanisms. *Journal of Theoretical Biology*, *551-552*, 111244. https://doi.org/10.1016/j.jtbi.2022.111244

Huang, X., Sun, L., Ji, S., Zhao, T., Zhang, W., Xu, J., Zhang, J., Wang, Y., Wang, X., Franzini-Armstrong, C., Zheng, M., & Cheng, H. (2013). Kissing and nanotunneling mediate intermitochondrial communication in the heart. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(8), 2846–2851. https://doi.org/10.1073/pnas.1300741110

Iborra, F. J., Kimura, H., & Cook, P. R. (2004). The functional organization of mitochondrial genomes in human cells. *BMC Biology*, *2*(1), 9. https://doi.org/10.1186/1741-7007-2-9

Insalata, F., Hoitzing, H., Aryaman, J., & Jones, N. S. (2022). Stochastic survival of the densest and mitochondrial DNA clonal expansion in aging [Publisher: Proceedings of the National Academy of Sciences]. *Proceedings of the National Academy of Sciences*, *119*(49), e2122073119. https://doi.org/10.1073/pnas.2122073119

Itoh, K., Weis, S., Mehraein, P., & Müller-Höcker, J. (1996). Cytochrome c oxidase defects of the human substantia nigra in normal aging. *Neurobiology of Aging*, *17*(6), 843–848. https://doi.org/10.1016/s0197-4580(96)00168-6

Jackson, S. E., Vernon, I., Liu, J., & Lindsey, K. (2020). Understanding hormonal crosstalk in arabidopsis root development via emulation and history matching [Publisher: De Gruyter]. *Statistical Applications in Genetics and Molecular Biology*, *19*(2). https://doi.org/10.1515/sagmb-2018-0053

James, D. I., Parone, P. A., Mattenberger, Y., & Martinou, J.-C. (2003). hFis1, a novel component of the mammalian mitochondrial fission machinery. *The Journal of Biological Chemistry*, *278*(38), 36373–36379. https://doi.org/10.1074/jbc.M303758200

Jasra, A., Holmes, C. C., & Stephens, D. A. (2005). Markov chain monte carlo methods and the label switching problem in bayesian mixture modeling. *Statistical Science*, *20*(1), 50–67. https://doi.org/10.1214/088342305000000016

Jellusova, J., Cato, M. H., Apgar, J. R., Ramezani-Rad, P., Leung, C. R., Chen, C., Richardson, A. D., Conner, E. M., Benschop, R. J., Woodgett, J. R., & Rickert, R. C. (2017). Gsk3 is a metabolic checkpoint regulator in b cells [Publisher: Nature Publishing Group]. *Nature Immunology*, *18*(3), 303–312. https://doi.org/10.1038/ni.3664

Jenuth, J. P., Peterson, A. C., Fu, K., & Shoubridge, E. A. (1996). Random genetic drift in the female germline explains the rapid segregation of mammalian mitochondrial DNA. *Nature Genetics*, *14*(2), 146–151. https://doi.org/10.1038/ng1096-146

Johnston, I. G., Burgstaller, J. P., Havlicek, V., Kolbe, T., Rülicke, T., Brem, G., Poulton, J., & Jones, N. S. (2015). Stochastic modelling, bayesian inference, and new in vivo measurements elucidate the debated mtDNA bottleneck mechanism (J. Nunnari, Ed.) [Publisher: eLife Sciences Publications, Ltd]. *eLife*, *4*, e07464. https://doi.org/10.7554/eLife.07464

Johnston, I. G., & Jones, N. S. (2016). Evolution of cell-to-cell variability in stochastic, controlled, heteroplasmic mtDNA populations [Publisher: Elsevier]. *The American Journal of Human Genetics*, *99*(5), 1150–1162. https://doi.org/10.1016/j.ajhg.2016.09.016

Kanki, T., & Klionsky, D. J. (2008). Mitophagy in yeast occurs through a selective mechanism. *The Journal of Biological Chemistry*, *283*(47), 32386–32393. https://doi.org/10.1074/jbc.M802403200

Kazak, L., Reyes, A., & Holt, I. J. (2012). Minimizing the damage: Repair pathways keep mitochondrial DNA intact. *Nature Reviews. Molecular Cell Biology*, *13*(10), 659–671. https://doi.org/10.1038/nrm3439

Kelly, R. D. W., Mahmud, A., McKenzie, M., Trounce, I. A., & St John, J. C. (2012). Mitochondrial DNA copy number is regulated in a tissue specific manner by DNA methylation of the nuclear-encoded DNA polymerase gamma a. *Nucleic Acids Research*, *40*(20), 10124–10138. https://doi.org/10.1093/nar/gks770

Kelsoe, G. (1996). Life and death in germinal centers (redux) [Publisher: Elsevier]. *Immunity*, *4*(2), 107–111. https://doi.org/10.1016/S1074-7613(00)80675-5

Kim, T.-Y., Wang, D., Kim, A. K., Lau, E., Lin, A. J., Liem, D. A., Zhang, J., Zong, N. C., Lam, M. P. Y., & Ping, P. (2012). Metabolic labeling reveals proteome dynamics of mouse mitochondria. *Molecular & cellular proteomics: MCP*, *11*(12), 1586–1594. https://doi.org/10.1074/mcp.M112.021162

King, M. P., & Attardi, G. (1989). Human cells lacking mtDNA: Repopulation with exogenous mitochondria by complementation. *Science (New York, N.Y.)*, *246*(4929), 500–503. https://doi.org/10.1126/science.2814477

Korr, H., Kurz, C., Seidler, T. O., Sommer, D., & Schmitz, C. (1998). Mitochondrial DNA synthesis studied autoradiographically in various cell types in vivo. *Brazilian Journal of Medical and Biological Research = Revista Brasileira De Pesquisas Medicas E Biologicas*, *31*(2), 289–298. https://doi.org/10.1590/s0100-879x1998000200012

Kowald, A., Dawson, M., & Kirkwood, T. B. L. (2014). Mitochondrial mutations and ageing: Can mitochondrial deletion mutants accumulate via a size based replication advantage? *Journal of Theoretical Biology*, *340*, 111–118. https://doi.org/10.1016/j.jtbi.2013.09.009

Kowald, A., & Kirkwood, T. B. L. (2013). Mitochondrial mutations and aging: Random drift is insufficient to explain the accumulation of mitochondrial deletion mutants in short-lived animals [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/acel.12098]. *Aging Cell*, *12*(4), 728–731. https://doi.org/10.1111/acel.12098

Kowald, A., & Kirkwood, T. B. L. (2014). Transcription could be the key to the selection advantage of mitochondrial deletion mutants in aging [Publisher: Proceedings of the National Academy of Sciences]. *Proceedings of the National Academy of Sciences*, *111*(8), 2972–2977. https://doi.org/10.1073/pnas.1314970111

Kraytsberg, Y., Kudryavtseva, E., McKee, A. C., Geula, C., Kowall, N. W., & Khrapko, K. (2006). Mitochondrial DNA deletions are abundant and cause functional impairment in aged human substantia nigra neurons. *Nature Genetics*, *38*(5), 518–520. https://doi.org/10.1038/ng1778

Krishnan, K. J., Reeve, A. K., Samuels, D. C., Chinnery, P. F., Blackwood, J. K., Taylor, R. W., Wanrooij, S., Spelbrink, J. N., Lightowlers, R. N., & Turnbull, D. M. (2008). What causes mitochondrial DNA deletions in human cells? *Nature Genetics*, *40*(3), 275–279. https://doi.org/10.1038/ng.f.94

Kruschke, J. K., & Liddell, T. M. (2018). The bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a bayesian perspective. *Psychonomic Bulletin & Review*, *25*(1), 178–206. https://doi.org/10.3758/s13423-016-1221-4

Kulkarni, S., Krell, M. M., Nabarro, S., & Moritz, C. A. (2022). Hardware-accelerated simulation-based inference of stochastic epidemiology models for COVID-19. *ACM Journal on Emerging Technologies in Computing Systems*, *18*(2), 1–24. https://doi.org/10.1145/3471188

Kunkel, T. A., & Loeb, L. A. (1981). Fidelity of mammalian DNA polymerases. *Science (New York, N.Y.)*, *213*(4509), 765–767. https://doi.org/10.1126/science.6454965

Kurtz, T. G. (1972). The relationship between stochastic and deterministic models for chemical reactions. *The Journal of Chemical Physics*, *57*(7), 2976–2978. https://doi.org/10.1063/1.1678692

Kuznetsov, A. V., Hermann, M., Saks, V., Hengster, P., & Margreiter, R. (2009). The cell-type specificity of mitochondrial dynamics. *The International Journal of Biochemistry & Cell Biology*, *41*(10), 1928–1939. https://doi.org/10.1016/j.biocel.2009.03.007

Lakshmanan, L. N., Yee, Z., Ng, L. F., Gunawan, R., Halliwell, B., & Gruber, J. (2018). Clonal expansion of mitochondrial DNA deletions is a private mechanism of aging in long-lived animals. *Aging Cell*, *17*(5), e12814. https://doi.org/10.1111/acel.12814

Langdahl, J. H., Larsen, M., Frost, M., Andersen, P. H., Yderstraede, K. B., Vissing, J., Dunø, M., Thomassen, M., & Frederiksen, A. L. (2018). Lecocytes mutation load declines with age in carriers of the m.3243a>g mutation: A 10-year prospective cohort. *Clinical Genetics*, *93*(4), 925–928. https://doi.org/10.1111/cge.13201

Lareau, C. A., Dubois, S. M., Buquicchio, F. A., Hsieh, Y.-H., Garg, K., Kautz, P., Nitsch, L., Praktiknjo, S. D., Maschmeyer, P., Verboon, J. M., Gutierrez, J. C., Yin, Y., Fiskin, E., Luo, W., Mimitou, E. P., Muus, C., Malhotra, R., Parikh, S., Fleming, M. D., … Ludwig, L. S. (2023). Single-cell multi-omics of mitochondrial DNA disorders reveals dynamics of purifying selection across human immune cells. *Nature Genetics*, *55*(7), 1198–1209. https://doi.org/10.1038/s41588-023-01433-8

Larsson, N. G., Tulinius, M. H., Holme, E., Oldfors, A., Andersen, O., Wahlström, J., & Aasly, J. (1992). Segregation and manifestations of the mtDNA tRNA(lys) a–>g(8344) mutation of myoclonus epilepsy and ragged-red fibers (MERRF) syndrome. *American Journal of Human Genetics*, *51*(6), 1201–1212. Retrieved March 11, 2025, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1682923/

Lecca, P. (2013). Stochastic chemical kinetics. *Biophysical Reviews*, *5*(4), 323–345. https://doi.org/10.1007/s12551-013-0122-2

Lee, J. E., Westrate, L. M., Wu, H., Page, C., & Voeltz, G. K. (2016). Multiple dynamin family members collaborate to drive mitochondrial division [Publisher: Nature Publishing Group]. *Nature*, *540*(7631), 139–143. https://doi.org/10.1038/nature20555

Legesse-Miller, A., Massol, R. H., & Kirchhausen, T. (2003). Constriction and dnm1p recruitment are distinct processes in mitochondrial fission. *Molecular Biology of the Cell*, *14*(5), 1953–1963. https://doi.org/10.1091/mbc.e02-10-0657

Lehmann, D., Tuppen, H. A. L., Campbell, G. E., Alston, C. L., Lawless, C., Rosa, H. S., Rocha, M. C., Reeve, A. K., Nicholls, T. J., Deschauer, M., Zierz, S., Taylor, R. W., Turnbull, D. M., & Vincent, A. E. (2019). Understanding mitochondrial DNA maintenance disorders at the single muscle fibre level. *Nucleic Acids Research*, *47*(14), 7430–7443. https://doi.org/10.1093/nar/gkz472

Lemasters, J. J. (2007). Modulation of mitochondrial membrane permeability in pathogenesis, autophagy and control of metabolism. *Journal of Gastroenterology and Hepatology*, *22 Suppl 1*, S31–37. https://doi.org/10.1111/j.1440-1746.2006.04643.x

Liu, P., Qian, L., Sung, J.-S., de Souza-Pinto, N. C., Zheng, L., Bogenhagen, D. F., Bohr, V. A., Wilson, D. M., Shen, B., & Demple, B. (2008). Removal of oxidative DNA damage via FEN1-dependent long-patch base excision repair in human cell mitochondria. *Molecular and Cellular Biology*, *28*(16), 4975–4987. https://doi.org/10.1128/MCB.00457-08

Longley, M. J., Graziewicz, M. A., Bienstock, R. J., & Copeland, W. C. (2005). Consequences of mutations in human DNA polymerase gamma. *Gene*, *354*, 125–131. https://doi.org/10.1016/j.gene.2005.03.029

Longley, M. J., Prasad, R., Srivastava, D. K., Wilson, S. H., & Copeland, W. C. (1998). Identification of 5 -deoxyribose phosphate lyase activity in human DNA polymerase and its role in mitochondrial base excision repair in vitro. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(21), 12244–12248. Retrieved November 28, 2024, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC22816/

López-Gallardo, E., López-Pérez, M. J., Montoya, J., & Ruiz-Pesini, E. (2009). CPEO and KSS differ in the percentage and location of the mtDNA deletion. *Mitochondrion*, *9*(5), 314–317. https://doi.org/10.1016/j.mito.2009.04.005

Ludwig, L. S., Lareau, C. A., Ulirsch, J. C., Christian, E., Muus, C., Li, L. H., Pelka, K., Ge, W., Oren, Y., Brack, A., Law, T., Rodman, C., Chen, J. H., Boland, G. M., Hacohen, N., Rozenblatt-Rosen, O., Aryee, M. J., Buenrostro, J. D., Regev, A., & Sankaran, V. G. (2019). Lineage tracing in humans enabled by mitochondrial mutations and single-cell genomics. *Cell*, *176*(6), 1325–1339.e22. https://doi.org/10.1016/j.cell.2019.01.022

Macmillan, C., Kirkham, T., Fu, K., Allison, V., Andermann, E., Chitayat, D., Fortier, D., Gans, M., Hare, H., Quercia, N., Zackon, D., & Shoubridge, E. A. (1998).

Pedigree analysis of french canadian families with t14484c leber's hereditary optic neuropathy. *Neurology*, *50*(2), 417–422. https://doi.org/10.1212/wnl.50.2.417

Mancuso, M., Orsucci, D., Angelini, C., Bertini, E., Carelli, V., Comi, G. P., Donati, M. A., Federico, A., Minetti, C., Moggio, M., Mongini, T., Santorelli, F. M., Servidei, S., Tonin, P., Toscano, A., Bruno, C., Bello, L., Caldarazzo Ienco, E., Cardaioli, E., … Siciliano, G. (2015). Redefining phenotypes associated with mitochondrial DNA single deletion. *Journal of Neurology*, *262*(5), 1301–1309. https://doi.org/10.1007/s00415-015-7710-y

Mao, L., Zabel, C., Wacker, M. A., Nebrich, G., Sagi, D., Schrade, P., Bachmann, S., Kowald, A., & Klose, J. (2006). Estimation of the mtDNA mutation rate in aging mice by proteome analysis and mathematical modeling. *Experimental Gerontology*, *41*(1), 11–24. https://doi.org/10.1016/j.exger.2005.09.012

Martin, W., & Müller, M. (1998). The hydrogen hypothesis for the first eukaryote [Publisher: Nature Publishing Group]. *Nature*, *392*(6671), 37–41. https://doi.org/10.1038/32096

McAuley, M. T., Kenny, R. A., Kirkwood, T. B., Wilkinson, D. J., Jones, J. J., & Miller, V. M. (2009). A mathematical model of aging-related and cortisol induced hippocampal dysfunction. *BMC Neuroscience*, *10*(1), 26. https://doi.org/10.1186/1471-2202-10-26

Mengersen, K. L., Robert, C. P., & Guihenneuc-Jouyaux, C. (1999, August 12). MCMC convergence diagnostics: A review. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 6: Proceedings of the sixth valencia international meeting june 6-10, 1998* (pp. 415–440). Oxford University Press. https://doi.org/10.1093/oso/9780198504856.003.0018

Menzies, R. A., & Gold, P. H. (1971). The turnover of mitochondria in a variety of tissues of young adult and aged rats. *The Journal of Biological Chemistry*, *246*(8), 2425–2429.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*(6), 1087–1092. https://doi.org/10.1063/1.1699114

Miller, F. J., Rosenfeldt, F. L., Zhang, C., Linnane, A. W., & Nagley, P. (2003). Precise determination of mitochondrial DNA copy number in human skeletal and cardiac muscle by a PCR-based assay: Lack of change of copy number with age. *Nucleic Acids Research*, *31*(11), e61. https://doi.org/10.1093/nar/gng060

Moeller, M. E., Mon Père, N. V., Werner, B., & Huang, W. (2024). Measures of genetic diversification in somatic tissues at bulk and single-cell resolution (J. Flegg & D. Weigel, Eds.) [Publisher: eLife Sciences Publications, Ltd]. *eLife*, *12*, RP89780. https://doi.org/10.7554/eLife.89780

Monnot, S., Gigarel, N., Samuels, D. C., Burlet, P., Hesters, L., Frydman, N., Frydman, R., Kerbrat, V., Funalot, B., Martinovic, J., Benachi, A., Feingold, J., Munnich, A., Bonnefont, J.-P., & Steffann, J. (2011). Segregation of mtDNA throughout human embryofetal development: M.3243a > g as a model system. *Human Mutation*, *32*(1), 116–125. https://doi.org/10.1002/humu.21417

Moslemi, A. R., Tulinius, M., Holme, E., & Oldfors, A. (1998). Threshold expression of the tRNA(lys) a8344g mutation in single muscle fibres. *Neuromuscular disorders: NMD*, *8*(5), 345–349. https://doi.org/10.1016/s0960-8966(98)00029-7

Mozdy, A. D., McCaffery, J. M., & Shaw, J. M. (2000). Dnm1p gtpase-mediated mitochondrial fission is a multi-step process requiring the novel integral membrane

component fis1p. *The Journal of Cell Biology*, *151*(2), 367. https://doi.org/10.1083/jcb.151.2.367

Nass, M. M. (1966). The circularity of mitochondrial DNA. *Proceedings of the National Academy of Sciences of the United States of America*, *56*(4), 1215–1222. Retrieved July 17, 2024, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC220045/

Nesbitt, V., Pitceathly, R. D. S., Turnbull, D. M., Taylor, R. W., Sweeney, M. G., Mudanohwo, E. E., Rahman, S., Hanna, M. G., & McFarland, R. (2013). The UK MRC mitochondrial disease patient cohort study: Clinical phenotypes associated with the m.3243a>g mutation–implications for diagnosis and management. *Journal of Neurology, Neurosurgery, and Psychiatry*, *84*(8), 936–938. https://doi.org/10.1136/jnnp-2012-303528

Ng, Y. S., Thompson, K., Loher, D., Hopton, S., Falkous, G., Hardy, S. A., Schaefer, A. M., Shaunak, S., Roberts, M. E., Lilleker, J. B., & Taylor, R. W. (2020). Novel MT-ND gene variants causing adult-onset mitochondrial disease and isolated complex i deficiency. *Frontiers in Genetics*, *11*. Retrieved October 31, 2023, from https://www.frontiersin.org/articles/10.3389/fgene.2020.00024

Noji, H., Yasuda, R., Yoshida, M., & Kinosita, K. (1997). Direct observation of the rotation of f1-ATPase [Publisher: Nature Publishing Group]. *Nature*, *386*(6622), 299–302. https://doi.org/10.1038/386299a0

Norris, J. R. (1997). *Markov chains*. Cambridge University Press. https://doi.org/10.1017/CBO9780511810633

Olivo, P. D., Van de Walle, M. J., Laipis, P. J., & Hauswirth, W. W. (1983). Nucleotide sequence evidence for rapid genotypic shifts in the bovine mitochondrial DNA d-loop. *Nature*, *306*(5941), 400–402. https://doi.org/10.1038/306400a0

Oyebamiji, O. K., Wilkinson, D. J., Jayathilake, P. G., Curtis, T. P., Rushton, S. P., Li, B., & Gupta, P. (2017). Gaussian process emulation of an individual-based model simulation of microbial communities. *Journal of Computational Science*, *22*, 69–84. https://doi.org/10.1016/j.jocs.2017.08.006

Pagliuso, A., Tham, T. N., Stevens, J. K., Lagache, T., Persson, R., Salles, A., Olivo-Marin, J.-C., Oddos, S., Spang, A., Cossart, P., & Stavru, F. (2016). A role for septin 2 in drp1-mediated mitochondrial fission. *EMBO reports*, *17*(6), 858–873. https://doi.org/10.15252/embr.201541612

Palade, G. E. (1953). An electron microscope study of the mitochondrial structure. *The Journal of Histochemistry and Cytochemistry: Official Journal of the Histochemistry Society*, *1*(4), 188–211. https://doi.org/10.1177/1.4.188

Pallotti, F., Binelli, G., Fabbri, R., Valentino, M. L., Vicenti, R., Macciocca, M., Cevoli, S., Baruzzi, A., DiMauro, S., & Carelli, V. (2014). A wide range of 3243a>g/tRNALeu(UUR) (MELAS) mutation loads may segregate in offspring through the female germline bottleneck. *PloS One*, *9*(5), e96663. https://doi.org/10.1371/journal.pone.0096663

Perkins, G., Renken, C., Martone, M. E., Young, S. J., Ellisman, M., & Frey, T. (1997). Electron tomography of neuronal mitochondria: Three-dimensional structure and organization of cristae and membrane contacts. *Journal of Structural Biology*, *119*(3), 260–272. https://doi.org/10.1006/jsbi.1997.3885

Pickett, S. J., Grady, J. P., Ng, Y. S., Gorman, G. S., Schaefer, A. M., Wilson, I. J., Cordell, H. J., Turnbull, D. M., Taylor, R. W., & McFarland, R. (2018). Phenotypic heterogeneity in m.3243a>g mitochondrial disease: The role of nuclear factors.

Annals of Clinical and Translational Neurology, *5*(3), 333–345. https://doi.org/10.1002/acn3.532

Pitceathly, R. D. S., Rahman, S., & Hanna, M. G. (2012). Single deletions in mitochondrial DNA–molecular mechanisms and disease phenotypes in clinical practice. *Neuromuscular disorders: NMD*, *22*(7), 577–586. https://doi.org/10.1016/j.nmd.2012.03.009

Plummer, M. (2003). JAGS: A program for analysis of bayesian graphical models using gibbs sampling.

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R news*, *6*(1), 7–11. https://journal.r-project.org/archive/

Plummer, M., Stukalov, A., & Denwood, M. (2024, August 19). *Rjags: Bayesian graphical models using MCMC* (Version 4-16). Retrieved March 18, 2025, from https://cran.r-project.org/web/packages/rjags/index.html

Pollizzi, K. N., Sun, I.-H., Patel, C. H., Lo, Y.-C., Oh, M.-H., Waickman, A. T., Tam, A. J., Blosser, R. L., Wen, J., Delgoffe, G. M., & Powell, J. D. (2016). Asymmetric inheritance of mTORC1 kinase activity during division dictates CD8+ t cell differentiation [Publisher: Nature Publishing Group]. *Nature Immunology*, *17*(6), 704–711. https://doi.org/10.1038/ni.3438

Rahman, S., Poulton, J., Marchington, D., & Suomalainen, A. (2001). Decrease of 3243 a→g mtDNA mutation from blood in MELAS syndrome: A longitudinal study. *American Journal of Human Genetics*, *68*(1), 238–240. Retrieved February 10, 2025, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1234919/

Rajasimha, H. K., Chinnery, P. F., & Samuels, D. C. (2008). Selection against pathogenic mtDNA mutations in a stem cell population leads to the loss of the 3243a–>g mutation in blood. *American Journal of Human Genetics*, *82*(2), 333–343. https://doi.org/10.1016/j.ajhg.2007.10.007

Rapaport, D., Brunner, M., Neupert, W., & Westermann, B. (1998). Fzo1p is a mitochondrial outer membrane protein essential for the biogenesis of functional mitochondria in *Saccharomyces cerevisiae**. *Journal of Biological Chemistry*, *273*(32), 20150–20155. https://doi.org/10.1074/jbc.273.32.20150

Rath, S., Sharma, R., Gupta, R., Ast, T., Chan, C., Durham, T. J., Goodman, R. P., Grabarek, Z., Haas, M. E., Hung, W. H. W., Joshi, P. R., Jourdain, A. A., Kim, S. H., Kotrys, A. V., Lam, S. S., McCoy, J. G., Meisel, J. D., Miranda, M., Panda, A., … Mootha, V. K. (2021). MitoCarta3.0: An updated mitochondrial proteome now with sub-organelle localization and pathway annotations. *Nucleic Acids Research*, *49*, D1541–D1547. https://doi.org/10.1093/nar/gkaa1011

Rausser, S., Trumpff, C., McGill, M. A., Junker, A., Wang, W., Ho, S.-H., Mitchell, A., Karan, K. R., Monk, C., Segerstrom, S. C., Reed, R. G., & Picard, M. (2021). Mitochondrial phenotypes in purified human immune cell subtypes and cell mixtures. *eLife*, *10*, e70899. https://doi.org/10.7554/eLife.70899

Reynolds, D. (2009). Gaussian mixture models. In S. Z. Li & A. Jain (Eds.), *Encyclopedia of biometrics* (pp. 659–663). Springer US. https://doi.org/10.1007/978-0-387-73003-5_196

Rickett, L. M., Pullen, N., Hartley, M., Zipfel, C., Kamoun, S., Baranyi, J., & Morris, R. J. (2015). Incorporating prior knowledge improves detection of differences in bacterial growth rate. *BMC systems biology*, *9*, 60. https://doi.org/10.1186/s12918-015-0204-9

Robberson, D. L., Kasamatsu, H., & Vinograd, J. (1972). Replication of mitochondrial DNA. circular replicative intermediates in mouse l cells. *Proceedings of the National Academy of Sciences of the United States of America*, *69*(3), 737–741. https://doi.org/10.1073/pnas.69.3.737

Roberts, G. O., & Rosenthal, J. S. (2001). Optimal scaling for various metropolis-hastings algorithms [Publisher: Institute of Mathematical Statistics]. *Statistical Science*, *16*(4), 351–367. https://doi.org/10.1214/ss/1015346320

Roberts, G. O., & Rosenthal, J. S. (2004). General state space markov chains and MCMC algorithms [Publisher: Institute of Mathematical Statistics and Bernoulli Society]. *Probability Surveys*, *1*, 20–71. https://doi.org/10.1214/154957804100000024

Rocha, M., Grady, J., Grünewald, A., Vincent, A., Dobson, P., Taylor, R., Turnbull, D., & Rygiel, K. (2015). A novel immunofluorescent assay to investigate oxidative phosphorylation deficiency in mitochondrial myopathy: Understanding mechanisms and improving diagnosis [Publisher: Sci Rep]. *Scientific reports*, *5*. https://doi.org/10.1038/srep15037

Rocha, M. C., Rosa, H. S., Grady, J. P., Blakely, E. L., He, L., Romain, N., Haller, R. G., Newman, J., McFarland, R., Ng, Y. S., Gorman, G. S., Schaefer, A. M., Tuppen, H. A., Taylor, R. W., & Turnbull, D. M. (2018). Pathological mechanisms underlying single large-scale mitochondrial DNA deletions. *Annals of Neurology*, *83*(1), 115–130. https://doi.org/10.1002/ana.25127

Roger, A. J., Muñoz-Gómez, S. A., & Kamikawa, R. (2017). The origin and diversification of mitochondria. *Current Biology*, *27*(21), R1177–R1192. https://doi.org/10.1016/j.cub.2017.09.015

Rossignol, R., Faustin, B., Rocher, C., Malgat, M., Mazat, J.-P., & Letellier, T. (2003). Mitochondrial threshold effects. *The Biochemical Journal*, *370*, 751–762. https://doi.org/10.1042/BJ20021594

Rossignol, R., Malgat, M., Mazat, J.-P., & Letellier, T. (1999). Threshold effect and tissue specificity: IMPLICATION FOR MITOCHONDRIAL CYTOPATHIES * [Publisher: Elsevier]. *Journal of Biological Chemistry*, *274*(47), 33426–33432. https://doi.org/10.1074/jbc.274.47.33426

Rouzier, C., Bannwarth, S., Chaussenot, A., Chevrollier, A., Verschueren, A., Bonello-Palot, N., Fragaki, K., Cano, A., Pouget, J., Pellissier, J.-F., Procaccio, V., Chabrol, B., & Paquis-Flucklinger, V. (2012). The MFN2 gene is responsible for mitochondrial DNA instability and optic atrophy 'plus' phenotype. *Brain: A Journal of Neurology*, *135*, 23–34. https://doi.org/10.1093/brain/awr323

Rufer, N., Zippelius, A., Batard, P., Pittet, M. J., Kurth, I., Corthesy, P., Cerottini, J.-C., Leyvraz, S., Roosnek, E., Nabholz, M., & Romero, P. (2003). Ex vivo characterization of human CD8+ t subsets with distinct replicative history and partial effector functions. *Blood*, *102*(5), 1779–1787. https://doi.org/10.1182/blood-2003-02-0420

Russell-Buckland, J., Barnes, C. P., & Tachtsidis, I. (2019). A bayesian framework for the analysis of systems biology models of the brain [Publisher: Public Library of Science]. *PLOS Computational Biology*, *15*(4), e1006631. https://doi.org/10.1371/journal.pcbi.1006631

Sallusto, F., Lenig, D., Förster, R., Lipp, M., & Lanzavecchia, A. (1999). Two subsets of memory t lymphocytes with distinct homing potentials and effector functions [Publisher: Nature Publishing Group]. *Nature*, *401*(6754), 708–712. https://doi.org/10.1038/44385

Santos, R. X., Correia, S. C., Wang, X., Perry, G., Smith, M. A., Moreira, P. I., & Zhu, X. (2010). A synergistic dysfunction of mitochondrial fission/fusion dynamics and mitophagy in alzheimer's disease. *Journal of Alzheimer's disease: JAD*, *20 Suppl 2*, S401–412. https://doi.org/10.3233/JAD-2010-100666

Satoh, M., & Kuroiwa, T. (1991). Organization of multiple nucleoids and DNA molecules in mitochondria of a human cell. *Experimental Cell Research*, *196*(1), 137–140. https://doi.org/10.1016/0014-4827(91)90467-9

Schaefer, A. M., McFarland, R., Blakely, E. L., He, L., Whittaker, R. G., Taylor, R. W., Chinnery, P. F., & Turnbull, D. M. (2008). Prevalence of mitochondrial DNA disease in adults. *Annals of Neurology*, *63*(1), 35–39. https://doi.org/10.1002/ana.21217

Schon, E. A., Rizzuto, R., Moraes, C. T., Nakase, H., Zeviani, M., & DiMauro, S. (1989). A direct repeat is a hotspot for large-scale deletion of human mitochondrial DNA. *Science (New York, N.Y.)*, *244*(4902), 346–349. https://doi.org/10.1126/science.2711184

Schultz, R. A., Swoap, S. J., McDaniel, L. D., Zhang, B., Koon, E. C., Garry, D. J., Li, K., & Williams, R. S. (1998). Differential expression of mitochondrial DNA replication factors in mammalian tissues. *The Journal of Biological Chemistry*, *273*(6), 3447–3451. https://doi.org/10.1074/jbc.273.6.3447

Schwarz, G. (1978). Estimating the dimension of a model [Publisher: Institute of Mathematical Statistics]. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Sebastián, D., Hernández-Alvarez, M. I., Segalés, J., Sorianello, E., Muñoz, J. P., Sala, D., Waget, A., Liesa, M., Paz, J. C., Gopalacharyulu, P., Orešič, M., Pich, S., Burcelin, R., Palacín, M., & Zorzano, A. (2012). Mitofusin 2 (mfn2) links mitochondrial and endoplasmic reticulum function with insulin signaling and is essential for normal glucose homeostasis. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(14), 5523–5528. https://doi.org/10.1073/pnas.1108220109

Shenkar, R., Navidi, W., Tavaré, S., Dang, M. H., Chomyn, A., Attardi, G., Cortopassi, G., & Arnheim, N. (1996). The mutation rate of the human mtDNA deletion mtDNA4977. *American Journal of Human Genetics*, *59*(4), 772–780.

Shervegar, M. V., & Bhat, G. V. (2018). Heart sound classification using gaussian mixture model. *Porto Biomedical Journal*, *3*(1), e4. https://doi.org/10.1016/j.pbj.0000000000000004

Shoffner, J. M., Lott, M. T., Lezza, A. M., Seibel, P., Ballinger, S. W., & Wallace, D. C. (1990). Myoclonic epilepsy and ragged-red fiber disease (MERRF) is associated with a mitochondrial DNA tRNA(lys) mutation. *Cell*, *61*(6), 931–937. https://doi.org/10.1016/0092-8674(90)90059-n

Shoffner, J. M., Lott, M. T., Voljavec, A. S., Soueidan, S. A., Costigan, D. A., & Wallace, D. C. (1989). Spontaneous kearns-sayre/chronic external ophthalmoplegia plus syndrome associated with a mitochondrial DNA deletion: A slip-replication model and metabolic therapy. *Proceedings of the National Academy of Sciences of the United States of America*, *86*(20), 7952–7956. https://doi.org/10.1073/pnas.86.20.7952

Shoubridge, E. A., & Wai, T. (2007). Mitochondrial DNA and the mammalian oocyte. *Current Topics in Developmental Biology*, *77*, 87–111. https://doi.org/10.1016/S0070-2153(06)77004-1

Sicheritz-Pontén, T., Kurland, C. G., & Andersson, S. G. (1998). A phylogenetic analysis of the cytochrome b and cytochrome c oxidase i genes supports an origin of mitochondria from within the rickettsiaceae. *Biochimica Et Biophysica Acta*, *1365*(3), 545–551. https://doi.org/10.1016/s0005-2728(98)00099-1

Sithamparanathan, S., Rocha, M. C., Parikh, J. D., Rygiel, K. A., Falkous, G., Grady, J. P., Hollingsworth, K. G., Trenell, M. I., Taylor, R. W., Turnbull, D. M., Gorman, G. S., & Corris, P. A. (2018). Skeletal muscle mitochondrial oxidative phosphorylation function in idiopathic pulmonary arterial hypertension: In vivo and in vitro study. *Pulmonary Circulation*, *8*(2), 2045894018768290. https://doi.org/10.1177/2045894018768290

Smirnova, E., Griparic, L., Shurland, D.-L., & van der Bliek, A. M. (2001). Dynamin-related protein drp1 is required for mitochondrial division in mammalian cells [_eprint: https://doi.org/10.1091/mbc.12.8.2245]. *Molecular Biology of the Cell*, *12*(8), 2245–2256. https://doi.org/10.1091/mbc.12.8.2245

Spelbrink, J. N., Li, F. Y., Tiranti, V., Nikali, K., Yuan, Q. P., Tariq, M., Wanrooij, S., Garrido, N., Comi, G., Morandi, L., Santoro, L., Toscano, A., Fabrizi, G. M., Somer, H., Croxen, R., Beeson, D., Poulton, J., Suomalainen, A., Jacobs, H. T., … Larsson, C. (2001). Human mitochondrial DNA deletions associated with mutations in the gene encoding twinkle, a phage t7 gene 4-like protein localized in mitochondria. *Nature Genetics*, *28*(3), 223–231. https://doi.org/10.1038/90058

Stan Development Team. (2020). RStan: The r interface to stan. http://mc-stan.org/

Stewart, J. B. (2021). Current progress with mammalian models of mitochondrial DNA disease [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jimd.12324]. *Journal of Inherited Metabolic Disease*, *44*(2), 325–342. https://doi.org/10.1002/jimd.12324

Su, T., Grady, J. P., Afshar, S., McDonald, S. A., Taylor, R. W., Turnbull, D. M., & Greaves, L. C. (2018). Inherited pathogenic mitochondrial DNA mutations and gastrointestinal stem cell populations. *The Journal of Pathology*, *246*(4), 427–432. https://doi.org/10.1002/path.5156

Sue, C. M., Quigley, A., Katsabanis, S., Kapsa, R., Crimmins, D. S., Byrne, E., & Morris, J. G. (1998). Detection of MELAS a3243g point mutation in muscle, blood and hair follicles. *Journal of the Neurological Sciences*, *161*(1), 36–39. https://doi.org/10.1016/s0022-510x(98)00179-8

Sun, L., Su, Y., Jiao, A., Wang, X., & Zhang, B. (2023). T cells in health and disease [Publisher: Nature Publishing Group]. *Signal Transduction and Targeted Therapy*, *8*(1), 1–50. https://doi.org/10.1038/s41392-023-01471-y

Sun F, Huo X, Zhai Y, Wang A, Xu J, Su D, Bartlam M, & Rao Z. (2005). Crystal structure of mitochondrial respiratory membrane protein complex II [Publisher: Cell]. *Cell*, *121*(7). https://doi.org/10.1016/j.cell.2005.05.025

Sunnåker, M., Busetto, A. G., Numminen, E., Corander, J., Foll, M., & Dessimoz, C. (2013). Approximate bayesian computation [Publisher: Public Library of Science]. *PLOS Computational Biology*, *9*(1), e1002803. https://doi.org/10.1371/journal.pcbi.1002803

Tang, Y., Schon, E. A., Wilichowski, E., Vazquez-Memije, M. E., Davidson, E., & King, M. P. (2000). Rearrangements of human mitochondrial DNA (mtDNA): New insights into the regulation of mtDNA copy number and gene expression - PubMed. *Molecular Biology of the Cell*, *11*(4), 1471–1485. https://doi.org/10.1091/mbc.11.4.1471

Tarasenko, T. N., Pacheco, S. E., Koenig, M. K., Gomez-Rodriguez, J., Kapnick, S. M., Diaz, F., Zerfas, P. M., Barca, E., Sudderth, J., DeBerardinis, R. J., Covian, R., Balaban, R. S., DiMauro, S., & McGuire, P. J. (2017). Cytochrome c oxidase activity is a metabolic checkpoint that regulates cell fate decisions during t cell activation and differentiation. *Cell Metabolism*, *25*(6), 1254–1268.e7. https://doi.org/10.1016/j.cmet.2017.05.007

Thompson, K., Collier, J. J., Glasgow, R. I. C., Robertson, F. M., Pyle, A., Blakely, E. L., Alston, C. L., Oláhová, M., McFarland, R., & Taylor, R. W. (2020). Recent advances in understanding the molecular genetic basis of mitochondrial disease. *Journal of Inherited Metabolic Disease*, *43*(1), 36–50. https://doi.org/10.1002/jimd.12104

Tsukihara T, Aoyama H, Yamashita E, Tomizaki T, Yamaguchi H, Shinzawa-Itoh K, Nakashima R, Yaono R, & Yoshikawa S. (1996). The whole structure of the 13-subunit oxidized cytochrome c oxidase at 2.8 a [Publisher: Science]. *Science (New York, N.Y.)*, *272*(5265). https://doi.org/10.1126/science.272.5265.1136

Twig, G., Elorza, A., Molina, A. J. A., Mohamed, H., Wikstrom, J. D., Walzer, G., Stiles, L., Haigh, S. E., Katz, S., Las, G., Alroy, J., Wu, M., Py, B. F., Yuan, J., Deeney, J. T., Corkey, B. E., & Shirihai, O. S. (2008). Fission and selective fusion govern mitochondrial segregation and elimination by autophagy. *The EMBO journal*, *27*(2), 433–446. https://doi.org/10.1038/sj.emboj.7601963

Tyynismaa, H., Mjosund, K. P., Wanrooij, S., Lappalainen, I., Ylikallio, E., Jalanko, A., Spelbrink, J. N., Paetau, A., & Suomalainen, A. (2005). Mutant mitochondrial helicase twinkle causes multiple mtDNA deletions and a late-onset mitochondrial disease in mice. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(49), 17687–17692. https://doi.org/10.1073/pnas.0505551102

Tyynismaa, H., & Suomalainen, A. (2009). Mouse models of mitochondrial DNA defects and their relevance for human disease. *EMBO Reports*, *10*(2), 137–143. https://doi.org/10.1038/embor.2008.242

Urata, M., Wada, Y., Kim, S. H., Chumpia, W., Kayamori, Y., Hamasaki, N., & Kang, D. (2004). High-sensitivity detection of the a3243g mutation of mitochondrial DNA by a combination of allele-specific PCR and peptide nucleic acid-directed PCR clamping. *Clinical Chemistry*, *50*(11), 2045–2051. https://doi.org/10.1373/clinchem.2004.033761

Van Goethem, G., Dermaut, B., Löfgren, A., Martin, J. J., & Van Broeckhoven, C. (2001). Mutation of POLG is associated with progressive external ophthalmoplegia characterized by mtDNA deletions. *Nature Genetics*, *28*(3), 211–212. https://doi.org/10.1038/90034

Van Laar, V. S., & Berman, S. B. (2009). Mitochondrial dynamics in parkinson's disease. *Experimental Neurology*, *218*(2), 247–256. https://doi.org/10.1016/j.expneurol.2009.03.019

van Riesen, A. K. J., Antonicka, H., Ohlenbusch, A., Shoubridge, E. A., & Wilichowski, E. K. G. (2006). Maternal segmental disomy in leigh syndrome with cytochrome c oxidase deficiency caused by homozygous SURF1 mutation. *Neuropediatrics*, *37*(2), 88–94. https://doi.org/10.1055/s-2006-924227

Vardhana, S. A., Hwee, M. A., Berisa, M., Wells, D. K., Yost, K. E., King, B., Smith, M., Herrera, P. S., Chang, H. Y., Satpathy, A. T., van den Brink, M. R. M., Cross, J. R., & Thompson, C. B. (2020). Impaired mitochondrial oxidative phosphorylation limits the self-renewal of t cells exposed to persistent antigen [Publisher: Nature

Publishing Group]. *Nature Immunology*, *21*(9), 1022–1033. https://doi.org/10.1038/s41590-020-0725-2

Vats, D., Flegal, J. M., & Jones, G. L. (2019). Multivariate output analysis for markov chain monte carlo. *Biometrika*, *106*(2), 321–337. Retrieved September 30, 2024, from https://ideas.repec.org//a/oup/biomet/v106y2019i2p321-337..html

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved r-hat for assessing convergence of MCMC (with discussion). *Bayesian Analysis*, *16*(2). https://doi.org/10.1214/20-BA1221

Veitia, R. A. (2018). How the most common mitochondrial DNA mutation (m.3243a>g) vanishes from leukocytes: A mathematical model. *Human Molecular Genetics*, *27*(9), 1565–1571. https://doi.org/10.1093/hmg/ddy063

Verbist, K. C., Guy, C. S., Milasta, S., Liedmann, S., Kamiński, M. M., Wang, R., & Green, D. R. (2016). Metabolic maintenance of cell asymmetry following division in activated t lymphocytes. *Nature*, *532*(7599), 389–393. https://doi.org/10.1038/nature17442

Verma, K., Ogonek, J., Varanasi, P. R., Luther, S., Bünting, I., Thomay, K., Behrens, Y. L., Mischak-Weissinger, E., & Hambach, L. (2017). Human CD8+ CD57- TEMRA cells: Too young to be called "old" [Publisher: Public Library of Science]. *PLOS ONE*, *12*(5), e0177405. https://doi.org/10.1371/journal.pone.0177405

Vernon, I., Goldstein, M., & Bower, R. (2014). Galaxy formation: Bayesian history matching for the observable universe [Publisher: Institute of Mathematical Statistics]. *Statistical Science*, *29*(1), 81–90. Retrieved June 9, 2025, from https://www.jstor.org/stable/43288453

Victora, G. D., & Nussenzweig, M. C. (2022). Germinal centers. *Annual Review of Immunology*, *40*, 413–442. https://doi.org/10.1146/annurev-immunol-120419-022408

Vincent, A. E., Chen, C., Gomes, T. B., Di Leo, V., Laalo, T., Pabis, K., Capaldi, R., Marusich, M. F., McDonald, D., Filby, A., Fuller, A., Lehmann Urban, D., Zierz, S., Deschauer, M., Turnbull, D., Reeve, A. K., & Lawless, C. (2024). A stagewise response to mitochondrial dysfunction in mitochondrial DNA maintenance disorders. *Biochimica Et Biophysica Acta. Molecular Basis of Disease*, *1870*(5), 167131. https://doi.org/10.1016/j.bbadis.2024.167131

Vincent, A. E., Rosa, H. S., Pabis, K., Lawless, C., Chen, C., Grünewald, A., Rygiel, K. A., Rocha, M. C., Reeve, A. K., Falkous, G., Perissi, V., White, K., Davey, T., Petrof, B. J., Sayer, A. A., Cooper, C., Deehan, D., Taylor, R. W., Turnbull, D. M., & Picard, M. (2018). Subcellular origin of mitochondrial DNA deletions in human skeletal muscle. *Annals of Neurology*, *84*(2), 289–301. https://doi.org/10.1002/ana.25288

Vincent, A. E., White, K., Davey, T., Philips, J., Ogden, R. T., Lawless, C., Warren, C., Hall, M. G., Ng, Y. S., Falkous, G., Holden, T., Deehan, D., Taylor, R. W., Turnbull, D. M., & Picard, M. (2019). Quantitative 3d mapping of the human skeletal muscle mitochondrial network. *Cell Reports*, *26*(4), 996–1009.e4. https://doi.org/10.1016/j.celrep.2019.01.010

Wai, T., Teoli, D., & Shoubridge, E. A. (2008). The mitochondrial DNA genetic bottleneck results from replication of a subpopulation of genomes. *Nature Genetics*, *40*(12), 1484–1488. https://doi.org/10.1038/ng.258

Walker, M. A., Lareau, C. A., Ludwig, L. S., Karaa, A., Sankaran, V. G., Regev, A., & Mootha, V. K. (2020). Purifying selection against pathogenic mitochondrial DNA

in human t cells. *The New England Journal of Medicine*, *383*(16), 1556–1563. https://doi.org/10.1056/NEJMoa2001265

Wallace, D. C. (1989). Mitochondrial DNA mutations and neuromuscular disease. *Trends in genetics: TIG*, *5*(1), 9–13. https://doi.org/10.1016/0168-9525(89)90005-x

Wanrooij, S., & Falkenberg, M. (2010). The human mitochondrial replication fork in health and disease. *Biochimica Et Biophysica Acta*, *1797*(8), 1378–1388. https://doi.org/10.1016/j.bbabio.2010.04.015

Warren, C., McDonald, D., Capaldi, R., Deehan, D., Taylor, R. W., Filby, A., Turnbull, D. M., Lawless, C., & Vincent, A. E. (2020). Decoding mitochondrial heterogeneity in single muscle fibres by imaging mass cytometry. *Scientific Reports*, *10*(1), 15336. https://doi.org/10.1038/s41598-020-70885-3

Weber, K., Wilson, J. N., Taylor, L., Brierley, E., Johnson, M. A., Turnbull, D. M., & Bindoff, L. A. (1997). A new mtDNA mutation showing accumulation with time and restriction to skeletal muscle. *American Journal of Human Genetics*, *60*(2), 373–380.

Wilkinson, D. J. (2007). Bayesian methods in bioinformatics and computational systems biology. *Briefings in Bioinformatics*, *8*(2), 109–116. https://doi.org/10.1093/bib/bbm007

Wilkinson, D. J. (2009). Stochastic modelling for quantitative description of heterogeneous biological systems. *Nature Reviews. Genetics*, *10*(2), 122–133. https://doi.org/10.1038/nrg2509

Wilkinson, D. J. (2018, December 7). *Stochastic modelling for systems biology, third edition* (0th ed.). Chapman; Hall/CRC. https://doi.org/10.1201/9781351000918

Williamson, D., Goldstein, M., Allison, L., Blaker, A., Challenor, P., Jackson, L., & Yamazaki, K. (2013). History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble. *Climate Dynamics*, *41*(7), 1703–1729. https://doi.org/10.1007/s00382-013-1896-4

Wilson, I. J., Carling, P. J., Alston, C. L., Floros, V. I., Pyle, A., Hudson, G., Sallevelt, S. C. E. H., Lamperti, C., Carelli, V., Bindoff, L. A., Samuels, D. C., Wonnapinij, P., Zeviani, M., Taylor, R. W., Smeets, H. J. M., Horvath, R., & Chinnery, P. F. (2016). Mitochondrial DNA sequence characteristics modulate the size of the genetic bottleneck. *Human Molecular Genetics*, *25*(5), 1031–1041. https://doi.org/10.1093/hmg/ddv626

Yamashita, S., Nishino, I., Nonaka, I., & Goto, Y.-I. (2008). Genotype and phenotype analyses in 136 patients with single large-scale mitochondrial DNA deletions. *Journal of Human Genetics*, *53*(7), 598. https://doi.org/10.1007/s10038-008-0289-8

Yang, D., Oyaizu, Y., Oyaizu, H., Olsen, G. J., & Woese, C. R. (1985). Mitochondrial origins. *Proceedings of the National Academy of Sciences of the United States of America*, *82*(13), 4443–4447. https://doi.org/10.1073/pnas.82.13.4443

Yau, C., & Campbell, K. (2019). Bayesian statistical learning for big data biology. *Biophysical Reviews*, *11*(1), 95–102. https://doi.org/10.1007/s12551-019-00499-1

Yi, J. S., Holbrook, B. C., Michalek, R. D., Laniewski, N. G., & Grayson, J. M. (2006). Electron transport complex i is required for CD8+ t cell function1. *The Journal of Immunology*, *177*(2), 852–862. https://doi.org/10.4049/jimmunol.177.2.852

Yoneda, M., Tanno, Y., Horai, S., Ozawa, T., Miyatake, T., & Tsuji, S. (1990). A common mitochondrial DNA mutation in the t-RNA(lys) of patients with myoclonus epilepsy associated with ragged-red fibers. *Biochemistry International*, *21*(5), 789–796.

Young, M. J., & Copeland, W. C. (2016). Human mitochondrial DNA replication machinery and disease. *Current Opinion in Genetics & Development*, *38*, 52–62. https://doi.org/10.1016/j.gde.2016.03.005

Youngblood, B., Hale, J. S., Kissick, H. T., Ahn, E., Xu, X., Wieland, A., Araki, K., West, E. E., Ghoneim, H. E., Fan, Y., Dogra, P., Davis, C. W., Konieczny, B. T., Antia, R., Cheng, X., & Ahmed, R. (2017). Effector CD8 t cells dedifferentiate into long-lived memory cells [Publisher: Nature Publishing Group]. *Nature*, *552*(7685), 404–409. https://doi.org/10.1038/nature25144

Zheng, L., Zhou, M., Guo, Z., Lu, H., Qian, L., Dai, H., Qiu, J., Yakubovskaya, E., Bogenhagen, D. F., Demple, B., & Shen, B. (2008). Human DNA2 is a mitochondrial nuclease/helicase for efficient processing of DNA replication and repair intermediates. *Molecular Cell*, *32*(3), 325–336. https://doi.org/10.1016/j.molcel.2008.09.024

Zhu, J., Vinothkumar, K. R., & Hirst, J. (2016). Structure of mammalian respiratory complex i - PubMed. *Nature*. https://doi.org/10.1038/nature19095