

Imperial College London
Department of Mathematics

Stochastic modelling and inference for evolution in ageing and infectious diseases

Ferdinando Insalata
December 2022

Supervised by Prof. Nick S. Jones

Submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy in Mathematics

Declaration of originality

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

Copyright

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

To Angelina Moscato,

Abstract

Counting processes are an important class of stochastic processes with numerous interdisciplinary applications, including to evolving biological systems. Counting processes can be the basis for forward generative models, which can help develop a mechanistic understanding of a phenomenon, or for inferential methods.

In this thesis we develop a bottom-up microscopic model of a phenomenon in mitochondrial biology and ageing. Mitochondria are organelles that possess their own DNA, involved in crucial physiological functions. A typical eukaryotic cell hosts a population of mitochondrial DNA molecules, where mutations can expand and cause dysfunctions. With age, skeletal muscle suffers a reduction in strength and functionality; in mammals, this has been connected to the clonal expansion of mitochondrial deletion mutations. The mechanism driving this phenomenon remains poorly understood despite intense research.

We develop a stochastic population dynamics model corresponding to a novel evolutionary mechanism, termed stochastic survival of the densest, and we show that it can account for the expansion of mitochondrial mutations in skeletal muscle through a literature-parameterised model. We predict that a species can invade a system in a wave-like fashion, without having an explicit replicative advantage and even if preferentially eliminated. We establish that this mechanism is driven by the combined effect of stochasticity, differences in carrying capacity (or density) and spatial structure.

Part of the work for this thesis coincided with the COVID-19 pandemic. We look at another application of counting processes, modelling data collected within the SIREN study, part of the national response to the pandemic. We estimate parameters of counting processes modelling SARS-CoV-2 infection events, that correspond to the reduction in rate of infection associated with COVID-19 vaccination (vaccine effectiveness), with a previous infection, or both (hybrid protection). We highlight short-term vaccine effectiveness that subsequently wanes, and confirm the immune escape of the Omicron variant.

Contents

Abstract	iv
1 Introduction	1
1.1 Mitochondria in health and disease	1
1.1.1 The powerhouse of the cell, with handbook	1
1.1.2 Mitochondrial DNA mutations and pathology	2
1.1.3 Expansion of mtDNA mutations in skeletal muscle	4
1.2 Mathematical modelling in mitochondrial biology	5
1.2.1 Stochastic processes in science and biology	5
1.2.2 Existing accounts of the clonal expansion of deletions in skeletal muscles	7
1.3 Stochastic processes and immunity from COVID-19 infection	10
1.3.1 The SIREN study and the importance of real-world data	10
1.3.2 Survival analysis, practical tools with a rich theoretical basis	11
1.4 Thesis outline	13
2 Stochastic Survival of the Densest	17
2.1 Introduction	18

2.2	The building block: single-unit system	19
2.2.1	Deterministic (ODE) treatment	20
2.2.2	Carrying capacity of the whole system and replication rate increase with heteroplasmy	22
2.2.3	A stochastic treatment exhibits selection reversal	24
2.2.4	Derivation of the effective SDE for the CM-constrained system	29
2.3	Two-unit system and noise-driven increase in mean proportion of mutants . . .	35
2.4	Noise-driven travelling waves of mutants in a chain of units	39
2.5	A noise-driven wave equation from the single-unit model.	43
2.6	Other feedbacks and modelling frameworks produce SSD	45
2.7	Implications for the evolution of altruism	47
2.8	Further links to the wider evolutionary literature	50
2.9	Discussion	52
3	Stochastic survival of the densest and ageing of skeletal muscle	54
3.1	Introduction	55
3.2	SSD matches clonal expansion in skeletal muscle fibres	56
3.3	Linking ageing, wave speed and copy number	60
3.4	Revisited estimates of mutant loads in skeletal muscle fibres	62
3.5	SSD can account for mutational load in short-lived mammals	64
3.6	Discussion	68

4	Estimating COVID-19 vaccine effectiveness	70
4.1	Introduction	71
4.2	Overview of survival analysis	72
4.2.1	Censoring is a defining aspect of survival analysis	72
4.2.2	Mathematical foundations	74
4.2.3	Non-parametric estimate of the survival function	75
4.2.4	Parametric survival models and multivariate survival analysis	77
4.2.5	Likelihood and censoring	79
4.2.6	Semiparametric Approach: Cox proportional hazards model	80
4.2.7	Assumptions and extensions of the Cox model	81
4.3	Proportional hazards models and the Poisson Point process.	83
4.3.1	Example: Cox regression on survival times generated from a Poisson process	84
4.4	Protection associated with primary course and/or primary infection	86
4.4.1	Background	86
4.4.2	Data collection and event of interest	87
4.4.3	Stratified Cox model with time-varying predictors	89
4.4.4	Results	92
4.4.5	Assessing the impact of potential biases	97
4.4.6	Discussion, strengths and limitations	100
4.5	Protection associated with booster and/or primary infection	102

4.5.1	Background	102
4.5.2	Data and modelling	103
4.5.3	Results	106
4.5.4	Discussion, strengths and limitations	112
5	Conclusions and future work	115
5.1	Stochastic survival of the densest as a new paradigm	115
5.2	Further characterisation of skeletal muscle ageing	118
5.3	Modelling further epidemiological and serological data from SIREN	121
5.4	Tools to tackle societal challenges	124
	Bibliography	124
A	Appendix to Chapter 2	158
A.1	Phenomenological formula for wave speed	158
A.2	Parameter values for simulations shown in the chapter	159
A.2.1	Fig. 2.1	160
A.2.2	Fig. 2.2	161
A.2.3	Fig. 2.3	162
A.2.4	Fig. 2.4	162
A.2.5	Fig. 2.5	163
A.3	Statistical Methods	163
A.4	Code availability	164

B	Appendix to Chapter 3	165
B.1	Additional heteroplasmy profiles	165
B.2	Realistic estimates of parameter values and setup of numerical simulations . . .	166
B.3	A combination of mutant higher death and replication rate yields the same Fisher-Kolmogorov wave speed	170
B.4	Estimation of fixation probabilities in Fig. 3.3	171
B.5	Relationship between steepness and speed of a travelling wave	173
B.6	Rhesus monkeys data collection	175
B.7	Statistical methods	175
B.8	Code availability	177
C	Appendix to Chapter 4	178
C.1	Supplementary Tables for Section 4.4	179
C.1.1	Full model outputs	179
C.1.2	Sensitivity analyses for depletion-of-susceptible bias	182
C.1.3	Sensitivity analyses for delayed-entry bias	186
C.1.4	Sensitivity analyses to assess the effect of discarding participants without a reliable date of primary infection	189
C.2	Supplementary Tables, Section 4.5	193
C.2.1	Full model outputs	193
C.2.2	Sensitivity analyses for depletion-of-susceptible bias	202
C.2.3	Sensitivity analyses for delayed-entry bias	209

Chapter 1

Introduction

1.1 Mitochondria in health and disease

1.1.1 The powerhouse of the cell, with handbook

Mitochondria are cytoplasmic organelles present in most eukaryotic cells. Their main – but not sole – function is to synthesise the molecule adenosine triphosphate (ATP)¹. Most of the ATP produced by cells is synthesised in mitochondria, in particular in the electron transport chain which can be thought of as the last stage of cellular respiration. ATP can be considered the energy currency of the cell, being used to drive fundamental physiological processes that require energy. Through their production of ATP, mitochondria provide approximately 90% of the energy to eukaryotic organisms, including ourselves. For this reason, mitochondria are often referred to as the powerhouse of the cell, and can be thought of as factories that produce batteries (the ATP) powering cellular functions. Moreover, they are involved in several other important processes, such as calcium-signaling², apoptosis³ and others, which makes them central also to the physiology of non-aerobic cells.

Importantly, mitochondria have their own DNA – the mitochondrial DNA (mtDNA) – of approximately 17 kilo-bases. The mtDNA consists of 13 protein-coding genes, 2 ribosomal RNAs and 22 tRNAs required for the translations of these proteins⁴. This is accounted for by the

endosymbiotic theory^{5,6} of the origin of eukaryotic cells, which holds that mitochondria were free-living proteobacteria engulfed by another cell circa 2 billion years ago. It is speculated that the bacterium provided energy to the cell, making it more successful at surviving and replicating, and the cell provided protection from the environment. Each mitochondrion contains 2–10 copies of mtDNA, and each eukaryotic cell (with some exceptions, like red blood cells) contains up to hundreds of mitochondria. Therefore, a eukaryotic cell can contain $\simeq 10^2 - 10^3$ molecules of mtDNA, distinct from the nuclear DNA of the cell. MtDNA is packed into *nucleoids*, structures held together by DNA-binding proteins, each containing 1-2 mtDNA molecules.

DNA can acquire mutations, and the rate of mutation during mtDNA replication is 100 times larger than that of nuclear DNA⁷. Replication of mtDNA is important in the cell cycle, as mtDNA copy number – the number of mtDNA molecules in a cell – approximately doubles before cell division. However, mtDNA replication also happens in post-mitotic cells (such as neurons and skeletal muscle). Mutations in mtDNA are acquired not only upon replication error, but also because of damage inflicted by reactive oxygen species (ROS)^{8,9}, chemical compounds generated upon leakage of electrons from the electron transport chain, that can also damage the mitochondrial membrane. Moreover, mitochondrial mutations can be inherited and hence be present from birth. In most species, including humans, mitochondria from the male gamete are degraded after conception, so that the embryo derives the entire mitochondrial genome from the oocyte. This is referred to as maternal inheritance of mtDNA. *Mitophagy* is the mechanism whereby cells remove malfunctioning (damaged or mutation-laden) mitochondria. A final important aspect is that mitochondria form a highly dynamic network in the cell, undergoing frequent fission and fusion. Research on the function of mitochondrial networks is ongoing^{10,11}.

1.1.2 Mitochondrial DNA mutations and pathology

Mutations in mtDNA can affect the functioning of mitochondria, and in turn of the cell. Given the large number of mtDNA molecules in a typical cell, when a mutation occurs, the mtDNA that carries it will coexist with the remaining wildtype molecules. *Heteroplasmy* is the presence

of more than one species of mtDNA in a single tissue, cell or mitochondrion, in contrast to *homoplasmy* where all copies of mtDNA are identical (wildtype or mutated). A large number of wildtype mtDNA can keep the cell working normally if the number of mutants is small. A general feature of mitochondrial physiology is the *threshold effect*: when the proportion of mutant mtDNA exceeds a critical value, cellular function and health is impaired^{12–18}. The threshold is specific to mutations and health conditions, but typical values are 60–90%. Moreover, co-existing mutations have been associated with adverse health effects and genetic instability in a mouse model, even if not pathological at homoplasmy¹⁹. Clearly, how the mutant load changes over time (in a cell or tissue) and the mechanism behind these variations are questions of great clinical relevance.

Given the key tasks fulfilled by mitochondria, it is not surprising that their malfunction can have serious consequences for cellular and organismal health. Mitochondrial diseases are genetic conditions stemming from inherited or *de novo* mutations in mtDNA or nuclear DNA which disrupt the functions of mitochondrially-encoded proteins. Some of the most studied mitochondrial diseases are Leber’s Hereditary Optic Neuropathy, Leigh’s Syndrome and Pearson Syndrome. It is estimated that mitochondrial diseases affect roughly 1 in 4300 adult humans in the world²⁰. No cure is available, diagnosis is often made difficult by inter-individual phenotypical variability and the aetiopathogenesis is often debated²¹. It is estimated that in the UK there are around 150 births per year from women who might pass mitochondrial diseases to their offspring²⁰. In late 2016, the UK Human Fertilisation and Embryology Authority licensed a mitochondrial replacement therapy, also called “mitochondrial transplant”²². In this procedure, the nucleus of woman’s mtDNA-mutation-heavy oocyte is inserted into a donor’s enucleated oocyte containing healthier mitochondria. This attracted considerable media attention, with some commentators speaking of “three-parent babies”. This procedure is ethically controversial^{23–26} and long-term risks have not been ruled out^{27–29}. Its approval speaks to the gravity of these health conditions, that can lead to the early death of the newborns³⁰.

Mitochondria have also been linked to diseases such as Parkinson’s disease³¹, cancer³², diabetes³³ and Alzheimer’s disease³⁴. Moreover, these organelles have been widely implicated in ageing^{35–37}. As the proportion of elderly and people affected by these conditions is increas-

ing in our modern societies, understanding the mechanism underlying their development is of paramount importance to inform interventions aimed at increasing healthy lifespan³⁸.

1.1.3 Expansion of mtDNA mutations in skeletal muscle

With age, our skeletal muscles accumulate damage, leading to a reduction in muscle mass and strength through atrophy and breakage of the fibres, termed sarcopenia³⁹, that has a knock-on effect on the health of elderly people, contributing to mounting pressures on healthcare systems. Sarcopenia has been linked, in human and other mammals, to the *clonal expansion of mtDNA deletions in skeletal muscle fibres*^{17,40–43}. Skeletal muscle fibres are long, multinucleated single cells containing up to thousands of nuclei distributed along their length. Each nucleus is surrounded by densely packed mitochondria. These fibres are postmitotic cells, no longer replicating, therefore the replication of mitochondria is independent of the cell cycle. Given the high energy demand due to muscular function, the number of mitochondria is higher than most other cell types and has been estimated at approximately 3500 per nucleus in healthy human skeletal muscle fibres⁴⁴. Deletion mutations are mtDNA molecules that lack a part of the DNA, typically thousands of bases. The expansion is *clonal*: damaged zones of the muscles are taken over by a single deletion, as consistently reported in single-cell studies in rats^{17,45}, rhesus monkeys^{46,47} and humans⁴⁸. The accumulation of mutants has distinctive spatial features¹⁶, with regions completely taken over by mutants flanked by transition zones to low or zero mutant loads. The fact that different affected fibres generally have different deletions suggests that founder mutations occur during the lifetime of an individual and then expand clonally⁴⁸, also moving along the fibre^{49,50}.

Other types of genetic variants are point mutations, where a single nucleotide base is changed. The expansion of point mutations is more common in mitotic tissue, and models based on random genetic drift seem to account satisfactorily for their expansion⁵¹. On the contrary, mtDNA deletions seem to expand preferentially in post-mitotic tissues, not only skeletal muscle fibres but also neurons from the substantia nigra, hippocampus, striatum and spinal cord⁵¹. Deletions tend to be more harmful to mitochondrial functions, given the large scale of the

alteration to the mtDNA. The mechanism behind the expansion of deletions remains poorly understood, which makes it difficult to propose therapies to delay it or prevent it.

One of the main contributions of this thesis is introducing a model and mechanism that account for experimental observations and important aspects of the clonal expansion of deletions. We will discuss existing mathematical models and introduce ours in the next section.

1.2 Mathematical modelling in mitochondrial biology

1.2.1 Stochastic processes in science and biology

It is hard to find a scientific discipline where probability and stochasticity are not relevant. One of the most successful and long-standing theories of Physics is thermodynamics. Statistical Mechanics can be seen as the mathematical formulation of thermodynamics⁵², and is based on probabilistic and stochastic modelling. Reinforcement Learning, based on stochastic control theory, is considered by some a potential path to artificial general intelligence⁵³. In the financial industry, stochastic models, for example of asset and option prices, have become widespread⁵⁴. Some of the most immediately policy-relevant mathematical research has concerned the spread and impact of COVID-19 under different levels of restrictions. In March 2020 predictions based on a stochastic model⁵⁵ played a significant role in the UK government's decision to impose the first nation-wide lockdown.

The above list is far from being exhaustive, and one could wonder what is behind this wide applicability of – and need for – stochastic modelling. Informally, a stochastic process can be thought of as a collection of random variables, each associated with an element of a (discrete or continuous) set⁵⁶. Many systems occurring in nature and society are subject to a myriad of influences and interactions, that can be unknown or hard to quantify. The unknown or uncertain elements of a system can be formalised mathematically via random variables, allowing modellers to account for the uncertainty. Biological systems burst with complexity, processes and interactions that are hard to observe and quantify, and events that cannot be perfectly predicted⁵⁷.

For instance, eukaryotic cells are highly dynamic and incredibly crowded environments, where a plethora of random processes occur, such as gene expression⁵⁸, diffusion of chemicals⁵⁹ and DNA mutations⁶⁰. In ecology, assuming that the number of offspring and times of reproduction of each individual of a population are fixed and deterministic would be unrealistic⁶¹. This is an example of *intrinsic* or *demographic* noise, i.e. uncertainty about events in the phenomenon under study^{62,63}. The interactions among species or between individuals and the ecosystems⁶⁴, or the amount of food available,⁶⁵ are unlikely to be precisely measurable by a single value of a parameter. The latter is an example of *extrinsic* noise or environmental stochasticity, namely uncertainty about interactions in the system or the environment⁶⁶. In a population of billions of cancer cells^{67,68}, or of thousands of mitochondria, there will be intrinsically stochastic sources of heterogeneity⁶⁹.

The mitochondrial threshold effect mentioned in the previous section is an interesting example of the importance of a stochastic approach. Say that a population of cells (e.g. a tissue) has a Gaussian across-cell distribution of the proportion of mtDNA mutants. The average value of this distribution could stay constant in time and be below the threshold value for pathology. However, the variance of the distribution might increase over time, so that an increasing number of cells has a mutant proportion above the threshold. Therefore, having the time evolution of a distribution has been crucial in obtaining a mathematical characterisation of the threshold effect⁷⁰.

Realistic models of biological systems, therefore, depend on appropriately accounting for the multiple sources of uncertainty, noise and heterogeneity that occur at different levels in the system. However, stochastic modelling is not only a toolbox to account for uncertainty and “propagate” it to predictions, that are typically time-varying distributions (i.e. indexed by time) of the value of a quantity. Genetic drift, the change in allele frequency over time due to chance that can have dramatic effect on a population, has been central to mathematical evolutionary theory since its inception^{71,72}. Stochastic models, moreover, can also exhibit behaviours that are qualitatively new and surprising – in that they are not observed in a deterministic setting – but have explanatory power. Some examples are noise-induced pattern formation⁷³ and cycling of species⁷⁴.

Another puzzling effect of this kind is stochastic selection reversal⁷⁵: in the competition for resources in a system, a species that deterministically would go extinct can, instead, prevail in the presence of stochasticity. This effect has been linked to one of the most debated questions in evolutionary biology, the evolution and spread of altruistic traits^{75–78}. Organisms that bring a benefit to others in the system, for example via public good production⁷⁵ or by a more economical use of common resources⁷⁶, may pay a cost for it, that can be encoded in a higher degradation rate. This corresponds to a definition of biological altruism^{79,80}. In a deterministic setting, the cost of altruism leads to extinction; stochasticity can offset this cost and drive the proliferation and expansion of altruistic species. We will see that this counterintuitive stochastic effect, initially identified *in nuce* in the mitochondrial literature^{81–83}, has deep connections with our account of the expansion of mitochondrial deletions.

1.2.2 Existing accounts of the clonal expansion of deletions in skeletal muscles

The temporal dynamics of the mitochondrial population in a cell can be modelled as a stochastic process, with stochastic replication and degradation of mtDNA molecules and random occurrence of mutations. If one does not model the occurrence of *de novo* mutation, a birth-death process⁸⁴ applies: each molecule can replicate or be degraded with a given rate, that can be constant or a function of time and/or population size. Models built upon this basic framework have been used to study the temporal dynamics of the population of mtDNA during development^{85,86}, in ageing^{81,83,87} and in the study of the dynamics of the mitochondrial network^{88–91}.

As we mentioned, the expansion of deleted mtDNA in skeletal muscle fibre is a phenomenon that has baffled scientists for decades, despite extensive research. Numerous mechanistic explanations have been proposed, some backed by mathematical modelling. All have attractive features but none has been widely accepted as a mechanism that can satisfactorily account for observations.

Survival of the slowest: mitochondrial mutants, less efficient at cellular respiration, generate

less of the membrane-damaging ROS and therefore might have a lower degradation rate. They would therefore have a higher net growth rate⁹². However, evidence shows that dysfunctional mitochondria are often preferentially degraded⁹³, rather than conserved. Moreover, this mechanism works only if mitochondria are separate entities, each with its own membrane, but they form a network undergoing fusion and fission⁹⁴.

Survival of the tiny: Deletion mutants have a smaller genome size, which can result in a shorter replication time leading to higher replication and net growth rates^{95,96}. The process of replication of a strand of mtDNA with a deletion can be up to 4.6 times faster than that of wild type. This seems to have an appreciable effect in fast-proliferating cells⁹⁷ and those recovering their mtDNA content after a heavy depletion⁹⁸. However, the time required for replication of an mtDNA molecule (1-2 h) is only a tiny fraction ($< 1\%$) of the typical time between replications (1-3 weeks) in post-mitotic cells. This makes a size-based replication advantage less important⁹⁹. Simulations show that a size-based advantage can account for the accumulation of deletions in skeletal muscle only for very long-lived species, and not for mice and rats, where comparable mutant loads accumulate in a short life-span¹⁰⁰. Moreover, this mechanism would predict a correlation between the size of the deletion and the speed of mutant accumulation, but there is evidence¹⁰¹ that small and large mtDNA deletions accumulate similarly.

Survival of the sick: defective mitochondria produce more ROS. It was proposed that ROS signal a deficiency of respiratory chain complexes back to the local mtDNA, which replicates in response, activating mitochondrial biogenesis¹⁰². In this way, mitochondrial mutants that produce more ROS would be selectively amplified. The main problem with this hypothesis is that it does not explain the predominance of deletions over point mutations (substitution in a single DNA basis)¹⁰³. Moreover, it predicts a central involvement of ROS in the ageing process, which has been questioned recently¹⁰⁴.

According to the *vicious cycle* hypothesis, mitochondrial mutants expand because, being defective, they generate more ROS which, in turn, cause more mutations¹⁰⁵. This account is not considered satisfactory because it implies that one should see many different mtDNA mutants in a single fibre, quite the contrary of what has been consistently observed in single-cell

studies^{16,17,103}.

More recently, it has been proposed that deleted mtDNA has a higher replication rate because most deletions remove genes that are responsible for down-regulating replication^{100,103}. The authors also suggest a subset of genes potentially involved in this feedback. Through simulations it has been shown that this mechanism can reproduce the observed mutant loads for a diverse lifespans. However, further examination of mtDNA deletion databases demonstrated that no one gene is always deleted and indeed deletions occur that do not remove any of the suggested genes^{51,106}. Moreover, this model is applied to a well-mixed population of mtDNA, without taking into account the spatial structure of muscle fibres. Finally, important parameters such as the magnitude of the replicative advantage and the *de novo* mutation rate are tuned to have simulations reproduce experimental observations, and not derived by independent data.

The above accounts all assume a mechanism whereby mtDNA deletions have a higher growth rate or are amplified in some way. However, no definitive explanation of this assumption is available and the evidence in post-mitotic tissue, such as skeletal muscle, is mixed¹⁰¹. Furthermore, there is evidence that mtDNA deletions are preferentially eliminated^{88,93,107–111}, which makes the phenomenon more puzzling. An influential set of studies using stochastic modelling^{81–83} argue that the clonal expansion instead takes place by *random drift* alone: by random fluctuations a macroscopic percentage of cells exceed the threshold of mutation level associated with pathology. Elson et al. have simulated⁸³ a mathematical model for a lifespan of 120 years and for biologically plausible parameters, obtaining good agreement with experimental data in aged tissues. The challenge with this account is twofold. First, these models do not reproduce the *clonal* expansion of a single mtDNA variant with a mutation, but an accumulation of *distinct* mutants through random drift and continuous addition of *de novo* mutations. Second, according to a more recent simulation study¹¹², in order to reproduce the mutant loads in short-lived animals such as rodents, very high *de novo* mutation rates are required. These high rates predict a myriad of different mtDNA mutants, which are not observed in experiments. Finally, the spatial features of the phenomenon are not considered. In this thesis we will look more in detail at this important series of studies.

A major contribution of our work is developing and testing a novel stochastic model of the expansion of mtDNA deletions in skeletal muscle fibres, without assuming a replicative advantage for mutants and without tuning free parameters. We will use a continuous-time stochastic process, specifically a Poisson point process, simulated via the Gillespie algorithm. Our model takes into account the distinctive spatial characteristics of these cells and recapitulates numerous experimentally observed features of the phenomenon. We will see that our model exhibits stochastic selection reversal and can be interpreted as a model of the spread of altruistic traits.

1.3 Stochastic processes and immunity from COVID-19 infection

1.3.1 The SIREN study and the importance of real-world data

With the outbreak of the COVID-19 pandemic, my research took another direction, keeping the underlying theme of the application of stochastic modelling to biomedical research. The opportunity arose to work on urgent and crucial questions that contributed to inform national public health policies, at the same time exploring novel scientific areas and acquiring complementary skills. Since July 2021, I have been responsible for the statistical modelling of data collected within an ongoing nation-wide study called SIREN (**S**arscov2 **I**mmunity **R**einfection **E**valuation**N**), part of the national response to the COVID-19 pandemic. The primary aim of the study, which started in June 2020, was to estimate the degree of immunity to a COVID-19 reinfection associated with a previous infection. When vaccines were introduced in December 2020, the scope of the SIREN study was extended to assessing protection associated with vaccination.

SIREN is a *prospective* cohort study. Prospective means that participants start being followed by researchers before the occurrence of the outcome of interest (in this case COVID-19 infection). This minimises the chance of selection bias, which can lead to a sample that is not representative of the population. Moreover, the follow-up events are at regular intervals, which

reduces the impact of recall error. The participants of the study are staff who work in hospitals from all over the UK; this includes not only clinical staff but also workers who do not have close contact with patients, such as cleaners, receptionists and office-based staff.

Importantly, SIREN gathers *real-world* observational data, collected while the study participants conduct their lives. Although vaccines are approved after clinical trials show safety and evidence of protection, real-world evidence after deployment remains crucial. Initial trials are usually funded by the vaccine manufacturers^{113–116}, hence independent studies and analysis are needed. Trials are conducted under strict research conditions, and differences from real world studies in the estimates of protection have been observed for COVID-19 vaccines. For instance, real-world data from Brazil led to an estimate of 36.8% for the protection against COVID-19 symptoms associated with a full course of the CoronaVac vaccine¹¹⁷, while the estimate from a phase 3 trial in Turkey was 83.5%¹¹⁸.

It is important to understand how, and to what extent, COVID-19 vaccines can help manage the pandemic, and this cannot be achieved in the controlled setting of a clinical trial, especially if under lockdown. Larger samples are needed to assess the effectiveness of the vaccine against different severity of illness, especially for a disease like COVID-19 that rarely leads to very severe outcomes in the healthy population. It is also necessary to assess the effect of vaccines in populations typically underrepresented in clinical trials, such as ethnic minorities, and those living with underlying health conditions. Long-term monitoring of vaccine safety is also needed, again with an emphasis on comorbidities. Finally, estimating mid- and long-term protection to identify possible waning is vital to inform adjustments to the vaccination programmes and further policies to tackle an evolving pandemic.

1.3.2 Survival analysis, practical tools with a rich theoretical basis

The statistical modelling of the SIREN data was carried out using techniques of *survival analysis*, one of the oldest branches of Statistics that dates back to the beginning of the actuarial and demographic sciences in the 17th century. Survival analysis is a set of statistical techniques used when the outcome variable is the *time until an event of interest*, often called *survival time*.

Examples of events of interest are death of a clinical trial participant, failure of a machine, occurrence or recurrence of a health condition. Survival analysis is used in numerous disciplines. In engineering, survival analysis is called reliability theory or analysis, and an application could be analysing the time until a component of a mechanical system breaks down. In sociology, where the topic is also called event history analysis, one could analyse the age people get married in a given society. In economics, one could investigate the time it takes new graduates to find their first full-time job in a specific job market. A criminologist might use survival analysis for research on recidivism, and insurance companies use it to price their policies. Survival analysis is also heavily used in ageing biology, where interventions and therapies are studied in terms of their effect on lifespan¹¹⁹.

This non-exhaustive list of applications speaks to the importance of survival analysis in scientific research and society at large, an importance that has grown since its early developments. The name “survival analysis” originates from early applications to the analysis of mortality rates; an alternative and more general name is *time-to-event analysis*. In a clinical setting, typical questions amenable to time-to-event analysis are: “what is the proportion of a population of patients which will survive 5 years after diagnosis?”, “what is the rate of death of patients?”, “how do particular circumstances (treatment) or characteristics (age of patient) affect the probability of survival?”.

A popular class of survival analysis models is *proportional hazards* models, which can be seen as generalised linear models of the rate at which the event of interest occurs. We will fit proportional hazards models with the event of interest being SARS-CoV-2 infection and with covariates encoding whether a person has been vaccinated, previously infected and how long ago. This will essentially tell us if, and by how much, being vaccinated or infected affects the rate of future infections. My work with these statistical techniques was facilitated by the experience with stochastic modelling gained in the first part of my PhD. Indeed, survival analysis has deep connections with stochastic modelling and, specifically, counting processes^{120–122}, such as the Poisson process, that will form the basis of the simulations in Chapters 2 and 3. In Chapter 4 we will give an overview of survival analysis, hinting at these connections with conceptual arguments in Section 4.3. We will see that the coefficients of proportional hazards models

are identifiable as parameters of a counting process that models infection events, and give a concrete example with simulated data.

1.4 Thesis outline

In this thesis, we first develop a forward stochastic model based on the Poisson point process and show that it can account for the expansion of deletions in skeletal muscle fibres. We then apply statistical methods from survival analysis, rooted in the theory of counting processes, to estimate the degree of protection from SARS-CoV-2 infection associated with previous infection or vaccination.

In **Chapter 2** we formulate a spatially structured stochastic model that can be schematised as two species competing for resources in a chain of demes, with hopping of individuals between neighbouring demes. This model exhibits an intriguing effect that we call *stochastic survival of the densest* and we introduce as a novel selection mechanism, dependent on the combined effect of stochasticity, spatial structure and differences in carrying capacities – density – between the two species. We see how these elements give rise to a travelling wave of the denser species invading the system, without the need of a higher net growth rate and even in the presence of a lower growth rate. We provide a phenomenological formula detailing how the speed of this wave scales with the parameters of the models, noting analogies and differences with the Fisher-Kolmogorov wave speed.

In **Chapter 3** we argue that stochastic survival of the densest can account for the clonal expansion of mitochondrial deletion mutants in skeletal muscle fibres with age. We see how a model based on stochastic survival of the densest can quantitatively account for experimental observations, using parameter values estimated from independent experimental data.

In **Chapter 4** we discuss the statistical modelling of data collected in the SIREN study, part of the UK national response to the COVID-19 pandemic. We first give an overview of survival analysis, with an emphasis on the proportional hazards models, and more detail on the SIREN study. We then apply a flexible proportional hazards model (Cox model) to estimate immunity

from SARS-CoV-2 (Delta and Omicron variant) associated with vaccination, previous infection, and both. We discuss biases that can affect studies that use survival data in general and vaccine effectiveness studies in particular, arguing that they are unlikely to affect our analyses based on the design of the SIREN study and on the results of sensitivity analyses. Finally, we explore some of the connections between survival analysis and counting processes through conceptual arguments and applying inference to data simulated from a Poisson point process.

The material in Chapters 2 and 3 is published in *Proceedings of the National Academy of Sciences of the United States of America*¹²³. I was single first author, performed most of the research (computational, analytical and data analysis) and wrote and edited the manuscript. Additional work was performed by Hanne Hoitzing, who first observed stochastic survival of the densest and revisited estimates of mitochondrial mutant loads, and Juvid Aryaman, who pointed me to the right analytical techniques and gave feedback on drafts and on the visualisation of results. Additionally, I co-supervised two MSc-level projects based on this work (Cecilia Fruet and Jasmin Daley); the main results of one project will be briefly discussed in the Conclusions (Chapter 5).

As for Chapter 4, most of the material in Section 4.4 is published in *The New England Journal of Medicine*¹²⁴. I was joint first author of this paper. I performed data cleaning and pre-processing, developed the statistical models and interpreted results, wrote the methods and result sections and edited the whole manuscript. This work was a large collective effort, with two more joint first authors and numerous other authors who reviewed, collected and managed data, among many other tasks crucial to running a nation-wide epidemiological study such as SIREN. The material in Section 4.5 is part of a manuscript that is in preparation, where I will also be joint first author and my contribution has been analogous. Within SIREN I have authored another paper¹²⁵, that will be touched upon in the Conclusions. The plots and tables in Chapter 4 and the supplementary tables in Appendix C are joint work with the other two first authors of Ref. [124] (Sarah Foulkes and Victoria Hall): we discussed at length how to best present the results in each case.

During the course of my PhD, I have also authored three papers that apply stochastic modelling

to questions in Statistical Mechanics^{126–128} and one paper that applies physics-inspired neural network architectures to magnetic resonance imaging data¹²⁹. These are not discussed here.

Publications and contributions

Stochastic survival of the densest and mitochondrial DNA clonal expansion in ageing *Proceedings of the National Academy of Sciences of the United States of America*, 119(49), e2122073119 (2022).

<https://www.pnas.org/doi/10.1073/pnas.2122073119>

Ferdinando Insalata, Hanne Hoitzing, Juvid Aryaman, Nick S. Jones

Contributions: Coded the model and performed large-scale simulations, observed and characterised the noise-induced waves, performed the analytical calculations and statistical analyses, obtained new estimates of the *de novo* mutation rate, interpreted findings, visualised results, wrote and edited the paper and supplement, answered reviewers' comments over three rounds.

Appearance in thesis: Chapters 2 and 3

Protection against SARS-CoV-2 after Covid-19 Vaccination and Previous Infection *New England Journal of Medicine*, 386(13), 1207-1220 (2022).

<https://doi.org/10.1056/NEJMoa2118691>

Victoria Hall*, Sarah Foulkes*, **Ferdinando Insalata***, ..., Andre Charlett and Susan Hopkins, for the SIREN Study Group

Contributions: Shared first author, denoted by * symbol. Cleaned and pre-processed data, developed and coded statistical models, performed sensitivity analyses, interpreted the results, wrote statistical and results sections of the manuscript, edited all the manuscript, answered reviewers' comments with a focus on the statistical methodology.

Appearance in thesis: Chapter 4.

Effectiveness of BNT162b2 mRNA booster doses and hybrid-immunity in protecting against SARS-CoV-2 infections during the Delta and Omicron variant waves; the UK SIREN study September 2021 to February 2022 *In preparation*

Victoria Hall*, Sarah Foulkes*, **Ferdinando Insalata***, ..., Andre Charlett and Susan Hopkins, for the SIREN Study Group

Contributions: Shared first author, as denoted by * symbol. Cleaned and pre-processed data, developed and coded statistical models, performed sensitivity analyses, interpreted the results, wrote the technical and results sections and edited whole paper.

Appearance in thesis: Chapter 4.

Chapter 2

Stochastic Survival of the Densest

Summary

In this chapter we first introduce the phenomenon of the expansion of mitochondrial deletion mutations in skeletal muscle fibres. This phenomenon has puzzled scientists, including bio-mathematicians, for more than 30 years, in that no convincing account of experimental observations is available. This scientific challenge motivates us to introduce a class of mathematical models that exhibit – we argue – a novel mechanism of evolution, whereby a species can outcompete another thanks to the combined effect of spatial structure, stochasticity and a higher carrying capacity (or density). This mechanism, that we call *stochastic survival of the densest*, predicts that the denser species can take over the system through a wave-like expansion, without an explicitly larger growth rate and even if subject to preferential elimination. We contrast our research with the concurrent lines of work in evolutionary theory, before moving to the next chapter where we apply the novel mechanism to the expansion of deletions in skeletal muscle.

2.1 Introduction

As we noticed in the Introduction, the mechanism behind the clonal expansion of mitochondrial deletion mutants in skeletal muscle fibres remains unclear despite more than three decades of research^{81–83,92,103,112,130–133}. When in a system one species (here the mtDNA deletions) outcompetes another species (here the wildtype mtDNA), it is often concluded that the former has a replicative advantage. This has been the approach followed by numerous authors in trying to account for the expansion of mtDNA deletions, and we have reviewed influential studies. However, there is no definitive explanation of why mtDNA mutants would have a replicative advantage, and, furthermore, there is evidence that mtDNA deletions can be preferentially eliminated^{88,93,107–111}. Other studies have modelled the phenomenon without assigning a replicative advantage to mutants^{81–83}, arguing that random drift alone is sufficient to explain the expansion. At the end of Section 1.2.2 we explained why these models can potentially work only for very long-lived species¹¹². We notice that, importantly, most previous attempts have modelled muscle fibres as a single isolated and well-mixed (unstructured) population of mtDNA molecules, without taking into account their distinctive structure as long multinucleated cells.

This scientific challenge led us to formulate and analyse the model we introduce in this chapter. We develop a model where a species can outcompete another without having a replicative advantage and, surprisingly, even if preferentially eliminated from the system. Our model predicts that the species with higher carrying capacity can invade the system through a noise-driven travelling wave. We provide a phenomenological formula for the speed of this wave. We justify the counterintuitive predictions of our model with heuristic arguments, and, in simple or limiting cases, analytical calculations. We argue that the mechanism we unveil may be the driving force behind other counterintuitive phenomena, such as the evolution and spread of altruistic traits^{79,80}.

In its basic form, our model is a special case of the competitive Lotka-Volterra model¹³⁴. However, our main findings are robust to other modelling frameworks: for example they can be reproduced using discrete non-overlapping generations as in the Moran or Wright-Fisher models. Moreover, although we use a simple linear form of the replication rate, we verify that

our core predictions are robust to other functional forms.

In the following we will introduce and analyse our mathematical model without direct references to mitochondrial genetics. This underlines the simplicity and generality of the mechanism we observe. In the next chapter, we apply our model the expansion of mitochondrial deletion mutations in skeletal muscle fibres.

2.2 The building block: single-unit system

The fundamental building block of our model is a single deme (unit) hosting a well-mixed population composed of individuals from two species, evolving under a birth-death process. This constitutes the building block for spatially structured models explored in the next sections. In 2.3, we will introduce a minimal spatial structure coupling together two units, and in 2.5 we will look at a chain of units.

In our model, there are two species, generically denoted as wildtypes w and mutants m . The symbols w and m denote their numbers, which in general are functions of time and space. They may be individuals from two species, or molecules with and without a certain mutation. Every molecule is degraded at a constant rate μ and replicates at rate $\lambda(w, m)$ given by

$$\lambda(w, m) = \max[0, \mu + c(N_{ss} - w - \delta m)], \quad (2.1)$$

i.e., the larger between 0 and $\mu + c(N_{ss} - w - \delta m)$, with parameters c, N_{ss}, δ . The replication rate is scaled by the distance of a count, $w + \delta m$, of the current population size from a target effective population N_{ss} . The parameter δ sets the relative density of mutants, whose carrying capacity is $1/\delta$ that of wildtypes (N_{ss}/δ against N_{ss} , see 2.2.2). For $0 < \delta < 1$, which will always be the case in this work, mutants have a higher carrying capacity and can live at higher copy numbers in the system. Therefore *mutants are denser*, as encoded by $0 < \delta < 1$. The parameter c quantifies the strength of attraction to the steady state. The larger c , the more strongly the system is penalised when away from steady state ($w + \delta m \neq N_{ss}$), i.e., when the

terms in parentheses on the right-hand-side (RHS) in Eq. (2.1) is not equal to 0 and hence $\lambda \neq \mu$. In other words, a larger value of $c > 0$ will cause a stronger push toward the steady state.

Apart from transients that could arise for particular initialisations of the system, for all biologically relevant values of the parameters $\mu + c(N_{ss} - w - \delta m)$ is practically always positive. Therefore, for simplicity we will mainly refer to the simpler form of the replication rate

$$\lambda(w, m) = \mu + c(N_{ss} - w - \delta m). \quad (2.2)$$

Because of the linear relationship between the per capita replication rate λ and the copy number we name this replication rate *linear feedback control*. In the following we study deterministic and stochastic versions of the model defined by the above rates.

2.2.1 Deterministic (ODE) treatment

In a deterministic setting, the above growth rate give rise to the following system of ODEs:

$$\begin{aligned} \frac{dw}{dt} &= cw(N_{ss} - w - \delta m) \\ \frac{dm}{dt} &= cm(N_{ss} - w - \delta m). \end{aligned} \quad (2.3)$$

The analysis of this deterministic dynamical system is useful to understand the behaviour of stochastic version of the model. We limit our analysis to $w > 0, m > 0$ given the meaning of the variables as copy numbers. We start by noticing the existence of the trivial, unstable fixed point $(0, 0)$. We then notice a set of attractive fixed points which form the straight line of equation $w + \delta m = N_{ss}$ in the (m, w) plane. We refer to this as the center manifold (CM) or steady state ($\mu = \lambda$) line. The CM includes the configurations in which there is only a single species, namely the wildtype fixed point $(N_{ss}, 0)$ and the mutant fixed point $(0, N_{ss}/\delta)$. The *carrying capacity* of a species is the population size of this species at its fixed point, at which the species exists in isolation. We see that in our model the carrying capacities are N_{ss} for wildtypes and

N_{ss}/δ for mutants. Another way of obtaining the wildtype (mutant) carrying capacity is setting $\lambda(w, m) = \mu$ – equivalent to the condition of fixed point – and $m = 0$ ($w = 0$).

When $0 < \delta < 1$ – which is always the case in this work – the mutant carrying capacity is larger than that of the wildtype. Another interpretation of the condition $0 < \delta < 1$ is that mutants are less tightly controlled or sensed by the system, as a change in the number of mutants entails a smaller variation in $\lambda(w, m)$ than an equal change in the number of wildtypes. Hence, in our model mutants are the *denser* and *less tightly controlled* species. We define the *heteroplasmy* h as the proportion of mutants:

$$h = \frac{m}{m + w}. \quad (2.4)$$

This quantity is conserved in the dynamical system of Eq. (2.3), i.e.

$$\frac{dh}{dt} = 0, \quad (2.5)$$

which can be verified by direct calculation. We use “heteroplasmy” as we will later deploy this modelling framework for populations of mitochondrial DNA. The meaning of “proportion of mutants” is common in studies that apply mathematical modelling to mitochondrial biology^{88,135}.

Eq. (2.5) is equivalent to the ratio m/w being constant, meaning that any line passing through the origin is a constant-heteroplasmy line. Hence, a way to summarise the behaviour of the system is:

- If the system is in steady state, i.e., if $w + \delta m = N_{ss}$, the dynamics stop.
- If the system is not in steady state, it will move toward the CM along a line connecting the initial condition to the origin.

The same behaviour can be recovered from the full solution of the system, which is

$$\begin{aligned}
w(t) &= N_{ss} \frac{w_0}{(N_{ss} - w_0 - \delta m_0)e^{-cN_{ss}t} + w_0 + \delta m_0} \\
m(t) &= N_{ss} \frac{m_0}{(N_{ss} - w_0 - \delta m_0)e^{-cN_{ss}t} + w_0 + \delta m_0},
\end{aligned} \tag{2.6}$$

where $w_0 = w(0)$ and $m_0 = m(0)$ are the initial conditions. From Eq. (2.6) it follows that

$$w(t) + \delta m(t) = N_{ss} \frac{w_0 + \delta m_0}{(N_{ss} - w_0 - \delta m_0)e^{-cN_{ss}t} + w_0 + \delta m_0}. \tag{2.7}$$

For $t \rightarrow \infty$, as $e^{-cN_{ss}t} \rightarrow 0$, $w(t) + \delta m(t) \rightarrow N_{ss}$. This shows that the parameter c is connected to how fast the state of the system moves to the CM, justifying the interpretation of c as control strength.

In summary, in the deterministic single-unit model the system evolves toward the CM of equation $w + \delta m = N_{ss}$ with the condition $\frac{dh}{dt} = 0$. We remark on the difference in carrying capacity of the two species, being modulated by the parameter δ .

2.2.2 Carrying capacity of the whole system and replication rate increase with heteroplasmy

The global carrying capacity K of the system is the value of the total population $n = w + m$ at steady state. K can be expressed as a function of h . Since $m = hn$ and $w = (1 - h)n$ the steady state condition $w + \delta m = N_{ss}$ becomes

$$(1 - h)n + \delta hn = (1 - h)K + \delta hK = N_{ss}, \tag{2.8}$$

where the first equality holds because we have defined K as the value of n at steady state. This leads to

$$K \equiv K(h) = \frac{N_{ss}}{1 - (1 - \delta)h}. \tag{2.9}$$

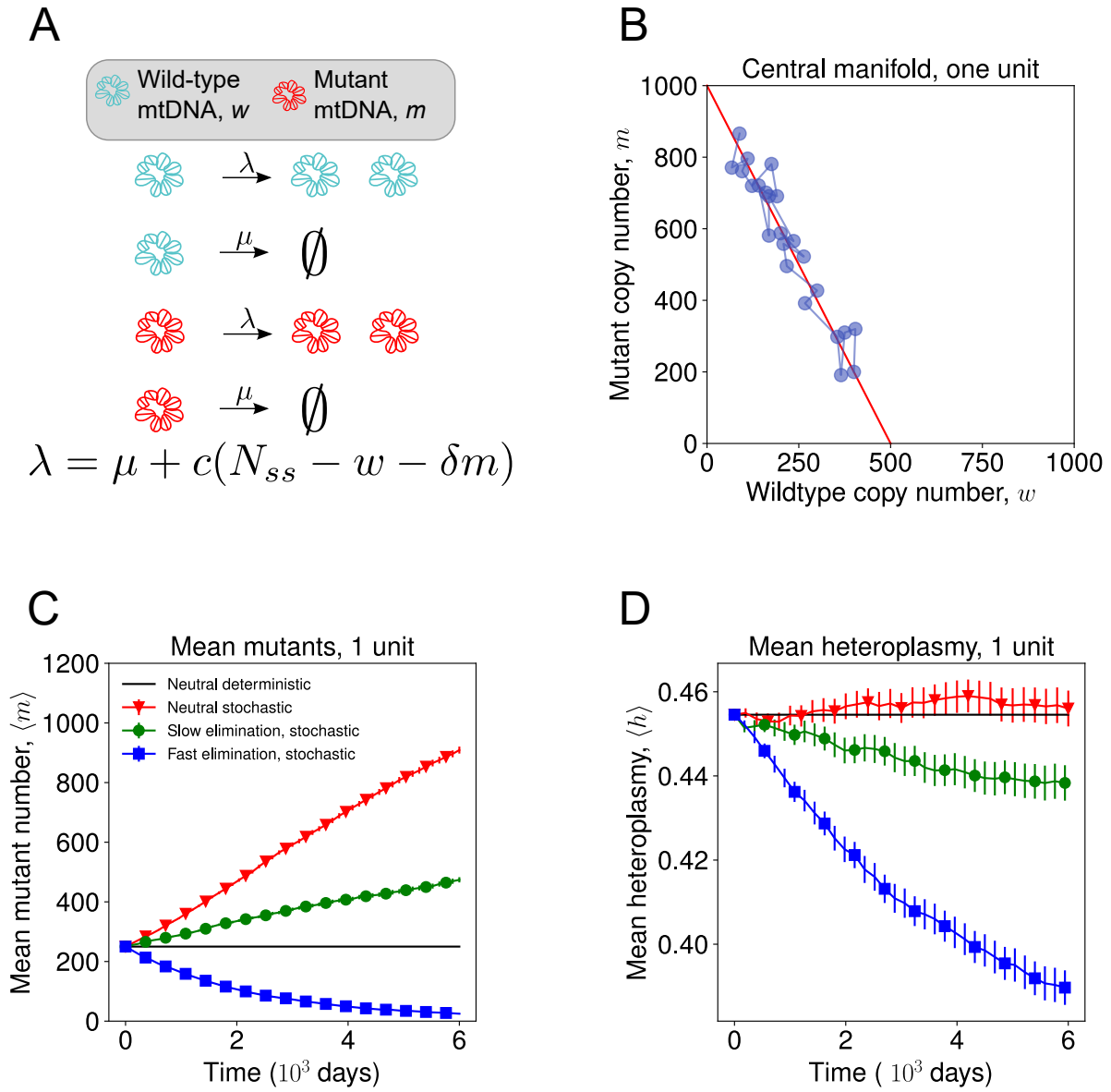


Figure 2.1: *In the single-unit system, the mean number of denser mutants increases stochastically even if they are preferentially degraded.* (A) In the single-unit system the possible events are birth and death of a wildtype (first two), birth and death of a mutant (third and fourth). Deaths happen at constant rate μ , births at rate $\lambda(w, m)$ (defined at the bottom). (B) The single-unit system fluctuates around the center manifold (CM) or steady state line (red line), and average proportion of mutant $\langle h \rangle$ stays constant. Example of a steady state line for $N_{ss} = 500$ and $\delta = 0.5$. (C) In the stochastic single-unit system, mean mutant copy number increases in the neutral model (red) and in presence of weak selective elimination of mutants (green). This is a consequence of higher carrying capacity for mutants ($\delta < 1$) and noise. In the corresponding deterministic neutral model, mutant copy number remains constant (black). Parameter values can be found in Section A.2. Error bars are standard error of the mean (SEM), averaged over 10^4 stochastic replicates. (D) In the single-unit system, heteroplasmy is constant in the neutral deterministic system, as well as mean heteroplasmy in the neutral stochastic system. Any selective elimination of mutants in the stochastic system causes a decrease in mean heteroplasmy (red and blue lines), while mean number of mutants can still increase (green line in panel B). Parameter values in Section A.2. Error bars are SEM, averaged over 10^4 stochastic replicates.

For $\delta < 1$, $K(h)$ is an increasing function of h : the higher the proportion of mutants, the larger the population that the system can sustain. When $h = 1$, i.e., mutants are in isolation, the mutant carrying capacity $\frac{N_{ss}}{\delta}$ is recovered. One way to interpret this is that mutants are using the resources of the system in a more economical way. The economical use of a limited resource is considered one of the earliest and simplest forms of altruism^{76,80}, which brings benefits to all the other individuals in the system regardless of their species.

Another way of seeing the benefit that mutants bring to all the other molecules in the system is in terms of enhanced replication rate. By expressing w, m in terms of h, n as above, Eq. (2.2) can be written as

$$\lambda(h, n) = \mu + c[N_{ss} - (1 + h - \delta h)n], \quad (2.10)$$

that shows that, for a given population size n , the common replication rate is an increasing function of h when $\delta < 1$. Notice that here, differently from Eq. (2.9), $n \neq K$ since we are not assuming that the system is in the steady state.

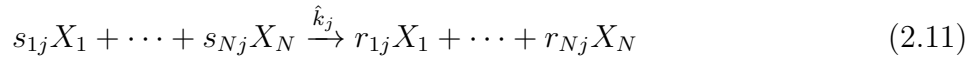
If mutants are subject to a higher degradation rate and if $\delta < 1$, they can be seen, in a very general sense, as agents that benefit all others in the population paying a cost (the higher degradation rate). This aligns with the definition of biological altruism^{77,80}, hence the mutants of our model can be considered an altruistic species. The connection of our work to the debate on the evolution of altruism is discussed in 2.7.

2.2.3 A stochastic treatment exhibits selection reversal

In the stochastic formulation of the model, the *per capita* degradation and replication rates μ and λ are interpreted as instantaneous probability (in an infinitesimal time interval) that each molecule is degraded or replicates. In this setting, w, m and h are random variables, and we ask questions about their probability distributions and moments.

Stochastic population dynamics models can be formalised as chemical reactions networks, consisting, in the terminology of chemical systems, of a set of reactants, reactions and products. Consider a general chemical system consisting of N distinct chemical species (X_i) interacting

via R chemical reactions, where the j^{th} reaction is of the form



where s_{ij} and r_{ij} are stoichiometric coefficients. We define $\hat{k}_j$ as the per molecule rate of the j^{th} reaction happening. A chemical reaction network can be described by the associated $N \times R$ stoichiometry matrix $S_{ij} = r_{ij} - s_{ij}$ and the set of rates $\hat{k}_j$. The reactants are the molecules of the two species. In our model the reactions are birth and death and the possible products are i) another molecule of the same species (birth), or ii) $\emptyset$ for the degradation of a molecule (death reaction).

The chemical reaction network for the single-unit model is



This is a network with $N = 2$ species and $R = 4$ reactions, with stoichiometry matrix given by

$$S = \begin{bmatrix} -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{bmatrix} \quad (2.13)$$

The global rates for wildtypes or mutants are found by multiplying the per molecule rates μ, λ by w or m . This is true only in the case $S_{ij} \leq 1$; more details for the general case can be found in Ref. [88]. The reaction network is summarised in Fig. 2.1A. The system will fluctuate around the deterministic CM of equation $w + \delta m = N_{ss}$. An example for a system with $N_{ss} = 500$ and $\delta = 0.5$, is plotted in Fig. 2.1B. These reactions and rates can be used to set up a Chemical Master Equation (CME), a system of coupled ODEs in $P(w, m, t)$, the probability that the system is in the state (w, m) at time t . In principle, solving the CME would give the probability that the system is found in any state at any given time. However,

given the non-linearity of the global rates, the CME cannot be solved exactly. One can explore the behaviour of the stochastic models by simulating the CME, which can be done through Gillespie's stochastic simulation algorithm^{136,137}. This algorithm is exact, in that each Gillespie simulation represents an exact sample from $P(w, m, t)$. The simulations allowed us to observe the increase in mean mutant copy number $\langle m \rangle$ for $\delta < 1$ (Fig. 2.1C, red line). Intuitively, while fluctuating around the steady state line the system on average drifts toward regions of higher m .

It is worthwhile having an analytical account of this effect, in order to establish how the increase in mutant copy number depends on the parameters of the model. We have obtained an effective description of the system, in the form of an approximate stochastic differential equation (SDE) that shows an increase in the number of mutants for $\delta < 1$. The steps are the following:

1. Apply the Kramers-Moyal expansion to the CME, to obtain a Fokker-Planck equation (a PDE for the probability distribution $P(w, m, t)$).
2. Convert the Fokker-Planck Equation into a system of two coupled SDEs in the variables $w(t)$ and $m(t)$.
3. Apply a stochastic dimensionality reduction procedure that exploits the fact that the system fluctuates around a center manifold (CM), to obtain a single SDE for $m(t)$ that shows a positive drift for $\delta < 1$.

The first two steps are standard^{138,139}. From the chemical reaction network in Eq. (2.12), we obtain the system of SDEs:

$$\begin{aligned} dw &= cw(N_{ss} - w - \delta m) dt + \{w[c(N_{ss} - w - \delta m) + 2\mu]\}^{1/2} dW_1 \\ dm &= cm(N_{ss} - w - \delta m) dt + \{m[c(N_{ss} - w - \delta m) + 2\mu]\}^{1/2} dW_2, \end{aligned} \quad (2.14)$$

where dW_1 and dW_2 are two i.i.d. Wiener increments, i.e., Gaussian noise with zero mean and variance dt (see also 2.32). The third step relies on a recently developed technique^{75,140,141}, which works specifically for systems fluctuating around a CM. In Section 2.2.4 we detail each

step, while here we only present and interpret the results. The final, effective SDE is

$$dm = \frac{2(1-\delta)\mu}{N_{ss}}m \left(1 - \frac{\delta m}{N_{ss}}\right) dt + \frac{1}{N_{ss}} [2m\mu(N_{ss} - \delta m)(N_{ss} + m - \delta m)]^{1/2} dW, \quad (2.15)$$

with dW a Wiener increment. The first term on the RHS is the drift, and it corresponds to a logistic growth with carrying capacity $\frac{N_{ss}}{\delta}$. When the drift is positive, since $\langle dW \rangle = 0$, the copy number increases on average. The drift is positive for $0 < \delta < 1$, i.e., when mutants have a higher carrying capacity or density, and $\delta m < N_{ss}$. Because of noise the final value of the mutant copy number will not be $\frac{N_{ss}}{\delta}$. The dynamics stop either when $m = 0$ or $m = \frac{N_{ss}}{\delta}$, the only values for which $dm = 0$. This approximation relies on the system fluctuating closely around the CM, which is true for large values of c (strong control). The parameter c is not present in Eq. (2.15) because this equation holds exactly for $c \rightarrow \infty$. Indeed, when simulating Eq. (2.15) using the Euler scheme, the agreement with the results of the exact Gillespie simulations is better for larger values of c (not shown). By applying Itô's formula (see Section 2.2.4) it is possible to find the SDE for h , which is

$$dh = \left[2\mu \frac{h(1-h)}{n} \right]^{\frac{1}{2}} dW, \quad (2.16)$$

where $n = w + m$. We notice in this equation the absence of a drift term, meaning that

$$\frac{d\langle h \rangle}{dt} = 0, \quad (2.17)$$

i.e., mean heteroplasmy is constant in the single-unit stochastic model, as shown in Fig. 2.1D, red line.

A heuristic way to understand why the mean number of mutants $\langle m \rangle$ increases and mean heteroplasmy $\langle h \rangle$ is constant is the following. The per capita rates in Eq. (2.12) are linear, hence the system is a stochastic Lotka-Volterra model, for which each individual has the same chance of generating a lineage that will take over the whole system¹⁴². Hence, this probability must be $1/(w_0 + w_0)$. Since there are m_0 mutants, the probability that a mutant will take over the whole population is $m_0/(w_0 + w_0) = h_0$: the mutant fixation probability coincides with

the initial fraction of mutants. At mutant fixation, the steady state mutant population is $\frac{N_{ss}}{\delta}$ and the mutant fraction is (obviously) 1. Based on this, one can conclude that, for $t \rightarrow \infty$, $\langle m \rangle \rightarrow h_0 \frac{N_{ss}}{\delta}$ and $\langle h \rangle \rightarrow h_0 \cdot 1 = h_0$. Taking the limit $t \rightarrow \infty$ means that we wait long enough for fixation to happen and for the population to reach the steady state. Recall that this differs from the deterministic case, in which there is a locus of fixed points of equation $w + \delta m = N_{ss}$ (the CM) and the system equilibrates at the point of the CM for which heteroplasmy is h_0 . For $0 < \delta < 1$, $h_0 \frac{N_{ss}}{\delta} > m_0$ always, hence $\langle m \rangle$ will increase to reach its final value. Conversely, the final value of $\langle h \rangle$ coincides with the initial value h_0 , giving intuition as to why $\langle h \rangle$ is constant. It is also illuminating to consider the case in which $\delta \approx 0$, namely when the mutants are scarcely controlled. When a wildtype dies at steady state this increases the birth rate. Either a mutant or a wildtype will be the next birth, but if a mutant is born this leaves the elevated birth rate (caused by the degradation of the wildtype) almost unchanged, encouraging further possible mutant births. Interestingly, even for $\delta \approx 0$ and in the extreme, unrealistic case of infinite mutant density $\delta = 0$, $\langle h \rangle$ stays constant. Indeed, Eq. (2.17) and the above heuristic argument are valid for any value of δ .

Importantly, the stochastic system can exhibit an increase in $\langle m \rangle$ even if we introduce preferential elimination of mutants (Fig. 2.1C, green line). Mathematically, this is described by replacing the second reaction in Eq. (2.12) by



with $\epsilon > 0$, leaving all other reactions unchanged. In the next section we will see that, through the techniques in Refs. [75, 140, 141], it is possible to derive an SDE for mutant copy number that is valid for $\epsilon \ll \mu$:

$$dm = \left[\frac{2(1-\delta)\mu}{N_{ss}} - \epsilon \right] m \left(1 - \frac{\delta m}{N_{ss}} \right) dt + \frac{1}{N_{ss}} [2m\mu(N_{ss} - \delta m)(N_{ss} + m - \delta m)]^{1/2} dW. \quad (2.19)$$

The drift in m is now positive for $\delta m < N_{ss}$ and $\epsilon < \frac{2(1-\delta)\mu}{N_{ss}}$. In this case, the combined effect of stochasticity and differences in carrying capacity (density) produces a stochastic reversal of the direction of deterministic selection⁷⁵. Indeed, in the deterministic case, for any $\epsilon > 0$

and for any initial condition with non-zero w_0 the system tends to the point $(N_{ss}, 0)$, the only stable fixed point, corresponding to wildtype fixation (mutant extinction). Conversely, in the stochastic case, there is a critical $\epsilon_M = \frac{2(1-\delta)\mu}{N_{ss}}$ such that, for $\epsilon < \epsilon_M$, $\langle m \rangle$ still increases, meaning that mutant fixation is more likely than wildtype. Inspection of the expression for ϵ_M , interpretable as the largest amount of additional degradation rate that still allows mutant to expand, reveals the following interesting points:

- ϵ_M increases with μ , meaning that the noisier the system, the higher the differential degradation rate that mutants can tolerate and still expand.
- For $0 < \delta < 1$, ϵ_M is a decreasing function of δ . Hence, the higher the difference in density, the higher the degradation rate that mutants can overcome.
- ϵ_M is a decreasing function of N_{ss} . In particular, $\epsilon_M \rightarrow 0$ when $N_{ss} \rightarrow \infty$. This is common to stochastic effects, whose intensities decrease with the system size, vanishing in the deterministic limit for which the system size tends to infinity.

2.2.4 Derivation of the effective SDE for the CM-constrained system

We now derive the effective SDE in the single-unit system Eq. (2.15), proving that it is an effective, approximate description of the system defined by the chemical reaction network in Eq. (2.12) with the replication rate Eq. (2.2). We provide details for each of the steps listed in Section 2.2. The approach used here is adapted from standard texts^{138,139} and from Refs. [140, 141]. We strive to provide a self-contained treatment, hoping to make our derivation and the set of techniques accessible to a wider readership. Our exposition partly follows Ref. [88], where the same procedure is used for a related problem.

1 - From the CME to a Fokker-Planck equation via Kramers-Moyal expansion.

Let us denote $\mathbf{n} = (w, m)$, a vector specifying the system state. The CME can be written in compact form as

$$\frac{dP(\mathbf{n}, t)}{dt} = \sum_{\mathbf{n}' \neq \mathbf{n}} [T(\mathbf{n}|\mathbf{n}')P(\mathbf{n}', t) - T(\mathbf{n}'|\mathbf{n})P(\mathbf{n}, t)] \quad (2.20)$$

$T(\mathbf{n}|\mathbf{n}')$ is the transition rate from state $\mathbf{n}'$ to $\mathbf{n}$. Given the reaction network in Eq. (2.12), the corresponding (global) transition rates are

$$\begin{aligned} T(w-1, m|w, m) &= w\mu \\ T(w, m-1|w, m) &= m\mu \\ T(w+1, m|w, m) &= w\lambda(w, m) \\ T(w, m+1|w, m) &= m\lambda(w, m) \end{aligned} \tag{2.21}$$

The first step is to write the CME in a continuous form. We replace copy numbers $\mathbf{n}$ by the abundances $\mathbf{x} = \frac{\mathbf{n}}{N_{ss}}$ relative to the effective population size. The variable $\mathbf{x}$ can be considered continuous for large N_{ss} . The CME becomes

$$\frac{\partial P(\mathbf{x}, t)}{\partial t} = \int_{-\infty}^{\infty} d\mathbf{x}' [T(\mathbf{x}|\mathbf{x}')P(\mathbf{x}', t) - T(\mathbf{x}'|\mathbf{x})P(\mathbf{x}, t)]. \tag{2.22}$$

We now proceed by expanding the CME through the second-order multivariate Kramers-Moyal expansion¹³⁸, that can be written as

$$\frac{\partial P(\mathbf{x}, t)}{\partial t} \approx \int_{-\infty}^{\infty} \left(-\nabla (T(\mathbf{x}'|\mathbf{x})P(\mathbf{x}, t))^T \cdot (\mathbf{x}' - \mathbf{x}) + \frac{1}{2}(\mathbf{x}' - \mathbf{x})^T \cdot \mathbf{H}_{\mathbf{x}'}(\mathbf{x}) \cdot (\mathbf{x}' - \mathbf{x}) \right) d\mathbf{x}' \tag{2.23}$$

where $\mathbf{H}_{\mathbf{x}'}(\mathbf{x})$ is the Hessian matrix of $T(\mathbf{x}'|\mathbf{x})P(\mathbf{x})$, whose general expression is

$$\mathbf{H}_{\mathbf{x}'}(\mathbf{x}) := \begin{pmatrix} \frac{\partial^2}{\partial x_1^2} & \cdots & \frac{\partial^2}{\partial x_1 \partial x_N} \\ \vdots & & \vdots \\ \frac{\partial^2}{\partial x_N \partial x_1} & \cdots & \frac{\partial^2}{\partial x_N^2} \end{pmatrix} T(\mathbf{x}'|\mathbf{x})P(\mathbf{x}). \tag{2.24}$$

A transition $\mathbf{x} \rightarrow \mathbf{x}'$ corresponds to some reaction j which moves the state from $\mathbf{x}$ to $\mathbf{x}'$. Since we know how each reaction affects state $\mathbf{x}$ through the constant stoichiometry matrix S_{ij} , and since the global rate of each reaction is independent of $\mathbf{x}'$ itself (see Eq. (2.21)), we may transition from a notation involving $\mathbf{x}$ and $\mathbf{x}'$ into a notation involving $\mathbf{x}$ and the reaction j that affects $\mathbf{x}$. We may therefore define $T_j(\mathbf{x}) := T(\mathbf{x}'|\mathbf{x})$, and let $\mathbf{H}_{\mathbf{x}'}(\mathbf{x}) \rightarrow \mathbf{H}_j(\mathbf{x})$.

We now wish to re-write Eq. (2.23) as a Fokker-Planck equation. Since the integral in Eq. (2.23) is over $\mathbf{x}'$, and for a given $\mathbf{x}$ every $\mathbf{x}'$ corresponds to a reaction j , we may interpret the integral in Eq. (2.23) as a sum over all reactions: $\int d\mathbf{x}' \rightarrow \sum_{j=1}^R$. Hence, for the j^{th} reaction, $[(\mathbf{x}' - \mathbf{x})]_i = S_{ij}/N_{ss}$. With these observations, we may write the first term in the integral in Eq. (2.23) as

$$\begin{aligned} \int_{-\infty}^{\infty} -\nabla(T(\mathbf{x}'|\mathbf{x})P(\mathbf{x},t))^T \cdot (\mathbf{x}' - \mathbf{x}) d\mathbf{x}' &= \int_{-\infty}^{\infty} -\nabla(T_j(\mathbf{x})P(\mathbf{x},t))^T \cdot (\mathbf{x}' - \mathbf{x}) d\mathbf{x}' \\ &= -\sum_{j=1}^R \sum_{i=1}^N \frac{\partial}{\partial x_i} (T_j(\mathbf{x})P(\mathbf{x},t)) S_{ij} \\ &= -\sum_{i=1}^N \frac{\partial}{\partial x_i} A_i P(\mathbf{x},t) \end{aligned} \quad (2.25)$$

where

$$\mathbf{A} := \mathbf{S}\mathbf{T}, \quad (2.26)$$

with $\mathbf{A}$ a vector of length N , S the $N \times R$ stoichiometry matrix and $\mathbf{T}$ the vector of transition rates, of length R . To re-write the second integral of Eq. (2.23), we write an element of the Hessian $\mathbf{H}_j$ in Eq. (2.24) as

$$H_{jlm} = \frac{\partial^2}{\partial x_l \partial x_m} T_j(\mathbf{x}) P(\mathbf{x},t) \quad (2.27)$$

where $j = 1, \dots, R$ and $l, m = 1, \dots, N$. Thus, we may write

$$\begin{aligned} \int_{-\infty}^{\infty} \frac{1}{2} (\mathbf{x}' - \mathbf{x})^T \cdot \mathbf{H}_{\mathbf{x}'}(\mathbf{x}) \cdot (\mathbf{x}' - \mathbf{x}) d\mathbf{x}' &= \frac{1}{2} \sum_{j=1}^R \sum_{l=1}^N \sum_{m=1}^N S_{lj} H_{jlm} S_{mj} \\ &= \frac{1}{2} \sum_{j=1}^R \sum_{l=1}^N \sum_{m=1}^N S_{lj} \frac{\partial^2}{\partial x_l \partial x_m} T_j P(\mathbf{x},t) S_{mj} \\ &= \frac{1}{2} \sum_{l=1}^N \sum_{m=1}^N \frac{\partial^2}{\partial x_l \partial x_m} \left(\sum_{j=1}^R S_{lj} T_j S_{mj} \right) P(\mathbf{x},t) \\ &= \frac{1}{2} \sum_{i,m=1}^N \frac{\partial^2}{\partial x_i \partial x_m} B_{im} P(\mathbf{x},t) \end{aligned} \quad (2.28)$$

where

$$\mathbf{B} := \mathbf{S} \cdot \text{Diag}(\mathbf{T}) \cdot \mathbf{S}^T, \quad (2.29)$$

and $\mathbf{B}$ is an $N \times N$ matrix, and $\text{Diag}(\mathbf{T})$ is a diagonal matrix whose main diagonal is the vector

T. We may therefore re-write Eq. (2.23) as a Fokker-Planck equation for the state vector $\mathbf{x}$ of the form

$$\frac{\partial P(\mathbf{x}, t)}{\partial t} \approx - \sum_{i=1}^N \frac{\partial}{\partial x_i} [A_i(\mathbf{x}) P(\mathbf{x}, t)] + \frac{1}{2} \sum_{i,m=1}^N \frac{\partial^2}{\partial x_i \partial x_m} [B_{im}(\mathbf{x}) P(\mathbf{x})]. \quad (2.30)$$

From a system of coupled ODEs for $P(w, m, t)$ (the CME of Eq. (2.20)) we have obtained a single PDE for $P(\mathbf{x}, t)$ approximating copy numbers as continuous variables.

2 - From Fokker-Planck to a system of Itô SDEs.

In general, the Fokker-Planck equation in Eq. (2.30) is equivalent¹³⁹ to the following Itô stochastic differential equation (SDE)

$$d\mathbf{x} = \mathbf{A} dt + \sqrt{\sigma} \mathbf{G} d\mathbf{W} \quad (2.31)$$

where $\mathbf{G}\mathbf{G}^T \equiv \mathbf{B}$ (where $\mathbf{G}$ is an $N \times R$ matrix) and $d\mathbf{W}$ is a vector of length R of independent Wiener increments, that satisfy

$$\int_0^t dW := W(t), \quad P(W, t) \equiv \frac{1}{\sqrt{2\pi t}} e^{-W^2/(2t)}, \quad (2.32)$$

and σ is a constant that controls the strength of the noise. For our chemical reaction network Eq. (2.12), the corresponding system of Itô SDEs is given in Eq. (2.14).

3 - From a system of Itô SDEs to a single effective SDE for a system forced onto the center manifold.

What follows is an adaptation from Refs. [140, 141], to which we refer for a proof and interpretation of the functions introduced. We start from Eq. (2.31). The procedure can be applied when the system admits a manifold on which $\mathbf{A} = 0$, the center manifold (CM). We thought that it would be a helpful complement to Refs. [140, 141], to make more explicit the sequence of steps involved for the specific system at hand.

- Identify the CM.

- Find the Jacobian of $\mathbf{A}$, namely J such that

$$J_{ij} = \frac{\partial A_i}{\partial x_j},$$

and evaluate it on the CM, obtaining J_{CM} .

- Find the eigenvalues $\lambda_1, \dots, \lambda_N$ of J_{CM} . The CM is the kernel of J_{CM} , hence the multiplicity of the zero eigenvalue is the dimensionality of the CM.
- Compute the decomposition of J_{CM} , writing

$$J_{CM} = W\Lambda W^{-1},$$

where $W = (\mathbf{w}_1, \dots, \mathbf{w}_N)$ is the matrix of eigenvectors with those corresponding to the zero eigenvalues written first.

- Compute J_{CM}^+ , defined as

$$J_{CM}^+ = W^{-1}\Lambda^+W,$$

where Λ^+ is the diagonal matrix with diagonal elements $\lambda_1^+, \dots, \lambda_N^+$ defined as

$$\lambda_i^+ = \begin{cases} 1/\lambda_i & \text{if } \lambda_i \neq 0 \\ 0 & \text{if } \lambda_i = 0. \end{cases}$$

- For each element of $\mathbf{A}$, compute the Hessian H_i as

$$H_{ijk} = \frac{\partial^2 A_i}{\partial x_j \partial x_k},$$

and evaluate it on the CM, obtaining $H_{i_{CM}}$.

- Compute the matrix $\mathcal{P}$ as

$$\mathcal{P} = I - J^+J,$$

where I is the identity matrix, and the matrices Q_i as

$$Q_i = - \sum_{l=1}^N [J_{il}^+ \mathcal{P}^T H_l \mathcal{P} + \mathcal{P}_{il} (J^{+T} H_l \mathcal{P} + \mathcal{P}^T H_l J^+)]$$

For notational simplicity, in the last two equations we wrote J to mean J_{CM} .

- Calculate the vector $\mathbf{g}(\mathbf{x})$, whose elements are

$$g_i(\mathbf{x}) = \frac{1}{2} \text{Tr} [\mathbf{G}^T Q_i \mathbf{G}].$$

- Finally, effective SDEs for the variable $\mathbf{z}$, namely the variable $\mathbf{x}$ constrained to the CM, are

$$d\mathbf{z} = \sigma \mathbf{g}(\mathbf{z}) dt + \sqrt{\sigma} \mathcal{P} \mathbf{G}(\mathbf{z}) d\mathbf{W}, \quad (2.33)$$

where the equations are now *uncoupled* because all the functions are evaluated on the CM.

This procedure allows one to obtain Eq. (2.15) as an effective description of the neutral stochastic model defined in Eq. (2.12). We provide a Mathematica notebook (Section A.4) to obtain Eq. (2.31) from the system of SDEs in Eq. (2.14). This technique is extended in Ref. [140] to the more general case in which the stochastic system is equivalent to a system of Itô SDEs of the form

$$d\mathbf{x} = (\mathbf{A} + \epsilon \mathbf{h}) dt + \sqrt{\sigma} \mathbf{G} d\mathbf{W}, \quad (2.34)$$

where $\mathbf{h} \equiv \mathbf{h}(\mathbf{x})$ is another vector-valued function of $\mathbf{x}$ and $\epsilon \ll \sigma$ is a small parameter, such that for $\epsilon = 0$ the system admits a CM. In this case, Eq. (2.33) becomes

$$d\mathbf{z} = [\epsilon \mathcal{P} \mathbf{h}(\mathbf{z}) + \sigma \mathbf{g}(\mathbf{z})] dt + \sqrt{\sigma} \mathcal{P} \mathbf{G}(\mathbf{z}) d\mathbf{W}. \quad (2.35)$$

The stochastic system in which mutants are subject to preferential elimination ϵ (Eq. (2.18)) corresponds to the case $\mathbf{h}(\mathbf{x}) = (0, x_2)$ (recall that $x_2 = m/N_{ss}$). Inserting this into Eq. (2.35), one obtains Eq. (2.19), which describes the noise and density-induced selection reversal.

Change of variable through Itô's formula to obtain an SDE for heteroplasmy.

Itô's formula states that, for an arbitrary function $y(\mathbf{x}, t)$ where $\mathbf{x}$ satisfies Eq. (2.31), we may write the following SDE:

$$dy(\mathbf{x}, t) = \left\{ \frac{\partial y}{\partial t} + (\nabla y)^T \mathbf{A} + \frac{1}{2} \text{Tr} [\mathbf{G}^T \mathbf{H}'(\mathbf{x}) \mathbf{G}] \right\} dt + (\nabla y)^T \mathbf{G} d\mathbf{W}, \quad (2.36)$$

where $\mathbf{H}'(\mathbf{x})$ is the Hessian matrix of $y(\mathbf{x}, t)$ with respect to x (see Eq. (2.24), where $T(\mathbf{x}'|\mathbf{x})P(\mathbf{x})$ should be replaced with $y(\mathbf{x}, t)$). Applying this rule to Eq. (2.15) for

$$y(\mathbf{x}, t) = (h(\mathbf{x}), n(\mathbf{x})),$$

where h is the heteroplasmy and n is the total copy number, one obtains Eq. (2.16)

2.3 Two-unit system and noise-driven increase in mean proportion of mutants

In this section we formulate the deterministic and stochastic two-unit model, the latter being the simplest system exhibiting stochastic survival of the densest. Moreover, we provide an intuitive understanding of the effect.

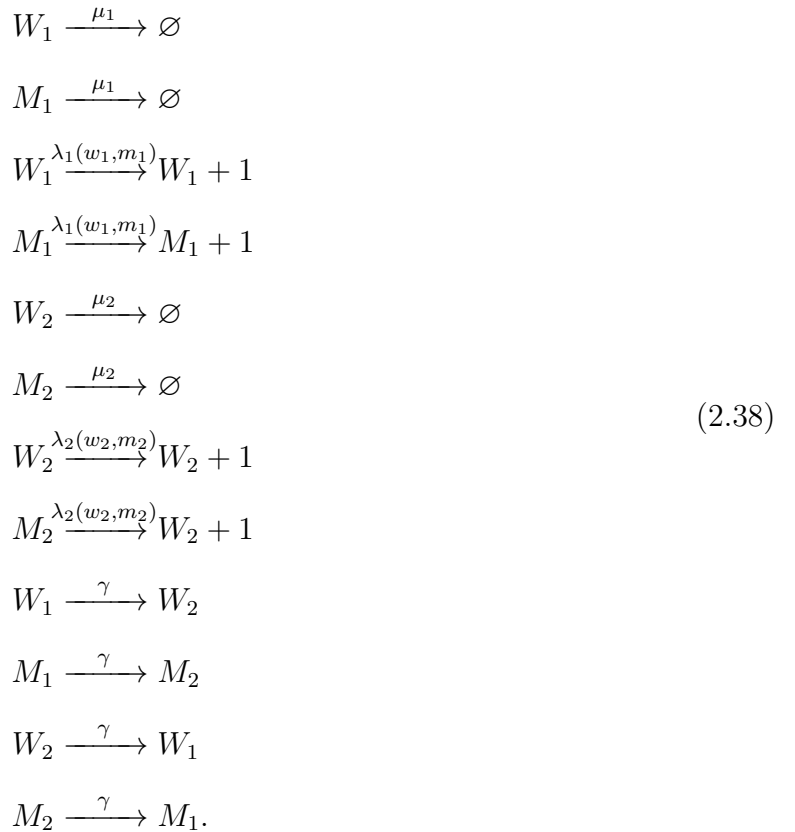
We introduce a minimal spatial structure coupling two units, allowing exchange of molecules between them. In each unit there is a population composed of wildtypes and mutants, evolving under Eq. (2.2). The deterministic two-unit model is formalised via the ODE system

$$\begin{aligned} \dot{w}_1 &= cw_1[N_{ss} - (w_1 + \delta m_1)] + \gamma(w_2 - w_1) \\ \dot{m}_1 &= cm_1[N_{ss} - (w_1 + \delta m_1)] + \gamma(m_2 - m_1) \\ \dot{w}_2 &= cw_2[N_{ss} - (w_2 + \delta m_2)] + \gamma(w_1 - w_2) \\ \dot{m}_2 &= cm_2[N_{ss} - (w_2 + \delta m_2)] + \gamma(m_1 - m_2). \end{aligned} \quad (2.37)$$

The addition of the second term on the RHS, proportional to the constant hopping rate γ

accounts for diffusion of the molecules between neighbouring units. The first term on the RHS of each equation means that each unit seeks to maintain its own separate target population, i.e. the control is local. This system does not admit an analytical solution. However, it has a CM given by $w_1 + \delta m_1 = w_2 + \delta m_2 = N_{ss}$, $w_1 = w_2$ (and consequently $m_1 = m_2$). It is easy to see that if these conditions are satisfied, all the time derivatives (the left-hand-sides) are zero and the dynamics stop, i.e., steady state is reached. Afterwards, heteroplasmy stays constant, as shown in Fig. 2.2B (black line). A similar system of ODEs can be used to describe chains of units of arbitrary length (see Section A.1).

The stochastic formulation of the model is obtained by replicating the reaction network in Eq. (2.38) for w_2, m_2 and adding the reactions accounting for hopping of the molecules between two units, which models diffusion when space is discrete. The death rate is still constant and common, and each unit has its own independently controlled replication rate given by $\lambda_i = \mu + c(N_{ss} - w_i - \delta m_i)$, for $i = 1, 2$. The full chemical reaction network is



In this setting of stochastic dynamics with higher mutant density ($0 \leq \delta < 1$) and spatial struc-

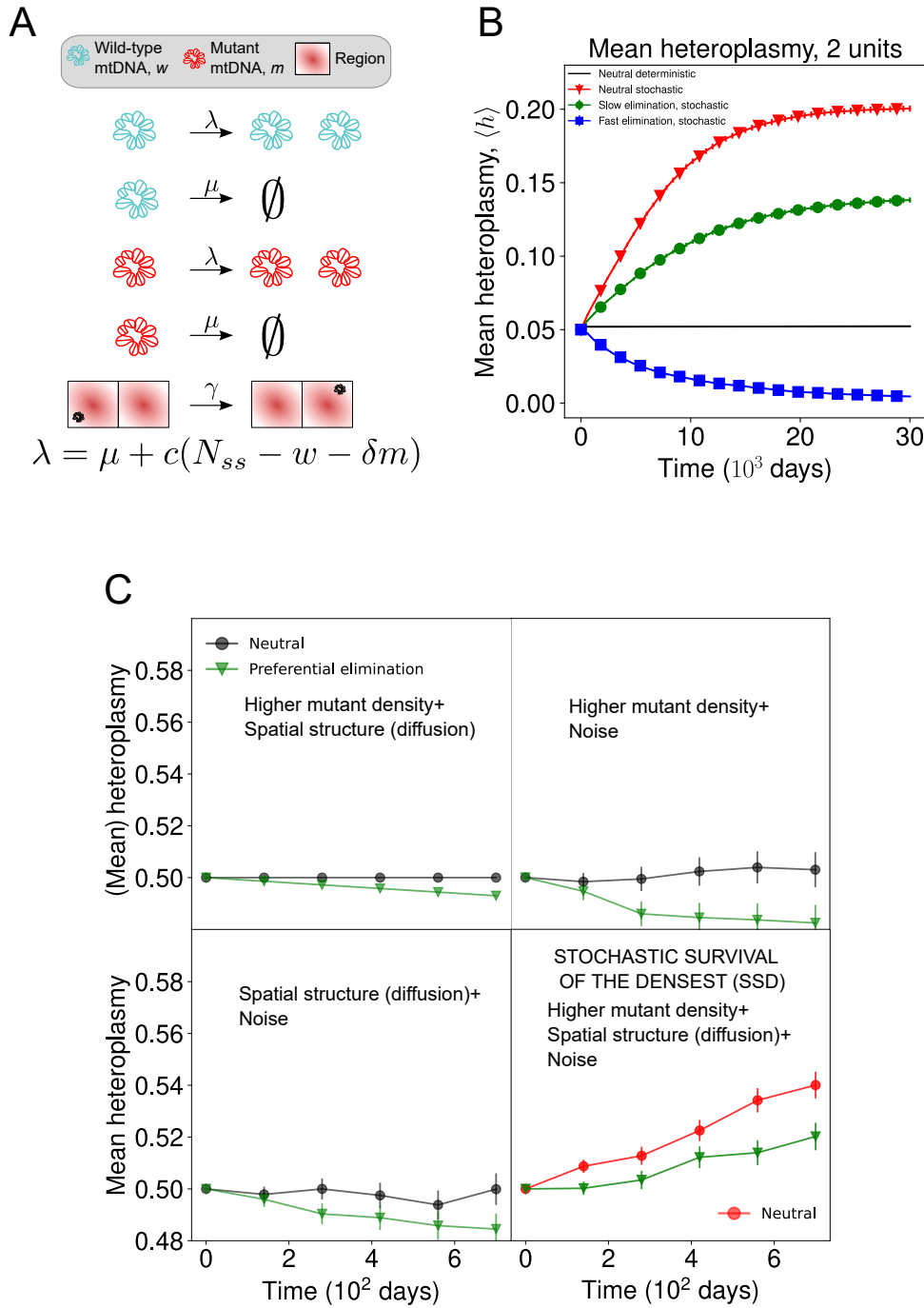


Figure 2.2: *Stochastic survival of the densest, increase in mean proportion of mutants, is observed in a spatially extended system in the presence of stochasticity and higher mutant density, even if mutants are preferentially eliminated.* (A) The minimal spatially extended system is composed of two units with same internal dynamics and with random hopping of molecules at constant rate γ . (B) In the stochastic two-unit system with higher density of mutants, mean mutant fraction (heteroplasmy) increases in the neutral model (red) and even if mutants are preferentially eliminated (green). Parameter values in Section A.2. Error bars are SEM. (C) Noise, spatial structure (with diffusion) and higher density of mutants are necessary and sufficient for stochastic survival of the densest (bottom right subpanel). If any of the factors is missing, as in the other subpanels, (mean) heteroplasmy stays constant in a neutral model (black line) or decreases with preferential elimination of mutants (green). Parameters in Section A.2. Error bars are SEM.

ture we crucially observe *increase in mean heteroplasmy*, as shown in Fig. 2.2B (red line). Fig. 2.2C proves that 1) spatial structure with diffusion of molecules between units, 2) stochasticity, 3) higher density of mutants ($0 < \delta < 1$) are necessary and sufficient to observe the increase in mean heteroplasmy. Indeed, it is observed if the three elements are present simultaneously (bottom right subpanel), and not observed if any of the three is missing (other subpanels). For this reason, we term this novel mechanism *stochastic survival of the densest* (SSD): mutants outcompete wildtypes in a stochastic setting by virtue of being denser. Importantly, the increase in mean heteroplasmy is observed even if mutants are subject to a slightly higher degradation rate (Fig. 2.2D, green line).

The dimensionality reduction technique from Ref. [140], used to obtain Eq. (2.19), is not viable in this case to obtain an SDE showing the increase in $\langle h_1 \rangle$ (or $\langle h_2 \rangle$) when $0 < \delta < 1$. The technique works by projecting the system onto the CM, where the two units are identical; however, the increase in mean heteroplasmy is driven by fluctuations that move *one* of the units away from the CM condition, after which the system relaxes to steady state with an average increase in h . Therefore the approximation involved in the technique is too crude to capture the effect.

The increase in mean heteroplasmy can be cast in terms of an increase of the fixation probability of mutants in the whole two-unit system. Consider two identical units coupled, so that heteroplasmy is the same at time 0 and $\langle h \rangle$ is approximately the same over time in the two units. Simulating multiple stochastic replicates of this system for a long time, we observe that $\langle h \rangle$ saturates to a constant value (Fig. 2.2B, red line). This value represents the fraction of replicates in which mutants take over the whole system. Indeed, a constant value means that in all replicates there has been either wildtype or mutant fixation, and the ensemble average of the fraction of mutants at that time is $\langle h \rangle$. The final value is greater than the initial one (for $0 < \delta < 1$) because the fixation probability of mutants *for the whole two-unit system* is greater than their initial fraction, which instead coincides with the fixation probability in the single-unit system (compare red lines in Fig. 2.1D Fig. 2.2B). Hence, the increase in mean heteroplasmy is closely related to the increase in the whole-system fixation probability of mutants.

Keeping in mind the connection between SSD and the increase in mutant fixation probability, one can gain an intuitive understanding of SSD considering the limit $\gamma \ll \mu, \lambda$, namely the limit in which diffusion is much slower than birth and death (the demographic processes). In this limit the fixation time in a single unit is much less than the typical time between hops of molecules. Consider two coupled units at carrying capacity, one containing only N_{ss} wildtypes and the other containing exclusively N_{ss}/δ mutants ($0 < \delta < 1$). At some point, one molecule will hop to the neighboring region and it will have time to either fix or go extinct before another hopping event takes place (the low hopping rate limit). Because of the larger number of mutants, the rate of mutant hopping into the wildtype unit is higher than that of wildtype hopping into the mutant region. Likewise because of the larger number of mutants, the fixation probability of a mutant hopping into the wildtype-only unit ($1/N_{ss}$) is higher than that of a wildtype entering the mutant-only unit (δ/N_{ss}). Because of the higher probabilities of hopping and fixation for mutants, the whole-system fixation probability of mutants is higher than that of wildtypes. Considering the above connection between fixation probability and the increase in mean heteroplasmy, this argument gives an intuition for SSD and clarifies the crucial role of diffusion and higher mutant density. It is also interesting to notice that in the opposite limit, i.e., $\gamma \gg \mu, \lambda$, SSD is not observed in a finite system because extremely fast diffusion would make multiple units behave like a single one.

2.4 Noise-driven travelling waves of mutants in a chain of units

The increase in mean heteroplasmy driven by SSD, conveniently highlighted in the two-unit system, is key to the appearance of a travelling wave of mutants in a chain of units. As before, each unit contains a population of mutants and wildtypes evolving according to Eq. (2.2). Any two neighbouring units exchange molecules at constant rate γ , which models diffusion of particles along the chain when space is discretised. In this setting the high-heteroplasmy front diffuses away without noise (Fig. 2.3D, left) or without higher mutant density (Fig. 2.3D,

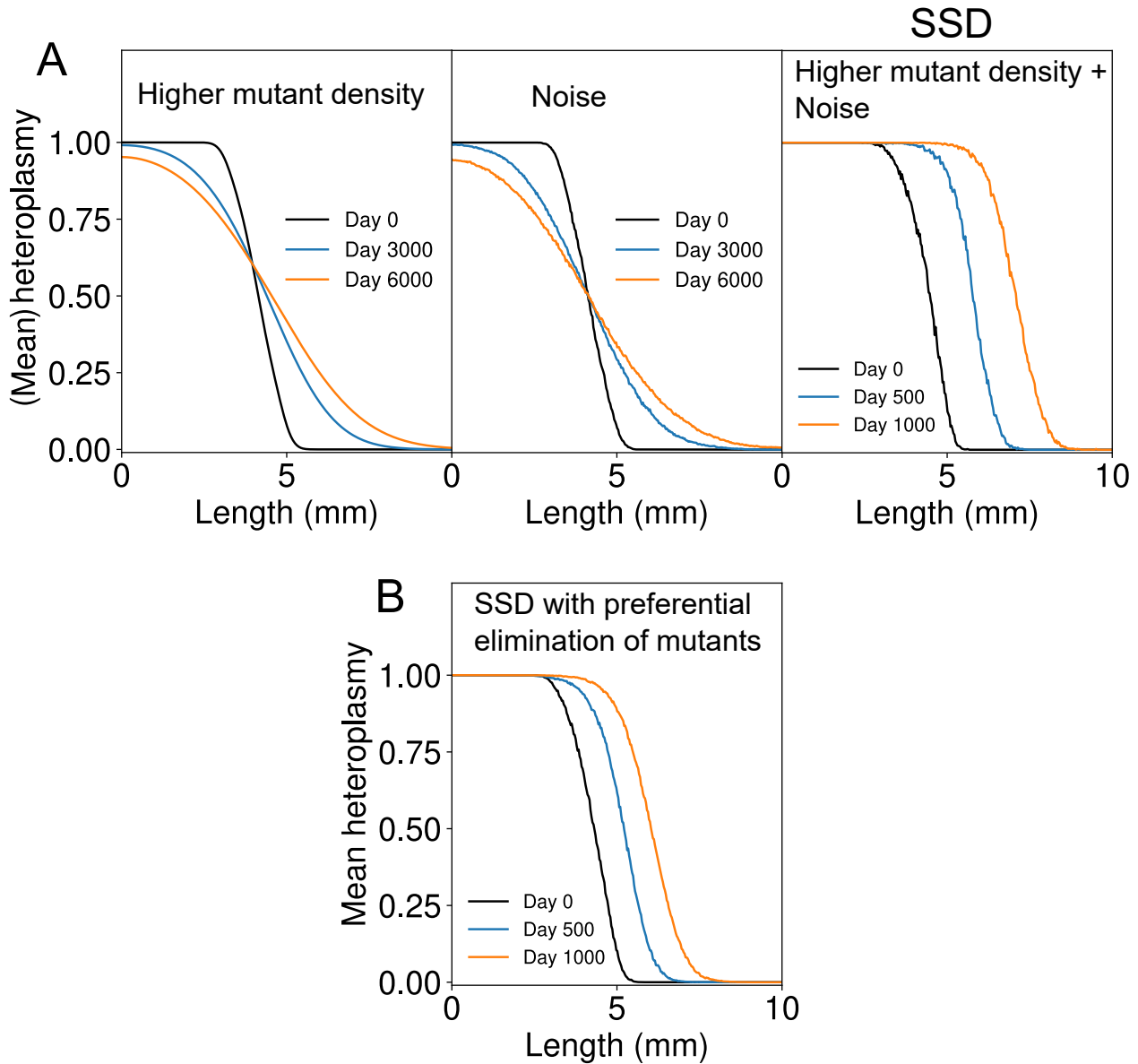


Figure 2.3: *Stochastic survival of the densest produces a travelling wave of mutants in a chain of units, even if mutants are preferentially eliminated.* (A) In a spatially structured model (chain with migration between neighbouring units), SSD drives a travelling wave of mutants only in the presence of noise and higher mutant density (rightmost subpanel), while the high-heteroplasmy front diffuses away if either noise (left, ODE system) or higher mutant density (middle, $\delta = 1$) is missing. “Length” on the horizontal axis stands for the position along the chain of units. Averaged over $2 \cdot 10^3$ stochastic replicates. (B) For the same model as in A, if mutants are preferentially degraded an SSD-driven invasive wave can still occur. Parameter values and details of simulations for both panels can be found in Section A.2.

middle), and advances only in the presence of noise and higher density mutants (Fig. 2.3D, right), i.e., when driven by SSD. In other words, the expanding travelling wave of mutants, driven by SSD, requires spatial structure with diffusion, noise and higher mutant density. In the two-unit model, the increase in mean heteroplasmy can be observed even if mutants are subject to (weak) preferential elimination. Analogously, a chain of units can exhibit a travelling wave of mutants despite their preferential elimination, as shown in Fig. 2.3B.

By measuring the simulated wave speed from SSD for 110 combinations of parameters, we found that it is well described by the phenomenological formula:

$$v \simeq 2\sqrt{kD}, \quad (2.39)$$

where D is the diffusion coefficient of molecules along the chain, and

$$k = \frac{\sqrt{(1-\delta)^2\mu\gamma}}{N_{ss}^{\frac{2}{3}}}. \quad (2.40)$$

Eq. (2.39) is analogous to the wave speed formula for the wave of advance of advantageous mutants introduced by Fisher and Kolmogorov^{143,144}. In our case k can be seen as an effective selective advantage for mutants induced by SSD, in contrast with the replicative advantage – or more generally difference in net growth rate – driving the original Fisher-Kolmogorov (FK) waves. In light of this analogy, SSD can be seen as a reaction-diffusion phenomenon where the reaction component emerges from the combined effect of noise and higher mutant density. One fundamental difference between the FK and SSD wave, is that Eq. (2.39) predicts that wave speed decreases when N_{ss} increases, in contrast to FK waves⁷⁶ for which the speed is Eq. (2.39) with k being the explicit difference in growth rates between the two species. Intuitively, in SSD the expansion of the denser species is driven by stochastic fluctuations, whose effect generally becomes smaller in larger systems. This observation will be important for the next chapter, where we compare the predictions of SSD and of a replicative-advantage model to the phenomenology of the expansion of mitochondrial deletion mutations in skeletal muscle fibres. In section A.1 and Fig.A.1 we provide more details about how Eq. (2.39) was

obtained.

In the limit of low diffusion (hopping) considered in the previous section 2.3, one can gain an intuitive understanding of how SSD leads to a travelling wave of mutants taking over the system. In this situation, there will be a sharp separation between the mutant-only and wildtype-only regions, with a step consisting of a unit containing N_{ss} wildtypes adjacent to a unit containing N_{ss}/δ mutants (recall $0 < \delta < 1$). Here, the higher probabilities of hopping and fixation for mutants – justified at the end of section 2.3 – produce a net movement of the mutant front into the wildtype-only region.

2.5 A noise-driven wave equation from the single-unit model.

In Ref. [76], a stochastic continuous-space model is presented, in the form of partial differential equations, describing a noise-driven wave of mutants, possibly subject to higher death rates than wildtypes. The paper is an extension of the classical Fisher-Kolmogorov PDE to a stochastic setting. In this section, we show how our microscopically interpretable bottom-up model, based on the simple Lotka-Volterra dynamics, gives rise to the model in Ref. [76] for $\delta \rightarrow 1$ and in a continuous-space limit.

Let us reconsider the deterministic system of Eq. (2.3) and change variable to (n, h) , where $n = w + m$ is the total population. We already observed that $\frac{dh}{dt} = 0$. For n we have

$$\frac{dn}{dt} = cn[N_{ss} - w - \delta m] \quad (2.41)$$

Let us write $\delta = 1 - \alpha$:

$$\frac{dn}{dt} = cn[N_{ss} - w - m + \alpha m] \quad (2.42)$$

The constancy of h allows us to write $m = hn$, hence

$$\frac{dn}{dt} = cn[N_{ss} - n(1 - \alpha h)] = cN_{ss}n \left[1 - \frac{n}{N_{ss}/(1 - \alpha h)} \right]. \quad (2.43)$$

Considering the case $\alpha \rightarrow 0$, equivalent to $\delta \rightarrow 1$, we can write

$$\frac{1}{1 - \alpha h} = 1 + \alpha h + \mathcal{O}(\alpha^2).$$

Neglecting $\mathcal{O}(\alpha^2)$ terms we obtain

$$\frac{dn}{dt} = cN_{ss}n \left[1 - \frac{n}{K(h)} \right], \quad (2.44)$$

with $K(h) = N_{ss}(1 + \alpha h)$. The total copy number follows a logistic growth with a carrying

capacity $K(h)$ that is a linearly increasing function of heteroplasmy h . In order to model a spatially-extended system such as a chain of units, we give spatial dependence to copy number, i.e., $n \equiv n(x)$. The hopping of molecules between neighbouring units of the chain is described in this case as a diffusion term, proportional to the second spatial derivative of $\frac{\partial^2 n(x)}{\partial x^2}$. In order to model a chain, we can therefore write

$$\frac{\partial n(x, t)}{\partial t} = D \frac{\partial^2 n(x, t)}{\partial x^2} + cN_{ss}n(x, t) \left[1 - \frac{n(x, t)}{K(h(x, t))} \right], \quad (2.45)$$

where we have explicitly written the time-dependence. D is the diffusion coefficient, for which $D = \gamma L^2$, where L is the inter-unit spacing.

Eq. (2.45), derived from the special case of the simple single-unit Lotka-Volterra model, is the starting point of the work in Ref. [76], where a noise-driven wave equation for (transformed) heteroplasmy is derived. The steps followed are:

- Write the equation for mutant copy number, which is analogous to 2.45 because mutants and wildtypes have identical death and replication rates.
- By change of variable, transform the equation for mutant copy number into an equation for heteroplasmy (mutant proportion).
- Simplify the equation for heteroplasmy by taking the low-diffusion limit $D \rightarrow 0$ ($\gamma \rightarrow 0$).
- Add a noise term to the resulting equation for heteroplasmy. In Ref. [76], the noise term is multiplicative and assumed as generated by sampling of individuals for reproduction in the Wright-Fisher model.
- Replace heteroplasmy with a non-linear function of it, that reduces to heteroplasmy itself in the limit $\delta \rightarrow 1$ (in which we are working).

The change of variable in the last point above, performed according to Itô's rule – leads to a diffusion equation with the addition of a noise-induced drift, which has the form of a logistic

growth term. One therefore ends up with a noisy reaction-diffusion equation for the transformed heteroplasmy $g(x, t)$, which in dimensionless variables (ξ, τ) reads:

$$\frac{\partial g(\xi, \tau)}{\partial \tau} = \frac{\partial^2 g(\xi, \tau)}{\partial \xi^2} + g(1 - g) + \text{noise} \quad (2.46)$$

The above stochastic partial differential equation (SPDE) is a noisy Fisher-Kolmogorov equation. It is known that its deterministic part admits an analytical wave-like solution^{143,144}. Due to the non-linear multiplicative noise¹⁴⁵, a solution and analytical wave-speed for the full SPDE are not available. However, simulations¹⁴⁵ show a wave-like expansion with a speed that decreases when population size increases. This is in agreement with the Gillespie simulations of our microscopic spatially-structured model and our phenomenological wave-speed formula Eq. (2.39), that applies also outside of the limits $D \rightarrow 0$, $\delta \rightarrow 1$.

In summary, our fundamental single-unit model leads to an analytic wave-equation, for limiting values of the parameters D and δ . Remarkably, from one of the simplest models of population genetics, the Lotka-Volterra model, it is possible to derive an equation for a travelling wave of mutants driven by stochastic survival of the densest without positing a replicative advantage, albeit in some limit.

2.6 Other feedbacks and modelling frameworks produce SSD

The travelling wave driven by SSD is not specific to the linear feedback control of Eq. (2.2). In Fig. 2.4C, D we show that other types of feedback control produce an analogous travelling wave of heteroplasmy without the need of a replicative advantage for mutants. The two additional types of controls are the *quadratic* control, namely

$$\lambda(w, m) = \mu + a[N_{ss}^2 - (w + \delta m)^2], \quad (2.47)$$

and the *reciprocal* control, given by

$$\lambda(w, m) = \mu - b \left(\frac{1}{N_{ss}} - \frac{1}{w + \delta m} \right), \quad (2.48)$$

where a and b are control strength parameters. These forms of the replication rate produce a non-monotonic dependence of the steady state overall copy number with heteroplasmy, unlike the linear feedback control (see Eq. (2.9)). This feature has been observed in leukocytes of mitochondrial diseases patients^{146,147}, which has motivated the study of these feedback forms^{11,135,148}. In Fig. 2.4 we show that, when $0 < \delta < 1$, these control mechanisms also produce a wave-like expansion. Moreover, we have verified that the speeds decrease when N_{ss} increases, a fundamental difference between our model and models based on a net replicative advantage for mutants.

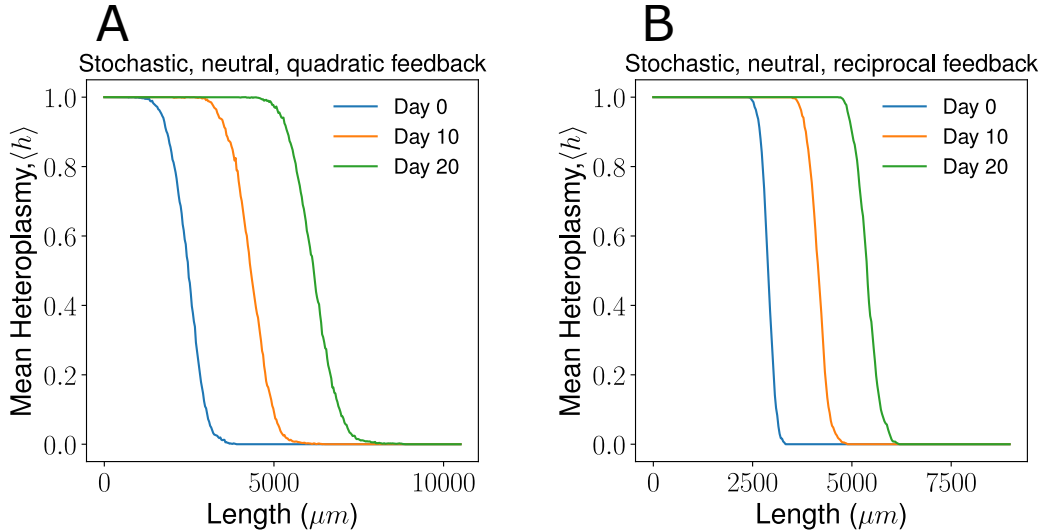


Figure 2.4: *Different choices of feedback control still produce a wave of mutants when stochasticity is present.* (A) Quadratic and (B) reciprocal feedback control (see Section A.2.4 for parameters) produce a qualitatively similar wave-like expansion of mutants in a stochastic model. Averaged over 400 stochastic replicates.

Moreover, we have verified that SSD is produced even in quite a different modelling framework, namely in coupled units evolving according to the Wright-Fisher model, with discrete non-overlapping generations and binomial sampling from parental generation and for molecules which hop to a neighbouring unit in each generation. In Fig. 2.5 we show that this model reproduces the four subpanels of Fig. 2.1C, that prove that the three elements (spatial structure, noise and increased mutant density) are necessary and sufficient to have SSD. Selective

elimination ϵ is reflected in a lower effective growth rate for mutants. Reproducing SSD in this alternative modelling framework was an idea of the anonymous reviewer of Ref. [123], who also coded up the dynamics of the system. We added stochastic averages, and the option of decreasing the effective growth rate of mutants, which is equivalent to adding selective elimination of mutants in the main continuous-time model used in this thesis.

In the light of these results, we argue that SSD is not tied to the mathematical details of the Poisson Point Process and the linear-feedback model presented in this work, provided that the three key factors of higher mutant density, stochasticity and spatial structure (with diffusion) are present.

2.7 Implications for the evolution of altruism

An altruistic trait has been defined as one that benefits others while costing its carriers^{79,80}. In Section 2.2.2, specifically in Eq. (2.9), we show that the global carrying capacity of the system – the total population it can sustain – increases with the proportion of mutants. In this sense, mutants are a density-increasing species. In Eq. (2.10) we show that, equivalently, the replication rate of both wildtypes and mutants is an increasing function of the proportion of mutants. If we assign a higher degradation rate to mutants, then our model can be linked to the specific definition of altruism in Refs. [79, 80]: “an organism behave altruistically when its behaviour benefits other organisms, at a cost to itself”. Indeed, the increase in carrying capacity and replication rate can be seen as a benefit that mutants bring to the whole population, including wildtypes, at the cost of a higher degradation rate. An increase in the system’s carrying capacity is present also in mathematical models of public good production^{75,149} and economical use of limited resources⁸⁰. These two behaviours are observed in microbial communities and have been described as early examples of altruism in the history of life^{76,150}. Here we highlight how our work contributes to the debate on the evolution of altruism, distinguishing SSD from concurrent lines of research referring to Section 2.8 for further links to the wider evolutionary biology literature.

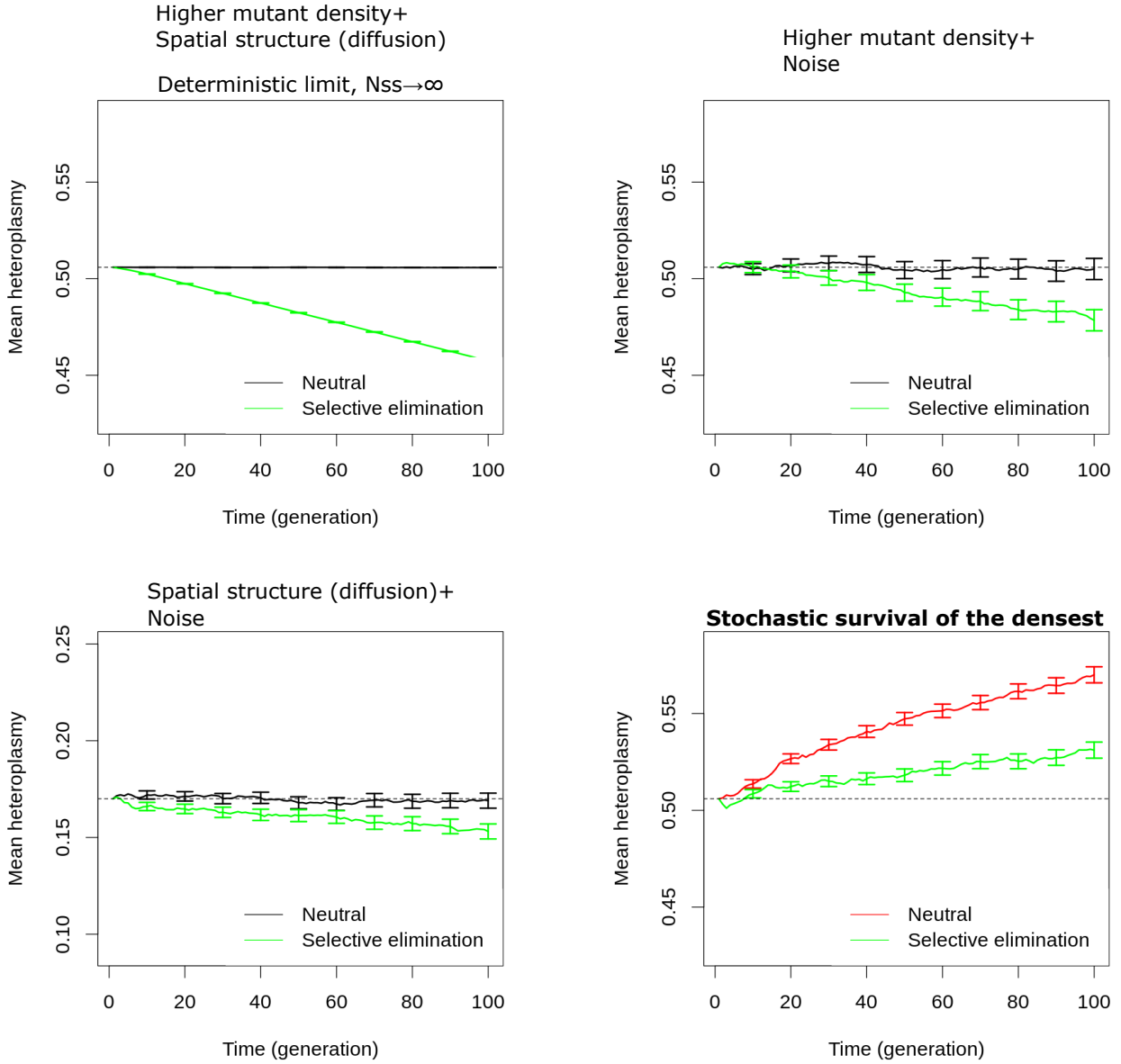


Figure 2.5: A simulation of two units based on Wright-Fisher model, with binomial sampling of molecules that hop between units, reproduces Fig. 2.1C, confirming that SSD is observed in the presence of noise, higher mutant density and spatial structure with diffusion, and that it is not observed if any of the three is missing. The deterministic simulation (upper left panel, originally the system of ODEs in Eq. (S36)) is replaced by the deterministic limit of the stochastic model $N_{ss} \rightarrow +\infty$, which effectively suppresses noise. We keep parameter values identical across panels, except for switching off a single element of SSD per panel. Parameter values are detailed in Section A.2; see Section A.4 for code. Averaged over 4000 replicates, error bars are SEM.

The two most prominent accounts for the evolution of altruism are kin selection and group selection¹⁵¹. Kin selection theory argues that altruism directed towards genetically related individuals can increase the overall success of the altruistic gene, which is likely to be shared by relatives of the carriers¹⁵². In group selection, altruism spreads through advantages conferred to the group; for instance, a group whose members are ready to sacrifice themselves for other members will likely prevail over other groups¹⁵³. These approaches seek a *deterministic* advantage for altruistic traits. In our model, the spread of density-increasing mutants is *driven by noise* and has no deterministic counterpart. Therefore our work is separate from kin and group selection.

A first, important merit of our model consists in describing the spread of the density-increasing trait in a spatially extended system using a fundamental and widely-used model of population genetics, the Lotka-Volterra model, which uses linear per molecule transition rates. The Moran model is another fundamental model of stochastic population genetics, dealing with the less general case of finite fixed-size populations. Previously, the study in Ref. [77] has shown that in a modified Moran Model with nonlinear rates, where the assumption of fixed population size is relaxed, density-increasing mutants (termed altruists by the author) can have a higher fixation probability despite a higher death rate. Analogous results are presented in Ref. [78], which is based on a modification of the Wright-Fisher model, retaining the feature of discrete, non-overlapping populations. Our continuous-time model uses simpler (linear) rates, and the relevance of the Lotka-Volterra model, widely used in population genetics, indicates that SSD might be a widespread effect.

The increase in mean mutant copy number in the single-unit system in a neutral model was highlighted 20 years ago in the field of mitochondrial biology, in the context of the relaxed replication model formulated in Refs.^{81–83,154} and further analysed in Refs.^{135,154}. Models yielding an increase in mean *copy number* of a denser mutant species have found that space can amplify the effect of randomness^{75,140,141}, in the sense that the denser mutants can tolerate a higher selective elimination in a spatially structured system. Crucially, our work explicitly shows that spatial structure can lead to a qualitatively different result, namely an increase in the *proportion* of the density-increasing mutants, that in turn leads to a travelling wave invading the system.

Another contribution of our study is the use of a physically motivated and biologically interpretable microscopic model, directly applied to the expansion of mitochondrial deletion mutations in skeletal muscles (in the next chapter). The study in Ref. [76] presents a phenomenological, macroscopic model that exhibits a noise-driven wave of (what the author calls) altruists. In Section 2.5 we have shown that the model in Ref. [76] is a special case of ours, in the limit of continuous space. We have further been able to obtain a (phenomenological) formula for the wave speed covering a range of biologically and clinically relevant regimes not covered by other studies, namely without taking $\delta \rightarrow 1$ and without imposing the quasi-static limit, i.e., very slow diffusion (see Section A.1).

2.8 Further links to the wider evolutionary literature

In this section we give an overview of, and distinguish SSD from, previous work on density-dependent selection and the impact of spatial structure on the evolution of altruism. The study of density-dependent selection has attracted interest across five decades^{155–158}, including in systems where stochasticity is important^{159–165}. The theory of r/K selection^{155,166}, one of the most influential ideas in evolutionary biology^{167–169}, examines how selection shapes the strategies of species in the presence (or in absence) of density effects. Density-dependent selection refers to situations in which the fitnesses of the different species within a population depend differently on population density (or size)¹⁶². In SSD, the two species have identical per capita death and replication rates. The dependence on population size is obviously identical for the two species. Hence SSD is not an example of density-dependent selection. Per capita degradation rate of mutants can be higher by a constant, but this does not introduce any dependence on population size (and is not a necessary element of SSD). Another fundamental difference is that SSD is dependent on stochasticity, while density-dependent selection can equally apply to stochastic and deterministic sections.

Frequency-dependent selection¹⁷⁰ is distinct, but related, to density-dependent selection. Numerous definitions of frequency dependent-selection have been used over the decades, a classic

one being that the relative fitness of a species varies with the relative frequencies of other species¹⁷¹. Our model satisfies a strong mathematical definition of frequency-*independent* selection, based on the per capita growth rate (PCGR, the difference between per capita birth and death rates) of the species. The PCGR can be defined as a vector-valued function of copy number variables, with an output for each species. Frequency-independent selection is assured when the Jacobian of the PCGR is independent of copy numbers¹⁷². This is clear for our linear feedback model (and all generalised Lotka-Volterra models), since the replication rate in Eq. (2.2) is linear in w and m , leading to a constant Jacobian. Therefore, SSD is not a case of frequency-dependent selection.

In the context of a single-species population, the Allee effect can be described as a positive dependence of the PCGR on the density or size of the population for some interval of the values of population size¹⁷³. This contrasts with theories in which the PCGR always decreases with population size because of competition for resources or space among members of the same species (intra-specific competition), as in logistic growth. When the dynamics of a population exhibits the Allee effect, it means that there is some other mechanism at play that counterbalances the general limiting effect of competition on the PCGR. An example is that of plants that reproduce through pollination, whose rate is increased by the density of plants in the environments. The Allee effect is connected to altruism, in that the efficiency of some fitness-increasing behaviours, e.g., defence and feeding, is enhanced by a large population size when the associated tasks are carried out cooperatively¹⁷⁴. The difference between SSD and the Allee effect is two-fold. The Allee effect refers to the possibility that PCGR in a population can be an increasing function of population size (for some interval) despite the increase in intra-specific competition that generally comes with higher density. Our work instead models inter-specific (between species) competitions, showing that the species that is *denser in isolation* can prevail despite a higher degradation rate and no replicative advantage. Moreover, the PCGR of our neutral model is proportional to $N_{ss} - w - \delta m$ for both species; in the case of a higher degradation rate for mutants, their PCGR is proportional to $N_{ss} - w - \delta m$. In both cases we have linearly decreasing functions of the population size, with no positive dependence on it. Therefore, SSD is distinct from the Allee effect.

Other studies have explored how the spatial structure of a population can favour^{175,176} or hinder¹⁷⁷ the evolution of altruistic behaviour. In these studies, the *continuous* spatial structure of the population influences the way its members interact, with interaction possible only within spatial neighbourhoods. When altruists cluster together they share the benefits of their behaviour mainly among themselves, and therefore the clustering can favour the spread of altruism. Our work is different in that, within each unit, the population is well-mixed and everyone benefits from the mutants' altruism in the same way, essentially mixing with all other particles at each point in space.

2.9 Discussion

In this chapter, we have introduced stochastic survival of the densest, a mechanism whereby a species can outcompete another in a spatially structured system without a replicative advantage and even if subject to preferential elimination. This counterintuitive effect is driven by stochasticity, spatial structure and density difference between the two species, with the denser species taking over. In an extended system, stochastic survival of the densest can drive wave of advance of the denser species. In Eq. (2.39) we have provided a phenomenological formula for the wave speed in the case of linear replication rate (Eq. (2.2)). Importantly, this speed decreases when population size increases. Analytical results for the single-unit model have been presented, which are useful to understand some of the features of stochastic survival of the densest. Eq. (2.35) highlights that the largest preferential elimination ϵ_M that mutants can tolerate and still expand also decreases with the (effective) population size N_{ss} . We have also shown that a model introduced in a previous study⁷⁶, where an analytical equation for a noise-driven wave is derived, is a continuous-space approximation of our model, for some limiting values of parameters. We have verified that our core predictions are robust to the particular choice of replication rate and to details of the modelling framework.

In the next chapter, we apply our model to the expansion of mitochondrial deletion mutants in skeletal muscle fibres. We believe that stochastic survival of the densest might have appli-

cations elsewhere. For example, a classic setting for the wave-like spread of a trait is in the uptake of agriculture^{178–181}. However, agriculture’s benefits to health and net fertility have been debated and it might not impart an explicit replicative advantage, possibly leading to higher death rates^{182–185}, but nonetheless spread due to an increase in density of the “farmer species”, namely an increase in carrying capacity of the farmed land. We believe that our work and the simplicity of the microscopic model we use might pave the way for an increased recognition of this intriguing mechanism in evolutionary biology, especially in light of the connections we have drawn with the evolution of altruistic behaviour.

Chapter 3

Stochastic survival of the densest and ageing of skeletal muscle

Summary

Here we apply stochastic survival of the densest (SSD) to the expansion of mitochondrial deletions in skeletal muscle fibres. We show that a literature-parameterised model based on SSD can reproduce a) the spatial profile of data on heteroplasmy in skeletal muscle b) the observed speed of expansion, c) its relationship with mtDNA population size and d) the level of mutations accumulated in a typical life-span. In our model, mutants expand without requiring a higher replication rate than wild-types, and can do so even if subject to preferential degradation. We argue that the spatial structure of skeletal muscle fibres is an important element that has been neglected by previous neutral models based on random drift. We show that a standard model with a replicative advantage for deletions produces an unrealistically fast expansion. We explore potential strategies to delay the expansion of deleterious mutants based on this novel understanding of the phenomenon.

3.1 Introduction

In this chapter we argue that stochastic survival of the densest (SSD) can account for the clonal expansion of mitochondrial DNA (mtDNA) mutation deletions in skeletal muscle fibres. This phenomenon has been causally implicated in sarcopenia, the loss of skeletal muscle mass and strength with age^{17,40,43}. We have presented the expansion of mtDNA deletions in the Introduction to this thesis, where we explained how existing mathematical models, based on a replicative advantage for mutations or on neutral random drift, have failed to give a coherent account of experimental observations. This lack of understanding concerns specifically the expansion of mtDNA deletions, commonly observed in post-mitotic tissue such as skeletal muscle. The expansion of point mutations, typically observed in mitotic tissue, is instead accounted for by models based on random drift⁵¹.

The high levels of mitochondrial deletions observed in aged skeletal muscle are attributed less to a high mutation rate than to the *clonal expansion* of these mutations. Indeed, clonality is the defining feature of this expansion. Single-cell studies have consistently reported that fibres in damaged zones of the muscles are taken over by a single deletion. This has been observed in rats^{17,45}, rhesus monkeys^{46,47} and humans⁴⁸. As different affected fibres generally have different deletions and coexist with adjacent unaffected fibres^{16,49,186}, it is believed that the founder mutation occurs and expand clonally during lifespan⁴⁸.

Another widely reported feature of the clonal expansion of mitochondrial deletions is that regions of muscle fibre taken over by the mutations present an approximately fivefold increase^{16,17,40,49,187} in absolute mtDNA copy number. A region of a fibre taken over by mutants contains a larger number mtDNA molecules than a region of the same length containing only wildtypes. This can be phrased in terms of carrying capacity of mutant mtDNA, or “density”: number of molecule per spatial unit. In this sense, mtDNA deletions are the denser species. Molecular mechanisms to explain the increased mutant density have been previously proposed, including the “maintenance of wildtype” hypothesis^{81,188}, and homeostasis on ATP production^{189,190}, proteome status^{82,191} and copy number¹⁹². These mechanisms essentially hypothesise that malfunctioning mutant mtDNA molecules cause an overall increase in

mtDNA replication as a nuclear response to restore some compromised cellular function, such as ATP production or protein synthesis. This brings about a higher absolute number of mtDNA molecules when the proportion of mutants is higher, leading to a higher mutant density. Our application of SSD to mtDNA in skeletal muscles relies on the above-mentioned higher density of mutants, but is independent of the detailed molecular mechanism behind it.

We deploy the spatial stochastic model introduced in the previous chapter to describe the spatio-temporal dynamics of mtDNA populations in skeletal muscle fibres. We will see that our model takes into account the fundamental features of the phenomenon and recapitulates experimental observations using only four parameters, with each parameter being constrained by auxiliary experimental data, without assuming any replicative advantage. We are able to reproduce the spatial features of the distribution of the mutant load along fibres, which are compatible with the travelling-wave of mutants driven by SSD. Our model predicts a speed of expansion that is compatible with observations, while a standard model based on a replicative advantage of mutants produces an implausibly fast clonal expansion. The SSD-driven wave has a speed that decreases as population size – here mtDNA copy number in the fibre – increases (see Eq. (2.39)); we corroborate this prediction on further experimental data. Finally, we show that an SSD-based model can predict the mutant loads observed in muscle fibres with reasonable *de novo* mutation rate, whereas previous models have overestimated these rate.

3.2 SSD matches clonal expansion in skeletal muscle fibres

Skeletal muscle fibres are long, multinucleated cells, with each nucleus surrounded by a mitochondrial population under its control. The clonality of the expansion of deletion mutations^{14,17,193} supports the idea that mtDNA molecules move along the fibres, as it is unlikely that the same mutation event occurred multiple times in distinct but adjacent regions of the same fibre. Mitochondria are tightly packed in skeletal muscle and the mitochondrial network is hyperfused^{194,195}. Moreover, it has been observed that mtDNA nucleoids move diffusively in

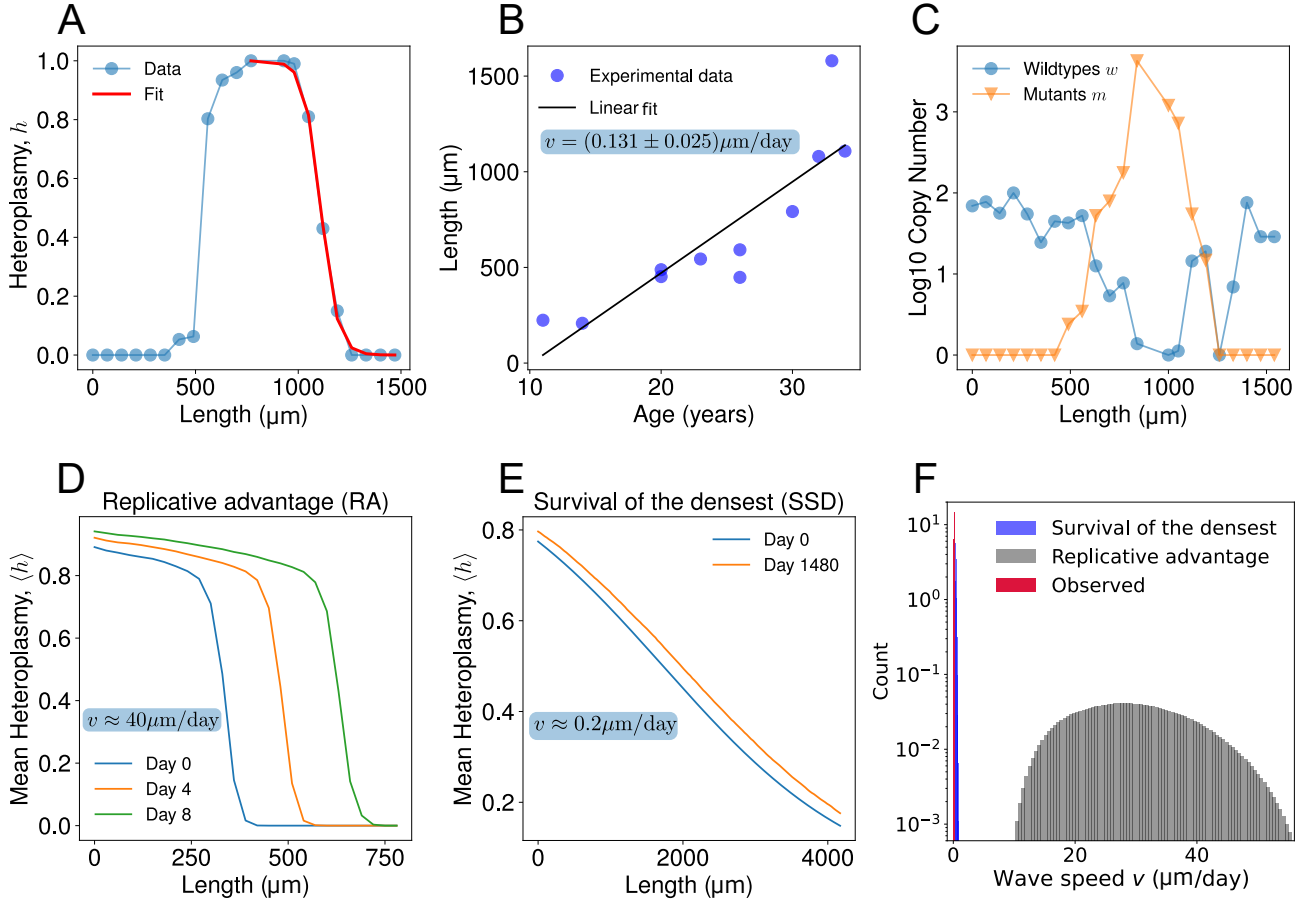


Figure 3.1: *Stochastic survival of the densest predicts a wave-like expansion of mtDNA mutants at a speed in agreement with experimental observations, while a standard linear replicative-advantage model (Eq. (B.1)) predicts a speed a factor of ≈ 300 too large.* A) Spatial profile of mutant fraction (heteroplasmy) along a human skeletal muscle fibre¹⁶. The heteroplasmy profile follows a sigmoid (Section B.7), the shape expected for a travelling wave. (B) Experimental data on the length of abnormal regions of muscle fibres against age in rhesus monkeys, from Ref. [193] (see Section B.6 for data processing). An approximate linear relationship is found ($R^2 = 0.76, p = 5 \cdot 10^{-4}$), compatible with a wave-like expansion with speed $(0.131 \pm 0.025) \mu\text{m}/\text{day}$ (linear fit). (C) Spatial structure of copy number for wildtype (blue) and mutants (orange) for the same muscle fibre as in panel A. The heteroplasmic regions present a higher absolute copy number, i.e., mutants are present at a higher density. (D) Stochastic simulations of a spatially extended model with a standard replicative advantage for mutants (Eq. (B.1)), with our best estimate of the model parameters (see Section B.2) for muscle fibres of rhesus monkeys predicts a wave-like expansion with a speed of $\simeq 40 \mu\text{m}/\text{day}$, 300 times faster than the observed speed. Averaged over 500 stochastic replicates, with $\mu = 0.07/\text{day}$, $\gamma = 0.1/\text{day}$, $N_{ss} = 3500$, $c = 2 \cdot 10^{-4}/\text{day}$, no density difference and replicative advantage $k_{RA} = 4cN_{ss}$ (see (Eq. (B.1))). (E) Simulations of survival of the densest yields a mutant wave speed of $\simeq 0.2 \mu\text{m}/\text{day}$ for the fibres of rhesus monkeys, which is comparable with experimental observations (see panel B). No replicative advantage and $\delta = 0.2$; other parameters are as in panel D. Averaged over 800 stochastic replicates. (F) Inserting probabilistic estimates of the model parameters (see Section B.2) into Eq. (2.39), we find that survival of the densest predicts a distribution for the wave speed (red histogram) compatible with observations (blue), whereas a wave driven by replicative advantage with the same model parameters is two orders of magnitude faster (grey).

mitochondria¹⁹⁶. Therefore we assume that mtDNA molecules diffuse along muscle fibres.

We model a fibre as a chain of discrete units, each being a system as introduced in section 2.2. Each unit contains a nucleus and a population of mtDNA molecules, composed of wildtypes w and mutants (deletions) m . This population is assumed to evolve notionally under the control of the nucleus, according to the stochastic dynamics introduced in 2.2.3, with the linear per-molecule replication rate of Eq. (2.2) (linear feedback control). We rewrite the replication rate here for simplicity:

$$\lambda(w, m) = \mu + c(N_{ss} - w - \delta m). \quad (3.1)$$

The dependence of the replication rate on w and m encodes a homeostatic control of the mitochondrial population. There is indeed extensive evidence that the mtDNA population of a cell exchanges signals with the nucleus, and that this signalling regulates mitochondrial content¹⁹⁷. As mitochondria are organised in networks, that are densely packed in skeletal muscle^{194,195}, we assume that the unit of homeostasis is the single mtDNA molecule and not the organelle: the nucleus exerts control over the population of mtDNA molecules.

The observed higher density of mtDNA deletion mutants^{16,17,40,49,187} is encoded in the parameter $0 < \delta < 1$. This also means that a change in the number of mutants will entail a smaller variation in $\lambda(w, m)$ than an equal change in the number of wildtypes. Mutants are less tightly controlled, i.e. the nucleus senses changes in the number of mutants less strongly. The quantity $w + \delta m$ can therefore be considered an effective population, i.e. the actual population sensed by the nucleus, and N_{ss} is the effective mtDNA population at steady state, when $\lambda(w, m) = \mu$. The parameter $c > 0$ quantifies the strength of the nuclear control towards homeostasis. Although the linear form of the replication rate might seem restrictive, if the system is close to steady state ($w + \delta m - N_{ss} \simeq 0$), a linear relationship approximates non-linear models. We will verify that our core results are robust to different, non-linear forms of the feedback mechanisms.

There is evidence of diffusive movement of mtDNA in mitochondria¹⁹⁶, and the clonality of the expansion of deletions suggests that mitochondria move along fibres. We model this movement

allowing random hopping of mtDNA molecules between neighboring units at constant rate γ . We use the above rates in a Poisson process simulated via the Gillespie algorithm. Therefore we have the three elements of SSD: density difference ($\delta \neq 1$), spatial structure with migration and stochasticity. From the previous chapter, we know that in the above setting SSD predicts an expansion of the denser mutants through a travelling wave. Experimental data (Fig. 3.1A, from Ref. [16]), show the characteristic spatial profile of heteroplasmy data, with high-heteroplasmy (OXPHOS-defective) regions flanked by transition regions to low or zero heteroplasmy. This heteroplasmy profile is well described by a sigmoid (Section B.7), the shape expected for a travelling wave (Section B.5), in a manner compatible with the expansion being a wave-like phenomenon. We report more profiles in Fig.3.2A, C and in Section B.1.

We have estimated the speed of this expansion by analysing data from Ref. [193] on the length of abnormal regions in rhesus monkeys (see Section B.6 for data processing). Regressing the lengths against age (Fig. 3.1B), we observe a relationship ($p = 5 \cdot 10^{-4}$) which is approximately linear ($R^2 = 0.76$) and corresponds to an average wave speed of $(0.131 \pm 0.025) \mu\text{m}/\text{day}$. Fig. 3.1C, relative to the same fibre as in 3.1A¹⁶, highlights the key fact that the absolute copy number in high-heteroplasmy regions is larger than in normal regions^{17,40,49,81,187}): mutants are the denser species.

In Section B.2 we explain our estimates of the parameter values from published experimental data for a typical mtDNA deletion mutation in rhesus monkeys fibres. With these estimates, simulations of a standard linear model (Eq. (B.1)) based on replicative advantage, in which the increase in mutant density is explained by a higher replication rate as expressed in Eq. (B.2), predict a wave-like expansion of $\approx 40 \mu\text{m}/\text{day}$ (Fig. 3.1D), 300 times faster than the observed speed. In contrast, SSD predicts a wave-like expansion of deletions with a speed of $\approx 0.2 \mu\text{m}/\text{day}$ (Fig. 3.1 E). Different mtDNA deletion mutations will have different characteristics, including increase in density or (supposed) replicative advantage. Other parameters of the system, like the diffusion constant, are also uncertain. Eq. (2.39) allowed us to verify robustness to uncertainty in parameter values. We obtained probability distributions for the wave speeds predicted by SSD and a replicative-advantage model, by inserting draws from the distributions of parameter values (given in Section B.2) into Eq. (2.39), with the appropriate

interpretation of k for the two models. The predicted distributions are plotted in Fig. 3.1F, together with the distribution of the experimentally observed wave speed obtained via linear fit (Fig. 3.1B). After accommodating this parametric uncertainty, SSD remains much superior to replicative advantage at reproducing the observed speed.

In Section B.3 we show that allowing for both higher mutant death and replication rate would not change the replicative-advantage wave speed distribution of Fig. 3.1F. We cannot exclude a situation in which an SSD *and* a replicative advantage (and possibly selective elimination) coexist, but our analysis suggests that replicative advantage is not necessary and – on its own and in its standard form – not viable. In summary, we have presented a null model showing that SSD can account for the fundamental features of the clonal expansion of mtDNA deletion mutations in skeletal muscle fibres.

3.3 Linking ageing, wave speed and copy number

As Eq. (2.39) states, SSD predicts a wave speed that decreases when copy number (per nucleus) increases. Indeed, the expansion of mutants is driven by stochastic fluctuations, whose effect generally becomes smaller for larger population size (see also Eq. (2.15)). In contrast, a standard model based on replicative advantage predicts a wave speed that increases with population size^{76,143}. This allows us to test SSD and replicative-advantage models against other experimental observations. It has been found that copy number depletion caused by antiretroviral therapy¹⁹⁸ or AKT2 deficiency¹⁹⁹ is associated with enhanced sarcopenia. Likewise, statins are well known for increasing the risk of sarcopenia^{200–202} and have consistently been associated with reduction in mitochondrial copy number^{203–205}. Conversely, increase in mtDNA content through exercise^{206,207} or overexpression of TFAM²⁰⁸ and Parkin²⁰⁹ have been found to protect against sarcopenia and muscle atrophy.

Skeletal muscle fibres can be broadly classified into Type 1 (oxidative) and Type 2 (glycolytic) fibres²¹⁰. The former rely on OXPHOS to function and typically have twice as many mitochondria as the latter^{193,211,212}, that depend on glycolysis. It is known that Type 2 fibres are

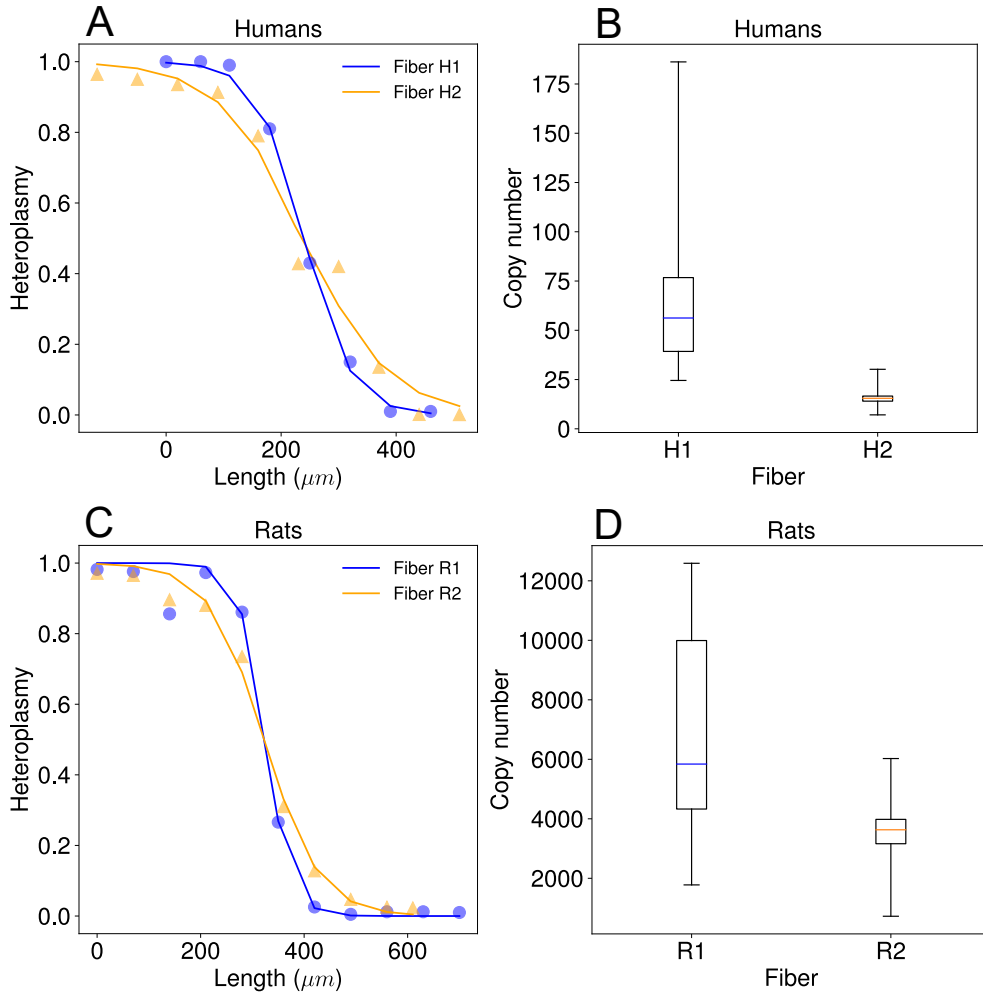


Figure 3.2: *A steeper wave of mutants propagates more slowly in fibres that have a higher copy number per unit length, in agreement with the predictions of stochastic survival of the densest.* Data are from previous experimental studies^{16,17}, our analysis is in Section B.7. (A) Significant difference (disjoint confidence intervals) in the steepness τ of wavefronts of two human muscle fibres H1 and H2: $\tau_{H1} = (2.46 \pm 0.11) \cdot 10^{-2} / \mu\text{m}$, 99% CI = $[2.13, 2.90] \cdot 10^{-2} / \mu\text{m}$ for H1 (blue, steeper) and $\tau_{H2} = (1.36 \pm 0.16) \cdot 10^{-2} / \mu\text{m}$, 99% CI = $[0.97, 1.99] \cdot 10^{-2} / \mu\text{m}$ for H2. MLE fit with curve $1/(e^{\tau x + b} + 1)$ and profile-likelihood CIs, data from Ref. [16]. According to the mathematics of travelling waves (Section B.5), the steeper wave is slower. (B) Comparison between the corresponding N_{ss} of the two fibres H1 and H2. We have found evidence ($p = 10^{-4}$, $d = 1.48$, one-sided Welch's t -test) that the average copy number per unit length (a slice of length $20 \mu\text{m}$ here) in normal regions of the two fibres H1 and H2 (N_{ss} in our model), is higher for the steeper, and hence slower, wave (H1, blue). This is in qualitative agreement with stochastic survival of the densest, which predicts that wave speed decreases with copy number. In contrast, a model based on replicative advantage predicts that speed increases with copy number. (C) Two muscle fibres in rats, R1 and R2, present waves of mutants with significantly different steepness τ of the waveform, as seen from the disjoint 99% CIs: $\tau_{R1} = (3.97 \pm 0.61) \cdot 10^{-2} / \mu\text{m}$, 99% CI = $[2.53, 8.30] \cdot 10^{-2} / \mu\text{m}$ for R1 (blue, steeper) and $\tau_{R2} = (1.88 \pm 0.16) \cdot 10^{-2} / \mu\text{m}$, 99% CI = $[1.46, 2.49] \cdot 10^{-2} / \mu\text{m}$ for R2. MLE fit with curve $1/(e^{\tau x + b} + 1)$ and profile-likelihood CIs, data from Ref. [17]. (D) We have found an indication ($p = 0.06$, $d = 1.00$, one-sided Welch's t -test) that the average copy number per unit length in normal regions of the two fibres R1 and R2, is higher for the steeper wave (R1, blue).

more affected by sarcopenia with ageing^{40,186,193,213–215}. Small, short-lived animals like rodents show sarcopenia on a time scale of years (from ≈ 2 years) and have a larger proportion of Type 2 fibres compared to long-lived animals such as rhesus monkeys and humans¹⁹⁸, that exhibit sarcopenia on a longer time scale (decades). While there are other physiological differences between the fibre types, this is a further link between smaller mtDNA copy numbers and faster mutant expansion.

The leading edge of a faster wave is flatter than that of a slower wave (e.g., Ref. [216], summarised in Section B.5). By exploiting this property, it is possible to compare the speeds of two waves by examining their shapes. Previously published data on muscle fibres for humans¹⁶ (Fig. 3.2A, B) and rats¹⁷ (Fig. 3.2C, D) show that flatter – hence faster – waves of mutants propagate along fibres with smaller copy numbers. All these observations support our model’s prediction of an inverse relationship between copy number and wave speed, opposite to that predicted by a standard replicative-advantage model⁷⁶.

Importantly, the travelling wave of mutants with inverse relationship between speed and copy number is observed not only in the case of linear feedback control (Eq. (2.2)), but also for other controls encoding a higher density of mutants (see Section 2.6), provided that stochasticity and spatial structure (with diffusion) are present.

3.4 Revisited estimates of mutant loads in skeletal muscle fibres

One important question about the expansion of mtDNA mutants in skeletal muscle fibres is the value of the *de novo* mutation rate for deletions. Previous neutral models have required unreasonably high mutation rates to reproduce the observed accumulation of mutants^{100,103}. In this section, joint work with Dr. Hanne Hoitzing, we argue for a different interpretation of available data on mutant loads, that takes into account the spatial extension of muscle fibres. In the next section we show that our model, thanks to the SSD-driven expansion of mutants,

is able to recapitulate observed mutant loads assuming a realistic *de novo* mutation rate.

Most existing modelling studies^{83,103,112} for humans and rodents have tried to match predictions to the following end-of-life target mutant loads: 5–10% fibres must have reached a heteroplasmy level of $\geq 60\%$. We refer to the former quantity as *threshold cell-fraction* (TCF) and to the latter as *threshold heteroplasmy* (TH). The TH of 60% comes from a study on cybrid HeLa cells, stating that a 5196 base pair deletion needs to reach a cell-fraction of $> 60\%$ before the cell shows a reduction in COX activity¹². However, TH is cell-dependent²¹⁷ and the 60% value for HeLa cells needs not be relevant for skeletal muscle fibres in humans and rodents. The TCF of 5%^{83,112} is often justified on the basis of three studies^{48,218,219}. In the first study, 5% is the upper bound of the range (0.1-5)%⁴⁸. The second study found a maximum of 0.37% COX-negative fibres in human limb muscle over a large age range²¹⁸. The last study²¹⁹ did observe high percentages of COX-negative muscle fibres ($\sim 40\%$) but, it refers to a patient affected by a mitochondrial deletion disease, whereas we are interested in muscle ageing in healthy subjects.

We propose new estimates after examining data on skeletal muscle fibres of aged healthy mammals^{14,16,48,193} and considering the spatial structure of muscle fibres. In these studies, hundreds of serial sections of length $\sim 10\mu\text{m}$ were cut along muscle fibres and stained for cytochrome *c* oxidase (COX) and succinate dehydrogenase (SDH) activities. COX is a mitochondrial-encoded enzyme that is fundamental to cellular respiration, and defects in this process are often revealed through COX deficiency. Succinate dehydrogenase refers to complex II of the respiratory chain, the only complex fully encoded by the nucleus. Therefore, negative SDH activities indicate a nuclear defect, whereas normal or hyperactive activities *and* negative COX activities point towards mtDNA defects. An example of these data, from Ref. [16], is Fig. 2A, where additionally the heteroplasmy is measured for each section. Notice that, for simplicity, we have reported the spatial dimension as continuous, while in reality the measurements (the blue circles in the plot) refer to the section as a whole. We suggest a TH of 90%, based on the fact that in Refs.^{14,48} regions of skeletal muscle fibres presenting mitochondrial dysfunctions were shown to contain $> 90\%$ of deletion mutations.

We now propose a new interpretation of the TCF. Previous mathematical studies modelled

muscle fibres as a single units without spatial extension, with an unstructured and well-mixed mitochondrial population around a single nucleus. These models define TCF as the fraction of single-unit systems in which heteroplasmy exceeds TH. Having introduced a spatial model, we give an estimate of TCF by referring to experimental studies measuring heteroplasmy levels along the length of muscle fibres. In Ref. [16], of 10652 human vastus lateralis (VL) muscle fibres of aged (92 years old) subjects 98 fibres presented an abnormal region, approximately 1% of the total. In Ref. [193] 2115 muscle fibres of 34-year-old rhesus monkeys vastus lateralis (VL) were analysed, and 51 abnormal regions were found, corresponding to $\approx 2\%$ of the total. These estimates for rhesus monkeys and humans refer to subjects at the end of their typical lifespans. We define TCF as the fraction of cells (i.e., fibres) that show a macroscopic abnormal region (where $TH > 90\%$) and, on the basis of the above studies, we propose an end-of-life value of $TCF \approx 1 - 2\%$.

3.5 SSD can account for mutational load in short-lived mammals

Previous efforts to establish the ability of neutral genetic models to account for mammalian ageing have often focussed on understanding the compatibility between mutant loads and R_{mut} through mathematical modelling. The approach is to develop a model predicting a certain mutational burden as a function of mutation rate, and then choose the R_{mut} that matches experimentally observed (i.e., target) mutant loads. In Ref. [83] a neutral stochastic model of degradation and replication of mtDNA that used a mutation rate $R_{mut} = 5 \times 10^{-5}$ per mtDNA replication, predicted that $\leq 4\%$ of fibres will be COX-deficient due to high amount of mtDNA deletions by the 80th birthday of the subject, in agreement with experimental observations of healthy aged human muscle^{48,218}. However, it was later argued that this mechanism cannot explain mutant loads for short-lived animals like rodents¹¹², which show similar accumulation patterns on much shorter time scales (≈ 3 years)^{17,45}. For the model to predict the observed mutant loads in short-lived animals such as mice and rats, R_{mut} must be so high that it would

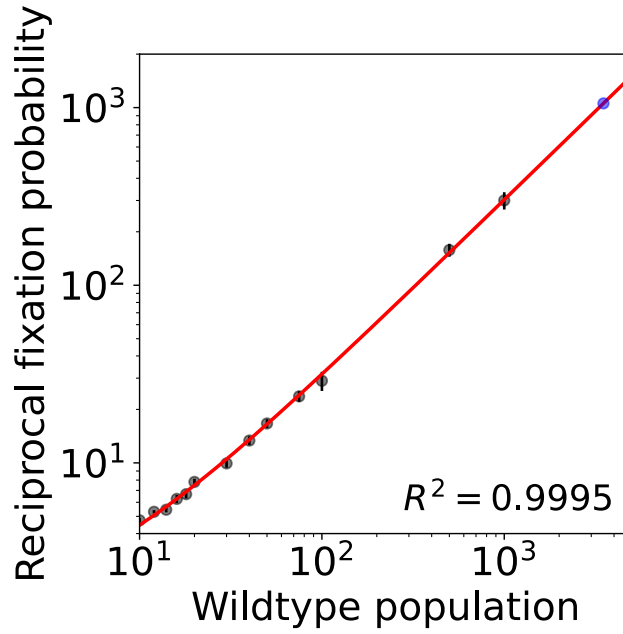


Figure 3.3: *Fixation probability of a founder mutation in a muscle fibre depends on local numbers of mtDNA per nucleus (N_{ss} in our model) and is independent of the total number of mtDNA in the fibre.* The reciprocal of the fixation probability P_f increases linearly with N_{ss} (see Section B.4), and hence $P_f \approx \alpha/N_{ss}$, with $\alpha = (3.316 \pm 0.002)$ for $N_{ss} \gg 1$. The black dots represent estimates of this probability for $10 < N_{ss} < 1000$, obtained via stochastic simulations (details in Section B.4, error bars are standard deviations). For $N_{ss} = 3500$, a plausible value for humans, the predicted fixation probability is $P_f = (9.47 \pm 0.06) \cdot 10^{-4}$ (blue dot).

produce an unrealistically high mutational diversity, whereas experimental studies report a single deletion that clonally expands¹⁰⁰. It has therefore often been assumed that neutral stochastic models cannot explain the mutational load in short-lived animals, and therefore models where mutants have a replicative advantage are needed.

In this chapter we have shown that SSD induces a wave-like clonal expansion of mutants without the need of a replicative advantage. In Fig. 3.3 we have plotted the reciprocal of the probability that a single founder mutation (arising in a muscle fibre modelled as a chain of units with $w = N_{ss}$ wildtype each) takes over the whole system, against N_{ss} . (details in Section B.4). We have simulated data in the range $10 \leq N_{ss} < 1000$, using our best estimates from the literature for the remaining parameter values (Section B.2). The data are well described ($R^2 = 0.9995$) by $a + bN_{ss}$, with $b = 0.301 \pm 0.002$, $a = 1.5 \pm 0.6$ (linear fit). Based on this regression analysis, we estimate a fixation probability for humans of $P_f = (9.47 \pm 0.06) \cdot 10^{-4}$ using $N_{ss} = 3500$ (a plausible value for humans⁴⁴). In Section 3.4 we referenced experimental

studies^{14,16,48,193} reporting that in abnormal regions of muscle fibres, heteroplasmy is higher than 90%. In the following we will consider that abnormal regions have heteroplasmy 100%, i.e., that mutants have fixed. This approximation is not drastic and allows for a mathematically transparent argument.

The fraction of fibres that show an abnormal region – *threshold cell fraction* (TCF), see previous Section 3.4 – can be related to the number M of sufficiently early mutations that occur by the formula $MP_f = \text{TCF}$ (we expect P_f to be very small). By “sufficiently early” we mean that these M mutations arise early enough to have time to take over a macroscopic region of the system and be experimentally detectable, if they have not, alternatively, been lost due to random degradation. In a chain of n units each containing N_{ss} molecules that can mutate at a rate R_{mut} , we expect to see $M = nN_{ss}R_{mut}\Delta T$ mutations arising in a time interval ΔT . If we define that mutations occurring in the first ΔT of life are occurring sufficiently early to be experimentally detectable (if successful), we have

$$R_{mut}nN_{ss}\Delta TP_f = \text{TCF} \implies R_{mut}^{SSD} = \frac{\text{TCF}}{nN_{ss}\Delta TP_f}, \quad (3.2)$$

where by R_{mut}^{SSD} we denote the estimate of R_{mut} based on SSD.

In experimental studies^{16,17,40} muscle fibres are cut into slices of 20 μm each, where heteroplasmy level are then measured. At the speed of approximately 0.2 $\mu\text{m}/\text{day}$ estimated in this study, a travelling wave of mutants would take over a 20 μm slice in less than 1 year. Therefore, a mutation that arises and gives rise to a wave in approximately the first 10^2 years of life (comparable to a lifetime) would still have time to take over a macroscopic region and become detectable. This means that $\Delta T \sim 10^2$. We take the extremely conservative position that it takes an entire lifetime for mtDNA mutations to accumulate to a detectable level and that all mutations that are detected are created in only the first year of life, i.e $\Delta T = 1$ year. We measure ΔT in multiples of the characteristic time for a single mtDNA molecule to replicate. Considering that, on average, a mtDNA molecule replicates every 10 days in a steady state situation (see estimate of μ in Section B.2), $\Delta T = 3.6$ replications. Typical skeletal muscle fibres are 3 – 12cm long, which means $n \sim 1000 - 4000$ (see estimate of L in Section B.2).

Inserting these estimates and $P_f = 9.47 \cdot 10^{-4}$, $\text{TCF} = 0.01$, $N_{ss} = 3500$ in Eq. (3.2), we obtain $R_{mut} = 4.1 \cdot 10^{-8} - 1.6 \cdot 10^{-6}$ per replication, two-to-three orders of magnitude smaller than the estimate in the reference study⁸³, despite our highly conservative assumption on ΔT . The value of TCF is estimated in section 3.4.

The above argument is meant to give an order-of-magnitude estimate, to quantify the intuition that SSD lowers the required R_{mut} needed to observe a given mutant load, in comparison with models that do not exhibit such a marked expansion of mutants. The values quantities in Eq. (3.2) need not be precise estimates to yield an R_{mut} that is much lower than previously thought. Moreover, a less conservative value of ΔT would have led to an even lower value of R_{mut} .

The above argument is valid under the following assumptions:

1. A fibre can be divided into units, the building block of our model. A mutation can arise in a single unit and can expand into neighbouring ones.
2. An abnormal region is caused by a single mutation event. It might be that while a deletion is expanding, an *identical* one arises in the same region of the muscle fibre. We assume that this does not happen in our argument, as mutations are rare event and most of them are lost due to random degradation.
3. A mutation must arise in the first ΔT of life in order to have enough time to expand and become detectable, if successful.
4. A mutation event cannot give rise to more than one abnormal region. Although it may be that distinct but nearby abnormal regions in a single fibre are the result of a single mutation event, this was not observed in the studies we investigated.

In summary, neglecting spatial structure caused early models to miss SSD and to require unrealistically high values of R_{mut} . As SSD increases the probability that a single founder mutation takes over a long fibre, it also dramatically decreases the *de novo* mutation rate required (i.e., number of mutations) to yield observed mutant loads.

More generally, what determines the possible progression of sarcopenia, is not only R_{mut} , as assumed in early studies, but rather the speed of the wave of mutants. Different animals can have different mtDNA turnover rates, copy numbers, type of mutation and diffusion rate that, according to Eq. (2.39) and Eq. (2.40), affect the wave speed. Different species might thus be expected to experience sarcopenia on different time scales. In particular, short-lived animals are predicted to have higher wave speeds for a larger proportion of fibres, since they have more glycolytic, low-copy number fibres¹⁹⁸. For these two reasons – drastically decreased R_{mut} and possibly faster waves – SSD has the potential to account for sarcopenia also in short-lived animals.

3.6 Discussion

We have argued that our model, based on stochastic survival of the densest (SSD), recapitulates the features of the clonal expansion of mitochondrial deletion mutations in skeletal muscle fibres. Ours is the first spatial model to account for the spread of mutant mtDNA, schematising skeletal muscle fibres as a chain of nuclei each surrounded by a population of mitochondria that can move diffusively along the fibre. Previously, the clonal expansion of mitochondrial deletions has been modelled via a replicative advantage^{51,103,112,133,220}. We have discovered that a standard literature-parameterised linear model of this type (Eq. (B.1)) produces an expansion that is 300 times faster than observed speeds. Our model predicts instead a speed that is of the same order of magnitude as observations, and this is robust to the uncertainty in parameter values. To the best of our knowledge, this is the first phenomenon with a compelling description as SSD.

Our phenomenological formula for the speed of the SSD-driven wave (Eq. (2.39)) has implications for therapy, since existing drugs allow us to modulate some of the parameters influencing the propagation of mutants. The SSD wave speed decreases when copy number increases; therefore increasing copy number via interventions such as increased physical exercise^{206,207}, calorie restriction²²¹ or both²²², and administering nicotinamide riboside²²³ or niacin²²⁴ has

the potential to slow down the clonal expansion of mutants. Decreasing mtDNA diffusion (by fragmenting the mitochondrial network) or turnover seems less viable, being associated with negative consequences and health challenges^{225–227}.

Models based on SSD exhibit a wave speed that decreases when copy number increases, unlike models based on replicative advantage. By reviewing existing data on how copy number affects the steepness of the wavefront (which is connected to wave speed) and the likelihood of developing sarcopenia, we have corroborated the SSD-based prediction of Eq. (2.39) and therefore shown evidence that SSD drives the clonal expansion. Future experiments investigating the relationship between copy number and the speed of the expansion (directly or through wavefront steepness) could provide additional evidence in favour of SSD, and possibly against mechanisms based on a replicative advantage. Manipulating mitochondrial turnover via administration of urolithin A^{228,229}, a mitophagy stimulator that increases mitochondrial degradation (μ in our model), could also help probe Eq. (2.39). Finally, we have shown that SSD, together with a re-evaluation of the current estimates of the mutant loads, lowers the *de novo* mutation rate needed to reproduce the observed mutant loads in humans by three orders of magnitude compared to previous models based on neutral stochastic drift^{81–83}. As a consequence, a *de novo* mutation rate lower than previously thought can yield pathology.

Future work might explore the interaction between the model presented here and mitochondrial network dynamics in skeletal muscle fibres, as done in Ref. [220] for a model based on replicative advantage. Moreover, our model may be applicable to a wider range of cells in the presence of sub-cellular control on sub-populations of mitochondria. Neurons might be a candidate, although their directed mitochondrial trafficking²³⁰ is different from random diffusion. There is also evidence of exchange of mtDNA between cells²³¹, but the scale and nature of this *in vivo* is under debate.

Our core claim is that the expansion of mtDNA carrying deletions in skeletal muscle fibres is not necessarily driven by a replicative advantage, or by conventional neutral drift. Rather, mtDNA deletion mutants might have an effective selective advantage induced by SSD, namely by the combined effect of stochasticity, higher density and spatial structure with diffusion.

Chapter 4

Estimating COVID-19 vaccine effectiveness

Summary

In this chapter, I present my ongoing work on estimating the effectiveness of two COVID-19 vaccines from data collected within the SIREN study. The chapter starts with a concise introduction to survival analysis, with a focus on the Cox proportional hazards model. The connection between proportional hazards models and counting processes is also explored, with an example of inference on data from a forward stochastic simulation analogous to those used in the previous chapters. After giving more details on the SIREN study, I then discuss the results of my research on SARS-CoV-2 vaccine effectiveness and protection from previous infection, for the Delta (B.1.617.2) and Omicron (B.1.1.529) variant. I also review challenges and potential biases of the analyses.

4.1 Introduction

On 31/12/2019, the Chinese authorities informed the World Health Organisation (WHO) that “cases of pneumonia of unknown etiology” had emerged in the country. Days later, Chinese scientists identified a new coronavirus, called *severe acute respiratory syndrome coronavirus 2* (SARS-CoV-2), causing a disease named COVID-19. On 20/01/2020, the first COVID-19 patients tested positive in the UK. The WHO declared a pandemic on 12/03/2020. The first UK-wide lockdown was ordered on March 23/03/2020. As of November 2022, there have been more than 625 million *confirmed* cases of COVID-19 and more than 6.5 million related deaths worldwide, of which more than 23 million cases and 190,000 deaths in the UK (data from <https://covid19.who.int>).

Since the early days of the pandemic, scientists have asked if COVID-19 infection protects from future reinfections, and for how long. One of the scientific studies aimed at answering this question is the SIREN (**S**arscov2 **I**mmunity **R**Einfection **E**valuation**N**) study, part of the UK national response to the pandemic and led by Public Health England (now UKHSA). SIREN was launched in June 2020, is a multi-center, prospective cohort study that enrolled healthcare workers in the United Kingdom. With the rollout of vaccines, which started on 8/12/2020 in the UK, SIREN’s primary goal became estimating protection from vaccination, previous infection or both.

I have worked on estimating these protections, having been involved in the SIREN study since July 2021 as the lead statistician of the team. This chapter is about my ongoing research on the topic, which uses methods from survival analysis on data from the SIREN study. The plots and tables in this chapter and in Appendix C are joint work with Sarah Foulkes and Victoria Hall, the joint first authors of¹²⁴. We discussed at length how to best present the results in each case. Therefore, for me The COVID-19 pandemic led to an opportunity to explore complementary mathematical and scientific areas in my PhD, and most importantly to give a contribution to the response to this global crisis.

4.2 Overview of survival analysis

In the introduction to the thesis we gave an overview of the applications of survival analysis, hinting at its connections with stochastic processes. What follows is a succinct and conceptual introduction to survival analysis, based largely on Refs. [232, 233], aimed at making this thesis reasonably self-contained. We start by introducing censoring, a defining feature of survival data. We then lay out the mathematical foundations of survival analysis, and introduce non-parametric, parametric and *semi*-parametric survival models, with an emphasis on the Cox proportional hazards model that we used in our work on COVID-19 vaccine effectiveness. More details can be found in the textbook Ref. [234].

4.2.1 Censoring is a defining aspect of survival analysis

Suppose you have two populations of bacteria from two different strains on two petri dishes, and that you want to investigate and compare how long each population can survive after introducing a lethal chemical, that we are sure will destroy all bacteria on a time-scale of days. Suppose that you can observe the two populations until all bacteria die out and that you can record the time each bacteria dies. In this setting, you have full information on the *time to event* of interest (death here), or *survival time* T . After you have observed the death of all bacteria, a number of statistical methods can be used to answer the question: “which strain of bacteria has survived the longest?” For example, one could inspect the distribution of survival times for the two strains by plotting histograms, and if normality conditions are satisfied a t -test can be used to compare their means.

Now, suppose you run a study with two groups of subject, one group that previously received the flu vaccine and the other unvaccinated, with research question “is the vaccine effective at preventing flu?” The event of interest here is infection from flu virus. Suppose the study runs for two months and starts at the beginning of the flu season. By the end of the study, it is likely that a) some subjects will have been infected, b) some will have left the study before the event of interest, c) some will have stayed in the study until its end and not get infected. The

survival time T of subjects in a) is the time the spent in the study until infection. Subjects in b) might have withdrawn from the study, or might be *lost to follow-up*, for example they moved country and stopped answering queries or were lost because of an error in the database system. In general, they stopped contributing to the study for a reason, possibly unknown, different from the event of interest. For subjects in b) and c), however, a lower bound l is known such that $T \geq l$; in this case l is the time spent in the study at their last follow-up (the last time subjects were seen, which coincides with the end of the study for c) subjects). This situation, in which the survival time is not known exactly but there is some information about it, is called *censoring*, and is a form of missing data problem. Censoring is a defining aspect of survival analysis, that makes it challenging and interesting, and requires special methods of analysis. This type of censored data, common in clinical and epidemiological studies, makes this situation markedly different from the above stylised example on strains of bacteria, preventing the use of standard graphical methods and statistical tests.

We underline that the above is an example of *right-censoring*: the event did not occur during the observation, we have a lower bound l on the survival time and no information about censored subjects after the end of the observation. A right-censored subject might or might not experience the event of interest after being censored. *Left* and *interval* censoring are also possible. Left censoring happens when the event of interest has occurred before observation starts, so that only an upper bound on the survival time is available. Interval censoring occurs when the time of the event of interest is known to lie in an interval, so that we have a lower and upper bound on the survival time. In this thesis we only deal with right-censored data, the most common form occurring in survival analysis, and we will use “censor” with the meaning of “right-censor.”

Standard survival analysis techniques, including those used in this thesis, are valid under the assumption of *non-informative* censoring. Non-informative censoring means that censored subjects have a censoring time that is statistically independent from failure time. In other words, we are assuming that knowing that a subject is censored does not give us any information on the likelihood and time of the event of interest. An example of informative censoring would a study investigating the time and occurrence of complications after surgery in hospital patients,

where discharged patients are right-censored. Being discharged is usually associated with good health conditions and a lower probability of complications compared to patients who are still in hospital, therefore censoring at discharge would be informative. Non-informative censoring means that censored and uncensored subjects have comparable survival prospects, or that the study cohort keeps being a representative random sample despite the censoring events. In medical terms, censoring is non-informative if it is unrelated to prognosis. Techniques to deal with or control for informative censoring include imputation for missing data and sensitivity analyses with best and worst case scenarios²³⁵.

4.2.2 Mathematical foundations

Two central concepts in survival analysis are the *survival probability* and the *hazard function*. Considering the time T of the event of interest a random variable, the survival probability $S(t)$, also called survival function, is the probability that the T is later than some time t , namely:

$$S(t) = P(T > t), \quad (4.1)$$

where “P” stands for probability. It is usually assumed that $S(0) = 1$ and that $S(\infty) = 0$, and the survival function is non-decreasing.

The hazard function $h(t)$ is the event rate at time t conditional on survival until at least t , i.e. $T \geq t$. Expressing the occurrence of an event at time t as $t \leq T < t + dt$ for infinitesimal dt , we have

$$h(t) = \lim_{dt \rightarrow 0} \frac{P(t \leq T < t + dt)}{dt \cdot S(t)} = -\frac{S'(t)}{S(t)} = -\frac{d}{dt}[\log S(t)]. \quad (4.2)$$

The hazard function is essentially the rate at which the event of interest occurs. It is sometimes called *force of failure* or *mortality*. The hazard rate is non-negative, and, in contrast with the survival function, can be non-monotonic and/or discontinuous. The *cumulative hazard function* $H(t)$ is the integral of the hazard $H(t) = \int h(t) dt$, and from 4.2 one can see that

$$H(t) = -\log S(t).$$

Survival analysis allows us to learn about the survival and hazard function from data. The set of techniques generally fall into three categories:

1. Non-parametric techniques do not make any assumption about the distribution of survival times or, equivalently, about the hazard function.
2. Parametric techniques assume a specific functional form for the hazard function, and hence a distribution for survival times.
3. Semi-parametric techniques do not assume a specific form for the hazard function, but make assumption on how characteristics of the subjects influence the hazard function.

4.2.3 Non-parametric estimate of the survival function

The Kaplan-Meier (KM) method, also called product-limit method, is a way of estimating the survival function non-parametrically from observed survival times, and can accommodate for right censoring. Suppose that events of interest happen at times $t_1 < t_2 < \dots < t_k$ (with $t_0 = 0$), and are assumed to be conditionally independent from one another. Denote by d_j the number of events at time t_j and by n_j the number of subjects under observation just before t_j . $S(t_j)$ can be calculated from $S(t_{j-1})$ as

$$S(t_j) = S(t_{j-1}) \left(1 - \frac{d_j}{n_j} \right), \quad (4.3)$$

with $S(0) = 1$. $S(t)$ is a step function whose value stays constant between any two events, and decreases at each event. Plotting $S(t)$ against time, one obtains the KM survival *curve*. In the absence of censoring, the KM estimate of $S(t)$ reduces to the ratio of subjects with no event at time t to the initial number of subjects. Reconsidering the example of the medical trial of the flu vaccine, one can estimate the survival function of the two groups (patient who received the vaccine and patients who did not) and statistical tests are available to establish if there is

a significant difference between the functions. The most widely used method is the log-rank test²³⁶, that tests the null hypothesis that there is no difference between the populations in the probability of an event of interest at any time point. The log-rank test can be also used in settings where there are multiple groups. An example of an analysis using KM survival curves and log-rank test is shown in Fig. 4.1.

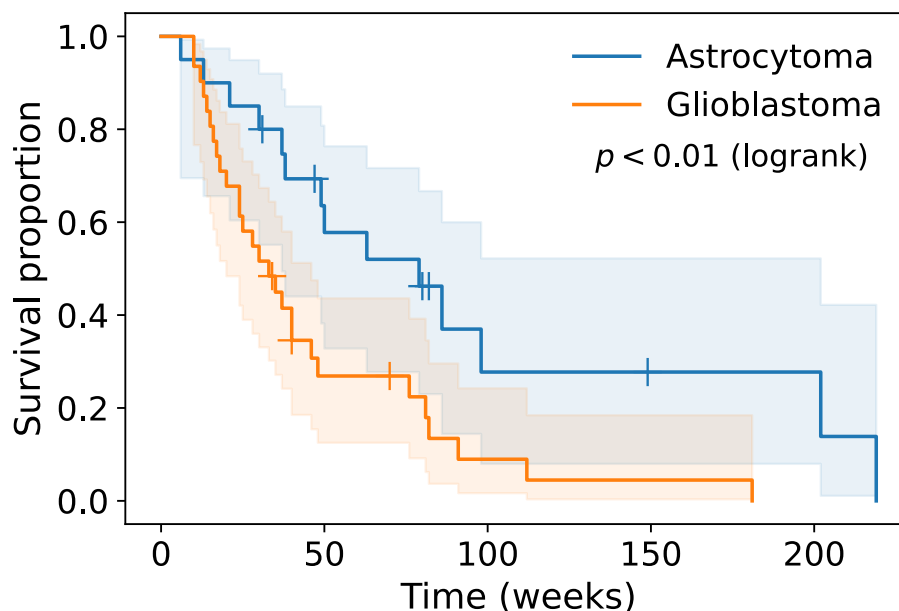


Figure 4.1: Kaplan-Meier survival curves for 51 adult women with recurrent brain tumour, split in two groups based on the type of tumor (astrocytoma and glioblastoma). Event of interest is death. A vertical segment on the curve indicates that an observation has been censored at that time point. Shaded regions are 95% confidence intervals, computed using Greenwood’s exponential formula²³⁷. The two groups can be compared via log-rank test that gives $p < 0.01$ here, i.e. evidence against the null hypothesis that the event of interest at any time point happens with the same probability in the two groups. Data from Ref. [238].

The KM method and log-rank test are versatile in that they do not require knowledge of, or make assumptions, about the shape of the survival curves and the distribution of survival times; they only assume independence of survival times and non-informative censoring. They also allow for risk to vary with time quite generally, e.g. the risk of respiratory infection dropping in the summer. However, they have limitations. First, they cannot provide a clinically interpretable estimate of the size of the difference (if any) between groups, i.e. a directly interpretable effect size (and associated confidence interval). Moreover, the survival prospects between groups might be influenced by other factors: in our flu vaccine example, the probability of infection might depend on the age and occupation of the study participant, not only on whether or

not they received the vaccine. In other words, a multi-variate analysis might be needed, and failure to adjust for relevant patient-related predictors might lead to spurious results, especially in case of confounding. Moving away from non-parametric approaches helps overcome these limitations.

Non-parametric estimates of the hazard function $h(t)$ are less straightforward. For the the *cumulative hazard* $H(t) = -\log S(t)$, integral of the hazard function, the Nelson-Aalen estimator is widespread and well understood, and an estimate of $h(t)$ can be derived by applying a kernel smoother to the increments. Parametric modelling, instead, assumes a parametric form for $h(t)$ and allows the estimation of parameters from data.

4.2.4 Parametric survival models and multivariate survival analysis

Having a fully specified form for $h(t)$ and $S(t)$ makes predictions possible, such as mean and median survival time or the proportion of subjects that have experienced the event of interest at any given time, even beyond the duration of observation. The simplest example of a parametric survival model is that with a constant hazard function: $h(t) = \lambda$. This yields an exponential distribution for the survival times, $S(t) = \exp^{-\lambda t}$, leading to the exponential survival model. Maximum likelihood can be used to estimate λ from data. Other widespread parametric models, with a time-dependent hazard and therefore more complex survival functions, are the Weibull, Gompertz and log-normal models. In the following we will discuss piecewise exponential model, with a piecewise constant hazard. The usual challenge associated with parametric modelling is establishing whether the data are described well enough by the assumed functional form; this can be assessed with the help of non-parametric estimates; model selection can be performed with standard techniques, such as the Akaike information criterion.

A common $h(t)$ for each subject under observation corresponds to an assumption of homogeneity of the population. Importantly, parametric modelling allows the accommodation of multiple factors that might affect the survival prospects of different subjects, whose relevance can be statistically tested. In order to do this, further assumptions on the form of the hazard functions are necessary. Under the *proportional hazard* (PH) assumption, the hazard for a subject i with

a vector of covariates $\mathbf{x}_i = (x_1, x_2, \dots, x_p)$ is

$$h(t|\mathbf{x}_i) = h_0(t) \exp(\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p), \quad (4.4)$$

where the coefficients β_i s are the parameters of interest to estimate, in addition to the parameters of $h_0(t)$. The hazard is factorised into the *baseline* hazard $h_0(t)$ and a term that describes the effect of the covariates. Notice that when $\mathbf{x}_i = \mathbf{0}$, $h(t|\mathbf{x}_i) = h_0(t)$, hence the name of baseline hazard for $h_0(t)$. A PH model is in essence multiple linear regression of the logarithm of the hazard, with the baseline hazard that can be viewed as a (time-varying) intercept. The quantity $\exp(\beta_j)$ is called the *hazard ratio* (HR), as it coincides with the ratio of the hazards of two subjects that differ only for a unit value of the covariate x_j . An HR of $\exp(\beta_j) > 1$ (equivalent to $\beta_j > 0$), indicates that as the value of x_j increases, the hazard increases and thus the length of survival decreases. For instance, if x_j is age, then HR of $\exp(\beta_j) = 1.03$ means that the hazard increases by 3% for every additional year of age of the subject. For a binary variable, e.g. $x_j = 1$ for smokers and $x_j = 0$ for non smokers, $\exp(\beta_j) = 0.54$ would mean that smoking is associated with a 46% reduction in the hazard. In this sense, the coordinates act multiplicatively on the $h(t)$. The name “proportional hazards” is used because the hazard for a subject with covariates $\mathbf{x}_i$ is a constant (time-independent) multiple of the hazard for any other subject with a different covariate vector, with proportionality constant given by a product of HRs. Importantly, the effect of covariates on the hazard is independent of time. It is necessary to verify that the PH assumption holds before using a PH model; statistical tests are available, in the following we will apply a test based on Schoenfeld residuals²³⁹.

An alternative to PH models are *accelerated failure time* models, which assume that the effect of a covariate is to accelerate or decelerate the life-course of a subject, i.e. shrink or dilate the time axis acting multiplicatively the time variable in $S(t)$. In this work we will not focus on accelerated failure time model.

4.2.5 Likelihood and censoring

It is interesting to see how the defining features of survival analysis, censoring, is reflected in the construction of the likelihood function. Here we give an example in the case of right censoring, using the definitions given in 4.2.2.

In random censoring, for each individual i in the population there is a potential lifetime T_i and a potential censoring time C_i . For each unit, we observe either a death (failure) or a censoring and we know which of the two events it is. In mathematical terms, we observe $Y_i = \min\{C_i, T_i\}$ and an indicator variable d_i , that takes the value 1 if unit i failed and 0 otherwise. We assume that these random variables are independent. If the individual i dies at time t_i , then the contribution to the likelihood is $L_i = S(t_i)h(t_i)$, namely the unit survived until time t_i and then died. If the unit is still alive at time t_i , then, under the assumption of non-informative censoring, we only know that $T_i > t_i$; therefore the contribution of a censored observation to the likelihood is $L_i = S(t_i)$. Recalling the meaning of d_i as the indicator variable for the death event, the likelihood can be written as

$$L = \prod_{i=1}^n L_i = \prod_{i=1}^n h(t_i)^{d_i} S(t_i). \quad (4.5)$$

Recalling that for the cumulative hazard $H(t)$ we have $H(t) = -\log S(t)$, and taking logarithms:

$$\log L = \sum_{i=1}^n [d_i \log h(t_i) - H(t_i)]. \quad (4.6)$$

The expressions in Eq. (4.5) and Eq. (4.6) represent the likelihood and log-likelihood in the presence of right censoring. Let us now see an example of the maximum likelihood estimation (MLE) procedure for the exponential survival model. As mentioned in 4.2.4, in the exponential survival model one assumes a constant hazard function $h(t) = \lambda$, hence $H(t) = \lambda t$. Using these in Eq. (4.6) we obtain

$$\log L = \sum_{i=1}^n [d_i \log \lambda - \lambda t_i].$$

By denoting $D = \sum_{i=1}^n d_i$ (total number of failures) and $T = \sum_{i=1}^n t_i$ (total follow-up time),

the log-likelihood can be written as

$$\log L = D \log \lambda - \lambda T. \quad (4.7)$$

Differentiating with respect to λ and setting to zero, the MLE $\hat{\lambda}$ for λ is

$$\hat{\lambda} = \frac{D}{T}, \quad (4.8)$$

namely the total number of deaths over the total follow-up time. In the absence of censoring, $D = n$ and the MLE of λ reduces to the reciprocal of the sample mean, which is the familiar MLE estimate for the rate of the exponential distribution.

4.2.6 Semiparametric Approach: Cox proportional hazards model

In fitting proportional hazards (PH) models, researchers are often interested exclusively in the parameters β_j , and the corresponding hazard ratios (HR) $\exp(\beta_j)$. The initial aim of the SIREN study, e.g., was to establish if a previous SARS-CoV-2 infection is associated with protection from future reinfections¹⁴⁵ and, if so, quantify this protection. A problem of this kind can be approached using a PH model where the event of interest is “SARS-CoV-2 infection” and with a binary predictor x_j , such that $x_j = 1$ if the participant was previously infected and $x_j = 0$ otherwise. An HR of $\exp(\beta_j) < 1$ would mean that previous infection is associated with a reduction in the hazard (hence probability) of infection.

David Cox^{240,241} showed that, assuming proportional hazards as in Eq. (4.4), it is possible to estimate the HRs without making any assumption on the baseline hazard $h_0(t)$, that need not be estimated. This is possible because the PH assumption in (Eq. (4.4)) gives rise to a likelihood function where the baseline hazard $h_0(t)$ cancels out^{240,241}. The resulting function, termed *partial likelihood*, depends only on the coefficients β_i s that are estimated by maximisation.

The Cox PH model, often called simply Cox model, is very popular. It retains the intuitive interpretation of HRs, that are often the relevant information researchers are after. At the same

time, it does not require any assumption on the baseline hazard $h_0(t)$ that might be difficult to model or might be affected by unmeasured covariates. The disadvantage is that it is less powerful in making predictions, as the hazard and survival functions are not estimated, being suitable when researchers are essentially interested in HRs. If a functional form for the baseline hazard is applicable to the problem at hand, then a parametric PH model is more efficient – narrower confidence intervals in the estimates of HRs – and powerful. Moreover, despite the flexibility of an unspecified $h_0(t)$, the non-trivial and strong PH assumption is still central and needs to be tested for each predictor. Below we look at extensions of the PH models, that make the model more flexible and also help deal with predictors that do not satisfy the PH assumption.

4.2.7 Assumptions and extensions of the Cox model

Time-varying predictors. During the observation, the values of covariates of subjects may change. In the flu vaccine example we are using, we might include a binary predictor for whether a subject works in close contact with other people, as the level of daily inter-personal contact might affect the probability of being infected. Some participants might change type of occupation during the study, or switch from office to remote work. In the Cox model a similar situation can be easily accommodated, and this depends on how the HRs are estimated in the presence of an unspecified $h_0(t)$ (partial likelihood, for more detail see Ref. [241]). In the analysis of SIREN data we extensively used time-varying covariates which are piecewise constant, by splitting the follow-up time into chunks which the covariate is constant. More details and examples will be provided when presenting results. Predictors that vary continuously over time can be included too, as discussed in Ref. [242].

Stratification. The PH assumption implies that the effect of a covariate on the hazard $h(t)$ is constant over time. An example of violation of this assumption is geographical region during the COVID-19 pandemic: the virus started spreading in some countries before others, therefore in modelling data from, say, both China and UK, it would be inappropriate to assume that being in the UK has a time-independent effect on the hazard of being infected. If a categorical predictor

has a time-varying effect on $h(t)$ – which can be tested using Schoenfeld residuals – including it in the Cox model can lead to spurious results. Instead, one can allow for a different baseline hazard $h_{0k}(t)$ for each category k of the variables (where for example k is an index denoting geographical location). This means assuming that a subject with covariate $\mathbf{x}_i$ and in the k th category of the PH-violating variable has hazard $h_k(t|\mathbf{x}_i) = h_{0k}(t) \exp(\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p)$. Estimation and interpretation of the hazard ratios is unchanged, and it is possible to stratify over more than one categorical variable. However, stratification by a non-proportional variable prevents one from estimating whether a covariate is relevant and the strength of the possible association. We used a stratified Cox model for analysis of the SIREN data, stratifying on (among others) the geographic region of the UK.

Time varying coefficients. If a covariate violates the PH assumption and we are interested in estimating its effect on the hazard in a PH-like framework, a possible approach is to introduce a time-varying β coefficient in Eq. (4.4) to model this effect and how it changes over time. For instance, researchers might study the efficacy of a novel antibiotic against a bacterial infection, with a group of patients administered the antibiotic and a control-group treated with the standard of care. The event of interest might be an adverse effect of the infection, e.g. sepsis. The efficacy of the novel drug might change over time due to potential antimicrobial resistance. If the novel drug might lose its efficacy, then it is crucial to know when and how fast the drug becomes ineffective. It is not possible to stratify on the type of drug administered, as this would prevent the study from addressing the main research question. The change in efficacy would also not be modelled by the varying baseline hazard rate, as the novel drug is administered only to one group. The Cox model with a time-varying coefficient for the treatment difference may shed light on these issues and help in exploring optimal treatment strategies taking into account the potential of acquired resistance to the drug. We did not use this extension of the Cox model in our work within the SIREN study so we will not focus on it. More details and worked examples can be found in Refs. [243, 244].

Finally, we mentioned that a generic PH model can be seen as a multivariate linear regression of the logarithm of the hazard. This assumption must also be tested, and a widespread approach is plotting the Martingale residuals against continuous covariates. We will not focus on this

aspect, as in our applications we used only categorical covariates.

4.3 Proportional hazards models and the Poisson Point process.

The general proportional hazards (PH) model of Eq. (4.4) reduces to the exponential survival model if we assume constant baseline hazard and same hazard for all subjects. Let us reconsider Eq. (4.7), the log-likelihood of the exponential survival model (i.e. constant hazard rate). Up to a constant term that does not depend on λ ,¹ this expression is the log-likelihood obtained assuming the total number of failures $D = \sum_{i=1}^n d_i$ follows a Poisson distribution with mean λT , with $T = \sum_{i=1}^n t_i$ being the total exposure time²⁴⁵. Therefore, modelling the number of failures as Poisson random variable with mean λT is equivalent (same estimate of λ) to modelling the exposure (censoring or survival) times as observations from an exponential distribution with rate λ . In other words, the failures and censoring governed by an exponential survival model can be modelled via a Poisson point process.

These observations point to an interesting connection the Poisson point process and PH models in survival analysis, which is not restricted to the special case of constant hazard. While the assumption of constant baseline hazard is restrictive, a much more flexible assumption is that of a piecewise (PW) constant baseline hazard, corresponding to a PW exponential baseline survival function. Splitting the observation time with cut-points $0 = \tau_0 < \tau_1 < \dots < \tau_M < \infty$, we can write

$$h_0(t) = \lambda_j, \quad t \in [\tau_{j-1}, \tau_j), \quad (4.9)$$

with parameters λ_j to be estimated. The number and locations of the cut-points can be seen as hyper-parameters. With appropriate choices, this type of model can be very powerful in approximating baseline hazard. The above correspondence to the Poisson distribution is generalised to this case by considering a Poisson random variable for each interval $[\tau_{j-1}, \tau_j)$,

¹Recall that d_i is the indicator variable for the event “failure of unit i ” and t_i is the corresponding survival or censoring time.

with outcome being the number of failures occurred in that interval²⁴⁶. Intuitively, the hazard ratios $\exp(\beta x)$ of the full hazard function do not enter the estimation of the rates λ_j (and are estimated independently) due to the factorised form of the general PH model (Eq. (4.4)).

In Cox PH model, no assumption is made on the baseline hazard; the model is therefore non-parametric in term of the baseline hazard. A sufficiently flexible PW constant hazard, with enough cut-points and sufficiently small intervals, can also be seen as a non-parametric model. Interestingly, a PW exponential model where the baseline hazard is constant *between consecutive failure events* returns exactly the same estimates of hazard ratios as the Cox PH model^{247,248}, that is therefore a limiting case of the PW exponential model. In summary, the Cox PH model is a special PW exponential model, and the exponential survival model is closely related to the Poisson process. This connection has been formalised^{120–122} and has brought about theoretical developments in the field of survival analysis²⁴⁹.

4.3.1 Example: Cox regression on survival times generated from a Poisson process

Here we exemplify the connection between the Poisson process and Cox PH models with simulated data. We simulate survival times from an inhomogenous Poisson point process with deaths only and two species, wildtypes and mutants. The chemical reaction network is simply



where $h_0(t)$ is the death rate of wildtypes and e^β a parameter that multiplies it to give the mutant death rate. We are under the proportional hazards assumption: the wildtype death rate is the baseline hazard and the parameter e^β is the hazard ratio of the mutant species. The wildtype death rate is $h_0(t) = 2 \cdot 10^{-3} \cdot \max[1, \sin(t/5) + 2 + \mathcal{N}(0, 1/10)]/\text{day}$, sinusoidal with added Gaussian noise and bounded below by a positive value. We start with 10^4 wildtypes and

10^4 mutants and simulate the process for a total time $T_{tot} = 150$ days, right censoring particles that are still in the system at T_{tot} . We repeat the procedure 40 times, each time with a different value of e^β , going from 0.05 to 2 in equally spaced steps, and each time estimating e^β via Cox regression.

In Fig. 4.2 we see that the Cox estimates of the hazard ratios closely follow the values set in the simulation, the plot being a straight line without intercept with slope $= 1.003 \pm 0.003$ and $R^2 = 1.000$. We also fit a straight line with an intercept, verifying that the intercept is not significantly different from zero ($p = 0.96$) and obtaining almost identical slope and goodness of fit. The choice of a noisy and oscillating $h_0(t)$ emphasises the independence of the Cox estimation procedure on the form of the baseline hazard. We have also tried other forms, such as constant and quadratic functions with and without noise, with analogous results.

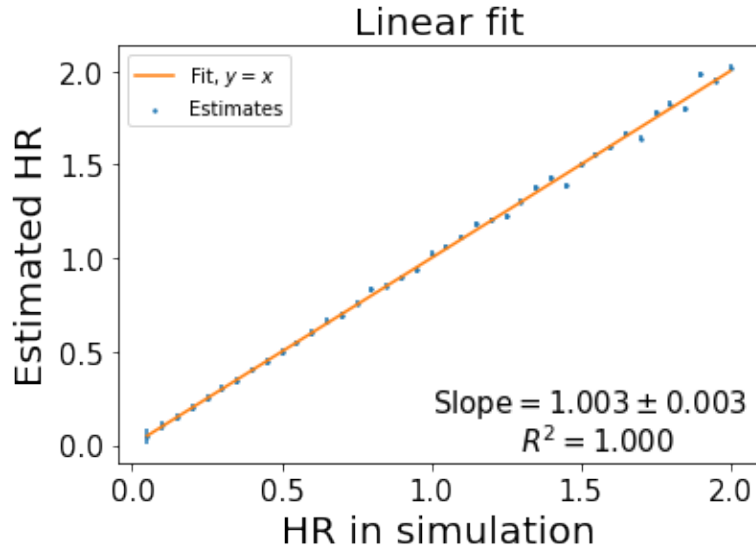


Figure 4.2: *Cox regression accurately estimates ratio of wildtype to mutant death rates used to generate survival times via Poisson point process.* Plot of hazard ratios (blue dots) estimated via Cox regression against the values of the parameter e^β set in the simulation of the Poisson process of Eq. (4.10). Error bars are standard error. The estimates closely follows the parameter values set in the Gillespie simulation (linear fit $y = x$), exemplifying the connection between point processes and proportional hazards models.

4.4 Protection associated with primary course and/or primary infection

4.4.1 Background

In December 2020, the first two vaccines against COVID-19 were approved in the United Kingdom, namely the BNT162b2 vaccine, manufactured by Pfizer–BioNTech, and the ChAdOx1 nCoV-19 vaccine, manufactured by AstraZeneca. These two vaccines became the most widely used in the vaccination campaign, including in the vaccination of the SIREN cohort. From now on we will refer to them as Pfizer and AZ vaccine. The research presented here is available at Ref. [124], in which I was a joint first author and led the data modelling and presentation of results.

In the Introduction to the thesis we argued for the importance of real-world studies of vaccine effectiveness, sometimes referred to as “phase 4” trials. In the current COVID-19 pandemic, real world data on vaccine effectiveness is especially important for a number of reasons. For a rapidly-mutating RNA virus like SARS-CoV-2, with different variants outcompeting the previous dominant strain on a time-scale of months, it is essential to estimate the effectiveness of vaccines against each variant. Different intervals between doses are observed to produce different immune responses in terms of levels of antibody produced²⁵⁰, but this does not automatically translates different levels of protections. Finally, assessing *hybrid* (combined) immunity from vaccination and previous infection (PI) has become a central research question, as it is expected that most of the World population will be infected regardless of vaccination, given the high rates of infections in vaccinated individuals (vaccine breakthroughs) seen with the latest SARS-CoV-2 variant.

In this section we apply the Cox model to investigate the effectiveness of one or two doses of Pfizer and AZ vaccines against the Alpha and Delta variant, dominant in the UK between December 2020 and September 2021 (10 calendar months), when most of the SIREN participants received the primary course of vaccination (first two doses).

4.4.2 Data collection and event of interest

The SIREN study is an ongoing, multicentre, prospective cohort study involving 44,546 health care workers (≥ 18 years of age) from 135 publicly-funded hospitals in the United Kingdom. The study received approval from the Berkshire Research Ethics Committee. The participants underwent polymerase-chain-reaction (PCR) testing every 2 weeks if they showed no symptoms of COVID-19. In the case of symptoms, participants were tested immediately and repeatedly, supplemented by lateral-flow testing. Moreover, all participants underwent antibody testing at enrollment, to find if they were previously infected, and on a monthly basis while in the study. The serum samples for the antibody investigation were tested at the U.K. Health Security Agency (formerly Public Health England) central testing laboratory at Porton Down with the use of the semiquantitative Elecsys Anti-SARS-CoV-2 nucleocapsid (N) protein assay and the fully quantitative Elecsys Anti-SARS-CoV-2 spike (S) protein assay (both manufactured by Roche Diagnostics). Every 2 weeks, participants completed questionnaires that included questions about symptoms. Vaccination data (the type of vaccine and dates of administration) were obtained through personal identifiers from each health administration, linked to the national vaccination register for England (Second Generation Surveillance System), and directly from the participants in questionnaires completed every 2 weeks. The dosing interval was categorized as “short” if the second dose was administered up to 6 weeks after the first dose and “long” if the second dose was administered 6 weeks or more after the first dose.

Two key strengths of the SIREN study are the regular asymptomatic PCR and antibody testing, which make missing the infections unlikely. In the event of a false negative of an infected but asymptomatic SIREN participant, antibody testing would still reveal the infection showing seroconversion. Importantly, as the mRNA vaccines (such as Pfizer and AZ) only cause seroconversion for the spike protein S, seroconversion in anti-N antibody can reveal a missed infection even after vaccination. In the following we will see how missing infections can bias results.

Follow-up began on December 7, 2020 (the day before COVID-19 vaccination was introduced in the United Kingdom), and continued until September 21, 2021, a period that covered 10

calendar months. All the participants who were enrolled on or before December 7, 2020, were followed from that date onward. Participants who were enrolled after December 7, 2020, (i.e., those with delayed entry) were followed from the date of their enrollment.

At enrolment, each participant was assigned to one of two cohorts: participants with no history of SARS-CoV-2 infection (the previously uninfected cohort, *naive participants*) and those who had ever received a PCR test result or an antibody test result consistent with previous SARS-CoV-2 infection (the previously infected cohort, PI cohort). For participants assigned to the PI cohort on the basis of a positive antibody test, the date of the primary infection was set to the date they started experiencing symptoms compatible with COVID-19. Cohort assignments were revised by clinicians working at the hospitals and by clinical scientists and epidemiologists part of the core SIREN team, assessing the historical information for each participant. The primary outcome and event of interest in the statistical model was a PCR-confirmed SARS-CoV-2 infection, irrespective of the participant's symptom status. This outcome was defined as:

- *primary infection* if the participant is in the naive cohort;
- *reinfection* if the participants is in the PI cohort.

Participants who joined the study in the naive cohort and who were infected during follow-up and before vaccination were moved (i.e. “rejoined” the study) to the PI cohort 90 days after their primary infection positive PCR, at which point they were considered at risk of reinfection. The 90-day delay rules out cases of persistently positive PCR tests caused by a single infection episode. For the same reason, for a participant who was in the PI cohort on the basis of positive antibody test results, a positive PCR was considered a reinfection only if occurred ≥ 28 days after the first antibody-positive result. In synthesis, a reinfection is defined as a new positive PCR result at least 90 days after a previous positive PCR or, in the absence of the PCR of primary infection, at least 28 days after the earliest positive antibody result. For individual participants, the survival time was the date of a PCR-confirmed SARS-CoV-2 infection, or they were censored on the date of the last PCR-negative test available. Given the constant

testing schedule, we assume that censoring is non-informative. Most censoring happens with participants who are actively engaged and have not been infected on the last day of analysis time. When participants stopped providing data before that, it means that they withdrew from the study not because of COVID-19 events, that are closely monitored.

Participants were excluded from this analysis if the cohort assignment could not be accurately completed or if the outcome could not be determined (e.g. participants who did not adhere to the testing schedule), or if the date of onset of the PI was not available (i.e. participants assigned to the PI cohort based on antibody results who did not recall COVID-19 symptoms). Moreover, we excluded participants who, before joining the study, had a PI *after* their first vaccine dose; likewise, we did not move to the PI cohort naive participants who were infected during follow-up but after vaccination. This was motivated by the potentially different immune response compared to participants infected before vaccination and the small numbers of participants infected after vaccination, that did not allow us to test for differences. Participants were also excluded if they had received a vaccine other than Pfizer or AZ for any of the two doses, or if they had received one dose of Pfizer and one of AZ, due to the small number of people in these situations. In summary, the analysis estimates protection from infection only, from vaccination only and from infection followed by vaccination in participants administered Pfizer or AZ.

4.4.3 Stratified Cox model with time-varying predictors

We used a Cox PH model where the outcome variable was time to PCR-positive SARS-CoV-2 infection, which is the event of interest. The unspecified baseline hazard of the Cox model is crucial here, as it accounts for the irregularly varying rates of infection in the population during the 10 months of follow-up (e.g. waves caused by Delta and Alpha variant, usual spring and summer lull of a respiratory virus infections). The model was stratified according to age group, geographic region, workplace setting, and frequency of exposure to patients with SARS-CoV-2 in the workplace. We chose stratification based on these categorical predictors because they improved the fit when included in the model (likelihood ratio test) but did not satisfy the proportional-hazards assumption (Schoenfeld test).

The main predictors are time since vaccination (vaccination status) and time since previous infection (PI status). The predictors are categorical. Time since first vaccine dose can take on the values (in days): 0-20, 21-27, 28-41, 42-55, 56+. Time since second vaccine dose can take on the values (in days): 0-13, 14-73, 74-133, 134-193 and 194+. For both, the first category corresponds to the time needed for the full immune response to develop, and we do not focus on the effectiveness of vaccine in this initial period. For the second dose, the other intervals correspond to approximately two months each and the last corresponds to more than six months after the second dose. We also grouped PI status into three categories: before primary infection, 3-12 months after the primary infection, and more than 12 months after the primary infection. The first 90 days (~ 3 months) after a primary infection are excluded, as two positive PCR tests less than 90 days apart are considered persistence of the same infection episode. For a given participant, these two predictors take on the above values as time passes, until infection or censoring (i.e. last negative PCR test available). This is incorporated in the Cox model by splitting the follow-up time of participants into chunks, each with a constant value of the predictor. Each participant can therefore contribute time at risk to more than one category, so we speak of “unvaccinated follow-up time”, or “unvaccinated exposure” “exposure after infection/first dose” and so on. Examples are given in Tables 4.1 and 4.2. We also controlled for gender and ethnic group because these categorical predictors improved the fit (likelihood ratio test) and satisfied the proportional-hazards assumptions. We estimated the HR and report vaccine effectiveness and protection associated with primary infection calculated as $1 - \text{HR}$, along with Wald confidence intervals. The study participants are partitioned by site, namely the hospital where they work. There might be a correlation among workers in the same site, for example if there is an outbreak of infection in a hospital. Therefore we used robust variance estimates, that are unbiased even in the presence of correlation at the hospital level²⁵¹.

We conducted two main analyses:

1. *Protection associated with vaccination only, in naive cohort.* We focussed on all pre-infection (naive) follow-up time, before and after vaccination with Pfizer and AZ. In this analysis, post-vaccination intervals were further categorized according to manufacturer

Start time	End time	Vaccination status	Exposure (days)	Outcome
3	21	Unvaccinated	18	No infection
21	41	0-20 days after dose 1	20	No infection
41	46	21-27 days after dose 1	5	No infection
46	59	0-13 days after dose 2	13	No infection
59	119	14-73 days after dose 2	60	No infection
119	170	74-133 days after dose 2	51	Infection

Table 4.1: *Example of splitting of the follow-up (exposure) of a participant, needed to fit a Cox model with vaccine status as time varying predictor.* These (hypothetical) participants enrolled three days after the start of the analysis time (day 0), received the first dose on day 21 and the second dose on day 46. At day 170, while being between 74 and 133 days after the second dose, the participant was infected (Outcome column). The participants contributed follow-up time (days of exposure) to the listed categories of the “vaccination status” predictor. In our model, vaccination status is further stratified on vaccine manufacturer, that we did not show here for simplicity.

Start time	End time	Vaccination status	PI status	Exposure (days)	Outcome
0	32	Unvaccinated	No PI	32	Infection
32	122	-	0-3 months after PI	<i>Excluded</i>	-
122	150	≥ 56 days after dose 1	3-9 months after PI	28	No infection
150	163	0-13 days after dose 2	3-9 months after PI	13	No infection
163	223	14-73 days after dose 2	3-9 months after PI	60	No infection
223	283	74-133 days after dose 2	3-9 months after PI	60	No infection
283	305	134-193 days after dose 2	3-9 months after PI	22	No infection
305	343	134-193 days after dose 2	9-15 months after PI	38	No infection
343	397	≥ 194 days after dose 2	9-15 months after PI	54	Censoring

Table 4.2: *Example of splitting of the follow-up (exposure) of a participant, needed to fit a Cox model with two time varying predictors.* The (hypothetical) participant was enrolled at the start of analysis time (day 0). Before receiving any vaccine, they had a primary infection (PI) confirmed by a positive PCR test at day 32. The following 90 days were excluded, as two positive PCR tests less than 90 days apart are considered persistence of a single infection episode (Section 4.4.2). Then, at day 122, the participant joined the positive cohort (PI cohort), while being more than 56 days after dose 1. At day 150 the participant received their second dose and then went through successive post-dose-2 vaccine categories. At day 305, nine months passed since primary infection, so the PI status changed. At day 397, the participants was censored: no infection, and last available PCR test. Differently from Table 4.1, here we have split on vaccination status and primary infection (PI) status (time since primary infection, if any). This is needed to fit a Cox model where vaccination status and PI status are both time-varying predictors.

(Pfizer or AZ) and dosing interval (short or long), the latter to explore differences in protection in participants who received the second dose closer in time to their first dose. The baseline group in this analyses was composed of unvaccinated follow-up time, so the HRs are related to the protection (decrease in hazard) associated with vaccination.

2. *Protection associated with PI and hybrid protection.* We focussed on time before any

vaccination and after Pfizer vaccination, and on follow-up time before and after PI. In the latter case the event of interest is reinfection. We fitted a model with combination of the time since PI and the time since vaccination, with the baseline being unvaccinated and naive follow-up time (as in 1 above). This allowed us to estimate protection associated with PI alone and hybrid protection (PI and vaccination). Follow-up time after AZ vaccine and the categorization according to dosing interval for the Pfizer vaccine were not considered because of small numbers in the PI cohort.

Our main results are estimates from the stratified Cox model adjusted for gender and ethnic group, but we also reported estimates from an unadjusted model, without stratifying or controlling for any predictor other than the vaccination and PI status. Estimation was performed optimising the partial likelihood, by using numerical routines of statistical programming languages. Goodness of fit was assessed with the use of the likelihood ratio test (against the null model). The widths of the confidence intervals have not been adjusted for multiplicity of inference. All the analyses were conducted by myself with the use of Stata software, version 15.1 (StataCorp). The results were replicated under my instructions by another member of the SIREN team with independent code written in R, version 4.1.1, survival package v.3.2-13 (R Foundation for Statistical Computing). Annotated code is available at <https://github.com/SIREN-study/SARS-CoV-2-Immunity>.

4.4.4 Results

4.4.4.1 Descriptive results

A total of 44,546 participants were enrolled between June 18, 2020, and April 23, 2021, from 135 sites across the United Kingdom; 35,768 met the inclusion criteria for this analysis. Most participants were women (84%), and the median age was 46 years (interquartile range, 36 to 54). At the beginning of the analysis, we assigned 26,280 participants to the naive cohort and 9488 to the PI cohort. By the end of the analysis, 94.9% of the participants had received two doses of vaccine: 78.5% had received the Pfizer vaccine with a long interval between doses, 8.6% had

received the Pfizer vaccine with a short interval between doses, and 7.8% had received the AZ vaccine. Follow-up time varied according to participant, with a total of 7,482,388 participant person-days, of which 998,270 before vaccination and 6,430,118 after vaccination (from the date of the first dose). Primary infections were noted in 2747 participants during follow-up, and reinfections were seen in 210, with cases peaking at the end of December 2020, declining by March and April 2021, and increasing in May 2021, a pattern that mirrored national trends. A detailed description of our cohort with summarising tables, mainly work by Sarah Foulkes, can be found in the original publication Ref. [124].

4.4.4.2 Analysis 1: vaccine effectiveness in naive cohort

Among the naive participants, with no history of previous SARSCoV- 2 infection, two doses of Pfizer vaccine administered with a long interval between doses were associated with a decrease in the risk of infection of 85% (95% confidence interval [CI], 72% to 92%) (i.e., the adjusted vaccine effectiveness in the first two months after the development of the full immune response, 14 to 73 days after the second dose) (Table 4.3, Fig. 4.3A). Over time, the adjusted vaccine effectiveness declined but remained high, at 68% (95% CI, 54% to 77%), 134 to 193 days after the second dose. At a median of 201 days (interquartile range, 197 to 205) after the second dose, we observed evidence of waning of protection, with an adjusted vaccine effectiveness of 51% (95% CI, 22% to 69%). A similar trend was observed in the participants who received a second dose of Pfizer vaccine with a short interval between doses (Fig. 4.3B), with high protection at 14 to 73 days (adjusted vaccine effectiveness, 89%; 95% CI, 78% to 94%) that decreased to 53% (95% CI, 28% to 69%) at a median of 238 days (interquartile range, 220 to 249) after the second dose. We found no significant difference between the Pfizer vaccine participants who had a long interval and those who had a short interval between doses with respect to protection after the second dose, with a hazard ratio for infection of 1.34 (95% CI, 0.58% to 3.10%) at 14 to 73 days with the use of the short interval as the reference group. The adjusted effectiveness of two doses of the AZ vaccine was 58% (95% CI, 23% to 77%) 14 to 73 days after the second dose. The effectiveness did not differ considerably with longer periods of time after the second dose, with overlapping confidence intervals of vaccine effectiveness reflecting the small number

Two doses in naive cohort

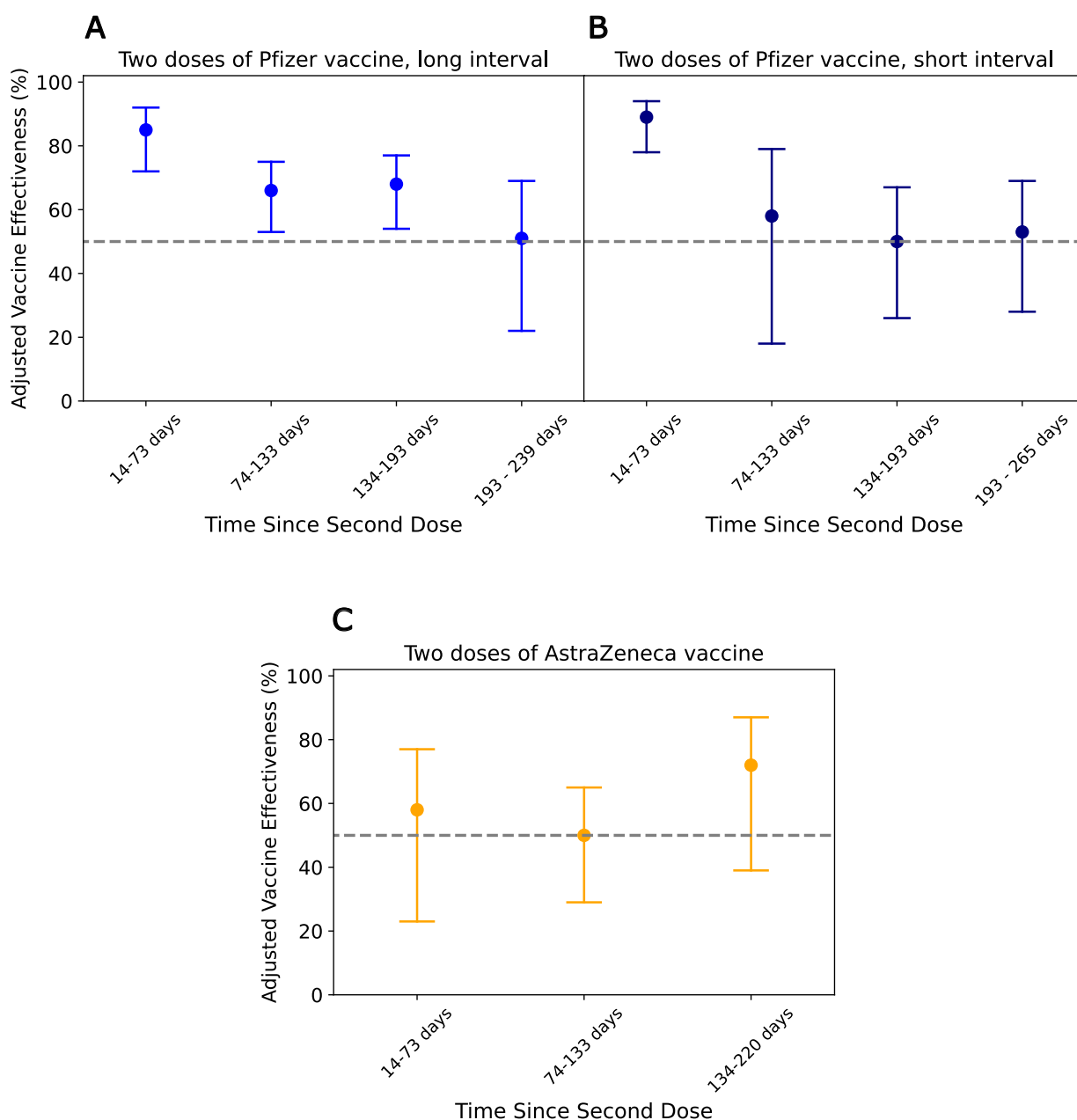


Figure 4.3: *Adjusted vaccine effectiveness over time in previously uninfected participants.* A) Adjusted vaccine effectiveness of two doses of Pfizer vaccine with a long interval between doses, B) Pfizer vaccine with a short interval between doses and C) AstraZeneca vaccine regardless of dosing interval, in participants without previous infection (naive cohort). Data are for the period from December 7, 2020, through September 21, 2021. Bars indicate 95% confidence intervals.

of participants contributing to this estimate (Table 4.3 and Fig. 4.3C). At 14 to 73 days after the second dose, the Pfizer vaccine with a short interval between doses was 74% more effective (95% CI, 36 % to 89 %) and the Pfizer vaccine with a long interval between doses was 65%

more effective (95% CI, 21% to 85%) than the AZ vaccine. The Wald chi-square of the model was 371.46 (31 degrees of freedom), with an Akaike information of 15,367.

Vaccine status	Participants	Days	All infections (symptomatic & asymptomatic)			
			Infections	Crude rate (per 10 ⁴ days)	VE (95% CI)	aVE (95% CI)
Unvaccinated	18094	649643	1038	15.98	ref	ref
Vaccinated 1 dose						
<i>Pfizer</i>						
Dose 1 21d - 27d	15549	102894	52	5.05	0.59 (0.44-0.71)	0.59 (0.42-0.71)
Dose 1 28d - 41d	15247	201531	60	2.98	0.64 (0.47-0.76)	0.66 (0.52-0.76)
Dose 1 42d - 55d	15691	207857	29	1.40	0.71 (0.56-0.81)	0.70 (0.54-0.81)
Dose 1 >56d	16376	341183	53	1.55	0.67 (0.53-0.77)	0.63 (0.46-0.75)
<i>AstraZeneca</i>						
Dose 1 21d - 27d	1471	10204	2	1.96	0.63 (-0.61-0.92)	0.63 (-0.80-0.92)
Dose 1 28d - 41d	1495	20496	1	0.49	0.87 (0.13-0.98)	0.85 (0.16-0.97)
Dose 1 4d - 55d	1494	20445	3	1.47	0.42 (-0.66-0.80)	0.32 (-0.87-0.75)
Dose 1 >56d	1470	38308	10	2.61	0.24 (-0.56-0.63)	0.09 (-0.87-0.55)
Vaccinated 2 doses						
<i>Pfizer long</i>						
Dose 2 14d - 73d	18562	1063102	16	0.15	0.85 (0.71-0.93)	0.85 (0.72-0.92)
Dose 2 74d-133d	17332	950734	264	2.78	0.70 (0.60-0.78)	0.66 (0.53-0.75)
Dose 2 134d-193d	13539	528245	479	9.07	0.73 (0.64-0.79)	0.68 (0.54-0.77)
Dose 2 >193d	2261	20774	81	38.99	0.46 (0.19-0.64)	0.51 (0.22-0.69)
<i>Pfizer short</i>						
Dose 2 14d - 73d	2259	118505	10	0.84	0.85 (0.70-0.92)	0.89 (0.78-0.94)
Dose 2 74d-133d	2238	130389	6	0.46	0.62 (0.19-0.82)	0.58 (0.18-0.79)
Dose 2 134d-193d	2122	118192	47	3.98	0.58 (0.39-0.70)	0.50 (0.26-0.67)
Dose 2 >193d	1706	69352	87	12.54	0.62 (0.45-0.74)	0.53 (0.28-0.69)
<i>AstraZeneca</i>						
Dose 2 14d - 73d	1414	79806	15	1.88	0.52 (0.15-0.73)	0.58 (0.23-0.77)
Dose 2 74d-133d	1213	59593	51	8.56	0.54 (0.32-0.68)	0.50 (0.29-0.65)
Dose 2 >133d	715	16936	26	15.35	0.67 (0.40-0.82)	0.72 (0.39-0.87)

Table 4.3: *Incidence of SARS-CoV-2 infection and effectiveness of COVID-19 vaccines against symptomatic and asymptomatic infection in participants without previous infection, December 2020 - September 2021.* The unadjusted vaccine effectiveness (VE) model was adjusted for the time since vaccination only (combined with the dosing interval and type of vaccine). The adjusted vaccine effectiveness (aVE) model includes additional predictors (gender, or ethnic group) and is stratified across workplace setting, frequency of contact with patients with COVID-19, geographic area of the participant's workplace, and age. The hazard ratios associated with the additional predictors are given in the supplementary tables in Section C.1.1.

4.4.4.3 Analysis 2: protection from PI; hybrid protection

A total of 6169 participants in the PI cohort were followed in the unvaccinated follow-up period and up to 1 year after a primary infection. These participants were predominantly infected in the spring of 2020 and were followed in the period before emergence of the Delta (B.1.617.2)

variant. The risk of reinfection among these participants was 86% (95% CI, 81% to 89%) lower than the risk of primary infection among the unvaccinated participants in the naive cohort (Table 4.4 and Fig. 4.4). There was evidence of considerable waning of protection more than 1 year after infection, with a reduction to 69% (95% CI, 38% to 84%); protection during the first year after infection was 54% (95% CI, 3% to 78%) higher than that after more than 1 year.

In the PI cohort, vaccination is associated with a stronger and longer-lasting protection than infection-acquired immunity only. We used naive and unvaccinated participants the reference group, in order to obtain an absolute estimate of the combined (hybrid) protection. The hybrid protection estimates was more than 90%, after either the first or second dose of vaccination. Waning of protection was not observed more than 1 year after infection or more than six months after vaccination. The Wald chi-square of the model was 789.68 (30 degrees of freedom), with an Akaike information of 14,841.

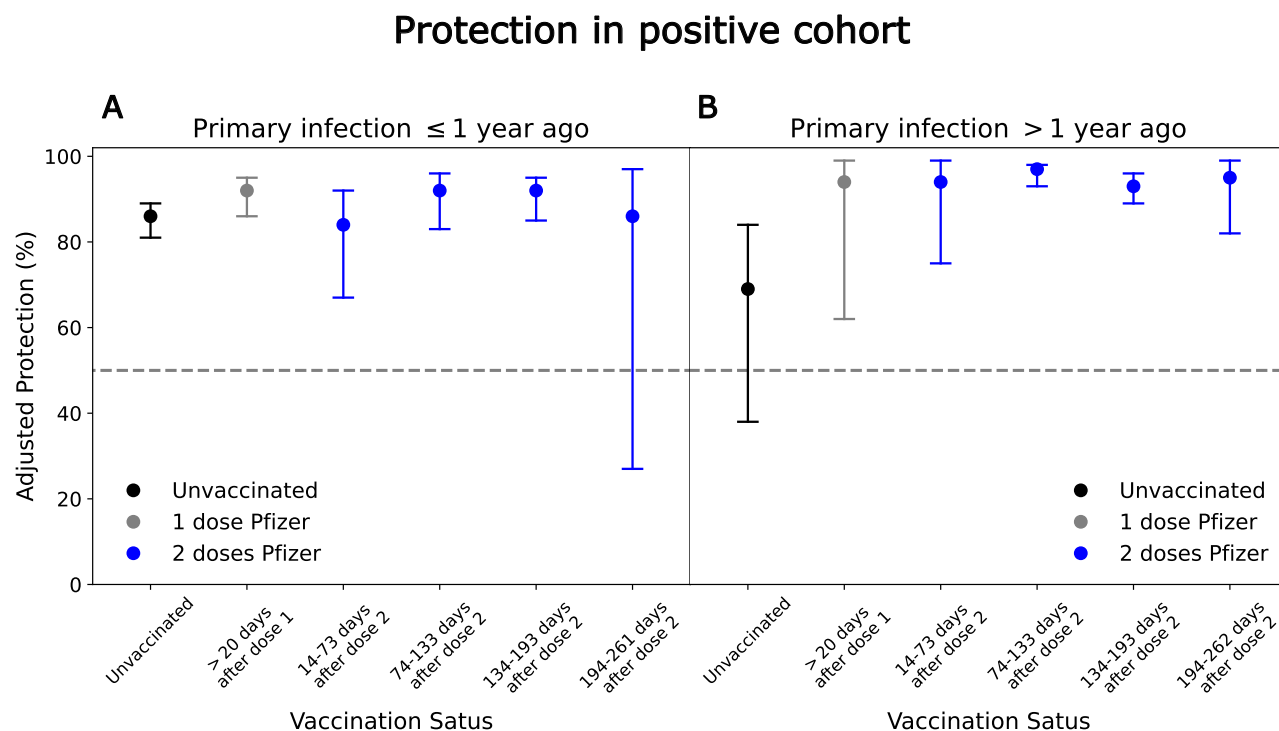


Figure 4.4: Protection associated with primary infection and hybrid protection (vaccine + primary infection) against reinfection up to 18 months after the primary infection. The baseline are participants without a previous infection and without any vaccination, so that the estimates represent absolute protection associated with primary infection only (black) or the absolute hybrid protection (grey and blue). Data are for the period from December 2020 through September 2021 and only for Pfizer recipients. Bars indicate 95% confidence intervals.

Vaccine and PI status	Participants	Days	All reinfections (symptomatic & asymptomatic)				
			Reinfections	Crude rate (per 10 ⁴ days)	VE (95% CI)	aVE (95% CI)	
Primary infection ≤1 year							
Unvaccinated	6169	258088	58	2.25	0.82 (0.76-0.87)	0.86 (0.81-0.89)	
Dose 1 21+ days	7381	303281	13	0.43	0.91 (0.84-0.95)	0.92 (0.86-0.95)	
Dose 2							
14 – 73 days	5075	201580	8	0.40	0.81 (0.60-0.91)	0.84 (0.67-0.92)	
74 – 133 days	2480	119013	12	1.01	0.90 (0.82-0.95)	0.92 (0.83-0.96)	
134 – 193 days	1533	51893	13	2.51	0.91 (0.85-0.95)	0.92 (0.85-0.95)	
>193 days	192	3346	3	8.97	0.75 (-0.19-0.95)	0.86 (0.27-0.97)	
Primary infection >1 year							
Unvaccinated	486	50041	12	2.40	0.71 (0.42-0.85)	0.69 (0.38-0.84)	
Dose 1 21+ days	1642	38422	2	0.52	0.90 (0.60-0.97)	0.94 (0.62-0.99)	
Dose 2							
14 – 73 days	4852	234484	2	0.09	0.93 (0.72-0.98)	0.94 (0.75-0.99)	
74 – 133 days	4970	261549	9	0.34	0.96 (0.92-0.98)	0.97 (0.93-0.98)	
134 – 193 days	3772	137473	18	1.31	0.95 (0.91-0.97)	0.93 (0.89-0.96)	
>193 days	654	15808	2	1.27	0.96 (0.84-0.99)	0.95 (0.82-0.99)	

Table 4.4: *Incidence of SARS-CoV-2 reinfection and protection associated with the Pfizer Vaccine and primary infection (hybrid protection), against symptomatic and asymptomatic reinfection, December 2020 - September 2021.* In order to provide an estimate of absolute protection, we defined the reference group as the unvaccinated participants in the previously uninfected cohort. For the unvaccinated groups we give protection associated with previous infection only. The adjusted protection (aVE) was estimated including additional predictors (gender and ethnic group) and stratifying across workplace setting, frequency of contact with patients with COVID-19, geographic area of the participant’s workplace, and age. The hazard ratios associated with the additional predictors are given in the supplementary tables in Section C.1.1.

4.4.5 Assessing the impact of potential biases

4.4.5.1 Delayed entry of participants

Often, participants enter a clinical study after the natural origin of the time measurement, which might also be different for each participant. In some cases, these two elements might invalidate the assumptions of common survival analysis models. For example, a study might analyse the time until first seizure for patients who suffered brain injury. It is unlikely that all participants are enrolled right after they suffered the injury. If subjects enroll some random time after their brain injury, researchers will observe only survival (or censoring) times that exceed the time to study entry. As a consequence, the observed survival times will not constitute a random sample. In the SIREN study, the natural origin of time measurement can be considered the same for all participants, namely when the virus started circulating in the UK.

In some studies, researchers allow participants to enrol continuously through the duration of the study. A known example is the study of the health effects of the Hiroshima atomic bomb on survivors²⁵². Enrollment started in 1968 but continued until 1992. As in the above example, survival times are biased upwards, as participants must have survived until 1968. Moreover, those who joined after 1968 might have been influenced by their health status²⁵³. People with serious consequences from exposure to radiation might be less likely to participate. In the SIREN study, enrollment started in June 2020 and continued for one year. We tested whether this introduced any bias by repeating the analysis including only participants who enrolled in June 2020 (Section C.1.3). As the results were comparable for all models, we concluded that our analysis is unlikely to be affected by bias due to the delayed entry of participants. If a bias is observed, researchers can resort to modelling²⁵⁴.

4.4.5.2 Depletion-of-susceptible bias

Absolute VE is estimated by comparing the differential rates, or hazards of disease in vaccinated groups to those in unvaccinated groups. Cox model adjusts the estimates derived from these comparisons for a varying and unspecified baseline hazard. If vaccines are effective and the groups are similarly exposed to infection, then the unvaccinated individuals will have higher rates of incidence and will therefore gain infection-acquired immunity (partial or total) at a faster rate. Over time, the rate of infection in the unvaccinated group will decrease. If some infections are missed by researchers, then the unvaccinated group will accumulate, at a faster rate, individuals who are considered susceptible but are in reality protected by a previous infection. As a consequence, the difference in hazards will appear to decrease over time. Equivalently, vaccine effectiveness will appear to wane over time or the waning of protection will be overestimated. Researchers may miss infections in a number of situations, for instance if some infections are asymptomatic and testing is not frequent, if participants do not adhere to a testing schedule or if they are less likely to seek testing if symptoms are mild and similar to those of other conditions.

This phenomenon is called depletion-of-susceptibles^{255,256}. The resulting bias is caused by

the depletion of (highly) susceptible individuals at different rates from the vaccinated and unvaccinated group. This phenomenon can also affect fully randomised studies, and in extreme cases the incidence of the event (infection) of interest in the vaccinated group can seem higher than in the reference group. Recently, this challenge of vaccine research has attracted attention in the context of rotavirus^{257,258} and seasonal influenza^{255,259}

Since waning of VE is one of our major findings, we asked whether this might be caused by a biased estimator or if it might be overestimated due to depletion of susceptibles. In the SIREN study, it is unlikely that infections are missed. Regular (bi-weekly) testing is in place, regardless of symptoms. Participants, who are healthcare workers, are highly likely to seek testing if any COVID-like symptoms occurs. In addition, sera samples are taken on a monthly basis, and the non-vaccine anti-N antibody titres are assessed, that can indicate a past infection even if the participant has been vaccinated. The use of serological data has been recommended²⁶⁰ as a strategy to avoid this bias. Recognising that we cannot completely rule out the possibility that some infections are missed, for example if participants do not adhere to the serology testing schedule, we performed a sensitivity analysis. Our follow-up went from December 2020 to September 2021. We repeated the analysis stopping follow-up time at 31 July, 2021. As we observed similar results to the main analyses, including comparable waning effects (Section C.1.2), we concluded that there is no evidence that our estimates are affected by the depletion-of-susceptibles bias.

4.4.5.3 Previously infected participants without reliable date of primary infection

Some participants (3432) were found to have antibodies associated with a previous infection (anti-N) at enrolment, without a reliable date of infection being known. In our main analysis, we excluded those participants. Given the large number, we performed a sensitivity analysis (Section C.1.4) including them and considering them at risk of reinfection 28 days after their first antibody-positive serum sample. We found that the estimates of protection associated with infection and hybrid protection were not significantly different from the main analysis, with confidence intervals largely overlapping.

4.4.6 Discussion, strengths and limitations

A total of 18 months after the emergence of SARS-CoV-2 and 10 months after the rapid deployment of COVID-19 vaccines, we assessed the durability of protection against SARS-CoV-2 infection conferred by both infection-acquired and vaccine-acquired immunity. Most of our cohort of 26,280 previously uninfected health care workers received two doses of Pfizer vaccine, which was administered with a long interval between doses; this regimen was associated with a considerably reduced risk of infection over the first six months that peaked in the first two months, with an adjusted vaccine effectiveness between 72% and 92%. However, we found evidence of considerable waning of immunity, with protection declining to between 22% and 69% after six months. We found no significant differences in the risk of infection when the Pfizer vaccine was administered with a short or long interval between doses. We found considerably lower protection after two doses of the AZ vaccine than after two doses of the Pfizer vaccine. The period of waning of protection coincided with the period when the Delta variant was the predominant circulating strain; this may account for the more pronounced waning of protection in our cohort, given the reported reduced vaccine effectiveness against the Delta variant²⁶¹. Among unvaccinated participants, the risk of infection was between 81% and 89% lower up to a year after infection among those who were previously infected than among those who were previously uninfected, but we found evidence of waning of protection more than 1 year after infection. Vaccination after previous infection (PI) appeared to boost and extend immunity, and we found no indication of waning of this immunity even more than 1 year after primary infection.

We found that protection conferred by primary infection remained high at up to 1 year but then began to wane. Most follow-up investigations of unvaccinated, previously infected participants occurred before the Delta variant wave, with most of this cohort infected in the spring of 2020 and vaccinated by the end of January 2021. Our ability to study infection-acquired immunity in unvaccinated persons at longer intervals was limited given the very small number of participants in our cohort who remained unvaccinated. It is possible that the sustained infection-acquired protection in our cohort was affected by repeated low-dose occupational exposure to SARS-CoV-

2²⁶² and that it is therefore less generalisable to populations with lower exposure. Because of the limited length of follow-up, it remains unclear how long immune protection will last after PI; however, some studies have suggested that protection could last for up to 61 months, and others have shown protection ranging from 5 to 12 months^{263–266}.

Although infection-acquired immunity is associated with a high level of protection, it wanes after 1 year in unvaccinated persons. In keeping with previous studies, we found an additional benefit of vaccination in previously infected participants^{267–269}, and our finding of high levels of protection associated with immunity from infection plus vaccination has also been observed previously²⁷⁰.

The key strengths of our study include the size of the cohort of participants who underwent frequent testing, independent of disease status, with an average PCR test interval of 16.6 days in the unvaccinated follow-up period and 14.5 days in the vaccinated follow-up period, supplemented by the widespread use of lateral-flow testing, which means we can be confident that most infections were detected. We were able to simultaneously investigate vaccination and PI status in this well-defined cohort and adjust for important confounders, including workplace exposures. Our design overcomes several possible, including underestimation of the proportion of participants with PI. The strengths of our study design and the speed of vaccine deployment probably limited the effect of depletion-of-susceptibles bias^{255–257,271}, which particularly challenges studies on waning of immunity from vaccination, but was not apparent in our sensitivity analysis (Section C.2.2). Importantly, our finding of waning of protection associated with vaccination in naive participants accrues the evidence base on a crucial epidemiological question, with high-quality data.

One limitation of our study is the relatively small number of participants who contributed follow-up data on some important vaccination exposures categories; these participants included those who were unvaccinated, those who received the AZ vaccine, and those who received the Pfizer vaccine with a short interval between doses. This small number of participants particularly affected the precision of our estimates and our ability to assess potential waning after two doses of the AZ vaccine. Given the relatively young and healthy profile of our cohort

and the rarity of severe disease observed in this study, we are unable to assess protection against severe outcomes.

The key messages of this analysis on the SIREN cohort are:

- Two doses of Pfizer vaccine were associated with at least $\approx 85\%$ reduction in the risk SARSCoV-2 infection (asymptomatic and symptomatic) in the short term (two months).
- After six months from the second Pfizer dose, this protection waned to $\approx 50\%$, during a period when the Delta variant predominated.
- Protection associated with two doses of AZ vaccine was $\approx 50\%$, considerably lower than that associated with the Pfizer vaccine overall.
- Protection from reinfection stayed high ($\approx 85\%$) in the first year after a primary infection, with signs of waning after that.
- The highest ($\approx 90\%$) and most durable protection was observed in participants who received one or two doses of vaccine after a primary infection.

4.5 Protection associated with booster and/or primary infection

4.5.1 Background

Results from the SIREN study, including those presented in the previous section, were part of the body of research that led the UK health authorities to roll out COVID-19 booster vaccinations (third doses) in Autumn 2021. The reasons were the strong evidence for waning of vaccine protection against the Delta variant and the expected increase in the circulation of respiratory viruses during winter. The Joint Committee on Vaccination and Immunisation (JCVI) set the priority for booster vaccination on adults aged 50 years and over, and those in clinical risk groups and in front-line healthcare workers, therefore including SIREN participants.

The uptake of the third dose was rapid in our cohort, with 80% of participants having received it in about a month. The vaccination programme was later extended to the entire adult population and some groups of teenagers.

On 26 November 2021 the World Health Organization classified the new B.1.1.529 variant, commonly called Omicron, which was spreading rapidly in a number of countries. From a genetic point of view, the Omicron variant is significant for the numerous mutations affecting the spike protein with respect to the wildtype virus, five-fold compared to the Delta variant. The spike protein mediates entry of SARS-CoV-2 into human cells²⁷², hence it is under evolutionary pressure when most individuals in a population have been vaccinated and/or previously infected. High rates of immune escape by the Omicron variant were anticipated, based on the large number of mutations and preliminary data from South Africa (where Omicron was first detected). In response to this, JCVI the booster vaccination programme was accelerated.

These fast developments had consequences for our work on vaccine effectiveness in the SIREN study, as we set out to estimate protection from boosters. It was estimated²⁷³ that in England the SARS-CoV-2 cases caused by the Omicron variant were half of the total by mid-December 2021 and 97% of the total by 1st January 2022. This rapid take-over by Omicron led us to obtain separate estimates of protection against Delta and Omicron variant. Importantly, it was not possible to obtain absolute vaccine effectiveness estimate, since by the time the boosters' rollout started virtually all participants had received two doses. Therefore we estimates *relative* vaccine effectiveness, namely the additional protection conferred by a booster compared to two doses. A manuscript relative to this analysis is in preparation. As for¹²⁴, I am joint first author and led the data pre-processing, modelling and presentation of results. Preliminary results were orally presented at the 32nd European Congress of Clinical Microbiology and Infectious Diseases (Lisbon, Portugal).

4.5.2 Data and modelling

Most of the data collection and modelling is analogous to the previous section 4.4. The following is meant to be read after 4.4.2 and focuses on the differences from the previous analysis.

We estimate the additional protection conferred by a Pfizer booster for participants who received two doses of Pfizer or two doses of AZ for their primary course. We excluded other boosters (Moderna or AZ) because of very small numbers in our cohort. Baseline groups were participants who had received their second dose 8+ months ago for Pfizer primary course, and 6+ months ago for AZ primary course. For AZ we were unable to use 8+ months due to very small follow-up time in that category, but we verified that using 6+ months for Pfizer gave very similar results. In summary, we estimate the relative protection conferred by a Pfizer booster compared to the waned protection associated with a full primary course (first two doses).

We estimated protection separately against Delta and Omicron. For Delta, we considered data from September 7, 2021 until November 30, 2021. The start date is the day the first participant entered the baseline group, the end date corresponds to the start of the spread of Omicron in the UK. For Omicron, the start day is December 13, 2022 and we stopped collecting data on February 28, 2022. December 13 was chosen because more than half of a sample of the positive PCR tests were caused by Omicron. We did not use data from December 1 to December 12, as Delta was still predominant; we tried to exclude data until January 2022, to minimise coexistence of strands further, but this would have caused a major loss of power in some vaccine status categories.

We defined three post-booster time intervals for the Omicron period: 0-2 months, 2-4 (upper bounds included) and 4+ months. For the Delta period we have a single interval approximately corresponding to the first two months after the booster (median 55 days, interquartile range 21 days), before Omicron set in. We neglected the first 6 days after administration of the booster, during which the immune response develops. We defined 6 previous infection (PI) categories: no primary infection, 3-6 months, 6-9, 9-12, 12-15 (upper bounds included) and 15+ months after PI.

We conducted three main analyses:

1. *Relative protection associated with Pfizer booster, in naive cohort.* We fitted the model in the previously uninfected cohort with vaccination status as a predictor. The relative vaccine effectiveness (rVE) estimates represent the additional protection offered by a

Pfizer booster, relative to a waned primary course. AZ boosters were excluded because of the small number of participants.

2. *Protection from PI.* We then considered both naive and PI participants. We fitted a model where time since primary infection and time since vaccination are added as marginal predictors, without interactions. With this model we estimate protection associated with a previous infection, assuming that it is independent from vaccine status. In order to test this assumption, we fitted a model with interaction between the two predictors and verified that interaction terms are not significant. For this interaction model we split time since primary infection in only two categories, 3-12 and more than 12 months ago, to ensure convergence of hazard ratios on all combinations. We can assume that if interactions are not significant on a coarse scale, then they would not appear in a more granular one.
3. *Hybrid protection.* Finally, we focussed on the hybrid (relative) protection from booster vaccination and previous infection. We focussed on the recipients of the Pfizer vaccine, excluding AZ recipients because of small numbers in the PI cohort. We fitted a model with combination of the time since the primary infection (3-12 and more than 12 months ago) and the time since vaccination (table 3). The baseline here is a waned primary course of Pfizer vaccination without any PI. The HRs from this model represent the combined effect associated with booster and PI.

Each of these three analyses was performed separately for the Omicron and Delta variants, resulting in six models in total. We stratified according to age group, geographic region, and frequency of exposure to hospital patients with SARS-CoV-2. We chose stratification based on these categorical predictors because they were statistically significant when controlled for but did not satisfy the proportional-hazards assumption (Schoenfeld test, according to predictor and global fit). We also controlled for gender, ethnic group and workplace setting, because we observed that these predictors were statistically significant, led to a significant increase in the likelihood value (likelihood ratio test) and Wald statistic, and satisfied the proportional-hazards assumptions. As done in Section 4.4, we report the adjusted estimates (where we control for these additional predictors), and also the estimates from an unadjusted

model, without stratifying or controlling for any predictor other than the vaccination and PI status. In estimating the protection from PI (main model of analysis 2), for the Delta variant we considered additional vaccine status categories (see Section C.2.1 for the full output) in order to have enough follow-up time to obtain estimates. Annotated code is available at <https://github.com/SIREN-study/SARS-CoV-2-Immunity-Omicron>.

4.5.3 Results

4.5.3.1 Descriptive Results

We included a total of 19,614 individuals who were participating in regular SIREN testing between September 2021 and February 2022. The cohort were predominantly female (84%), of white ethnicity (89%), with a median age of 47 (interquartile range 16). On 15 September 2021, 29% had a history of SARS-CoV-2 infection, of which 1,901 (65%) were presumed infected in the wildtype period, 688 (24%) in the Alpha variant period and 8 (0.3%) in the Delta variant period. Vaccine booster uptake was 95% by the end of the analysis period.

During the period of dominance of the Delta variant (07 September 2021 to 30 November 2021, Delta period henceforth) 18,429 participants contributed follow-up time; during and Omicron period (13 December 2021 to 28 February 2022) 16,458 participants contributed. There were 278 primary infections and 85 reinfections during the Delta period, 2467 primary infections and 881 reinfections during the Omicron period. Crude rates of incidence were higher for Omicron (43/10,000-days for primary infection, 32/10,000-days for reinfections) than Delta (4/10,000-days primary, 1/10,000-days reinfections).

4.5.3.2 Analysis 1: additional protection from Pfizer boosters in naive cohort

In the infection-naive cohort, Pfizer boosters increased protection against both variants relative to waned dose-2 (Pfizer or AZ), but less so against Omicron (Table 4.5 and Fig. 4.5). The rVE against Delta was 63% (CI 40%-77%) relative to Pfizer primary course, 77% (CI 37%-91%)

Booster effectiveness in naive cohort

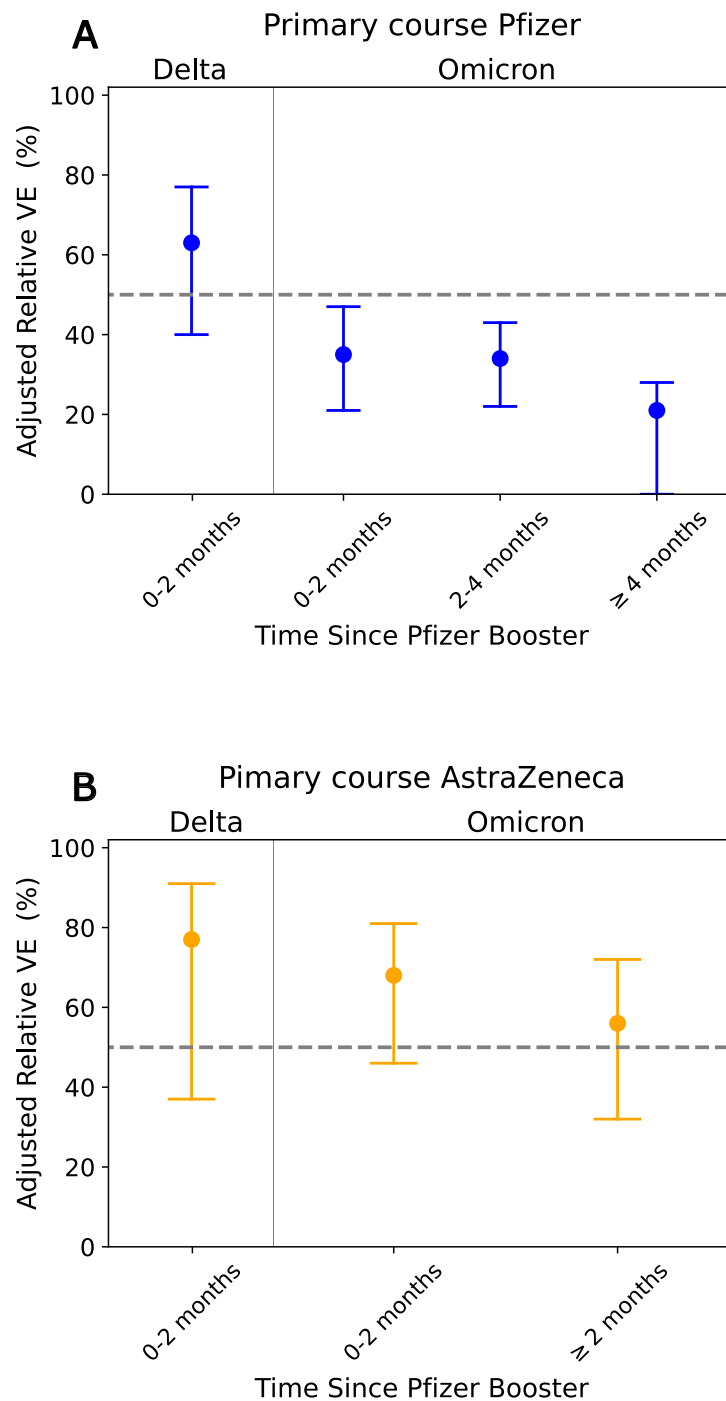


Figure 4.5: *Relative additional protection from Pfizer booster vaccination in participants without a previous infection (naive cohort). Data from September 2021 to February 2022. A) Relative vaccine effectiveness following third Pfizer dose after two Pfizer doses (primary course), against Delta-variant and Omicron-variant infections. The baseline is the waned primary course of two Pfizer doses. B) Same as panel A), but in participants who received two doses of AstraZeneca for their primary course. The baseline is the waned primary course of two AstraZeneca doses. The estimates are not directly comparable between panels, as they represent the boosts in protection with respect to different primary courses.*

relative to AZ primary course. We only have a post-booster time interval in the Delta period, roughly two months after booster. Against Omicron, the boost in protection compared to primary course Pfizer was 35% (CI 21%-47%) in the first two months, 34% (CI 22%-43%) 2-4 months and 21% (CI 0%-38%) 4+ months after booster. Relative to AZ primary course, the rVE was 68% (CI 46%-81%) in the first two months and 56% (CI 32%-72%) after 2+ months.

Vaccine status	Participants	Days	All primary infections (symptomatic & asymptomatic)			
			Infections	Crude rate (per 10 ⁴ days)	VE (95% CI)	aVE (95% CI)
DELTA						
Primary course Pfizer						
Two doses ≥ 8 months ago	1870	31680	29	9.15	ref	ref
Pfizer booster ~ 0 – 2 months ago	11232	500852	177	3.53	0.69 (0.52-0.80)	0.63 (0.40-0.77)
Primary course AstraZeneca						
Two doses ≥ 6 months ago	483	9127	15	16.43	ref	ref
Pfizer booster ~ 0 – 2 months ago	540	18649	7	3.75	0.77 (0.28-0.92)	0.77 (0.37-0.91)
OMICRON						
Primary course Pfizer						
Two doses ≥ 8 months ago	698	20416	151	73.96	ref	ref
<i>Pfizer booster</i>						
0-2 months	4590	68350	249	36.43	0.43 (0.30-0.53)	0.35 (0.21-0.47)
2-4 months	9179	389980	1627	41.72	0.46 (0.37-0.54)	0.34 (0.22-0.43)
>4months	4093	54258	297	54.74	0.40 (0.24-0.53)	0.21 (0.00-0.38)
Primary course AstraZeneca						
Two doses ≥ 6 months ago	143	3954	32	80.93	ref	ref
<i>Pfizer booster</i>						
0-2 months	485	10596	28	26.43	0.66 (0.44-0.80)	0.68 (0.46-0.81)
>2 months	484	20072	72	35.87	0.60 (0.39-0.73)	0.56 (0.32-0.72)

Table 4.5: *Booster effectiveness in the previously uninfected cohort against Delta and Omicron infections, September 2021 – February 2022.* We have provided an estimate of relative protection, defining the reference group as participants previously uninfected who had received only two vaccines, and had received their second dose more than 8 months ago (Pfizer primary course) or more than 6 months ago (AstraZeneca primary course). The adjusted vaccine effectiveness (aVE) model includes additional predictors (gender, ethnic group, workplace setting) and is stratified across age group, geographic region and frequency of exposure to patients with COVID-19. In the post-vaccination intervals, the upper time bound is included. The hazard ratios associated with the additional predictors are given in the supplementary tables in Section C.2.1.

4.5.3.3 Analysis 2: protection associated with primary infection

In agreement with the results in section 4.4, the estimated protection from PI against the Delta variant was high, 90% (CI 42%-98%) 3-6 months after PI, remaining at 85% (CI 81%-89%) ≥ 15 months after PI. Against Omicron, PI protection was estimated at 67% (CI 56%-75%) 3-6 months after, but then decreases substantially, levelling off around 25%-30% from 9 months after PI onward. Figs. 4.6A, B make this difference evident. We report all estimates in Table 4.6. We found no evidence of interaction between protection associated with boosters and protection associated with PI. We ran Wald tests on all interaction terms, finding $p = 0.71$ for Delta period and $p = 0.67$ for Omicron period; every interaction term on its own had $p \geq 0.15$ (not shown).

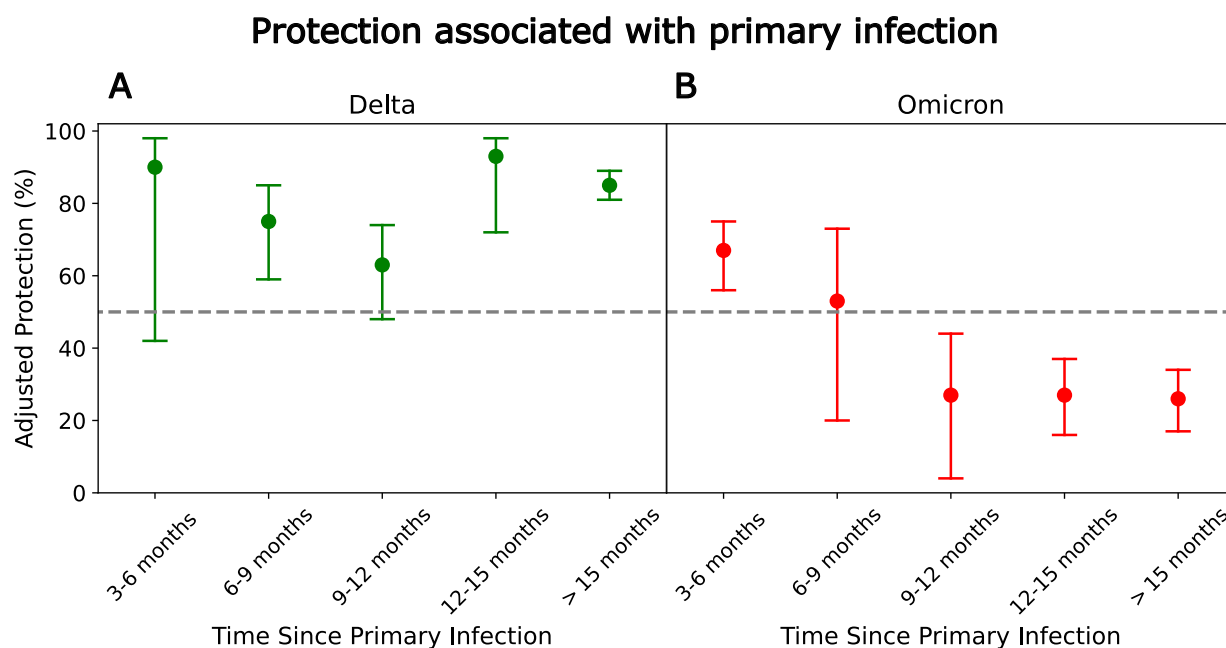


Figure 4.6: *Protection from reinfection associated with primary infection, during Delta-variant and Omicron-variant periods. Baselines are participants without a previous infection. Controlled for vaccine status, but without including interaction terms (found to be not statistically significant, Wald test). A) Protection associated with primary infection against a reinfection with Delta variant, data from September 2021 to November 2021. B) Protection associated with primary infection against a reinfection with Omicron variant, data from December 2021 to February 2022.*

Primary infection status	Participants	Days	All (re)infections (symptomatic & asymptomatic)			
			(Re)infections	Crude rate (per 10 ⁴ days)	Protection (95% CI)	Adjusted Protection (95% CI)
DELTA						
No previous infection	19103	1935033	1125	5.81	Ref	ref
Previous infection						
3-6 months	718	19590	1	0.51	0.90 (0.39-0.98)	0.90 (0.42-0.98)
6-9 months	2006	89006	13	1.46	0.77 (0.61-0.86)	0.75 (0.59-0.85)
9-12 months	2515	143138	34	2.38	0.61 (0.44-0.72)	0.63 (0.48-0.74)
12-15 months	1953	59925	2	0.33	0.93 (0.71-0.98)	0.93 (0.72-0.98)
≥ 15months	4595	398639	35	0.88	0.85 (0.81-0.89)	0.85 (0.81-0.89)
OMICRON						
No previous infection	11074	569635	2467	43.31	ref	ref
Previous infection						
3-6 months	784	30397	47	15.46	0.66 (0.55-0.74)	0.67 (0.56-0.75)
6-9 months	243	5855	14	23.91	0.50 (0.11-0.72)	0.53 (0.20-0.73)
9-12 months	970	23871	80	33.51	0.19 (-0.07-0.38)	0.27 (0.04-0.44)
12-15 months	1660	61798	214	34.63	0.22 (0.10-0.33)	0.27 (0.16-0.37)
≥ 15 months	3099	150276	526	35.00	0.20 (0.10-0.29)	0.26 (0.17-0.34)

Table 4.6: *Protection from reinfection associated with a primary infection (PI). Controlled for vaccine status (not shown) but without interaction, during Delta-variant and Omicron-variant periods.* We have controlled for vaccine status without interaction terms between previous infection and vaccine status, after having verified that interaction terms are not significant (Wald tests: $p = 0.71$ for Delta period, $p = 0.67$ for Omicron period). The adjusted model includes additional predictors (gender, ethnic group, workplace setting) and is stratified across age group, geographic region and frequency of exposure to patients with COVID-19. In the post-infection intervals, the upper time bound is included. The hazard ratios associated with the additional predictors are given in the supplementary tables in Section C.2.1.

4.5.3.4 Analysis 3: hybrid protection

Using participants with two waned Pfizer doses and no previous infections as the reference group, we saw that a PI and a Pfizer booster are associated with stronger protection than booster only (analysis 1 above). Against Delta, a Pfizer booster and previous infection were associated with around 90% additional protection, even more than one year after primary infection (Fig. 4.7A, B). Against Omicron, the combined additional protection remained around 70% for previous infection less than one year ago and around 50% for previous infection more than one year ago. Despite wide confidence intervals in the 4-month post-booster category (reflecting small numbers of participants), there is an indication that the additional protective effect associated with previous infection wanes over time against Omicron (Fig. 4.7B, right). All estimates can be found in Table 4.7.

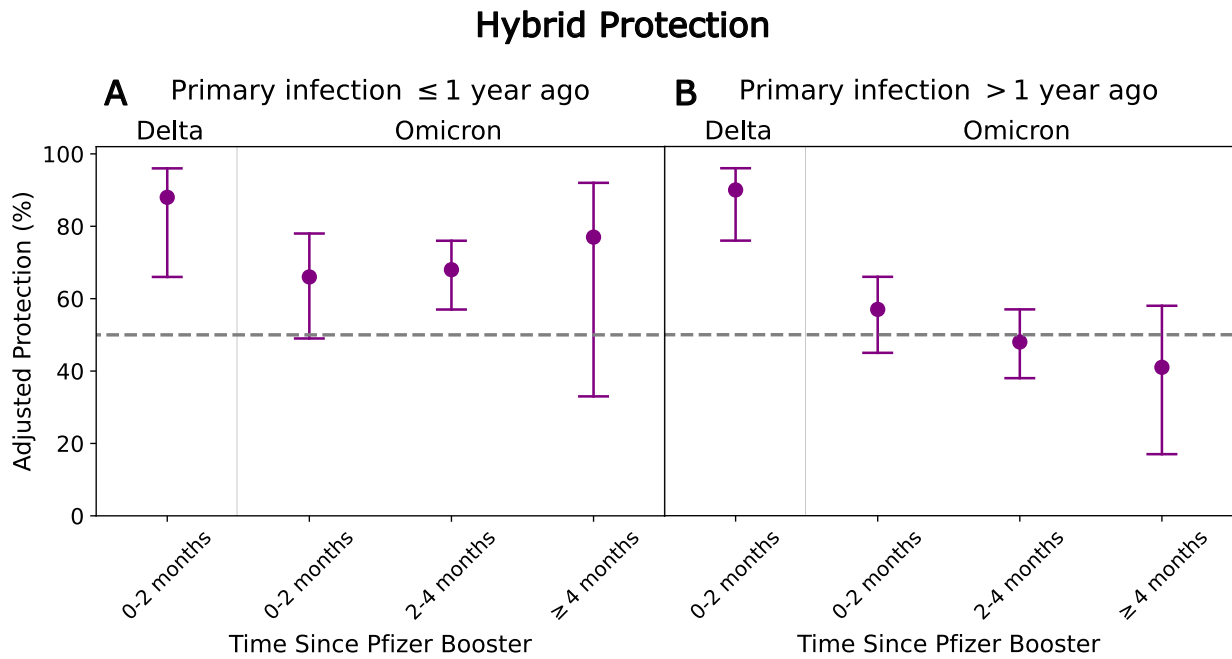


Figure 4.7: *Hybrid protection, from previous infection and Pfizer booster, including only participants who received primary course Pfizer. Baseline is made up of participants without a previous infection and with a waned (8+ months) primary course of Pfizer. A) Hybrid protection in participants who had their primary infection up to one year ago, against Delta and Omicron reinfections. B) Same as panel A), but for participants who had their previous infection more than one year ago.*

Vaccine and PI status	Participants	Days	All (re)infections (symptomatic & asymptomatic)			
			(Re)infections	Crude rate (per 10 ⁴ days)	VE (95% CI)	aVE (95% CI)
DELTA						
Two doses >8 months ago and no previous infection	1870	31680	29	9.15	ref	ref
Infection <1 year ago, booster ~ 0 – 2 months ago	1715	55523	4	0.72	0.94 (0.81-0.98)	0.88 (0.66-0.96)
Infection >1 year ago, booster ~ 0 – 2 months ago	2760	16249	2	1.23	0.94 (0.88-0.97)	0.90 (0.76-0.96)
OMICRON						
Two doses >8 months ago and no previous infection	698	20416	151	73.96	ref	ref
<i>Infection <1 year ago, and booster:</i>						
0-2 months	901	14825	31	20.91	0.70 (0.56-0.79)	0.66 (0.49-0.78)
2-4 months	1080	31993	70	21.88	0.72 (0.63-0.79)	0.68 (0.57-0.76)
>4 months	232	2698	6	22.24	0.76 (0.43-0.90)	0.77 (0.33-0.92)
<i>Infection >1 year ago, and booster:</i>						
0-2 months	1909	34166	94	27.51	0.60 (0.48-0.69)	0.57 (0.45-0.66)
2-4 months	3273	128921	457	35.45	0.54 (0.45-0.62)	0.48 (0.38-0.57)
>4 months	1313	16579	69	41.62	0.53 (0.37-0.66)	0.41 (0.17-0.58)

Table 4.7: *Protection from infection and booster vaccination (hybrid protection) during Delta-variant and Omicron-variant periods, September 2021 - February 2022, Pfizer recipients. We have provided an estimate of relative protection, defining the reference group as participants previously uninfected who had received only two Pfizer vaccines, with second dose more than 8 months ago. In the post-vaccination intervals, the upper time bound is included. The adjusted vaccine effectiveness (aVE) model includes additional predictors (gender, ethnic group, workplace setting) and is stratified across age group, geographic region and frequency of exposure to patients with COVID-19. The hazard ratios associated with the additional predictors are given in the supplementary tables in Section C.2.1.*

4.5.4 Discussion, strengths and limitations

After almost two years from the emergence of the wildtype SARS-CoV-2 virus, the Omicron (B.1.1.529) variant caused a detrimental shift to the epidemiological scenario. Our analysis and concurrent studies found that Omicron partially – but significantly – escapes protection from available vaccines and from previous infection. In individuals without any history of previous infection, the relative effectiveness of three Pfizer doses – compared to two “waned” doses administered 8+ months before – against Omicron was about 35%, significantly lower than against Delta (about 63%). Moreover, 4+ months after a Pfizer booster we did not find a significant difference in protection from two waned Pfizer doses. The additional protection from a Pfizer booster was higher for participants who had received two doses of AZ, but those participants were less protected against infection than recipients of two Pfizer jabs, as found by ourselves (see 4.4) and other studies. Previous infection and three doses of vaccine were associated with higher protection, but still lower for Omicron and especially for early previous infections (more than year before) The findings of decreased vaccine effectiveness against Omicron compared to Delta are common to other studies, which focussed on symptomatic disease and severe outcomes and used alternative designs^{274,275}.

Likewise, we found that a previously infected participant is approximately three times more likely to be reinfected by Omicron than by Delta, when they suffered a first infection 9+ months ago. Other studies also found an increased risk of reinfection associated with Omicron^{276,277}, also in line with the early report²⁷⁸ by the Imperial College COVID-19 response team, that significantly influenced UK health policies in December 2021. We were unable to control for the variant of PI, because of the strong confounding with time since PI. However, in the Omicron period a PI up to 9 month ago (recent) is largely coincident with a Delta infection. These recent infections are associated with a higher protection ($\geq 50\%$) than early (9+ months) infections, associated with $\approx 25\%$ protection. A Delta primary infection might therefore be more protective than Alpha or wildtype against a Omicron reinfection.

One limitation of our analysis is the limited follow-up time in some important vaccine status categories, mainly involving participants having received AZ primary course. We were unable

to assess protection from AZ booster, but AZ boosters are also rare in the general population of the UK. On the contrary, Moderna boosters have been more widely used, but we were still unable to estimate a protection from our dataset because of small sample size. These limitations stem from our participants being administered mainly the Pfizer vaccine, and by cohort attrition. Participants started gradually leaving the study from June 2021 onward, one year after the earliest enrollments. The initial duration of follow-up was indeed set at 12 months from enrollment, with the offer of an extension for 12 more months taken up by approximately half of the participants. The small numbers of participants who remained unvaccinated and the rapid uptake of the booster dose prevented us from estimating absolute vaccine effectiveness of three doses and the effectiveness of two doses against Omicron. A waned primary course was associated with little-to-no protection against Omicron symptomatic infection²⁷⁴. Protection from asymptomatic infection is likely to be very low or null.

The key strength of our work is the ability to assess the effect of previous infection, at a time in the pandemic when the depletion of susceptible individuals²⁵⁶ has a strong potential of biasing estimates. In almost two years since the start of the pandemic, any cohort whose individuals do not undergo constant testing risks accumulating a large number of missed infections, responsible for the bias²⁵⁷. Among our findings, the most sensitive to this bias is the rapid waning of booster effectiveness against Omicron, that we observed 4+ months after the third jab. At the same time, this result is highly important, as reliable estimates of the long-term effectiveness of the third and future doses will be crucial in the long-term management of the pandemic. In our cohort, participants adhered to the scheduled regular testing regardless of symptoms. This is especially relevant for Omicron, which has been associated with less severe or even asymptomatic infections. Moreover, as for the analysis presented in 4.4, we ran the sensitivity analyses described in 4.4.5 and did not find evidence of the depletion-of-susceptible or of delayed-entry bias (Section C.2.2). We also ran sensitivity analyses to assess if delayed entry of participants is biasing our estimates, finding no important difference with the main analysis (Section C.2.3).

Taken together, these findings show a considerable immune escape by the Omicron variant, in line with what was anticipated on the basis of the high number of mutations in the spike

protein. Fortunately, the reduced severity of Omicron infections²⁷⁹ limited the impact on public health. SARS-CoV-2 continues evolving. Following the original Omicron variant (also called BA.1), four more genetically distinct sub-variants have been identified as of September 2022. Preliminary results²⁸⁰ showed that previous infection with an early Omicron sub-variant (BA.1 and BA.2) is associated with high (75% - 80%) protection against the otherwise more infectious BA.4 and BA.5, identified in South Africa in April 2022²⁸¹. In summer 2022, the UK and the USA government agencies MHRA and FDA approved the “updated” boosters, that target the Omicron BA.1 variant spike protein as well as the wildtype.

The complexities of this rapidly evolving scenario underline the importance of continuous monitoring through large-scale epidemiological studies. Given the increasing number of immunisation events (infections or vaccinations) to which many of us have been exposed, the design and analysis of future research studies will be increasingly challenging and nuanced. The potential emergence of immune-escape variants will likely require multiple adjustments to public health strategies. The ascertained waning of vaccine-acquired protection warrants periodic assessments of the need for booster doses, especially in groups that are highly exposed or at risk due to comorbidities, and possibly with updated formulations based on recently circulating variants.

Chapter 5

Conclusions and future work

In this thesis we have:

- Developed stochastic models, based on counting processes, that unearth a novel evolutionary mechanism, termed *stochastic survival of the densest*, and account for the clonal expansion of mitochondrial deletion mutants in skeletal muscle fibres.
- Applied proportional hazards model to data from a large-scale epidemiological study (SIREN), essentially inferring parameters of counting processes, to estimate protection from future SARS-CoV-2 infection associated with vaccination and/or previous infections.

Our results open numerous questions and motivate further experimental work and data analysis. As of October 22, the SIREN study continues collecting relevant high-quality data about participants, and I am currently involved in two lines of research. Below we summarise the results presented in this thesis and review current and future work.

5.1 Stochastic survival of the densest as a new paradigm

In Chapter 2 we have introduced stochastic survival of the densest (SSD), an intriguing evolutionary mechanism whereby a species with higher carrying capacity – denser species – in the

presence of noise can outcompete another and invade a spatially structured system in a wave-like fashion. This noise-driven take-over can happen even if the denser species has a higher death rate, or if the competing species has a higher replication rate; the combined effect of noise, spatial structure and higher density can therefore offset the effect of a lower net growth rate, that instead dominates in a deterministic setting. In our treatment, we have first focussed on a system without spatial structure; although SSD is not observed in an unstructured system, we have derived analytical results and provided heuristic arguments that elucidate the interaction of stochasticity and density differences. Moving to a spatially structured system, we have observed an increase in mean mutant proportion in a minimally extended system and a travelling wave of mutants in a larger system. We have provided heuristic arguments for this expansion and shown that, in continuous space and in some limit of parameter values, our model gives rise to a noise-driven wave equation derived analytically in a recent study⁷⁶. By large-scale numerical simulations we have found a phenomenological equation (Eq. (2.39)) for the wave speed for a wider range of parameter values, and verified that our core results are not dependent on the mathematical details of our model (specific forms of replication rate) and the modelling frameworks (continuous-time counting process). Finally, we have explored the implications of our findings for the field of evolutionary biology, in particular for modelling the spread of altruistic traits, a counterintuitive topic subject to intense debate.

Our work may pave the way for an increased recognition of SSD in evolutionary biology. Enhancing our theoretical and mathematical understanding of this intriguing effect is therefore worthwhile for future applications. An expression for the temporal dynamics of the proportion of mutants (heteroplasmy) in a minimally extended system, such as the two units of Section 2.3, could provide a deeper understanding of the phenomenon by highlighting how the rate of increase depends on model parameters. A line of attack could be finding a conserved quantity in the dynamical system Eq. (2.37) and then applying the techniques in Refs. [75, 140]. An approach based on conserved quantities might be easily generalisable to different functional forms of the replication rate, and possibly to larger numbers of coupled units. Ideally, one would have an expression for how heteroplasmy changes over time in an infinite chain of units. In Section 2.3 we have argued that the increase in mean heteroplasmy is connected to an higher

fixation probability of mutants. An analytical expression for the mutant fixation probability, in two or more units, would also complement our understanding of SSD.

A further line of research aims to obtain an analytical justification for the expression Eq. (2.39), the speed of the wave driven by SSD. In Section 2.2.4 we approximate the chemical master equation by a system of Itô stochastic differential equations, and apply this to the single-unit system in Eq. (2.2.3). This procedure can be applied to a chain of n units, obtaining a system of $2n$ coupled differential equations. At this point one could study the continuous-space limit of this system, which would lead to a stochastic partial differential equation, and test solutions that exhibit a wave-like phenomenon from which a wave speed could be derived. This is likely to require approximations, to be assessed taking into account the ranges of parameter values correspondent to specific phenomena. We have already noticed that the aforementioned system $2n$ coupled differential equations exhibit SSD not only in the form given by the procedure of Section 2.2.4, but also when noise terms simpler than those derived analytically are used (not discussed in the thesis). This simplification could help with analytical progress.

Questions on fixation probabilities are also relevant to understanding the SSD-driven wave. Let us consider a single founder mutation arising in a unit of the chain. For a wave of mutants to arise, the mutation must escape a “stochastic” phase where it can be lost due to (possibly enhanced) degradation, and then expand along the fibre. Once a sufficiently large number of mutations is present in a set of units of the chain, the wave-like expansion can be seen as “deterministic” (despite being SSD-driven), as the extinction of mutants becomes very unlikely at this point. SSD not only drives the travelling wave, but also increases the probability that there is a kind of “local fixation” (escape of the stochastic phase), which approximately corresponds to the probability that a mutation gives rise to a travelling wave of mutants. It would be useful to make analytical progress on this probability, especially because – as we argue below for skeletal muscle – it is useful for predicting observed mutant loads.

In term of applications, at the end of Chapter 2 we mentioned that SSD might contribute to a mechanistic explanation the uptake of agriculture in Europe, a classic wave-like phenomenon^{178–181}. Indeed, the carrying capacity of the farmed land increases and it is not ob-

vious that farmers have a higher net growth rate, as there is evidence that their death rate is higher^{182–185}. It is plausible to develop a physically motivated model of the competition between farmers and hunter and gatherers and try to estimate relevant parameters (namely migration rates, differences in death and reproductive rates and carrying capacity) from the literature. Likewise, there is evidence that the early stages of cancer are characterised by an increase in cellular density^{282,283}. Indeed, cancer cells modify their micro-environment so that it can sustain more cells. This similarity to the mutants of our model points to the possibility that SSD can be a component in the spread of cancer. For these purpose it would be interesting to study SSD in two and three dimensional systems, instead of the chain of demes we use in this thesis, something that could be of theoretical interest on its own.

More generally, when in nature a species *sensu lato* outcompetes another, one could ask whether this is led by a higher net growth rate or if SSD can (contribute to) account for the phenomenon. If SSD is relevant, the connection we have drawn with the spread of altruism (Section 2.7) could shed a new light on the specific phenomenon: the successful species might be winning the competition not because it selfishly occupies the system and uses resources, but in virtue of an altruistic strategy. In this sense SSD might introduce a new paradigm in evolutionary biology and social sciences.

5.2 Further characterisation of skeletal muscle ageing

In Chapter 3 we have shown that SSD can account for the clonal expansion of mitochondrial deletion mutations in skeletal muscle fibres, deriving the parameters of our model from previously published experimental data. Using a model based SSD we have reproduced the spatial profiles of heteroplasmy and speed of expansion of deletion mutants observed in experimental single-cell studies on skeletal muscle. We have applied the phenomenological formula presented in Chapter 2 to check that our conclusions are robust to uncertainty in parameter values. By exploiting the connection between wave speed and wave front steepness we have corroborated our prediction that faster waves propagate in fibres with lower mitochondrial copy number,

and vice versa. Importantly, we have explicitly shown that our model, which does *not* assume a higher replication rate for mutants (i.e. is neutral), can account for the observed mutant loads with *de novo* mutation rates three orders of magnitude lower than previously thought. Previous neutral models, which neglected spatial structure and missed SSD, have been criticised for requiring unrealistically high *de novo* mutation rates in short-lived animals such as rodents. This had led numerous researchers to propose a mechanism whereby mutants would have a higher replication rates, and related mathematical models. A major consequence of our findings is that a higher mutant replication rate is not needed, and furthermore we showed that a standard model based on this cannot account for experimental data.

In Section 2.6 we have shown that the core predictions of SSD are robust to the specific form of the replication rate $\lambda(w, m)$, which encodes the control exerted by the nucleus on the mtDNA population. However, experiments that look at how a population of mitochondria returns to steady state after a perturbation could help specify the form of the feedback control and provide further justification for our model. Although our results are robust to uncertainty in parameter values (Fig. 3.1F), better estimates of the parameters could strengthen the evidence in favor of SSD driving the clonal expansion. In particular, we did not find a way for directly estimate the control strength c , which, being related to how quick a mtDNA population goes back to steady state, could be estimated with the above mentioned experimental work. Microscopy experiments that track single mitochondria in a population are underway in the research group.

Strong evidence in favour of SSD has the main driver of clonal expansion could come from data like those shown in Fig. 3.2. Systematic measures of copy number and heteroplasmy in skeletal muscle fibres with a macroscopic region taken over by mutants could further corroborate the SSD-specific relationship between copy number and speed of expansion, through the connection between speed and steepness of a wave front. If this relationship is confirmed, it would also be evidence against models based on replicative advantage, as those show the opposite dependence of speed on copy number. Another potential way of discerning between replicative advantage and SSD as the driver of the clonal expansion (and possibly other phenomena) could come from tracking lineages of replicating mtDNA and looking at phylogenetic trees. Via simulation one could identify distinctive phylogenies specific to each mechanism, and look for evidence of

either in experimental data.

The probability that a mutation gives rise to a travelling wave of mutants, namely the local fixation probability mentioned in the previous section, is important to fit SSD-based models to experimental data. In Section 3.5 we estimate the *de novo* mutation rate required to produce the observed mutant loads based on a computational estimate of this fixation probability (Fig. 3.3) obtained with our best estimates of the parameter values for humans. Having an expression would allow us to apply the argument of Section 3.5 to other species and to check for robustness to uncertainty in parameter values.

The application of SSD to muscle fibres is challenged by the large number of mitochondrial DNA molecules per nucleus ($N_{ss} \approx 3500$) for humans), which causes Gillespie simulations for realistic spatially extended systems to be very slow. For instance, the data in Figs. 3.1E and 3.3 were produced by hundreds of processors running in parallel for up to nine months. In a MSc project that I co-supervised, Jasmin Daley showed that approximating the Chemical Master Equation as a system of stochastic differential equations (Section 2.2.4) and simulating this system via Euler method reproduced Fig. 3.1E and the associated speed of the expansion. The simulation in the MSc project took a few hours on a standard laptop, saving millions of hours of compute. If this approximation is further validated, it could allow computational studies that are currently unfeasible, such as exploring wide biologically-relevant region of the parameter space and using SSD-based model for approximate Bayesian computation.

This experimental and computational work has the potential to confirm the clonal expansion of mitochondrial deletions in skeletal muscle as the first phenomenon with a compelling description as SSD. This could inform future therapies and recommendations to delay muscle ageing, as those discussed at the end of Chapter 3. If SSD is recognised as the driving force behind the emergence of other diseases, our work will be the basis for an increased scientific understanding as well as new therapeutic avenues.

5.3 Modelling further epidemiological and serological data from SIREN

The mechanistic models developed and tested in Chapters 2 and 3 are based on counting processes. The experience with this class of models turned out to be relevant for urgent research in public health conducted with the UK Health Security Agency. Chapter 4 presents statistical modelling of data from the SIREN study, a prospective cohort study part of the national response to the COVID-19 pandemic. Applying proportional hazards model, a class of models from survival analysis, we have estimated the protection from future SARS-CoV-2 infections associated with vaccination, previous infection, or both. These estimates are derived as hazard ratios of the well-established Cox proportional hazards models; these hazard ratios can be seen as parameters of an underlying counting process that models infection events, as we argue and exemplify via simulated data in Section 4.3.

In summary, we have found that two doses of Pfizer or AstraZeneca vaccine provide high short-term protection against the SARS-CoV-2 Delta variant, but that there is a waning of protection after 6+ months from the second dose; previous infection is associated with high protection for more than one year, and a combination of previous infection and 2 doses of vaccine are associated with high and sustained protection. With the emergence of the Omicron variant, the protection associated with three doses of vaccine was still sizeable but lower, and the benefit of a booster dose seemed to subside 4+ months after administration. Importantly, previous infection with Delta, Alpha variant or wildtype was less protective against Omicron, with previously infected participants up to three times as likely to be reinfected by Omicron than Delta.

A natural follow-up of the work presented in this thesis is a further analysis of protection associated with vaccines and previous infection, with data collected until late autumn or early winter 2022. This will allow us to estimate a) the effect of a second booster (fourth dose) and b) the protection against the current Omicron sub-variants associated with infection with one of the Omicron sub-variant. The answer to question a) is crucial for the future of the pandemic,

as the fourth dose is the first to be administered with an updated vaccine formula, based on the spike protein of the early Omicron variant. If these new jabs provide high protection, the update could potentially be performed periodically, offering a long-term management strategy against SARS-CoV-2. In other words, if Omicron-based vaccines are successful, it means that current vaccine research is able to keep up with this virus.

Point b) above is also very relevant for public health, given the large numbers of people infected or reinfected during the first Omicron wave in winter 2021-2022. Early analysis have been published²⁸⁰, but the strengths of the SIREN study will allow us to contribute with high-quality evidence. An analysis secondary to those presented in Section 4.5 has shown mixed evidence for the benefits of a single booster (third dose) in participants with two doses and a recent previous infection. It is therefore plausible that, if Omicron sub-variants continue to dominate, people with a history of Omicron (re)infection and three vaccine doses could have very little benefit from a fourth dose, that in UK is currently recommended for people over 50 or at high risk. With disparities in access to vaccines at a global level^{284,285} and even within high-income countries²⁸⁶ identifying groups who do not need (at least for now) an additional dose of vaccine would help protecting the people at highest risk as soon as possible.

Data from epidemiological studies as those analysed in Chapter 4 lead to estimates of pooled efficacy of vaccines, namely aggregate values of the protection associated with a given vaccination status. However, the next stages of management of the COVID-19 pandemic will need more fine-grained assessments, ideally at an individual level. In other words, clinicians will need to know the risk factors for a SARS-CoV-2 infection. In this context, the search for a *correlate of protection* against the virus has been intense^{125,287–291}.

In general terms, a correlate of protection (CoP) is the level of immune response that a vaccination or an infection triggers and that is statistically correlated with protection against a future infection. Although immune responses are multifaceted, single components such as the antibody or T-cell levels elicited can correlate with the level of protection. For other diseases, most of the currently accepted CoPs are based on antibody concentrations (titres)²⁹². There is evidence that *neutralising* antibody titres^{287,288,293} are a causal correlate of protection for

infection against COVID-19, but numerous questions remain open. First, the CoPs will likely need to be stratified by severity of outcome²⁸⁹, as the immune response required to protect against infection is probably different from that against severe symptoms. Moreover, the CoP can differ on virus variant, given the specificity of antibodies; CoP can also vary by the previous immunisation events (infection, vaccination or both): for example the CoP for infection can be different from that for reinfection. It would be ideal to identify an *absolute* CoP, a specific level of response that is highly correlated with protection and can be interpreted as a true protective threshold²⁹².

The SIREN study is contributing to the search for reliable CoPs. I am involved in the mathematical and statistical modelling of serological data from the study, co-authored a publication that identified a pre-vaccination CoP¹²⁵ and currently working on a post-vaccination CoP.

It is hard to overstate the importance of finding reliable CoPs against SARS-CoV-2. Estimating risk at an individual level would allow focussed vaccination programmes and principled patient management. Importantly, immuno-compromised individuals could be protected much more efficiently, recommending boosters if titers are too low, or focussing on non-pharmaceutical interventions if a sufficient immune response is absent²⁹⁰. New vaccines need to be licensed to ensure worldwide access. Estimating the effectiveness of new vaccines via epidemiological trials will become increasingly challenging, as a growing proportion of the world population is being exposed to multiple immunisation events. Repeated clinical trials could be too expensive or slower than the emergence of new variants²⁸⁹. As effective vaccines becomes available placebo-controlled trials will become less ethically justifiable. For all these reasons, a reliable CoP is needed in order to rapidly assess the effectiveness of existing and new vaccines, against current and future (sub)variants, by measuring the immune response they elicit. Finally, combining mathematical modelling of the decay of antibody levels and knowledge of a CoP, researchers could predict how protection changes over time.

5.4 Tools to tackle societal challenges

Increasing life expectancy comes with a heavier burden of the diseases of later life³⁸, that poses a systemic threat to our societies. The COVID-19 pandemic has been an epochal event and shock to global systems, but also a stark reminder of the impact of age-related health decline^{294,295} and a further challenge to ageing societies²⁹⁶.

Thanks to the ongoing data explosion in biomedical research^{297,298}, there is an opportunity to transform our understanding of human disease to successfully face these global challenges. However, principled models are required to account for the complex details and intrinsic randomness of biological systems that is being unveiled⁵⁷. Breakthroughs in biomedical sciences, such as the COVID-19 vaccines that saved tens of millions of lives²⁹⁹, are increasingly dependent on mathematical and statistical modelling. This thesis represents our contribution to these challenging, but exciting and meaningful, endeavours.

Bibliography

- [1] W. Kühlbrandt, Structure and function of mitochondrial membrane protein complexes *BMC biology*, vol. 13, no. 1, pp. 1–11, 2015.
- [2] R. Rizzuto, D. De Stefani, A. Raffaello, and C. Mammucari, Mitochondria as sensors and regulators of calcium signalling *Nature reviews Molecular cell biology*, vol. 13, no. 9, pp. 566–578, 2012.
- [3] C. Wang and R. J. Youle, The role of mitochondria in apoptosis *Annual review of genetics*, vol. 43, p. 95, 2009.
- [4] S. Anderson, A. T. Bankier, B. G. Barrell, M. H. de Bruijn, A. R. Coulson, J. Drouin, I. C. Eperon, D. P. Nierlich, B. A. Roe, F. Sanger, *et al.*, Sequence and organization of the human mitochondrial genome *Nature*, vol. 290, no. 5806, p. 457, 1981.
- [5] L. Sagan, On the origin of mitosing cells *Journal of theoretical biology*, vol. 14, no. 3, pp. 225–IN6, 1967.
- [6] W. Martin and M. Mentel, The origin of mitochondria *Nature Education*, vol. 3, no. 9, p. 58, 2010.
- [7] H. R. Lee and K. A. Johnson, Fidelity of the human mitochondrial DNA polymerase *Journal of Biological Chemistry*, vol. 281, no. 47, pp. 36236–36240, 2006.
- [8] S. Papa and V. Skulachev, Reactive oxygen species, mitochondria, apoptosis and aging *Detection of mitochondrial diseases*, pp. 305–319, 1997.

- [9] G. Lenaz, Mitochondria and reactive oxygen species. Which role in physiology and pathology? *Advances in mitochondrial medicine*, pp. 93–136, 2012.
- [10] H. Hoitzing, I. G. Johnston, and N. S. Jones, What is the function of mitochondrial networks? A theoretical assessment of hypotheses and proposal for future research *BioEssays*, vol. 37, no. 6, pp. 687–700, 2015.
- [11] H. Hoitzing, *Controlling mitochondrial dynamics: population genetics and networks*. PhD thesis, Imperial College London, 2017.
- [12] J.-L. Hayashi, S. Ohta, A. Kikuchi, M. Takemitsu, Y.-i. Goto, and I. Nonaka, Introduction of disease-related mitochondrial DNA deletions into HeLa cells lacking mitochondrial DNA results in mitochondrial dysfunction *Proceedings of the National Academy of Sciences*, vol. 88, no. 23, pp. 10614–10618, 1991.
- [13] K. Nakada, K. Inoue, T. Ono, K. Isobe, A. Ogura, Y.-I. Goto, I. Nonaka, and J.-I. Hayashi, Inter-mitochondrial complementation: mitochondria-specific system preventing mice from expression of disease phenotypes by mutant mtDNA *Nat. Med.*, vol. 7, no. 8, p. 934, 2001.
- [14] C. T. Moraes and E. A. Schon, Detection and analysis of mitochondrial DNA and RNA in muscle by in situ hybridization and single-fiber PCR *Methods in enzymology*, vol. 264, pp. 522–540, 1996.
- [15] R. Rossignol, B. Faustin, C. Rocher, M. Malgat, J. Mazat, and T. Letellier, Mitochondrial threshold effects *Biochem. J.*, vol. 370, no. Pt 3, pp. 751–762, 2003.
- [16] E. Bua, J. Johnson, A. Herbst, B. DeLong, D. McKenzie, S. Salamat, and J. M. Aiken, Mitochondrial DNA-deletion mutations accumulate intracellularly to detrimental levels in aged human skeletal muscle fibers *Am. J. Hum. Genet.*, vol. 79, no. 3, pp. 469–480, 2006.
- [17] A. Herbst, J. W. Pak, D. McKenzie, E. Bua, M. Bassiouni, and J. M. Aiken, Accumulation of mitochondrial DNA deletion mutations in aged muscle fibers: evidence for a causal role

- in muscle fiber loss *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, vol. 62, no. 3, pp. 235–245, 2007.
- [18] J. Aryaman, I. G. Johnston, and N. S. Jones, Mitochondrial DNA Density Homeostasis Accounts for the Threshold Effect in Human Mitochondrial Disease *bioRxiv*, p. 078519, 2016.
- [19] M. S. Sharpley, C. Marciniak, K. Eckel-Mahan, M. McManus, M. Crimi, K. Waymire, C. S. Lin, S. Masubuchi, N. Friend, M. Koike, *et al.*, Heteroplasmy of mouse mtDNA is genetically unstable and results in altered behavior and cognition *Cell*, vol. 151, no. 2, pp. 333–343, 2012.
- [20] G. S. Gorman, A. M. Schaefer, Y. Ng, N. Gomez, E. L. Blakely, C. L. Alston, C. Feeney, R. Horvath, P. Yu-Wai-Man, P. F. Chinnery, *et al.*, Prevalence of nuclear and mitochondrial DNA mutations related to adult mitochondrial disease *Annals of neurology*, vol. 77, no. 5, pp. 753–759, 2015.
- [21] G. S. Gorman, P. F. Chinnery, S. DiMauro, M. Hirano, Y. Koga, R. McFarland, A. Suomalainen, D. R. Thorburn, M. Zeviani, and D. M. Turnbull, Mitochondrial diseases *Nature reviews Disease primers*, vol. 2, no. 1, pp. 1–22, 2016.
- [22] R. J. Castro, Mitochondrial replacement therapy: the UK and US regulatory landscapes *Journal of Law and the Biosciences*, vol. 3, no. 3, p. 726, 2016.
- [23] S. Reardon, Three-parent baby’s claim raises hopes—and ethical concerns *Nature*, vol. 28, 2016.
- [24] S. Reardon, Reports of ‘three-parent babies’ multiply *Nature*, 2016.
- [25] N. Dahiya and S. Garg, Three-parent baby: Is it ethical? *Indian journal of medical ethics*, vol. 3, no. 2, pp. 169–169, 2017.
- [26] E. Y. Adashi and I. G. Cohen, Mitochondrial replacement therapy: unmade in the USA *JAMA*, vol. 317, no. 6, pp. 574–575, 2017.

- [27] E. Kang, J. Wu, N. M. Gutierrez, A. Koski, R. Tippner-Hedges, K. Agaronyan, A. Platero-Luengo, P. Martinez-Redondo, H. Ma, Y. Lee, *et al.*, Mitochondrial replacement in human oocytes carrying pathogenic mitochondrial DNA mutations *Nature*, vol. 540, no. 7632, pp. 270–275, 2016.
- [28] M. Yamada, V. Emmanuele, M. J. Sanchez-Quintero, B. Sun, G. Lallo, D. Paull, M. Zimmer, S. Pagett, R. W. Prosser, M. V. Sauer, *et al.*, Genetic drift can compromise mitochondrial replacement by nuclear transfer in human oocytes *Cell stem cell*, vol. 18, no. 6, pp. 749–754, 2016.
- [29] T. Ishii, Reproductive medicine involving mitochondrial DNA modification: evolution, legality, and ethics *REPRODUCTIVE HEALTH*, 2018.
- [30] S. Eom, H. N. Lee, S. Lee, H.-C. Kang, J. S. Lee, H. D. Kim, and Y.-M. Lee, Cause of death in children with mitochondrial diseases *Pediatric Neurology*, vol. 66, pp. 82–88, 2017.
- [31] V. S. Van Laar and S. B. Berman, Mitochondrial dynamics in Parkinson’s disease *Experimental neurology*, vol. 218, no. 2, pp. 247–256, 2009.
- [32] S. Grandemange, S. Herzig, and J.-C. Martinou, Mitochondrial dynamics and cancer in *Seminars in cancer biology*, vol. 19, pp. 50–56, Elsevier, 2009.
- [33] M. Liesa, M. Palacín, and A. Zorzano, Mitochondrial dynamics in mammalian health and disease *Physiological reviews*, vol. 89, no. 3, pp. 799–845, 2009.
- [34] X. Zhu, G. Perry, M. A. Smith, and X. Wang, Abnormal mitochondrial dynamics in the pathogenesis of Alzheimer’s disease *Journal of Alzheimer’s Disease*, vol. 33, no. s1, pp. S253–S262, 2013.
- [35] D. Harman, The Biologic Clock: The Mitochondria? *Journal of the American Geriatrics Society.*, vol. 20, no. 4, pp. 145–147, 1972.
- [36] K. J. Krishnan, L. C. Greaves, A. K. Reeve, and D. Turnbull, The ageing mitochondrial genome *Nucleic acids research*, vol. 35, no. 22, pp. 7399–7405, 2007.

- [37] T. E. Kauppila, J. H. Kauppila, and N.-G. Larsson, Mammalian mitochondria and aging: an update *Cell Metab.*, vol. 25, no. 1, pp. 57–71, 2017.
- [38] L. Partridge, J. Deelen, and P. E. Slagboom, Facing up to the global challenges of ageing *Nature*, vol. 561, pp. 45–56, Sept. 2018.
- [39] Y. Rolland, S. Czerwinski, G. A. van Kan, J. E. Morley, M. Cesari, G. Onder, J. Woo, R. Baumgartner, F. Pillard, Y. Boirie, W. M. C. Chumlea, and B. Vellas, Sarcopenia: Its assessment, etiology, pathogenesis, consequences and future perspectives *The Journal of Nutrition Health and Aging*, vol. 12, pp. 433–450, Sept. 2008.
- [40] E. A. Bua, S. H. McKiernan, J. Wanagat, D. McKenzie, and J. M. Aiken, Mitochondrial abnormalities are more frequent in muscles undergoing sarcopenia *Journal of Applied Physiology*, vol. 92, no. 6, pp. 2617–2624, 2002.
- [41] F. Bellanti, A. Lo Buglio, and G. Vendemiale, Mitochondrial Impairment in Sarcopenia *Biology*, vol. 10, no. 1, p. 31, 2021.
- [42] S. Alway, J. Mohamed, and M. Myers, Mitochondria Initiate and Regulate Sarcopenia *Exercise and Sport Sciences Reviews*, vol. 45, pp. 58–69, Apr. 2017.
- [43] E. Ferri, E. Marzetti, R. Calvani, A. Picca, M. Cesari, and B. Arosio, Role of Age-Related Mitochondrial Dysfunction in Sarcopenia *International Journal of Molecular Sciences*, vol. 21, no. 15, p. 5236, 2020.
- [44] F. J. Miller, F. L. Rosenfeldt, C. Zhang, A. W. Linnane, and P. Nagley, 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*, vol. 31, no. 11, pp. e61–e61, 2003.
- [45] Z. Cao, J. Wanagat, S. H. McKiernan, and J. M. Aiken, Mitochondrial DNA deletion mutations are concomitant with ragged red regions of individual, aged muscle fibers: analysis by laser-capture microdissection *Nucleic acids research*, vol. 29, no. 21, pp. 4502–4508, 2001.

- [46] N. G. Gokey, Z. Cao, J. W. Pak, D. Lee, S. H. McKiernan, D. McKenzie, R. Weindruch, and J. M. Aiken, Molecular analyses of mtDNA deletion mutations in microdissected skeletal muscle fibers from aged rhesus monkeys *Aging Cell*, vol. 3, no. 5, pp. 319–326, 2004.
- [47] S. H. McKiernan, R. Colman, M. Lopez, T. Mark Beasley, R. Weindruch, and J. M. Aiken, Longitudinal analysis of early stage sarcopenia in aging rhesus monkeys *Experimental Gerontology*, vol. 44, pp. 170–176, Mar. 2009.
- [48] E. J. Brierley, M. A. Johnson, R. N. Lightowlers, O. F. James, and D. M. Turnbull, Role of mitochondrial DNA mutations in human aging: implications for the central nervous system and muscle *Annals of neurology*, vol. 43, no. 2, pp. 217–223, 1998.
- [49] A. E. Vincent, H. S. Rosa, K. Pabis, C. Lawless, C. Chen, A. Grünewald, K. A. Rygiel, M. C. Rocha, A. K. Reeve, G. Falkous, *et al.*, Subcellular origin of mitochondrial DNA deletions in human skeletal muscle *Annals of neurology*, vol. 84, no. 2, pp. 289–301, 2018.
- [50] A. E. Vincent, K. White, T. Davey, J. Philips, R. T. Ogden, C. Lawless, C. Warren, M. G. Hall, Y. S. Ng, G. Falkous, *et al.*, Quantitative 3D mapping of the human skeletal muscle mitochondrial network *Cell reports*, vol. 26, no. 4, pp. 996–1009, 2019.
- [51] C. Lawless, L. Greaves, A. K. Reeve, D. M. Turnbull, and A. E. Vincent, The rise and rise of mitochondrial DNA mutations *Open Biology*, vol. 10, no. 5, p. 200061, 2020.
- [52] J. R. Waldram, The theory of thermodynamics *The Theory of Thermodynamics*, 1985.
- [53] D. Silver, S. Singh, D. Precup, and R. S. Sutton, Reward is enough *Artificial Intelligence*, vol. 299, p. 103535, 2021.
- [54] J. Janssen, R. Manca, and E. Volpe, *Mathematical finance: deterministic and stochastic models*. John Wiley & Sons, 2013.
- [55] P. G. Walker, C. Whittaker, O. J. Watson, M. Baguelin, P. Winskill, A. Hamlet, B. A. Djafaara, Z. Cucunubá, D. Olivera Mesa, W. Green, *et al.*, The impact of COVID-19 and

- strategies for mitigation and suppression in low-and middle-income countries *Science*, vol. 369, no. 6502, pp. 413–422, 2020.
- [56] E. Parzen, *Stochastic processes*. SIAM, 1999.
- [57] G. Constable, Y. Krumbeck, and T. Rogers, An Invitation to Stochastic Mathematical Biology *Notices of the American Mathematical Society*, vol. 68, no. 11, 2021.
- [58] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain, Stochastic gene expression in a single cell *Science*, vol. 297, no. 5584, pp. 1183–1186, 2002.
- [59] P. C. Bressloff and J. M. Newby, Stochastic models of intracellular transport *Reviews of Modern Physics*, vol. 85, no. 1, p. 135, 2013.
- [60] S. Uphoff, N. D. Lord, B. Okumus, L. Potvin-Trottier, D. J. Sherratt, and J. Paulsson, Stochastic activation of a DNA damage response causes cell-to-cell mutation rate variation *Science*, vol. 351, no. 6277, pp. 1094–1097, 2016.
- [61] R. Lande, S. Engen, B.-E. Saether, *et al.*, *Stochastic population dynamics in ecology and conservation*. Oxford University Press on Demand, 2003.
- [62] L. G. Shoemaker, L. L. Sullivan, I. Donohue, J. S. Cabral, R. J. Williams, M. M. Mayfield, J. M. Chase, C. Chu, W. S. Harpole, A. Huth, *et al.*, Integrating the underlying structure of stochasticity into community ecology *Ecology*, vol. 101, no. 2, p. e02922, 2020.
- [63] B. E. Kendall and G. A. Fox, Unstructured individual variation and demographic stochasticity *Conservation biology*, vol. 17, no. 4, pp. 1170–1172, 2003.
- [64] B. S. Fleishman, Stochastic theory of ecological interactions *Ecological Modelling*, vol. 17, no. 2, pp. 65–73, 1982.
- [65] B. Tenhumberg, A. J. Tyre, and B. Roitberg, Stochastic variation in food availability influences weight and age at maturity *Journal of theoretical biology*, vol. 202, no. 4, pp. 257–272, 2000.
- [66] M. Fujiwara and T. Takada, Environmental stochasticity *Els*, pp. 1–8, 2017.

- [67] A. Marusyk, V. Almendro, and K. Polyak, Intra-tumour heterogeneity: a looking glass for cancer? *Nature Reviews Cancer*, vol. 12, no. 5, p. 323, 2012.
- [68] C. E. Hayford, D. R. Tyson, C. J. Robbins III, P. L. Frick, V. Quaranta, and L. A. Harris, An in vitro model of tumor heterogeneity resolves genetic, epigenetic, and stochastic sources of cell state variability *PLoS biology*, vol. 19, no. 6, p. e3000797, 2021.
- [69] J. Aryaman, I. G. Johnston, and N. S. Jones, Mitochondrial heterogeneity *Frontiers in genetics*, vol. 9, p. 718, 2019.
- [70] H. Hoitzing, I. G. Johnston, and N. S. Jones, Stochastic models for evolving cellular populations of mitochondria: disease, development, and ageing in *Stochastic Processes, Multi-scale Modeling, and Numerical Methods for Computational Cellular Biology*, pp. 287–314, Springer, 2017.
- [71] R. Lande, Natural selection and random genetic drift in phenotypic evolution *Evolution*, pp. 314–334, 1976.
- [72] J. Masel, Genetic drift *Current Biology*, vol. 21, no. 20, pp. R837–R838, 2011.
- [73] O. Hallatschek, P. Hersen, S. Ramanathan, and D. R. Nelson, Genetic drift at expanding frontiers promotes gene segregation *Proceedings of the National Academy of Sciences*, vol. 104, no. 50, pp. 19926–19930, 2007.
- [74] A. J. McKane and T. J. Newman, Predator-prey cycles from resonant amplification of demographic stochasticity *Physical review letters*, vol. 94, no. 21, p. 218102, 2005.
- [75] G. W. Constable, T. Rogers, A. J. McKane, and C. E. Tarnita, Demographic noise can reverse the direction of deterministic selection *Proceedings of the National Academy of Sciences*, vol. 113, no. 32, pp. E4745–E4754, 2016.
- [76] O. Hallatschek, Noise Driven Evolutionary Waves *PLOS Computational Biology*, vol. 7, no. 3, pp. 1–9, 2011.
- [77] B. Houchmandzadeh and M. Vallade, Selection for altruism through random drift in variable size populations *BMC Evolutionary Biology*, vol. 12, p. 61, May 2012.

- [78] B. Houchmandzadeh, Fluctuation driven fixation of cooperative behavior *BioSystems*, vol. 127, p. 60, Jan. 2015.
- [79] S. Okasha, Biological Altruism in *The Stanford Encyclopedia of Philosophy* (E. N. Zalta, ed.), Metaphysics Research Lab, Stanford University, summer 2020 ed., 2020.
- [80] J.-U. Kreft, Biofilms promote altruism *Microbiology*, vol. 150, no. 8, pp. 2751–2760, 2004.
- [81] P. F. Chinnery and D. C. Samuels, Relaxed replication of mtDNA: a model with implications for the expression of disease *Am. J. Hum. Genet.*, vol. 64, p. 1158, 1999.
- [82] G. J. Capps, D. C. Samuels, and P. F. Chinnery, A model of the nuclear control of mitochondrial DNA replication *J. Theor. Biol.*, vol. 221, no. 4, pp. 565–583, 2003.
- [83] J. Elson, D. Samuels, D. Turnbull, and P. Chinnery, Random intracellular drift explains the clonal expansion of mitochondrial DNA mutations with age *The American Journal of Human Genetics*, vol. 68, no. 3, pp. 802–806, 2001.
- [84] E. Renshaw, *Stochastic population processes: analysis, approximations, simulations*. OUP Oxford, 2015.
- [85] I. G. Johnston, J. P. Burgstaller, V. Havlicek, T. Kolbe, T. Rülcke, G. Brem, J. Poulton, and N. S. Jones, Stochastic modelling, Bayesian inference, and new in vivo measurements elucidate the debated mtDNA bottleneck mechanism *eLife*, vol. 4, p. e07464, 2015.
- [86] I. G. Johnston, Varied mechanisms and models for the varying mitochondrial bottleneck *Frontiers in Cell and Developmental Biology*, vol. 7, p. 294, 2019.
- [87] A. Kowald and E. Klipp, Mathematical models of mitochondrial aging and dynamics in *Prog. Mol. Biol. Transl. Sci.*, vol. 127, pp. 63–92, Elsevier, 2014.
- [88] J. Aryaman, C. Bowles, N. S. Jones, and I. G. Johnston, Mitochondrial network state scales mtDNA genetic dynamics *Genetics*, vol. 212, no. 4, pp. 1429–1443, 2019.
- [89] J. Aryaman, Stochastic modelling and inference for mitochondria in health and disease 2018.

- [90] A. Kowald, F. P. Kemeth, and T. B. Kirkwood, An explanation of how mutant and wild-type mitochondria might stably co-exist in inherited mitochondrial diseases *PNAS Nexus*, 2022.
- [91] R. C. Glastad and I. Johnston, Mitochondrial network structure controls cell-to-cell mtDNA variability generated by cell divisions *bioRxiv*, 2022.
- [92] A. D. N. J. D. Grey, A proposed refinement of the mitochondrial free radical theory of aging *BioEssays*, vol. 19, no. 2, pp. 161–166, 1997.
- [93] G. Twig, A. Elorza, A. J. Molina, H. Mohamed, J. D. Wikstrom, G. Walzer, L. Stiles, S. E. Haigh, S. Katz, G. Las, *et al.*, Fission and selective fusion govern mitochondrial segregation and elimination by autophagy *EMBO J.*, vol. 27, no. 2, pp. 433–446, 2008.
- [94] A. Kowald and T. B. Kirkwood, Evolution of the mitochondrial fusion–fission cycle and its role in aging *Proc. Natl. Acad. Sci. USA*, vol. 108, no. 25, pp. 10237–10242, 2011.
- [95] D. C. Wallace, Mitochondrial genetics: a paradigm for aging and degenerative diseases? *Science*, vol. 256, no. 5057, p. 628, 1992.
- [96] C. M. Lee, M. E. Lopez, R. Weindruch, and J. M. Aiken, Association of age-related mitochondrial abnormalities with skeletal muscle fiber atrophy *Free Radical Biology and Medicine*, vol. 25, no. 8, pp. 964–972, 1998.
- [97] O. M. Russell, I. Fruh, P. K. Rai, D. Marcellin, T. Doll, A. Reeve, M. Germain, J. Bastien, K. A. Rygiel, R. Cerino, A. W. Sailer, M. Lako, R. W. Taylor, M. Mueller, R. N. Lightowlers, D. M. Turnbull, and S. B. Helliwell, Preferential amplification of a human mitochondrial DNA deletion in vitro and in vivo *Scientific Reports*, vol. 8, pp. 1–10, Jan. 2018.
- [98] F. Diaz, M. P. Bayona-Bafaluy, M. Rana, M. Mora, H. Hao, and C. T. Moraes, Human mitochondrial DNA with large deletions repopulates organelles faster than full-length genomes under relaxed copy number control *Nucleic Acids Res.*, vol. 30, no. 21, pp. 4626–4633, 2002.

- [99] A. D. De Grey, *The mitochondrial free radical theory of aging*. Citeseer, 1999.
- [100] A. Kowald and T. B. Kirkwood, Resolving the Enigma of the Clonal Expansion of mtDNA Deletions *Genes*, vol. 9, Feb. 2018.
- [101] G. Campbell, K. J. Krishnan, M. Deschauer, R. W. Taylor, and D. M. Turnbull, Dissecting the mechanisms underlying the accumulation of mitochondrial DNA deletions in human skeletal muscle *Hum. Mol. Genet.*, vol. 23, no. 17, pp. 4612–4620, 2014.
- [102] N. Lane, Mitonuclear match: optimizing fitness and fertility over generations drives ageing within generations *Bioessays*, vol. 33, no. 11, pp. 860–869, 2011.
- [103] A. Kowald and T. B. Kirkwood, Transcription could be the key to the selection advantage of mitochondrial deletion mutants in aging *Proceedings of the National Academy of Sciences*, vol. 111, no. 8, pp. 2972–2977, 2014.
- [104] J. Lapointe and S. Hekimi, When a theory of aging ages badly *Cellular and Molecular Life Sciences*, vol. 67, no. 1, pp. 1–8, 2010.
- [105] B. Bandy and A. J. Davison, Mitochondrial mutations may increase oxidative stress: implications for carcinogenesis and aging? *Free Radical Biology and Medicine*, vol. 8, no. 6, pp. 523–539, 1990.
- [106] J. Damas, J. Carneiro, A. Amorim, and F. Pereira, MitoBreak: the mitochondrial DNA breakpoints database *Nucleic acids research*, vol. 42, no. D1, pp. D1261–D1268, 2014.
- [107] G. S. Shadel and D. A. Clayton, Mitochondrial DNA maintenance in vertebrates *Annual review of biochemistry*, vol. 66, no. 1, pp. 409–435, 1997.
- [108] P. Fernández-Silva, J. A. Enriquez, and J. Montoya, Replication and Transcription of Mammalian Mitochondrial Dna *Experimental Physiology*, vol. 88, no. 1, pp. 41–56, 2003.
- [109] D.-F. Suen, D. P. Narendra, A. Tanaka, G. Manfredi, and R. J. Youle, Parkin overexpression selects against a deleterious mtDNA mutation in heteroplasmic cybrid cells *Proc. Natl. Acad. Sci. USA*, vol. 107, no. 26, pp. 11835–11840, 2010.

- [110] S. J. Goldman, R. Taylor, Y. Zhang, and S. Jin, Autophagy and the degradation of mitochondria *Mitochondrion*, vol. 10, no. 4, pp. 309–315, 2010.
- [111] I. Kim and J. J. Lemasters, Mitophagy selectively degrades individual damaged mitochondria after photoirradiation *Antioxidants & redox signaling*, vol. 14, no. 10, pp. 1919–1928, 2011.
- [112] A. Kowald and T. B. Kirkwood, Mitochondrial mutations and aging: random drift is insufficient to explain the accumulation of mitochondrial deletion mutants in short-lived animals *Aging Cell*, vol. 12, no. 4, pp. 728–731, 2013.
- [113] P. M. Folegatti, K. J. Ewer, P. K. Aley, B. Angus, S. Becker, S. Belij-Rammerstorfer, D. Bellamy, S. Bibi, M. Bittaye, E. A. Clutterbuck, *et al.*, Safety and immunogenicity of the ChAdOx1 nCoV-19 vaccine against SARS-CoV-2: a preliminary report of a phase 1/2, single-blind, randomised controlled trial *The Lancet*, vol. 396, no. 10249, pp. 467–478, 2020.
- [114] L. R. Baden, H. M. El Sahly, B. Essink, K. Kotloff, S. Frey, R. Novak, D. Diemert, S. A. Spector, N. Rouphael, C. B. Creech, *et al.*, Efficacy and safety of the mRNA-1273 SARS-CoV-2 vaccine *New England journal of medicine*, 2020.
- [115] F. P. Polack, S. J. Thomas, N. Kitchin, J. Absalon, A. Gurtman, S. Lockhart, J. L. Perez, G. P. Marc, E. D. Moreira, C. Zerbini, *et al.*, Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine *New England journal of medicine*, 2020.
- [116] S. Chalkias, C. Harper, K. Vrbicky, S. R. Walsh, B. Essink, A. Brosz, N. McGhee, J. E. Tomassini, X. Chen, Y. Chang, *et al.*, A bivalent Omicron-containing booster vaccine against Covid-19 *New England Journal of Medicine*, 2022.
- [117] M. D. Hitchings, O. T. Ranzani, M. S. S. Torres, S. B. de Oliveira, M. Almiron, R. Said, R. Borg, W. L. Schulz, R. D. de Oliveira, P. V. da Silva, *et al.*, Effectiveness of CoronaVac among healthcare workers in the setting of high SARS-CoV-2 Gamma variant transmission in Manaus, Brazil: A test-negative case-control study *The Lancet Regional Health-Americas*, vol. 1, p. 100025, 2021.

- [118] M. D. Tanrioer, H. L. Doğanay, M. Akova, H. R. Güner, A. Azap, S. Akhan, Ş. Köse, F. Ş. Erdiñç, E. H. Akalın, Ö. F. Tabak, *et al.*, Efficacy and safety of an inactivated whole-virion SARS-CoV-2 vaccine (CoronaVac): interim results of a double-blind, randomised, placebo-controlled, phase 3 trial in Turkey *The Lancet*, vol. 398, no. 10296, pp. 213–222, 2021.
- [119] Y. Liang, C. Liu, M. Lu, Q. Dong, Z. Wang, Z. Wang, W. Xiong, N. Zhang, J. Zhou, Q. Liu, *et al.*, Calorie restriction is the most reasonable anti-ageing intervention: a meta-analysis of survival curves *Scientific reports*, vol. 8, no. 1, pp. 1–9, 2018.
- [120] O. Aalen, Nonparametric inference for a family of counting processes *The Annals of Statistics*, pp. 701–726, 1978.
- [121] P. K. Andersen, Ø. Borgan, N. L. Hjort, E. Arjas, J. Stene, and O. Aalen, Counting process models for life history data: A review [with discussion and reply] *Scandinavian Journal of Statistics*, pp. 97–158, 1985.
- [122] P. K. Andersen, O. Borgan, R. D. Gill, and N. Keiding, *Statistical models based on counting processes*. Springer Science & Business Media, 2012.
- [123] F. Insalata, H. Hoitzing, J. Aryaman, and N. S. Jones, Stochastic survival of the densest and mitochondrial DNA clonal expansion in aging *Proceedings of the National Academy of Sciences*, vol. 119, no. 49, p. e2122073119, 2022.
- [124] V. Hall, S. Foulkes, F. Insalata, P. Kirwan, A. Saei, A. Atti, E. Wellington, J. Khawam, K. Munro, M. Cole, *et al.*, Protection against SARS-CoV-2 after Covid-19 vaccination and previous infection *New England Journal of Medicine*, vol. 386, no. 13, pp. 1207–1220, 2022.
- [125] A. Atti, F. Insalata, E. J. Carr, A. D. Otter, J. Castillo-Olivares, M. Wu, R. Harvey, M. Howell, A. Chan, J. Lyall, *et al.*, Antibody correlates of protection from SARS-CoV-2 reinfection prior to vaccination: a nested case-control within the SIREN study *Journal of Infection*, 2022.

- [126] F. Insalata, F. Corberi, L. Cugliandolo, and M. Picco, Coarsening and percolation in the Ising Model with quenched disorder in *Journal of Physics: Conference Series*, vol. 956, p. 012018, IOP Publishing, 2018.
- [127] F. Corberi, L. F. Cugliandolo, F. Insalata, and M. Picco, Fractal character of the phase ordering kinetics of a diluted ferromagnet *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2019, no. 4, p. 043203, 2019.
- [128] R. Agrawal, F. Corberi, F. Insalata, and S. Puri, Asymptotic states of Ising ferromagnets with long-range interactions *Physical Review E*, vol. 105, no. 3, p. 034131, 2022.
- [129] D. Unyi, F. Insalata, P. Veličković, and B. Gyires-Tóth, Utility of Equivariant Message Passing in Cortical Mesh Segmentation *arXiv preprint arXiv:2206.03164*, 2022.
- [130] D. C. Wallace, Mitochondrial DNA mutations and neuromuscular disease *Trends Genet.*, vol. 5, pp. 9–13, 1989.
- [131] N. Arnheim and G. Cortopassi, Deleterious mitochondrial DNA mutations accumulate in aging human tissues *Mutation Research/DNAging*, vol. 275, pp. 157–167, Sept. 1992.
- [132] M. Yoneda, A. Chomyn, A. Martinuzzi, O. Hurko, and G. Attardi, Marked replicative advantage of human mtDNA carrying a point mutation that causes the MELAS encephalomyopathy *Proceedings of the National Academy of Sciences*, vol. 89, no. 23, pp. 11164–11168, 1992.
- [133] A. Kowald, M. Dawson, and T. B. L. Kirkwood, Mitochondrial mutations and ageing: Can mitochondrial deletion mutants accumulate via a size based replication advantage? *Journal of Theoretical Biology*, vol. 340, pp. 111–118, Jan. 2014.
- [134] C. Zhu and G. Yin, On competitive Lotka–Volterra model in random environments *Journal of Mathematical Analysis and Applications*, vol. 357, no. 1, pp. 154–170, 2009.
- [135] I. G. Johnston and N. S. Jones, Evolution of cell-to-cell variability in stochastic, controlled, heteroplasmic mtDNA populations *Am. J. Hum. Genet.*, vol. 99, no. 5, pp. 1150–1162, 2016.

- [136] D. T. Gillespie, A general method for numerically simulating the stochastic time evolution of coupled chemical reactions *J. Comp. Phys.*, vol. 22, no. 4, pp. 403–434, 1976.
- [137] D. T. Gillespie, Exact stochastic simulation of coupled chemical reactions *The journal of Physical Chemistry*, vol. 81, no. 25, pp. 2340–2361, 1977.
- [138] C. Gardiner, *Stochastic Methods: A Handbook for the Natural and Social Sciences*. Springer Series in Synergetics, Springer Berlin Heidelberg, 1985.
- [139] K. Jacobs, *Stochastic processes for physicists: understanding noisy systems*. Cambridge University Press, 2010.
- [140] T. L. Parsons and T. Rogers, Dimension reduction for stochastic dynamical systems forced onto a manifold by large drift: a constructive approach with examples from theoretical biology *J. Phys. A*, vol. 50, no. 41, p. 415601, 2017.
- [141] G. W. A. Constable and A. J. McKane, Exploiting Fast-Variables to Understand Population Dynamics and Evolution *Journal of Statistical Physics*, vol. 172, pp. 3–43, July 2018.
- [142] P. Czuppon and A. Traulsen, Fixation probabilities in populations under demographic fluctuations *Journal of Mathematical Biology*, vol. 77, pp. 1233–1277, Oct. 2018.
- [143] R. A. Fisher, The Wave of Advance of Advantageous Genes *Annals of Eugenics*, vol. 7, no. 4, pp. 355–369, 1937.
- [144] A. Kolmogorov, I. Petrovskii, and N. Piskunov, Investigation of the Diffusion Equation Involving Growth in the Amount of Matter, and Its Application to a Biological Problem *Byul. Moskov. Gos. Univ. Sek. A*, vol. 1, no. 6, pp. 1–26, 1937.
- [145] V. J. Hall, S. Foulkes, A. Charlett, A. Atti, E. J. Monk, R. Simmons, E. Wellington, M. J. Cole, A. Saei, B. Oguti, *et al.*, SARS-CoV-2 infection rates of antibody-positive compared with antibody-negative health-care workers in England: a large, multicentre, prospective cohort study (SIREN) *The Lancet*, vol. 397, no. 10283, pp. 1459–1469, 2021.

- [146] C.-S. Liu, C.-S. Tsai, C.-L. Kuo, H.-W. Chen, C.-K. Lii, Y.-S. Ma, and Y.-H. Wei, Oxidative stress-related alteration of the copy number of mitochondrial DNA in human leukocytes *Free radical research*, vol. 37, no. 12, pp. 1307–1317, 2003.
- [147] C.-S. Liu, W.-L. Cheng, C.-F. Lee, Y.-S. Ma, C.-Y. Lin, C.-C. Huang, and Y.-H. Wei, Alteration in the copy number of mitochondrial DNA in leukocytes of patients with mitochondrial encephalomyopathies *Acta neurologica scandinavica*, vol. 113, no. 5, pp. 334–341, 2006.
- [148] H. Hoitzing, I. G. Johnston, and N. S. Jones, Stochastic Models for Evolving Cellular Populations of Mitochondria: Disease, Development, and Ageing in *Stochastic Processes, Multiscale Modeling, and Numerical Methods for Computational Cellular Biology*, pp. 287–314, Springer, 2017.
- [149] M. A. Brockhurst, A. Buckling, D. Racey, and A. Gardner, Resource supply and the evolution of public-goods cooperation in bacteria *BMC biology*, vol. 6, no. 1, p. 20, 2008.
- [150] T. Pfeiffer, S. Schuster, and S. Bonhoeffer, Cooperation and competition in the evolution of ATP-producing pathways *Science*, vol. 292, no. 5516, pp. 504–507, 2001.
- [151] M. Uyenoyama and M. W. Feldman, Theories of kin and group selection: A population genetics perspective *Theoretical Population Biology*, vol. 17, pp. 380–414, June 1980.
- [152] A. S. Griffin and S. A. West, Kin selection: fact and fiction *Trends in Ecology & Evolution*, vol. 17, pp. 15–21, Jan. 2002.
- [153] C. R. Darwin, *The descent of man, and selection in relation to sex*. London: John Murray. vol. 1. John Murray, London, 1871.
- [154] H. Hoitzing, P. A. Gammage, L. Van Haute, M. Minczuk, I. G. Johnston, and N. S. Jones, Energetic costs of cellular and therapeutic control of stochastic mitochondrial DNA populations *PLoS Computational Biology*, vol. 15, no. 6, p. e1007023, 2019.
- [155] B. Clarke, Density-dependent selection *The American Naturalist*, vol. 106, no. 947, pp. 1–13, 1972.

- [156] L. D. Mueller, Theoretical and empirical examination of density-dependent selection *Annual Review of Ecology and Systematics*, vol. 28, no. 1, pp. 269–288, 1997.
- [157] T. E. Farkas and G. Montejó-Kovacevich, Density-dependent selection closes an eco-evolutionary feedback loop in the stick insect *Timema cristinae* *Biology letters*, vol. 10, no. 12, p. 20140896, 2014.
- [158] J. Bertram and J. Masel, Density-dependent selection and the limits of relative fitness *Theoretical population biology*, vol. 129, pp. 81–92, 2019.
- [159] M. Slatkin and J. M. Smith, Models of coevolution *The Quarterly Review of Biology*, vol. 54, no. 3, pp. 233–263, 1979.
- [160] M. Turelli and D. Petry, Density-dependent selection in a random environment: an evolutionary process that can maintain stable population dynamics *Proceedings of the National Academy of Sciences*, vol. 77, no. 12, pp. 7501–7505, 1980.
- [161] D. G. Heckel and J. Roughgarden, A species near its equilibrium size in a fluctuating environment can evolve a lower intrinsic rate of increase *Proceedings of the National Academy of Sciences*, vol. 77, no. 12, pp. 7497–7500, 1980.
- [162] R. Lande, S. Engen, and B.-E. Sæther, An evolutionary maximum principle for density-dependent population dynamics in a fluctuating environment *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1523, pp. 1511–1518, 2009.
- [163] T. L. Parsons and C. Quince, Fixation in haploid populations exhibiting density dependence I: the non-neutral case *Theoretical population biology*, vol. 72, no. 1, pp. 121–135, 2007.
- [164] T. L. Parsons and C. Quince, Fixation in haploid populations exhibiting density dependence II: The quasi-neutral case *Theoretical population biology*, vol. 72, no. 4, pp. 468–479, 2007.
- [165] R. Oizumi, T. Kuniya, and Y. Enatsu, Reconsideration of r/K selection theory using

- stochastic control theory and nonlinear structured population models *PLoS One*, vol. 11, no. 6, p. e0157715, 2016.
- [166] R. H. MacArthur and E. O. Wilson, *The theory of island biogeography*. Princeton University Press, 1967.
- [167] E. R. Pianka, On r-and K-selection *The american naturalist*, vol. 104, no. 940, pp. 592–597, 1970.
- [168] M. S. Boyce, Restitution of r-and K-selection as a model of density-dependent natural selection *Annual review of ecology and systematics*, vol. 15, pp. 427–447, 1984.
- [169] D. Reznick, M. J. Bryant, and F. Bashey, r-and K-selection revisited: the role of population regulation in life-history evolution *Ecology*, vol. 83, no. 6, pp. 1509–1520, 2002.
- [170] F. J. Ayala and C. A. Campbell, Frequency-dependent selection *Annual review of Ecology and systematics*, vol. 5, no. 1, pp. 115–138, 1974.
- [171] M. Heino, J. A. Metz, and V. Kaitala, The enigma of frequency-dependent selection *Trends in Ecology & Evolution*, vol. 13, no. 9, pp. 367–370, 1998.
- [172] P. A. DeBenedictis, The meaning and measurement of frequency-dependent competition *Ecology*, vol. 58, no. 1, pp. 158–166, 1977.
- [173] W. C. Allee, Animal aggregations *The Quarterly Review of Biology*, vol. 2, no. 3, pp. 367–398, 1927.
- [174] F. Courchamp, T. Clutton-Brock, and B. Grenfell, Inverse density dependence and the Allee effect *Trends in ecology & evolution*, vol. 14, no. 10, pp. 405–410, 1999.
- [175] M. A. Nowak and R. M. May, Evolutionary games and spatial chaos *Nature*, vol. 359, no. 6398, pp. 826–829, 1992.
- [176] S. Lion and M. v. Baalen, Self-structuring in spatial evolutionary ecology *Ecology letters*, vol. 11, no. 3, pp. 277–295, 2008.

- [177] C. Hauert and M. Doebeli, Spatial structure often inhibits the evolution of cooperation in the snowdrift game *Nature*, vol. 428, no. 6983, pp. 643–646, 2004.
- [178] A. J. Ammerman and L. L. Cavalli-Sforza, Measuring the rate of spread of early farming in Europe *Man*, pp. 674–688, 1971.
- [179] A. J. Ammerman, A population model for the diffusion of early farming in Europe *The Explanation of Culture Change.*, pp. 343–357, 1973.
- [180] A. Ammerman and L. Cavalli-Sforza, The wave of advance model for the spread of early farming *Transformations: mathematical approaches to culture change. New York: Academic Press. p*, pp. 275–293, 1979.
- [181] R. Pinhasi, J. Fort, and A. J. Ammerman, Tracing the origin and spread of agriculture in Europe *PLoS biology*, vol. 3, no. 12, 2005.
- [182] A. J. Robson, A bioeconomic view of the Neolithic transition to agriculture *Canadian Journal of Economics/Revue canadienne d'économique*, vol. 43, no. 1, pp. 280–300, 2010.
- [183] P. M. Lambert, Health versus fitness: competing themes in the origins and spread of agriculture? *Current Anthropology*, vol. 50, no. 5, pp. 603–608, 2009.
- [184] M. N. Cohen, *Health and the Rise of Civilization*. Yale University Press, 1989.
- [185] G. J. Armelagos and M. N. Cohen, *Paleopathology at the Origins of Agriculture*. Academic Press Orlando, FL, 1984.
- [186] J. Wanagat, Z. Cao, P. Pathare, and J. M. Aiken, Mitochondrial DNA deletion mutations colocalize with segmental electron transport system abnormalities, muscle fiber atrophy, fiber splitting, and oxidative damage in sarcopenia *FASEB J.*, vol. 15, no. 2, pp. 322–332, 2001.
- [187] C. Barthélémy, H. Ogier De Baulny, J. Diaz, M. Armelle Cheval, P. Frachon, N. Romero, F. Goutieres, M. Fardeau, and A. Lombès, Late-onset mitochondrial DNA depletion: DNA copy number, multiple deletions, and compensation *Annals of neurology*, vol. 49, no. 5, pp. 607–617, 2001.

- [188] S. E. Durham, D. C. Samuels, L. M. Cree, and P. F. Chinnery, Normal levels of wild-type mitochondrial DNA maintain cytochrome *c* oxidase activity for two pathogenic mitochondrial DNA mutations but not for m.3243A→G *Am. J. Hum. Genet.*, vol. 81, no. 1, pp. 189–195, 2007.
- [189] E. A. Shoubridge, G. Karpati, and K. E. Hastings, Deletion mutants are functionally dominant over wild-type mitochondrial genomes in skeletal muscle fiber segments in mitochondrial disease *Cell*, vol. 62, no. 1, pp. 43–49, 1990.
- [190] P. M. Garcia-Roves, M. E. Osler, M. H. Holmström, and J. R. Zierath, Gain-of-function R225Q mutation in AMP-activated protein kinase γ 3 subunit increases mitochondrial biogenesis in glycolytic skeletal muscle *Journal of Biological Chemistry*, vol. 283, no. 51, pp. 35724–35734, 2008.
- [191] S. A. Dogan, C. Pujol, P. Maiti, A. Kukat, S. Wang, S. Hermans, K. Senft, R. Wibom, E. I. Rugarli, and A. Trifunovic, Tissue-specific loss of DARS2 activates stress responses independently of respiratory chain deficiency in the heart *Cell metabolism*, vol. 19, no. 3, pp. 458–469, 2014.
- [192] B. L. Gitschlag, C. S. Kirby, D. C. Samuels, R. D. Gangula, S. A. Mallal, and M. R. Patel, Homeostatic responses regulate selfish mitochondrial genome dynamics in *C. elegans* *Cell metabolism*, vol. 24, no. 1, pp. 91–103, 2016.
- [193] M. E. Lopez, N. L. Van Zeeland, D. B. Dahl, R. Weindruch, and J. M. Aiken, Cellular phenotypes of age-associated skeletal muscle mitochondrial abnormalities in rhesus monkeys *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, vol. 452, no. 1, pp. 123–138, 2000.
- [194] J. Avellaneda, C. Rodier, F. Daian, N. Brouilly, T. Rival, N. M. Luis, and F. Schnorrer, Myofibril and mitochondria morphogenesis are coordinated by a mechanical feedback mechanism in muscle *Nature communications*, vol. 12, no. 1, pp. 1–18, 2021.
- [195] B. Glancy, L. M. Hartnell, D. Malide, Z.-X. Yu, C. A. Combs, P. S. Connelly, S. Subra-

- maniam, and R. S. Balaban, Mitochondrial reticulum for cellular energy distribution in muscle *Nature*, vol. 523, no. 7562, pp. 617–620, 2015.
- [196] F. J. Iborra, H. Kimura, and P. R. Cook, The functional organization of mitochondrial genomes in human cells *BMC Biol.*, vol. 2, no. 1, p. 9, 2004.
- [197] P. M. Quirós, A. Mottis, and J. Auwerx, Mitonuclear communication in homeostasis and stress *Nature reviews Molecular cell biology*, vol. 17, no. 4, pp. 213–226, 2016.
- [198] M. A. Pellegrino, M. Canepari, R. Rossi, G. D’Antona, C. Reggiani, and R. Bottinelli, Orthologous myosin isoforms and scaling of shortening velocity with body size in mouse, rat, rabbit and human muscles *The Journal of Physiology*, vol. 546, no. 3, pp. 677–689, 2003.
- [199] S. A. Abdul Aziz, M. Mcstea, N. S. Ahmad Bashah, M. L. Chong, S. Ponnampalavanar, S. F. Syed Omar, H. Sulaiman, I. Azwa, M. P. Tan, A. Kamarulzaman, R. Rajasuriar, and S. B. Kamaruzzaman, Assessment of sarcopenia in virally suppressed HIV-infected Asians receiving treatment *AIDS (London, England)*, vol. 32, no. 8, pp. 1025–1034, 2018.
- [200] S. L. Di Stasi, T. D. MacLeod, J. D. Winters, and S. A. Binder-Macleod, Effects of statins on skeletal muscle: a perspective for physical therapists *Physical Therapy*, vol. 90, pp. 1530–1542, Oct. 2010.
- [201] R. Ramachandran and A. S. Wierzbicki, Statins, Muscle Disease and Mitochondria *Journal of Clinical Medicine*, vol. 6, July 2017.
- [202] A. Bielecka-Dabrowa, J. Fabis, D. P. Mikhailidis, S. von Haehling, A. Sahebkar, J. Rysz, and M. Banach, Prosarcopenic Effects of Statins May Limit Their Effectiveness in Patients with Heart Failure *Trends in Pharmacological Sciences*, vol. 39, no. 4, pp. 331–353, 2018.
- [203] H. Päivä, K. M. Thelen, R. V. Coster, J. Smet, B. D. Paepe, K. M. Mattila, J. Laakso, T. Lehtimäki, K. v. Bergmann, D. Lütjohann, and R. Laaksonen, High-dose statins and skeletal muscle metabolism in humans: A randomized, controlled trial *Clinical Pharmacology & Therapeutics*, vol. 78, no. 1, pp. 60–68, 2005.

- [204] H. A. J. Stringer, G. K. Sohi, J. A. Maguire, and H. C. F. Côté, Decreased skeletal muscle mitochondrial DNA in patients with statin-induced myopathy *Journal of the Neurological Sciences*, vol. 325, pp. 142–147, Feb. 2013.
- [205] A. Cirigliano, A. Amelina, B. Biferali, A. Macone, C. Mozzetta, M. M. Bianchi, M. Mori, B. Botta, E. Pick, R. Negri, and T. Rinaldi, Statins interfere with the attachment of *S. cerevisiae* mtDNA to the inner mitochondrial membrane *Journal of Enzyme Inhibition and Medicinal Chemistry*, vol. 35, pp. 129–137, Jan. 2020.
- [206] S.-Z. Yoo, M.-H. No, J.-W. Heo, D.-H. Park, J.-H. Kang, S. H. Kim, and H.-B. Kwak, Role of exercise in age-related sarcopenia *Journal of Exercise Rehabilitation*, vol. 14, pp. 551–558, Aug. 2018.
- [207] N. T. Theilen, G. H. Kunkel, and S. C. Tyagi, The Role of Exercise and TFAM in Preventing Skeletal Muscle Atrophy *Journal of Cellular Physiology*, vol. 232, pp. 2348–2358, Sept. 2017.
- [208] N. T. Theilen, N. Jeremic, G. J. Weber, and S. C. Tyagi, TFAM overexpression diminishes skeletal muscle atrophy after hindlimb suspension in mice *Archives of Biochemistry and Biophysics*, vol. 666, pp. 138–147, May 2019.
- [209] J.-P. Leduc-Gaudet, O. Reynaud, S. N. Hussain, and G. Gouspillou, Parkin overexpression protects from ageing-related loss of muscle mass and strength *The Journal of Physiology*, vol. 597, no. 7, pp. 1975–1991, 2019.
- [210] R. L, Propriétés et structures différentes des muscles rouges et des muscles blanc, chez les lapins et chez les raies. *CR Hebd Acad Sci*, vol. 77, p. 1030–1043, 1873.
- [211] T. Ogata and Y. Yamasaki, Scanning electron-microscopic studies on the three-dimensional structure of mitochondria in the mammalian red, white and intermediate muscle fibers *Cell and Tissue Research*, vol. 241, pp. 251–256, Aug. 1985.
- [212] L. He, P. F. Chinnery, S. E. Durham, E. L. Blakely, T. M. Wardell, G. M. Borthwick, R. W. Taylor, and D. M. Turnbull, Detection and quantification of mitochondrial DNA

- deletions in individual cells by real-time PCR *Nucleic Acids Research*, vol. 30, p. e68, July 2002.
- [213] L. LARSSON, Histochemical characteristics of human skeletal muscle during aging *Acta physiologica scandinavica*, vol. 117, no. 3, pp. 469–471, 1983.
- [214] A. Herbst, J. Wanagat, N. Cheema, K. Widjaja, D. McKenzie, and J. M. Aiken, Latent mitochondrial DNA deletion mutations drive muscle fiber loss at old age *Aging cell*, vol. 15, no. 6, pp. 1132–1139, 2016.
- [215] M. Picard, D. Ritchie, M. M. Thomas, K. J. Wright, and R. T. Hepple, Alterations in intrinsic mitochondrial function with aging are fiber type-specific and do not explain differential atrophy between muscles *Aging Cell*, vol. 10, no. 6, pp. 1047–1055, 2011.
- [216] J. D. Murray, *Mathematical Biology: I. An Introduction*. Interdisciplinary Applied Mathematics, Mathematical Biology, New York: Springer-Verlag, 3 ed., 2002.
- [217] R. H. Hämäläinen, T. Manninen, H. Koivumäki, M. Kislin, T. Otonkoski, and A. Suomalainen, Tissue-and cell-type-specific manifestations of heteroplasmic mtDNA 3243A>G mutation in human induced pluripotent stem cell-derived disease model *Proceedings of the National Academy of Sciences*, vol. 110, no. 38, pp. E3622–E3630, 2013.
- [218] J. Müller-Höcker, Cytochrome c oxidase deficient fibres in the limb muscle and diaphragm of man without muscular disease: an age-related alteration *Journal of the neurological sciences*, vol. 100, no. 1, pp. 14–21, 1990.
- [219] D. Cottrell, P. Ince, E. Blakely, M. Johnson, P. Chinnery, M. Hanna, and D. Turnbull, Neuropathological and histochemical changes in a multiple mitochondrial DNA deletion disorder *Journal of Neuropathology & Experimental Neurology*, vol. 59, no. 7, pp. 621–627, 2000.
- [220] Z. Y. Tam, J. Gruber, B. Halliwell, and R. Gunawan, Context-dependent role of mitochondrial fusion-fission in clonal expansion of mtDNA mutations *PLoS Comp. Biol.*, vol. 11, no. 5, p. e1004183, 2015.

- [221] G. López-Lluch, N. Hunt, B. Jones, M. Zhu, H. Jamieson, S. Hilmer, M. Cascajo, J. Al-lard, D. K. Ingram, P. Navas, *et al.*, Calorie restriction induces mitochondrial biogenesis and bioenergetic efficiency *Proceedings of the National Academy of Sciences*, vol. 103, no. 6, pp. 1768–1773, 2006.
- [222] A. E. Civitarese, S. Carling, L. K. Heilbronn, M. H. Hulver, B. Ukropcova, W. A. Deutsch, S. R. Smith, and E. Ravussin, Calorie restriction increases muscle mitochondrial biogenesis in healthy humans *PLoS medicine*, vol. 4, no. 3, p. e76, 2007.
- [223] N. A. Khan, M. Auranen, I. Paetau, E. Pirinen, L. Euro, S. Forsström, L. Pasila, V. Velagapudi, C. J. Carroll, J. Auwerx, *et al.*, Effective treatment of mitochondrial myopathy by nicotinamide riboside, a vitamin B 3 *EMBO molecular medicine*, vol. 6, no. 6, pp. 721–731, 2014.
- [224] E. Pirinen, M. Auranen, N. A. Khan, V. Brilhante, N. Urho, A. Pessia, A. Hakkarainen, J. Kuula, U. Heinonen, M. S. Schmidt, *et al.*, Niacin cures systemic NAD⁺ deficiency and improves muscle performance in adult-onset mitochondrial myopathy *Cell Metabolism*, vol. 31, no. 6, pp. 1078–1090, 2020.
- [225] A. B. Knott, G. Perkins, R. Schwarzenbacher, and E. Bossy-Wetzel, Mitochondrial fragmentation in neurodegeneration *Nature Reviews Neuroscience*, vol. 9, no. 7, pp. 505–518, 2008.
- [226] M. N. Serasinghe and J. E. Chipuk, Mitochondrial fission in human diseases *Pharmacology of Mitochondria*, pp. 159–188, 2016.
- [227] A. Terman, T. Kurz, M. Navratil, E. A. Arriaga, and U. T. Brunk, Mitochondrial turnover and aging of long-lived postmitotic cells: the mitochondrial–lysosomal axis theory of aging *Antioxidants & redox signaling*, vol. 12, no. 4, pp. 503–535, 2010.
- [228] R. Filograna, M. Mennuni, D. Alsina, and N.-G. Larsson, Mitochondrial DNA copy number in human disease: the more the better? *FEBS letters*, vol. 595, no. 8, pp. 976–1002, 2021.

- [229] D. Ryu, L. Mouchiroud, P. A. Andreux, E. Katsyuba, N. Moullan, A. A. Nicolet-dit Félix, E. G. Williams, P. Jha, G. Lo Sasso, D. Huzard, *et al.*, Urolithin A induces mitophagy and prolongs lifespan in *C. elegans* and increases muscle function in rodents *Nature Medicine*, vol. 22, no. 8, pp. 879–888, 2016.
- [230] T. L. Schwarz, Mitochondrial trafficking in neurons *Cold Spring Harbor Perspectives in Biology*, vol. 5, no. 6, p. a011304, 2013.
- [231] D. Torralba, F. Baixauli, and F. Sánchez-Madrid, Mitochondria know no boundaries: mechanisms and functions of intercellular mitochondrial transfer *Front. Cell Dev. Biol.*, vol. 4, p. 107, 2016.
- [232] T. G. Clark, M. J. Bradburn, S. B. Love, and D. G. Altman, Survival analysis part I: basic concepts and first analyses *British Journal of Cancer*, vol. 89, no. 2, pp. 232–238, 2003.
- [233] M. J. Bradburn, T. G. Clark, S. B. Love, and D. G. Altman, Survival analysis part II: multivariate data analysis—an introduction to concepts and methods *British Journal of Cancer*, vol. 89, no. 3, pp. 431–436, 2003.
- [234] D. G. Kleinbaum, M. Klein, *et al.*, *Survival analysis: a self-learning text*, vol. 3. Springer, 2012.
- [235] W. J. Shih, Problems in dealing with missing data and informative censoring in clinical trials *Current Controlled Trials in Cardiovascular Medicine*, vol. 3, no. 1, pp. 1–7, 2002.
- [236] R. Peto and J. Peto, Asymptotically efficient rank invariant test procedures *Journal of the Royal Statistical Society: Series A (General)*, vol. 135, no. 2, pp. 185–198, 1972.
- [237] S. Sawyer, The Greenwood and exponential Greenwood confidence intervals in survival analysis *Applied Survival Analysis: Regression Modeling of Time to Event Data*, pp. 1–14, 2003.
- [238] R. C. Rostomily, A. M. Spence, D. Duong, K. McCormick, M. Bland, and M. S. Berger, Multimodality management of recurrent adult malignant gliomas: results of a phase

- II multiagent chemotherapy study and analysis of cytoreductive surgery *Neurosurgery*, vol. 35, no. 3, pp. 378–388, 1994.
- [239] D. Schoenfeld, Partial residuals for the proportional hazards regression model *Biometrika*, vol. 69, no. 1, pp. 239–241, 1982.
- [240] D. R. Cox, Regression models and life-tables *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 187–202, 1972.
- [241] D. R. Cox, Partial likelihood *Biometrika*, vol. 62, no. 2, pp. 269–276, 1975.
- [242] Z. Zhang, J. Reinikainen, K. A. Adeleke, M. E. Pieterse, and C. G. Groothuis-Oudshoorn, Time-varying covariates and coefficients in Cox regression models *Annals of Translational Medicine*, vol. 6, no. 7, 2018.
- [243] L. Tian, D. Zucker, and L. Wei, On the Cox model with time-varying regression coefficients *Journal of the American Statistical Association*, vol. 100, no. 469, pp. 172–183, 2005.
- [244] C. A. Bellera, G. MacGrogan, M. Debled, C. T. De Lara, V. Brouste, and S. Mathoulin-Pélissier, Variables with time-varying effects and the Cox model: some statistical concepts illustrated with a prognostic factor study in breast cancer *BMC Medical Research Methodology*, vol. 10, no. 1, pp. 1–12, 2010.
- [245] G. Rodríguez, Multilevel generalized linear models in *Handbook of multilevel analysis*, pp. 335–376, Springer, 2008.
- [246] T. R. Holford, The analysis of rates and of survivorship using log-linear models *Biometrics*, pp. 299–305, 1980.
- [247] N. Laird and D. Olivier, Covariance analysis of censored survival data using log-linear analysis techniques *Journal of the American Statistical Association*, vol. 76, no. 374, pp. 231–240, 1981.
- [248] N. Breslow, Covariance analysis of censored survival data *Biometrics*, pp. 89–99, 1974.

- [249] O. O. Aalen, P. K. Andersen, Ø. Borgan, R. D. Gill, and N. Keiding, History of applications of martingales in survival analysis *arXiv preprint arXiv:1003.0188*, 2010.
- [250] R. P. Payne, S. Longet, J. A. Austin, D. T. Skelly, W. Dejnirattisai, S. Adele, N. Meardon, S. Faustini, S. Al-Taei, S. C. Moore, *et al.*, Immunogenicity of standard and extended dosing intervals of BNT162b2 mRNA vaccine *Cell*, vol. 184, no. 23, pp. 5699–5714, 2021.
- [251] R. L. Williams, A note on robust variance estimation for cluster-correlated data *Biometrics*, vol. 56, no. 2, pp. 645–646, 2000.
- [252] E. B. Douple, K. Mabuchi, H. M. Cullings, D. L. Preston, K. Kodama, Y. Shimizu, S. Fujiwara, and R. E. Shore, Long-term radiation-related health effects in a unique human population: lessons learned from the atomic bomb survivors of Hiroshima and Nagasaki *Disaster Medicine and Public Health Preparedness*, vol. 5, no. S1, pp. S122–S133, 2011.
- [253] R. A. Betensky and M. Mandel, Recognizing the problem of delayed entry in time-to-event studies: better late than never for clinical neuroscientists *Annals of Neurology*, vol. 78, no. 6, pp. 839–844, 2015.
- [254] M. Matsuura and S. Eguchi, Modeling late entry bias in survival analysis *Biometrics*, vol. 61, no. 2, pp. 559–566, 2005.
- [255] M. Lipsitch, E. Goldstein, G. T. Ray, and B. Fireman, Depletion-of-susceptibles bias in influenza vaccine waning studies: how to ensure robust results *Epidemiology & Infection*, vol. 147, 2019.
- [256] M. Lipsitch, Challenges of vaccine effectiveness and waning studies 2019.
- [257] B. A. Lopman and V. E. Pitzer, Waxing understanding of waning immunity *The Journal of Infectious Diseases*, vol. 217, no. 6, pp. 851–853, 2018.
- [258] E. T. Rogawski, J. A. Platts-Mills, E. R. Colgate, R. Haque, K. Zaman, W. A. Petri, and B. D. Kirkpatrick, Quantifying the impact of natural immunity on rotavirus vaccine

- efficacy estimates: a clinical trial in Dhaka, Bangladesh (PROVIDE) and a simulation study *The Journal of Infectious Diseases*, vol. 217, no. 6, pp. 861–868, 2018.
- [259] G. T. Ray, N. Lewis, N. P. Klein, M. F. Daley, M. Lipsitch, and B. Fireman, Depletion-of-susceptibles bias in analyses of intra-season waning of influenza vaccine effectiveness *Clinical Infectious Diseases*, vol. 70, no. 7, pp. 1484–1486, 2020.
- [260] R. Kahn, S. J. Schrag, J. R. Verani, and M. Lipsitch, Identifying and alleviating bias due to differential depletion of susceptible people in postmarketing evaluations of COVID-19 vaccines *American Journal of Epidemiology*, vol. 191, no. 5, pp. 800–811, 2022.
- [261] J. L. Bernal, N. Andrews, C. Gower, E. Gallagher, R. Simmons, S. Thelwall, J. Stowe, E. Tessier, N. Groves, G. Dabrera, *et al.*, Effectiveness of Covid-19 vaccines against the B. 1.617. 2 (Delta) variant *New England Journal of Medicine*, 2021.
- [262] L. Swadling, M. O. Diniz, N. M. Schmidt, O. E. Amin, A. Chandran, E. Shaw, C. Pade, J. M. Gibbons, N. Le Bert, A. T. Tan, *et al.*, Pre-existing polymerase-specific T cells expand in abortive seronegative SARS-CoV-2 *Nature*, vol. 601, no. 7891, pp. 110–117, 2022.
- [263] J. P. Townsend, H. B. Hassler, Z. Wang, S. Miura, J. Singh, S. Kumar, N. H. Ruddle, A. P. Galvani, and A. Dornburg, The durability of immunity against reinfection by SARS-CoV-2: a comparative evolutionary study *The Lancet Microbe*, vol. 2, no. 12, pp. e666–e675, 2021.
- [264] G. Milne, T. Hames, C. Scotton, N. Gent, A. Johnsen, R. M. Anderson, and T. Ward, Does infection with or vaccination against SARS-CoV-2 lead to lasting immunity? *The Lancet Respiratory Medicine*, vol. 9, no. 12, pp. 1450–1466, 2021.
- [265] Z. Wang, F. Muecksch, D. Schaefer-Babajew, S. Finkin, C. Viant, C. Gaebler, H.-H. Hoffmann, C. O. Barnes, M. Cipolla, V. Ramos, *et al.*, Naturally enhanced neutralizing breadth against SARS-CoV-2 one year after infection *Nature*, vol. 595, no. 7867, pp. 426–431, 2021.

- [266] C. H. Hansen, D. Michlmayr, S. M. Gubbels, K. Mølbak, and S. Ethelberg, Assessment of protection against reinfection with SARS-CoV-2 among 4 million PCR-tested individuals in Denmark in 2020: a population-level observational study *The Lancet*, vol. 397, no. 10280, pp. 1204–1212, 2021.
- [267] S. Gazit, R. Shlezinger, G. Perez, R. Lotan, A. Peretz, A. Ben-Tov, E. Herzel, H. Alapi, D. Cohen, K. Muhsen, *et al.*, SARS-CoV-2 Naturally Acquired Immunity vs. Vaccine-induced Immunity, Reinfections versus Breakthrough Infections: a Retrospective Cohort Study. *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America*, pp. ciac262–ciac262, 2022.
- [268] A. M. Cavanaugh, K. B. Spicer, D. Thoroughman, C. Glick, and K. Winter, Reduced risk of reinfection with SARS-CoV-2 after COVID-19 vaccination—Kentucky, May–June 2021 *Morbidity and Mortality Weekly Report*, vol. 70, no. 32, p. 1081, 2021.
- [269] M. Murugesan, P. Mathews, H. Paul, R. Karthik, J. J. Mammen, and P. Rupali, Protective effect conferred by prior infection and vaccination on COVID-19 in a healthcare worker cohort in South India *PLoS One*, vol. 17, no. 5, p. e0268797, 2022.
- [270] L. J. Abu-Raddad, H. Chemaitelly, H. H. Ayoub, H. M. Yassine, F. M. Benslimane, H. A. Al Khatib, P. Tang, M. R. Hasan, P. Coyle, Z. Al Kanaani, *et al.*, Association of prior SARS-CoV-2 infection with risk of breakthrough infection following mRNA vaccination in Qatar *Jama*, vol. 326, no. 19, pp. 1930–1939, 2021.
- [271] T. De Smedt, E. Merrall, D. Macina, S. Perez-Vilar, N. Andrews, and K. Bollaerts, Bias due to differential and non-differential disease-and exposure misclassification in studies of vaccine effectiveness *PLoS One*, vol. 13, no. 6, p. e0199180, 2018.
- [272] J. Shang, Y. Wan, C. Luo, G. Ye, Q. Geng, A. Auerbach, and F. Li, Cell entry mechanisms of SARS-CoV-2 *Proceedings of the National Academy of Sciences*, vol. 117, no. 21, pp. 11727–11734, 2020.
- [273] R. S. Paton, C. E. Overton, and T. Ward, The rapid replacement of the SARS-CoV-2

- Delta variant by Omicron (B.1.1.529) in England *Science Translational Medicine*, vol. 14, no. 652, p. eabo5395, 2022.
- [274] N. Andrews, J. Stowe, F. Kirsebom, S. Toffa, T. Rickeard, E. Gallagher, C. Gower, M. Kall, N. Groves, A.-M. O’Connell, *et al.*, Covid-19 vaccine effectiveness against the Omicron (B. 1.1. 529) variant *New England Journal of Medicine*, vol. 386, no. 16, pp. 1532–1546, 2022.
- [275] A. S. Luring, M. W. Tenforde, J. D. Chappell, M. Gaglani, A. A. Ginde, T. McNeal, S. Ghamande, D. J. Douin, H. K. Talbot, J. D. Casey, *et al.*, Clinical severity of, and effectiveness of mRNA vaccines against, covid-19 from omicron, delta, and alpha SARS-CoV-2 variants in the United States: prospective observational study *BMJ*, vol. 376, 2022.
- [276] S. P. Andeweg, B. de Gier, D. Eggink, C. van den Ende, N. van Maarseveen, L. Ali, B. Vlaemynck, R. Schepers, S. J. Hahné, C. Reusken, *et al.*, Protection of COVID-19 vaccination and previous infection against Omicron BA. 1 and Delta SARS-CoV-2 infections, the Netherlands, 22 November 2021–19 January 2022 *medRxiv*, 2022.
- [277] J. R. Pulliam, C. van Schalkwyk, N. Govender, A. von Gottberg, C. Cohen, M. J. Groome, J. Dushoff, K. Mlisana, and H. Moultrie, Increased risk of SARS-CoV-2 reinfection associated with emergence of Omicron in South Africa *Science*, vol. 376, no. 6593, p. eabn4947, 2022.
- [278] N. Ferguson, A. Ghani, A. Cori, A. Hogan, W. Hinsley, and E. Volz, Report 49: Growth, population distribution and immune escape of Omicron in England *Imperial College London*, vol. 10, 2021.
- [279] T. Nyberg, N. M. Ferguson, S. G. Nash, H. H. Webster, S. Flaxman, N. Andrews, W. Hinsley, J. L. Bernal, M. Kall, S. Bhatt, *et al.*, Comparative analysis of the risks of hospitalisation and death associated with SARS-CoV-2 Omicron (B. 1.1. 529) and Delta (B. 1.617. 2) variants in England: a cohort study *The Lancet*, vol. 399, no. 10332, pp. 1303–1312, 2022.

- [280] H. N. Altarawneh, H. Chemaitelly, H. Ayoub, M. R. Hasan, P. Coyle, H. M. YASSINE, H. A. Al Khatib, F. Benslimane, Z. Al-Kanaani, E. Al Kuwari, *et al.*, Protection of SARS-CoV-2 natural infection against reinfection with the BA. 4 or BA. 5 Omicron subvariants *medRxiv*, 2022.
- [281] H. Tegally, M. Moir, J. Everatt, M. Giovanetti, C. Scheepers, E. Wilkinson, K. Subramoney, Z. Makatini, S. Moyo, D. G. Amoako, *et al.*, Emergence of SARS-CoV-2 Omicron lineages BA. 4 and BA. 5 in South Africa *Nature medicine*, pp. 1–6, 2022.
- [282] A. Neri, D. Welch, T. Kawaguchi, and G. L. Nicolson, Development and biologic properties of malignant cell sublines and clones of a spontaneously metastasizing rat mammary adenocarcinoma *Journal of the National Cancer Institute*, vol. 68, no. 3, pp. 507–517, 1982.
- [283] P. Gerlee and A. R. Anderson, The evolution of carrying capacity in constrained and expanding tumour cell populations *Physical Biology*, vol. 12, no. 5, p. 056001, 2015.
- [284] A. D. Usher, A beautiful idea: how COVAX has fallen short *The Lancet*, vol. 397, no. 10292, pp. 2322–2325, 2021.
- [285] V. Pilkington, S. M. Keestra, and A. Hill, Global COVID-19 vaccine inequity: failures in the first year of distribution and potential solutions for the future *Frontiers in Public Health*, vol. 10, 2022.
- [286] A. Chhibber, A. Kharat, K. Duong, R. E. Nelson, M. H. Samore, F. A. Wilson, and N. Chaiyakunapruk, Strategies to minimize inequity in COVID-19 vaccine access in the US: Implications for future vaccine rollouts *The Lancet Regional Health–Americas*, vol. 7, 2022.
- [287] D. S. Khoury, D. Cromer, A. Reynaldi, T. E. Schlub, A. K. Wheatley, J. A. Juno, K. Subbarao, S. J. Kent, J. A. Triccas, and M. P. Davenport, Neutralizing antibody levels are highly predictive of immune protection from symptomatic SARS-CoV-2 infection *Nature medicine*, vol. 27, no. 7, pp. 1205–1211, 2021.

- [288] K. A. Earle, D. M. Ambrosino, A. Fiore-Gartland, D. Goldblatt, P. B. Gilbert, G. R. Siber, P. Dull, and S. A. Plotkin, Evidence for antibody as a protective correlate for COVID-19 vaccines *Vaccine*, vol. 39, no. 32, pp. 4423–4428, 2021.
- [289] S. S. A. Karim, Vaccines and SARS-CoV-2 variants: the urgent need for a correlate of protection *The Lancet*, vol. 397, no. 10281, pp. 1263–1264, 2021.
- [290] F. Krammer, A correlate of protection for SARS-CoV-2 vaccines is urgently needed *Nature medicine*, vol. 27, no. 7, pp. 1147–1148, 2021.
- [291] T. Koch, S. C. Mellinghoff, P. Shamsrizi, M. M. Addo, and C. Dahlke, Correlates of vaccine-induced protection against SARS-CoV-2 *Vaccines*, vol. 9, no. 3, p. 238, 2021.
- [292] S. A. Plotkin, Correlates of protection induced by vaccination *Clinical and Vaccine Immunology*, vol. 17, no. 7, pp. 1055–1065, 2010.
- [293] A. Atti, M. Ferrari, J. Castillo-Olivares, E. Monk, R. Gopal, M. Patel, K. Hoschler, M. Cole, A. Semper, J. Hewson, *et al.*, Serological profile of first SARS-CoV-2 reinfection cases detected within the SIREN study *Journal of Infection*, vol. 84, no. 2, pp. 248–288, 2022.
- [294] D. J. M. Fernández-Ayala, P. Navas, and G. López-Lluch, Age-related mitochondrial dysfunction as a key factor in COVID-19 disease *Experimental Gerontology*, vol. 142, p. 111147, 2020.
- [295] S.-J. Kang and S. I. Jung, Age-related morbidity and mortality among patients with COVID-19 *Infection & Chemotherapy*, vol. 52, no. 2, p. 154, 2020.
- [296] D. R. Petretto and R. Pili, Ageing and COVID-19: what is the role for elderly people? *Geriatrics*, vol. 5, no. 2, p. 25, 2020.
- [297] C. S. Greene and O. G. Troyanskaya, Chapter 2: Data-driven view of disease biology *PLoS Computational Biology*, vol. 8, no. 12, p. e1002816, 2012.
- [298] V. Marx, The big challenges of big data *Nature*, vol. 498, no. 7453, pp. 255–260, 2013.

- [299] O. J. Watson, G. Barnsley, J. Toor, A. B. Hogan, P. Winskill, and A. C. Ghani, Global impact of the first year of COVID-19 vaccination: a mathematical modelling study *The Lancet Infectious Diseases*, vol. 22, no. 9, pp. 1293–1302, 2022.
- [300] I. G. Johnston, B. C. Rickett, and N. S. Jones, Explicit tracking of uncertainty increases the power of quantitative rule-of-thumb reasoning in cell biology *Biophysical Journal*, vol. 107, no. 11, pp. 2612–2617, 2014.
- [301] N. J. Gross, G. S. Getz, and M. Rabinowitz, Apparent turnover of mitochondrial deoxyribonucleic acid and mitochondrial phospholipids in the tissues of the rat *Journal of Biological Chemistry*, vol. 244, no. 6, pp. 1552–1562, 1969.
- [302] D. Neubert, E. Oberdisse, and R. Bass, *Biochemical aspects of the biogenesis of mitochondria*. Adriatica Editrice, 1969.
- [303] F. Diaz and C. T. Moraes, Mitochondrial biogenesis and turnover *Cell Calcium*, vol. 44, pp. 24–35, July 2008.
- [304] J. P. Burgstaller, I. G. Johnston, N. S. Jones, J. Albrechtová, T. Kolbe, C. Vogl, A. Futschik, C. Mayrhofer, D. Klein, S. Sabitzer, *et al.*, MtDNA segregation in heteroplasmic tissues is common in vivo and modulated by haplotype differences and developmental stage *Cell Rep.*, vol. 7, no. 6, pp. 2031–2041, 2014.
- [305] J. Bruusgaard, K. Liestøl, M. Ekmark, K. Kollstad, and K. Gundersen, Number and spatial distribution of nuclei in the muscle fibres of normal mice studied in vivo *The Journal of Physiology*, vol. 551, no. 2, pp. 467–478, 2003.
- [306] R. Erban, J. Chapman, and P. Maini, A practical guide to stochastic simulations of reaction-diffusion processes *arXiv preprint arXiv:0704.1908*, 2007.

Appendix A

Appendix to Chapter 2

A.1 Phenomenological formula for wave speed

In section 2.5 we have shown that how our model, in the limits $\gamma \rightarrow 0$ and $\delta \rightarrow 1$, leads to an analytical description of a noise-driven wave. Moreover, through numerical simulations we have shown that the wave-like expansion is exhibited by our model without having to assume the above limits. We have performed simulations of a 201-unit chain with 110 combination of the parameter values, with $\delta \in [0.1, 1]$, $N_{ss} \in [2, 35]$, $\gamma \in [0, 0.25]$, $\mu \in [0.01, 0.15]$. For each parameter configuration we have observed a wave-like expansion, except, as expected, when $\delta = 1$ or $\gamma = 0$. We have calculated the wave speed plotting the wavefront at different times and measuring the distance covered, measuring the degree of shift between the two wavefronts. In some cases, the overlap could not be total since the wavefront's steepness changed slightly over time: in those cases we have looked for an overlap of the midpoint of the curves (i.e., the point for which $\langle h \rangle = 0.5$). In all cases, it has been verified that the speed is constant, by measuring it for several different time intervals. We have found that v can be predicted with excellent accuracy through the formula

$$v = 2\sqrt{kD} + \beta. \tag{A.1}$$

Apart from the intercept, Eq. (A.1) is analogous to the formula for the wave speed of FK waves, with k being an effective reaction rate induced by SSD, given by

$$k = k_{SSD} = \frac{\sqrt{(1-\delta)^2 \mu \gamma}}{N_{ss}^{\frac{2}{3}}}, \quad (\text{A.2})$$

and $\beta = 0.207 \pm 0.007$ is calculated via MLE (uncertainty is standard deviation). In Fig. A.1 we show the fit, for which $R^2 = 0.99$, and we observe residuals without any clear pattern. However, the presence of an intercept $\beta \neq 0$ means that Eq. (A.1) is not accurate for $x \rightarrow 0$, for which $v = 0$.

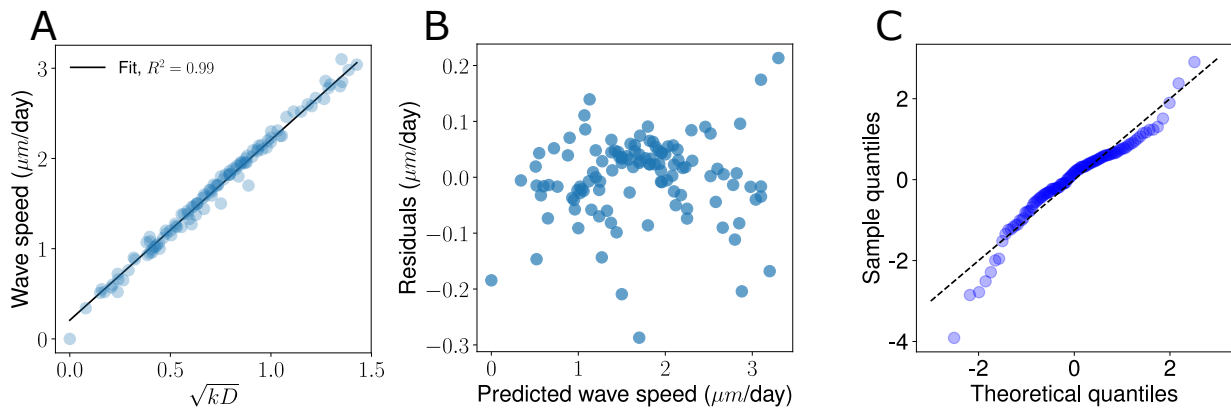


Figure A.1: *An empirical formula accurately predicts the survival of the densest wave speed obtained through stochastic simulations for a wide range of parameter values.* A) The blue dots are 110 values of wave speed measured through stochastic simulations of the linear feedback control for varying values of the parameters N_{ss}, μ, γ and $0 < \delta < 1$. The black line is the linear fit against the variable $x = \sqrt{kD} = (1-\delta)(\mu\gamma^3)^{\frac{1}{4}}/N_{ss}^{\frac{1}{3}}$. The excellent fit ($R^2 = 0.99$) shows that the wave speed v is well predicted by Eq. (A.1), except for $x \rightarrow 0$. B) Although the residuals do not follow a Gaussian distribution (Shapiro-Wilk, $p < 10^{-4}$), the absence of a clear pattern in the plot suggests that Eq. (A.1) is a satisfactory approximation in the parameter range considered. C) The Q-Q plot of the residuals against a normal distribution confirms the absence of a clear pattern.

A.2 Parameter values for simulations shown in the chapter

The parameter values used in the simulations of Chapter 2 are not derived from the scientific literature and are not meant to describe a realistic system. Rather, they are chosen to show

the increase in mean mutant copy number and the increase in mean heteroplasmy with a limited computational effort in simple systems. Realistic parameter values for simulating skeletal muscle fibres modelled as a chain of units are discussed and used in Chapter 3.

A.2.1 Fig. 2.1

All panels in Fig. 2.1 refer to systems with linear replication rate $\lambda(w, m)$ given by Eq. (2.1) and as reported in panel A.

Fig. 2.1B depicts the steady state line (red) for a single-unit deterministic system (Eq. (2.3)) with $N_{ss} = 500, \delta = 0.5$ (red line). In blue we depict a possible trajectory of the stochastic system (Eq. (2.12)).

Fig. 2.1C refers to the single-unit system. Data are obtained by simulating the deterministic ODE system Eq. (2.3) (black line) and the chemical reaction network Eq. (2.12), with linear replication rate λ given by Eq. (2.1), via the Gillespie algorithm (other lines) with parameters $\mu = 7 \cdot 10^{-2}/\text{day}$, $\delta = 10^{-1}$, $c = 2.5 \cdot 10^{-3}/\text{day}$, $N_{ss} = 325$. The system is initialised at steady state with $w_0 = 300, m_0 = 250$ corresponding to an initial heteroplasmy $h_0 \approx 0.45$. For the deterministic (black) and neutral stochastic (red) cases, mutants and wildtypes have the same degradation rate μ . For the cases of slow and fast preferential elimination, the degradation rate of mutants is increased by a constant ϵ (see Eq. (2.18)), with value $\epsilon = 2.935 \cdot 10^{-4}/\text{day}$ for slow elimination (green) and $\epsilon = 1.1735 \cdot 10^{-3}/\text{day}$ for fast elimination (blue). For the deterministic system we plot mutant copy number m ; for the stochastic systems we plot $\langle m \rangle$, averaging over an ensemble of 10^4 simulations. Error bars are standard error of the mean (SEM).

Fig. 2.1D is relative to the same simulation, with same replication rate and same parameter values, of the same system as panel B, but plotting (mean) heteroplasmy. For the deterministic system we plot heteroplasmy h ; for the stochastic systems we plot mean heteroplasmy $\langle h \rangle$, averaging over an ensemble of 10^4 simulations. Error bars are standard error of the mean (SEM).

A.2.2 Fig. 2.2

All panels refer to systems with linear replication rate $\lambda(w, m)$ (Eq. (2.1)), reported also at the bottom of panel A.

Fig. 2.2B refers to the two-unit system. One unit is initialised with zero heteroplasmy $w_{10} = 455, m_{10} = 0$; the other with $w_{20} = 450, m_{20} = 50$, corresponding to an initial heteroplasmy of 0.1. The black line refers to the deterministic ODE system Eq. (2.37). Other lines refer to the chemical reaction network Eq. (2.38). Common parameters of all lines are $\gamma = \mu = 7 \cdot 10^{-2}/\text{day}$, $\delta = 10^{-1}$, $c = 10^{-3}/\text{day}$, $N_{ss} = 455$. The chemical reaction network is simulated via the Gillespie algorithm. For the deterministic (black) and neutral stochastic (red) cases mutants and wildtypes have the same degradation rate μ . For the cases of slow and fast preferential elimination, the degradation rate of mutants is increased by a constant ϵ (see Eq. (2.18)), with value $\epsilon = 1.5 \cdot 10^{-3}/\text{day}$ for slow elimination (green) and $\epsilon = 6.25 \cdot 10^{-3}/\text{day}$ for fast elimination (blue). For the deterministic system the quantity plotted is $h_{mean} = (h_1 + h_2)/2$, the mean of the heteroplasmy of the two units. For the stochastic system, we plot $\langle h_{mean} \rangle = (\langle h_1 + h_2 \rangle)/2$, averaging over an ensemble of $3.6 \cdot 10^4$ simulations. Error bars are standard error of the mean (SEM).

In Fig. 2.2C, all simulations are run for $7 \cdot 10^2$ days. All systems have $\mu = 7 \cdot 10^{-2}/\text{day}$, $c = 10^{-3}/\text{day}$. All units are initialised (in steady state) with copy number $(w_0, m_0) = 100$, corresponding to an initial heteroplasmy $h_0 = 0.5$. In all subpanels, the black dots refer to the neutral system, the green dots refer to the system with preferential elimination of mutants (Eq. (2.18)), with $\epsilon = 10^{-3}\mu$. In the bottom-right subpanel, the neutral system is plotted in red to highlight the occurrence of SSD. We plot the (mean) heteroplasmy of one unit; the other unit is identical in the deterministic system or follows closely in the stochastic cases. Parameters that vary by panel are specified below. Averages have been taken over $2 \cdot 10^3$ realisations. Error bars are SEM.

The top-left subpanel refers to the two-unit deterministic system (Eq. (2.37)), with the above parameters and $N_{ss} = 150, \delta = 0.5, \gamma = 7 \cdot 10^{-4}/\text{day}$. The top-right subpanel refers to the

single-unit stochastic system (Eq. (2.12)), with $N_{ss} = 150, \delta = 0.5$. The bottom-left subpanel refers to the two-unit stochastic system (Eq. (2.38)), with $N_{ss} = 200, \delta = 1, \gamma = 3.5 \cdot 10^{-2}/\text{day}$. The bottom-right subpanel refers to the two-unit stochastic system (Eq. (2.38)), with $N_{ss} = 150, \delta = 0.5, \gamma = 3.5 \cdot 10^{-2}/\text{day}$.

A.2.3 Fig. 2.3

Fig. 2.3 refers to a chain of 340 units, with linear replication rate $\lambda(w, m)$ (Eq. (2.1)), where $N_{ss} = 4, \mu = 7 \cdot 10^{-2}/\text{day}, c = 0.4/\text{day}$ and δ varies by subpanel as explained below. Diffusion is parameterised by the hopping rate $\gamma = 0.12/\text{day}$. Ensemble averages are taken over $2 \cdot 10^3$ stochastic replicates. We did not plot error bars for clarity, as here we are focusing on the advancing wavefront.

The left subpanel refers to a deterministic system with $\delta = 2/3$, and is obtained simulating an ODE system analogous to Eq. (2.37). The mid subpanel refers to a stochastic system, obtained simulating a chemical reaction network analogous to Eq. (2.38) via Gillespie algorithm, with $\delta = 1$. Same for the the right subpanel, but with $\delta = 2/3$.

Fig. 2.3B is analogous to the right subpanel of Fig. 2.3A and with same parameter values, except that preferential elimination of mutants (Eq. (2.18)) has been added, with $\epsilon = 7.5 \cdot 10^{-2}\mu$.

A.2.4 Fig. 2.4

The two panels in Fig. 2.4 are obtained via stochastic simulations of a chain of units, again modelled via a chemical reaction network analogous to Eq. (2.38) and simulated via Gillespie algorithm. Ensemble averages are taken over 400 stochastic replicates. The difference from Fig. 2.3 is that the replication rate is not the linear feedback control of Eq. (2.2), but is Eq. (2.47) and Eq. (2.48) for panels A and B respectively. For panel A, $\gamma = 0.12/\text{day}$, and parameter values in Eq. (2.47) are $\mu = 7 \cdot 10^{-2}/\text{day}, \delta = 2/3, a = 0.4/\text{day}, N_{ss} = 4$. For panel B $\gamma = 10^{-2}/\text{day}$, and parameter values in Eq. (2.48) are $\mu = 7 \cdot 10^{-2}/\text{day}, \delta = 0.1, b = 1.5/\text{day}, N_{ss} = 3$.

A.2.5 Fig. 2.5

This Figure is a reproduction of Fig. 2.1C, proving that stochastic survival of the densest is observed in a different modelling framework than the continuous-time Poisson Point Process used throughout the thesis. The system is made up of two coupled units (with hopping of molecules) evolving according to the Wright-Fisher model, with discrete non-overlapping generations and binomial sampling from parental generation and for molecules which hop to a neighbouring unit in each generation. This was an idea of the anonymous reviewer of Ref. [123].

Parameters are identical between panels, except for a single parameter values that switches off one element of SSD at a time, with all elements present in the bottom right panel. Specifically: $c = 0.5$; $N_{ss} = 100$, except in top left subpanel where $N_{ss} = 5 \cdot 10^5$ (deterministic limit, effectively no noise); $\gamma = 0.1$, except in the top right where $\gamma = 0$ (no hopping = no spatial structure); $\delta = 0.2$, except in the bottom left where $\delta = 1$ (no higher density). Selective elimination is $\epsilon = 0$ for black/red lines and $\epsilon = 0.002$ for green lines. The hopping rate *gamma* is here the trial success probability of the binomial used to sample molecules that hop to the neighbouring unit. Initial condition fixed by $w_0 + \delta m_0 = N_{ss} = 100$ (steady state). Averaged over 4000 replicates.

A.3 Statistical Methods

In Fig. 2.1B, C and D, data from numerical simulations are plotted. Error bars are SEM and averages are taken over a sample of 10^4 simulations. Details in Section A.2.

In Fig. 2.4, synthetic data is plotted, averaged over an ensemble of 400 simulations; more details in Section 2.6.

In Fig. A.1A, there are 110 points, and linear least squares was used to fit the expression $2\sqrt{kD} + \beta$, obtaining $\beta = 0.207 \pm 0.007$, with effect size $R^2 = 0.99$. Uncertainty reported is the standard error. The variable k is defined in Chapter 2 (Section A.1) and D is the diffusion

coefficient. The only parameter estimated is β .

A.4 Code availability

Code is freely available at <https://github.com/ferdinando17/Survival-of-the-Densest>. The C code was used to produce results from stochastic simulations. The parameter values currently specified allows the reproduction of Fig. 2.2B, the simplest system in which survival of the densest is observed. The number of units is tuneable ($n \geq 1$), and other plots can be reproduced with the parameter and instructions specified in this work (Sections A.2, B.2).

A Python 3 script is available as Jupyter Notebook, producing stochastic survival of the densest in a minimal two-unit system and only requires NumPy to run and Matplotlib for plotting.

An R script, based on the code by the anonymous reviewer of Ref. [123], is available as a Jupyter Notebook to produce Fig. 2.5.

Finally, we provide a Mathematica notebook of the stochastic dimensionality reduction procedure detailed in Section 2.2.4.

Appendix B

Appendix to Chapter 3

B.1 Additional heteroplasmy profiles

In Figs. 3.1A, 3.2A, C we plot data from experimental studies in humans¹⁶ and rats¹⁷, showing that the profiles of the high-heteroplasmy region in skeletal muscle fibres are compatible with a sigmoidal shapes, the form expected for a traveling wave. In Fig. B.1 we report more experimental data from these studies, showing that the profiles are compatible with a sigmoidal shape. We also include the fibers in Figs. 3.1A, 3.2A, C for comparison. We reported and fitted a sigmoid (via MLE) for profiles with at least three points in the transition regions (from high to low heteroplasmy), as fewer than three points would trivially be compatible with any shape.

The two pairs of fibres we have chosen for the quantitative comparison in Fig. 3.2 are those with the best measurements of copy number in the wildtype-only region of the fibre. This allows us to estimate N_{ss} more reliably. We notice that one of the rat fibres branches into three at some point along its length (rat fibre 3, fibre three in¹⁷; here we plot one branch). It is not a typical fibre and we deemed it unsuitable for a quantitative comparison, as different branches would have different steepness value.

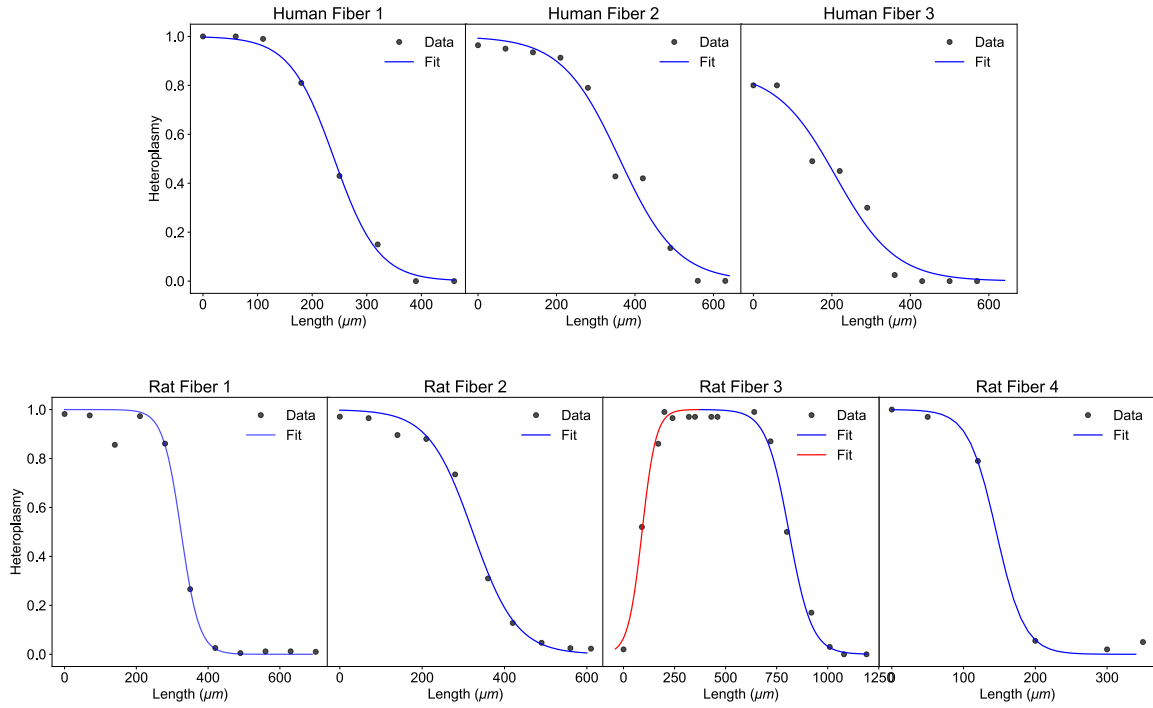


Figure B.1: Data on the heteroplasmy profiles in skeletal muscle fibres from published studies in humans¹⁶ and rats¹⁷ are compatible with a sigmoidal shape. We fit a sigmoid to the transition region from all-mutant to all-wildtype. We did not consider profiles with fewer than three measurements in the transition region. In rat fiber 3 we fit one sigmoid for each side (red and blue) as we have enough measurements.

B.2 Realistic estimates of parameter values and setup of numerical simulations

In this section, we report and justify our best estimates of the parameter values used to simulate the spread of mtDNA deletions in the skeletal muscle fibres of rhesus monkeys. Specifically, we discuss the point estimates used to obtain Figs. 3.1D, E, and the estimated ranges used to obtain Figs. 3.1F. We derive these values from the scientific literature, in order to simulate a realistic travelling wave of mutants and compare the wave speed predicted by our simulations with that estimated from experiments (Fig. 3.1B).

Our model has a total of five parameters, all having an immediate and clear biological interpretation. We do not tune any of these parameters. Instead, we analyse experimental data to estimate parameter values that apply to skeletal muscle fibres of rhesus monkeys and use point estimates for our simulations (Figs. 3.1D, E). In Fig. 3.1D we show that stochastic simulations

(Gillespie algorithm) of a standard linear model based on a net replicative advantage k_{RA} for mutants (Eq. (B.1)) predicts a travelling wave of heteroplasmy with a wave speed that is approximately 300 times faster than the observed speed. Crucially, in Fig. 3.1E we show that stochastic simulations of our model for the same system predict a wave of heteroplasmy with a speed of expansion comparable with observations. Setting up this simulation required some care, due to the slow dynamics caused by the large values of N_{ss} . The wavefront of a travelling wave has the shape of a sigmoid whose steepness depends on its speed (see Section B.5), therefore one needs to initialise the wavefront with the correct steepness such that the wavefront remains approximately stationary. For all the other simulations presented in the thesis, this is not an issue since the heteroplasmy wavefront assumes the steepness corresponding to its constant speed in a short time transient (faster dynamics), that can then be neglected. Instead, for Fig. 3.1E we established an appropriate approximate initial steepness after a series of trials, by verifying that the shape did not change appreciably after running the simulation for a sufficient number of days (around 50). The chosen initial heteroplasmy profile is $h_0(x) = 1/(e^{\tau x - b} + 1)$ for $x > 0$, with $\tau = 7.15 \cdot 10^{-4}/\mu\text{m}$ and $b = 3.915$, over 500 units of length $L = 30\mu\text{m}$. We also introduced approximations at the boundaries of the system. In order to simulate the effect of a macroscopic region of the muscle fibre taken over by mutants we fixed the value of the heteroplasmy of the leftmost unit of the chain at 1. At the right boundary, we truncated the values of heteroplasmy lower than $5 \cdot 10^{-3}$ to zero, to rule out the possibility that edge effects alter the speed of expansion of the heteroplasmy wavefront. We verified that the rightmost units are mutant-free up to the longest simulated time.

In addition, in Fig. 3.1F we provide evidence that our conclusions are robust to the uncertainty in parameter values. We have drawn parameter values from the distributions inferred from experimental data (see discussion below for details). We have then obtained distributions of the wave speeds predicted by a SSD and replicative-advantage model (FK waves), inserting parameter values in the corresponding formulae for wave speed, namely Eq. (A.1) with the appropriate interpretation of k for each case (setting $\beta = 0$ for FK waves). The approach used to obtain the distributions of the predicted wave speeds in is detailed in Ref. [300] and can be easily reproduced using the online tool at

<http://caladis.org/> and the parameter estimates given below. The distributions plotted in Fig. 3.1F confirm that SSD is strongly favoured to account for the expansion.

The parameters μ, γ, c, N_{ss} are common to SSD and to the replicative-advantage model, whereas the net mutant replicative advantage k_{RA} is replaced by δ in our model. We estimate the degradation rate μ from the half-life of mitochondria. Reported values for the half-life in the literature range from a few to 100 days^{135,301–304}. We therefore assume that the half-life is uniformly distributed on the range (2,100) days. As a point estimate, we have chosen 10 days for the half-life, corresponding to a degradation rate of $\mu = 0.07/\text{day}$.

N_{ss} represents the average copy number per nucleus when only wildtypes are present. A study⁴⁴ reports $N_{ss} \simeq 3500$ for healthy human skeletal muscle fibres. We therefore suppose that for rhesus monkeys muscle fibres our uncertainty on N_{ss} has a uniform distribution on the range (2000,5000). We have chosen $N_{ss} = 3000$ since macaques are smaller animals. According to our empirical formula in Eq. (2.39) and Eq. (2.40), changing N_{ss} from 3000 to 4000 would reduce the wave speed by about 10%, Therefore, we do not expect the uncertainty on N_{ss} to affect our results significantly.

An important quantity in the study of muscle ageing, although not directly a parameter of our model, is the internuclear distance L . It has been reported that in skeletal muscle of healthy mice around 30 nuclei are present in 1mm of fibre length, i.e., ~ 1 nucleus per 30 μm ³⁰⁵. We therefore take $L = 30 \mu\text{m}$ as a point estimate, while assuming a uniform distribution on the range (25, 35) μm .

The hopping rate γ is linked to the diffusion coefficient of mtDNA molecules in the muscle fibre, through the relationship $D = \gamma L^2$ ³⁰⁶. Individual organelles in muscle fibres appear to be tightly packed¹⁹⁴. However, a study, Ref. [196], reported that the movement of nucleoids within organelles in these tissues can be effectively described as diffusion with a value of D corresponding to $\gamma = 0.1/\text{day}$. Therefore we choose this value as a point estimate of the individual mtDNA molecules. We quantify the uncertainty about the value of γ from a uniform distribution on the range (0.05,0.15)/day.

The control strength c is the only parameter for which we do not have an optimal measurement. Our estimate is based on the fact that, in replicative cells, during a cell-cycle (≈ 1 day) the number of mtDNA roughly doubles. In this situation, in which a fast mitochondrial biogenesis is needed, we suppose that population is growing exponentially at a rate $\lambda(w=0, m=0)$. For a doubling time of 1 day, this corresponds to

$$c = \frac{\ln 2/\text{day} - \mu}{N_{ss}},$$

that for $N_{ss} = 3000$ and $\mu = 0.07/\text{day}$ gives $c = 2 \cdot 10^{-4}/\text{day}$. Since c is not an independent parameter of our model, it is not present in the phenomenological wave speed formula Eq. (2.39).

The parameters δ and k_{RA} correspond to the two ways of reproducing the increase in mutant carrying capacity (or density), respectively for our model and for a model based on a replicative advantage. From the experimental data¹⁶ reported in Fig. 3.1C, in high-heteroplasmy regions density is approximately increased by a factor $\mathcal{F} = 5$, as reported also in Ref. [187]. We therefore choose 5 as the point estimate of $\mathcal{F}$ and we assume that $\mathcal{F}$ has a uniform distribution on the range (2,8). As for δ , we notice that in our model the carrying capacity m^* for mutants is obtained by setting in Eq. (2.1) $\lambda = \mu$ and $w = 0$ gives, leading to $m^* = N_{ss}/\delta$. The carrying capacity for wildtypes is obtained by setting $\lambda = \mu$ and $m = 0$ in Eq. (2.1), leading to $w^* = N_{ss}$. Therefore, $\delta = 1/\mathcal{F}$ and we chose $\delta = 1/5$ as a point estimate.

In a replicative-advantage model, the mutants have a replication rate given by Eq. (2.1) (with $\delta = 1$) and the addition of a net replicative advantage k_{RA} , namely

$$\lambda_m(w, m) = \mu + c(N_{ss} - w - m) + k_{RA}. \quad (\text{B.1})$$

In order to reproduce an $\mathcal{F}$ -fold increase of mutant carrying capacity, it must be $k_{RA} = cN_{ss}(\mathcal{F} - 1)$. Indeed, setting $\lambda_m = \mu$ and $w = 0$ yields $m^* = k_{RA}/c + N_{ss}$. Hence, we assume k_{RA} to be uniformly distributed on the range $(1, 7)cN_{ss}$ and we choose $k = 4cN_{ss}$ as a point estimate, yielding

$$\lambda_m(w, m) = \mu + c(5N_{ss} - w - m) \quad (\text{B.2})$$

as the mutant replication rate in a replicative-advantage model.

B.3 A combination of mutant higher death and replication rate yields the same Fisher-Kolmogorov wave speed

One could ask if a model that allows for both higher mutant and death rate can account for the observed wave speed.

For $\delta = 1$ (no SSD), the five-fold increase in carrying capacity fixes the *net* replicative advantage, namely differential growth rate of mutants over wildtypes. In turn, this determines the speed of the Fisher-Kolmogorov wave.

Consider mutant species with $\delta = 1$ and both higher death and replication rates μ_m and λ_m :

$$\begin{aligned}\mu_m &= \mu + k_1 \\ \lambda_m(w, m) &= \mu + c(N_{ss} - w - m) + k_2,\end{aligned}\tag{B.3}$$

where $k_1, k_2 \geq 0$ are the increases in mutant death and replication rate over wildtypes'.

Following the reasoning of Section B.2, a five-fold increase in mutant carrying capacity yields $k_2 - k_1 = 4cN_{ss}$.

The speed of the Fisher-Kolmogorov traveling wave of mutants, that applies to this situation, is $v = 2\sqrt{kD}$, where $k = k_2 - k_1$. Therefore allowing for both higher mutant death and replication rates (and therefore introducing a free parameter) does not alter the wave-speed.

B.4 Estimation of fixation probabilities in Fig. 3.3

In Fig. 3.3 we have regressed the reciprocal of the fixation probabilities (of a founder mutation in a chain of units) as a function of the size of the steady-state wildtype population in which the founder mutation arises, which coincides with N_{ss} . We estimated the fixation probabilities via Gillespie simulation, with best estimates for the values of remaining parameters μ, δ, γ, c (Section B.2). The analysis has shown a convincing linear relationship ($p = 2 \cdot 10^{-19}$, $R^2 = 0.9995$).

We have estimated the fixation probabilities for $10 \leq N_{ss} \leq 10^3$. For values $10 \leq N_{ss} \leq 100$, we have simulated a large ensemble of $N_{sim} \approx 2.5 \cdot 10^4$ muscle fibres, modelled as chains of hundreds of units with a founder mutation in the middle of the chain. We have run the simulations until, in every fibre, the mutation had either died or taken over ($h = 1$) a macroscopic region of the fibre (~ 50 -100 units). In the latter case, we have assumed that the mutants would have taken over the entire fibre with more time, excluding extremely unlikely fluctuations that could have led to the extinction of the entire mutant population (of size $\sim 10^3$ here) occupying a large fraction of the fibre. The fixation probability P_f has then been estimated as N_{mut}/N_{sim} , where N_{mut} is the number of fibres mutants took over. In every case, we have ensured that at the boundaries of the chain were free or almost free of mutants, tolerating a heteroplasmy at the boundaries of at most 1/100. This is to avoid finite-size affecting the fixation probability, and makes our estimates of P_f independent of the length of the chains. Muscle fibres are indeed thousands of units long, and a deletion mutation expanding and taking over a macroscopic region is unlikely to reach the end of the fibre.

For $N_{ss} \geq 500$, simulations with fibres long enough that mutants do not reach the boundaries are extremely slow. We have therefore estimated the fixation probability as shown in Fig. B.2. We have run a large number N_{sim} of simulations ($N_{sim} \approx 25000$) with a founder mutation at time 0, and recorded the number of simulations where mutants are present over time. We noticed that this number decreases exponentially. Fitting an exponential decay to the data allows us to estimate the height of the horizontal asymptote of the curve, namely the stationary value of the number of simulations with surviving mutants. Since this asymptotic value coincides

with N_{mut} , we use it to estimate P_f , again as N_{mut}/N_{sim} . As above, we are ignoring extremely rare events, where a large number of mutants ($\sim 10^4$ here), that have taken over a macroscopic fraction of the fibre (hundreds of units here), would go extinct. We stopped the simulations when running them for longer, and adding data for longer times to the exponential fits in Fig. B.2, changed the estimate of N_{mut} – i.e., the horizontal asymptotes – by less than one unit. More detail on the exponential fit are given in Section B.7.

We estimated the standard deviation σ of P_f modelling the phenomenon as a binomial experiment, namely as $\sigma = [P_f(1-P_f)/N_{sim}]^{1/2}$. In Fig. 3.3 we plotted the reciprocal of the fixation probabilities, i.e., $1/P_f = N_{sim}/N_{mut}$. For the uncertainties on the values of $1/P_f$, which are the error bars plotted in Fig. 3.3, we used for each value of N_{ss} the standard deviation of the random variables $Y = N_{ss}/X$, where $X \sim \text{Binomial}(N_{ss}, P_f)$. We estimated this standard deviation simulating 10^7 samples of Y for each value of N_{ss} . Notice that error bars corresponding to these standard deviations are not visible for $N_{ss} = 500, 1000$ as they are smaller than the corresponding dot. In Table B.1, we list the values of N_{mut} , N_{sim} and the corresponding estimates for each value of N_{ss} used.

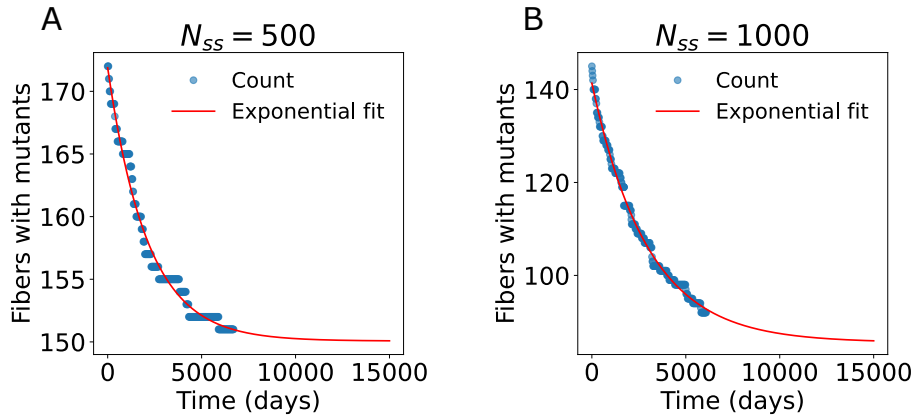


Figure B.2: The above exponential fits allow us to estimate the fixation probability P_f of a founder mutation in a chain of 500 (A) and 1000 (B) units. We have initialised an ensemble of $N_{sim} \approx 25000$ fibres with a founder mutation. Fitting an exponential to the points, we have estimated the stationary value N_{mut} of the number of simulations with surviving mutants, that coincides with the number of simulations in which mutants take over the fibre (neglecting extremely unlikely extinction events). We then estimate $P_f = N_{mut}/N_{sim}$.

N_{ss}	N_{sim}	N_{mut}	P_f	σ
10	4000	842	0.211	0.006
12	9000	1697	0.189	0.004
14	4000	734	0.184	0.006
16	2000	319	0.160	0.008
18	2000	300	0.150	0.008
20	4000	512	0.128	0.005
30	3000	302	0.101	0.005
40	3994	299	0.075	0.004
50	5000	300	0.060	0.003
75	4999	211	0.042	0.003
100	2000	69	0.035	0.004
500	23683	150	0.0063	0.0005
1000	25834	86	0.0033	0.0004

Table B.1: Estimates of P_f and standard deviation σ , with corresponding values of N_{mut} , N_{sim} , for different values of N_{ss} . Obtained via Gillespie simulation, with best estimates for the values of remaining parameters μ, δ, γ, c (Section B.2).

B.5 Relationship between steepness and speed of a travelling wave

This section is based on Secs. 13.2 and 13.3 of Ref. [216]. Let us consider the classical Fisher-Kolmogorov reaction-diffusion equation

$$\frac{\partial h}{\partial t} = kh(1 - h) + D \frac{\partial^2 h}{\partial x^2}, \quad (\text{B.4})$$

where h is heteroplasmy, k is the reaction rate and D is the diffusion coefficient. This equation admits travelling wave solutions moving in the positive x -direction with speed c given by

$$h(x, t) = h(z) = h(x - ct),$$

having defined $z = x - ct$. If the mutants of our system had a replicative advantage k , Eq. (B.4) would describe the wave-like expansion of the high-heteroplasmy front in the continuous-space approximation. If the system is initialised such that

$$h(x, 0) \sim e^{-ax} \quad \text{as} \quad x \rightarrow \infty,$$

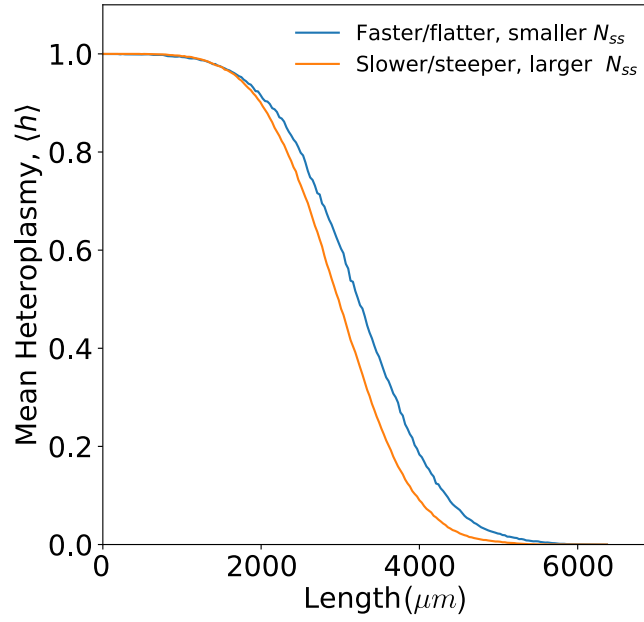


Figure B.3: The two curves are the travelling wavefronts of two systems that differ only in the average wildtype copy number N_{ss} . The smaller the N_{ss} , the faster the wave, according to Eq. (A.1) and Eq. (2.40). The blue curve is the wavefront propagating in the system with the smaller $N_{ss}^{fast} = 2$, while the orange curve is relative to $N_{ss}^{slow} = 6$. The plot shows that faster waves have a flatter wavefront, as it is the case for FK waves. Average over an ensemble of 4000 simulations.

for arbitrary $a > 0$, the heteroplasmy wavefront will evolve into an approximately sigmoidal shape with a steepness τ linked to the wave speed through

$$c = \frac{k}{\tau}, \quad (\text{B.5})$$

meaning that the slower the wave, the steeper the wavefront (larger τ).

We speculate that waves driven by stochastic survival of the densest can be described by an equation analogous to Eq. (B.4), with an effective reaction rate induced by stochastic survival of the densest, that we hypothesise is given by Eq. (2.40). Therefore, also for our model we expect that steeper waves will be slower. In absence of a formal proof, we have verified this intuition computationally. In Fig. B.3 we plot the travelling waves of heteroplasmy for two systems that differ only in copy number N_{ss} . The blue curve corresponds to a system with a smaller N_{ss} and hence, according to Eq. (2.39) presenting a faster wave; the orange curve corresponds to a larger N_{ss} and hence a slower wave. As expected, the slower wave has a steeper wavefront.

In Fig. 3.2 we have reported experimental data on human (panels A and B) and rat (panels C and D) muscle fibres, showing that the steeper wavefront corresponds to the fibre with a larger N_{ss} . We can conclude that the steeper wavefront corresponds to a slower wave. Hence, the data reported in Fig. 3.2 support the prediction of our model that slower waves travel in fibres with a larger N_{ss} .

The data plotted in Fig. B.3 are obtained simulating a chain of 550 units evolving under linear feedback control (Eq. (2.1)), with parameters $\mu = 7 \cdot 10^{-2}/\text{day}$, $c = 0.4/\text{day}$, $\gamma = 0.12/\text{day}$, $\delta = 2/3$. The faster (flatter) blue curve corresponds to $N_{ss} = 6$, while the slower (steeper) orange one to $N_{ss} = 2$.

B.6 Rhesus monkeys data collection

The data plotted in Fig. 3.1B, on the length of the abnormal regions in muscle fibres of rhesus monkeys, were originally published in Ref. [193]. The measurements were performed on skeletal muscle tissues of 11 different animals and the data set consists of the lengths of several abnormal regions for each subject. For each animal, we only use the length of the longest abnormal region, that we assume has originated from a mutation at birth. We consider shorter regions to be the results of mutations that have arisen later in life. Although these assumption are unlikely to be precisely true, our analysis is enough to give an indicative order-of-magnitude estimate, that can be considered as a moderately tight lower bound on the speed of expansion.

B.7 Statistical methods

In Fig. 3.1A, the functional form fitted is a logistic function (sigmoid) $1/(1 + e^{\tau x})$. The variable x is the position along the muscle fibre, centred at the point where $h = 1$. The parameter is $\tau = (2.46 \pm 0.11) \cdot 10^{-2}/\mu\text{m}$. The values are obtained using MLE fit with normal likelihood (bblmle package in R), and the uncertainty is the standard error. In Fig. 3.1B there are 11 points. Via ordinary least squares with normal likelihood (coinciding with MLE) we obtain

$R^2 = 0.758$, $p = 5 \cdot 10^{-4}$, $t = 5.3$ for a line of equation $y = a + bx$, with $a = (-482 \pm 229) \mu\text{m}$ and $b = 0.131 \pm 0.025 \mu\text{m/day}$ (uncertainty is standard error). The wave speed relative to Figs. 3.1D, E have been calculated as explained in Section A.1. In Fig. 3.1F $9 \cdot 10^6$ values have been drawn for each of the three histograms.

For the fits in Figs. 3.2A, C MLE with normal likelihood is used (bbmlmle package in R). The values obtained are reported in the caption of Fig. 3.2, the uncertainty is standard error and we report profile-likelihood confidence interval for better coverage. Notice that one of the fiber in Fig. 3.2A is the same as that in Fig. 3.1A. The statistical test performed on the data in Figs. 3.2B, D is a one-tailed Welch's test (unequal-variance t -test). For data in Fig. 3.2B, the two samples are of size 15 and 7, with measurements taken from different samples, meaning distinct slices of the same muscle fibres; means for copy numbers are $\bar{x}_1 = 67$, $\bar{x}_2 = 16$, standard deviations $\sigma_1 = 39$, $\sigma_2 = 6.5$. Welch's test gives $p = 10^{-4}$, $t = 4.69$, $d = 1.48$, where the effect size d is the unequal-variance Cohen's d . For data in Fig. 3.2B, the two samples are of size 6 and 5, with measurements taken from different samples, meaning distinct slices of the same muscle fibres; means for copy numbers are $\bar{x}_1 = 6875$, $\bar{x}_2 = 3505$, standard deviations $\sigma_1 = 3840$, $\sigma_2 = 1700$. Welch's test gives $p = 0.06$, $t = 1.76$, $d = 1.00$ (Cohen's d). The elements of the box plots are minimum value (lower whisker), first quartile, median, third quartile and maximum value (upper whisker).

In Fig. 3.3, we have 13 points, with error bars being standard deviations, obtained as explained in Section B.4. By linear least squares, we fit a line of equation $a + bN_{ss}$, with $b = 0.301 \pm 0.002$, $a = 1.5 \pm 0.6$, with $p = 10^{-20}$ (for b) and effect size $R^2 = 0.9995$ (square of the Pearson's correlation coefficient r). The uncertainty reported is standard error.

In Fig. B.2, non-linear least squares was used, fitting the function $y = ae^{-bt} + c$. For $N_{ss} = 500$, the estimated values are $a = 21.8 \pm 0.1$, $b = (4.7 \pm 0.1) \cdot 10^{-4}/\text{day}$, $c = 150.1 \pm 0.1$. For $N_{ss} = 1000$, the estimated values are $a = 55.8 \pm 0.3$, $b = (3.37 \pm 0.06) \cdot 10^{-4}/\text{day}$, $c = 85.6 \pm 0.4$. Uncertainty reported is standard deviation. See Section B.4 for more details on how the synthetic data was obtained.

In Fig. B.3, the average is over 4000 simulations; more details in Section B.5.

B.8 Code availability

The SSD simulations reported in Chapter 3 (Figs. 3.3 and 3.1E) have been obtained with the code freely available at <https://github.com/ferdinando17/Survival-of-the-Densest>, same C code as Chapter 2. However, these plots required hundreds of processors running in parallel for 10 months.

Appendix C

Appendix to Chapter 4

Here we report tables that are supplementary to the results presented in Chapter 4. We give the full output of the models, including the hazard ratios for all predictors (not only estimates of protection as in Chapter 4). Moreover, we report the results of the sensitivity analyses mentioned in Chapter 4.

List of abbreviations:

- HR: hazard ratios
- CI: confidence interval
- PF: Pfizer vaccine
- d1: dose 1
- d2: dose 2
- long/short: long/short interval in Pfizer dosing schedule.
- d1 21-56: between 21 and 56 days after dose (example).
- AZ: AstraZeneca vaccine
- PI: previous infection

- PI < 1yr : previous infection less than one year ago (example).

C.1 Supplementary Tables for Section 4.4

C.1.1 Full model outputs

Full model output, negative cohort

	HR	95% CI
Vaccine status		
Unvaccinated	Ref	-
d1 0-20 PF	0.72	(0.63-0.83)
d1 21-27 PF	0.41	(0.29-0.58)
d1 28-41 PF	0.34	(0.24-0.48)
d1 42-55 PF	0.30	(0.19-0.46)
d1 56+ PF	0.37	(0.25-0.54)
d1 0-20 AZ	0.45	(0.29-0.70)
d1 21-27 AZ	0.37	(0.08-1.80)
d1 28-41 AZ	0.15	(0.03-0.84)
d1 42-55 AZ	0.68	(0.25-1.87)
d1 56+ AZ	0.91	(0.45-1.87)
d2 0-13 PF_long	0.21	(0.10-0.44)
d2 14-73 PF_long	0.15	(0.08-0.28)
d2 74-133 PF_long	0.34	(0.25-0.47)
d2 134-193 PF_long	0.32	(0.23-0.46)
d2 194+ PF_long	0.49	(0.31-0.78)
d2 0-13 PF_short	0.25	(0.14-0.43)

Table C.1 continued from previous page

	HR	95% CI
d2 14-73 PF_short	0.11	(0.06-0.22)
d2 74-133 PF_short	0.42	(0.21-0.82)
d2 134-193 PF_short	0.50	(0.33-0.74)
d2 194+ PF_short	0.47	(0.31-0.72)
d2 14-73 AZ	0.42	(0.23-0.77)
d2 74-133 AZ	0.50	(0.35-0.71)
d2 134+ AZ	0.28	(0.13-0.61)
Gender		
Male	Ref	-
Female	0.91	(0.79-1.03)
Non-binary	0.49	(0.07-3.61)
Prefer not to say	0.30	(0.03-2.67)
Ethnic group		
White	Ref	-
Mixed	0.89	(0.68-1.16)
Asian	1.26	(1.08-1.47)
Black	1.79	(1.37-2.33)
Other	0.84	(0.59-1.21)
Prefer not to say	1.38	(0.64-3.01)

Full model output, negative and positive

	HR	95% CI
Vaccine and PI status		
Unvaccinated	Ref	-

Table C.2 continued from previous page

	HR	95% CI
d1 0-20 naive	0.95	(0.83-1.09)
d1 21+ naive	0.47	(0.38-0.59)
d2 0-13 naive	0.29	(0.17-0.48)
d2 14-73 naive	0.13	(0.07-0.23)
d2 73-133 naive	0.35	(0.26-0.46)
d2 133-193 naive	0.37	(0.29-0.49)
d2 193+ naive	0.53	(0.38-0.75)
Unvaccinated PI <1yr	0.14	(0.11-0.19)
d1 0-20 PI <1yr	0.14	(0.10-0.21)
d1 21+ PI <1yr	0.08	(0.05-0.14)
d2 0-13 PI <1yr	0.12	(0.04-0.39)
d2 14-73 PI <1yr	0.16	(0.08-0.33)
d2 73-133 PI <1yr	0.08	(0.04-0.17)
d2 133-193 PI <1yr	0.08	(0.05-0.15)
d2 193+ PI <1yr	0.14	(0.03-0.73)
Unvaccinated PI >1yr	0.31	(0.16-0.62)
d1 21+ PI >1yr	0.06	(0.01-0.38)
d2 0-13 PI >1yr	0.24	(0.03-1.83)
d2 14-73 PI >1yr	0.06	(0.01-0.25)
d2 73-133 PI >1yr	0.03	(0.02-0.07)
d2 133-193 PI >1yr	0.07	(0.04-0.11)
d2 193+ PI >1yr	0.05	(0.01-0.18)

Table C.2 continued from previous page

	HR	95% CI
Gender		
Male	Ref	-
Female	0.90	(0.78-1.03)
Non-binary	0.41	(0.06-2.98)
Prefer not to say	0.26	(0.03-2.22)
Ethnic group		
White	Ref	-
Mixed	0.93	(0.70-1.23)
Asian	1.24	(1.04-1.48)
Black	1.68	(1.30-2.16)
Other	0.87	(0.60-1.27)
Prefer not to say	1.30	(0.56-2.98)

C.1.2 Sensitivity analyses for depletion-of-susceptible bias

We ran a sensitivity analysis to assess if our estimates suffer from depletion-of-susceptible bias, that can cause the observation of a spurious waning of protection. We stopped follow up time at July 31, 2021. Results are very similar to those presented in Section 4.4, in the sense that show comparable waning. We therefore conclude that there is no evidence that our results are affected by the depletion-of-susceptible bias.

The table below is relative to the naïve cohort only.

	HR	95% CI
Vaccine status		
d1 0-20 PF	0.72	(0.62-0.82)

Table C.3 continued from previous page

	HR	95% CI
d1 21-27 PF	0.40	(0.29-0.57)
d1 28-41 PF	0.32	(0.22-0.45)
d1 42-55 PF	0.27	(0.17-0.42)
d1 56+ PF	0.30	(0.18-0.49)
d1 0-20 AZ	0.44	(0.29-0.68)
d1 21-27 AZ	0.36	(0.07-1.83)
d1 28-41 AZ	0.14	(0.03-0.79)
d1 42-55 AZ	0.62	(0.22-1.70)
d1 56+ AZ	1.05	(0.46-2.41)
d2 0-13 PF long	0.17	(0.08-0.37)
d2 14-73 PF long	0.11	(0.05-0.24)
d2 74-133 PF long	0.38	(0.25-0.58)
d2 134-193 PF long	0.38	(0.24-0.60)
d2 0-13 PF_short	0.24	(0.14-0.43)
d2 14-73 PF_short	0.10	(0.05-0.21)
d2 74-133 PF_short	0.35	(0.16-0.76)
d2 134-193 PF_short	0.60	(0.36-1.00)
d2 194+ PF_short	0.48	(0.23-1.01)
d2 14-73 AZ	0.42	(0.23-0.76)
d2 74-133 AZ	0.52	(0.28-0.95)
d2 134+ AZ	0.86	(0.05-14.41)
Gender		
Female	0.90	(0.78-1.04)
Non-binary	0.58	(0.08-4.34)
Prefer not to say	0.57	(0.07-4.87)

Table C.3 continued from previous page

	HR	95% CI
Ethnic group		
Mixed	0.75	(0.54-1.05)
Asian	1.33	(1.14-1.57)
Black	1.88	(1.44-2.47)
Other	0.88	(0.58-1.33)
Prefer not to say	0.97	(0.42-2.27)

The table below is relative to the naïve and positive cohort, and AstraZeneca recipients are excluded from the model. The estimates of some coefficients did not converge, flagged by NC.

	HR	95% CI
Vaccine and PI status		
d1 0-20 naïve	0.94	(0.82-1.08)
d1 21+ naïve	0.44	(0.35-0.56)
d2 0-13 naïve	0.27	(0.16-0.46)
d2 14-73 naïve	0.12	(0.06-0.21)
d2 73-133 naïve	0.39	(0.28-0.54)
d2 133-193 naïve	0.45	(0.31-0.64)
d2 193+ naïve	0.58	(0.29-1.16)
Unvaccinated PI<1yr	0.15	(0.11-0.19)
d1 0-20 PI<1yr	0.14	(0.10-0.21)
d1 21+ PI<1yr	0.09	(0.05-0.15)
d2 0-13 PI<1yr	0.13	(0.04-0.41)
d2 14-73 PI<1yr	0.19	(0.09-0.37)

Table C.4 continued from previous page

	HR	95% CI
d2 73-133 PI<1yr	0.10	(0.04-0.23)
d2 133-193 PI<1yr	0.14	(0.03-0.57)
d2 193+ PI<1yr	NC	-
Unvaccinated PI>1yr	0.43	(0.15-1.19)
d1 21+ PI>1yr	NC	-
d2 0-13 PI>1yr	0.31	(0.04-2.44)
d2 14-73 PI>1yr	0.07	(0.02-0.31)
d2 73-133 PI>1yr	0.05	(0.02-0.09)
d2 133-193 PI>1yr	0.11	(0.04-0.28)
d2 193+ PI>1yr	NC	-
Gender		
Female	0.90	(0.77-1.05)
Non-binary	0.49	(0.07-3.70)
Prefer not to say	0.45	(0.05-3.76)
Ethnic group		
Mixed	0.78	(0.55-1.12)
Asian	1.34	(1.10-1.62)
Black	1.76	(1.34-2.32)
Other	0.89	(0.58-1.37)
Prefer not to say	0.86	(0.31-2.42)

C.1.3 Sensitivity analyses for delayed-entry bias

We ran sensitivity analyses to assess if having delayed entry in our cohort-study is biasing our estimates. For this sensitivity analysis we only consider participants who were enrolled on or before the beginning of analysis time (December 7, 2020). We obtain comparable (but more uncertain) estimates, concluding that delayed entry is not affecting our results.

The table below is relative to the naïve cohort only.

	HR	95% CI
Vaccine status		
d1 0-20 PF	0.74	(0.64-0.85)
d1 21-27 PF	0.42	(0.30-0.60)
d1 28-41 PF	0.33	(0.22-0.47)
d1 42-55 PF	0.27	(0.16-0.46)
d1 56+ PF	0.29	(0.16-0.52)
d1 0-20 AZ	0.44	(0.28-0.69)
d1 21-27 AZ	0.40	(0.07-2.16)
d1 28-41 AZ	0.16	(0.03-0.90)
d1 42-55 AZ	0.60	(0.21-1.75)
d1 56+ AZ	0.82	(0.40-1.68)
d2 0-13 PF_long	0.22	(0.08-0.61)
d2 14-73 PF_long	0.15	(0.05-0.42)
d2 74-133 PF_long	0.26	(0.18-0.39)
d2 134-193 PF_long	0.32	(0.22-0.46)
d2 194+ PF_long	0.61	(0.33-1.11)
d2 0-13 PF_short	0.25	(0.14-0.45)
d2 14-73 PF_short	0.11	(0.06-0.23)
d2 74-133 PF_short	0.18	(0.04-0.81)

Table C.5 continued from previous page

	HR	95% CI
d2 134-193 PF_short	0.46	(0.25-0.86)
d2 194+ PF_short	0.37	(0.22-0.61)
d2 14-73 AZ	0.32	(0.12-0.84)
d2 74-133 AZ	0.45	(0.27-0.75)
d2 134+ AZ	0.23	(0.09-0.56)
Gender		
Female	0.94	(0.81-1.10)
Non-binary	0.62	(0.09-4.30)
Prefer not to say	0.52	(0.05-5.86)
Ethnic group		
Mixed	0.87	(0.65-1.18)
Asian	1.31	(1.11-1.55)
Black	1.88	(1.42-2.49)
Other	0.89	(0.62-1.27)
Prefer not to say	1.43	(0.61-3.34)

The table below is relative to the naïve and positive cohort, and AstraZeneca recipients are excluded from the model. The estimates of some coefficients did not converge, flagged by NC.

	HR	95% CI
Vaccine status		
d1 0-20 PF	0.74	(0.64-0.85)
d1 21-27 PF	0.42	(0.30-0.60)
d1 28-41 PF	0.33	(0.22-0.47)
d1 42-55 PF	0.27	(0.16-0.46)

Table C.6 continued from previous page

	HR	95% CI
d1 56+ PF	0.29	(0.16-0.52)
d1 0-20 AZ	0.44	(0.28-0.69)
d1 21-27 AZ	0.40	(0.07-2.16)
d1 28-41 AZ	0.16	(0.03-0.90)
d1 42-55 AZ	0.60	(0.21-1.75)
d1 56+ AZ	0.82	(0.40-1.68)
d2 0-13 PF_long	0.22	(0.08-0.61)
d2 14-73 PF_long	0.15	(0.05-0.42)
d2 74-133 PF_long	0.26	(0.18-0.39)
d2 134-193 PF_long	0.32	(0.22-0.46)
d2 194+ PF_long	0.61	(0.33-1.11)
d2 0-13 PF_short	0.25	(0.14-0.45)
d2 14-73 PF_short	0.11	(0.06-0.23)
d2 74-133 PF_short	0.18	(0.04-0.81)
d2 134-193 PF_short	0.46	(0.25-0.86)
d2 194+ PF_short	0.37	(0.22-0.61)
d2 14-73 AZ	0.32	(0.12-0.84)
d2 74-133 AZ	0.45	(0.27-0.75)
d2 134+ AZ	0.23	(0.09-0.56)
Gender		
Female	0.94	(0.81-1.10)
Non-binary	0.62	(0.09-4.30)
Prefer not to say	0.52	(0.05-5.86)
Ethnic group		

Table C.6 continued from previous page

	HR	95% CI
Mixed	0.87	(0.65-1.18)
Asian	1.31	(1.11-1.55)
Black	1.88	(1.42-2.49)
Other	0.89	(0.62-1.27)
Prefer not to say	1.43	(0.61-3.34)

C.1.4 Sensitivity analyses to assess the effect of discarding participants without a reliable date of primary infection

We ran a sensitivity analysis to assess the impact of having excluded positive cohort participants without a reliable date of primary infection, whose positivity status has been ascertained via antibody testing. In the analysis we included these participants, using the date of their first positive antibody test as a proxy for the date of primary infection. Results are comparable with those presented in Section 4.4, but generally more uncertain.

The table below is relative to the naïve cohort only.

	HR	95% CI
Vaccine status		
d1 0-20 PF	0.72	(0.63-0.83)
d1 21-27 PF	0.42	(0.30-0.58)
d1 28-41 PF	0.35	(0.25-0.49)
d1 42-55 PF	0.30	(0.19-0.46)
d1 56+ PF	0.39	(0.27-0.56)
d1 0-20 AZ	0.48	(0.30-0.77)
d1 21-27 AZ	0.38	(0.08-1.85)
d1 28-41 AZ	0.15	(0.03-0.85)

Table C.7 continued from previous page

	HR	95% CI
d1 42-55 AZ	0.71	(0.26-1.94)
d1 56+ AZ	0.88	(0.43-1.80)
d2 0-13 PF_long	0.21	(0.10-0.45)
d2 14-73 PF_long	0.15	(0.08-0.29)
d2 74-133 PF_long	0.35	(0.26-0.48)
d2 134-193 PF_long	0.33	(0.24-0.47)
d2 194+ PF_long	0.51	(0.31-0.82)
d2 0-13 PF_short	0.24	(0.14-0.43)
d2 14-73 PF_short	0.11	(0.06-0.22)
d2 74-133 PF_short	0.43	(0.22-0.85)
d2 134-193 PF_short	0.51	(0.34-0.74)
d2 194+ PF_short	0.49	(0.32-0.75)
d2 14-73 AZ	0.42	(0.23-0.78)
d2 74-133 AZ	0.50	(0.35-0.72)
d2 134+ AZ	0.29	(0.13-0.63)
Gender		
Female	0.91	(0.80-1.04)
Non-binary	0.44	(0.06-3.26)
Prefer not to say	0.30	(0.03-2.61)
Ethnic group		
Mixed	0.88	(0.67-1.15)
Asian	1.26	(1.08-1.47)
Black	1.78	(1.37-2.31)
Other	0.87	(0.62-1.24)

Table C.7 continued from previous page

	HR	95% CI
Prefer not to say	1.39	(0.63-3.06)

The table below is relative to the naïve and positive cohort, and AstraZeneca recipients are excluded from the model.

	HR	95% CI
Vaccine and PI status		
d1 0-20 naïve	0.95	(0.83-1.09)
d1 21+ naïve	0.47	(0.38-0.58)
d2 0-13 naïve	0.28	(0.17-0.48)
d2 14-73 naïve	0.12	(0.07-0.22)
d2 73-133 naïve	0.35	(0.27-0.45)
d2 133-193 naïve	0.37	(0.28-0.48)
d2 193+ naïve	0.52	(0.37-0.74)
Unvaccinated PI<1yr	0.16	(0.13-0.20)
d1 0-20 PI<1yr	0.15	(0.10-0.22)
d1 21+ PI<1yr	0.11	(0.07-0.17)
d2 0-13 PI<1yr	0.17	(0.06-0.50)
d2 14-73 PI<1yr	0.21	(0.12-0.37)
d2 73-133 PI<1yr	0.09	(0.05-0.16)
d2 133-193 PI<1yr	0.07	(0.04-0.12)
d2 193+ PI<1yr	0.10	(0.02-0.57)
Unvaccinated PI>1yr	0.25	(0.13-0.51)
d1 21+ PI>1yr	0.05	(0.01-0.29)

Table C.8 continued from previous page

	HR	95% CI
d2 0-13 PI>1yr	0.20	(0.03-1.51)
d2 14-73 PI>1yr	0.05	(0.01-0.21)
d2 73-133 PI>1yr	0.05	(0.03-0.09)
d2 133-193 PI>1yr	0.06	(0.04-0.10)
d2 193+ PI>1yr	0.04	(0.01-0.14)
Gender		
Female	0.90	(0.78-1.04)
Non-binary	0.38	(0.05-2.73)
Prefer not to say	0.25	(0.03-2.15)
Ethnic group		
Mixed	0.92	(0.69-1.22)
Asian	1.21	(1.01-1.44)
Black	1.71	(1.31-2.23)
Other	0.84	(0.57-1.23)
Prefer not to say	1.24	(0.54-2.87)

C.2 Supplementary Tables, Section 4.5

Further abbreviations used in the tables of this Section:

- d3: three vaccine doses.
- 3PF: three doses of Pfizer vaccine
- 2AZ/1PF : primary course AstraZeneca, 1 Pfizer booster.

C.2.1 Full model outputs

We report results for the same models as those presented in the Section 4.5, but including the estimates of hazard ratios (HR) for additional predictors. We report only HR (not vaccine effectiveness) for all predictors.

C.2.1.1 Analysis 1, booster effectiveness in negative cohort

Delta period, the estimates of some coefficients did not converge (NC):

	HR	95% CI
Vaccination status		
d3 0-6 3PF	0.65	(0.33-1.27)
d3 7+ 3PF	0.37	(0.23-0.60)
d2 194+ AZ	1.58	(0.76-3.28)
d3 0-6 2AZ/1PF	1.91	(0.69-5.30)
d3 7+ 2AZ/1PF	0.44	(0.18-1.08)
Gender		
Female	0.62	(0.47-0.83)
Other	NC	-

Table C.9 continued from previous page

	HR	95% CI
Ethnic group		
Asian	0.84	(0.50-1.41)
Black	0.56	(0.14-2.25)
Mixed	0.75	(0.26-2.19)
Other	0.44	(0.11-1.72)
Prefer not to say	NC	-
Staff type		
Doctor	1.37	(0.84-2.23)
Nursing	1.29	(0.81-2.03)
Healthcare Assistant	1.79	(1.03-3.11)
Midwife	1.80	(0.76-4.27)
Healthcare Scientist	1.69	(0.94-3.03)
Pharmacist	1.00	(0.48-2.06)
Physiotherapist/Occupational Therapist/SALT	1.09	(0.58-2.05)
Student (Medical/Nursing/Midwifery/Other)	1.30	(0.48-3.48)
Estates/Porters/Security	1.62	(0.72-3.61)
Other - no patient contact	1.43	(0.90-2.28)
Other - patient contact	1.15	(0.44-3.04)

Omicron period:

	HR	95% CI
Vaccination status		
d3 0-6 3PF	1.13	(0.65-1.97)
d3 7-66 3PF	0.65	(0.53-0.79)
d3 67-126 3PF	0.66	(0.57-0.78)
d3 127+ 3PF	0.79	(0.62-1.00)
d2 194+ AZ	1.22	(0.81-1.84)
d3 7-66 2AZ/1PF	0.39	(0.27-0.56)
d3 67+ 2AZ/1PF	0.54	(0.40-0.72)
Gender		
Female	1.13	(1.00-1.28)
Other	0.98	(0.36-2.69)
Ethnic group		
Asian	1.11	(0.92-1.33)
Black	1.17	(0.83-1.65)
Mixed	0.97	(0.68-1.39)
Other	1.24	(0.86-1.79)
Prefer not to say	1.53	(0.68-3.43)
Staff type		
Doctor	0.70	(0.59-0.83)
Nursing	0.80	(0.72-0.90)
Healthcare Assistant	1.12	(0.94-1.34)
Midwife	0.89	(0.66-1.21)
Healthcare Scientist	0.72	(0.58-0.89)
Pharmacist	0.64	(0.47-0.87)

Table C.10 continued from previous page

	HR	95% CI
Physiotherapist/Occupational Therapist/SALT	0.87	(0.69-1.11)
Student (Medical/Nursing/Midwifery/Other)	0.92	(0.70-1.19)
Estates/Porters/Security	1.28	(0.92-1.78)
Other - no patient contact	0.77	(0.65-0.91)
Other - patient contact	0.75	(0.54-1.05)

C.2.1.2 Analysis 2, protection associated with primary infection

Delta period, the estimates of some coefficients did not converge (NC):

	HR	95% CI
Primary infection status		
3-6 mth ago	0.10	(0.02-0.58)
6-9 mth ago	0.25	(0.15-0.41)
9-12 mth ago	0.37	(0.26-0.52)
12-15 mth ago	0.07	(0.02-0.28)
15+ mth ago	0.15	(0.11-0.19)
Vaccination status		
d2 194-253 PF	1.33	(1.08-1.65)
d2 254+ PF	1.48	(0.99-2.21)
d3 0-6 3PF	1.04	(0.69-1.56)
d3 7+ 3PF	0.57	(0.43-0.77)
d2 194+ AZ	2.42	(1.43-4.10)
d3 0-6 2AZ/1PF	2.81	(1.19-6.67)
d3 7+ 2AZ/1PF	0.73	(0.32-1.67)

Table C.11 continued from previous page

	HR	95% CI
Gender		
Female	0.71	(0.61-0.84)
Other	NC	-
Ethnic group		
Asian	0.95	(0.72-1.25)
Black	1.03	(0.60-1.76)
Mixed	0.94	(0.58-1.53)
Other	0.67	(0.37-1.23)
Prefer not to say	0.60	(0.08-4.47)
Staff type		
Doctor	1.02	(0.81-1.29)
Nursing	1.06	(0.86-1.31)
Healthcare Assistant	1.36	(1.06-1.75)
Midwife	1.20	(0.71-2.02)
Healthcare Scientist	1.18	(0.91-1.53)
Pharmacist	0.93	(0.64-1.33)
Physiotherapist/Occupational Therapist/SALT	1.14	(0.84-1.56)
Student (Medical/Nursing/Midwifery/Other)	1.14	(0.75-1.72)
Estates/Porters/Security	0.98	(0.63-1.51)
Other - no patient contact	1.08	(0.83-1.40)
Other - patient contact	1.26	(0.75-2.10)

Omicron period:

	HR	95% CI
Vaccination status		
d3 0-6 3PF	0.93	(0.55-1.59)
d3 7-66 3PF	0.63	(0.53-0.74)
d3 67-126 3PF	0.67	(0.59-0.77)
d3 127+ 3PF	0.81	(0.66-0.99)
d2 194+ AZ	1.08	(0.81-1.44)
d3 0-6 2AZ/1PF	0.42	(0.11-1.64)
d3 7-66 2AZ/1PF	0.45	(0.35-0.59)
d3 67+ 2AZ/1PF	0.51	(0.40-0.64)
Primary infection status		
3-6 mth ago	0.33	(0.25-0.44)
6-9 mth ago	0.47	(0.27-0.80)
9-12 mth ago	0.73	(0.56-0.96)
12-15 mth ago	0.73	(0.63-0.84)
15+ mth ago	0.74	(0.66-0.83)
Gender		
Female	1.12	(1.00-1.27)
Other	0.84	(0.32-2.21)
Ethnic group		
Asian	1.12	(0.95-1.31)
Black	1.18	(0.88-1.57)
Mixed	1.08	(0.81-1.44)
Other	1.20	(0.90-1.61)
Prefer not to say	1.37	(0.61-3.08)

Table C.12 continued from previous page

	HR	95% CI
Staff type		
Doctor	0.75	(0.64-0.87)
Nursing	0.85	(0.76-0.94)
Healthcare Assistant	1.04	(0.89-1.21)
Midwife	0.84	(0.64-1.10)
Healthcare Scientist	0.77	(0.62-0.95)
Pharmacist	0.65	(0.51-0.85)
Physiotherapist/Occupational Therapist/SALT	0.84	(0.68-1.04)
Student (Medical/Nursing/Midwifery/Other)	0.92	(0.74-1.14)
Estates/Porters/Security	1.14	(0.86-1.51)
Other - no patient contact	0.76	(0.65-0.89)
Other - patient contact	0.83	(0.61-1.11)

C.2.1.3 Analysis 3, hybrid protection

Delta period, the estimate of numerous coefficients did not converge (NC):

	HR	95% CI
Vaccination status		
d3 0-6 3PF PI ≤1yr ago	0.19	(0.02-2.11)
d3 7+ 3PF PI ≤1yr ago	0.12	(0.04-0.34)
d3 0-6 3PF PI >1yr ago	0.22	(0.04-1.06)
d3 7+ 3PF PI >1yr ago	0.10	(0.04-0.24)
Gender		
Female	1.38	(0.55-3.43)
Other	NC	-

Table C.13 continued from previous page

	HR	95% CI
Ethnic group		
Asian	0.24	(0.02-2.36)
Black	NC	-
Mixed	NC	-
Other	NC	-
Prefer not to say	NC	-
Staff type		
Doctor	0.76	(0.17-3.42)
Nursing	0.83	(0.21-3.27)
Healthcare Assistant	0.58	(0.06-5.46)
Midwife	NC	-
Healthcare Scientist	NC	-
Pharmacist	0.88	(0.10-8.15)
Physiotherapist/Occupational Therapist/SALT	0.25	(0.02-2.66)
Student (Medical/Nursing/Midwifery/Other)	0.96	(0.08-10.87)
Estates/Porters/Security	2.46	(0.48-12.68)
Other - no patient contact	0.50	(0.12-2.13)
Other - patient contact	NC	-

Omicron period, the estimate of one coefficient did not converge (NC):

	HR	95% CI
Vaccination status		
d2 194-253 PF PI ≤ 1 yr	0.79	(0.38-1.67)
d2 254+ PF PI ≤ 1 yr	0.45	(0.20-0.99)

Table C.14 continued from previous page

	HR	95% CI
d3 0-6 3PF PI ≤1yr	0.59	(0.16-2.13)
d3 7-66 3PF PI ≤1yr	0.34	(0.22-0.51)
d3 67-126 3PF PI ≤1yr	0.32	(0.24-0.43)
d3 127+ 3PF PI ≤1yr	0.23	(0.08-0.67)
d2 194-253 PF PI >1yr	0.83	(0.46-1.48)
d2 254+ PF PI >1yr	0.68	(0.50-0.91)
d3 0-6 3PF PI >1yr	0.44	(0.10-1.87)
d3 7-66 3PF PI >1yr	0.43	(0.34-0.55)
d3 67-126 3PF PI >1yr	0.52	(0.43-0.62)
d3 127+ 3PF PI >1yr	0.59	(0.42-0.83)
Gender		
Female	1.09	(0.85-1.39)
Other	NC	-
Ethnic group		
Asian	1.16	(0.93-1.43)
Black	1.33	(0.83-2.15)
Mixed	1.16	(0.73-1.83)
Other	1.07	(0.72-1.58)
Prefer not to say	0.66	(0.07-6.40)
Staff type		
Doctor	0.71	(0.56-0.89)
Nursing	0.86	(0.70-1.06)
Healthcare Assistant	0.89	(0.65-1.21)
Midwife	0.92	(0.57-1.47)

Table C.14 continued from previous page

	HR	95% CI
Healthcare Scientist	0.88	(0.59-1.30)
Pharmacist	0.75	(0.46-1.22)
Physiotherapist/Occupational Therapist/SALT	0.74	(0.52-1.04)
Student (Medical/Nursing/Midwifery/Other)	0.89	(0.56-1.39)
Estates/Porters/Security	0.98	(0.61-1.59)
Other - no patient contact	0.69	(0.56-0.84)
Other - patient contact	1.08	(0.57-2.04)

C.2.2 Sensitivity analyses for depletion-of-susceptible bias

C.2.2.1 Delta period

We run a sensitivity analysis to investigate if the depletion of susceptible bias affected our estimates of protection against the Delta variant.

The only model to which this analysis applies is the model estimating protection from primary infection. Models investigating booster's vaccine effectiveness do not have time-dependent, as only one time interval was considered for lack of follow-up time, the interval being approximately 0-2 months after booster. As no waning can be identified with a single time-interval, the depletion of susceptible bias cannot be ascertained.

We stopped the study at November 1st 2021, instead of December 1st 2021. For the estimates of protection from primary infection, results are more uncertain, but compatible with those obtained considering the full dataset. We conclude that there is no evidence of the depletion of susceptible bias in this dataset. Some estimated did not converge, flagged by NC.

	HR	95% CI
Primary infection status		
3-6 mth ago	0.18	(0.03-1.01)
6-9 mth ago	0.25	(0.15-0.42)
9-12 mth ago	0.42	(0.29-0.61)
12-15 mth ago	NC	-
15+ mth ago	0.13	(0.09-0.18)
Vaccine status		
d2 134-193 PF	Ref.	-
d2 194-253 PF	1.36	(1.10-1.69)
d2 254+ PF	1.24	(0.73-2.11)
d3 0-6 3PF	1.04	(0.70-1.54)
d3 7+ 3PF	0.67	(0.46-0.98)
d2 194+ AZ	1.98	(0.76-5.11)
d3 0-6 2AZ/1PF	2.53	(0.95-6.75)
d3 7+ 2AZ/1PF	1.08	(0.27-4.42)
Gender		
Female	0.75	(0.63-0.90)
Other	NC	-
Ethnic group		
Asian	1.03	(0.76-1.39)
Black	0.98	(0.52-1.85)
Mixed	1.00	(0.60-1.69)
Other	0.77	(0.40-1.46)
Prefer not to say	0.75	(0.10-5.54)

Table C.15 continued from previous page

	HR	95% CI
Staff type		
Doctor	0.93	(0.72-1.20)
Nursing	1.00	(0.79-1.26)
Healthcare Assistant	1.24	(0.94-1.63)
Midwife	1.22	(0.71-2.09)
Healthcare Scientist	1.07	(0.81-1.43)
Pharmacist	0.87	(0.58-1.31)
Physiotherapist/Occupational Therapist/SALT	1.11	(0.79-1.56)
Student (Medical/Nursing/Midwifery/Other)	1.13	(0.70-1.82)
Estates/Porters/Security	0.86	(0.51-1.44)
Other- patient contact	1.00	(0.75-1.35)
Other - no patient contact	1.31	(0.76-2.28)

C.2.2.2 Omicron period

In order to investigate the depletion-of-susceptible bias in the Omicron period, we stopped the study at January 31, 2022, instead of February 28, 2022. Here the sensitivity analysis applies to all models: naive protection, hybrid protection and protection from primary infection. Results are more uncertain, but compatible with those obtained considering the full dataset. We conclude that there is no evidence of the depletion of susceptible bias.

Effectiveness of booster in negative cohort:

	HR	95% CI
Vaccine status		
d3 0-6 3PF	1.19	(0.67-2.11)

Table C.16 continued from previous page

	HR	95% CI
d3 7-66 3PF	0.67	(0.55-0.82)
d3 67-126 3PF	0.67	(0.57-0.79)
d3 127+ 3PF	1.13	(0.64-2.00)
d2 194+ AZ	1.36	(0.88-2.09)
d3 7-66 2AZ/1PF	0.38	(0.26-0.56)
d3 67+ 2AZ/1PF	0.57	(0.41-0.80)
Gender		
Female	1.14	(0.98-1.32)
Other	0.35	(0.06-2.22)
Ethnic group		
Asian	1.04	(0.83-1.31)
Black	1.19	(0.83-1.71)
Mixed	0.73	(0.45-1.21)
Other	1.28	(0.81-2.03)
Prefer not to say	1.39	(0.56-3.46)
Staff type		
Doctor	0.68	(0.55-0.84)
Nursing	0.77	(0.68-0.88)
Healthcare Assistant	1.08	(0.89-1.32)
Midwife	1.00	(0.71-1.41)
Healthcare Scientist	0.76	(0.60-0.96)
Pharmacist	0.63	(0.45-0.87)
Physiotherapist/Occupational Therapist/SALT	0.79	(0.60-1.04)
Student (Medical/Nursing/Midwifery/Other)	0.78	(0.55-1.10)

Table C.16 continued from previous page

	HR	95% CI
Estates/Porters/Security	1.21	(0.87-1.69)
Other - no patient contact	0.74	(0.61-0.91)
Other - patient contact	0.87	(0.61-1.23)

Hybrid protection:

	HR	95% CI
Vaccine status		
d2 194-253 PF PI ≤1yr	0.78	(0.38-1.61)
d2 254+ PF PI ≤1yr	0.48	(0.22-1.07)
d3 0-6 3PF PI ≤1yr	0.62	(0.17-2.21)
d3 7-66 3PF PI ≤1yr	0.31	(0.21-0.48)
d3 67-126 3PF PI ≤1yr	0.31	(0.22-0.43)
d3 127+ 3PF PI ≤1yr	NC	-
d2 194-253 PF PI >1yr	0.88	(0.46-1.69)
d2 254+ PF PI >1yr	0.70	(0.51-0.96)
d3 0-6 3PF PI >1yr	0.44	(0.10-1.90)
d3 7-66 3PF PI >1yr	0.41	(0.32-0.53)
d3 67-126 3PF PI >1yr	0.51	(0.42-0.63)
d3 127+ 3PF PI >1yr	0.87	(0.31-2.44)
Gender		
Female	1.13	(0.86-1.50)
Other	NC	-
Ethnic group		
Asian	1.14	(0.89-1.47)

Table C.17 continued from previous page

	HR	95% CI
Black	1.19	(0.72-1.96)
Mixed	1.27	(0.78-2.07)
Other	1.18	(0.78-1.77)
Prefer not to say	0.78	(0.08-8.00)
Staff type		
Doctor	0.64	(0.50-0.82)
Nursing	0.78	(0.63-0.96)
Healthcare Assistant	0.95	(0.69-1.31)
Midwife	0.99	(0.62-1.58)
Healthcare Scientist	0.77	(0.50-1.17)
Pharmacist	0.82	(0.52-1.29)
Physiotherapist/Occupational Therapist/SALT	0.70	(0.49-1.01)
Student (Medical/Nursing/Midwifery/Other)	0.82	(0.50-1.35)
Estates/Porters/Security	0.94	(0.60-1.48)
Other - no patient contact	0.64	(0.51-0.80)
Other - patient contact	1.09	(0.54-2.19)

Protection associated with primary infection:

	HR	95% CI
Primary infection status		
3-6 mth ago	0.28	(0.20-0.40)
6-9 mth ago	0.67	(0.37-1.20)
9-12 mth ago	0.72	(0.54-0.95)
12-15 mth ago	0.70	(0.61-0.81)
15+ mth ago	0.76	(0.66-0.86)

Table C.18 continued from previous page

	HR	95% CI
Vaccination status		
d3 0-6 3PF	0.97	(0.57-1.66)
d3 7-66 3PF	0.63	(0.53-0.75)
d3 67-126 3PF	0.68	(0.59-0.78)
d3 127+ 3PF	1.12	(0.68-1.84)
d2 194+ AZ	1.15	(0.86-1.54)
d3 0-6 2AZ/1PF	0.43	(0.11-1.69)
d3 7-66 2AZ/1PF	0.43	(0.33-0.55)
d3 67+ 2AZ/1PF	0.53	(0.42-0.68)
Ethnic group		
Asian	1.08	(0.87-1.33)
Black	1.14	(0.84-1.54)
Mixed	0.94	(0.67-1.33)
Other	1.30	(0.90-1.86)
Prefer not to say	1.29	(0.52-3.22)
Staff type		
Doctor	0.71	(0.58-0.85)
Nursing	0.81	(0.71-0.91)
Healthcare Assistant	1.04	(0.88-1.23)
Midwife	0.91	(0.68-1.21)
Healthcare Scientist	0.77	(0.61-0.96)
Pharmacist	0.64	(0.49-0.85)
Physiotherapist/Occupational Therapist/SALT	0.79	(0.62-1.01)
Student (Medical/Nursing/Midwifery/Other)	0.83	(0.62-1.11)

Table C.18 continued from previous page

	HR	95% CI
Estates/Porters/Security	1.06	(0.79-1.42)
Other - no patient contact	0.73	(0.61-0.88)
Other - patient contact	0.94	(0.68-1.29)

C.2.3 Sensitivity analyses for delayed-entry bias

We ran sensitivity analyses to assess the effect of having delayed entry in our studies.

C.2.3.1 Delta period

For the Delta period, we reran the models including only participants who did not enroll in the second-year extension.

Booster effectiveness in negative cohort:

	HR	95 % CI
Vaccine status		
d3 0-6 3PF	0.79	(0.36-1.70)
d3 7+ 3PF	0.41	(0.23-0.73)
d2 194+ AZ	1.71	(0.73-4.01)
d3 0-6 2AZ/1PF	2.73	(0.45-16.61)
d3 7+ 2AZ/1PF	0.40	(0.08-1.95)
Gender		
Female	0.66	(0.42-1.04)
Other	NC	-

Table C.19 continued from previous page

	HR	95 % CI
Ethnic group		
Asian	1.10	(0.57-2.13)
Black	0.80	(0.23-2.77)
Mixed	1.35	(0.43-4.27)
Other	0.44	(0.05-3.87)
Prefer not to say	NC	-
Staff type		
Doctor	1.58	(0.77-3.23)
Nursing	1.88	(0.96-3.69)
Healthcare Assistant	1.89	(0.76-4.71)
Midwife	1.36	(0.20-9.05)
Healthcare Scientist	2.54	(1.25-5.18)
Pharmacist	0.91	(0.35-2.41)
Physiotherapist/Occupational Therapist/SALT	1.88	(0.82-4.32)
Student (Medical/Nursing/Midwifery/Other)	NC	-
Estates/Porters/Security	1.47	(0.35-6.12)
Other - no patient contact	2.18	(1.05-4.53)
Other - patient contact	2.95	(0.77-11.28)

Hybrid protection. The estimates of numerous coefficients did not converge (NC).

	HR	95% CI
Vaccine and PI status		
d3 0-6 3PF PI ≤1yr ago	NC	-
d3 7+ 3PF PI ≤1yr ago	0.08	(0.01-0.69)
d3 0-6 3PF PI >1yr ago	NC	-

Table C.20 continued from previous page

	HR	95% CI
d3 7+ 3PF PI >1yr ago	0.06	(0.01-0.31)
Gender		
Female	1.11	(0.16-7.69)
Other	NC	-
Ethnic group		
Asian	0.98	(0.11-8.84)
Black	NC	-
Mixed	NC	-
Other	NC	-
Prefer	NC	-
Staff type		
Doctor	0.37	(0.05-2.92)
Nursing	0.79	(0.22-2.87)
Healthcare Assistant	0.51	(0.05-5.69)
Midwife	NC	-
Healthcare Scientist	NC	-
Pharmacist	NC	-
Physiotherapist/Occupational Therapist/SALT	0.47	(0.03-6.40)
Student (Medical/Nursing/Midwifery/Other)	NC	-
Estates/Porters/Security	NC	-
Other - no patient contact	0.50	(0.10-2.59)
Other - patient contact	NC	-

Protection associated with primary infection. The estimate of some coefficients did not converge (NC):

	HR	95% CI
Primary infection status		
3-6 mth ago	0.17	(0.03-0.89)
6-9 mth ago	0.29	(0.14-0.59)
9-12 mth ago	0.38	(0.23-0.63)
12-15 mth ago	NC	-
15+ mth ago	0.14	(0.08-0.23)
Vaccination status		
d2 194-253 PF	1.29	(0.96-1.72)
d2 254+ PF	1.59	(1.00-2.53)
d3 0-6 3PF	1.28	(0.71-2.30)
d3 7+ 3PF	0.67	(0.44-1.03)
d2 194+ AZ	2.56	(1.23-5.33)
d3 0-6 2AZ/1PF	3.87	(0.77-19.37)
d3 7+ 2AZ/1PF	0.72	(0.17-3.00)
Gender		
Female	0.62	(0.51-0.75)
Other	NC	-
Ethnic group		
Asian	1.00	(0.71-1.41)
Black	1.09	(0.58-2.07)
Mixed	1.01	(0.52-1.97)
Other	0.79	(0.37-1.69)

Table C.21 continued from previous page

	HR	95% CI
Prefer not to say	0.80	(0.11-5.90)
Staff type		
Doctor	0.95	(0.71-1.27)
Nursing	1.17	(0.89-1.55)
Healthcare Assistant	1.47	(1.06-2.02)
Midwife	0.81	(0.36-1.85)
Healthcare Scientist	1.24	(0.93-1.65)
Pharmacist	0.88	(0.51-1.52)
Physiotherapist/Occupational Therapist/SALT	1.01	(0.65-1.58)
Student (Medical/Nursing/Midwifery/Other)	1.54	(0.79-3.02)
Estates/Porters/Security	0.95	(0.53-1.71)
Other - no patient contact	1.17	(0.81-1.68)
Other - patient contact	1.26	(0.54-2.92)

C.2.3.2 Omicron period

For the Omicron period, we reran the models including only participants who did enroll in the second-year extension. This is an alternative way to assess the potential impact of delayed-entry bias. An analysis with only participants who did not enroll in the extension has been attempted as well, but has been found to be severely under powered. The reason is that almost all participants who were in SIREN during the Omicron period were those who had enrolled in the second-year extension.

Relative effectiveness of booster in negative cohort:

	HR	95% Ci
Vaccination status		
d3 0-6 3PF	1.13	(0.48-2.69)
d3 7-66 3PF	0.67	(0.48-0.93)
d3 67-126 3PF	0.70	(0.53-0.94)
d3 127+ 3PF	0.85	(0.59-1.21)
d2 194+ AZ	1.46	(0.81-2.62)
d3 7-66 2AZ/1PF	0.42	(0.25-0.68)
d3 67+ 2AZ/1PF	0.62	(0.42-0.91)
Gender		
Female	1.18	(1.00-1.39)
Other	1.26	(0.45-3.48)
Ethnic group		
Asian	1.21	(0.98-1.48)
Black	1.08	(0.62-1.85)
Mixed	1.09	(0.72-1.66)
Other	1.30	(0.78-2.19)
Prefer not to say	1.59	(0.43-5.90)
Staff type		
Doctor	0.72	(0.57-0.90)
Nursing	0.82	(0.70-0.96)
Healthcare Assistant	1.07	(0.84-1.37)
Midwife	0.90	(0.64-1.26)
Healthcare Scientist	0.81	(0.61-1.08)
Pharmacist	0.62	(0.40-0.95)
Physiotherapist/Occupational Therapist/SALT	0.81	(0.60-1.11)

Table C.22 continued from previous page

	HR	95% CI
Student (Medical/Nursing/Midwifery/Other)	1.02	(0.76-1.36)
Estates/Porters/Security	1.40	(0.95-2.08)
Other - no patient contact	0.80	(0.65-1.00)
Other - patient contact	0.86	(0.58-1.26)

Hybrid protection. The estimates of some coefficients did not converge (NC).

	HR	95% CI
Vaccination and PI status		
d2 194-253 PF PI ≤1yr	0.79	(0.38-1.67)
d2 254+ PF PI ≤1yr	0.45	(0.20-0.99)
d3 0-6 3PF PI ≤1yr	0.59	(0.16-2.13)
d3 7-66 3PF PI ≤1yr	0.34	(0.22-0.51)
d3 67-126 3PF PI ≤1yr	0.32	(0.24-0.43)
d3 127+ 3PF PI ≤1yr	0.23	(0.08-0.67)
d2 194-253 PF PI >1yr	0.83	(0.46-1.48)
d2 254+ PF PI >1yr	0.68	(0.50-0.91)
d3 0-6 3PF PI >1yr	0.44	(0.10-1.87)
d3 7-66 3PF PI >1yr	0.43	(0.34-0.55)
d3 67-126 3PF PI >1yr	0.52	(0.43-0.62)
d3 127+ 3PF PI >1yr	0.59	(0.42-0.83)
Gender		
Female	1.09	(0.85-1.39)
Other	NC	-
Ethnic group		

Table C.23 continued from previous page

	HR	95% CI
Asian	1.16	(0.93-1.43)
Black	1.33	(0.83-2.15)
Mixed	1.16	(0.73-1.83)
Other	1.07	(0.72-1.58)
Prefer not to say	0.66	(0.07-6.40)
Staff type		
Doctor	0.71	(0.56-0.89)
Nursing	0.86	(0.70-1.06)
Healthcare Assistant	0.89	(0.65-1.21)
Midwife	0.92	(0.57-1.47)
Healthcare Scientist	0.88	(0.59-1.30)
Pharmacist	0.75	(0.46-1.22)
Physiotherapist/Occupational Therapist/SALT	0.74	(0.52-1.04)
Student (Medical/Nursing/Midwifery/Other)	0.89	(0.56-1.39)
Estates/Porters/Security	0.98	(0.61-1.59)
Other - no patient contact	0.69	(0.56-0.84)
Other - patient contact	1.08	(0.57-2.04)

Protection associated with primary infection:

	HR	95% CI
Primary Infection Status		
3-6 mth ago	0.36	(0.26-0.52)
6-9 mth ago	0.47	(0.25-0.87)
9-12 mth ago	0.73	(0.54-1.00)
12-15 mth ago	0.82	(0.69-0.98)

Table C.24 continued from previous page

	HR	95% CI
15+ mth ago	0.80	(0.71-0.91)
Vaccination status		
d3 0-6 3PF	0.83	(0.41-1.66)
d3 7-66 3PF	0.64	(0.49-0.83)
d3 67-126 3PF	0.70	(0.56-0.88)
d3 127+ 3PF	0.82	(0.62-1.09)
d2 194+ AZ	1.06	(0.73-1.53)
d3 0-6 2AZ/1PF	0.33	(0.04-2.72)
d3 7-66 2AZ/1PF	0.46	(0.33-0.65)
d3 67+ 2AZ/1PF	0.54	(0.40-0.73)
Gender		
Female	1.18	(1.01-1.37)
Other	0.95	(0.36-2.51)
Ethnic group		
Asian	1.20	(0.99-1.46)
Black	1.07	(0.73-1.57)
Mixed	1.22	(0.89-1.66)
Other	1.21	(0.83-1.77)
Prefer not to say	1.09	(0.30-3.95)
Staff type		
Doctor	0.79	(0.66-0.95)
Nursing	0.85	(0.74-0.97)
Healthcare Assistant	0.94	(0.77-1.16)

Table C.24 continued from previous page

	HR	95% CI
Midwife	0.76	(0.56-1.03)
Healthcare Scientist	0.82	(0.62-1.07)
Pharmacist	0.66	(0.47-0.92)
Physiotherapist/Occupational Therapist/SALT	0.79	(0.60-1.04)
Student (Medical/Nursing/Midwifery/Other)	1.00	(0.78-1.28)
Estates/Porters/Security	1.28	(0.95-1.73)
Other - no patient contact	0.80	(0.66-0.97)
Other - patient contact	0.89	(0.61-1.28)