Mathematical modelling of complete recessive lethals: Adaptive dynamics across varying genetic and population structures

Dissertation

zur
Erlangung des Doktorgrades (Dr. rer. nat.)
der
Mathematisch-Naturwissenschaftlichen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Luis Aniello La Rocca

aus
Malsch

Bonn, September 2024

Angefertigt mit Genehmigung der

Mathematisch-Naturwissenschaftlichen Fakultät der

Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Gutachter: Prof. Dr. Anton Bovier

2. Gutachter: Prof. Dr. Peter Krawitz

Tag der Promotion:

Erscheinungsjahr:

Abstract

The study of recessive lethal diseases in populations presents significant challenges, from estimating key parameters - such as the number of genes involved and the mutation rates driving genetic degeneration - to understanding the increased prevalence of autosomal recessive intellectual disorders (ARID) in the offspring of consanguineous unions.

To address these challenges, we developed a mathematical model based on a diploid individual-based framework of adaptive dynamics. This model allowed us to make several important discoveries.

First, we showed that the higher disease burden for ARID observed in consanguineous unions is a transient phenomenon associated with rapidly expanding population sizes. This finding highlights the need for widespread carrier screening as the drop in prevalence in randomly mating populations is associated with an increased mutation burden.

Second, we extended the drift-barrier hypothesis, which states that the ability of natural selection to refine traits is limited by genetic drift. We introduced a new parameter - the recessive gene count. We found that populations with a higher gene count face a similar barrier to that imposed by an increased mutation rate. In addition, our analysis provides a new perspectives on Muller's ratchet, a classic concept in population genetics that describes the irreversible accumulation of deleterious mutations in the absence of recombination. Our results show how mutations accumulate rapidly after a long period of stability, and how the population finds its way back to stability after the emergence of clusters of highly correlated genes.

Finally, we have implemented a simulation framework based on Gillespie's algorithm, which allows exact stochastic simulations of our model. This framework allows the study of the dynamics of complex interacting systems. The tool is flexible, scalable, and designed to facilitate further studies.

Acknowledgements

Without the support of many people, this thesis would not have been possible. I would like to take this opportunity to express my gratitude.

First and foremost, I want to extend my deepest thanks to my advisor, Anton Bovier. Your unwavering belief in me and constant support, even when I doubted myself, have been pivotal throughout my PhD journey. I have greatly benefited from your vast mathematical knowledge, remarkable intuition, and thoughtful guidance. Thank you for always making time for me, for your patience when I lost sight of the bigger picture, and for giving me the confidence to persevere.

I would also like to express my heartfelt thanks to Peter Krawitz, who dedicated countless hours to our papers and helped develop the intuition for our model. His contributions have been crucial in shaping this work, and I am deeply grateful for his mentorship.

Special thanks also go to Anna Kraut, whose constant encouragement and support throughout the course of my PhD helped me stay grounded.

A warm thank you goes to Mei-Ling, the heart and soul of the wor group. Her kindness, care, and tireless efforts in supporting all of us have made the journey so much smoother. Mei-Ling, your dedication to the group and to each of us personally has been invaluable, and I am incredibly grateful.

I would also like to acknowledge the entire stochastic group for creating such a welcoming and stimulating environment. In particular, I am grateful to the fellow PhD students and colleagues with whom I shared this journey.

I am immensely grateful to my wife, Aline, and to my family for being by my side during all the highs and lows, and for enduring my bad moods and crises with such grace.

Additionally, I would like to thank my badminton family at 1.BC Beuel. The camaraderie and time spent on the court provided much-needed relief and balance during this intense period.

I want to express my appreciation for the financial support provided by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy GZ 2047/1, Projekt-ID 390685813 and GZ 2151, ProjectID 390873048 and through the Priority Programme 1590 "Probabilistic Structures in Evolution" and the Bonn International Graduate School in Mathematics (BIGS). Being part of these programs and engaging with the broader research community has been an enriching experience.

Finally, I want to thank the members of my committee, Anton Bovier, Peter Krawitz, Alexander Effland, and Kevin Thurley, for taking the time to review my work.

Contents

List of Figures

1 Introduction

Mathematical models have long been essential for understanding, quantifying and predicting natural phenomena. From the first logistic models of population growth developed by Malthus to the groundbreaking models of population genetics developed by Wright, Fisher and Haldane in the 1920s and 1930s, mathematics has been a crucial tool in unravelling the fundamental principles of evolution. The neutral models of Kimura in the 1950s highlighted the importance of genetic drift and randomness in evolution. The introduction of adaptive dynamics models in the 1990s further pushed the boundaries of how we approach evolutionary processes.

The advent of genome sequencing - beginning with the 1000 Genomes Project and continuing into the era of third-generation sequencing technologies - has led to a more detailed understanding of complex genetic processes. This deeper knowledge has driven the need for more sophisticated models that can integrate multiple factors and isolate the most critical elements influencing evolutionary dynamics. Inevitably, any biological model will reproduce certain aspects of nature with greater fidelity than others: "All models are wrong, but some are useful". [27] It is also important to recognise that no model can reproduce the full complexity of natural phenomena. It is therefore up to the researcher to identify which information is essential to answer the questions posed and which elements introduce unnecessary noise: "Sensing which assumptions may be critical and which are irrelevant to the question at hand is the art of modelling". [94]

The work that led to the results presented in this thesis can be divided into three distinct phases. The first step is to generate a mathematical model that accurately describes the biological, genetic or evolutionary process under consideration. This requires a deep understanding of the theoretical biological framework that governs the natural phenomenon. Once we have a clear and rigorous mathematical formulation of the model, we implement it using stochastic simulation algorithms. The second phase focuses on developing an intuition for the dynamics of the model and understanding its behaviour. Where possible, we use mathematical analysis, such as solving ordinary differential equations (ODEs), but this is often challenging due to the high dimensionality and complexity of the models. As a result, much of the analysis is numerical. We systematically compare different versions of the model at three levels of complexity: first, by changing the underlying dynamics (e.g. comparing adaptive dynamics with population genetics); second, by changing internal model mechanisms (e.g. different mating schemes in populations); and third, by varying key parameter regimes (e.g. mutation rates). Through this comparative approach, we develop an intuition about which mechanisms and parameters are most relevant to the biological questions being addressed. The final step is to integrate these insights into the applied fields. This requires not only a solid understanding of existing theory, but also the ability to translate the mathematical insights into meaningful contributions to biological and genetic research.

The results help to generate new research questions, both in the applied sciences and in the development of mathematical methods.

This work builds on a strong foundation of mathematical modelling to develop and explore complex systems. The focus is on understanding the spread and persistence of recessive autosomal diseases across different genetic and population structures.

The prevalence of severe autosomal recessive diseases (i.e. disease frequency within the population) and mutation burden (average number of deleterious mutations per individual) are analysed under varying population sizes and mating conditions, while holding mutation rates and genetic architecture constant. The simulations start with a population of 500 individuals free of deleterious mutations and expand to $10,000$ after an initial equilibrium is reached. The models compare random and consanguineous mating patterns, the latter influenced by family size (κ) and the probabilities (α, β) of mating within close or extended family structures. In randomly mating populations, there is a temporary decrease in disease prevalence after population expansion, followed by a long-term increase in mutation load. It takes more than 500 generations for disease prevalence to stabilise, while mutation burden remains elevated. In contrast, consanguineous populations show stable prevalence and mutation burden during expansion because their mating patterns, constrained by family size, are unaffected by population growth.

The effects of mutation rate (μ), the number of recessive genes (N, recessive gene count) and recombination (r) on population stability were also investigated. In the absence of recombination, the loss of mutation-free haplotypes leads to mutation fixation and potential extinction. This effect is particularly pronounced at high mutation rates or recessive gene counts. Recombination stabilises the population by reducing the mutation load, allowing tolerance to higher recessive gene counts. The findings could be used to refine the drift-barrier hypothesis, an evolutionary theory, by including the recessive gene count as a parameter that influences the genetic drift of a population.

In Chapter 1.1 we first introduce basic genetic concepts that will serve as the basis for modelling these situations. We then examine some of the core mathematical models in evolutionary biology in Chapter 1.2, before introducing the primary model that underpins this thesis in Chapter 1.3. Finally, in Chapters 2 and 3 we adapt and apply this model to different scenarios in order to answer specific research questions related to recessive diseases.

The complexity of these models often pushes mathematical analysis to its limits. To address this challenge, we implement a simulation framework based on the Gillespie algorithm. This simulation approach allows us to efficiently study our high-dimensional models and gain insights that would otherwise be difficult to obtain using purely analytical methods. Chapter 4 serves to preserve the simulation framework, thereby facilitating adaptation and further development of the algorithm by subsequent researchers.

1.1 General biological background

In order to bridge the disciplines by modelling genetic processes, it is essential to have a clear understanding of the mechanism that occur in nature. This chapter establishes a general biological foundation for understanding the applications of the mathematical models discussed in this thesis and introduces a basic vocabulary to facilitate comprehension of the results presented in Chapters 2 and 3. These results are addressed to a genetic audience rather than a mathematical one. This chapter starts with an exposition of the fundamental genetic principles that underpin molecular biology. We then proceed to examine the mechanisms that govern evolutionary processes, delving deeper into the intricacies of inheritance and the evolution of sexual mating. We conclude this chapter with an exploration of recessive diseases, which represent the primary application of the models discussed in this thesis.

1.1.1 Genetic background

This section begins with a short investigation of the fundamental structural components of living organisms at the cellular level, subsequently progressing to an analysis of the mechanisms of reproduction. For a more comprehensive exploration of molecular cell biology, we recommend the textbook by Lodish [144].

All living organisms are composed of cells, which are the fundamental units of life, capable of performing all essential life functions. Organisms are broadly categorised into two main types: *Prokaryotes* and *eukaryotes*, which differ in their structural composition. Prokaryotes are unicellular organisms, comprising a single cell. In the absence of a defined nucleus, the genetic material of prokaryotic cells is located freely within the inside of the cell. These cells are typically smaller and possess a reduced number of internal structures in comparison to those observed in eukaryotic cells. The domain of prokaryotes is subdivided into two distinct groups: bacteria and archaea. In contrast, eukaryotic cells are found in more complex organisms, including plants, animals, fungi, and protists. They are more advanced, larger in size and possess a well-defined nucleus that contains the cell's genetic material. Additionally, eukaryotic cells possess other specialized structures, known as organelles, which include mitochondria, the endoplasmic reticulum, and chloroplasts in plant cells.

The *genome* represents the fundamental unit of biological processes, encompassing the complete set of genetic information present within an organism. The genome is composed of deoxyribonucleic acid (DNA), which is a molecule comprising two long chains of nucleotides twisted into a double helix. Each nucleotide comprises a sugar molecule, a phosphate group, and one of four nitrogenous bases: adenine (A), thymine (T), cytosine (C), or guanine (G). The specific sequence of these bases encodes all the instructions necessary for the construction and maintenance of an organism.

In eukaryotic cells, deoxyribonucleic acid (DNA) is organised into structures known as *chromosomes*. Each chromosome is composed of a single, long DNA molecule that is wrapped around proteins called histones. The number of chromosomes differs between organisms. For example, the human genome comprises 46 chromosomes, organised into 23 pairs (Human Genome Project, 2001; [138]), whereas the fruit fly genome has only eight chromosomes

[187]. Within the DNA, specific sequences of base pairs (bp) form *genes*, which are the fundamental units of heredity. The size of a gene can vary, with measurements typically expressed in 1000 base pairs (kb). For example, in humans the average gene comprises 10 to 15 kb, but can range from approximately 0.2 kb (tyrosine tRNA gene) to over 2,500 kb (dystrophin gene) [194]. A gene contains the instructions for the synthesis of a specific protein or set of proteins, which in turn perform a range of functions within the organism. Each gene is located at a specific position on a chromosome known as a *locus*. Different versions, of a gene are called *alleles*. To illustrate, a gene that determines flower colour in a plant may possess one allele that codes for purple flowers (P) and another that codes for white flowers (p). The combination of alleles that an organism possesses for a specific gene is defined as its *genotype*. The genotype is defined as the genetic makeup of an organism, specifically the set of alleles that an individual possesses for a particular gene or set of genes. The term *phenotype* is used to describe the observable characteristics or traits of an organism that result from the interaction of its genotype with the environment. Hence, in the context of the flower colour gene, the genotype would be defined as the alleles present (e.g. one purple allele and one white allele), whereas the phenotype would be the actual colour of the flower. The relationship between genotype and phenotype is not always straightforward. In some cases, alleles may be designated as dominant, which implies that they can obscure the effects of other alleles at the same locus. In the case of recessive alleles, the trait is only expressed when two copies of the same allele are present. Within the example, in pea plants, the allele responsible for purple flowers (P) is dominant over the allele that causes white flowers (p). Therefore, a plant with the genotype PP or Pp will exhibit purple flowers, whereas only a plant with the genotype pp will display white flowers [157]. An organism is described as *homozygous* for a particular gene if both alleles at a locus are identical (e.g. PP or pp for flower colour). An organism is *heterozygous* if the two alleles at a locus are different (e.g. Pp).

The majority of eukaryotic organisms undergo sexual reproduction, during which their cells exist in two forms: haploid and diploid. *Haploid* cells contain a single complete set of chromosomes (n). In humans, haploid cells are *gametes*, specifically sperm and egg cells, which contain 23 chromosomes. *Diploid* cells contain two complete sets of chromosomes (2n), with one set inherited from each parent. Hence, human body cells for example are diploid and possess 46 chromosomes (23 pairs). Sexual reproduction entails the combination of genetic material from two parents through the processes of meiosis and fertilisation. During *meiosis*, the number of chromosomes is reduced by half, resulting in the production of haploid gametes (sperm and eggs). The gametes unite during *fertilisation* to form a diploid *zygote*. This introduces genetic diversity, as the offspring inherit a unique combination of alleles from both parents.

In 1865, Gregor Mendel was the first to observe how traits are transferred from one generation to the next through his experiments with pea plants [157]. His work, however, went largely unnoticed until it was rediscovered in 1900 by Hugo de Vries [56], Carl Correns [50], and Erich Tschermak [201]. Mendel's observations were subsequently formulated into the well-known Mendelian rules of inheritance, which consist of three fundamental laws. First, the *Law of Independent Segregation* states that during meiosis, the two copies of a gene segregate from each other, and each gamete carries only one allele for each gene. Second, the *Law of Independent Assortment* explains that the segregation of alleles for one gene occurs

independently of the segregation of alleles for other genes. This means that different traits are inherited independently of each other, leading to a variety of genetic combinations in the offspring. Lastly, the *Law of Dominance and Uniformity* states that some alleles are dominant, while others are recessive. This law explains why, in the case of heterozygous pairs, the dominant allele will mask the expression of the recessive allele, leading to uniform expression of the dominant trait in the offspring.

With advances in genome sequencing, the once-clear boundaries of Mendel's laws have become increasingly blurred. Research has revealed that certain genes do not conform to the traditional paradigms of dominance and recessiveness; instead, they may display phenomena such as incomplete dominance or co-dominance. In these cases, the expression of both alleles results in a novel phenotype that is a blend of the parental traits, challenging the simplicity of Mendelian inheritance models [195]. Moreover, the emergence of epigenetics has further complicated our understanding of gene segregation. It has become evident that gene expression is not solely determined by genetic makeup but is also significantly influenced by environmental factors and intricate regulatory mechanisms. These influences can disrupt the assumption of independence in gene segregation, suggesting that the interplay between genetics and environment is far more complex than previously acknowledged [61, 23]. Despite these advancements and the complexities they introduce, Mendel's foundational hypotheses continue to underpin many biological models. His laws provide a fundamental framework for understanding inheritance, serving as a starting point from which the intricate tapestry of modern genetics can be explored. Contemporary genetic research builds upon these principles, integrating insights from genome sequencing, epigenetics, and molecular biology. This synthesis not only enriches our comprehension of genetic inheritance but also lays the groundwork for more sophisticated models in evolutionary biology and population genetics.

During meiosis, *recombination* may occur, which involves the exchange of genetic material between paired chromosomes. This phenomenon further increases genetic variation by shuffling alleles prior to their transmission to the offspring. Another mode of reproduction is clonal (asexual) reproduction, whereby offspring are produced by a single parent without the involvement of gametes. The offspring are genetically identical to the parent (clones), except for rare *mutations*, which are random changes in the DNA sequence. A common example of clonal reproduction occurs in bacteria, which reproduce by binary fission, a process whereby a single bacterium divides into two identical daughter cells. All cells are subject to mutation, with the frequency of mutation dependent on a number of factors, including radiation, age and other environmental variables. The impact of a mutation can be highly variable. In some cases, a mutated gene may still result in the same protein, while in others, the functionality may be completely lost (loss of function).

In Chapter 1.5, we take a closer look at mutation rates across the genome, examining their variability and the underlying mechanisms driving these differences. In Chapter 1.6, we dive deeper into the mechanisms of recombination, investigating how genetic material is exchanged between chromosomes during meiosis.

1.1.2 Theory of evolution

Evolution shapes biological diversity and the adaptation of organisms to their environments, and modelling these processes effectively requires a solid grasp of evolutionary theory. This section provides a brief introduction to the theory of evolution, exploring its historical origins and laying out the essential concepts necessary for mathematical modelling

When thinking of the theory of evolution, most people immediately recall Charles Darwin's seminal work *On the Origin of Species*, published in 1859 [54]. Darwin's idea of natural selection - the survival of the fittest and best-adapted individuals within a species - sparked significant controversy among biologists, especially in the early 20th century. While much of the criticism of Darwin's theory arose from non-scientific circles, debates also occurred within the scientific community about the nature of evolution.

With the rediscovery of Mendelian inheritance laws in 1900, Darwin's idea that evolutionary changes happen gradually and incrementally faced opposition. Some biologists began proposing that evolutionary changes occur in more pronounced leaps rather than through continuous, small adjustments. It was not until the 1920s and 1930s that the founding fathers of population genetics - Sewall Wright, Ronald Fisher, and J.B.S. Haldane - developed a unifying model that combined the gradualism of Darwinism with the rules of heredity provided by Mendelism [68, 97, 213]. These groundbreaking population genetic models, which we will discuss in section 1.2.2, provided a framework for understanding the evolutionary process more comprehensively.

In this section, we explore the evolutionary process. For readers seeking a broader historical perspective, we refer to Provine's work for further classification and insights into the development of population genetics [181].

As Darwin noted in his original work, three primary mechanisms drive evolution. The first is *heredity*, which refers to the process of reproduction in which individuals pass their traits on to their offspring. Secondly, there is *variation*, meaning that traits differ between offspring, and heredity is not perfect. Finally, there is *natural selection*, which acts on the variations mentioned above. Different traits confer different fitness levels, meaning that some traits have a higher probability of being passed on to the next generation due to a higher reproductive or survival rate.

One of the main criticisms of Darwin's theory of evolution concerned the origin of variation. At the time, it was thought that the traits of offspring were a blending of parental traits, which, in a randomly mating population, would lead to a gradual loss of variation over generations. Without variation, natural selection would have nothing to act upon. However, variation was observable within populations, and this suggested that the traits of offspring must differ from those of their parents. This idea countered the argument that offspring of parents favoured by natural selection would automatically enjoy the same selective advantage. Indeed, we now know that separation fosters evolution. Sub-populations in isolated habitats, such as the famous Galápagos finches studied by Charles Darwin himself, tend to be more adapted to their specific environments than panmictic populations.

In Darwin's time, mutation was the only known source of variation, though today we understand that there are many others, such as recombination or horizontal gene transfer (HGT), the transfer of genetic material between individuals rather than inheritance from parent to offspring. This phenomenon is observed in bacteria, where it plays a major role in adaptation, but has also been documented in some eukaryotes, expanding our understanding of how genetic material can influence evolution outside traditional inheritance models [112, 169].

Natural selection arises from interactions between individuals and their environment, as well as with other individuals. These interactions can be competitive, such as competition between individuals of the same or different species for resources or mating partners. Alternatively, they can be dependent interactions, like predator-prey or parasite-host relationships. Additionally, symbiotic relationships can also emerge, where cooperation between species results in mutual benefits.

It wasn't until the 1950s that the DNA molecule, now known to be the basis of heredity, was discovered by James Watson and Francis Crick [207], Maurice Wilkins [210], and Rosalind Franklin [75]. Unlike the phenotype, which can vary throughout an individual's life due to environmental factors, the DNA - or genotype - of an individual remains unchanged, except for errors that may occur during replication. Another opposition to Darwin's theory was Lamarck's idea of inheritance, which is often illustrated with the image of a giraffe. He hypothesized that traits acquired during an organism's lifetime could be passed on to its offspring [137]. For example, a giraffe that stretches its neck to reach high leaves would pass this elongated neck to its offspring, leading to the evolution of long necks over generations. While Lamarck's theory has been widely refuted, some aspects of his ideas have regained attention with the advances in epigenetics. In certain traits, it is not just the availability of the appropriate genetic material that is important, but also which parts of the DNA get expressed. These epigenetic changes - modifications in gene expression gained throughout an individual's lifetime - can, in some cases, be passed on to offspring [143].

Ultimately, as we now understand, heredity ensures that traits are passed down, while variation introduces new differences upon which selection can act. These mechanisms form the foundation of evolutionary change and have been greatly clarified by advances in genetics. For more details on the evolutionary process, we refer readers to the well-curated *Encyclopedia Britannica*, which provides an extensive and insightful exploration of evolutionary biology and its key concepts [10].

1.1.3 Evolution of sexual mating

In the overwhelming majority of eukaryotic multicellular organisms, sexual reproduction represents the exclusive means of reproduction [205]. The evolution of sexual reproduction has presented a significant challenge to the field of biology for many years. While some progress has been made and a number of hypotheses have been proposed, a unifying theory remains elusive. Two principal questions are put forth for consideration. The first question pertains to the origin and evolutionary history of sexual reproduction. In prokaryotes, such as bacteria and archaea, one can also find modes of exchange or transfer of genetic material from one individual to another. Such processes include conjunction, transformation and

translation. Nevertheless, it remains unclear whether these are the point of origin of sexual reproduction in eukaryotes [183]. The second question refers to the maintenance of sexual reproduction: After sexual mating evolved, how does it persist in such a highly competitive world, especially across such a large class of organisms? Sexual reproduction introduces several costs compared to asexual reproduction, such as the twofold cost of producing males and the energy required to find a mate (see below, [51, 142, 192]). Yet, despite these disadvantages, sexual reproduction persists widely across species, including most plants and animals.

Clonal reproduction, often seen in organisms that reproduce asexually, indeed appears to be more efficient than sexual reproduction in several respects. Sexual reproduction requires a significant amount of energy, primarily due to the complex cellular processes involved [141]. In sexual reproduction, meiosis is a key process where a diploid cell undergoes division to produce haploid gametes. This process is not only energy-intensive but also time-consuming. After meiosis, sexual reproduction involves the fusion of gametes (fertilization), where the sperm and egg unite to form a zygote. This step is necessary to restore the diploid state in the offspring and again requires precise and energy-intensive cellular machinery. Following gamete fusion, the nuclei of the gametes must also fuse, combining genetic material from both parents. This process is another layer of complexity that requires additional energy and time.

In contrast, clonal reproduction, such as binary fission in prokaryotes or mitosis in eukaryotes, is much simpler. The genetic material is directly duplicated and divided between two daughter cells, without the need for meiosis or gamete fusion. Because clonal reproduction bypasses the elaborate steps required in sexual reproduction, it consumes far less energy. This makes it a more efficient process at the cellular level, especially in environments where resources are scarce.

In sexually reproducing species, finding a mate can be a significant challenge. The process of searching for and selecting a mate involves the time and energy spent locating and courting potential partners. Many species engage in elaborate mating rituals or displays to attract mates, such as the colourful plumage displays in birds or the production of pheromones in insects. These behaviours, while important for sexual selection, require additional energy and resources. In many species, individuals must compete with others to secure a mate. In some animals, such as certain mammals and birds, males may physically compete with each other for access to females, which can lead to injuries or even death. Males may need to invest in physical attributes (like antlers in deer) or produce elaborate displays to outcompete rivals, which can divert resources from other survival-related activities. The time and energy invested in finding and securing a mate can result in missed opportunities to forage for food, care for offspring, or avoid predators, which can have a direct impact on survival and reproductive success.

One of the key features of sexual reproduction is the recombination of genetic material during meiosis, which shuffles alleles and creates new combinations in offspring. While this can introduce genetic diversity, which is beneficial for adaptation to changing environments, it can also disrupt advantageous allele combinations that have been selected for in previous generations [166]. Over time, certain combinations of alleles can become highly adapted to

specific environmental conditions, leading to a well-functioning "gene complex." Recombination can break apart these co-adapted gene complexes, resulting in offspring that may be less well-adapted to their environment, thus reducing individual fitness.

In many species sexual reproduction requires close physical contact between individuals, which can facilitate the transmission of infectious diseases. This is particularly relevant in species where mating behaviours involve prolonged contact or where there are multiple partners [171]. The spread of diseases within a population can lead to increased mortality and reduced reproductive rates, which can have broader implications for population stability and growth. In extreme cases, it could even lead to population decline or extinction if the disease burden becomes too high.

Another well-known disadvantage of sexual reproduction is the *twofold cost of sex*. In asexual populations, all individuals can reproduce, effectively doubling the population size each generation. In contrast, sexual populations require two individuals (a male and a female) to produce offspring, which halves the per-capita birth rate compared to asexual population. Sexual reproduction requires resources to be allocated towards the production of males, which do not directly produce offspring but are necessary for fertilization. This allocation can be seen as inefficient when compared to asexual reproduction, where all individuals contribute directly to the next generation [192].

Theories and models have been proposed to explain the evolution and maintenance of sexual reproduction in a vast array of species [16, 88, 173]. Below is a brief overview of some of the most prominent theories:

1. **The Red-Queen hypothesis** [15, 203]
   This theory, named after the Red-Queen's race in Alice in Wonderland, suggests that species must continuously evolve to survive in a world where their environment, including their predators, parasites, and competitors, is also constantly evolving. Sexual reproduction provides a mechanism for generating genetic diversity, which allows populations to adapt more rapidly to these changing conditions. One of the key examples supporting the Red-Queen hypothesis is the co-evolution of hosts and parasites. Parasites evolve to exploit common host genotypes, while sexual reproduction in hosts shuffles genes, creating novel genotypes that are more resistant to parasite infections.

2. **Muller's ratchet** [67, 162]
   In asexual populations, harmful mutations can accumulate over time because there is no mechanism to eliminate them without some form of genetic recombination. This process, known as Muller's Ratchet, leads to a gradual decrease in fitness. Sexual reproduction, through recombination, can bring together multiple beneficial mutations while purging deleterious ones, thus maintaining the overall health and fitness of a population. In section 1.6.2 we discuss this theory in detail, including mathematical models that underpin the hypothesis.

3. **Tangled bank hypothesis** [147, 59, 77]
   The Tangled Bank Hypothesis proposes that sexual reproduction generates diversity, which allows organisms to exploit a variety of ecological niches. This diversity can

reduce competition among offspring by enabling them to specialize in different niches within the environment, thus improving the survival chances of the species as a whole.

4. **Fisher-Muller Hypothesis** [69, 161]
   This hypothesis suggests that sexual reproduction allows for the combination of beneficial mutations from different individuals. In asexual populations, beneficial mutations must occur sequentially in the same lineage to combine, which can be a slow process. In contrast, sexual reproduction can bring together different beneficial mutations from separate lineages in a single generation, accelerating the process of adaptation.

5. **Mutational deterministic hypothesis** [124, 125]
   This theory posits that sexual reproduction is beneficial when deleterious mutations interact synergistically, allowing for more effective purging of harmful alleles. While it provides a robust framework for understanding the advantages of sex under high mutation rates, empirical support for synergistic epistasis is limited, and the model's assumptions may not always hold.

No single theory fully explains the dominance of sexual reproduction, and each theory has its strengths and limitations. It's likely that the selective advantages of sexual reproduction are context-dependent, with different mechanisms being more relevant in different ecological and evolutionary scenarios.

Some species have evolved to possess both modes of reproduction. Some of these species even coexist in sexual and asexual lineages, either alternatively through the life cycle [198] or in spatially or temporally isolated populations [186].

To develop a more unified understanding of sexual reproduction, it is necessary to construct and analyse more complex population models that can integrate multiple theories. Such models would need to account for various factors, including genetic diversity, environmental variability, mutation rates, and species interactions [102, 209].

1.1.4 Complete recessive lethals

In autosomal recessive inheritance, a genetic condition manifests only when an individual inherits two copies of a pathogenic variant, one from each parent. Since the gene in question is located on one of the autosomes (the non-sex chromosomes), the inheritance pattern is independent of the individual's sex. Individuals with only one copy of the variant (heterozygotes) are typically unaffected by the disorder but are known as *carriers*. Carriers have a 50% chance of passing the mutated gene to their offspring, but the condition only presents itself if both parents pass on the mutated gene. In such cases there is a 25% chance that the child inherits two mutated alleles and expresses the disorder, a 50% chance that the child inherits one mutated allele and becomes a carrier and a 25% chance that the child inherits no mutated alleles and is unaffected.

A subset of autosomal recessive disorders is classified as lethal, meaning they result in early death or prevent the individual from reproducing. One notable group within this subset is *autosomal recessive intellectual disabilities* (ARID), which encompass a wide range of

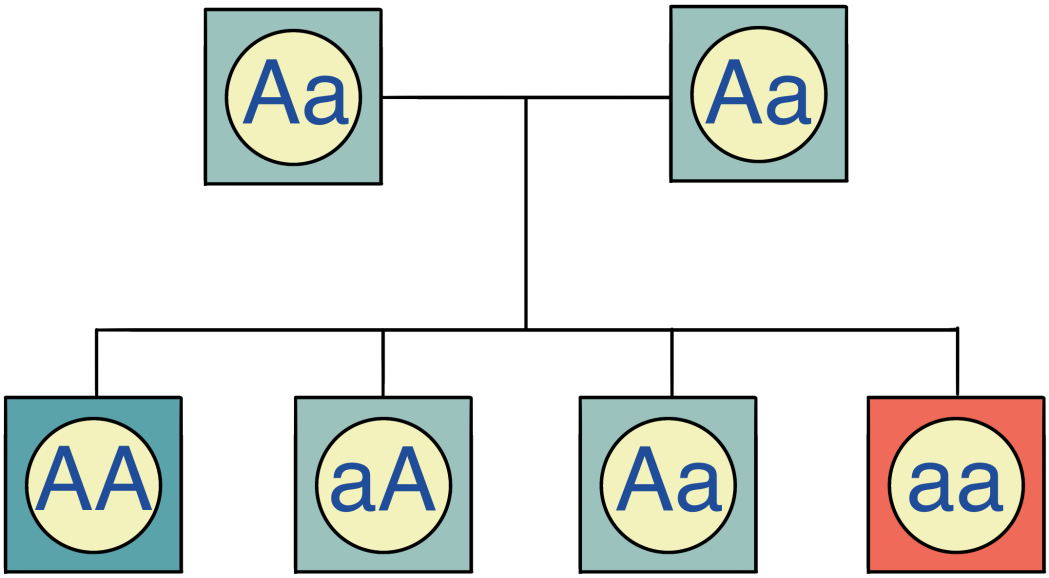

Figure 1.1: Recessive inheritance. The potential outcomes of the inheritance of two alleles from two carriers are as follows. In one case, the offspring inherits both the wild type alleles. In two cases, the offspring is again a carrier, having a heterogeneous allele combination. In one case, the individual inherits two mutated alleles, thus expressing the disease.

neurodevelopmental disorders characterized by impaired cognitive functioning and adaptive behaviours [93].

This class of disorders may also be caused by a *compound heterozygous* mutation. Compound heterozygousity introduces further complexity into the inheritance and manifestation of autosomal recessive disorders. In this scenario, an individual inherits two different pathogenic variants in the same gene, one from each parent. Although each parent may carry a different mutation, the combination can still result in a recessive disorder.

The mathematical models discussed in this thesis describe the dynamics of such disorders within populations. By considering factors like mutation rates, the number of ARID genes, these models try to predict the spread and maintenance of recessive alleles in a population. The mutation rate and the number of recessive genes are challenging to estimate, even with the aid of modern sequencing techniques [107]. This makes them one of the principal variables in our analysis.

The *de novo* mutation rate is the rate at which new deleterious mutations appear in the human genome. Current estimates suggest that new mutations occur at a rate of approximately 1.2×10^{-8} per base pair (bp) per generation. However, the coding sequences of genes, which are most relevant to the expression of ARID, vary significantly in length, ranging from about 500 to 10 000 bp. This variability implies that the *mutation rate per gene* can differ considerably depending on the gene's length [119, 122].

Advances in sequencing technologies have identified over 600 genes associated with ARID, when mutated [122]. However, this figure is likely an underestimation. Many cases of ARID remain undiagnosed, and it is hypothesized that a significant number of ARID-related genes have yet to be discovered, particularly in cases of rare or undiagnosed intellectual disabilities. The total number of ARID genes is estimated to be between 2 500 and 3 000, although this figure is subject to a high degree of uncertainty. This uncertainty poses a challenge for the models, as the true number of ARID genes directly influences the predictions of disease prevalence and the mutation burden within populations [106, 163].

The difficulty in accurately estimating the de novo mutation rate and the number of ARID genes underscores the importance of sensitivity analysis in the models. By varying these parameters within plausible ranges, the models can provide a more comprehensive understanding of how these factors influence the spread and maintenance of recessive alleles in a population.

1.2 Mathematical models in evolutionary biology

This section presents an overview of the most commonly used mathematical models representing population evolution. Each group of models is tailored to capture different aspects of nature more effectively and is better suited for specific research questions. It places the models considered in this work within a broader context and provides readers with an opportunity to familiarize themselves with the literature on various models. The following list, although not exhaustive, is guided by the lecture notes of Bovier and Kraut [25], which we highly recommend as a starting point for further investigation into the topics of stochastic individual based models and scaling limits in this framework.

1.2.1 Population dynamics

As the name suggests, population dynamics models focus on the population as a whole and concentrate on the environmental factors that influence population growth or decline within a given setting. A fundamental initial model for population growth in a limited environment dates back to Thomas Malthus [150]. He posited that populations would grow exponentially in an unlimited environment, with this growth being constrained only by limited resources such as food or space. This theory now known as the Malthusian growth model, which can be described mathematically as a simple, deterministic differential equation that describes the population size $n(t)$ over time, namely

$$\frac{\mathrm{d}}{\mathrm{d}t}n(t) = rn(t) - cn(t)^2.$$

Here r denotes the exponential growth rate of the unrestrained population, which for example can be interpreted as the difference between birth and death rates, and $c > 0$ represents the competitive pressure for resources within the (monomorphic) population. As long as the initial population size $n(0)$ is positive and the growth rate r is also positive, the solution to the differential equation converges to a stable equilibrium $\frac{r}{c}$. However, if the growth rate is not positive $r \leq 0$, the population will eventually die out, and $n(t)$ will converge to zero. This equation can also be extended to a system of interacting populations.

$$\frac{\mathrm{d}}{\mathrm{d}t}n_i(t) = n_i(t)\left(r_i - \sum_{j=1}^{k} c_{ij}n_j(t)\right), \quad i = 1, \ldots, k$$

These types of equations are called competitive Lotka-Volterra equations and go back to Alfred Lotka [148] and Vito Volterra [204]. Here, the coefficients c_{ij} represent the interaction between subpopulation or species i and j. Notably, these interactions can have either positive or negative effects. A particularly well-known example of such models is the Lotka-Volterra equations, which describe predator-prey relationships. In this system, there are $k = 2$ subpopulations: the prey population, which has a size of $n_1(t)$, and the predator population, which has a size of $n_2(t)$ at time $t \geq 0$. The prey population grows at a rate of $r_1 > 0$ and declines at a rate of $c_{12} > 0$ in response to encounters with predators. Conversely, the

predator population increases at a rate of $c_{21} > 0$ in response to encounters with prey and decreases at a rate of $r_2 > 0$.

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}n_1(t) &= \quad r_1 n_1(t) - c_{12} n_1(t) n_2(t) \\ \frac{\mathrm{d}}{\mathrm{d}t}n_2(t) &= -r_2 n_2(t) + c_{21} n_1(t) n_2(t)\end{aligned}$$

This model is particularly illustrative because, with the right choice of parameters, it produces periodically oscillating solutions.

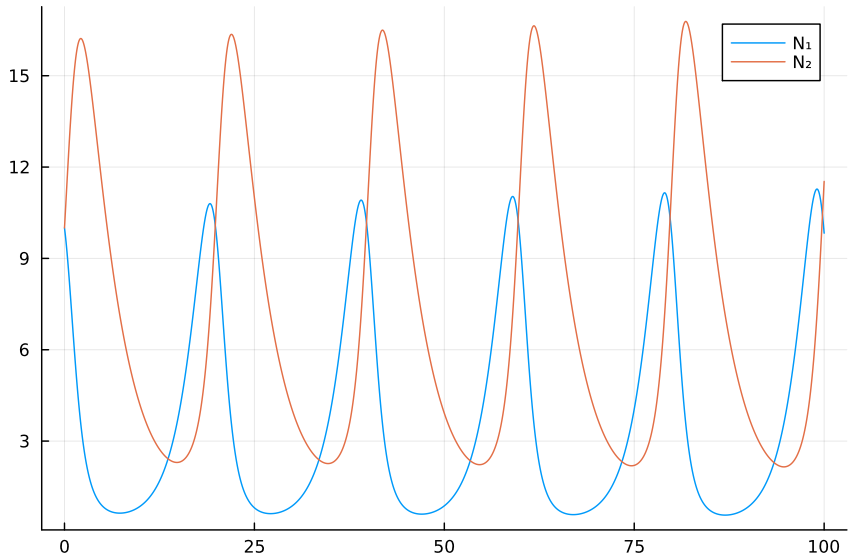

Figure 1.2: Predator-pray system. Example of a periodic solution of the two dimensional Lotka-Volterra system with $r_1 = 0.5, r_2 = 0.25, c_{12} = 0.07, c_{21} = 0.07$.

While the three-dimensional competitive Lotka-Volterra system is well analysed [219], the analysis becomes significantly more challenging as the number of dimensions increase. Indeed, for $k \geq 5$, the system is found to exhibit arbitrary complex long-term behaviour.[191]. The challenge of analysing high-dimensional systems of differential equations will reappear in Chapter 3. For a deeper insight into the field of population dynamics, we recommend the reading of Hofbauer and Sigmund [105].

1.2.2 Population genetics

In contrast to population dynamics, which focus more on the ecology, environmental influences are mostly omitted in models of population genetics. Here, the focus is primarily heredity and on changes in allele frequencies over time. However, as in the previous section, the emphasis is on the entire population rather than on individual organisms. The founders of this mathematical subfield and its initial models were Ronald Aymler Fisher [68], John Burdon Sanderson Haldane [97], and Sewall Green Wright [213], who mathematically combined and framed Mendelian inheritance and Darwinian evolution in the 1920s (see Section 1.1.2).

One of the most well-known and widely used models in population genetics is the Wright-Fisher model. In its simplest, neutral form, the model considers a population with a finite, constant number N of individuals. These individuals are characterized by a single gene with two alleles a and A. Time is measured in discrete units, or generations, where each generation replaces the previous one without overlap. In generation $t+1$, each individual selects an individual uniformly at random from generation t and inherits its trait. The process of allele frequencies of one of the two alleles $N_a(t)$ can then be described as a discrete time Markov chain on $0, \dots, N$ with transition rates.

$$\mathbb{P}\left(N_a(t+1) = n | N_a(t) = m\right) = \binom{N}{n} \left(\frac{m}{N}\right)^n \left(1 - \frac{m}{N}\right)^{N-n}$$

Hence the gene frequencies evolve due to random fluctuations within the binomial resampling of each consecutive generation. One of the two alleles a or A will eventually become fixed due to natural fluctuations, while the other gets lost, representing a fixed point of the system. This process of changes in allele frequencies that originate from random resampling is known as *genetic drift*. Since this is a neutral model, the allele frequency process $N_a(t)$ is a martingale, which is in line with a well known concept in population genetics, the Hardy-Weinberg theorem. This fundamental theorem states, that in the absence of disturbances such as mutations, selection, migration or other factors the genetic variation within an infinitely large population is conserved, hence the expected allele frequencies stay constant over time [101, 208].

In the Wright-Fisher model, as we rescale time by tN and let the population size N approach infinity, the model converges to the Wright-Fisher diffusion process $\left(X(t)\right)_{t\geq 0}$. It can be described as the solution of the stochastic differential equation.

$$\frac{\mathrm{d}}{\mathrm{d}t} X(t) = \sqrt{X(t)(1 - X(t))}\mathrm{d}B(t)$$

where B is a standard Brownian motion. This stochastic process, was first introduced by Motoo Kimura [115] and later formalized as a limit by Ethier and Norman [65]. This process describes the continuous limit of allele frequencies over time, capturing the evolutionary dynamics of allele frequencies in a large population.

The Wright-Fisher model can be extended to include additional features such as migration, mutation, or selection. It is important to note that in this context, selection impacts the genotype directly, without involving interactions between individuals or competition with the environment.

One classic example of an extended model is John Haigh's model [96] to quantify the effect of Muller's Ratchet, which describes the inevitable accumulation of deleterious mutations [161]. In his model, individuals are characterized by the number of deleterious mutations they carry, and those with fewer mutations are preferred when passing on their genes. Mutations occur randomly and uniformly at a constant rate μ. Thus, the relative fitness of an individual with k deleterious mutations is $(1-s)^k$, which is exponentially dependent on the number of mutations it carries. Here $s > 0$ is the selection coefficient, which models the strength of each individual mutation. Hence if $N_k(t)$ is the number of individuals in generation t with exactly k deleterious mutations, the state of the population in generation t is described by

the infinite dimensional vector $\mathbf{N}(t) = (N_0(t), N_1(t), \dots) \in \mathbb{N}^\infty$. The transition rates are given by multinomial sampling, namely

$$\mathbb{P}\left(\mathbf{N}(t+1) = \mathbf{n} | \mathbf{N}(t) = \mathbf{m}\right) = \frac{N!}{\prod_{k=0}^{\infty} n_k!} \prod_{k=0}^{\infty} p_k(t)^{n_k}$$

where $\mathbf{n}, \mathbf{m} \in \mathbb{N}^\infty$ with $\sum_{k=0}^{\infty} n_k = \sum_{k=0}^{\infty} m_k = N$ and with probabilities

$$p_k(t) = \sum_{j=0}^{k} \frac{m_{k-j}(1-s)^{k-j} e^{-\mu} \frac{\mu^j}{j!}}{\sum_{i=0}^{\infty} m_i (1-s)^i}$$

Note that the exponential term in the probabilities of the multinomial sampling come from the assumption that mutations are rare and hence the actual number of new mutations has a Poisson distribution with mean μ [117]. You can find more on the findings of Haigh and other mathematical frameworks to understand Muller's ratchet in Chapter 1.6.2.

Furthermore, there are additional adaptations of the Wright-Fisher model that address various aspects of population dynamics. Here, we briefly outline a few notable ones:

The Moran model This adaptation introduces a continuous-time version of the model, allowing for overlapping generations. In the Moran model, birth and death events occur at exponentially distributed times. Rather than sampling and replacing the entire population at each event, only one individual is replaced, thereby maintaining a constant population size throughout. This model provides a more realistic representation of certain population dynamics compared to discrete generation models [160].

The Canning model This model, also operating in discrete generations, adjusts the number of offspring per individual, assuming that the number of offspring per parent is exchangeable. The Canning model introduces more flexibility in modelling varying reproductive outputs and is useful for studying different reproductive scenarios within a discrete-time framework [34, 35].

The Fleming-Viot process Similar to the Wright-Fisher diffusion, the Fleming-Viot process arises as the infinite population limit of the Moran model. It is a continuous-time process that generalizes the Wright-Fisher model to include additional features such as mutation and selection, providing a rich framework for studying allele frequency dynamics in large populations [71].

The F-KPP equation The spread of a beneficial allele in a spatially structured population is described by the Fisher-KPP (Kolmogorov-Petrovsky-Piskunov) equation. This equation, highly popular in mathematical circles, models the spatial propagation of advantageous traits and has garnered significant interest due to its applications in various fields of population genetics and theoretical ecology [70, 123].

A highly valuable feature for mathematical analysis of the Wright-Fisher model is the ability to trace the genealogy of individuals and project it backward in time. This capability enables the analysis of the process in reverse - known as the coalescent process - and explores the duality between forwards and backwards in time processes. This approach not only addresses questions about the past, such as identifying the most recent common ancestor, but also aids in interpreting genetic data. One of the pioneers in this field was John Kingman, who provided a detailed description of the ancestry of the Wright-Fisher diffusion now known as Kingman's coalescent [118]. His work laid the foundation for modern analyses, which employ tools such as ancestral selection graphs and recombination graphs [132, 140]. These contemporary methods offer powerful frameworks for understanding genetic variation and evolutionary processes by modelling the genetic history and relationships within populations.

For interested readers, we particularly recommend the works of Evans [66] and Etheridge [63] for a comprehensive exploration of these topics, as well as Crow and Kimura for detailed insights into neutral mutations and the effects of genetic drift [52].

1.2.3 Adaptive dynamics

While population genetics can explain how certain traits gain or lose frequency within a population, it struggles to model the emergence of new species. This is where adaptive dynamics becomes crucial. Adaptive dynamics builds upon the interactions between populations observed in population dynamics and combines these with the principles of inheritance and mutation from population genetics.

Adaptive dynamics addresses the interactions between individuals and their environment, relaxing the assumption of a constant population size and a fixed fitness landscape. Instead, it considers that the fitness of an individual depends on the entire state of the population, allowing for the co-evolution of the environment with the population. This approach incorporates density-dependent selection, which considers how population density affects individual fitness.

Exciting questions in adaptive dynamics often revolve around the evolutionary impact of stochastic effects such as mutations. To rigorously analyse these processes, a crucial assumption is made: the evolutionary timescale is separated from the ecological timescale. This means that mutations occur so rarely that beneficial mutants can become fixed in the population before further mutations arise. However, this assumption may be overly simplistic for some real-world cases, as Metz, one of the founding fathers of adaptive dynamics, highlighted in his essay "Adaptive Dynamics" [158].

Mathematically, these models are rigorously formalized as individual-based Markov processes. Moreover, it is possible to derive a pathwise representation of the system in terms of Poisson point measures [72]. A primary focus is the analysis of limiting processes in large populations with rare mutations. If the limiting process is monomorphic, it is termed the Trait Substitution Sequence (TSS) [38], where the process jumps between consecutive fitter traits. Alternatively, if the process is polymorphic, it is called the Polymorphic Evolution Sequence (PES) [40].

Another interesting aspect arises in a continuous trait space, where, in addition to large populations and rare mutations, a small effect of mutations is considered. In this case, the limit is described by the canonical equation of adaptive dynamics (CEAD) [58, 39, 11].

Since most models discussed in this thesis are individual-based models of adaptive dynamics, we will discuss them in detail in Section 1.3. However, our focus is not on scaling limits in large populations but rather on stochastic effects in finite populations. For readers interested in scaling limits and the detailed mathematical framework underlying adaptive dynamics, we recommend the comprehensive book by Méléard [156] or the lecture notes by Bovier and Kraut [25].

1.2.4 Diploid models

The majority of models in adaptive dynamics consider haploid populations with clonal reproduction. While this theory effectively explains the emergence of new species, it falls short in explaining the effects of genetic variability and diversity that arises from processes involved in mating.

One key process is gametogenesis, where genetic material from a diploid cell is mixed through recombination and then divided into haploid germ cells. Another process is the fusion of two gametes to form a new zygote, merging the genetic material from two parents into a new offspring.

Mathematical models for diploid populations are well-established and extensively studied in population genetics [52, 164, 66, 31]. The first diploid models of adaptive dynamics, which account for fluctuating population sizes, were proposed by Kisdi [121]. Collet, Méléard and Metz later demonstrated the convergence to the Trait Substitution Sequence (TSS) under certain conditions in the diploid case [49]. Bovier, Neukirch and Coquille then showed within the framework of adaptive dynamics that diploid populations exhibit greater diversity [26, 168]. Specifically, they proved that suboptimal traits could persist longer as heterozygotes in the population, and the extinction of such traits is slower compared to the replacement of disadvantageous traits in haploid populations.

This dynamic, particularly in combination with mutation and an ever-changing environment, can lead to a more rapid adaptability of diploid populations.

The greatest challenges in modelling diploid populations are as follows. First, there is the need to distinguish between genotype and phenotype. In haploid populations within adaptive dynamics, mutations directly affect the phenotype, thus immediately altering traits that influence fitness. However, the presence of heterozygous genotypes for each trait now necessitates a clear distinction between genotype and phenotype.

Second, mating now involves the selection of two individuals. Consequently, not only does the fertility rate of a single individual play a role, but also the rate at which another individual is chosen for mating. These factors increase the complexity of the models and make mathematical analysis more challenging.

1.3 The core model of this thesis

In this section, we introduce the foundational model that will serve as the basis for the mathematical models discussed in detail later in this thesis. This model captures the essential dynamics we are interested in and forms the starting point for our exploration into more complex scenarios.

1.3.1 A diploid individual based model of adaptive dynamics

We start with the model of adaptive dynamics of Mendelian diploids studied by P. Collet, S. Méléard, J. Metz et al. [49]. The major adaptation we make is a finite, but high dimensional genotype space $\mathcal{X} \subset \mathbb{R}^N$ and a more general approach on the recombination and propagation mechanism of genotypes during a mating of individuals. A diploid individual is characterized by its genotype $\mathbf{x} = (x_1, x_2) \in \mathcal{X}^2$. In the following, we introduce the demographic parameters that encode the biology of the model. We assume that these parameters depend on the allele configuration through the phenotype. As the dependence of the phenotype on the genotype is assumed to be symmetrical, all coefficient functions defined are also assumed to be symmetric in the allele configuration.

(i) $b(x_1, x_2) \in \mathbb{R}_+$: the per birth rate of an individual with genotype (x_1, x_2). Furthermore, an individual with genotype (x_1, x_2) will be selected as a mating partner with probabilities proportional to $b(x_1, x_2)$.

(ii) $d(x_1, x_2) \in \mathbb{R}_+$: the intrinsic death rate of an individual with genotype (x_1, x_2).

(iii) $c(x_1, x_2, y_1, y_2) \in \mathbb{R}_+$: the competition pressure from an individual with genotype (y_1, y_2) exerted onto an individual with genotype (x_1, x_2).

(iv) $m(x_1, x_2, y_1, y_2, z_1, z_2) \in [0, 1]$: the mating and mutation measure gives the probability that the mating of an individual with genotype (x_1, x_2) with an individual with genotype (y_1, y_2) produces an offspring with genotype (z_1, z_2). It is assumed to satisfy

   (a) for each $\mathbf{x}, \mathbf{y} \in \mathcal{X}^2$ $m(\mathbf{x}, \mathbf{y}, \cdot)$ is a probability kernel on $\mathcal{X}^2$ and for each $\mathbf{z} \in \mathcal{X}^2$ the function $(\mathbf{x}, \mathbf{y}) \to m(\mathbf{x}, \mathbf{y}, \mathbf{z})$ is measurable.

   (b) for every $(x_1, x_2), (y_1, y_2), (z_1, z_2) \in \mathcal{X}^2$ the following symmetry properties

$$
\begin{aligned}
m(x_1, x_2, y_1, y_2, z_1, z_2) &= m(x_2, x_1, y_1, y_2, z_1, z_2) \\
m(x_1, x_2, y_1, y_2, z_1, z_2) &= m(x_1, x_2, y_2, y_1, z_1, z_2) \\
m(x_1, x_2, y_1, y_2, z_1, z_2) &= m(y_1, y_2, x_1, x_2, z_2, z_1)
\end{aligned}
$$

   The first two properties correspond to the fact that we do not want to make a difference between the two genotypes of an individual. Both are equally present in the production of the offspring genotype. The second property yields that the mating of two individuals has the same probabilities of producing a given pair of genotypes regardless the order of the mating.

For simplicity we ignore the existence of sexes within the population. Hence an individual chooses a mate with probabilities proportional to the birthrate of the partner. In Chapter 3 and Section 2.6.1, we challenge the assumption of panmixia and explore scenarios involving non-random mating, where individuals show a preference for mating with partners who exhibit certain traits. These traits, although not directly affecting birthrate, influence mate choice and can have significant implications for the genetic structure and evolution of the population.

At any point in time $t \geq 0$ we consider a finite number N_t of individuals. Denote their genotypes as $(x_1^1, x_2^1), \ldots, (x_1^{N_t}, x_2^{N_t}) \in \mathcal{X}^2$. The population state at time $t \geq 0$ is described by the point measure on $\mathcal{X}^2$

$$\nu_t = \sum_{i=1}^{N_t} \delta_{(x_1^i, x_2^i)(t)}$$

where $\delta_{(x_1,x_2)}$ is the Dirac measure at $(x_1, x_2) \in \mathcal{X}^2$. Let $\langle \nu, f \rangle$ denote the integral of a measurable function f with respect to the measure ν. Then $\langle \nu_t, 1 \rangle = N_t$ and for any $(x_1, x_2) \in \mathcal{X}^2$, the non-negative number $\langle \nu_t, \mathbb{1}_{\{(x_1,x_2)\}} \rangle$ is called the density of genotype (x_1, x_2) at time t. In an abuse of notation we define

$$\langle \nu_t, \mathbb{1}_x \rangle \coloneqq \langle \nu_t(x, dy), 1 \rangle + \langle \nu_t(dy, x), 1 \rangle$$

to be the density of the haplotype $x \in \mathcal{X}$ at time t. Let $\mathcal{M}\left(\mathcal{X}^2\right)$ denote the set of finite, nonnegative point measures on $\mathcal{X}^2$, equipped with the weak topology,

$$\mathcal{M}\left(\mathcal{X}^2\right) \coloneqq \left\{ \sum_{i=1}^{n} \delta_{(x_1^i, x_2^i)} \colon n \in \mathbb{N}_0, (x_1^1, x_2^1), \ldots, (x_1^n, x_2^n) \in \mathcal{X}^2 \right\}$$

An individual with genotype (x_1, x_2) in the population ν_t reproduces with an individual with genotype (y_1, y_2) at a rate $b(x_1, x_2)\frac{b(y_1,y_2)}{\langle \nu_t, b \rangle}$. The genotype of the offspring is chosen according to the mutation and mating measure $m(x_1, x_2, y_1, y_2, dz_1, dz_2)$. An individual with genotype (x_1, x_2) in the population ν_t dies at rate

$$d(x_1, x_2) + \langle \nu_t, c(x_1, x_2, dy_1, dy_2) \rangle$$

The population process $(\nu_t)_{t \geq 0}$ is defined as a $\mathcal{M}(\mathcal{X}^2)$-valued Markov process with the dynamics described above. These are encoded in the infinitesimal generator $\mathcal{L}$ of the process, which is defined for any bounded measurable function $f : \mathcal{M}(\mathcal{X}^2) \to \mathbb{R}$ and for all $\nu \in \mathcal{M}(\mathcal{X})$, by

$$\begin{aligned}(\mathcal{L}f)(\nu) = &\int_{\mathcal{X}^2} b(\mathbf{x}) \int_{\mathcal{X}^2} \frac{b(\mathbf{y})}{\langle \nu, b \rangle} \int_{\mathcal{X}^2} \left(f(\nu + \delta_{\mathbf{z}}) - f(\nu) \right) m(\mathbf{x}, \mathbf{y}, d\mathbf{z}) \nu(d\mathbf{y}) \nu(d\mathbf{x}) \\ &+ \int_{\mathcal{X}^2} \left(d(\mathbf{x}) + \int_{\mathcal{X}^2} c(\mathbf{x}, \mathbf{y}) \nu(d\mathbf{y}) \right) \left(f(\nu - \delta_{\mathbf{x}}) - f(\nu) \right) \nu(d\mathbf{x})\end{aligned}$$

The first term describes the mating and birth event. The second term describes the death of an individual. We ignore the unnatural fact that an individual can choose itself as a partner to mate as the probability of that event will become negligible as the population size increases.

Remark. Since we assume the model parameters b, d, c take finite, non-negative values, and the trait space $\mathcal{X}^2$ is finite we immediately get the existence and uniqueness of the process. Since if the population is of finite size n and in the state $\nu = \sum_{i=1}^{n} \delta_{\mathbf{x_i}}$ the total event rate is

$$R(\nu) = \sum_{i=1}^{n} b(\mathbf{x}) + d(\mathbf{x}) + \int_{\mathcal{X}^2} c(\mathbf{x}, \mathbf{y})\nu(d\mathbf{y}) \leq n \left(\max_{\mathbf{x} \in \mathcal{X}^2} \{b(\mathbf{x}) + d(\mathbf{x})\} \right) + n^2 \max_{\mathbf{x}, \mathbf{y} \in \mathcal{X}^2} c(x, y) < \infty$$

bounded from above as long as the population size is finite.

We see that this is true on finite time intervals as long as we start in a possibly random population with finite mean. Moreover we have

Lemma 1.1. Assume that there exist $\bar{b}, \bar{d}, \bar{c} < \infty$ such that for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}^2$

$$0 \leq b(\mathbf{x}) \leq \bar{b}, \quad 0 \leq d(\mathbf{x}) \leq \bar{d} \quad \text{and} \quad 0 \leq c(\mathbf{x}, \mathbf{y})$$

and that there exists $\underline{c} > 0$ such that, for all $\mathbf{x} \in \mathcal{X}^2$, $\underline{c} \leq c(x, x)$. Moreover, assume that $m(\mathbf{x}, \mathbf{y}, \cdot)$ is uniformly bounded for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}^2$ and that $\mathbb{E}\left[\langle \nu_0, 1 \rangle^2\right] < \infty$. Then, for any $T < \infty$,

$$\mathbb{E}\left[\sup_{t \leq T} \langle \nu_t, 1 \rangle^2\right] < \infty.$$

The proof can be adapted from [72] [Theorem 3.1 (ii)] to a diploid population.

1.3.1.1 The law of large numbers

While the random effects of the dynamics are also of interest, to gain insight into the process, it is also informative to understand the equilibrium, deterministic behaviour of the population. The aim of this section is therefore to examine the dynamics of the process as the population approaches infinity. It will be demonstrated that this results in a system of deterministic, ordinary differential equations (ODE) that can be solved in certain instances. However, due to the high-dimensional nature of the system, analysing these ODEs is typically very challenging.

In order to obtain a law of large numbers type result, it is necessary to apply the correct rescaling of the population. This is achieved by accelerating the birth and death events by a factor of K, while simultaneously scaling down the step size by a factor of $1/K$. Furthermore, we replace the function $c(\mathbf{x}, \mathbf{y})$ with the scaled function $\mathrm{c}\frac{1}{K}c(\mathbf{x}, \mathbf{y})$. Thus, as the value of K increases, the frequency of interaction between individuals decreases, while the population size simultaneously grows. The factor K is referred to as the *carrying capacity*, as it places a limit on the population size in the order of K. Consequently, it can be viewed as the quantity of resources accessible to the population within its environment. More precisely we consider the rescaled point measure

$$\nu_t^K = \frac{1}{K} \sum_{i=1}^{N_t} \delta_{(x_1^i, x_2^i)}$$

on the state space

$$\mathcal{M}^K(\mathcal{X}) \coloneqq \left\{ \frac{1}{K}\sum_{i=1}^{n} \delta_{(x_1^i, x_2^i)} \colon n \geq 0, (x_1^1, x_2^1), \ldots, (x_1^n, x_2^n) \in \mathcal{X}^2 \right\}$$

The rescaled generator is then given by

$$\begin{aligned}\left(\mathcal{L}^K f\right)\left(\nu^K\right) = &\int_{\mathcal{X}^2} b(\mathbf{x}) \int_{\mathcal{X}^2} \frac{b(\mathbf{y})}{\langle \nu, b\rangle} \int_{\mathcal{X}^2} \left(f\left(\nu + \frac{1}{K}\delta_{\mathbf{z}}\right) - f(\nu)\right) m(\mathbf{x}, \mathbf{y}, d\mathbf{z}) \nu(d\mathbf{y}) \nu(d\mathbf{x}) \\ &+ \int_{\mathcal{X}^2} \left(d(\mathbf{x}) + \frac{1}{K}\int_{\mathcal{X}^2} c(\mathbf{x}, \mathbf{y}) \nu(d\mathbf{y}) \right) \left(f\left(\nu - \frac{1}{K}\delta_{\mathbf{x}}\right) - f(\nu)\right) \nu(d\mathbf{x})\end{aligned}$$

The following theorem was proven by N. Fournier and S. S. Méléard in the case of a haploid, clonal population, but can be adapted easily to this setup.

Theorem 1.2. *Assume that the initial conditions ν_0^K converge, as $K \to \infty$, in law and for the vague topology on $\mathcal{M}(\mathcal{X}^2)$ to some deterministic finite measure $\xi_0 \in \mathcal{M}(\mathcal{X}^2)$ and that $\sup_K \mathbb{E}\left[\langle \nu_0^K, 1\rangle^2\right] < \infty$. Then for all $T > 0$, the sequence ν^K converges, as $K \to \infty$, in law, in $\mathbb{D}([0,T], \mathcal{M}(\mathcal{X}^2))$, to a deterministic continuous function $\xi \in C([0,T], \mathcal{M}(\mathcal{X}))$. This measure-valued function ξ is the unique solution, satisfying $\sup_{t\in[0,T]}\langle \xi_t, 1\rangle < \infty$, of the integro-differential equation written in its weak form: for all bounded and measurable functions, $h : \mathcal{X}^2 \to \mathbb{R}$,*

$$\begin{aligned}&\int_{\mathcal{X}^2} h(\mathbf{x})\xi_t(d\mathbf{x}) - \int_{\mathcal{X}^2} h(\mathbf{x})\xi_0(d\mathbf{x}) \\ &\qquad = \int_0^t \int_{\mathcal{X}^2} b(\mathbf{x}) \int_{\mathcal{X}^2} \frac{b(\mathbf{y})}{\langle \xi_s, b\rangle} \left(\int_{\mathcal{X}^2} h(\mathbf{z}) m(\mathbf{x}, \mathbf{y}, d\mathbf{z}) \right) \xi_s(d\mathbf{y}) \xi_s(d\mathbf{x}) ds \\ &\qquad - \int_0^t \int_{X^2} h(\mathbf{x}) \left(d(\mathbf{x}) + \int_{\mathcal{X}^2} c(\mathbf{x}, \mathbf{y}) \xi_s(d\mathbf{y}) \right) \xi_s(\mathbf{x}) ds\end{aligned}$$

Remark. Assume the finite trait space is countable and of size $|\mathcal{X}| = n$ and we have a numbering on that space, such that $\mathcal{X} = \{x_1, \ldots, x_n\}$. Then we can realize the process as a Markov process with state space $\mathbb{R}_+^{n\times n}$ with generator acting on functions $f : \mathbb{R}_+^{n\times n} \to \mathbb{R}$. Therefore, let $z = (z_{ij})_{1\leq i,j\leq n}$ be a non-negative $n \times n$ matrix. Then the generator is defined

as

$$
\begin{aligned}
(Lf)(z) = \sum_{i_1,i_2=1}^{n} z_{i_1 i_2} b(x_{i_1}, x_{i_2}) \sum_{j_1,j_2=1}^{n} \frac{z_{j_1 j_2} b(x_{j_1}, x_{j_2})}{\sum_{l_1,l_2=1}^{n} z_{l_1 l_2} b(x_{l_1}, x_{l_2})} \times \\
\times \sum_{k_1,k_2=1}^{n} m(x_{i_1}, x_{i_2}, x_{j_1}, x_{j_2}, x_{k_1}, x_{k_2}) \left(f(z + e_{k_1 k_2}) - f(z)\right) \\
+ \sum_{i_1,i_2=1}^{n} z_{i_1 i_2} \left(d(x_{i_1}, x_{i_2}) + \sum_{j_1,j_2=1}^{n} z_{j_1 j_2} c(x_{i_1}, x_{i_2}, x_{j_1}, x_{j_2}) \right) \left(f(z - e_{i_1 i_2}) - f(z)\right)
\end{aligned}
$$

where $e_{ij} \in \mathbb{R}^{n \times n}$ is the $n \times n$ matrix with zero everywhere besides at the position (i, j) where it takes the value one for some $1 \leq i, j \leq n$. In that case the test functions for the limiting integro-differential equation from the law of large numbers 1.2 can be limited to the indicator functions on elements on the trait space $\mathcal{X}^2 = \{x_1, \ldots, x_n\} \times \{x_1, \ldots, x_n\}$. If we set $\langle \xi_t, \mathbb{1}_{(x_i, x_j)} \rangle := z_{ij}(t)$ we get the n^2 ordinary differential equations

$$
\frac{d}{dt} z_{i_1 i_2}(t) = \sum_{j_1,j_2=1}^{n} z_{j_1 j_2}(t) b(x_{j_1}, x_{j_2}) \sum_{k_1,k_2=1}^{n} \frac{z_{k_1 k_2}(t) b(x_{k_1}, x_{k_2})}{\sum_{l_1,l_2=1}^{n} z_{l_1 l_2}(t) b(x_{l_1}, x_{l_2})} m(x_{j_1}, x_{j_2}, x_{k_1}, x_{k_2}, x_{i_1}, x_{i_2}) \tag{1.1}
$$

$$
- z_{i_1 i_2}(t) \left(d\left(x_{i_1}, x_{i_2}\right) + \sum_{j_1,j_2=1}^{n} z_{j_1 j_2}(t) c\left(x_{i_1}, x_{i_2}, x_{j_1}.x_{j_2}\right) \right) \tag{1.2}
$$

for $1 \leq i, j \leq n$.

Solving this system of ODEs (1.1) is far from straightforward, and depending on the parameters (in particular on the size of the trait space n), it can often be impossible to find analytic solutions. In the next step, we will adapt this base model to our specific context, focusing on the spread of autosomal recessive diseases. By incorporating the relevant biological and genetic factors into the equations, we aim to analyse the dynamics of how these diseases propagate within a population.

1.3.2 Complete recessive lethal diseases

We study the effect of a group of recessive genetic diseases that share the same structure. These diseases arise from genetic changes and are inherited from one generation to the next. As long as the degeneration of the genetic material is present on only one set of chromosomes, it has no effect on the fitness of the carrier. It is only when the disease is present in a homozygous state that it manifests. If one of the diseases is expressed, the individual loses the ability to reproduce, but the life expectancy of an affected individual remains unchanged. We consider N genes or gene segments in which these diseases can occur. Mutations in these gene segments trigger the diseases. We ignore both neutral or beneficial mutations and reversions that restore the genetic material. Each mutation potentially has a negative effect on the individual. Furthermore, multiple mutations can occur at different

positions within the same gene segment on the same chromosome without any additional effect. Therefore, it is only important whether a gene segment is mutated and not how many mutations it carries. This results in the following selection of parameters:

The trait space is $\mathcal{X} = \{0,1\}^N$ hence every individual is characterized by a $2 \times N$ matrix with values in $\{0,1\}$. Here zero represents the wild type and a one indicates that (at least one) mutation is present. Define the set $\mathcal{D}_N \subset \mathcal{X}^2$ as

$$\mathcal{D}_N := \left\{(x,y) \in \mathcal{X}^2 \colon \exists\, 1 \leq i \leq N \text{ such that } x_i = 1 = y_i\right\}.$$

Then for $x, y, z, w \in \mathcal{X}$ the birth, death and competition rates are given by

$$b(x,y) := \bar{b}\mathbb{1}_{\mathcal{X}^2 \setminus \mathcal{D}_N}(x,y) \quad \text{and} \quad d(x,y) := \bar{d} \quad \text{and} \quad c(x,y,z,w) := \bar{c}$$

for some finite $\bar{b}, \bar{d}, \bar{c} \in \mathbb{R}_+$. Moreover define $\mu > 0$ to be the mutation rate per gamete. Since usually the number of loci N is big and the mutation rate μ is small we assume that the number of mutation per birth is Poisson distributed with mean 2μ. The mutation location then is uniform distributed among all $2N$ possible positions. In Chapter 2, we introduce an alternative assumption, namely that the genes themselves vary in size and that the mutations occur in proportion to the size of the gene. We assume that during gamete formation, each gene is passed on independently, corresponding to the case of a fully recombining genome. This allows for maximum genetic variation and independence of loci. In Chapter 3, however, we relax this assumption and examine in more detail the effects of varying recombination rates. We also explore the scenario where genomes do not recombine at all, and gamete formation only involves segregation, meaning entire segments of the genome are inherited together. This change in recombination dynamics has significant implications for the evolution and spread of genetic traits, particularly recessive diseases, and will be analysed thoroughly. To model meiosis in the independent form we introduce the following function. Let $\tau = (\tau_1, \tau_2, \ldots, \tau_N) \in \{1,2\}^N$ be the choice of each gene, then define for $\mathbf{x} \in \{0,1\}^{2\times N}$

$$\phi_\tau(\mathbf{x}) = \left(x^1_{\tau_1}, \ldots, x^N_{\tau_N}\right)$$

In the general form we then define the mating and mutation probabilities as

$$m(\mathbf{x}, \mathbf{y}, d\mathbf{z}) = \frac{1}{2^{2N}} \sum_{\tau,\tau' \in \{1,2\}^N} \sum_{k=0}^{\infty} \frac{(2\mu)^k}{k!} e^{-2\mu} \frac{1}{Z_k} \sum_{m \in \diamondsuit_k^{2N}} \delta_{\left(\left(\phi_\tau(\mathbf{x}), \phi_{\tau'}(\mathbf{y})\right) + m\right) \wedge 1}(d\mathbf{z})$$

where $\diamondsuit_k^{2N} := \left\{m \in \mathbb{N}_+^{2N} \colon m_1 + \ldots + m_{2N} = k\right\}$ is the set of all lattice vectors in $\mathbb{N}_+^{2N}$ with one norm equal to k, moreover $Z_k = \sum_{j=0}^{2N} \binom{2N}{j} p_j(k)$ is the size of the set $\diamondsuit_k^{2N}$ and where $p_j(k)$ is the number of partitions of k into exactly j parts. For notational reasons define for $x \in \mathbb{R}^{2N}$ and $k \in \mathbb{R}$ the component wise maximum as $x \wedge k := (x_1 \wedge k, \ldots, x_{2N} \wedge k)$.

1.3.3 Prevalence and mutation burden

The state of the population at any given time can be represented as a high-dimensional vector, detailing the precise genetic configuration of every living individual. However, this

level of detail is often not particularly insightful for understanding the broader dynamics of the population. To effectively analyse and observe the stability of the population, we focus primarily on two key statistics: *prevalence* and *mutation burden* (also referred to as mutation load).

Prevalence refers to the proportion of individuals in the population expressing a severe recessive disease. In this context, prevalence is equivalent to incidence rates because individuals either express the disease or do not - there is no concept of infection or cure during their lifetime for such genetic conditions. Prevalence, therefore, is a measure of the phenotype and can be observed with greater accuracy in real-world studies, as it is based on visible traits.

On the other hand, mutation burden is a measure of the genotype and represents the average number of lethal equivalents per individual. This serves as a risk score, indicating the genetic load of potentially harmful mutations within the population. Unlike prevalence, mutation burden is much harder to measure in nature. It requires advanced genome sequencing techniques, and even then, many rare diseases often remain undetected. Thus, while mutation burden provides critical insight into the genetic health of the population, it presents significant challenges in terms of practical measurement and assessment.

Mathematically speaking, the two statistics, prevalence and mutation burden can be defined as follows: Let $\nu_t \in \mathcal{M}(\mathcal{X}^2)$ be the state of the population at time $t \geq 0$. Recall that $\mathcal{D}_N \subset \mathcal{X}^2$ is the set of all affected configurations. Hence, the prevalence is defined as

$$P(\nu_t) \coloneqq \frac{\langle \nu_t, \mathbb{1}_{\mathcal{D}_N} \rangle}{\langle \nu_t, 1 \rangle}$$

To define the mutation burden of the population, we first define the (absolut) mutation burden of a gamete. Let $x = (x^1, x^2, \ldots, x^N) \in \mathcal{X}$ then the mutation burden of the haploid configuration x is defined as the sum of all its mutations

$$B'(x) = \sum_{n=1}^{N} x^n$$

and the (relative) mutation burden of the whole population is then defined as

$$B(\nu_t) = \frac{1}{\langle \nu_t, 1 \rangle} \int_{\mathcal{X}^2} B'(x_1) + B'(x_2) \nu_t(\mathrm{d}\mathbf{x})$$

Although we will introduce additional statistics throughout this thesis to gain deeper insights into more complex phenomena, the prevalence and mutation burden - alongside the population size - will serve as our primary indicators.

1.4 Stochastic simulation algorithm

This section explores the necessity of stochastic simulations for the analysis of complex and high-dimensional models. Given the intricate nature of the dynamics, traditional mathematical analysis often falls short in providing clear insights. As a result, numerical analysis and stochastic simulations become indispensable tools for exploring these dynamics. One of the central tools we employ for simulating these complex systems is the Gillespie algorithm, also known as the stochastic simulation algorithm (SSA). We undertake a brief excursion into the field of computer science, wherein we elucidate the rationale behind our selection of the Julia programming language as the foundation for the simulation framework outlined in Chapter 4.

1.4.1 Gillespie algorithm

This section provides an overview of the main aspects of the methods used to simulate biological systems. For a more mathematical and comprehensive overview, we refer the reader to the following useful works [83, 174, 104, 91, 211].

The Gillespie algorithm, developed by Daniel T. Gillespie in 1976 [81], was originally created to address the problem of simulating the stochastic behavior of chemically reacting systems at the molecular level. Before its invention, most models used deterministic differential equations to describe reaction kinetics, which failed to capture the inherent randomness present in systems with small numbers of molecules [85]. Gillespie's algorithm introduced a way to model these systems probabilistically, accounting for the random timing and order of reaction events.

Initially applied in chemical physics, the Gillespie algorithm has since found widespread use in various fields, including population genetics, epidemiology, and systems biology. Today, it is employed to simulate complex stochastic processes, such as gene regulation, cell division, evolutionary dynamics, and ecological interactions, providing insights into systems where randomness plays a critical role [154, 155, 14, 47, 6].

The SSA is a Monte Carlo method used to simulate the time evolution of a Markov process. In the context of population genetics, these states could represent different genetic configurations in a population, and the transitions between states correspond to events like birth (that may include mutation or recombination) and or death.

The key steps of the Gillespie algorithm are:

1. **Initialization:** Set the initial state of the system (e.g., the genetic composition of the population).
2. **Reaction selection:** Calculate the propensity (or rate) of each possible reaction/event occurring next.
3. **Time step calculation:** Determine the time until the next event occurs, which is drawn from an exponential distribution based on the total propensity.

4. **Reaction execution:** Select and execute one of the possible reactions/events based on their relative propensities.
5. **Update:** Update the state of the system according to the reaction executed.
6. **Iteration:** Repeat steps 2-5 until the desired simulation time is reached or the system reaches a specified state.

This method is statistically exact, whereby a complete probability distribution would be constructed from an infinite number of simulations, resulting in a representation that is identical to the distribution of the underlying Markov process [80]. While this exactness property is undoubtedly beneficial, it is possible to argue that it is not a crucial factor. While the simulation results may precisely align with the model, the model itself will (of course) not be an exact representation of reality [27]. Furthermore, only an infinite number of simulations would result in an exact representation of the distribution of the model. However, in reality, computational time is a limiting factor.

One of the main disadvantages of Gillespie's algorithm is its computational intensity, particularly due to the need to generate two random numbers at every simulation step. For systems with many reactions or large populations, this results in a high computational cost, as the algorithm proceeds event-by-event, making it impractical for large-scale or long-duration simulations. There is a notable solution to improve the speed is the *τ-leap method*, which accelerates simulations by allowing multiple reactions to occur simultaneously over a small time interval τ, rather than handling one event at a time [82]. This method reduces the number of random number generations by approximating the number of reactions that occur during τ, balancing accuracy with computational efficiency. While the τ-leap method sacrifices some precision compared to Gillespie's exact algorithm, it greatly improves the simulation speed for large and complex systems. However, within the framework implemented for this thesis we do not make use of this technique. Besides the above method a number of algorithmic enhancements have been implemented with the objective of accelerating the processing speed [78]. In addition, a variety of alternative versions of the SSA have been developed [36, 216]. For further insight, two reviews are recommended [83, 84].

The Gillespie algorithm is a relatively simple and straightforward approach that can be readily implemented in a variety of scenarios. It serves as an excellent point of departure for those interested in exploring stochastic simulation. However, there are numerous alternative simulation methods, each with its own set of advantages and disadvantages. For a comprehensive overview of stochastic simulation methods, we refer to the mini-review paper by Székely, which provides a valuable introduction to the field of stochastic simulation algorithms [197].

1.4.2 The Julia programming language

In this section, we provide a brief overview of the advantages and disadvantages of the Julia programming language, focusing on its practical applications rather than delving deeply into the technical intricacies of computer science. Those wishing to gain a deeper understanding

of Julia are encouraged to consult the comprehensive documentation [18] and a selection of published works on the subject [139, 133].

The Julia programming language was first introduced in 2012 by a team of developers led by Jeff Bezanson, Stefan Karpinski, Viral B. Shah, and Alan Edelman [17]. The language was designed from the ground up to address the needs of high-performance numerical computing, combining the ease of use of Python and R with the performance of C and Fortran [30, 131]. Today, Julia is used in various fields, including machine learning, data science, and scientific computing. It has been downloaded over 45 million times as of August 2024 and is gaining popularity among data scientists as a potential "future language" for data science and artificial intelligence [19].

One of the principal advantages of the Julia programming language is its high level of performance. Julia's just-in-time (JIT) compilation allows it to achieve speeds comparable to C, making it particularly effective for CPU-intensive tasks. This performance advantage is crucial in scientific computing where efficiency is paramount. Furthermore, Julia is straightforward to use and learn, making it accessible to beginners and enabling researchers to test hypotheses on the move. Its syntax is clear and concise, resembling mathematical notation, which simplifies the coding process for scientists and engineers. The interactivity of Julia facilitates the development of models and the verification of conjectures. Users can easily build prototypes and modify them interactively using Julia's REPL (Read-Eval-Print Loop) interface or an interactive notebook using Pluto [202]. A distinctive attribute of Julia that facilitates the maintenance of code clarity is its multiple dispatch functionality. In contrast to other languages, Julia's multiple dispatch enables the language to determine the appropriate function to execute when a call is made, based on the types of all arguments, not just the initial one. This feature allows for more flexible and efficient function definitions, which can ultimately result in more maintainable code in scientific applications. Furthermore, Julia offers a comprehensive suite of libraries and tools specifically designed for scientific computing, including advanced packages for linear algebra (in the standard library, [18]), differential equations [182], and data visualization [53]. This ecosystem is continually expanding, making Julia an attractive option for researchers. Although Julia's ecosystem is expanding, it still lags behind Python in terms of the number of available libraries and community support. Moreover, some research groups and projects rely on packages and code available only in other languages. While the decision to transition to Julia is ultimately individual, there is the possibility to include and compile Python code in Julia, thus maintaining familiarity with the frameworks one is used to using [109].

In the context of initiating a novel project or evaluating a hypothesis rapidly, we strongly advise considering Julia as a potential solution [19].

1.4.3 Dense problems

Both the Gillespie algorithm and the Julia ecosystem offer considerable flexibility when simulating stochastic processes. In particular, the JumpProcess.jl Package in the Scientific Machine Learning Software Ecosystem in Julia provides a wide range of simulation methods, including the SSA [218]. The ecosystem is undergoing constant growth, and many problems

can be solved with relative ease using the tools provided therein. We strongly encourage the reader to consult the documentation. However, one limitation is that all possible reactants and interactions between these reactants have to be specified at the beginning of the simulation. In the context of a small number of types this is not the problem as for example in the classical *SIR* model [114], where there are three types of individuals. Suscepteble (S), infected (I) and recovered (R) individuals. With a given rate susceptable individuals get infected when they interact with infected individuals and then recover after some time.

$$\begin{aligned} S + I &\to 2I \\ I &\to R \end{aligned}$$

Here writing down all the possible interactions and types is easy. However, the models we discuss in this paper have up to 4^N possible types, where N can be as large as 1000. It is not feasible to document all interactions between the types in question. The same holds true for models where the trait of an individual is a real number $x \in \mathbb{R}$ (e.g.[57]). In that case - although at any time during the simulation there are only finitely many individuals - the number of potential traits that can appear during the course of a simulation is uncountable. It was thus necessary to devise a variant of Gillespie's algorithm that is more flexible and enables simulation in these dense models. This is achieved by calculating the birth and death rates of each individual at each time interval. These depend on the trait of the individual at that time. However, to ensure the minimum possible computation time, it is necessary to implement the birth and death rates function with great care.

In Section 1.7.3, we provide a more detailed account of our efforts to optimise the runtime of the simulation and to minimise memory usage. In Chapter 4, we present the simulation framework in the context of a number of use cases.

1.5 Non-Random mating and population growth

Previously, the dynamics of mutation load and incidence rates for lethal recessive diseases have been analysed for different demographic models including explosive population growth, but excluding different mating schemes [103]. In this section, we will therefore focus on models for random and consanguineous mating in a growing population. Here we lay the foundation for the research presented in Chapter 2. Part of this section is taken from an unpublished preprint [134], that was the predecessor of the published paper in Chapter 2. Here, we introduce the specific biological concepts that are central to the later discussion. Additionally, we adapt the core model from Chapter 1.3 to suit the particular use case, ensuring that the model accurately reflects the biological mechanisms and dynamics of a growing, non-randomly mating population.

1.5.1 Population size

In population genetics and evolutionary biology, the size of a population is a critical factor that influences the impact of stochastic effects. In finite populations, randomness can have a profound influence on evolutionary dynamics. Random events, such as genetic drift, can cause allele frequencies to fluctuate unpredictably, leading to outcomes that deviate from the predictions of deterministic models[74]. In large populations, these stochastic effects tend to average out, and the population's behaviour closely follows deterministic predictions. The relative magnitude of these fluctuations is inversely proportional to the population size.

Inbreeding occurs when closely related individuals mate (see below), which increases the probability that offspring will inherit identical alleles from both parents. This increased homozygosity can reduce the *effective population size* N_e, a concept that represents the size of an idealized population with the same genetic drift characteristics as the observed population [214, 213]. The effective population size is usually smaller than the actual census size N because factors like inbreeding, unequal sex ratios, and population substructure all reduce genetic diversity [206]. A smaller effective population size again means that genetic drift has a stronger influence, increasing the chances of allele fixation or loss and reducing genetic variability.

The effects of population growth on genetic diversity and evolutionary dynamics are complex and not fully understood. Rapid population growth increases the absolute number of individuals, potentially enhancing genetic diversity and reducing the relative impact of stochastic effects. However, if the growth is accompanied by bottlenecks or founder effects (where a small number of individuals establish a new population), the effective population size may still be small, maintaining strong genetic drift effects.

The human population has experienced exponential growth since the 1800s. This growth is expected to plateau within this century, leading to a transition from exponential growth to a stable or slowly declining population size. This adds another layer of complexity to understanding the dynamics of mutation load and prevalence of lethal recessive diseases [185].

1.5.2 Consanguineous mating

In this section, we explore the effects of different mating schemes, particularly focusing on consanguineous mating or inbreeding, where closely related individuals mate. Most population models assume random mating within a well-mixed population, which often suffices for many applications. However, this assumption can overlook the nuanced effects of mating preferences that are prevalent in various natural and human populations.

In the natural world, there are mating strategies where individuals select their partners based on the resemblance of their phenotypes (assortative mating or dissortative mating). However, this thesis will focus on mating strategies that are kinship-based, where the degree of relatedness between mating individuals plays a crucial role. Consanguineous mating refers to unions between closely related individuals, such as cousins or other family members. In contrast, outbreeding involves mating between unrelated or distantly related individuals.

To quantify the degree of relatedness between individuals, Sewall Wright introduced the *coefficient of relationship* r in 1922 [212]. This coefficient measures the proportion of the genome that two individuals share due to common ancestry. For example a parent and offspring share half of their genomes, so $r = 0.5$. Similarly siblings also share approximately half of their genomes, resulting in $r = 0.5$, whereas first cousins share about one-eighth of their genomes, so $r = 0.125$. A detailed overview of the coefficient of relationship can be found in the consanguinity table 1.6.

These values are theoretical considerations, reflecting the genetic similarity expected based on inheritance patterns. However, it's important to note that all humans share about 99.6%-99.9% of their genome [110], which means that the genetic variation affected by consanguinity is a small, but significant portion of the genome. Consanguineous matings are typically defined as those occurring between individuals who are second cousins or closer, corresponding to a coefficient of relationship $r \geq 0.03125$. This definition highlights the genetic closeness necessary for inbreeding to have notable effects on the offspring's genetic makeup.

Another way to measure consanguinity is through Wright's inbreeding coefficient $f \in [0, 1]$. It quantifies the probability that two alleles at a given locus in an individual are identical by descent (IBD). Two alleles are said to be *identical by state* (IBS) if they are identical in their nucleotide sequence. If they are also copies of a single ancestral allele, hence when both alleles were inherited from a common ancestor without any recombination events breaking the inheritance chain, they are said to be IBD. An individual with two homologous genes that are IBD is called *autozygous*, whereas it is called *allozygous* if the two alleles are from different origin or if a common origins is unknown due to incomplete pedigree information.

Thus, an individual with an inbreeding coefficient f has a probability f that the two genes at a given locus are IBD and a probability $1 - f$ that they are independent. In the case of independent genes, the allele frequencies are given by the Hardy-Weinberg equilibrium (see section 1.2.2, [101, 208]). Thus, for a locus with two alleles a and A with frequencies p and $(1 - p)$ in the population, the genotype frequencies are given by

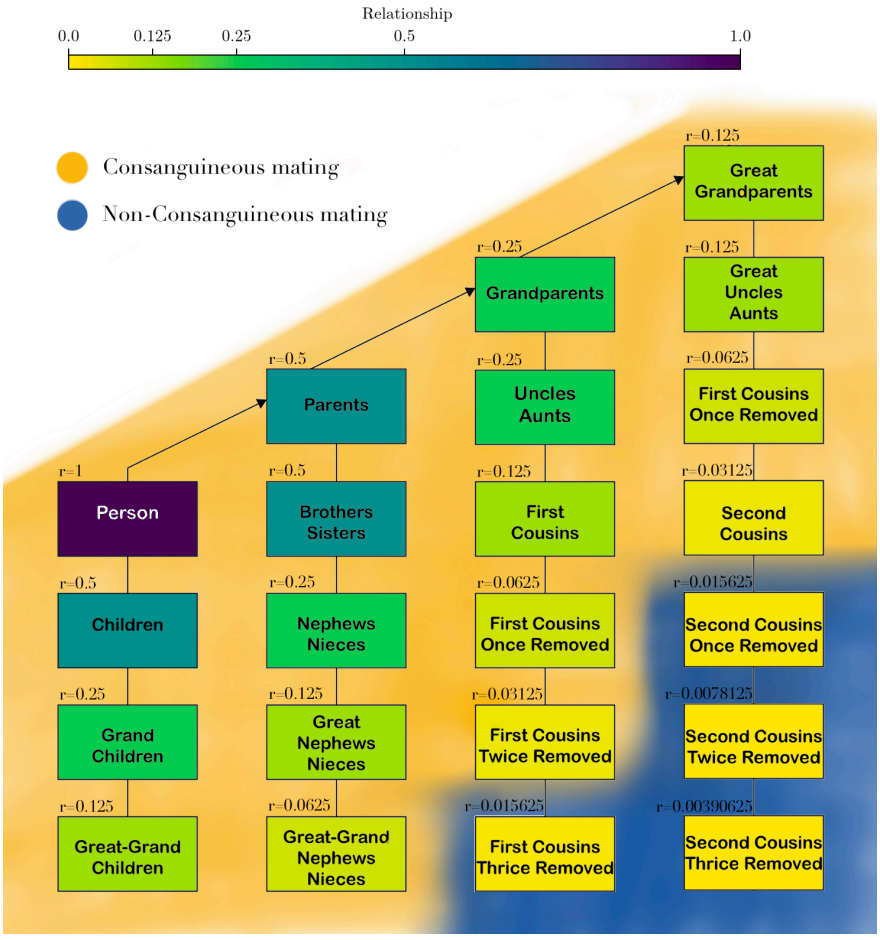

Figure 1.3: Table of consanguinity. Relationships at different levels are given together with the coefficient of correlation r. The orange shaded area is associated with consanguineous mating if two individuals from that area would mate, whereas the blue shaded area is no longer associated with consanguinity.

	allozygous		autozygous
aa	$p^2(1-f)$	$+$	pf
aA	$2p(1-p)(1-f)$		
AA	$(1-p)^2(1-f)$	$+$	$(1-p)f$
$\sum$	$1-f$		f

Unlike the coefficient of relationship, which measures the proportion of genes shared between two individuals due to common ancestry, the inbreeding coefficient measures the likelihood that two alleles in an individual are IBD due to inbreeding. For more insight into the effects of inbreeding and different mating patterns on allele frequencies in the context of population genetics, we recommend the comprehensive book by James Crow and Motoo Kimura [52].

Consanguineous marriages are more common in certain parts of the world. It is estimated that approximately 8.5% of all children worldwide are born from consanguineous matings

[20]. Moreover, around 20% of the global population resides in regions where consanguineous marriages are culturally preferred or socially accepted [159]. This type of union is particularly prevalent in parts of the Middle East, West Asia and North Africa, where such practices are often embedded in cultural or religious traditions [100]. In contrast, there are many countries where the law prohibits marriage or sexual relations between blood relatives [170].

Inbreeding increases the likelihood that offspring will inherit the same allele from both parents, raising the probability of homozygosity for recessive alleles. This can lead to an increase in the expression of recessive genetic disorders, which may otherwise remain hidden in heterozygous carriers. The empirical observation that consanguinity is associated with an increased risk of autosomal recessive disorders, has been made in many countries. Martin, et al. recently showed that the contribution of autosomal recessive developmental disorders is 31% in the current British population if the autozygosity is above 0.02 [153]. Likewise, in the Iranian population it is estimated that offspring from first-cousin unions have a probability for intellectual disabilities that is four times higher than in non-consanguineous partnerships [106, 111, 163]. Although most people probably agree that a lower burden of disease and child mortality is a desirable goal in a society, it is also clear that the occurrence and coexistence of different marriage patterns over many centuries cannot be understood by population genetics alone, especially as demographic, social and economic factors interact in a complex manner [22]. However, since there have been repeated attempts to motivate social conventions by genetic reasoning, we took a closer look at the validity of these arguments.

The European Court of Human Rights case of Stübing v. Germany concerned consanguine siblings who had four children following consensual intercourse, whereupon both siblings were charged with incest [2]. One of the siblings lodged a complaint, arguing that the legislature violated his right to sexual self-determination, his private and family life. The Court found that 24 out of 44 European States reviewed, criminalized consensual sexual acts between adult siblings, and all prohibited siblings from getting married. The German government argued that the law against incest partly aimed to protect against the significantly increased risk of genetic damage among children from an incestuous relationship. Motivating a law on avoiding a higher probability of disease can be viewed as eugenic. As the German Ethics Council opined after the judgment, no convincing argument can be derived from there being a risk of genetic damage, concurring with a statement from the German Society of Human Genetics that "The argument that reproduction needs to be thwarted in couples whose children possess an elevated risk for recessively inherited illnesses is an attack on the reproductive freedom of all" [3, 1]. Furthermore, as our work shows, the argument that there exists an increased risk of genetic damage, requires the definition of a reference population for comparison. However, there is neither agreement about a suitable reference nor an accurate measurement for mutation load.

1.5.3 Model modifications

Building on the core model introduced in Chapter 1.3, we incorporate several additional features to align the model more closely with the biological reality of autosomal recessive diseases in a growing, non-randomly mating population.

In many cases, the exact number of genes that can cause autosomal recessive diseases is not known, and this number can only be estimated approximately [122]. Additionally, these genes vary significantly in size, with some being as short as 500 base pairs (bp) and others extending over 10 000 bp. This variation in gene length affects the mutation probability for each gene. Larger genes, by virtue of having more nucleotides, are more likely to be affected by mutations than smaller genes. To accurately model this, we introduce a gene-specific mutation rate that accounts for gene length. We also consider the possibility of compound heterozygous mutations. In this scenario, even if two mutations occur at different positions within the same gene on each chromosome copy, the disease can still be expressed. By incorporating compound heterozygosity into the model, we account for a broader spectrum of genetic mutations that contribute to disease expression, which is often observed in real-world scenarios.

Furthermore, we assume that the genes are distributed across n_c chromosomes, which recombine freely during meiosis. We can easily adjust the population size in our adaptive dynamics model by spontaneously increasing the carrying capacity K. With more resources, the population grows naturally until it reaches a new equilibrium.

The most important modification we make is to allow for non-random mating. In our case mating preferences should not depend on the genotype of individuals, which is the only characteristic trait in the core model. Therefore, we need to introduce another trait that is independent of genotype, but rather relatedness. It is not feasible to trace the kinship of all living individuals in our model, so we introduce family flags, which do not represent the exact family background, but rather give an idea of the origin of the individual. Each individual is assigned two additional numbers $(f_1, f_2) \in \mathbb{N}^2$ as a trait. When two individuals with family traits mate, the offspring inherits a randomly selected trait from each parent. However, in the event that both parents carry the exact same family flags, the offspring will also inherit the exact same two flags. When choosing a mate, individuals then make their choice according to the combination of the two additional traits.

However, this has several flaws. First, there is the possibility that siblings do not share any family trait and are therefore completely unrelated in this model (see Figure 1.4). Furthermore, since the number of families is finite and no new families arise due to natural fluctuations, all but one family will eventually die out, leaving us with random mating again. However, we solve the latter problem by splitting up families when they become too large. We therefore introduce a maximum family size, and any group that exceeds this size is split equally at random into two subfamilies. One of the two subfamilies is then given a new homogeneous family trait that has not been given to any other family before.

A detailed mathematical description of the adapted model can be found in 2.6.1. However, we can see that, despite these limitations, consanguineous mating works well enough to answer the question of how recessive lethal diseases behave in a growing population. We obtain similar results when we compare the results with a simulation that is able to trace the exact pedigree information for each individual. For more details on this argument, see section 2.2.3.

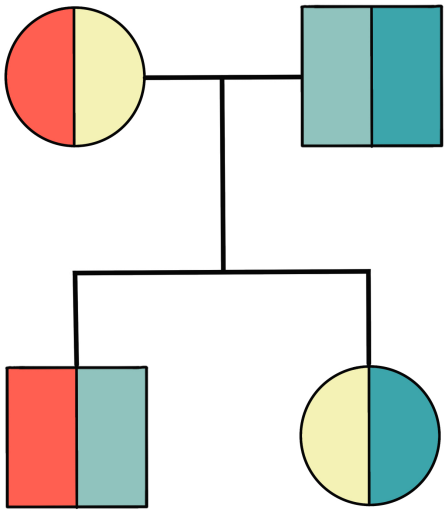

Figure 1.4: Problem with family flags. Family flags are represented by distinct colours. The possibility is presented where siblings born of the same parents do not possess any shared familial flag.

1.6 Recessive gene count and recombination

Despite significant advancements in genome sequencing and the development of sophisticated estimation methods, there remain considerable uncertainties surrounding the mutation rates that lead to autosomal recessive diseases [217, 4, 215]. This uncertainty also extends to the number of genes - referred to as the "recessive gene count" - that, when mutated, can result in lethal recessive diseases [106, 163, 122]. Given the uncertainties in mutation rates, recessive gene count, and additionally in population size, and recombination rates, our investigation focuses on exploring different parameter combinations to understand their effects on the dynamics of lethal recessive diseases

In this chapter, we introduce some genetic concepts that are fundamental to understanding the mechanisms of the models presented in Chapter 3. We take a closer look at the mechanism of recombination and its role in the emergence of complex organisms. Of particular interest is Muller's ratchet, a theory that explains how deleterious mutations accumulate in asexual, non-recombining populations. Additionally, we explore the evolution of mutation rates and how random genetic drift can impede the refinement of phenotypes, resulting in suboptimal traits.

1.6.1 Recombination

Recombination plays a pivotal role in the evolution of populations by introducing genetic variation, which is crucial for natural selection and adaptation. By reshuffling alleles between chromosomes, recombination creates new combinations of genes, allowing populations to respond more effectively to changing environments and selection pressures. The process of recombination can be broken down into two main types: interchromosomal recombination and intrachromosomal recombination.

Interchromosomal recombination occurs during meiosis, where homologous chromosomes (one from each parent) are randomly distributed into two haploid daughter cells. This random assortment of chromosomes results in gametes with different combinations of maternal and paternal chromosomes, contributing to genetic diversity. For example, if a diploid organism has n chromosomes, there are 2^n possible combinations of chromosomes that can be passed on to the offspring, depending on how the chromosomes are distributed during meiosis. As the number of chromosome pairs increases, the potential number of combinations increases exponentially, greatly enhancing genetic variation in the offspring.

Intrachromosomal recombination, occurs during meiosis when homologous chromosomes pair up and exchange segments of DNA through a process known as *crossing over*. This exchange happens at the chiasmata, points where the chromosomes physically overlap and swap genetic material. The result is the formation of recombinant chromosomes that contain a mix of alleles from both the maternal and paternal chromosomes. The exchanged segments can vary in size, ranging from a few base pairs to entire gene regions. The frequency and exact locations of crossovers are not uniform across the genome but are influenced by factors like chromosomal structure, sequence motifs, and the presence of specific proteins that guide the recombination machinery. This intrachromosomal recombination contributes significantly to

genetic diversity, such that almost surely no two gametes from one parent carry the exact same haploid genome. For more details on the mechanisms of recombination, we recommend the book by Bruce Alberts [5].

In our models, we previously considered only interchromosomal recombination. However, to better reflect the natural processes and the role of recombination in evolution, we must now incorporate intrachromosomal recombination. For more details on the modifications we are making to our core model to better capture recombination, see sections 1.6.4 and 3.6.3. As we will see in the next section, in the absence of recombination, populations could face significant challenges.

1.6.2 Muller's ratchet

Muller's ratchet is a concept in evolutionary genetics named after Hermann Joseph Muller. It describes a process by which genomes of an asexual population accumulate deleterious mutations in an irreversible manner due to the lack of recombination. This phenomenon has significant implications for understanding the evolution of sex, the degeneration of chromosomes, and the persistence of asexual lineages. We will discuss its implications and some mathematical models around the concept in the following section. For a deeper look into the topic we recommend to read the article [145] and explore its comprehensive glossary.

Hermann Joseph Muller introduced the concept of Muller's ratchet in 1932 [161, 162] to address a fundamental question in evolutionary biology: why do so many organisms reproduce sexually despite the apparent costs associated with sexual reproduction? Muller hypothesized that sexual reproduction provides an evolutionary advantage by allowing recombination, which helps eliminate deleterious mutations from the genome. In asexual populations, the absence of recombination means that once a deleterious mutation occurs, it can only be purged if the entire lineage carrying it is lost. Over time, stochastic events (genetic drift) can lead to the irreversible accumulation of these mutations, particularly in small populations. This process is akin to a ratchet mechanism, where the "clicks" represent the fixation of deleterious mutations, and the ratchet cannot turn back. Even though this is not the only advantage of sexual reproduction over clonal reproduction, it certainly is a piece in the puzzle of the evolution of sexual mating. In particular, Muller's ratchet provides a reasonable explanation for the degeneration of Y chromosomes in sexual organisms. Y chromosomes, which do not undergo recombination over most of their length, are especially susceptible to the accumulation of deleterious mutations. This accumulation can lead to the loss of functional genes and the shrinkage of the Y chromosome over evolutionary time [41, 42, 184, 45, 46].

It was John Haigh in 1978 who first quantified the effect of Muller's ratchet and proposed a mathematical model [96]. He formulated a Wright-Fisher model, where individuals are characterized by the number of deleterious mutations. Mutations are added at a constant rate according to a Poisson-distributed random variable, and the progeny of an individual with k deleterious mutations is proportional to $(1-s)^k$, where $s > 0$ is the selection coefficient, modelling the strength of the deleterious effect of individual mutations. For more details on the transition rates, see section 1.2.2.

Haigh divided the population into load classes, where each load class consists of individuals with the exact same number of deleterious mutations. The ratchet clicks when the least-loaded class - the class of individuals with the smallest number of deleterious mutations - goes extinct due to natural fluctuations. In the absence of recombination and back mutation, this extinction is final.

After the ratchet clicks, a new least-loaded class emerges, where every individual carries one more mutation than those in the load class that just went extinct. Within this framework, Haigh identified the stationary distributions of this process. Specifically, he found that a distribution $(n_k)_{k\in\mathbb{N}}$, where n_k represents the proportion of individuals with exactly k mutations has Poisson weights. The only stationary distribution with $n_0 > 0$ - so before the ratchet has clicked - is given by

$$n_k = Ne^{-\theta}\frac{\theta^k}{k!} \quad \text{,with} \quad \theta = \frac{\lambda}{s}$$

for $k = 0, 1, \dots$ and where $N \in \mathbb{N}$ is the population size and $\lambda \geq 0$ is the average number of new, deleterious mutations per birth. Moreover, Haigh found that shortly after the ratchet clicks, a new equilibrium is established, which takes the same shape as before. Namely if the ratchet has clicked $J \geq 1$ times one get the following equilibrium distribution

$$n_0 = n_1 = \cdots = n_{J-1} = 0\,,\; n_{J+k} = Ne^{-\theta}\frac{\theta^k}{k!} \quad \text{,for} \quad k = 0, 1, \dots$$

Therefore, the shape of the equilibrium load class distribution remains the same, but it shifts to the right with each click of the ratchet.

1.6.2.1 Click rate

Although the mathematical model of Muller's ratchet is well understood and many questions have been explicitly answered, some questions remain open. One intriguing question is whether the ratchet has an evolutionary effect. Specifically, how many generations will it take for a population to accumulate deleterious mutations, or mathematically speaking, what is the rate at which the ratchet clicks? Exact results for this question are still not available, even in the relatively simple model formulated by Haigh. However, there are good approximations and simulations for certain parameter regimes.

Haigh initially found a linear relationship between the size of the least-loaded class n_0 and the average time between clicks of the ratchet, though this was only for relatively small values of n_0. For larger values of n_0, the selection coefficient becomes more relevant, and rates can vary as simulations for parameter regimes that mimic large non-recombining chromosomal sections suggest [193, 90].

Besides numerical approximations also analytic approximations have been made. Etheridge, Pfaffelhuber and Wakolbinger suggest a diffusion approximation and found that the parameter

$$\gamma = \frac{N\lambda}{Ns\log(N\lambda)}$$

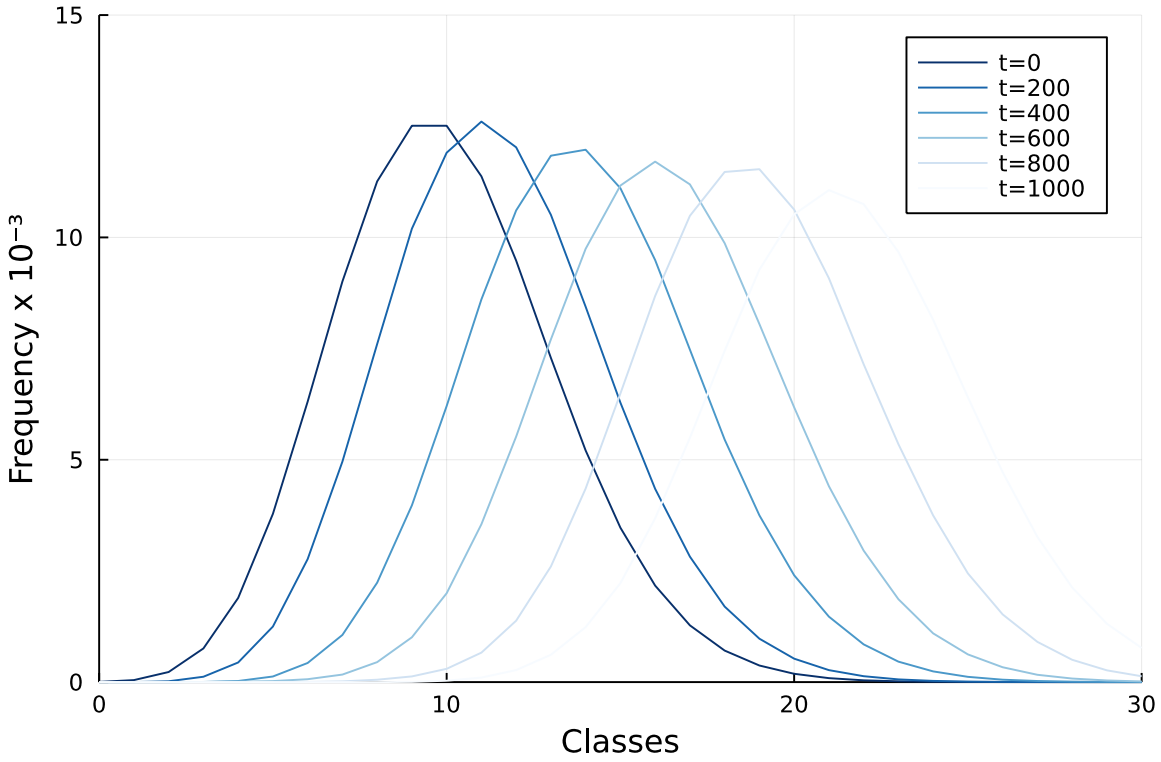

Figure 1.5: Load classes under Muller's ratchet. The load class distribution under the classical effect of Mullers ratchet, as described by Haigh, exhibits a consistent shape over time, with a gradual shift to the right.

plays an important role [64]. Note that they rescaled time in units of the population size N and therefore parameters like the selection coefficient s and the mutation rate λ get multiplied by N. They analysed the following Fleming-Viot diffusion, which is an infinite-dimensional version of the standard multi-dimensional Wright-Fisher diffusion. Let X_k be the frequency of individuals with exactly k mutations; then the diffusion can be expressed as a stochastic differential equation. For $k = 0, 1, 2, \ldots$

$$\mathrm{d}X_k = \Big(s\left(M_1(\mathbf{X}) - k - \lambda\right) X_{k-1}\Big)\,\mathrm{d}t + \sum_{j\neq k}\sqrt{\frac{1}{N}X_j X_k}\mathrm{d}W_{jk} \tag{1.3}$$

where $X_{-1} \coloneqq 0$ and $(W_{jk})_{j>k}$ is an array of independent standard Brownian Motion with $W_{kj} \coloneqq W_{jk}$. Moreover $M_1(\mathbf{X})$ is the mutation load, hence the average number of mutations per individual of the population $\mathbf{X} = (X_0, X_1, \ldots)$. Similar to Haigh's findings in the discrete model, they discovered a travelling wave solution for the problem, and the speed of the wave is determined by the variance of the profile. The main challenge is that, to analyse the dynamics of the least-loaded class, one must know the state of the entire, potentially infinite-dimensional population. Therefore, they sought to find a good approximation of the average number of mutations per individual, $M_1(\mathbf{X})$, based on the size of the least-loaded class. Simulations suggested a linear relationship between the two, through which they derived a one-dimensional diffusion approximation.

Finally, they were able to formulate the following rule of thumb: The rate of the ratchet is of order $N^{\gamma-1}\lambda^{\gamma}$ for $\gamma \in (1/2, 1)$, whereas it is exponentially slow in $(N\lambda)^{1-\gamma}$ for $\gamma < 1/2$.

One can easily see that the classical Muller's ratchet model from Haigh clicks almost surely in finite time, since with probability $(1 - e^{-\lambda})^N$, all individuals mutate in one generation, which necessarily leads to a click. Similarly, in the diffusion approximation, the probability that the least-loaded class goes extinct in finite time is one [9]. Therefore, the diffusion approximation from (1.3) is a reasonable approximation for the phenomenon of Muller's ratchet, as simulations had already predicted [64]. Moreover, Audiffren proved that the time to the first click has an exponential moment of some order, which depends on the parameters N, s and λ [9].

The interest in the mathematical analysis of the accumulation of deleterious mutations in non-recombining populations remains strong. Several modifications of the diffusion approximation have been explored, such as adding compensatory mutations [179], incorporating spatial structure [73], or considering different fitness landscapes [89].

Moreover, there has been significant interest in understanding the dynamics between two consecutive clicks. Recently, Pardoux proved the existence and uniqueness of quasi-stationary distributions for both finite and infinite-dimensional diffusion approximations [152].

1.6.2.2 Muller's ratchet in diplolids

In diploid organisms, Muller's ratchet acts both in clonal reproduction [151] and in mating with segregation. The case of diploid individuals that reproduce clonally behaves similar to a haploid population [45]. An important factor for the analysis of the effect of Muller's ratchet on diploids is the selective effect of heterozygous mutations. Hence the dominance coefficient h is added to the parameters that influence the strength and speed of the ratchet.

Dominance coefficient The dominance coefficient scales the effect of heterozygous mutations in diploid models. More precisely if there are two alleles a and A at a locus and we assume that A is the deleterious variant with the selection coefficient s, then the fitness effect of the diploid genetic configurations are

genotype	aa	aA	AA
fitness	1	$1 - hs$	$1 - s$

Hence there are different interpretations of the heterogeneous genotype depending on the value of h. These are

- **Complete dominance** ($h = 0$): The heterozygote aA has the same fitness as the homozygote aa, implying that allele a is completely dominant over A.
- **Complete recessiveness** ($h = 1$): the heterozygote aA has the same fitness as the homozygote AA, indicating that allele A is completely dominant over a.

- **Incomplete dominance** ($0 < h < 1$): The fitness of the heterozygote aA is intermediate between the two homozygotes, reflecting that neither allele is completely dominant.
- **Overdominance** ($h < 0$): The heterozygote aA has higher fitness than both homozygotes, a situation often referred to as heterozygote advantage.
- **Underdominance** ($h > 1$): The heterozygote aA has lower fitness that both homozygotes.

Deleterious mutations are often recessive ($h \approx 0$), meaning they are less likely to be expressed in heterozygous individuals and thus can persist at low frequencies in populations [180]. There is a detailed discussion of complete recessive diseases in section 1.3.2. The dominance coefficient is thus a key factor in understanding the dynamics of allele frequencies, the persistence of genetic variation, and the evolutionary potential of diploid populations. However, the actual distribution of dominance coefficients in natural populations remains under-explored, limiting the predictive power of population genomic approaches [21]

In the case of sexual mating with segregation but without recombination, it is not the extinction of the mutation-free diploid individual however, that triggers the ratchet and leads to an irreversible accumulation of deleterious mutations. Even after the loss of the mutation-free individual, it can be restored through mating between two parents, each providing a mutation-free gamete. Therefore, it is more the extinction of the mutation-free gamete that leads to a click of the ratchet. Consequently, even in a diploid population, the focus is on the extinction of the least-loaded haploid class. Moreover really recombination is the driving force that prevents the degeneration of a population through Muller's ratchet. Sexual mating by itself with only segregation is not enough [192].

Rapid fixation Most models in the context of Muller's ratchet assume an infinitely large genome, where each new mutation appears at a completely new locus. Within these models, the fate of an individual mutation cannot be determined, and the allele frequencies per site remain infinitesimal regardless of the ratchet's progression. However, Charlesworth argued that in a haploid population with a finite genome and a limited number of loci, after a click of the ratchet, the allele frequency for a single mutation will rise quickly, leading to the fixation of one particular mutation within the next least loaded class [45].

Charlesworth's theoretical arguments were based on simulation results in small populations, where this effect is expected to be more pronounced than in larger populations. The rate of fixation, therefore, depends not only on the mutation rate but also on the population size. The reason for this rapid fixation is that if, for example, the class of individuals without any mutations is lost, the class of individuals with exactly one mutation has no new input due to the lack of back mutation. Consequently, individuals in this class can only be lost due to additional mutations, leading to a fixation within the class with one mutation and high frequencies of this allele in classes with more than one mutation.

In a diploid population, the term fixation is not as clearly defined as in haploid populations. Depending on the dominance and selection coefficient, fixation can mean either that every haploid gamete carries the mutation, hence it is homozygous in every individual of that class,

or that every diploid individual carries the mutation on at least one of its chromosomal pairs. If the selection coefficient is high and dominance is low, fixation in the first sense would imply a significant reduction in fitness for this class, which could eventually lead to its extinction. However, if the dominance coefficient is sufficiently small, the fixation of a deleterious mutation can be decoupled from the advance of the ratchet.

Moreover, Charlesworth found that in some cases, populations become "crystallized" into segregating haplotypes, within each of which deleterious alleles are fixed [45]. This phenomenon will be explored in detail in Chapter 3 and Section 1.7.2, providing further insights into the complex interplay between mutation, selection, and genetic drift in shaping population genetics.

1.6.3 Drift-Barrier hypothesis

For a long time, there has been great interest in understanding the origins of mutation rates. In the 1990s, when mutation rate estimates and genetic data were limited, it was hypothesized that the genome-wide mutation rate would be constant across species, a concept that became known as Drake's rule [60]. However, with advancements in genome sequencing and mutation rate estimation, it became clear that this does not hold true across the entire tree of life. The genome-wide mutation rate varies significantly across different species, raising the question: why? Michael Lynch proposed a unifying theory on the evolution of mutation rates across different species, emphasizing the interplay between directional forces such as selection and random forces like genetic drift. This theory became known as the drift-barrier hypothesis, which we will explain in the following section. For more background and insight, we recommend reading the article [149] and exploring its comprehensive glossary.

At the core of the drift-barrier hypothesis lies the assumption that directional evolutionary forces work against random forces to create an evolutionary equilibrium. In particular, this means that whenever genetic drift is more pronounced (for example, in small populations), selection as its counterpart as a directional force will be weaker. This has several implications for evolution.

Firstly, consider the evolution of a phenotype in a fixed environment. If we assume mutations change a phenotype in a neutral way, meaning that some lead to a fitter and some to a less fit phenotype with equal probability, then selection will drive the population toward a better-adapted trait. This improvement comes to a halt when the selective advantage of new mutations becomes smaller and smaller, and the fixation of a slightly fitter mutant gets inhibited by the random fluctuations in the system. Hence, at some point, genetic drift prohibits the further refinement of a phenotype and keeps it below a theoretical optimum.

Additionally, the interplay between selection and genetic drift shapes the mutation rate itself. This means that the mutation rate for a population is not fixed, but rather evolves like any other trait under selection. In this context, selection primarily aims to improve the replication mechanism and, consequently, to lower the mutation rate. However, it needs to be balanced between the costs of deleterious mutations and the need for genetic diversity for adaptation.

In small populations, genetic drift is strong, meaning random fluctuations can dominate over selection. This allows for higher mutation rates because deleterious mutations are not efficiently purged by selection. In large populations, however, selection is more effective at removing deleterious mutations, favouring lower mutation rates. Thus, the mutation rate is kept low to minimize the accumulation of harmful mutations. This balance between mutation rates and the strength of selection versus genetic drift is a key aspect of the drift-barrier hypothesis, explaining the variability in mutation rates observed across different species.

The theory matches with genetic data from a wide range of species, ranging from bacteria and other prokaryotes to unicellular eukaryotes to complex, multicellular organisms like humans [149]. Although there have been findings of organisms that deviate from the trajectory drawn by the drift-barrier hypothesis, the vast majority of species fit the model [200].

A question that still remains is where the drift-barrier appears. As mentioned above it is shaped by the population size which affects the genetic drift. Lynch and co-authors also found that the relative frequency of mutator alleles versus anti-mutator alleles plays a significant role. Mutator alleles are genetic variants that increase the mutation rate, for example by suppressing repair mechanisms. They can be beneficial in rapidly changing environments by generating genetic diversity, but they can also increase the load of deleterious mutations. On the other hand, anti-mutators are genetic variants that decrease the mutation rate. They reduce the accumulation of deleterious mutations, enhancing genome stability, but may limit adaptive potential in changing environments.

We found in the model discussed in Chapter 3 that there is another parameter - the recessive gene count - that influences the genetic variation within a population and thereby also contributes to the drift-barrier hypothesis.

1.6.4 Model modifications

To better capture the effects of varying parameter regimes and to model recombination more realistically, we introduce a series of modifications to the core model presented in section 1.3. These modifications focus on refining our understanding of how recombination impacts the inheritance of genetic traits in the context of autosomal recessive diseases.

In the core model, the inheritance of each gene was treated as independent, meaning that during reproduction, genes were passed on without considering their relative positions on chromosomes. This simplification allowed us to model inheritance in a straightforward manner but did not fully capture the complexity of genetic recombination as it occurs in nature (see 1.6.1).

In earlier modifications (introduced in Section 1.5.3), we took a step towards realism by dividing the genome at predetermined positions, allowing for independence between segments while still maintaining some structure. This approach introduced the idea that recombination could occur between chromosomes, but it still assumed a relatively fixed structure.

Now, we introduce a new parameter, the *recombination rate* $r \in [0, 1]$, which adds a layer of complexity to how recombination is modelled. This parameter reflects the probability that, during meiosis, a recombination event occurs independently at each of the $N - 1$ possible

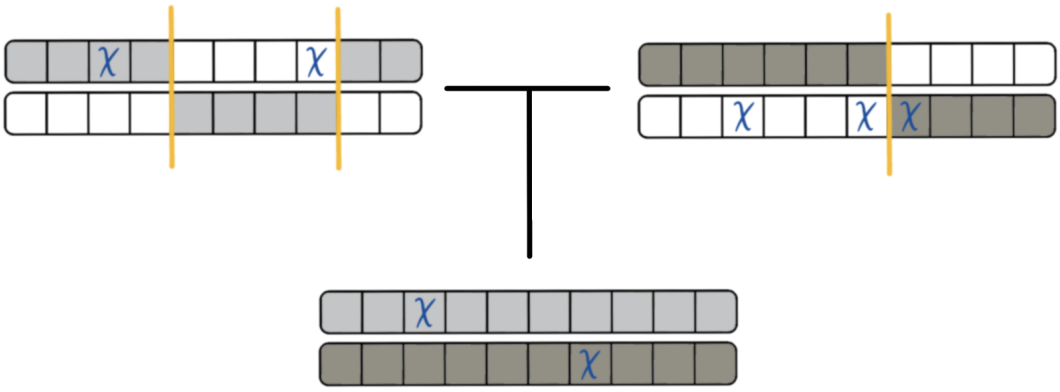

Figure 1.6: Mating with recombination. A schematic representation depicts the mating scheme with recombination, which ultimately forms a new diploid offspring. The orange vertical lines represent the recombination sites.

breaking points along the genome. The value of r can range from 0 to 1, where $r = 0$ refers to no recombination, meaning that only segregation happens, and the haploid genome is inherited as a single, unbroken unit from one parent. Whereas $r = 1$ refers to full recombination, resulting in independent inheritance of each gene, akin to the assumption in the core model. By varying r, we can simulate different levels of recombination, ranging from complete linkage (no recombination) to full independence (high recombination). This flexibility allows us to model a spectrum of genetic scenarios and observe how varying recombination rates affect the spread and persistence of recessive alleles within a population. The modified version of the mating probabilities with recombination can be found in (3.8).

We will see in section 1.7 that under theses extreme condition of complete recessive lethals (dominance coefficient $h = 0$, selection coefficient $s = 1$) the effect of Muller's ratchet is fatal. When the mutation free class (the class of genomes with the no mutations) is lost, the entire population rapidly heads toward extinction. To gain deeper insights into the mechanisms that drive population extinction under these conditions, we make a modification to our model: we adapt the birth rates of fit individuals to counterbalance the increasing disease prevalence. This adaptation allows us to maintain a relatively stable population size while observing the effects of Muller's ratchet more clearly. Thus we maintain a total birth rate of $N_t\bar{b}$, as if the whole population were healthy, and distribute it equally among all healthy individuals. This accelerates the birth rate of all healthy individuals and they compensate for the individuals lost due to the expression of the disease. The generator for the adapted model is augmented with an additional birth factor, resulting in the following form. Let $f : \mathcal{M}(\mathcal{X}^2) \to \mathbb{R}$ be any bounded, measurable function then for all $\nu \in \mathcal{M}(\mathcal{X})$, the generator

is given by

$$(\mathcal{L}f)(\nu) = \int_{\mathcal{X}^2} b(\mathbf{x}) \left(\frac{\nu\left(\mathcal{X}^2\right)}{\nu\left(\mathcal{X}^2\right) - \nu\left(\mathcal{D}_N\right)} \right) \int_{\mathcal{X}^2} \frac{b(\mathbf{y})}{\langle \nu, b \rangle} \int_{\mathcal{X}^2} \left(f\left(\nu + \delta_{\mathbf{z}}\right) - f\left(\nu\right) \right) m(\mathbf{x}, \mathbf{y}, d\mathbf{z}) \nu(d\mathbf{y}) \nu(d\mathbf{x})$$
$$+ \int_{\mathcal{X}^2} \left(d(\mathbf{x}) + \int_{\mathcal{X}^2} c(\mathbf{x}, \mathbf{y}) \nu(d\mathbf{y}) \right) \left(f\left(\nu - \delta_{\mathbf{x}}\right) - f\left(\nu\right) \right) \nu(d\mathbf{x})$$

It is important to note that this adaptation does not reflect the dynamics of a natural population. In reality, as the disease prevalence increases, we would expect the population size to decrease.

1.7 Outline, main results and open questions

The main body of this thesis is structured into three interconnected chapters, each of which can be read independently, though their contents are closely related. Chapters 2 and 3 give a concrete genetic application of the model introduced in section 1.3. Chapter 4 introduces a simulation framework used to perform the stochastic simulations and obtain the results in this thesis. It serves as a manual so that the framework can be used for other research questions involving complex systems. In the following, we summarise the contents and main results of the three chapters. Some results are extended with heuristics, conjectures and open questions.

1.7.1 Transient drop in prevalence for random mating during population expansion

In chapter 2 we analyse the mutation burden and prevalence for a recessive lethal disease in a growing, consanguineous population. We found that the observed reduced incidence rates for recessive diseases in randomly mating populations are a transient phenomenon induced by population expansion at the cost of increased mutation burden. This chapter was published in the American Journal of Medical Genetics as joint work with Julia Frank, Heidi Beate Bentzen, Jean Tori Pantel, Konrad Gerischer, Anton Bovier and Peter M. Krawitz [135],

> L. A. La Rocca, J. Frank, H. B. Bentzen, J. T. Pantel, K. Gerischer, A. Bovier, and P. M. Krawitz. Understanding recessive disease risk in multi-ethnic populations with different degrees of consanguinity. *American Journal of Medical Genetics*, 194(3):e63452, 2024

Chapter 2 contains the published version, with only minor changes to correct some typing errors and adapt the layout to the format of this thesis.

We have analysed the prevalence and mutation burden of severe autosomal recessive diseases. The mutation rate and genetic architecture - that is, the number and size of recessive genes and their distribution across a fixed number of chromosomes - were kept constant throughout the analysis and were based on estimates from human population data. There are two things we have changed in the course of this analysis. First, we start the simulation with a completely healthy, i.e. mutation-free, population of 500 individuals. After a short "burn-in" period to allow mutation-selection equilibrium to establish, we increased the carrying capacity to allow for approximately 10 000 individuals. We then continued the simulation to see the effect of the population increase. Secondly, we ran independent simulations with a number of different mating schemes. We compared a randomly mating population with consanguineous mating. We can model the strength of consanguinity by three parameters. The first is the maximum family size κ. In addition, two weights α and β represent the probability of mating within the close family and within the wider family group. The remaining proportion of individuals continue to mate at random.

We found that in a randomly mating population there is a sharp, transient drop in incidence rates during and after population expansion. However, this is associated with an increase in mutation load. It takes more than 500 generations after the expansion for the population to return to equilibrium. The prevalence level of before the increase is reached again, but

the mutation burden in the larger population is higher than before. In contrast, consanguineous populations do not experience this temporary decline. The prevalence level of the inbreeding population is approximately the same as that of randomly mating populations. The mutation load, however, is much lower. And unlike the random mating population, population expansion keeps both statistics constant. This is because, for individuals in a randomly mating population, the number of potential mates increases with population size. For a consanguineously mating individual, the number of potential mates is bounded by the maximum family size and therefore more or less independent of population growth (see Figure 2.1).

In addition, we compared our simulation framework of an adaptive dynamics model implemented with an exact stochastic algorithm with an evolutionary simulation framework called SLiM, which works internally with a Wright-Fisher model [98]. The big advantage of the latter simulation framework was that it works with exact pedigree information. By knowing exactly how many ancestors individuals have in common, consanguineous mating could be implemented more realistically (see Figure 2.4). However, the discrete, non-overlapping population of constant or deterministically increasing size of the Wright-Fisher model cannot model natural population growth as well as the adaptive dynamics model (see Figure 2.5). It was therefore intriguing to see that when we made all the parameters in both models as equal as possible, we got similar results, even though the backgrounds of the two simulation frameworks were quite different. Thus, we saw that the simple implementation of consanguineous mating with the addition of family flags was sufficient for the purpose of our research question.

Finally, we analysed the effect of family size and degree of consanguinity on the dynamics. The prevalence remained constant regardless of the degree of consanguinity, but the mutation load increased with family size. By increasing the family sizes, we obtained a gradual transition from a consanguineously mating population to a randomly mating population (see Figure 2.2). The same effect was observed in the statistics of mutation burden and prevalence. However, the effect of family size or, in a broader context, population size on mutation burden is not yet well understood. As well as the influence of the number of recessive genes and gene architecture on mutation burden and prevalence. We have analysed the latter in Chapter 3.

1.7.2 The role of recessive genes in genome stability and population collapse

In chapter 3 we analysed recessive lethal diseases from the perspective of different parameter regimes. We found that natural populations face a barrier when we increase the mutation rate and/or the number of recessive genes. We compared this with the drift barrier hypothesis (see section 1.6.3) and further investigated the effect of Muller's ratchet (see section 1.6.2) on complete recessive lethals. We found that the extinction of mutation-free gametes destabilises the population, and that stability is only restored by the formation of clusters of highly correlated genes. Chapter 3 is available as a preprint as joint work with Konrad Gerischer,

Anton Bovier and Peter M. Krawitz [136],

> L. A. La Rocca, K. Gerischer, A. Bovier, and P. M. Krawitz. Refining the drift barrier hypothesis: a role of recessive gene count and an inhomogeneous Muller's ratchet. https://arxiv.org/abs/2406.09094, 2024

Chapter 3 contains the preprint, with only minor changes to correct some typing errors and adapt the layout to the format of this thesis.

To produce the results of this research, we ran numerous independent simulations with different genome-wide mutation rates μ and different numbers of recessive genes N with a fully recombining genome ($r = 1$) or in the absence of recombination with only segregation ($r = 0$). We found that the haploid load classes $c_k(t)$, that is, the fraction of gametes in the population with exactly k lethal equivalents (defined in (3.3)), play a crucial role in the stability of the population. Without recombination, the mutation-free gametes $c_0(t)$ can go extinct, leading to mutation fixation, which in our setup leads to extinction of the entire population (see Figure 1.7).

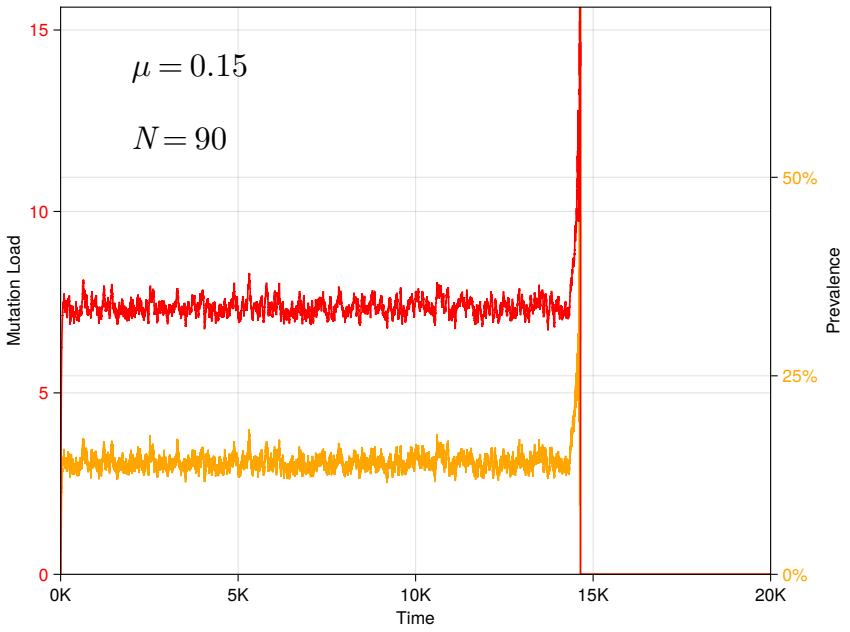

Figure 1.7: Population extinction in absence of recombination. In the absence of the fixation of the total birth rate in the adaptive dynamics model and in the absence of recombination, the extinction of the mutation-free gamete results in the rapid extinction of the entire population. The plot illustrates the mutation burden (red) and prevalence (orange) for $N = 90$ recessive genes and a genome-wide mutation rate of $\mu = 0.15$. Time is measured in 1000 generations.

The probability of extinction of the least loaded class, leading to a destabilisation of the population, depends strongly on three parameters. The mutation rate μ, the population size K and the number of recessive genes N. The drift-barrier hypothesis already finds a correlation between mutation rate and population size through a balance between natural

selection and genetic drift [149]. We were able to show that the number of recessive genes also influences the variation in the population, and therefore adds to the drift-barrier hypothesis as a trait that influences genetic drift (see Figure 3.4). Moreover, recombination effectively reduces this variance, allowing the organism to tolerate a higher number of recessive genes.

Switching to the stable population size model (for more details see section 1.6.4), we observed an inhomogeneous click rate for Muller's ratchet. After the initial extinction of the least loaded class c_0, which can take several thousand generations depending on the parameter combination, the next least loaded classes become extinct in a few hundred generations. Only after a significant mutational load has accumulated the click rate slows down and extinctions become less frequent (see Figure 3.2). In the classical Muller's ratchet, the quasi-stationary distribution of the haploid load classes retains its shape following each click, exhibiting only a rightward shift [96, 64]. However, our observations indicate not only a shift but also a complete change in shape, resulting in a distribution that resembles a juxtaposition of multiple independent distributions. Nevertheless, the shape and characteristics of the distribution are not maintained (see Figure 1.8).

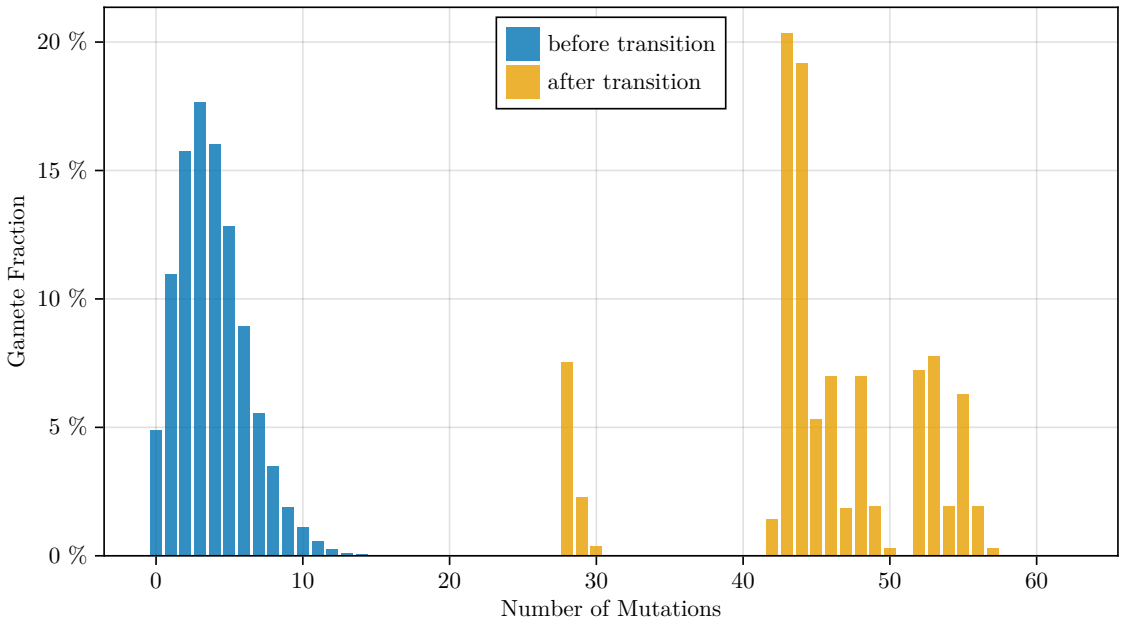

Figure 1.8: Haploid load class distribution before and after the transition. The initial equilibrium haploid load class distribution (blue) is compared with the load class distribution after the transition (orange). The simulation was conducted with $N = 500$ recessive genes and a genome-wide mutation rate of $\mu = 0.03$. Both distributions represent the average of approximately 20 000 generations, during which the mutation burden and prevalence remained constant.

Looking at the correlation between genes in the population (defined in 3.4), we see that the stabilisation of the population at a new level of prevalence and mutation burden is associated with the formation of highly correlated clusters of genes. Before the transition, all genes are uncorrelated. After the transition, we observe that there are clusters of genes that behave

almost identically, while the correlation between the clusters remains zero (see Figure 3.4 B). We hypothesize that after the extinction of the mutation-free load class c_0 due to lack of recombination, a lethal equivalent would fix with in the haploid load class c_1 with one mutation as described by Brian Charlesworth [45]. However, if the majority of the population has a mutation in the exact same recessive gene, this will have lethal consequences, as we have seen in our natural population model. The population will therefore destabilise, leading to a series of load class extinctions. This eventually slows down when two or more mutually exclusive haplotypes emerge. In this case, gametes belonging to one haploid cluster cannot successfully mate, but the probability of having a healthy offspring with a mate from another cluster is increased compared to the case where mutations are evenly distributed across the genome due to the extreme alleviated mutation burden.

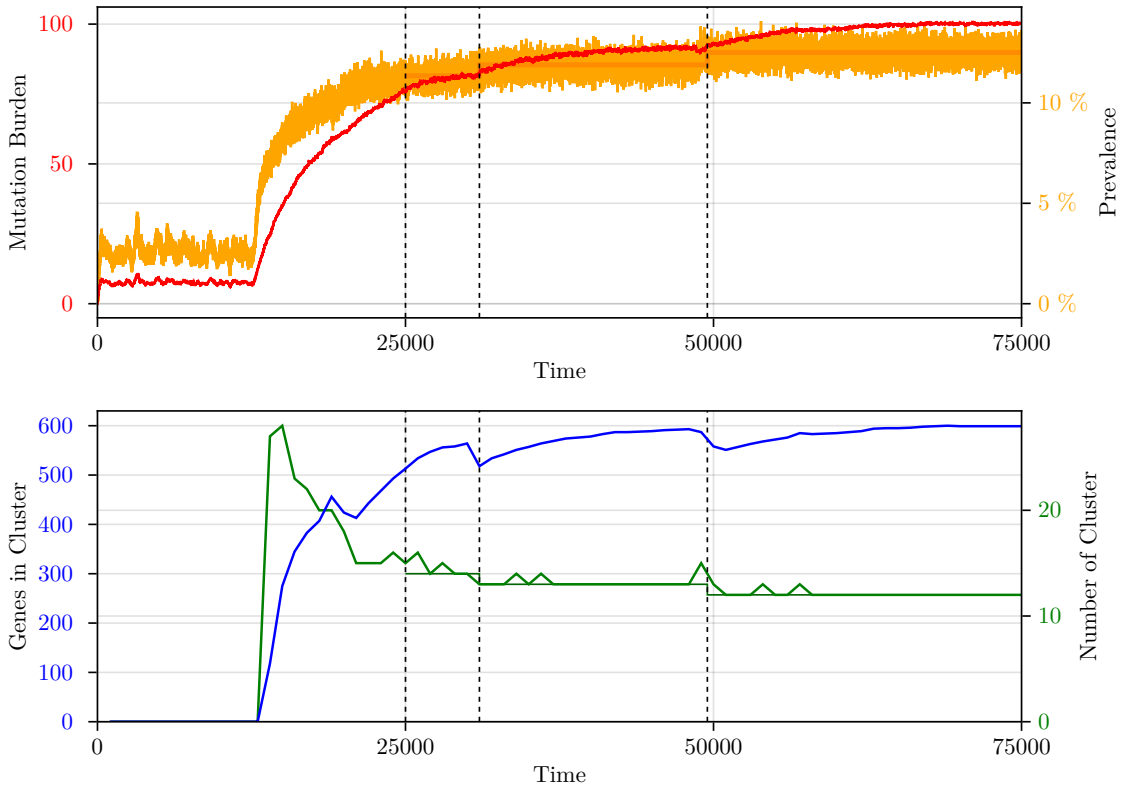

Figure 1.9: Dependence of prevalence and mutation load on cluster sizes. The bottom plot depicts the time trajectory of the top plot, thereby illustrating the correlation between the prevalence and mutation burden and the total number of genes associated with one of the clusters and the total number of distinct clusters. It can be observed that the prevalence level is correlated with the total number of clusters, whereas the mutation burden is contingent upon the number of bound genes.

Following the initial click, the dynamics are dependent on the size and number of these clusters. While not all genes are bound to one cluster and remain uncorrelated, the mutation burden increases gradually with the incorporation of a new gene into a cluster. However, the prevalence remains constant during this process. Conversely, when the number of clusters

changes, the prevalence also changes. If a cluster is outweighed by the others, it collapses, leading to an increase in prevalence. The genes that were previously bound to the cluster become free and are then associated with other clusters (see Figure 1.9).

Our findings indicate that the equilibrium allele frequency of the diseased allele is

$$\varphi = \sqrt{1 - e^{-\mu/N}},$$

as determined by considerations coming from a single gene $N = 1$ (see section 3.6.1 and [167]). Following the transition, all the uncorrelated genes maintain this average allele frequency, whereas the genes in the cluster exhibit a significantly increased allele frequency. As genes within a cluster exhibit near-identical behaviour, the allele frequency is also consistent across all genes in a cluster. The equilibrium frequency appears to be influenced by the size of the cluster. A larger cluster size tends to have a lower average allele frequency for the mutated allele. It can be reasonably assumed that additional factors must be considered in order to accurately determine the allele frequency, beyond the mere cluster size. This is evidenced by the existence of clusters with an identical number of recessive genes that exhibit disparate allele frequencies (see Figure 1.10).

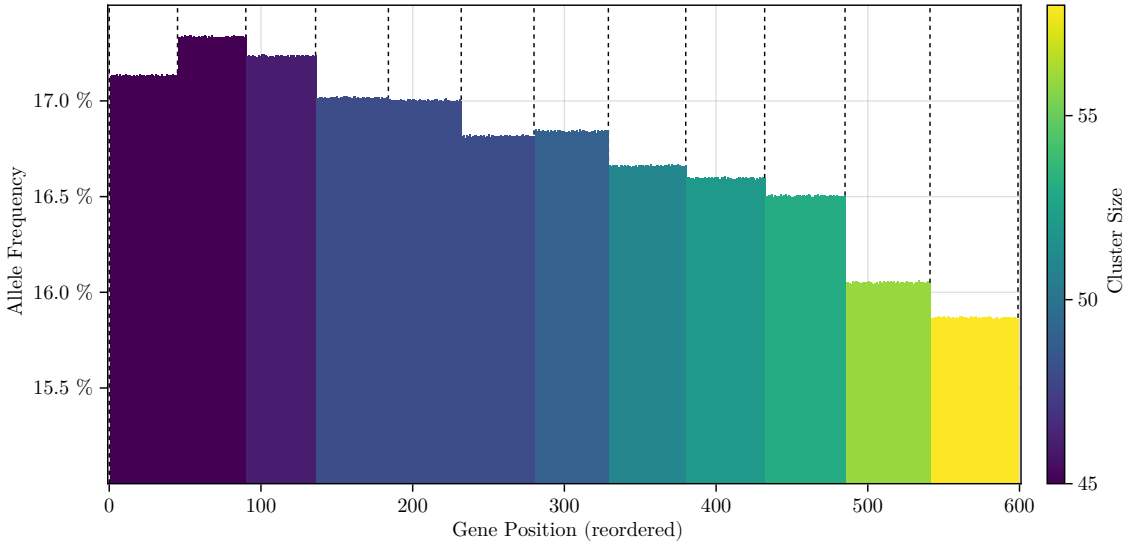

Figure 1.10: Allele frequencies after the transition. The average allele frequencies per gene have been calculated following the transition to higher prevalence levels and the emergence of a cluster. The gene positions have been reordered in a manner that facilitates comparison between clusters, with genes that belong to the same cluster being placed adjacent to one another and clusters with a smaller gene count beginning on the left. The colours have been used to indicate the size of the cluster, while the vertical lines separate the genes from different clusters. The average has been calculated over the last 20,000 generations, and the full simulation is presented in Figure 3.2.

1.7.2.1 Gamete model

The analysis of the diploid model focused on the number of gametes with a certain number of mutations. We looked at the population as a collection of gametes rather than individuals, and the union of gametes to form individuals became relevant only to determine the fitness of the two gametes as a combination of recessive genes. We therefore developed a haploid model, which simplifies the analysis but produces similar dynamics. By focusing only on "fit" gametes and excluding the creation of gametes pairs that would result in a diseased individual, we solve the paradox of simulating recessive diseases in a haploid population.

The trait of an "individual" in this population is a gamete and characterised by a finite sequence in $\mathcal{X} = \{0,1\}^N$. Define $\Lambda^i = \{x \in \mathcal{X} \colon x_i = 0\}$ and for $x \in \mathcal{X}$ define $I_x := \{i \in \{1,\dots,N\} \colon x_i = 1\}$. Every trait $x \in \mathcal{X}$ has a list of potential partners that (without mutation) would lead to a healthy individual. These are given as

$$\Lambda_x := \bigcap_{i \in I_x} \Lambda^i$$

Let ξ_t be the state of the population at time t given by the finite point measure

$$\xi_t = \sum_{i=1}^{M_t} \delta_{x_i}$$

where M_t is the population size at time t and $x_1, \dots, x_{M_t}$ are the gametes alive at time t in an arbitrary order. Let $\mathcal{M}(\mathcal{X})$ be the set of all finite point measures on $\mathcal{X}$. Assume for simplicity that only one mutation happens per birth event. The measure valued Markov process is characterized by the following generator. For any $\phi \colon \mathcal{M}(\mathcal{X}) \to \mathbb{R}$ bounded, measurable

$$\begin{aligned}\left(\mathcal{L}'\phi\right)(\xi) = &\int_{\mathcal{X}} b \frac{\xi(\Lambda_x)}{\langle \xi, 1\rangle} e^{-\mu} \left(\phi\left(\xi + \delta_x\right)\right) - \phi(\xi)\right) \xi(dx) \\ &+ \int_{\mathcal{X}} b \left(1 - e^{-\mu}\right) \int_{\mathcal{X}} \frac{\mathbb{1}_{\Lambda_x}}{\langle \xi, 1\rangle} \sum_{i=1}^{N} \frac{\mathbb{1}_{i \notin I_x \cup I_y}}{N} \left(\phi\left(\xi + \delta_{x+e_i}\right)\right) - \phi(\xi)\right) \xi(dy)\xi(dx) \\ &+ \int_{\mathcal{X}} \left(d + c\langle \xi, 1\rangle\right) \left(\phi\left(\xi - \delta_x\right)\right) - \phi(\xi)\right) \xi(dx)\end{aligned}$$

where $e_i \in \{0,1\}^N$ denotes the i-th base vector, that is zero everywhere, besides at the i-th position, where it takes the value one. The first term is clonal reproduction without mutation by choosing a compatible partner. The second term is birth with mutation by first choosing the right partner and then a mutation location that does not lead to a homogeneous mutation. The third term is death due to natural causes or due to competition. Again, we assumed that birth, death and competition rates were the same for all traits, to ensure that carriers did not feel the burden of the mutation and to get as close to the diploid model as possible. As for the diploid model the trait space $\mathcal{X} = \{0,1\}^N$ is high-dimensional, but finite. Therefore, as in the case of the diploid model, we obtain a finite total event rate and hence the existence and uniqueness of the process defined by the generator above (see 1.3.1). Simulation of this model demonstrates that the mutation load classes play a significant role

in this context as well. As observed in the diploid model, the extinction of mutation-free gametes results in a population destabilisation, manifested by an increase in prevalence and mutation burden. The cascade of extinction of load classes only decelerates when highly correlated clusters emerge. Consequently, we observe the same inhomogeneous click rate and stabilisation mechanism that we see in the diploid model. Despite the possibility of reducing the dimensionality of the trait space in the haploid model, classical mathematical analysis remains a significant challenge. In particular, determining an exact solution to the system of ODEs that arise as a rescaled, large population limit is not feasible. Nevertheless, by reducing the level of complexity, it is anticipated that subsequent analysis and research will become more straightforward using the gamete model over the diploid model. Nevertheless, it becomes evident that a certain degree of complexity is necessary to observe the phenomena that emerge from our models. A further reduction in the dimensionality of the trait space results in the loss of inhomogeneity in the click rate of the Muller ratchet, as will be demonstrated in the following section.

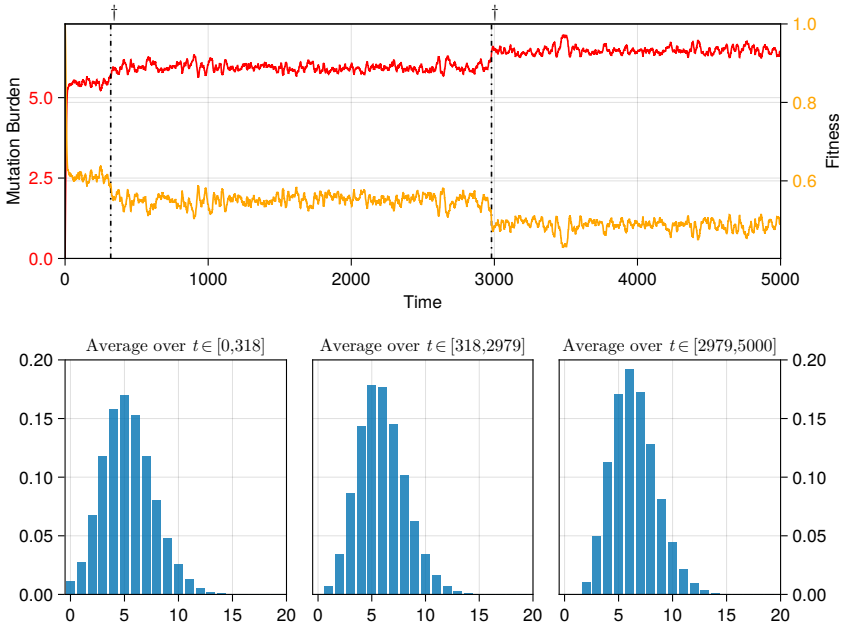

Figure 1.11: Classical Muller's ratchet in position-free model. The above plot illustrates the mutation burden and fitness over time, which is defined as one minus the prevalence. The dotted vertical lines and dagger at the top indicate the extinction times of the least loaded classes. It can be observed that the effect of Muller's ratchet is gradual. The bottom plot depicts the average of the haploid load class distribution between consecutive extinctions. The shape of the distribution appears to be maintained, while it shifts to the right.

1.7.2.2 Position-Free model

In this section, we present a version of the diploid model that neglects to consider the specific position of the mutation, instead focusing only on the total number of mutations per gamete.

We have seen that both an adaptive dynamics model with a stable total birth rate and a Wright-Fisher model produce comparable results. Therefore, we turn here to a Wright-Fisher dynamic.

Consider a diploid Wright-Fisher model. Suppose there are N known lethal recessive genetic diseases. Each individual can carry a certain number of these genetic diseases on each of its two sets of chromosomes. However, for the purposes of this model, we do not keep information about which specific diseases an individual is a carrier for. We only record the number of diseases per set of chromosomes and whether (at least) one of the diseases is manifested, i.e. present in a homozygous state. In such cases, the individual is excluded from the mating process and is considered unfit. During mating, a healthy individual chooses a mate from among all healthy individuals. During gamete formation, one of the two sets of chromosomes is passed on at random, and mutations occur at a constant rate. We assume that each gene degenerates with equal probability and that the the number of new mutations for a gamete carrying k lethal equivalents, during gamete formation follows a *Poisson* distribution with parameter $\mu(1 - k/N)$, where μ is the total gamete mutation rate as before. Thus, we assume that if a gene that already carries a mutation is hit by another mutation, the number of mutations for that individual will not increase. If the offspring is formed from two gametes, one carrying n diseases and the other m, we calculate the probability $p(n, m)$ that none of the n and m mutations will carry the same disease and would result in a homogeneous state in a diploid individual. This probability is given by

$$p(n,m) = \frac{\binom{N}{n+m}\binom{n+m}{m}}{\binom{N}{n}\binom{N}{m}} = \frac{\binom{N-n}{m}}{\binom{N}{m}} = \frac{(N-n)!(N-m)!}{N!(N-n-m)!} \tag{1.4}$$

This is precisely the probability that, when selecting n positions from the N possible positions and then independently selecting m positions from the N possible positions, no position is chosen more than once.

The type of an individual is given as triplet $\mathbf{x} = (x_1, x_2, x_3) \in \{0, 1, \ldots, N\}^2 \times \{0, 1\}$ where the first two entries are the number of mutations on each of the gametes and the third entry is the fitness of an individual. In this context, zero means that the individual is unfit, indicating the presence of at least one homozygous recessive genetic disease, while one means that the individual is fit. The next generation is formed as follows: Let K be the constant population size. Each individual from generation $t + 1$ independently selects two parents from generation t with fitness one. From each parent, it selects one gamete, adds a Poisson-distributed number of mutations, and finally, a Bernoulli-distributed random variable with parameter $p(n, m)$ determines whether the descendant is fit or unfit. Here, n and m represent the number of mutations per gamete after the addition of *de novo* mutations. More precisely the outcome of the mating of the two fit individuals $\mathbf{x}, \mathbf{y}$ with $x_3 = 1 = y_3$ is given as follows. Let $U_x, U_y \sim \mathrm{Ber}(1/2)$ be two independent random variables and let P_x, P_y be two independent *Poisson* random variables with parameters

$$\mu\left(1 - \frac{x_{U_x+1}}{N}\right) \quad \text{respectively} \quad \mu\left(1 - \frac{y_{U_y+1}}{N}\right)$$

and finally let Z be a *Bernoulli* random variable with success probability

$$p(x_{U_x+1} + P_x, y_{U_y+1} + P_y).$$

Then the offspring is given by the triplet $(x_{U_x+1} + P_x, y_{U_y+1} + P_y, Z)$ as seen in the following scheme

$$(x_1, x_2, x_3) \xrightarrow{\text{if } x_3=1} \left(x_{U_x+1} + P_x, y_{U_y+1} + P_y, Z\right) \xleftarrow{\text{if } y_3=1} (y_1, y_2, y_3)$$

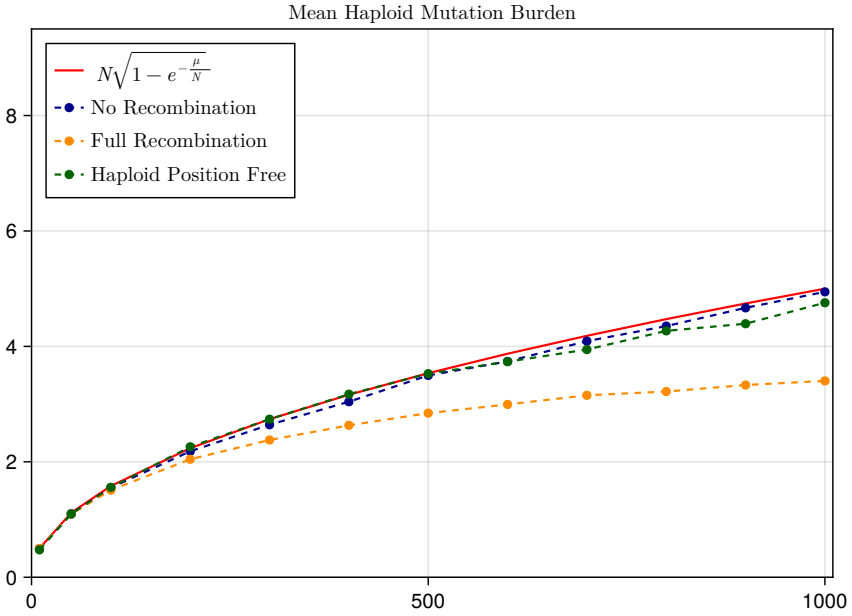

Figure 1.12: Mean haploid mutation burden. A comparison of the mean haploid mutation burden for the three models, with full recombination, without recombination and the position-free model, is presented as a function of the recessive gene count. While the position-free model and the model without recombination exhibit a similar trend, with both remaining close to §N$\sqrt{1-e^{-\mu/N}}$, the full recombination model displays a lower value.

Reduction to haploid model The same dynamics can be achieved once more through the utilisation of a haploid Wright-Fisher model, as was previously employed. In this instance, the focus is on haploid gametes, as opposed to diploid individuals. Once more, the pivotal point is to define an appropriate fitness function, given that a haploid gamete cannot manifest a recessive genetic disease initially. Consequently, the relative fitness of the gametes within the population is considered, with the aim of determining the likelihood of a gamete generating a fit diploid individual within the population. This probability determines the selection of the ancestor for the subsequent generation.

Let $X_k(t)$ be the number of gametes in generation t who carry k mutations and $\mathbf{X}(t) = (X_0(t), X_1(t), \ldots, X_N(t))$. Then the distribution of $\mathbf{X}(t+1)$ conditioned on $\mathbf{X}(t)$ will be multinomial with parameters K and $(p_k(t))_{k=0,1,\ldots,N}$ where

$$p_k(t) = \frac{M_k(\mathbf{X}(t))}{T(t)} \sum_{j=0}^{N} M_j(\mathbf{X}(t)) p(j,k)$$

where $M : \{0, 1, \ldots, N\} \to \{0, 1, \ldots, N\}$ shifts the current population by *de novo* mutations

$$M_k(\mathbf{X}) = \sum_{j=0}^{k} X_{k-j} e^{-\mu\left(1-\frac{k-j}{N}\right)} \frac{\mu^j \left(1 - \frac{k-j}{N}\right)^j}{j!}$$

and $T(t)$ is the total fitness of the population at time t after adding mutations

$$T(t) = \sum_{i,j=0}^{N} M_i(\mathbf{X}(t)) M_j(\mathbf{X}(t)) p(i, j).$$

Here $p(i, j)$ is again the probability form (1.4) that to gametes with i and j uniformly distributed mutations across N recessive genes express a disease.

Although this model still presents a challenge in terms of identifying stationary distributions, it offers a more promising avenue for analysis when compared to the models previously discussed. However, the inability to trace the positions of the mutations prevents the occurrence of fixation, leading us to revert to the classical homogeneous effect of Muller's ratchet: The least loaded class may also become extinct as a result of the lack of recombination. However, the effect is minor. Both the mutation burden and the prevalence increase gradually, and the haploid load class distribution shifts to the right while maintaining its shape (see Figure 1.11). These effects are analogous to those described by John Haigh for a population in which the fitness depends exponentially on the number of mutations [96].

It is therefore evident that monitoring the locations of mutations within the genome is of paramount importance. This introduces a further layer of complexity to the model, thereby rendering the application of conventional techniques difficult. The analysis of the click rate of Muller's ratchet, for instance, is typically conducted through the use of diffusion approximations (for further details, refer to Section 1.6.2 and [64, 89, 179, 152]). Consequently, it is necessary to be able to calculate the drift and diffusion terms, which depend on the fitness of an individual. In this case, however, the fitness of an individual is not solely dependent on the number of mutations, as in the classical model formulated by Haigh [96]. Rather, it also dependents on the state of the entire population. This renders the calculation of these terms and the application of the techniques more challenging. Therefore, we adopted a simulation-based and comparative approach, whereby we compared multiple modifications of the model to gain a deeper comprehension of the dynamics of Muller's ratchet in this context. We found that without this additional layer of complexity, it is not be possible to observe nor the intriguing phenomena of "crystallisation" as described by Brian Charlesworth [45] as the emergence of mutually exclusive haplotypes, nor an inhomogeneous click rate and a stationary distribution of Muller's ratchet, which undergo a change in shape following the initial transition. The behaviour of the system following the initial extinction represents a significant challenge, as does the analysis of the stationary distributions in the initial equilibrium, where the mutation-free load class persists. This will be explored in greater detail in the following subsection.

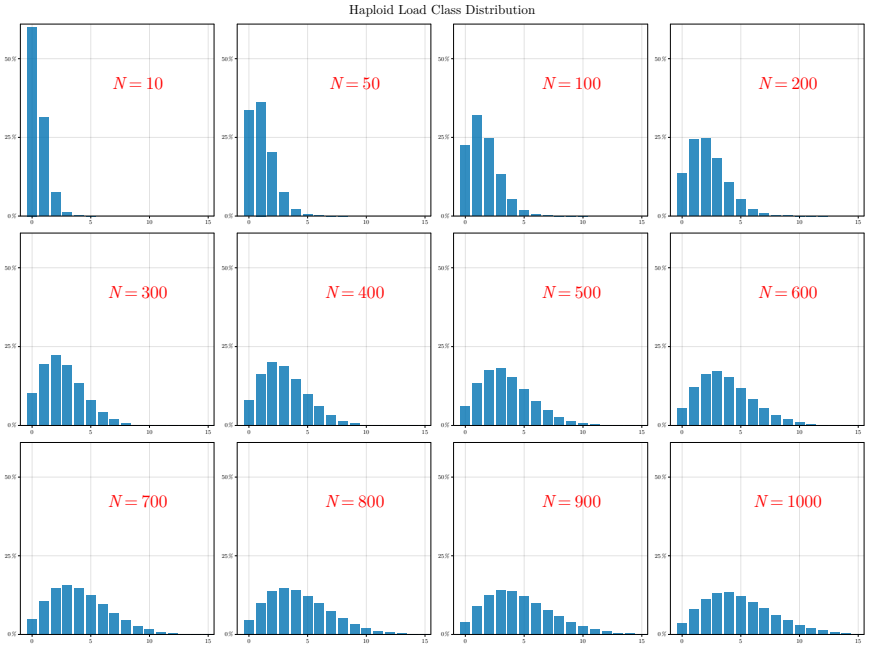

Figure 1.13: Haploid load class distribution for $r = 0$. Each frame illustrates the average haploid load class distribution $(c_k)_{k\in\mathbb{N}}$ in the initial equilibrium state prior to the first click of the ratchet. The number of recessive genes increases from the top left frame, where $N = 10$, to the bottom right frame, where $N = 1000$.

1.7.2.3 Initial quasi-stationary distribution

The identification of quasi-stationary distributions subsequent to the initial click of the ratchet seems out of reach. Similarly, characterizing the initial quasi-stationary distribution presents a significant challenge with numerous unanswered questions. Exact solutions of the deterministic system in the large population limit are only available for a small recessive gene count N (see for example 3.6.1 for $N = 1$). Subsequently, the number of potential traits and, consequently, the dimensionality of the system of ODEs increases exponentially. In a diploid system, there are 4^N potential configurations, whereas in a haploid system, there are 2^N. Since we are interested in genome sizes of several hundred genes, we are limited to numerical analysis paired with heuristics based on the exact results from small genome sizes.

As previously stated, the allele frequency of the mutated allele for $N = 1$ is $\varphi = \sqrt{1 - e^{-\mu}}$,. The genome-wide mutation rate μ for $N > 1$ is then dispersed across the N genes. Consequently, we hypothesise that the allele frequency per gene for a genome consisting of N genes would reduce to

$$\varphi_N = \sqrt{1 - e^{-\frac{\mu}{N}}}.$$

It can be immediately deduced that the equilibrium mutation burden is

$$B_N = 2N\varphi_N = 2N\sqrt{1 - e^{-\frac{\mu}{N}}}.$$

	$r = 0$	$r = 1$	Haploid Position Free Model
Average allele frequency φ	$\varphi = \sqrt{1 - e^{-\frac{\mu}{N}}}$	$\varphi < \sqrt{1 - e^{-\frac{\mu}{N}}}$ *(especially for N large)*	$\varphi = \sqrt{1 - e^{-\frac{\mu}{N}}}$
Mean haploid burden B	$B = N\varphi$	$B = N\varphi$	$B = N\varphi$
Variance in haploid burden σ^2	linear in N $\sigma^2 > B$	$\sigma^2 = N\varphi$ $\sigma^2 = B$	$\sigma^2 = N\varphi$ $\sigma^2 = B$
Haploid load class distribution $(c_k)_{k\in\mathbb{N}}$	~~Binomial~~ , ~~Poisson~~, ~~Normal~~	Poi(Nφ)	Poi(Nφ)

Figure 1.14: Comparison of initial quasi-stationary distributions. This overview collects presented information about the haploid load class distribution, its mean, and variance for the three models under consideration: full recombination, no recombination, and the position-free model.

Indeed, the results of our simulation indicate that this is actually the average allele frequency and the mutation burden for $N > 1$. However, only in the absence of recombination ($r = 0$). In contrast, for a fully recombining genome $r = 1$, the allele frequency is significantly lower, and the distance increases with N. Furthermore, φ_N is also the allele frequency that is obtained for the position-free model (see Figure 1.12).

When examining the variance of the haploid mutation burden σ^2 (defined in (3.2)) rather than the average, we observe that the variance for $r = 0$ increases linearly with the recessive gene count and exceeds the expectation. Conversely, the variance is equal to the mean for the fully recombining genome, supporting the hypothesis that the haploid load class distribution $(c_k)_{k\in\mathbb{N}}$ for $r = 1$ follows a *Poisson* distribution. This finding is consistent with the position-free model, where the haploid load class distribution is also a *Poisson* distribution. The average, however, for $r = 1$ is lower than that of the position-free model, as previously stated. Nevertheless, a closed form for the haploid load class distribution for $r = 0$ could not be found. In this case, the distribution is neither *Binomial*, *Normal* nor *Poisson*. It can be observed that as N increases, the right tail becomes heavier, indicating that the fractions of gametes with a high number of mutated genes rises. In particular, when N increases, the distribution deviates from a binomial distribution (see Figure 1.13). The table

(1.14) provides a comprehensive overview of the similarities and differences between the three versions of the model for recessive diseases, including those with and without recombination, as well as the model that do not remember the position of mutations.

The intriguing phenomenon observed in the absence of recombination, as well as the comparison between a fully recombining genome and one that experiences no recombination and only segregation, raises a number of open questions as seen above. A more detailed examination of the click rate and extinction probability would facilitate a more profound understanding of the heterogeneous nature of the rate. It would be valuable to dedicate future research to analysing the case of small recombination rates above zero. In this scenario, two competing rare events occur depending on the mutation and recombination rate. The first is the extinction of the least loaded class due to the accumulation of mutations. The second is the rebirth of the mutation-free class due to recombination of mutation-free gene segments. It would be interesting to observe the dynamics between these two events and whether the rebirth of the mutation-free gamete can stabilise the population before it crystallises into highly correlated clusters.

1.7.3 Simulation framework for dense problems

In Chapter 4, we present a simulation framework that implements a version of Gillespie's algorithm to address the high-dimensional challenges of the models discussed previously. A manual is provided to facilitate the use and adaptation of the core algorithm to a broad range of complex and dense problems. The simulation framework is publicly available as a Julia package,

> L.A. La Rocca. DenseGillespieAlgorithm.jl, 2024
> https://github.com/roccminton/DenseGillespieAlgorithm.jl

Chapter 4 contains the content of the manual, with only minor changes to adapt the layout to the format of this thesis.

Without delving too deeply into the technical specifics, we aim to delineate the principal challenges associated with the implementation of the framework and highlight potential bottlenecks for optimising both computational time and memory usage.

The simulation is governed by two principal functions: the rate function and the affect function. In the context of evolutionary biology, both functions can be understood as comprising two events: birth and death. The rate function calculates the total rate of a birth or death event, while the affect function then executes the event, following the draw of event time from random variables. The rates are calculated after each occurrence of an event, and the affect function is executed for each event. Therefore, during the course of a single simulation, both functions are called many times. It is therefore important to make them as efficient as possible; even minor improvements can significantly impact the total runtime of the algorithm. We encourage every developer to carefully evaluate the performance of these two functions and to carefully consider all performance tips given in the Julia documentation [19]. This entails considering or recalculating the rates from scratch after each event or updating them in accordance with modifications made by the affected function. It is unclear which method

is more efficient, recalculation or updating. Depending on the model, one may outperform the other. Consequently, it is essential to experiment with different functions and benchmark carefully to achieve optimal performance.

In particular, when the traits of individuals become more complex, the utilisation of memory becomes a crucial factor. In some cases, the high complexity of individual traits is necessary for calculating rates and determining the dynamics of the population. However, there is no interest in the exact configuration of every individual over time. In such cases, summary statistics (such as mutation burden and prevalence, as discussed in Section 1.3) provide more insight. It is therefore recommended that these summary statistics be calculated during the runtime of the algorithm, rather than saving the entire population history. It is thus possible to introduce not only custom rates and affect functions (which are mandatory), but also, if desired, custom statistics functions that calculate the required information from the population and save it as a time series for subsequent analysis. It should be noted, however, that in such a case the detailed population history is lost, necessitating the conduct of new simulations should different statistics be required.

In the case of highly complex individual traits, it is also necessary to consider the reuse of memory during the runtime of the simulation. In particular, when the total population size is expected to remain relatively constant, it is more efficient to initialise and store the traits for all individuals in an array and then simply save the indices of individuals that are part of the current population. Then free individual traits can be used and modified as required to match the characteristics of new offspring. This approach also allows for the implementation of high-dimensional models without exceeding the CPU capacity.

One of the key strengths of Julia is its inherent parallel computing capabilities. However, this is not a viable option when performing time series evolution simulations, as the recalculation of rates for subsequent events is contingent upon the occurrence of an initial event. Nevertheless, there are two potential applications of parallel computing in this context that warrant consideration. The first, and arguably less probable, scenario is that the process of updating and recalculating the rates is so computationally intensive that it would be more efficient to divide it into different threads. The second, and arguably more probable, scenario is that it is desired to run multiple simulations. Or, one may wish to compare different runs of the same parameter setting in order to ascertain the impact of randomness and to gain insight into the underlying distribution. Or, the effect of different parameter settings can be evaluated. In both instances, it is advised to utilise distinct kernels and leverage the computational efficiency of Julia in a parallel computing environment.

2 Understanding recessive disease risk in multi-ethnic populations with different degrees of consanguinity

This chapter was published in the American Journal of Medical Genetics as joint work with Julia Frank, Heidi Beate Bentzen, Jean Tori Pantel, Konrad Gerischer, Anton Bovier and Peter M. Krawitz [135],

> L. A. La Rocca, J. Frank, H. B. Bentzen, J. T. Pantel, K. Gerischer, A. Bovier, and P. M. Krawitz. Understanding recessive disease risk in multi-ethnic populations with different degrees of consanguinity. *American Journal of Medical Genetics*, 194(3):e63452, 2024

Population medical genetics aims at translating clinically relevant findings from recent studies of large cohorts into healthcare for individuals. Genetic counseling concerning reproductive risks and options is still mainly based on family history, and consanguinity is viewed to increase the risk for recessive diseases regardless of the demographics. However, in an increasingly multi-ethnic society with diverse approaches to partner selection, healthcare professionals should also sharpen their intuition for the influence of different mating schemes in non-equilibrium dynamics. We, therefore, revisited the so-called out-of Africa model and studied in forward simulations with discrete and not overlapping generations the effect of inbreeding on the average number of recessive lethals in the genome. We were able to reproduce in both frameworks the drop in the incidence of recessive disorders, which is a transient phenomenon during and after the growth phase of a population, and therefore showed their equivalence. With the simulation frameworks, we also provide the means to study and visualize the effect of different kin sizes and mating schemes on these parameters for educational purposes.

2.1 Introduction

Medical population genetics is dedicated to elucidating the role of genomic variation in susceptibility to diseases and requires expertise in medical genetics, population genetics, epidemiological genetics, and community genetics. This knowledge is usually distributed over many teams and labs and rarely integrated within a single institute, let alone a single person [86]. For the following work, therefore, we imagine a reader who is likely to excel in one of these areas but is only familiar with the foundations of others. We hope that the simulation frameworks we present will be so easy to use that many will end up using them to perform further analysis. In the following, we will motivate the choice of our parameter settings that are based on findings that became available due to recent genome-wide sequencing studies.

Sequencing of large cohorts confirmed estimates the number of recessive, lethal equivalents per genome which were previously based on epidemiological data of disease prevalences and stillbirths: On average, healthy individuals carry 0.5 to 2 heterozygous variants that would prevent reproduction if they occurred in a homozygous state [165, 22, 37, 76]. With respect to population genetics, it is irrelevant whether such variants cause a severe, lethal condition in the affected individual before reproductive age or simply result in complete sterility and are therefore also referred to as lethal equivalents. In simulations that aim at reproducing empirical findings, individuals who are homozygous for a lethal equivalent have a fitness of $s = -1$ and are removed from the gene pool. In contrast, heterozygous carriers of lethal equivalents have the same fitness as wildtypes, $s = 0$, and with respect to simulations, modeling the mating pattern is crucial for the dynamics in population genetics. However, the question of how the ancestral background and the degree of consanguinity affect the recessive lethal load per person is still vividly discussed because empirical data and predictions by theoretical population genetics are partially contradictory [13]: in the case of mutation-selection balance, the prevalence of recessive disorders should be the same regardless of ethnicity and mating scheme. However, in the Deciphering Developmental Disorders (DDD) cohort, the proportion of cases due to recessive coding variants was 3.6% in patients of European ancestry, compared to 31% in patients with Pakistani ancestry [153]. Even within the same population, e.g. in Iran, the probability for a recessive cause of intellectual disability is four times higher for offspring from first-cousin unions than for offspring of non-consanguineous partnerships [111, 106, 163]. To explain this discrepancy between the load of recessive lethal variants and the recessive disease burden, some authors recently argued that the unexpectedly high frequency of lethal equivalents might also be explained by an ascertainment bias, that is, some of the pathogenic mutations reached high frequency by chance and are therefore overreported [7]. However, since the assumption of mutation-selection balance is not justified, other authors studied the effect of different demographic dynamics including explosive population growth on mutation burden [103]. Expanding populations incur a mutation burden, also referred to as expansion load, which is a transient phenomenon but can persist for many generations depending on the mating scheme and the coefficients of selection and dominance [92, 178, 12].

In this work, we explore the influence of different mating schemes in nonequilibrium dynamics by means of two different simulation frameworks with distinct and overlapping generations. Each model had the advantage of handling certain aspects of population genetics particularly well. The first is an adaption of the classical Wright-Fischer model with discrete non-overlapping generations run in the forward genetic simulation framework SLiM [99, 68]. In the second model, generations can overlap because diploid individuals die and give birth at independent exponential times on a continuous timescale [7]. For random mating populations with two sexes, the equivalence of the effective population size was already delineated for overlapping generations [62]. In the following, we show that simulations of the discrete, as well as the overlapping model yield comparable results for an out-of-Africa scenario, suggesting that the existing modeling approaches can be used to fit empirical data that result from nonequilibrium dynamics [29].

2.2 Methods

Throughout this framework, the mutation burden is defined as the average number of lethal equivalents per individual. The lethal alleles in the genome are deleterious alleles that are disease-causing if both copies of a gene in an individual harbor at least one such variant. The totality of these pathogenic variants could also be regarded as the theoretical superset of an extended carrier screen [8]. By this means, we are able to focus on the incidence rate of severe recessive disorders with early onset that prevent reproduction almost with certainty. Likewise, we can study how the selection of a partner, which we refer to as a mating scheme, influences the disease prevalence and mutation burden and we are able to monitor these parameters in the population over time. This is done by counting the number of lethal equivalents that enter the gene pool due to a constant *de novo* mutation rate, or leave the gene pool due to selection. If the disease prevalence does not change any more, the population is in a steady state, that is a flux balance for lethal equivalents.

In population genetics, the lifespan of individuals that do not reproduce does not matter. In our simulations we therefore used the same age distribution for every individual, regardless of the number of lethal equivalents or the affection status. With the same life span in affected and unaffected individuals, disease prevalence and incidence are also equivalent and their rate is proportional to the amount of lethal equivalents removed from the gene pool per generation or time unit. In fact, the expected number of lethal equivalents that is lost by an affected individual that is not propagating is two. This is equivalent to the difference in the average mutation burden between affected and unaffected individuals and can also be derived from the simulations. An expansion of the population will affect prevalence and mutation burden as we will discuss in more detail in the following.

Consider a finite population of individuals where each individual is characterized by a diploid set of N gene segments of different sizes. Pathogenic variants appear at every gene independently with a rate that is proportional to its size. As long as an individual carries a pathogenic variant at only one gene, its fitness is unaffected. But as soon as both copies of a gene carry a pathogenic variant, the individual's reproductive fitness is reduced to zero. In this case, the individual will be excluded from the mating process and is not able to reproduce any more. Other than that, all individuals are equally fit, no matter how many recessive disorders they carry. Simulations always start with a small, healthy population. After a period of time in which a mutation selection balance is established, a logistic growth phase starts, that settles after a new population equilibrium is reached. We investigate changes of the dynamics of the mutation burden and the prevalence when the population applies different mating schemes. On one hand, random mating occurs, where individuals select their partner from all potential partners with non-zero fitness uniformly. On the other hand, a consanguineous mating scheme is employed, in which individuals exhibit a preference for mating with close relatives.

2.2.1 Discrete model

In the default setting, the simulation package from Haller and Messer [99] samples a diploid population evolution according to the standard Wright-Fisher model. Sexes were added such

that each sex is equally represented in the population at any time. In generation $n \geq 1$ there is a finite number of individuals $M_n \geq 0$ with a total of $2M_n$ genomes alive. In the initial phase the population size is held constant with $M_n = M_0$ for all generations $n \leq n_{grow}$ in order to establish a mutation selection balance ("burn-in"). Afterwards, the growth phase begins and the population size of each generation grows logistically with growth rate $r > 0$ until it approaches the carrying capacity K. Therefore, the population size of each generation is determined by the following formula

$$M_n = \left\lceil \frac{K}{1 + C_0 e^{-rK(n-n_{grow})}} \right\rceil \quad \text{for all } n \geq n_{grow},$$

where $C_0 = \frac{K-M_0}{M_0}$.

The two mating schemes - random and consanguineous mating - are introduced as following. To generate generation $(n+1)$, first select M_{n+1} females from generation n independently at random with replacement among all females with non-zero fitness. For the random mating scheme, each female then selects a male uniformly at random from the pool of potential partners who possess positive fitness. To implement the consanguineous mating scheme, utilize the pedigree information provided by SLiM for the last two generations, tracing backwards in time. For each individual, their parents and grandparents are known. In the consanguineous population, a female now selects a mate using a weighted uniform distribution from the set of all potential partners. This choice is influenced by weights α, and $\beta \in [0, 1]$ with $\alpha + \beta \leq 1$. The individual then chooses a male partner with non-zero fitness with

$$\begin{aligned} &\text{two common grandparents with probability } \alpha \\ &\text{one common grandparent with probability } \beta \\ &\text{no common grandparents with probability } 1 - (\alpha + \beta) \end{aligned}$$

Notice that having two grandparents is akin to a cousin relationship, while sharing one grandparent relates to a half-cousins relationship, as depicted in Figure 2.4A.

To start the simulation select N gene segments from the entire human genome. Each with an independently and uniformly distributed number of base pairs $w_1, \ldots, w_N \sim \mathcal{U}_{[a,b]}$, where $a, b > 0$, representing the minimum and maximum segment size, respectively. Furthermore, the entire genome is divided into n_c chromosomes. During birth, changes in the offspring's genetic information occur not only through mutation but also via recombination. For each chromosome, initiate an independent *Poisson* Process with rate $r_{\text{rec}} > 0$, which identifies the recombination breakpoints. Here r_{rec} represents the overall recombination rate. The discrete model was implemented in SLiM version 3.2.1.

2.2.2 Adaptive dynamics

We employ a diploid version of the adaptive dynamic models introduced by Fournier and Collet [72, 49]. A distinct characteristic of these models lies, firstly, in their foundation on a Poisson process. This entails that individuals produce offspring and undergo mortality at independent rates. Secondly, a noteworthy feature is the ongoing feedback between demographics and ecology due to the competition among individuals. This competitive pressure

for finite resources among individuals enables the modelling of a naturally fluctuating population with limited capacity. In the following, we outline the key features of the model. For a comprehensive mathematical description, please refer to the Appendix. We initiate the simulations within a small, entirely healthy population. This population not only settles into a mutation-selection equilibrium but also experiences fluctuations around a natural population size. This size is contingent upon birth and death rates, as well as the interplay of competition among individuals and the mutation rate. Following the initial burn-in phase, we decrease competition, thereby providing the population with more resources. This alteration triggers logistic population growth until the growth rate tapers off upon reaching the new population equilibrium. To simulate consanguineous mating, we equip each individual additionally to the genetic information with two family flags, aimed at indicating the origin of the individual. During each birth, the newly born individual inherits one randomly chosen family flag from each parent. If both parents possess the same family information, the offspring inherits an identical copy of this information. This modelling approach presents several challenges. Firstly, we must ensure that family groups do not become too large and should periodically disintegrate once the maximum family size of κ is reached. Secondly, this identification mechanism only partially mirrors actual families. For instance, in this model, it's possible that grandparents and their grandchildren do not belong to the same family. In the random mating scheme, individuals select partners randomly from the pool of fit individuals. On the other hand, in the consanguineous mating scheme, partner selection depends on family affiliation. We model the reproductive compatibility between two individuals such that, in an equilibrium population, the probability of selecting a partner with the same family flags is α as long as the family size fluctuates around $\kappa/2$. Conversely, the probability of selecting a partner who shares only one of the family flag with oneself is β. Finally, a partner outside the family is chosen with a probability of $1 - \alpha - \beta$. This holds assuming the population is in equilibrium and the relevant family has a size of $\kappa/2$. During each birth, a Poisson-distributed number of pathogenic variants is randomly distributed across the $2N$ gene segments. The pathogenic variants are allocated to the N genes using a weighted uniform distribution, where the weights correspond to the respective sizes of the genes. Each mutation contributes to the degeneration of the gene segment. There are no back mutations, beneficial mutations, or neutral mutations in this scenario. Instead of recombination, we employ a form of genetic information reshuffling. During each gamete formation, the genetic information is divided into n_c chromosomes, and from these, one copy is randomly selected. We have implemented the simulations in Python version 3.8 using a Gillespie algorithm.

2.2.3 Comparing both models

Both models, the discrete generation model implemented with SLiM and the adaptive dynamics model using the Gillespie algorithm, excel in different aspects of capturing nature. A prominent advantage of SLiM and the discrete model lies in the precise pedigree information generated for every individual. However, the adaptive model can only roughly cluster individuals into family groups and cannot differentiate among members within a single family, as depicted in Figure 2.4B. Nonetheless, a significant drawback of the discrete model is its non-overlapping generations. This limitation precludes the possibility of matings

between individuals on different pedigree levels, such as uncle-niece marriages. This constraint is overcome by the continuous-time model. As individuals independently give birth and die, different generations coexist due to varying ages. The discrete model, similar to the Wright-Fisher model, operates with constant or deterministically increasing population sizes. Conversely, the continuous model accommodates a fluctuating and naturally growing population, as depicted in Figure 2.5. It is worth noting that for large populations, the random fluctuations in population size are of order $\frac{1}{K}$, and the stochastic process converges in law to the solution of a deterministic logistic equation [72]. Recombination is also approached differently in the two models. SLiM operates with genuine interchromosomal recombination, while the adaptive model simply reshuffles parental chromosomes during gamete production. This distinction arises from SLiM's ability to record the precise base positions of mutations on the human genome. In contrast, the adaptive dynamics model possesses information only about the number of pathogenic variants per gene segment and lacks knowledge of their exact locations within each segment. Given the assumption that all genes are compound heterozygotes, the varying implementations of recombination do not impact the fitness of individuals. However, this reduction in information brings a significant advantage in terms of algorithm runtime.

Apart from all the differences outlined, a substantial effort has been made to ensure parameter equality between both simulations. This includes factors such as the number of gene segments N, the initial and equilibrium population size M_0 and K, and numerous other parameters. Additionally, in the continuous model, family sizes are calibrated to attain an approximate balance between the number of potential partners in the consanguineous setting of both models. Similarly, the birth rate in the continuous-time model is established at $b = 1$, ensuring that within a time interval of $t \in [n, n+1]$, corresponding to one discrete generation, there are M_{t+1} birth events, where M_t denotes the population size at that particular time. The only distinction lies in the discrete generation model having *exactly* M_{n+1} births in generation n, whereas the continuous-time model experiences *on average* that number of births.

2.3 Results

We initiate our simulations with a population of 500 individuals, allowing for approximately 500 generations to reach a steady state, that is no significant change in the mutation burden. A comparable size has also been suggested for the population that left the African continent 10,000 to 200,000 years ago [95, 199]. Following this out-of-Africa event, the population expands to a size of 10,000 individuals in approximately 130 generations. This corresponds to an estimated duration of around 2,500 years and an average growth of 1-2% per generation. The population expansion adheres to a logistic growth curve, which takes on the appearance of a step function (as depicted by the grey curve in Figure 2.1), due to the extensive duration of our complete simulations spanning 2,000 generations. All individuals have diploid genomes with 1,000 recessive genes that we deem crucial for reproductive success. Their coding sequence ranges between 500 and 10,000 base pairs (bp) per gene, novel alleles are introduced with a *de novo* mutation rate of 1.2×10^{-8} per bp, and one out of nine mutations is expected to be a lethal equivalent [130, 129]. The choice of these parameters are motivated

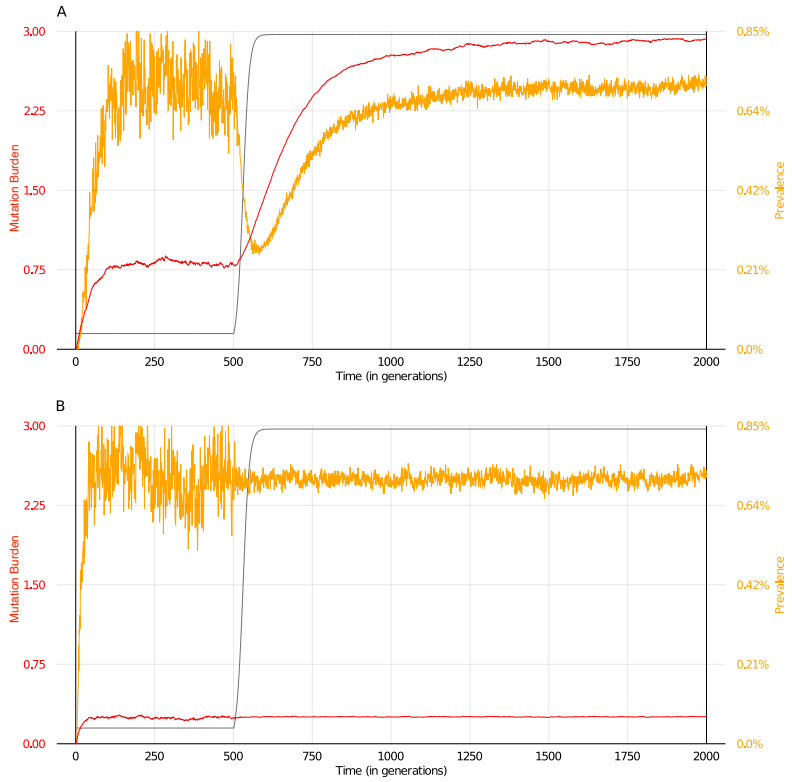

Figure 2.1: Dynamics of mutation load and prevalence for severe recessive disorders: A population expansion from 500 to 10 000 individuals (grey), starting in generation 500 does not affect prevalence (orange) nor mutation load (red) if partners are preferentially chosen within relatives (consanguineous mating scheme) (B). In contrast, in a random mating population, there is a transient drop of prevalence at the expense of an increasing mutation load (A). It takes more than 550 generations after the end of the growth phase, until the steady state is reached and the prevalence for both mating schemes are comparable again. The plots show the average of 50 exact trajectories of the stochastic process simulated with the Wright-Fisher model.

by the distribution of coding lengths and the deleteriousness scores for known autosomal recessive genes [119, 122]. Pairs for procreation are formed either randomly or based on their relatedness that is traced over the two most recent generations. In a highly consanguineous mating scheme, the number of potential partners is hardly affected by the population size, as most marriages happen within families. In our simulations, this mating scheme is realized as follows: 50% of all partnerships share two grandparents, 30% share one grandparent, and only 20% share no grandparent. In this scenario, the mutation burden and prevalence do not change during population growth (Figure 2.1 B). However, linkage disequilibrium suggests that out-of-Africa populations have only reached effective populations sizes of around 3k, thus this might be an overestimate [199]. In contrast, in a randomly mating population, there is a sharp transient drop of incidence rates during expansion at the expense of an increasing mutation burden (Figure 2.1 A). However, after the final size of the population is reached, it takes almost another 550 generations until the mutation burden reaches its

new plateau of approximately three pathogenic variants in 1,000 recessive disease genes. In contrast to the mutation burden, the prevalence is independent of effective population size and a function of mutation rate only. For constant μ, the prevalence returns to the initial value before the expansion. Since affected individuals in our simulation have the same life expectancy and only do not propagate, prevalence and incidence are the same and there are roughly 70 affected individuals per generation in a population of 10,000 or 0.7%.

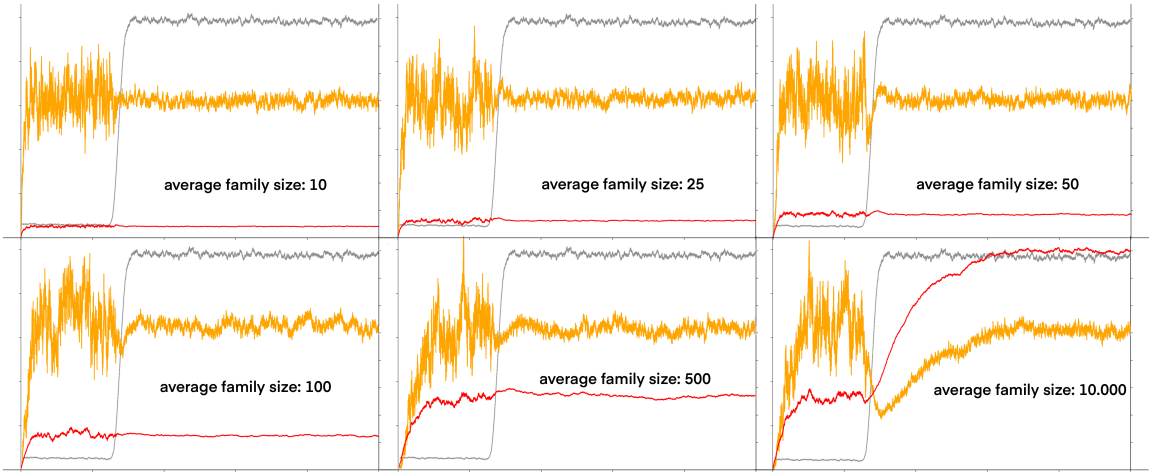

Figure 2.2: Influence of family size on mutation load and prevalence: The mating scheme is characterized by the family size and a probability function that describes how many of the partners are chosen within the family. In a preferentially consanguineous mating population the dynamics change when the maximum family size increases (upper left panel to lower right from 10, 25, 50, 100, 500 up to 10 000). The mutation load starts to increase considerably if mating is happening in tribes of 500 individuals. However, at this stage there is still only a minor effect of further population growth. In the lower right the maximum of the allowed family size is equivalent to the population size and thus, dynamics do not differ from a random mating scheme any more. The plots show the average of 10 exact trajectories of the stochastic process simulated with the individual-based model of adaptive dynamics model.

The mutation burden in the steady state increases in both mating schemes with the number of autosomal recessive genes, but with population size only for random mating (Figure 2.3 A,B). This is best explained by a limit of the effective number of available partners that the consanguineous mating scheme imposes, regardless of the final population size. In line with that argument, there is a transition from the dynamics of consanguineous to random when we incrementally increase family size, which would correspond to more potential mating partners (Figure 2.3). Although the phase of population growth lasts only 130 generations in our simulations, the time span to reach the new equilibrium for the mutation burden lasts much longer. In both simulation frameworks, we were able to achieve numbers of lethal equivalents that are in accordance with observations from the literature that are based on epidemiological data as well as population genetic data. In a recent study, Narasimhan et al. analyzed exomes of 3222 British adults of Pakistani heritage with a high parental relatedness and found a significantly lower number of homozygous knockout genotypes than expected from the summary statistics of a more outbred population. By this means, they were able

to compute an average number of 1.6 recessive-lethal equivalents per individual [165]. In mutation-selection balance, the number of recessive-lethal equivalents is only a function of genome architecture and the effective population, which the mating scheme influences. In the non-equilibrium dynamics, however, the choice of the partner has the greatest influence on the increase of recessive lethal equivalents. Since human societies almost mirror the unmanageable variety of mating systems in the mammalian kingdom it is noteworthy that with the discrete and adaptive simulations, different aspects of mating can be modeled [48]. In the adaptive framework, for example, we allowed partnerships between different generations and for each offspring the parents were selected anew (lottery polygyny) [33]. Despite the differences in the implementation details, both simulations yielded comparable dynamics when the extended family size κ and the autozygosity were adjusted. Over certain historic periods, the extended family size κ, which was the parameter used in the adaptive model, might be easier to delineate. Whereas kinship coefficients could be estimated with exact pedigrees and genomic data. We therefore extended the possibilities of how empirical data can be explained by population genetic simulations.

2.4 Discussion

The empirical observation that consanguinity is associated with an increased risk of autosomal recessive disorders, has been made in many countries but are only based on records of relatively few generations. Martin et al. showed that the contribution of autosomal recessive developmental disorders is 31% in the current British population if the autozygosity is above 0.02 [153]. Likewise, in the Iranian population it is estimated that offspring from first-cousin unions have a probability for intellectual disabilities that is four times higher than in non-consanguineous partnerships [111, 106, 163]. Although population genetics predicts these findings as a transient phenomenon in nonequilibrium dynamics, this literature is often not cited in the empirical works [103, 92, 178, 12, 120, 146, 87, 190]. In our work we studied how rapid changes in population size affect the expected number of lethal equivalents when generations overlap, and achieved similar results as in the Wright-Fisher model. By that means we addressed an outstanding question in nonequilibrium population genetics. We hypothesize that epidemiological data accumulated over a few centuries, which is a short time period with respect to recessive selection and a lack of knowledge in population genetics, might frame a biased risk perception that might even influence aspects of social norms. According to our simulations and previous work, the advantage of outbreeding is a transient phenomenon for a population that is initially in mutation-selection balance and that starts to grow. The lower prevalence compared to an inbred population lasts for many generations even after the expansion phase has ended, until mutation-selection balance is reached again with a higher count of lethal equivalents. We found it intriguing that e.g. first-cousin marriage in Europe was banned after several generations of population growth during the Roman empire and considerable migration and admixture [103]. While this continent clearly benefitted at that time point from a change of social conventions with respect to the recessive disease burden, the consequences of different mating schemes e.g. on the proportion of

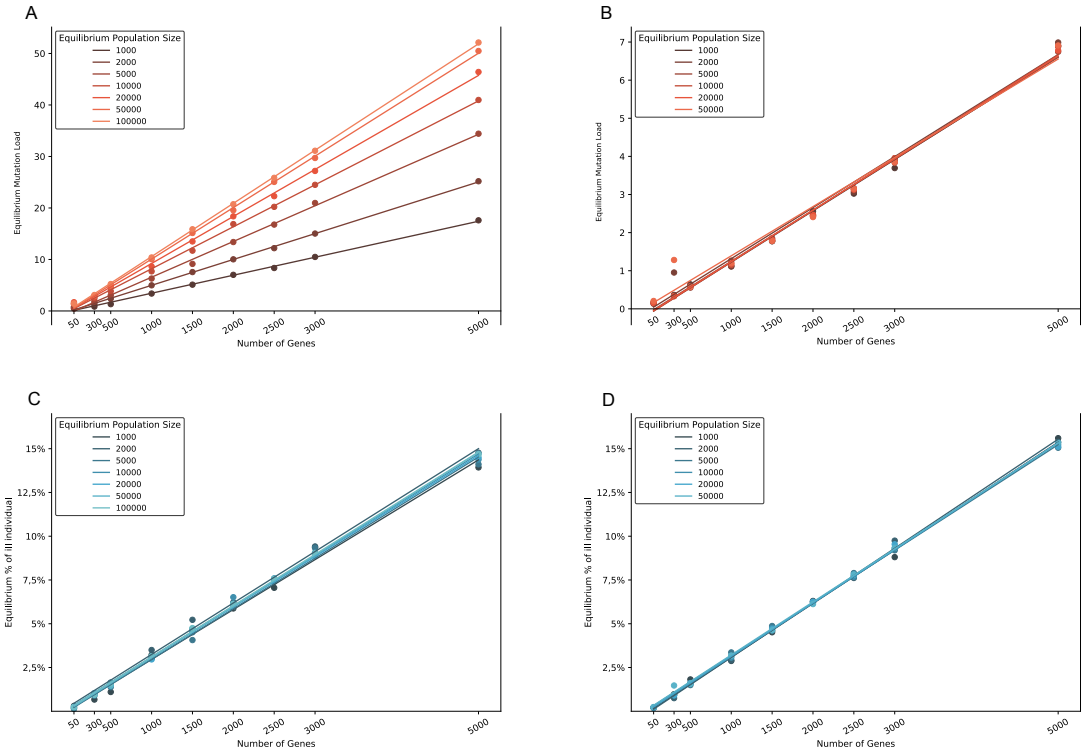

Figure 2.3: Influence of genomic architecture and population size: The capacity of the genome for deleterious mutations is larger in the random mating population. With an increasing number of genes and growing population size, deleterious mutations accumulate (A). In contrast, in the consanguineous mating scheme, family size limits the effective population size, and therefore mutation load is independent of the total number of individuals (B). Prevalence increases linearly in both mating schemes when the number of genes increases and is independent from population size, as regression analysis indicates (C,D).

congenital malformation are less prominent in populations that were more constant in size over a long period of time [37]. One of the most extreme examples of descendants of a small group might be the Hutterites, who increased in population size by more than a factor of 400 in less than 200 years from a founding population of less than 100 people [28]. This is comparable to a kin of 100 which is still too small to benefit from a drop in prevalence during growth as shown in Figure 2.2. The few initial lethal equivalents of the founders were amplified to high prevalence and are now also listed as recessive alleles of high frequency in the database of genetic disorders in Amish, Mennonite and Hutterite [177]. However, a transient reduction of recessive disease burden can be achieved by marriage that is colony exogamous, which is also most likely for that reason a social accepted mating scheme. The occurrence and coexistence of different marriage patterns over many centuries can certainly not be understood by population genetics alone since social, cultural and economic factors interact with demographics in a complex manner [103]. It is therefore concerning when

questionable genetic reasoning is used in the legislature. For instance, the European Court of Human Rights case of *Stübing v. Germany* concerned consanguineous siblings who had four children following consensual intercourse, whereupon both siblings were charged with incest [2]. One of the siblings lodged a complaint, arguing that the legislature violated his right to sexual self-determination, his private and family life. The Court found that 24 out of 44 European States reviewed, criminalized consensual sexual acts between adult siblings, and all prohibited siblings from getting married. The German government argued that the law against incest partly aimed to protect against the significantly increased risk of genetic damage among children from an incestuous relationship [32]. However, motivating a law to avoid a higher probability of disease can be viewed as eugenic: As the German Ethics Council opined after the judgement, no convincing argument can be derived from there being a risk of genetic damage [3]. The Council also pointed out that prohibiting procreation in non-consanguineous couples who carry a genetic burden, would not be allowed to be proposed or considered in any manner [3]. Any prohibition of consanguineous relationships should therefore build on non-genetic reasoning. The view of the German Ethics Council concurs with a statement vy the German Society of Human Genetics criticizing eugenic reasoning in a judgement by the German Federal Constitutional Court in 2008 on criminal liability of incest between siblings. The Society stated that "The argument that reproduction needs to be thwarted in couples whose children possess an elevated risk for recessively inherited illnesses is an attack on the reproductive freedom of all." ["Das Argument, es müsse in Partnerschaften, deren Kindern ein erhöhtes Risiko für rezessive erbliche Krankheiten haben, einer Fortpflanzung entgegengewirkt werden, ist ein Angriff auf die reproduktive Freiheit aller".] [1]. The Society added that apart from being factually incorrect, eugenic reasoning also encourages discrimination and should therefore be avoided by the courts [1].

Furthermore, as our work shows, the argument that there exists an increased risk of genetic damage requires the definition of a reference population for comparison. However, there is neither agreement about a suitable reference nor an accurate measurement for mutation burden [103]. When genetic counseling is sought, the predicted recessive disease burden that is communicated in the consultation might influence decisions e.g. about the choice of partners or family planning. Since this risk does not only depend on mating schemes but also on mutation burden it is important to measure this parameter as accurately as possible. In our simulations, an individual of the outbreed population had on average four times more lethal equivalents than an individual of the inbred population when the mutation-selection balance was reached again many generations after the growth phase ended.

Interestingly, these values and the range are comparable to what has also been described in the literature for real populations. With respect to the British subpopulations of Pakistani (PABI) and European (EABI) ancestry in Martin et al., this could mean that PABI with a considerably higher autozygosity and many first-cousin marriages are closer to mutation-selection balance than EABI. This would imply that the disease prevalence for recessive disorders will remain constant for PABI while it will approach that level for EABI in the following generations, given that the different mating schemes continue. In contrast, the higher mutation burden in the EABI subgroup due to the higher effective population size might already now contribute to a higher risk for autism spectrum disorders, which are also highly heritable but do not follow monogenic inheritance [108]. Since assessing recessive lethals based on family history is very challenging, genetic counseling should increasingly focus on

carrier testing in cases where individuals seek help to gain information to make their own decisions. Based on current ClinVar statistics, there are more than 150,000 pathogenic alleles known for recessive genes that cause severe disorders. In large German cohort of individuals with rare disorders, a diagnosis could be established in 125 cases due to homozygosity or compound heterozygosity of pathogenic variants. 94 of these causative variants would also have been classified as pathogenic in the healthy parents in a preconceptional exome analysis [188]. Expanded carrier screens can play an important role in genetic counselling in multi-ethnic populations with different degrees of consanguinity, and it should be discussed who should have access to this test to make their own informed decisions [189, 8].

2.5 Code availability

All scripts to reproduce our simulation results can be found in the following repository:

`https://github.com/roccminton/Diploid_Model_Two_Loci`

A video clip of our simulations can be found at:

`https://youtu.be/5hOgLyRqWPg`

2.6 Appendix

2.6.1 Adaptive dynamics model

The adaptive dynamics model is continuous in time, hence time is not measured in $n \in \mathbb{N}$ discrete generations, but on the positive real axis $t \in \mathbb{R}_+$. No exact pedigree are available for this continuous model, therefore introduce a new diploid family trait, which indicates the ancestry of an individual. Hence every individual is characterized by two diploid traits. The first refers to the family origin whereas the second gives insight in the genetical information of the individual. Introduce $\mathcal{F} \subset \mathbb{N}$ as the finite set of all possible family traits and $\mathbf{f} = (f^1, f^2) \in \mathcal{F}^2$ being the family trait of an individual in the current population. Moreover the diploid genetic information of an individual is a finite vector

$$\mathbf{x} = (x_1^1, \dots, x_1^N, x_2^1, \dots, x_2^N) \in \mathbb{N}_{\geq 0}^{2N}$$

where the entries x_1^i and x_2^i represent the number of pathogenic variants at the i^{th} gene segment in the first and second genome. Thus if there are $(\mathbf{f}_1, \mathbf{x}_1), \dots, (\mathbf{f}_{M_t}, \mathbf{x}_{M_t})$ individuals alive at time $t > 0$ in an arbitrary order, define the population state as a point measure on $\mathcal{X} := \left(\mathcal{F}^2 \times \mathbb{N}^{2N}\right)$

$$\nu_t(\cdot) := \sum_{i=1}^{M_t} \delta_{(\mathbf{f}_i, \mathbf{x}_i)}(\cdot)$$

Individuals give birth and die at exponential rates $b(\mathbf{f},\mathbf{x})$ resp. $d(\mathbf{f},\mathbf{x})$ which depend on the family trait and the chromosomal configuration of the individuals. One can think of every individual carrying two independent clocks, a birth and a death clock. If the birth clock rings first the individual makes its mating choice, reproduces and resets its clock. Whereas if the death clock rings the individual disappears from the population. Additional to the intrinsic death rate every individual sense the competition pressure of every other individual in the population. The term $C(\mathbf{f},\mathbf{x},\mathbf{g},\mathbf{y})$ gives the competition pressure executed by an individual of type $(\mathbf{g},\mathbf{y})$ and felt by an individual of type $(\mathbf{f},\mathbf{x})$. Hence the total death rate of an individual in population ν is increased by the term

$$\int_{\mathcal{X}} C(\mathbf{f},\mathbf{x},\mathbf{g},\mathbf{y})d\nu(\mathbf{g},\mathbf{y})$$

When an individual gives birth it chooses a partner at random from the population according to the partners birth rate and the reproductive compatibility between them. Notice that the reproductive compatibility $R_{\mathbf{f}}(\mathbf{g}) \in [0,1]$ of two individuals depends only on their family traits $\mathbf{f}$ and $\mathbf{g}$. After a mate was chosen the newborns family trait will be a uniform random combination of the four parental family traits unless both parents have the same traits. In this case the child inherits the exact same couple of traits. The genetic configuration of the newborn is not only a random combination of the parental alleles since the effect of mutation and the reshuffling of the parental chromosomes come into play. For an accurate definition of the reshuffling of the diploid parental chromosomes to a mixed haploid set, define first the sections on the genetic information, that form chromosomes. Let $n_c \in \mathbb{N}$ be the number of chromosomes for every individual. Introduce the chromosome breakpoints $\{c_1, c_2, \ldots, c_{n_c+1}\} \in \{1,\ldots,N\}$ with $0 = c_1 < c_2 < \cdots < c_{n_c} < c_{n_c+1} = N$. Divide the genetic information of every individual $\mathbf{x} = (x_1^1,\ldots,x_N^1,x_1^2,\ldots,x_N^2)$ into chromosomes of the same length in both copies

$$\mathbf{x} = (\underline{x}_1^1,\ldots,\underline{x}_{n_c}^1,\underline{x}_1^2,\ldots,\underline{x}_{n_c}^2) \quad \text{with} \quad \underline{x}_i^j = \left(x_{c_i+1}^j, x_{c_i+2}^j,\ldots,x_{c_{i+1}}^j\right)$$

for $j \in \{1,2\}$ and $i \in \{1,\ldots,n_c\}$. Finally get the reshuffled gamete from $\mathbf{x}$ via a selection variable $\tau : \{1,2,\ldots,n_c\} \longrightarrow \{1,2\}$ as follows

$$\mathbf{x}^\tau := \left(\underline{x}_1^{\tau(1)},\ldots,\underline{x}_{n_c}^{\tau(n_c)}\right)$$

Selecting τ among all possible assignments equals an uniform recombination of the diploid chromosomes into haploid set. At each birth a *Poisson* distributed number of pathogenic variants is added to every offspring. The expectation of these identically distributed and independent *Poisson* random variables equals the total mutation rate $2\mu\bar{w}$, where μ is the mutation rate per base pair and $\bar{w} = w_1 + \cdots + w_N$ is the sum of the length of the gene segments under consideration. After sampling the number of pathogenic variants per birth, these variants are distributed equally on the $2N$ gene segments according to their length. Finally a pathogenic variant at a given gene increases the genetic value by one. All of this is captured in the mutation with recombination operator $M^{\mathtt{rec}}_{\mathbf{f},\mathbf{x},\mathbf{g},\mathbf{y}}$ for a mating between

individuals $(\mathbf{f},\mathbf{x})$ and $(\mathbf{g},\mathbf{y})$

$$\left(M^{\mathtt{rec}}_{\mathbf{f},\mathbf{x},\mathbf{g},\mathbf{y}}\phi\right)(\nu)=\begin{cases}\dfrac{1}{2^{2n_c}}\displaystyle\sum_{\tau,\tau'\in\{1,2\}^{n_c}}\int_{\mathcal{X}}\left(\phi\left(\nu+\delta_{\mathbf{f},(\mathbf{x}^{\tau},\mathbf{y}^{\tau'})+h}\right)-\phi(\nu)\right)m\left((\mathbf{x},\mathbf{y}),dh\right) & \text{,for } \mathbf{f}=\mathbf{g}\\ \dfrac{1}{2^{2n_c+2}}\displaystyle\sum_{\substack{i,j\in\{1,2\}\\ \tau,\tau'\in\{1,2\}^{n_c}}}\int_{\mathcal{X}}\left(\phi\left(\nu+\delta_{f_i,g_j,(\mathbf{x}^{\tau},\mathbf{y}^{\tau'})+h}\right)-\phi(\nu)\right)m\left((\mathbf{x},\mathbf{y}),dh\right) & \text{,else}\end{cases}$$

and the mutation measure m is defined as

$$m\left((x,y),dh\right):=\sum_{k=0}^{\infty}\left(\frac{(2\mu\bar{w})^k e^{-2\mu\bar{w}}}{k!}\frac{1}{Z_k}\sum_{l\in\diamondsuit_k^{2N}}\left(\sum_{i=1}^{N}\left(w_i^{l_i}\mathbb{1}_{\{l_i>0\}}+w_i^{l_{i+N}}\mathbb{1}_{\{l_{i+N}>0\}}\right)\right)\delta_l(dh)\right)$$

where $\diamondsuit_k^{2N}:=\left\{l\in\mathbb{N}_+^{2N}\mid l_1+\cdots+l_{2N}=k\right\}$ is the set of all lattice vectors in $\mathbb{N}_+^{2N}$ with one norm equal to k moreover $Z_k>0$ is a normalizing constant depending on the size of $\diamondsuit_k^{2N}$. Let

$$\mathcal{M}(\mathcal{X})=\left\{\sum_{i=1}^{n}\delta_{\mathbf{f}_i,\mathbf{x}_i}\colon n\geq 0,(\mathbf{f}_1,\mathbf{x}_1),\ldots,(\mathbf{f}_n,\mathbf{x}_n)\in\mathcal{X}\right\}$$

be the set of all finite point measures on $\mathcal{X}$. The dynamics of the continuous time, $\mathcal{M}(\mathcal{X})$-valued jump process $(\nu_t)_{t\geq 0}$ can be described by the generator $\mathcal{L}$ defined for any bounded measurable function $\phi\colon\mathcal{M}(\mathcal{X})\longrightarrow\mathbb{R}$ as

$$\begin{aligned}\mathcal{L}\phi(\nu)=&\int_{\mathcal{X}}b(\mathbf{f},\mathbf{x})\left(\int_{\mathcal{X}}\frac{b(\mathbf{g},\mathbf{y})R_{\mathbf{f}}(\mathbf{g})}{\langle\nu,b\cdot R_{\mathbf{f}}\rangle}\left(M^{\mathtt{rec}}_{\mathbf{f},\mathbf{x},\mathbf{g},\mathbf{y}}\phi\right)(\nu)d\nu(\mathbf{g},\mathbf{y})\right)d\nu(\mathbf{f},\mathbf{x})\\ &+\int_{\mathcal{X}}\left(d(\mathbf{f},\mathbf{x})+\int_{\mathcal{X}}C(\mathbf{f},\mathbf{x},\mathbf{g},\mathbf{y})d\nu(\mathbf{g},\mathbf{y})\right)\left(\phi\left(\nu-\delta_{(\mathbf{f},\mathbf{x})}\right)-\phi(\nu)\right)d\nu(\mathbf{f},\mathbf{x})\end{aligned}$$

Assume the boundedness of the birth and death rates b and d as well as the boundedness of the competition kernel C. Starting in a initial state $\nu_0\in\mathcal{X}$ such that

$$\mathbb{E}\left[\nu_0,1\right]<\infty$$

existence and uniqueness in law of the process with infinitesimal generator $\mathcal{L}$ and initial condition ν_0 can be derived from [72].

Keep families small To ensure a stable mating scheme during the evolution of the population split families which became too big up into two subfamilies. Therefore introduce the following sequence of stopping times.

$$\begin{aligned}\theta_0&:=0\\ \theta_{k+1}&:=\inf\left\{t>\theta_k:\exists\mathbf{f}\in\mathcal{F}^2\text{ s.th. }\langle\nu,\mathbb{1}_{\mathbf{f}}\rangle>\kappa\right\}\end{aligned}$$

for some fixed $\kappa > 0$. Notice that at any stopping time it is always a unique family $\mathbf{f} \in \mathcal{F}^2$ that exceeds the maximum family size, since at any time there is at most one individual entering or exiting the population. At these random times the big family is split up at random into two subfamilies, where the size of each subfamily is binomial distributed with mean $\frac{1}{2}$. One family keeps the old family trait and the other one gets a completely new, homogeneous one. To make this precise associate a number to each individual in a family. Therefore let $H^{\mathbf{f}} = (H_1^{\mathbf{f}}, H_2^{\mathbf{f}}, \dots, H_k^{\mathbf{f}}, \dots) : \mathcal{M}(\mathcal{X}) \longrightarrow (\mathbb{N}^{2N})^{\mathbb{N}}$ defined by

$$H^{\mathbf{f}}\left(\sum_{i=1}^{n} \delta_{\mathbf{g}_i, \mathbf{x}_i}\right) = \left(\mathbf{x}_{\sigma(1)}, \mathbf{x}_{\sigma(2)}, \dots, \mathbf{x}_{\sigma(k)}, 0, 0, 0, \dots\right)$$

where the $\mathbf{x}_i$ for $i = 1, \dots, k$ are the genetic configurations of the individuals in $((\mathbf{g}_1, \mathbf{x}_1), \dots, (\mathbf{g}_n, \mathbf{x}_n))$ with $\mathbf{g}_i = \mathbf{f}$ and where $\mathbf{x}_{\sigma(1)} \preceq \cdots \preceq \mathbf{x}_{\sigma(k)}$ is the lexicographical order $\preceq$ on $\mathbb{N}^{2N}$ and $k = \langle \nu, \mathbb{1}_{\mathbf{f}} \rangle$ is the family size of $\mathbf{f}$. Then the splitting of a family with trait $\mathbf{f} \in \mathcal{F}^2$ can be expressed with the following operator for any bounded measurable function $\phi \colon \mathcal{M}_F(\mathcal{X}) \longrightarrow \mathbb{R}$

$$\left(S_{\mathbf{f}}\phi\right)(\nu) \coloneqq \frac{1}{2^{\langle \nu, \mathbb{1}_{\mathbf{f}} \rangle}} \sum_{\pi \in \{0,1\}^{\langle \nu, \mathbb{1}_{\mathbf{f}} \rangle}} \left(\phi\left(\nu + \Delta_{\nu,\mathbf{f}}(\pi)\right) - \phi(\nu)\right)$$

where $\Delta_{\nu,\mathbf{f}}(\pi)$ executes the splitting of the family $\mathbf{f}$ in ν into two with configuration π, hence

$$\Delta_{\nu,\mathbf{f}}(\pi) \coloneqq \sum_{i=1}^{\langle \nu, \mathbb{1}_{\mathbf{f}} \rangle} \pi(i) \left(-\delta_{\mathbf{f}, H_i^{\mathbf{f}}} + \delta_{\mathbf{f}^{\mathtt{new}}, H_i^{\mathbf{f}}}\right) d\nu(\mathbf{f}, \mathbf{y})$$

where $\mathbf{f}^{\mathtt{new}} = (f^{\mathtt{new}}, f^{\mathtt{new}})$ with $f^{\mathtt{new}} \in \{g \in \mathcal{F} \mid \langle \nu, \mathbb{1}_{(g,g')} \rangle = 0, \forall g' \in \mathcal{F}\}$ chosen deterministically, is a homogeneous family trait which is entirely new to the population. A possible way of choosing the new family trait at time θ_k is to set $f^{\mathtt{new}} = n_{\mathcal{F}} + k$ where $n_{\mathcal{F}}$ is the number of different families in the initial population ν_0. Hence the dynamics of the evolutionary process with splitting is given as the solution of the following martingale problem. Let $\nu_0 \in \mathcal{M}(\mathcal{X})$ be a initial population then for any real, continuous, bounded function ϕ on $\mathcal{M}(\mathcal{X})$ the process

$$M_t^{\phi} = \phi(\nu_t) - \phi(\nu_0) - \left(\int_0^t \mathcal{L}\phi(\nu_s) ds + \sum_{\mathbf{f} \in \mathcal{F}^2} \left(S_{\mathbf{f}}\phi\right)(\nu_t) \mathbb{1}_{\langle \nu_t, \mathbb{1}_{\mathbf{f}} \rangle > \kappa}\right)$$

is a martingale.

Choices of parameters Introduce the subset $\mathcal{D}_N \subseteq \mathcal{X}$ of traits having at least one pathogenic variant in the same gene segment on both copies of the chromosome as

$$\mathcal{D}_N \coloneqq \left\{ (\mathbf{f}, \mathbf{x}) \in \mathcal{X}^2 \,\middle|\, \exists n \in \{1, \dots, N\} \text{ s.th. } x_n^1 > 0 \text{ and } x_n^2 > 0 \right\}$$

and set the birth and death rate to constant unless an individual falls in the set of non-propagable types

$$b(\mathbf{f},\mathbf{x}) = \bar{b}\mathbb{1}_{(\mathbf{f},\mathbf{x})\notin\mathcal{D}_L} \quad , \quad d(\mathbf{f},\mathbf{x}) = \bar{d}$$

for $\bar{b} > \bar{d} > 0$. To ensure uniform competition among all individuals set the competition pressure constant to

$$C(\mathbf{f},\mathbf{x},\mathbf{g},\mathbf{y}) = \frac{\bar{b}-\bar{d}}{K}$$

for all individuals $(\mathbf{f},\mathbf{x})$ and $(\mathbf{g},\mathbf{y})$. Therefore the population size in equilibrium fluctuates around the carrying capacity K of the system. The different mating schemes are defined as follows. First the random mating, where $R^{\texttt{rnd}}_{\mathbf{f}}(\mathbf{g}) = 1$ for all $\mathbf{f},\mathbf{g} \in \mathcal{F}^2$ and the consanguineous mating scheme with

$$R^{\texttt{cng}}_{(f_1,f_2)}(g_1,g_2) \coloneqq \begin{cases} \frac{2\alpha}{\kappa} & \text{if } (f_1,f_2) = (g_1,g_2) \\ \frac{2\beta}{\kappa} & \text{if } f_i \in \{g_1,g_2\} \text{ for } i=1 \text{ or } i=2 \\ \frac{1-\alpha-\beta}{K-\frac{\kappa}{2}} & \text{else} \end{cases}$$

where $\alpha,\beta \in [0,1)$ and $\kappa > 0$ is the maximum family size. Therefore the probability of mating within their own family that is of size $\frac{\kappa}{2}$ in a population that is at its stable equilibrium size K is constant α. Note that for families with family size smaller than $\frac{\kappa}{2}$ the probability of mating within the family is slightly lower, whereas it gets bitter when the family size surpasses this size. Furthermore the probability increases at the beginning of the growth phase when the carrying capacity K gets uplifted and the population size starts to grow slowly. During this initial phase of expansion there will be more consanguineous mating overall. This imbalance levels off as soon as the population size approaches K.

Start the evolution with a small, healthy population of size $M_0 > 0$ in population equilibrium, where nobody carries any pathogenic variant. The clonal individuals are following divided into $n_{\mathcal{F}} \in \mathbb{N}_{>0}$ families with homogeneous family traits $(1,1),(2,2),\ldots,(n_{\mathcal{F}},n_{\mathcal{F}})$. After an initial phase during which a mutation selection balance is established raise the carrying capacity to generate a natural exponential population growth up to the new equilibrium. The population parameters we are particular interested in, the mutation burden and the prevalence rate for the disability can be formulated in terms of the population process. The relative mutation burden of a population ν is defined as

$$L(\nu) \coloneqq \frac{1}{\langle \nu,1\rangle}\int_{\mathcal{X}} l(\mathbf{x})d\nu(\mathbf{f},\mathbf{x}) \quad \text{with} \quad l(\mathbf{x}) \coloneqq \sum_{i\in\{1,2\}}\sum_{n=1}^{N} x_n^i$$

And the relative number of individuals in the population ν belonging to the set $\mathcal{D}_N$ is

$$I(\nu) \coloneqq \frac{\nu(\mathcal{D}_N)}{\langle \nu,1\rangle}$$

To generate stochastically correct trajectories of the population dynamics process we implemented a variation of Gillespie algorithm for the above model in Python.

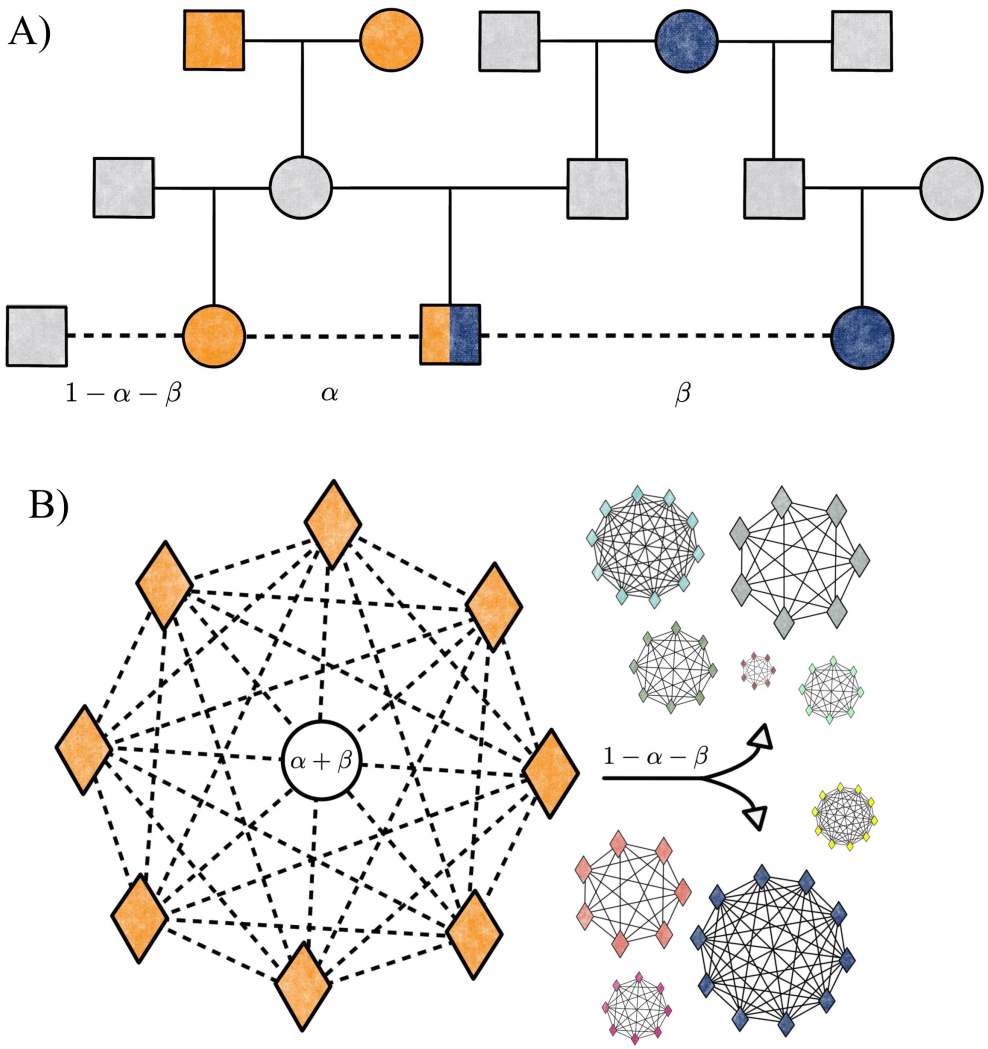

Figure 2.4: Comparison of Implementation of consanguineous mating scheme: The upper image (A) depicts a typical pedigree resulting from the implementation of consanguineous mating in SLiM. Precise inheritance up to two generations in the past are known. it is highly unlikely for two parents to have more than one child together since females independently choose partners for each mating. The pivotal factor in partner selection is the number of shared ancestors in the previous generation lineage. If two parents have two common ancestors, a mating occurs with probability α; if they share one common ancestor from tow generations ago, mating occurs with probability β; and if they lack any common ancestors, mating transpires with probability $1-\alpha-\beta$. In the lower image (B), a schematic visualization of consanguineous mating in the adaptive model is presented. Within families, no specific structures is retained. Mating within the family occurs with a probability of $\alpha+\beta$, while mating outside the family transpires with a probability of $1-\alpha-\beta$. In addition to the probabilities α, β the average family size $\kappa/2$ plays a decisive role here.

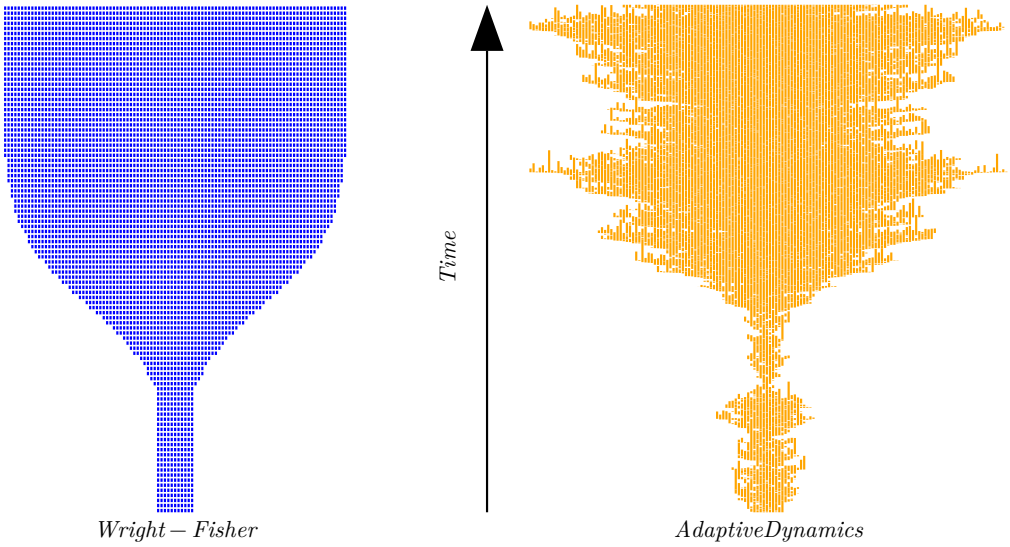

Figure 2.5: Comparison of population size and life spans of individuals: Each tick represents the lifespan of an individual. Time progesses from bottom to top. On the left side in blue, one can observe that all individual in the Wright-Fisher model share the same lifespan, generations do not overlap, and the population is of constant size initially, then grows deterministically until the new constant size is reached. In contrast, on the right side in orange, the adaptive model exhibits varying lifespans among individuals, leading to population fluctuations. Moreover, birth times of individuals are independent of each other, resulting in smoothly transitioning generations. At the point where the deterministic growth starts in the Writght-Fisher model the adaptive dynamics population was given more capacity which also leads to a logistic grow.

3 Refining the drift-barrier hypothesis: a role of recessive gene count and an inhomogeneous Muller‘s ratchet

This Chapter 3 is available as a preprint as joint work with Konrad Gerischer, Anton Bovier and Peter M. Krawitz [136],

> L. A. La Rocca, K. Gerischer, A. Bovier, and P. M. Krawitz. Refining the drift barrier hypothesis: a role of recessive gene count and an inhomogeneous Muller‘s ratchet. https://arxiv.org/abs/2406.09094, 2024

The drift-barrier hypothesis states that random genetic drift constrains the refinement of a phenotype under natural selection. The influence of effective population size and the genome-wide deleterious mutation rate were studied theoretically, and an inverse relationship between mutation rate and genome size has been observed for many species. However, the effect of the recessive gene count, an important feature of the genomic architecture, is unknown. In a Wright-Fisher model, we studied the mutation burden for a growing number of N completely recessive and lethal disease genes. Diploid individuals are represented with a binary $2 \times N$ matrix denoting wild-type and mutated alleles. Analytic results for specific cases were complemented by simulations across a broad parameter regime for gene count, mutation and recombination rates. Simulations revealed transitions to higher mutation burden and prevalence within a few generations that were linked to the extinction of the wild-type haplotype (least-loaded class). This metastability, that is, phases of quasi-equilibrium with intermittent transitions, persists over 100 000 generations. The drift-barrier hypothesis is confirmed by a high mutation burden resulting in population collapse. Simulations showed the emergence of mutually exclusive haplotypes for a mutation rate above 0.02 lethal equivalents per generation for a genomic architecture and population size representing complex multicellular organisms such as humans. In such systems, recombination proves pivotal, preventing population collapse and maintaining a mutation burden below 10. This study advances our understanding of gene pool stability, and particularly the role of the number of recessive disorders. Insights into Muller‘s ratchet dynamics are provided, and the essential role of recombination in curbing mutation burden and stabilizing the gene pool is demonstrated.

3.1 Introduction

The dependency of genome size and mutation rate was first noticed by Drake and further developed by Lynch into the drift-barrier hypothesis [60, 149, 196]. The basic idea of

this theory is that the refinement of a phenotype is ultimately limited by the noise of genetic drift, which is a consequence of effective population size and genome-wide deleterious mutation rate. The effective population size, K, assumes a sexual diploid population that mates randomly. Whereas the genome-wide deleterious mutation rate is the product of the base-substitution rate for deleterious mutations per nucleotide site per generation times an effective genome size that will evolve under the drift-barrier hypothesis. Drake's conjecture about an approximate constant of 0.003 [deleterious] mutations per genome per generation was based on a very small number of taxa and the first generation of sequencing technology. More than twenty years later, Sung et al. [196] refined this conjecture and showed an inverse relationship between the deleterious mutation rate and genome size over multiple orders of magnitude for viruses, eubacteria, and archaebacteria [149]. In order to apply the theory also to eukaryotes, the definition of an effective genome size was introduced, and the size of the coding DNA was used as a proxy. With the latest data from studies on the human population, the base substitution rate for humans could be further specified, and the effective genome size could encompass any region where deleterious mutations may occur. Thus, certainly the exome, but most probably also other non-coding elements such as enhancers.

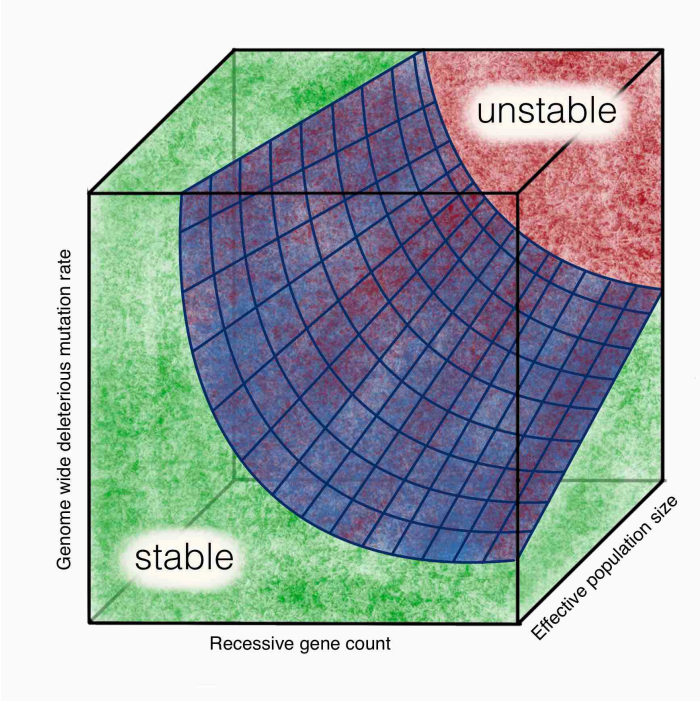

Figure 3.1: A shchematic representation of the Drift-Barrier in a three-dimensional parameter space. Below the hyperplane, populations can exist stably, whereas above it, the risk of extinction becomes too high. While the effect of the population size has already been described we observe an exponentially decreasing effect of the recessive gene count (see Figure 3.3). The combined effect has not yet been investigated. This is merely a sketch to visualize the extension of the drift-barrier hypothesis.

While the drift-barrier model was originally based on empirical results, there has also been extensive theoretical work to understand the stochastic nature of this phenomenon. In simple terms, the dynamics of a gene pool are either stable or unstable, depending on which side

of the barrier the system operates which is defined by its parameter settings (Figure 3.1). Most simulations in that field are based on the discrete Wright-Fisher Model or adaptive dynamics and both frameworks were shown to yield comparable results [136]. In our work, we consider a deleterious mutation rate μ, that is the number of lethal equivalents introduced per generation *de novo* into the recessive genes of a gamete. As an indicator of the fitness of a population or their gene pool, mutation burden $B(t)$ and disease incidence $P(t)$ were studied for populations of constant size K over a total time span of 100 000 generations [103]. The proportion of diploid individuals in the population that are biallelic for such lethal equivalents in at least one gene is the incidence or prevalence rate $P(t)$. The number of pathogenic alleles in the entire gene pool is defined as the mutation burden $B(t)$. Incidence rate and mutation burden are also the system parameters that we use to assess its stability. All haplotypes in the gene pool can be assigned to a class indicating the number of genes that harbour pathogenic alleles. By that definition, c_0 is the least loaded class containing only the haplotype without any mutations, c_1 consists of all haplotypes with exactly one affected gene, and so on. Due to maximal selection ($s = 1$) and total recessivity (dominance coefficient $h = 0$), a pathogenic or deleterious allele can also be referred to as a lethal equivalent that prevents propagation if both haplotypes of a gene are affected. Muller first studied a simple stochastic process for lethal mutations in haploid genomes without recombination and observed the irreversible loss of the least loaded class from the population which he described as clicks of a ratchet [96, 162]. By theoretical arguments, Charlesworth and Charlesworth argued that in diploid genomes the accumulation of pathogenic or lethal recessive alleles can result in the "crystalization" of the population, that is, the occurrence of haplotypes that are incompatible with each other [45]. In our work, we study the stochastic process of Muller's ratchet in the limit of strong selection on completely recessive diploid genomes in a parameter space with a variable number of genes that operates close to the drift-barrier. We find evidence for the crystalization phenomenon when increasing the genome-wide deleterious mutation rate or gene count beyond the drift-barrier. A particular focus of our work are the inhomogeneous or metastable dynamics that follow after the extinction of the mutation-free gametes until the emergence of haploid clusters, when the population regains stability after several thousand generations [24]. Unlike in other models, we do not reach an equilibrium every time between two successive clicks [152]. In agreement with all other models, we can show that recombination weakens the selective disadvantage of a pathogenic recessive allele by exporting mutations to other members of the population [44, 55, 116, 126, 128]. Therefore, recombination is also the force that helps to balance genetic load to a certain degree in genomes with an increasing number of recessive genes, which we introduced as a novel parameter of the drift-barrier.

3.2 Methods

Consider a diploid population where individuals are characterized by N diploid genes or gene sections that when mutated can carry and express a lethal disease. When an individual expresses one or more of such diseases it will be excluded from the mating process and thus cannot reproduce further. If a gene segment has already mutated once, further mutations on its same haplotype are neglected. Therefore, we can think of the genome of an individual

as a $2 \times N$ matrix with values in $\{0,1\}$, where a zero represents the wild type and a one indicates the presence of at least one mutation at that location. The fitness of an individual $x \in \{0,1\}^{2\times N}$ is optimal (fitness=1) unless it carries at least one mutation at each copy of at least one gene and hence expresses the disease. In this case the reproductive fitness is reduced to zero. In a classical Wright-Fisher Model with a constant population size of K individuals, in every generation for every offspring two parents are chosen according to their fitness and the offspring then inherits a combination of the parental genetic material. The creation of the offspring's genes is influenced by two parameters: the probability of recombination r and the mutation rate μ.

The recombination rate $r \in [0,1]$ denotes the probability that a potential crossover breakpoint occurs between neighbouring genes (see Supporting Information for details). The average number of potential breakpoints is hence Binomial distributed with parameter $N-1$ and r. In the course of the copying of the DNA to form a new gamete each of the two haploid sets of chromosomes is chosen independently with equal probability. After these recombination events, each newly born individual receives a gamete from each of their parents, creating their new diploid set of chromosomes. Lastly, *de novo* mutations are added independently with rate μ at every gene on either genome. Therefore, the diploid mutation rate is 2μ, and the probability of changing the wild type to a mutated site at a specific locus is approximately $\frac{\mu}{N}$. There are no back mutations. Due to the usually small mutation rates and comparatively large number of possible sites, it is reasonable to assume that the total number of mutations per birth is Poisson distributed with mean 2μ.

3.2.1 Model description

For notational reasons, we interpret the $2 \times N$ matrix $x \in \{0,1\}^{2\times N}$ as two binary numbers, where each number represents the maternal or paternal genome. Therefore, take an integer $i \in \{0,\dots,2^N-1\}$ and denote by $z_i = (z_1^i,\dots,z_N^i) \in \{0,1\}^N$ the N digits of the dual representation of i with leading zeros if necessary. Hence the values $z_1^i,\dots,z_N^i \in \{0,1\}$ are chosen such that

$$i = \sum_{n=1}^{N} z_n^i \cdot 2^{n-1}$$

The vector z_i is called haploid configuration or gamete. With this interpretation of the diploid configuration we can easily enumerate all configurations. For $i,j \in \{0,\dots,2^N-1\}$ denote by $x_{ij} \in \{0,1\}^{2\times N}$ the genetic configuration

$$x_{ij} = \begin{pmatrix} z_i \\ z_j \end{pmatrix} = \begin{pmatrix} z_1^i & z_2^i & \dots & z_N^i \\ z_1^j & z_2^j & \dots & z_N^j \end{pmatrix} \in \{0,1\}^{2\times N}$$

and denote by $X_{ij}(t)$ the number of individuals in generation t with configuration x_{ij}. Further, let $\boldsymbol{X}(t) = \big(X_{ij}(t)\big)_{0\le i,j\le 2^N-1}$ be the state of the population at time t. The reproductive fitness f of an individual with configuration x_{ij} is defined as

$$f\left(x_{ij}\right) \coloneqq \begin{cases} 0 & \text{, if } \exists\, n = 1,\dots,N\colon z_n^i = z_n^j = 1 \\ 1 & \text{, else} \end{cases}$$

The distribution of $\boldsymbol{X}(t+1)$ given $\boldsymbol{X}(t)$ is multinomial with parameters K and probabilities $(p_{ij}(t))_{0\leq i,j\leq 2^N-1}$ given by

$$p_{ij}(t) \coloneqq \sum_{h,h',k,k'=0}^{2^N-1} \frac{X_{hh'}(t)f(x_{hh'})X_{kk'}(t)f(x_{kk'})}{T(t)^2} m_{ij}(x_{hh'}, x_{kk'}; r, \mu) \tag{3.1}$$

with

$$T(t) \coloneqq \sum_{i,j=0}^{2^N-1} X_{ij}(t)f(x_{ij})$$

the total fitness of the population. The term $m_{ij}(x_{hh'}, x_{kk'}; r, \mu)$ denotes the probability of two configurations $x_{hh'}$ and $x_{kk'}$ producing an offspring with configuration x_{ij}. Note that, due to the inheritance rules described above, we obtain the following symmetries. First, there is no distinction between the maternal and paternal genomes, and second, the order of the partners in the composition of the genome is irrelevant. Hence for any $i, j, h, h', k, k' \in \{0, \dots, 2^N - 1\}$ we have

$$\begin{aligned} (i) \qquad & m_{ij}\left(x_{hh'}, x_{kk'}\right) = m_{ji}\left(x_{kk'}, x_{hh'}\right) \\ (ii) \qquad & m_{ij}\left(x_{hh'}, x_{kk'}\right) = m_{ij}\left(x_{h'h}, x_{k'k}\right) \end{aligned}$$

Subsequently, we introduce further statistics to investigate genetic phenomena within the population. Starting with the fraction of gametes $Z_i(t)$ in the population at time t with the haploid configuration $z_i \in \{0,1\}^N$, which is defined as

$$Z_i(t) \coloneqq \frac{1}{2K} \sum_{j=0}^{2^N-1} \left(X_{ij}(t) + X_{ji}(t)\right)$$

and $\mathbf{Z}(t) = (Z_i(t))_{0\leq i\leq 2^N-1}$ the gamete distribution at time t. For a gamete $z_i \in \{0,1\}^N$ we define the **mutation burden** $b(z_i)$ as the number of lethal equivalents on that specific haploid configuration, hence

$$b(z_i) \coloneqq \sum_{n=0}^{N} z_n^i$$

By some abuse of notation we set the mutation burden of a diploid configuration $x_{ij} \in \{0,1\}^{2\times N}$ to be $b(x_{ij}) \coloneqq b(z_i) + b(z_j)$. The mean haploid mutation burden $\beta(t)$ of the population $\boldsymbol{X}(t)$ in generation t is defined as the weighted mean of $\mathbf{Z}(t)$ with weights according to b, hence

$$\beta(t) \coloneqq \sum_{i=0}^{2^N-1} b(z_i)Z_i(t)$$

To measure fluctuations within the haploid mutation burden of a population, we also look at the weighted variance $\sigma_b^2(t)$ of $\mathbf{Z}(t)$ with weights b at time $t \geq 0$ and set

$$\sigma_b^2(t) \coloneqq \sum_{i=0}^{2^N-1} Z_i(t) \left(b(z_i) - \beta(t)\right)^2 \tag{3.2}$$

When we speak about mutation burden usually, we mean the average number of mutations per individual in the population at time $t \geq 0$. Hence the **mutation burden** $B(t)$ of the population at time $t \geq 0$ is defined as $B(t) \coloneqq 2\beta(t)$. Moreover, we define the **prevalence** $P(t)$ as the inverse of the relative fitness

$$P(t) \coloneqq 1 - \frac{T(t)}{K} = \frac{1}{K} \sum_{i,j=0}^{2^N-1} X_{ij}(t)(1 - f(x_{ij}))$$

At every haploid locus there are two possible alleles. The wild type (0) and the mutant allele (1). For $n = 1, \ldots, N$, we define the haploid **allele frequency** $\varphi_n(t)$ of the mutant allele at locus n at time t as

$$\varphi_n(t) \coloneqq \sum_{i=0}^{2^N-1} z_n^i \cdot Z_i(t)$$

and the average allele frequency across all loci as $\varphi(t) \coloneqq \frac{1}{N} \sum_{n=1}^{N} \varphi_n(t)$. For any $k \in \{0, \ldots, N\}$, define the **haploid load class** $c_k(t)$ as the fraction of gametes at time $t \geq 0$ with exactly k mutated genes

$$c_k(t) \coloneqq \sum_{i=0}^{2^N-1} Z_i(t) \mathbb{1}_{\{b(z_i)=k\}} \tag{3.3}$$

The discrete probability distribution $H(t)$ on $\{0, 1, \ldots, N\}$ with weights $\big(c_k(t)\big)_{k=0,\ldots,N}$ is called the haploid load class distribution. Note that by assigning each gamete uniquely to one of the $N + 1$ disjoint load classes, we obtain $\sum_{k=0}^{N} c_k(t) = 1$. The mean and variance of the haploid load class distribution are given by $\beta(t)$ and $\sigma_b^2(t)$ respectively. Lastly, we are interested in the level of linkage (dis-)equilibrium between different genes within the population, and we introduce the joint allele frequency $\varphi_{n,m}(t)$ of the mutated allele at the two distinct loci $n, m \in \{1, \ldots, N\}$ as

$$\varphi_{n,m}(t) \coloneqq \sum_{i=0}^{2^N-1} z_n^i z_m^i \cdot Z_i(t)$$

The square coefficient of correlation between the pair of loci is defined as

$$\rho_{n,m}^2(t) \coloneqq \frac{\big(\varphi_{n,m}(t) - \varphi_n(t)\varphi_m(t)\big)^2}{\varphi_n(t)\left(1 - \varphi_n(t)\right)\varphi_m(t)\left(1 - \varphi_m(t)\right)} \tag{3.4}$$

and consequently $\big(\rho_{n,m}(t)\big)_{1 \leq n,m \leq N}$ is called the **correlation matrix** between the loci at time $t \geq 0$. All notations are summarized in table 3.5 at the end of this paper.

3.3 Results

3.3.1 Mutation burden beyond the Drift-Barrier

The initial model that we analyzed, consisted of a genome with $2\mu = 0.05$, $N = 600$ recessive genes, and an effective population size of $K = 10\,000$ diploid individuals. Exact values are not

known, but empirical data suggest that the orders of magnitude should be correct [153, 215, 165, 92, 103]. In the simulation shown in Figure 3.2, mutation burden and prevalence remain constant over many generations. In this equilibrium, the count of deleterious mutations that are added to the gene pool, is equal to the number of such variants that are removed due to the affected individuals who do not procreate. To calculate these equilibria from the differential equations coming from the multinomial sampling in (3.1) is very hard. Already giving explicit formula for the probabilities $(p_{ij}(t))_{i,j=0,\dots,2^N-1}$ is quite challenging and in many cases does not give much insight. However in the simplest case for $N = 1$ explicit calculations are feasible (see Appendix). These match the considerations of Nei [167] for complete recessive lethals, which result in an allele frequency that is equal to the square root of the mutation probability. Extending these results to $N > 1$, assuming that in equilibrium the allele frequencies across all loci are equal yields an equilibrium allele frequency of

$$\varphi = \sqrt{1 - e^{-\frac{\mu}{N}}},$$

per gene, since the mutation probability for every individual gene with N total genes is equal to $1 - e^{-\frac{\mu}{N}}$. This leads to a mutation burden for an individual with $2N$ genes of

$$B = 2N\varphi = 2N\sqrt{1 - e^{-\frac{\mu}{N}}}$$

and a prevalence of

$$P = 1 - (1 - \varphi^2)^N = 1 - e^{-\mu}.$$

Note that the prevalence is independent from the number of loci N and equals the probability that a mutation appears on a gamete at birth. In Figure 3.4 A we see that in the case of no recombination ($r = 0$) the simulations match these considerations. In the case of full recombination ($r = 1$) however the allele frequency and hence also the haploid mutation burden is lower that what was expected by Nei [167]. However in the absence of recombination, these equilibria, while they may remain stable for several thousand generations, are fragile. For example, in Figure 3.2 at around generation 20k we observe a transition to a higher mutation burden and prevalence within a few generations. Further transitions follow over a larger time frame in a stochastic manner. On the level of haplotypes, these transitions are a consequence of the extinction of the least loaded classes, which has also been referred to as clicks of Muller‘s ratchet (Figure 3.2 B). After the extinction of the c_0 class the new least loaded class c_1 of gametes with exactly one lethal equivalent is left without influx, but rather looses gametes due to *de novo* mutations. This would lead to a rapid fixation of one mutation within the population, even under weaker selection coefficients [45]. Fixation however is not possible without the extinction of the whole population for recessive lethals. Instead we observe that the extinction of the mutation free gamete sets off a cascade of extinction events that is only reassured by the formation of clusters of similar haplotypes that are mutually exclusive. This phenomenon of “crystalization” was already predicted for low dominance coefficients by Charlesworth and Charlesworth [45]. In the period after the population has stabilized, there can be instances of further genes being incorporated into one of the clusters. This may or may not be accompanied by the extinction of the least loaded class. Moreover, there can be a reduction in the number of clusters as they compete with each other. This results not only in an increase in mutation burden but also in a rapid rise in prevalence (see Figure 3.2 at around 50k).

3.3.2 Influence of recessive gene count on metastability

The metastability that we observed for mutation burden and prevalence occurred on the time scale of 100k generations in all of our simulations but at different time points, indicating the stochastic nature of this process (Muller‘s ratchet). Under natural conditions, the transitions to a much higher mutation burden in the gene pool result in the extinction of the population for the combination of μ and K, which would be in accordance with the drift-barrier hypothesis (see Supporting Information for details). In the following experiments, we aimed to characterize the interplay of μ and the number of recessive genes. We, therefore, counted how often and when the first click of Muller‘s ratchet, which is the loss of the c_0 class, happened (Figure 3.3). We found that for a given K the drift-barrier depends not only on μ but also on the recessive gene count N. As long as the proportion of c_0 haplotypes in the gene pool remains above roughly $7.3 \cdot 10^{-3}$ a transition to higher mutation burden and prevalences is unlikely to occur within 100k generations (dotted line in Figure 3.3). For a genome with $N = 400$ recessive genes, there were still enough haplotypes in the c_0 class at a mutation rate of $2\mu = 0.03$. However, when the gene count increased to $N = 1000$, mutations started to accumulate and c_0 died out unless the mutation rate was lowered to roughly $2\mu < 0.015$. This suggests that Drake‘s Rule can also be formulated as the minimal proportion of the haplotypes without deleterious mutations, c_0, in the gene pool that is required to avoid Muller‘s ratchet.

3.3.3 Recombination can avoid the extinction of the least loaded class

Beyond the drift-barrier, the gene pool quickly acquires deleterious mutations that can eventually result in a collapse of the entire population. In finite populations this stochastic mechanism can be counteracted by amphimixis, that is sexual reproduction involving the fusion of two different gametes, that have to undergo meiosis [113, 128, 162, 172]. During meiosis, recombination can occur that counteracts linkage disequilibrium (LD) between deleterious mutations in different genes by negative selection. In our simulations, the recombination rate is the probability of a crossing over between genes. That is, for a recombination rate of $r = 0$, either the grandmaternal or grandpaternal haplotype is transmitted. In contrast, for a recombination rate of $r = 1$, the resulting haploid genome of the gamete would be a random sequence of the ancestral genes. We observed that in a population without recombination, metastability occurs for $2\mu = 0.05$ around $N = 500$ genes. That means the mutation burden in the gene pool after 100 000 generations deviates strongly from the beginning. The introduction of recombination, however, is able to control the mutation burden effectively and keep it below 10 for $N = 1000$, and probably above (Figure 3.4). Even in the initial equilibrium stage during which the mutation free gamete is present, the frequency of the mutated allele is reduced by recombination. This inevitably is followed by an - albeit small - reduction in the size of the least loaded class (Figure 3.4 B). And yet recombination prevents the extinction of the mutation free gamete, because of two reasons. First, recombination lowers the fluctuation within the population. On the one hand for full recombination the haploid load class distribution H is a *Poisson* distribution, the variance of the haploid mutation burden stays comparable low for large N, like the expectation. Whereas on the other hand we observe a linear growth in N for the variance in the case of no recombination (Figure 3.4

C). These higher variances due to the absence of recombination, similar to a reduction of the population size, favour the extinction of the least loaded class due to natural fluctuations. Second - and more important - recombination can restore the mutation free gamete after it got lost. To see that denote by $g_r : \{0,1\}^{2\times N} \to [0,1]$ the probability that a given genetic configuration with recombination rate $r \in [0,1]$ produces a mutation free gamete. Here, we consider only the effect of recombination during gamete formation, without the influence of *de novo* mutation, as they are added only to the newly formed diploid individual. Then

$$g_1(x_{ij}) = \begin{cases} 0 & \text{, if } f(x_{ij}) = 0 \\ \left(\frac{1}{2}\right)^{b(x_{ij})} & \text{, else.} \end{cases}$$

In particular note, that even if at some point in time $c_0 = 0$ the class of mutation free gametes may still have an influx with positive probability. That is true whenever $r > 0$. Only in the absence of recombination and only segregation we get

$$g_0(x_{ij}) = \begin{cases} 1 & \text{, if } i + j = 0 \\ \frac{1}{2} & \text{, if } i \cdot j = 0 \text{ and } i + j > 0 \\ 0 & \text{, else.} \end{cases}$$

In that case, if at some point $t_\dagger > 0$ the class of mutation free gametes goes extinct by natural fluctuations it will stay extinct for all $t \geq t_\dagger$ and every new gamete will carry at least one mutation, which is known as a click of Mullers ratchet [161]. Theoretical models widely acknowledge that even low recombination rates can decelerate the accumulation of mutations [67, 162].

3.4 Discussion

In this work we showed that the recessive gene count is another parameter that is required to decide whether a population can reach a mutation selection balance or in other words whether it is able to operate on the stable side of the drift-barrier (Figure 3.1). In fact, we observed the complex dynamics of the mutation burden beyond the barrier for the first time by coincidence in our previous work, in which we studied the effect of different mating schemes and demographic histories on the recessive disease risk and incidence rates [136]. We expected that selection would be sufficient as an opposing force to counteract the effect of deleterious mutations and genetic drift in the full range of population sizes that we simulated. However, for certain parameter settings, particularly many recessive genes that are linked on the same chromosome, we noticed a metastability of the mutation burden and investigated this phenomenon further. Prior to us, the puzzling observation of a population collapse had been made by theoretical biologists studying molecular evolution and e.g in the book the “Crumbling Genome”, Kondrashov theorised about the origins of sexual mating [127]. Amphimixis, that is reproduction of a diploid multicellular organism by means of haploid sperm or egg cells, is just one possibility how Nature implemented genetic recombination [128]. By recombination, a reconstruction of the wildtype sequence becomes possible, even if lethal equivalents affect multiple recessive genes. Kondrashov estimated that the average rate

of "contamination" cannot surpass 10 in a human genome, or otherwise mutation-selection balance would not be sustainable [127]. Kondrashov defined contamination as the sum of all heterozygous pathogenic variants in recessive genes weighted by their selection coefficient and coined the term "muller" for this unit. 10 mullers would be 10 lethal equivalents per genome, or 100 pathogenic variants with a selection coefficient of $s = -0.1$, and so on. In retrospect, we can confirm that passing this threshold in our previous work caused the metastability, which also explains the reproducibility issues that we had for different seeds in our simulations. However, with recombination we were able to keep the mutation burden below 10 effectively for $N = 1\,000$, $2\mu = 0.05$, and a population size of $K = 10\,000$.

In the present work, we aimed at approaching the drift-barrier by numerical simulations for wide parameter ranges in all three dimensions. We did so by means of adaptive dynamics that behave equivalently to classical Wright-Fisher population genetics. We could also confirm the emergence of multiple mutually exclusive haplotypes, and for smaller numbers of genes. This "crystallization" into two complementary segregating haplotypes, is a phenomenon predicted by Brian Charlesworth after studying earlier works from Pàlsson, et al. [43, 176, 175]. In fact, we see our work as a new part of a long sequence of findings ever since John Haigh established a mathematical model in 1978 to quantify certain effects of Muller's ratchet [96]. Particularly, the question of the click rate - that is, the speed at which successive least loaded classes become extinct - is of great interest. Recently, good approximations depending on effective population size have been achieved using diffusion approximation [64]. However, all these models observe time-homogeneous click rates. Interestingly, we observe highly inhomogeneous click rates and we hypothesized that this is due to the higher complexity of the genomic architecture that we modelled. This makes techniques like diffusion approximation or coalescent approaches from previous studies inapplicable [89, 9]. Indeed, after simplifying our model to only consider the number of deleterious mutations per gamete and assuming these mutations are always uniformly distributed across the genome in each step, we observed a homogeneous click rate again. Thus, the position of mutations is crucial, making mathematical analysis very challenging. The inhomogeneity is an interesting finding from a mere stochastic point of view and could be motivation enough to analyse the timespan between clicks and determine their distribution. Furthermore, for evolutionary biology it might be interesting to focus not only on the clicks of Muller's ratchet but also on the addition of recessive genes to a cluster - which correlates with an increase in mutation burden - and the extinction of clusters - which correlates with an increase in prevalence. By this means our model can also contribute to the understanding of multilocus dynamics, that have recently been shown to the associative overdominance and background selection [79]. Providing these measures in terms of the order of effective population size would be a significant challenge. We share Charlesworth's hypothesis that this evolution stops only when either the entire population becomes extinct or two exclusive clusters emerge, each encompassing the entire set of genes.

3.5 Code availability

The code that supports the findings of this study has been deposited in an open-access repository and can be accessed via the following link:

https://doi.org/10.5281/zenodo.10985649

The repository contains detailed instructions for code usage, dependencies, and any other relevant information to facilitate reproducibility of the results.

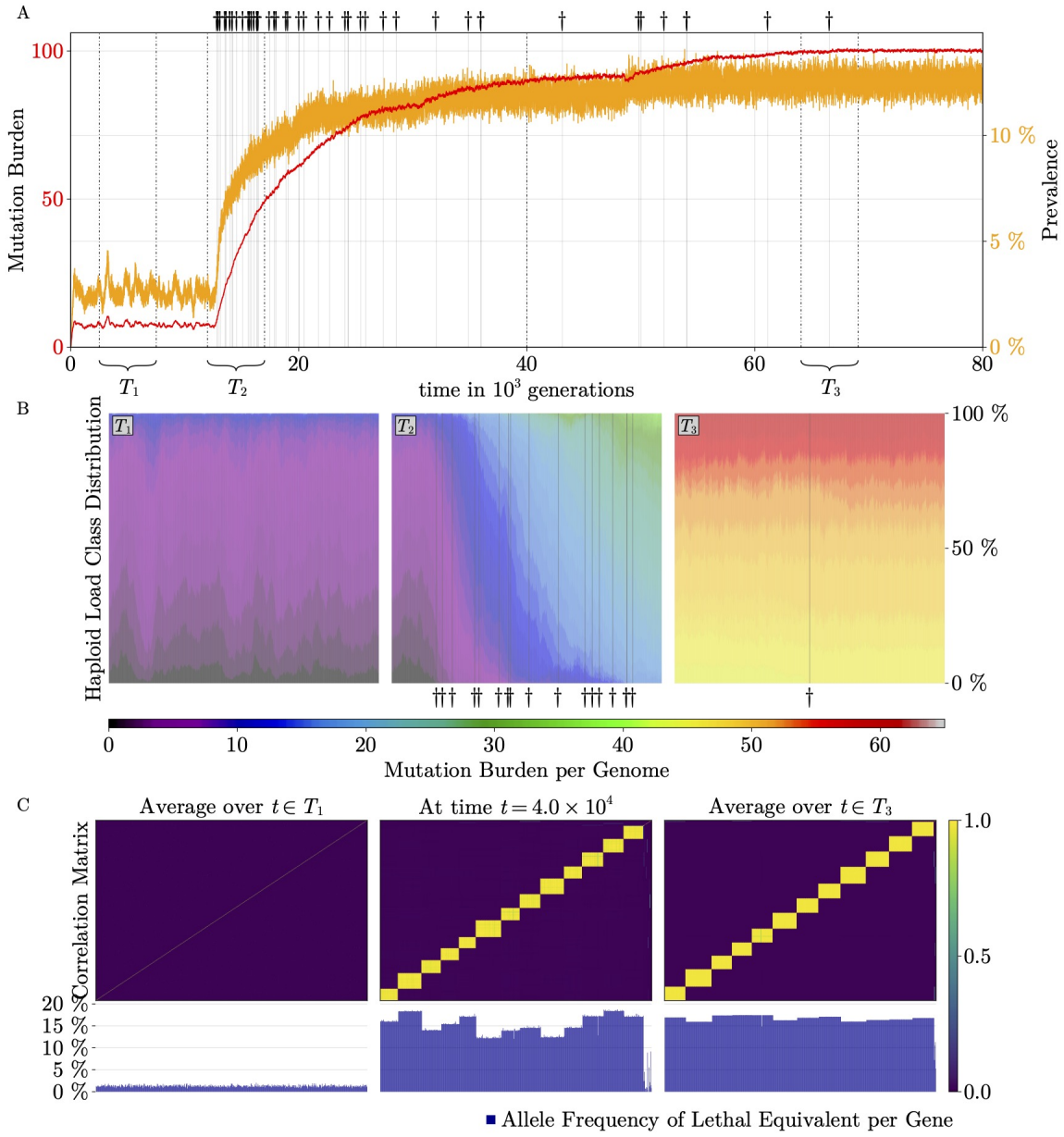

Figure 3.2: Metastability of mutation burden. For $K = 10\,000$, $N = 600$, $2\mu = 0.05$, the population operates close to the drift-barrier. A) Over a time span of 100 000 generations, several transitions to higher levels of mutation burden and prevalence can be observed that occur within a few generations and remain constant over many generations (metastability). B) The molecular cause for a transition to a higher level is the extinction of the least loaded class c_n, indicated by † (the first dagger is the loss of c_0, the class of haplotypes without any pathogenic mutation, and so on). C) The correlation or similarity matrix of the haplotypes changes over time. In the beginning, mutations are randomly distributed over the genes, and haplotypes are not correlated (T_1). At generation 40k, most haplotypes can already be assigned to one out of thirteen clusters with a heterozygote advantage. The cluster sizes, as well as the proportion of haplotypes assigned to clusters, increase over time, and the number of clusters decreases. In this simulation, presumably, a fixed state with twelve clusters was reached at T_3.

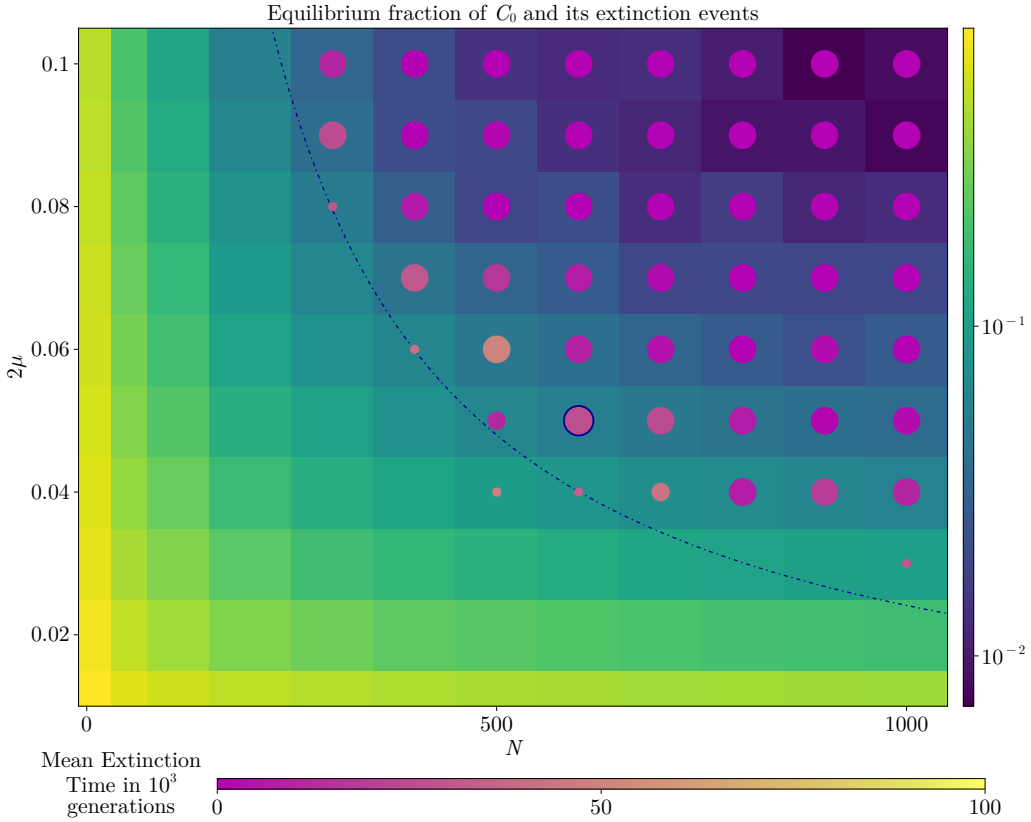

Figure 3.3: Influence of recessive gene count on the drift-barrier. If the extinction of the least loaded class occurs within 100 000 generations, metastability occurs. The grid visualizes the outcome of three iterations for each parameter combination, and the radius of the circles in the squares indicates the probability of extinction. E.g. for a deleterious mutation rate of $2\mu = 0.05$ and $N = 600$ genes, the least loaded class c_0 died out in all three simulations. In contrast, for $N = 200$, this event was not observed despite the same mutation rate. This indicates that the phase transition depends not only on the genome-wide mutation rate μ but also on N, and the recessive gene count becomes an additional parameter of the genomic architecture. The dotted line indicates the 99% quantile of extinction events.

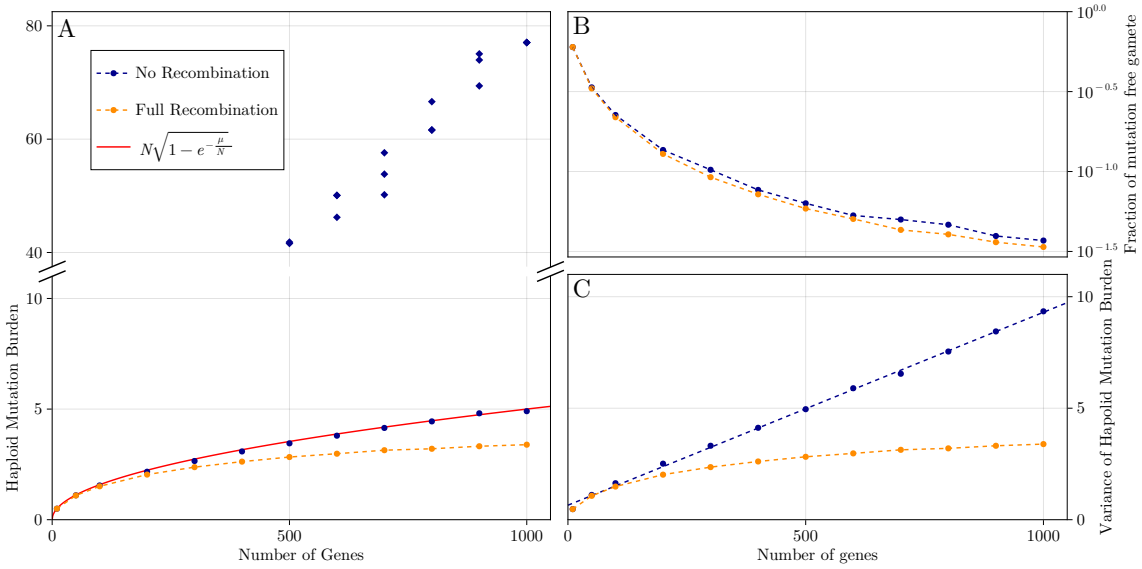

Figure 3.4: Recombination can effectively control mutation burden for higher gene counts. A) The number of recessive genes N and the genome-wide deleterious mutation rate μ affect the drift-barrier. For $2\mu = 0.05$ and $N = 500$, metastability would already occur (see Figure 3.3), and the average mutation burden after 100 000 generations deviates from the beginning of the simulations. The average haploid mutation burden before and after the transition is indicated by circles and diamonds. The mutation burden before any transition and without recombination can be described by $N\sqrt{1-e^{-\frac{\mu}{N}}}$, and the variance increases linearly (C). However, beyond the drift-barrier, the system becomes unstable, and mutation burdens after the transitions can differ. With recombination, the mutation burden in more than 500 recessive genes can remain below 10, which is considered an important threshold for lethal equivalents in the mutation-selection equilibrium. Likewise, the variance of the haploid mutation burden increases considerably slower with recombination (B).

Description	Notation
number of diploid genes	$N \in \mathbb{N}$
total / effective population size	$K \in \mathbb{N}$
recombination rate	$r \in [0,1]$
haploid genome mutation rate	$\mu \in \mathbb{R}_{\geq 0}$
individual / configuration / genome	$x_{ij} \in \{0,1\}^{2\times N}$
gamete / haploid configuration / haplotype	$z_i \in \{0,1\}^N$
integers which binary representation represents the haploid genome	$i, j = 0,\ldots,2^N - 1$
digits of the binary numbers representing the state of one haploid gen	$z_n^i \in \{0,1\}$
state of the population at time $t \in \mathbb{N}$	$\mathbf{X}(t) = \left(X_{ij}\right)_{i,j=0,\ldots,2^N-1} \in \mathbb{N}^{2^N}$
number of individuals with configuration $x_{ij} \in \{0,1\}^{2\times N}$ at time $t \in \mathbb{N}$	$X_{ij}(t) \in \{0,\ldots,K\}$
proportion of gametes with haploid configuration $z_i \in \{0,1\}^N$ at time $t \in \mathbb{N}$	$Z_i(t) \in [0,1]$
probabilities of multinomial distribution of $\mathbf{X}(t+1)$ given $\mathbf{X}(t)$	$\left(p_{ij}\right)_{i,j=0,\ldots,2^N-1} \in [0,1]^{2^N}$
reproduction probabilities - probability that a paring of $x_{hh'}$ and $x_{kk'}$ results in x_{ij} under the mutation rate μ and recombination rate r	$m_{ij}\left(x_{hh'}, x_{kk'}; r, \mu\right) \in [0,1]$
haploid mutation burden - number of mutations of the haploid configuration $z_i \in \{0,1\}^N$	$b\left(z_i\right) \in \{0,\ldots,N\}$
relative mutation burden of the population at time$t \in \mathbb{N}$	$B(t) \in \mathbb{R}_{\geq 0}$
prevalence of the population at time $t \in \mathbb{N}$	$P(t) \in [0,1]$
local allele frequency of the mutated allele at the n^{th} gene at time $t \in \mathbb{N}$	$\varphi_n(t) \in [0,1]$

Figure 3.5: Notation used within this paper.

3.6 Appendix

3.6.1 Only one gene

In general, providing explicit formulas for the probabilities (3.1) is not a fruitful endeavor, and identifying stationary distributions for the multinomial resampling is an intractable problem. The system of equations that one needs to solve grows exponentially with the number of genes. However, for very small genes, the equations can be obtained. More precisely, for $N = 1$, the system consists of 4 equations. Using symmetry properties, the dimensionality can be reduced even further. Symmetric configurations like x_{10} and x_{01} are indistinguishable in the sense that they have the same reproductive rates, hence there are three distinct configurations

$$x_{00} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad , \quad x_{01} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad , \quad x_{11} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

The terms $(m_{ij})_{i,j\in\{0,1\}}$, which describe the likelihood of a given mating resulting in an offspring of type ij, are given by the following table.

	$x_{00} \times x_{00}$	$x_{00} \times x_{01}$	$x_{01} \times x_{01}$
m_{00}	$e^{-2\mu}$	$\frac{1}{2}e^{-2\mu}$	$\frac{1}{4}e^{-2\mu}$
m_{01}	$2e^{-\mu}\left(1-e^{-\mu}\right)$	$e^{-\mu}\left(1-e^{-\mu}\right)+\frac{1}{2}e^{-\mu}$	$\frac{1}{2}e^{-\mu}\left(1-e^{-\mu}+\frac{1}{2}e^{-\mu}\right)$
m_{11}	$\left(1-e^{-\mu}\right)^2$	$\frac{1}{2}\left(1-e^{-\mu}\right)^2+\frac{1}{2}\left(1-e^{-\mu}\right)$	$\frac{1}{4}\left(1-e^{-\mu}\right)^2+\frac{1}{2}\left(1-e^{-\mu}\right)+\frac{1}{4}$

Given that individuals exhibiting the phenotype x_{11} possess a reproductive fitness of zero and are therefore excluded from the mating process, it follows that their probability of generating any configuration is effectively negligible. A particular distribution $\boldsymbol{k} = (k_{00}, k_{01}, k_{11})$ is stationary if

$$\mathbb{E}\left[\boldsymbol{X}(t+1)|\boldsymbol{X}(t) = K\boldsymbol{k}\right] = K\boldsymbol{k}$$

Since $\boldsymbol{X}(t+1)|\boldsymbol{X}(t) \sim \text{Multinomial}\left(K; p_{00}, p_{01}, p_{11}\right)$ we have that

$$\mathbb{E}\left[X_{ij}(t+1)|\boldsymbol{X}(t)\right] = Kp_{ij} \qquad \text{,for } 0 \le i \le j \le 1.$$

which yields

$$k_{00} = \frac{k_{00}^2 + k_{00}k_{01} + \frac{1}{4}k_{01}^2}{(k_{00}+k_{01})^2}e^{-2\mu} \tag{3.5}$$

$$k_{01} = \frac{k_{00}^2 + k_{00}k_{01} + \frac{1}{4}k_{01}^2}{(k_{00}+k_{01})^2}2e^{-\mu}\left(1-e^{-\mu}\right) + \frac{k_{00}k_{01} + \frac{1}{2}k_{01}^2}{(k_{00}+k_{01})^2}e^{-\mu} \tag{3.6}$$

$$k_{11} = \frac{k_{00}^2 + k_{00}k_{01} + \frac{1}{4}k_{01}^2}{(k_{00}+k_{01})^2}\left(1-e^{-\mu}\right)^2 + \frac{k_{00}k_{01} + \frac{1}{2}k_{01}^2}{(k_{00}+k_{01})^2}\left(1-e^{-\mu}\right) + \frac{\frac{1}{4}k_{01}^2}{(k_{00}+k_{01})^2} \tag{3.7}$$

It should be noted that there are in fact two distinct heterogeneous types, which introduces a factor of two into the mating of a heterogeneous couple.

Theorem 3.1. *The unique positive solution of the system (3.5- 3.7) is given by*

$$k_{00} = \left(1 - \sqrt{1 - e^{-\mu}}\right)^2, \quad k_{01} = 2\left(1 - \sqrt{1 - e^{-\mu}}\right)\sqrt{1 - e^{-\mu}}, \quad k_{11} = 1 - e^{-\mu}$$

Proof. Set

$$\rho = \frac{k_{00} + \frac{1}{2}k_{01}}{k_{00} + k_{01}}$$

then the system (3.5- 3.7) changes to

$$\begin{aligned} k_{00} &= \rho^2 e^{-2\mu} \\ k_{01} &= 2\rho e^{-\mu}\left(1 - \rho e^{-\mu}\right) \\ k_{11} &= \left(1 - \rho e^{-\mu}\right)^2 \end{aligned}$$

and we can solve for ρ in

$$\rho = \frac{\rho^2 e^{-2\mu} + \rho e^{-\mu}\left(1 - \rho e^{-\mu}\right)}{\rho^2 e^{-2\mu} + 2\rho e^{-\mu}\left(1 - \rho e^{-\mu}\right)} = \frac{1}{2 - \rho e^{-\mu}}$$

which results in

$$\rho = \frac{1 \pm \sqrt{1 - e^{-\mu}}}{e^{-\mu}}.$$

Therefore the only solutions with only positive entries is given by

$$k_{00} = \left(1 - \sqrt{1 - e^{-\mu}}\right)^2, \quad k_{01} = 2\left(1 - \sqrt{1 - e^{-\mu}}\right)\sqrt{1 - e^{-\mu}}, \quad k_{11} = 1 - e^{-\mu}$$

□

Remark. It is evident that the equilibrium population adheres to the Hardy-Weinberg principle, exhibiting frequencies for the mutated allele of

$$\varphi = \sqrt{1 - e^{-\mu}}$$

Moreover we find the equilibrium mutation burden and prevalence for $N = 1$ as

$$\hat{B} = 2\sqrt{1 - e^{-\mu}} \qquad \text{and} \qquad \hat{P} = 1 - e^{-\mu}$$

3.6.2 A diploid individual based model of adaptive dynamics

In addition to the Wright-Fisher model, we also implemented an adaptive dynamics model. Unlike Wright-Fisher models, the latter operates with a fluctuating population size and overlapping generations. Here, individuals reproduce and die at independent, exponentially distributed times, determined both by the individual's fitness and, in the case of the death rate, by the competitive pressure of the population. Accordingly, it can better capture effects that lead to growth or shrinkage of the population than population models with a fixed, constant population size. In the limit of large populations, we observe no differences however between the Wright-Fisher and the adaptive dynamics model for populations in equilibrium [136].

3.6.2.1 Model description

We use a variation of the model of adaptive dynamics of Mendelian diploids studied by P. Collet, S. Méléard, J. Metz et al. [49]. The major adaptation we make is a finite, but high dimensional genotype space $\mathcal{X} \subset \mathbb{R}^N$ and a more general approach on the recombination and propagation mechanism of genotypes during a mating of individuals. Here the dimension N corresponds to the number of gene segments under consideration. Hence a diploid individual is characterized by its genotype $\mathbf{x} = (x_1, x_2) \in \mathcal{X}^2$. In the following we introduce the demographic parameters that encode all of biology. We assume that these parameters are influenced by the allelic traits through the phenotypic trait. As this dependency is symmetrical, all coefficient functions defined are also assumed to be symmetric in the allelic traits.

(i) $b(x_1, x_2) \in \mathbb{R}_+$: on the one hand this is the birth rate of an individual with genotype (x_1, x_2) and on the other hand an individual with genotype (x_1, x_2) has probabilities proportional to $b(x_1, x_2)$ to be chosen as a mate during the birth event of another individual.

(ii) $d(x_1, x_2) \in \mathbb{R}_+$: the intrinsic death rate of an individual with genotype (x_1, x_2).

(iii) $c(x_1, x_2, y_1, y_2) \in \mathbb{R}_+$: the competition pressure from an individual with genotype (y_1, y_2) exerted onto an individual with genotype (x_1, x_2).

(iv) $m(x_1, x_2, y_1, y_2, z_1, z_2) \in [0, 1]$: the mating and mutation measure gives the probability that the mating of an individual with genotype (x_1, x_2) with an individual with genotype (y_1, y_2) produces an offspring with genotype (z_1, z_2). It is assumed to satisfy

(a) for each $\mathbf{x}, \mathbf{y} \in \mathcal{X}^2$

$$\int_{\mathcal{X}^2} m(\mathbf{x}, \mathbf{y}, d\mathbf{z}) = 1 \quad \text{and} \quad \int_{\mathbb{R}^{2N} \setminus \mathcal{X}^2} m(\mathbf{x}, \mathbf{y}, d\mathbf{z}) = 0.$$

Note that since $|\mathcal{X}^2| < \infty$ this means, that $m(\mathbf{x}, \mathbf{y}, \cdot)$ is a probability mass function with mass exclusively on $\mathcal{X}^2$.

(b) for every $(x_1, x_2), (y_1, y_2), (z_1, z_2) \in \mathcal{X}^2$ the following symmetry properties

$$\begin{aligned} m(x_1, x_2, y_1, y_2, z_1, z_2) &= m(x_2, x_1, y_1, y_2, z_1, z_2) \\ m(x_1, x_2, y_1, y_2, z_1, z_2) &= m(x_1, x_2, y_2, y_1, z_1, z_2) \\ m(x_1, x_2, y_1, y_2, z_1, z_2) &= m(y_1, y_2, x_1, x_2, z_2, z_1) \end{aligned}$$

The first two properties correspond to the fact, that we do not want to make a difference between the two genotypes of an individual. Both are equally present in the production of the offsprings genotype. Whereas the second property yields that the mating of two individuals has the same probabilities of producing a given pair of genotypes regardless the order of the mating.

For simplicity we ignore the existence of sexes and spacial structures within the population. Hence an individual chooses a mate with probabilities only proportional to the birthrate of the partner. At any point in time $t \geq 0$ we consider a finite number N_t of individuals. Denote their genotypes as $(x_1^1, x_2^1), \ldots, (x_1^{N_t}, x_2^{N_t}) \in \mathcal{X}^2$. The population state at time $t \geq 0$ is described by the point measure on $\mathcal{X}^2$

$$\nu_t = \sum_{i=1}^{N_t} \delta_{(x_1^i, x_2^i)}$$

where $\delta_{(x_1,x_2)}$ is the Dirac measure at $(x_1, x_2) \in \mathcal{X}^2$. Let $\langle \nu, f \rangle$ denote the integral of a measurable function f with respect to the measure ν. Then $\langle \nu_t, 1 \rangle = N_t$ and for any $(x_1, x_2) \in \mathcal{X}^2$, the non-negative number $\langle \nu_t, \mathbb{1}_{\{(x_1,x_2)\}} \rangle$ is called the density of genotype (x_1, x_2) at time t. In an abuse of notation we define

$$\langle \nu_t, \mathbb{1}_x \rangle \coloneqq \langle \nu_t(x, dy), 1 \rangle + \langle \nu_t(dy, x), 1 \rangle$$

to be the density of the haplotype $x \in \mathcal{X}$ at time t. Let $\mathcal{M}\left(\mathcal{X}^2\right)$ denote the set of finite, nonnegative point measures on $\mathcal{X}^2$, equipped with the weak topology,

$$\mathcal{M}\left(\mathcal{X}^2\right) \coloneqq \left\{ \sum_{i=1}^{n} \delta_{(x_1^i, x_2^i)} \colon n \geq 0, (x_1^1, x_2^1), \ldots, (x_1^n, x_2^n) \in \mathcal{X}^2 \right\}$$

An individual with genotype (x_1, x_2) in the population ν_t reproduces with an individual with genotype (y_1, y_2) at a rate $b(x_1, x_2) \frac{b(y_1,y_2)}{\langle \nu_t, b \rangle}$. The genotype of the offspring is chosen according to the mutation and mating measure $m(x_1, x_2, y_1, y_2, dz_1, dz_2)$. An individual with genotype (x_1, x_2) in the population ν_t dies at rate

$$d(x_1, x_2) + \langle \nu_t, c(x_1, x_2, dy_1, dy_2) \rangle$$

The population process $(\nu_t)_{t \geq 0}$ is defined as a $\mathcal{M}(\mathcal{X}^2)$-valued Markov process with the dynamics described above. These are encoded in the infinitesimal generator $\mathcal{L}$ of the process, which is defined for any bounded measurable function $f : \mathcal{M}(\mathcal{X}^2) \to \mathbb{R}$ and for all $\nu \in \mathcal{M}(\mathcal{X})$, by

$$\begin{aligned}(\mathcal{L}f)(\nu) = &\int_{\mathcal{X}^2} b(\mathbf{x}) \int_{\mathcal{X}^2} \frac{b(\mathbf{y})}{\langle \nu, b \rangle} \int_{\mathcal{X}^2} \left(f(\nu + \delta_{\mathbf{z}}) - f(\nu) \right) m(\mathbf{x}, \mathbf{y}, d\mathbf{z}) \nu(d\mathbf{y}) \nu(d\mathbf{x}) \\ &+ \int_{\mathcal{X}^2} \left(d(\mathbf{x}) + \int_{\mathcal{X}^2} c(\mathbf{x}, \mathbf{y}) \nu(d\mathbf{y}) \right) \left(f(\nu - \delta_{\mathbf{x}}) - f(\nu) \right) \nu(d\mathbf{x})\end{aligned}$$

The first term describes the mating and birth event. The second term describes the death of an individual. We ignore the unnatural fact that an individual can choose itself as a partner to mate as the probability of that event will become negligible as the population size increases.

Remark. Since we assume the model parameters b, d, c take finite, non-negative values, and the trait space $\mathcal{X}^2$ is finite we immediately get the existance and uniqueness of the process. Since if the population is of finite size n and in the state $\nu = \sum_{i=1}^{n} \delta_{\mathbf{x_i}}$ the total event rate is

$$R(\nu) = \sum_{i=1}^{n} b(\mathbf{x}) + d(\mathbf{x}) + \int_{\mathcal{X}^2} c(\mathbf{x}, \mathbf{y})\nu(d\mathbf{y}) \leq n \left(\max_{\mathbf{x} \in \mathcal{X}^2} \{b(\mathbf{x}) + d(\mathbf{x})\} \right) + n^2 \max_{\mathbf{x}, \mathbf{y} \in \mathcal{X}^2} c(x, y) < \infty$$

bounded from above as long as the population size is finite.

We see that this is true on finite time intervals as long as we start in a possibly random population with finite mean.

The trait space is $\mathcal{X} = \{0, 1\}^N$ hence every individual is characterized by a $2 \times N$ matrix with values in $\{0, 1\}$. Here zero represents the wild type and a one indicates that (at least one) mutation is present. Define the set $\mathcal{D}_N \subset \mathcal{X}^2$ as

$$\mathcal{D}_N := \left\{ (x, y) \in \mathcal{X}^2 \colon \exists\, 1 \leq i \leq N \text{ such that } x_i = 1 = y_i \right\}.$$

Then for $x, y, z, w \in \mathcal{X}$ the birth, death and competition rates are given by

$$b(x, y) := \bar{b} \mathbb{1}_{\mathcal{X}^2 \setminus \mathcal{D}_N}(x, y) \quad \text{and} \quad d(x, y) := \bar{d} \quad \text{and} \quad c(x, y, z, w) := \bar{c}$$

for some finite $\bar{b}, \bar{d}, \bar{c} \in \mathbb{R}_+$. Moreover define $\mu > 0$ to be the mutation rate per gamete. Since usually the number of loci N is big and the mutation rate μ is small we assume that the number of mutation per birth is Poisson distributed with mean 2μ. The mutation location then is uniform distributed among all $2N$ possible positions. During gamete formation, recombination events occur with constant rates. Let $r \in [0, 1]$ be the probability that a crossover breakpoint occurs between two adjacent gene sequences. At these points the genetic information is split and any copy is chosen at uniformly at random to produce the gamete. We assume that a crossover breakpoint occurs at any possible cutting point with equal probability c, independently of all other points. Knowing this the probabilities $m(x_1, x_2, y_1, y_2, z_1, z_2)$ can be calculated for any three pairs of genetic information $x_1, x_2, y_1, y_2, z_1, z_2 \in \mathcal{X}$ that the paring of (x_1, x_2) with (y_1, y_2) results in (z_1, z_2). First define the function $\gamma \colon \{1, \ldots, N-1\} \to \{0, 1\}$ that determines weather there is a crossover point between two genes or not in the sense that

$$\gamma(i) = \begin{cases} 1 & \text{if there is a crossover breakpoint between genes } i \text{ and } i+1 \\ 0 & \text{else.} \end{cases}$$

Then define the choice function $\tau_\gamma \colon \{1, \ldots, \|\gamma\|_1 + 1\} \to \{1, 2\}$ that chooses one of the two copies of each gene segments. Adding both together we define the function $\phi^\gamma_{\tau_\gamma} : \mathcal{X}^2 \to \mathcal{X}$ that determines the gamete of an individual with crossover points determined by $\gamma \in \{0, 1\}^{N-1}$ and chromosome selection $\tau_\gamma \in \{1, 2\}^{\|\gamma\|_1 + 1}$ as

$$\phi^\gamma_{\tau_\gamma}(x_1, x_2) = \left(x^1_{\tau_\gamma(1)}, x^2_{\tau_\gamma(\gamma(1)+1)}, \ldots, x^N_{\tau_\gamma(\gamma(1)+\ldots+\gamma(N-1)+1)} \right)_{k=1,\ldots,N}$$

Then we can define the mating and mutation probabilities in a general setting as

$$m(\mathbf{x},\mathbf{y},d\mathbf{z}) = \sum_{\gamma_x,\gamma_y\in\{0,1\}^{N-1}} r^{\|\gamma_x\|_1+\|\gamma_y\|_1}(1-r)^{2N-2-\|\gamma_x\|_1-\|\gamma_y\|_1}\frac{1}{2^{\|\gamma_x\|_1+\|\gamma_y\|_1+2}} \tag{3.8}$$
$$\times \sum_{\substack{\tau_x\in\{1,2\}^{\|\gamma_x\|_1+1}\\ \tau_y\in\{1,2\}^{\|\gamma_y\|_1+1}}} \sum_{k=0}^{\infty}\frac{(2\mu)^k}{k!}e^{-2\mu}\frac{1}{Z_k}\sum_{m\in\diamondsuit_k^{2N}}\delta_{\left(\left(\phi_{\tau_x}^{\gamma_x}(\mathbf{x}),\phi_{\tau_y}^{\gamma_y}(\mathbf{y})\right)+m\right)\wedge 1}(d\mathbf{z})$$

where $\diamondsuit_k^{2N} \coloneqq \left\{m\in\mathbb{N}_+^{2N}\colon m_1+\ldots+m_{2N}=k\right\}$ is the set of all lattice vectors in $\mathbb{N}_+^{2N}$ with one norm equal to k, moreover $Z_k=\sum_{j=0}^{2N}\binom{2N}{j}p_j(k)$ is the size of the set $\diamondsuit_k^{2N}$ and where $p_j(k)$ is the number of partitions of k into exactly j parts. For notational reasons define for $x\in\mathbb{R}^{2N}$ and $k\in\mathbb{R}$ the component wise maximum as $x\wedge k\coloneqq(x_1\wedge k,\ldots,x_{2N}\wedge k)$.

3.6.2.2 Results

In the initial equilibrium state, where the parameter c_0 is present, specific combinations of the parameters N and μ may result in a total population size that is less than the carrying capacity. Nevertheless, all relative statistics, such as prevalence and mutation burden, remain comparable to those of the constant size model. The reduction in population size associated with specific parameter combinations gives rise to heightened relative fluctuations, thereby increasing the likelihood of extinction for the least loaded class due to natural fluctuations. The extinction of c0 in this scenario precipitates a rapid escalation in mutation burden and prevalence. In contrast to the constant size model, the remaining healthy individuals are unable to effectively manage the rapidly rising prevalence, resulting in the population's rapid extinction. By modifying the model to maintain a constant overall birth rate distributed evenly among all healthy individuals, the exact dynamics observed in the Wright-Fisher model can be replicated.

3.6.3 Remark on Recombination

In this study, recombination is conducted in two stages. Firstly, potential crossover breakpoints are identified within the genome. Subsequently, the respective gene segments are selected with equal probability from either the maternal or paternal genome. Consequently, the mean number of true crossover breakpoints is approximately equal to the number of potential breakpoints, divided by two. A true crossover breakpoint occurs only when different origins are chosen for two consecutive genes, which occurs at each potential breakpoint with probability $\frac{1}{2}$. An alternative implementation, frequently encountered in the literature, combines these two processes into a single one. In this approach, crossover breakpoints are also determined for each individual with equal probability between each gene. However, these automatically result in switching between the maternal and paternal genomes, or vice versa. Consequently, a single starting genome is selected (maternal or paternal), and switching occurs automatically at each breakpoint. The two implementations are equivalent. However, since the selection process is already incorporated into the latter, the recombination rate,

which is the probability of a crossover breakpoint occurring between neighbouring genes, ranges from 0 to $\frac{1}{2}$ instead of $r \in [0, 1]$ as in the former. Accordingly, the case of full recombination, which is the focus of this study, is equivalent to a recombination rate of r = 1/2 for models with the second implementation of recombination.

4 DenseGillespieAlgorithm.jl

4.1 Home

This package implements a version of the Gillespies algorithm that performs exact stochastic simulations for dense problems. The Gillespie algorithm [80], introduced by Daniel Gillespie in 1976, is a fundamental tool for simulating the time evolution of systems with discrete, stochastic events, particularly in contexts like biochemical reactions and population dynamics. Its applications are particularly prevalent in contexts such as biochemical reactions and population dynamics. The Gillespie Algorithm is employed to simulate the behaviour of systems wherein reactions or events occur at random intervals. The algorithm generates a sequence of events and their timings by first calculating the rates at which different events or reactions occur. Subsequently, the time until the next event is determined based on these rates, and the type of event that occurs next is selected according to its probability. In the final step, the system state is updated based on the event, and the process is repeated.

The Gillespie algorithm is a highly renowned and widely utilised technique across diverse communities and ecosystems. A particularly efficient, flexible and comprehensive implementation can be found in the JumpProcess.jl package within the SciML ecosystem. We strongly recommend the use of this framework wherever feasible.

However, the majority of implementations of the Gillespie algorithm require prior knowledge of all potential types and all reactions between these types before the reaction commences. A classic illustration of this is the SIR model (see 4.3). The objective of our implementation in this package is to eliminate this restriction and permit the consideration of both high-dimensional systems, where the precise interactions between every conceivable combination are theoretically possible but practically infeasible, and additionally, systems where the trait space is uncountable, such as the real line. In both cases, the number of distinct traits that are present at any given time is finite, given that the population size is limited. However, new types emerge during the course of the simulation, and the interactions between these types are determined by their specific characteristics.

4.1.1 Manual Outline

- Manual
  - Installation
  - Setting up the model functions
  - Setting up the model parameter, population history and initial population

  - Execute the simulation
  - Customized statistics
- Examples
  - 1. SIR-Model
  - 2. Continuous trait space
  - 3. High-dimensional model
- Performance tips
  - Julia performance tips and benchmarking
  - Natural bottleneck in Gillespies Algorithm
  - Reuse memory space
  - Recalculate vs. update
  - Keep calm
- Public API
  - Detailed API

4.1.2 Index

- `DenseGillespieAlgorithm.sumsumdict`

4.2 Manual

The DenseGillespieAlgorithm framework is designed to assist researchers in simulating their complex models in an exact stochastic manner. It is the responsibility of the user to implement all model-specific functions, such as those pertaining to birth and death events or rate functions. Once this has been done, the framework executes the Gillespie Algorithm and saves the population history. The following section provides an overview of the main function of this package.

4.2.1 Installation

4.2.1.1 Install from GitHub

You can install the package directly from this GitHub repository:

```
using Pkg
Pkg.add("https://github.com/roccminton/DenseGillespieAlgorithm.jl")
```

4.2.1.2 Install from Julia

Once the package is registered in the official Julia package registry, you can install it via:

```
using Pkg
Pkg.add("DenseGillespieAlgorithm")
```

This will install the latest stable version of the package and all required dependencies.

Package dependencies When loading the package directly from GitHub, the following packages must be available `Random, Distributions, ProgressMeter, SparseArrays`

4.2.2 Setting up the model functions

The initial step is to define all interaction functions for the model. In population models, these are typically limited to two: birth and death. However, there is no upper limit on the number of interactions that can be included. As the number of fundamentally different interactions increases, the efficiency of the algorithm is reduced. The framework is speciallised to a small number of different events.

Subsequently, all the interaction functions should be incorporated into a single execute function. This function must accept three inputs: an index, the current population state, and the model parameter. The index specifies which of the defined events should be executed. The current population state is then modified by the event functions.

```
execute!(i,x,par)
```

Next we need to define the rates function. There must be as many rates as there are interaction events. Therefore the variable `initrates` is usually a `Vector` with as many entries as there are events. The rates function takes as input the initial rates, the current population state and the additional model parameters. The function should calculate the rates according to the population state and modify the `initrates` accordingly.

```
rates!(initrates,x,par)
```

Function names The function name may be designated as desired, as they are passed to the core function. The nomenclature is inconsequential.

Function signature Nevertheless, it is crucial to maintain the original function signature, which entails retaining the sequence and the number of arguments as they are called within the algorithmic structure.

Parameter variable There are no restrictions on the parameter variable `par`. Any additional information used to calculate rates and to change the current population state can be added to the parameter element that is passed through all functions. For example, if you want to know the current time of the simulation within the functions you run for time-inhomogeneous models, you could add this to your model parameter.

4.2.3 Setting up the model parameter, population history and initial population

The final step before running the simulation is to define the model parameters, including the time horizon of the simulation and the initial population state, as well as a blank population history.

The time horizon of the simulation is tipically a `UnitRange`, but can be anything that can be enumerated. The type of initial population state should correspond to the functionalities defiend in the function `rates!` and `execute!` as they use and modify this type. The empty population history should also match the type of population state, as it will be copied into the population history. In addition, if the population history is a vector or matrix, it should be at least as long as the time horizon. You can customise the saving process with your own Statistics! function. In this case, you will have to adapt the coupling history to the functionalities of this function. For more details, see Customized Statistics (4.2.5)

4.2.4 Execute the simulation

With everything in place, it is time to run the simulations. To do this, call the `run_gillespie!` function from the package.

```
run_gillespie!(
        time,n_0,par,
        execute!,rates!,
        initrates,
        population_history[,
        hstart=0,statistic!
        ])
```

Run a exact stochastic simulation, return and fill the `population_history`.

Arguments

- `time::AbstracVector`: time interval for the simulation
- `n0`: initial population state
- `par`: additional parameter (gets passed to ‘execute¡ and ‘rates¡)
- `execute!`: execute function
- `rates!`: rates function
- `initrates`: initial rates
- `population_history`: empty population history
- `hstart=0`: time shift for parameter change *(opitonal)*
- `statistic!`: additional statistic function *(optional)*

Extended help

- Note that `n0,initrates,population_history` all three get modified during the simulation.
- The algorithm expects the `execute!` function to have the following signature

  ```
  execute!(i::Number,n0,par)
  ```

  where the `i` is the event that gets executed and the population state `n0` gets modified accordingly. The only exception is when the `initrates` are given as a dictionary. In that case the signature is `execute!(i,trait,n0,initrates,par)`, where `trait` is the key that is modified.
- The algorithm expects the `rates!` function to have the following signature

  ```
  rates!(initrates,n0,par)
  ```

where the rates get modified according to the current population state given in `n0`.

- The algorithm expects the `statistic!` function to have the following signature

```
statistic!(population_hist,t,n0,par)
```

  where the population history gets modified at position `t` with the current population state `n0`.

- Note that the `population_history` needs to be accessible via index from 1 to `length(time)`, or if `hstart` is given from `1+hstart` to `length(time)+hstart`. Unless a specified `statistic!` function is given.
- Note that the initial population state `n0` must match the `population_history` in the sense that `population_history :: Vector{typeof(n0)}`. Unless a specified `statistic!` function is given.
- The parameter variable `par` is passed through all functions (`execute!`,`rates!`,`statistics!`), thereby affording the user additional flexibility.

Once the simulation has reached its conclusion, the modified population history is returned for further analysis.

4.2.5 Customized statistics

For many high-dimensional models, the exact configuration at any given time is too much information. In many cases only summery statistics are needed. To avoid accumulating too much data during the runtime of the algorithm that is not needed afterwards, you can define your own `statistic!` function. In this case, only the information you want to collect is stored for further analysis.

As for the `rates!` and `execute!` functions, the function signature is of particular significance. The function accepts as input the population history, which is modified by the function and the current time index, hence the index at which the statistics of the current population state are saved. Additionally, the current state and the model parameter are required.

```
statistics!(population_history,t,x,par)
```

4.3 Examples

Three illustrative examples are provided. The first is a minimal working example, designed to facilitate the initial implementation of the framework on the user's machine. The second is a slightly more advanced example, which illustrates the use of an uncountable trait space and caching for performance improvement in a particular use case. The third is a highly complex example, which demonstrates the comprehensive versatility of the package.

4.3.1 1. SIR-Model

The SIR model is a tree-dimensional model that is used to model infectious diseases. It is a simple model that assumes that individuals can be placed into one of three categories: susceptible, infected, or recovered. Infected individuals can infect susceptible individuals through random interactions. After becoming infected, individuals can recover and become immune. For more details, see for example here.

The initial step is to implement the fundamental interaction functions. In this scenario, two events are occurring: infection and recovery. The objective is to implement these functions in a manner that modifies the population state, which is represented as a vector with three entries, one for each possible state of an individual.

```
using Plots
using DenseGillespieAlgorithm

# Define the reactions
function infection!(x)
        x[1] += -1
        x[2] += 1
        nothing
end

function recovery!(x)
        x[2] += -1
        x[3] += 1
        nothing
end
```

The subsequent step is to combine the aforementioned two functions into a single execute function, which will subsequently be provided to the algorithm.

```
#Combine all reactions into one execute! function
function execute!(i,x,par)
        if i == 1
                infection!(x)
        elseif i == 2
                recovery!(x)
        else
                error("Unknown event number i = $i")
        end
                nothing
end
```

Furthermore, define the rate at which the events occur, which also depends on the population state. It should be noted that a modification of an existing variable that holds the current rates is necessary. In this case, as there are two events, namely infection and recovery, the rates variable will be a vector with two entries.

```
# Define the reactions (reaction rates and species interactions)
function rates!(rates,x,par)
        #rate of infection
        rates[1] = par.b * x[1] * x[2]
        #rate of recovery
        rates[2] = par.g * x[2]
        nothing
end
```

Prior to commencing the study, it is essential to define all relevant model parameters. These include the interaction rates, the initial population state and the time horizon for the simulation. Prior to commencing the simulation, it is essential to define all relevant model parameters. These include the interaction rates, the initial population state and the time horizon for the simulation. Additionally, it is necessary to provide the algorithm with an empty population history, which will be populated with data during runtime.

```
#Define the model parameter
par = (
        b = 0.000005,
        g = 0.005
)

# Define the initial state of the system
x0 = [9999,1,0]

# Define the time horizon for the simulation
t = 0:2000

# Initialize population history
hist = zeros(Int,(length(t),3))
```

At this point in the process, all the necessary components have been put in place, and the task can be handed over to the core function of the package. Once the simulation has been executed, the results are plotted.

```
# Run the simulation
run_gillespie!(
        t,x0,par,
        execute!,rates!,
        Vector{Float64}(undef,2),hist
)

# Analyze or plot the result (example with a simple print)
plot(hist,label=["S" "I" "R"])
```

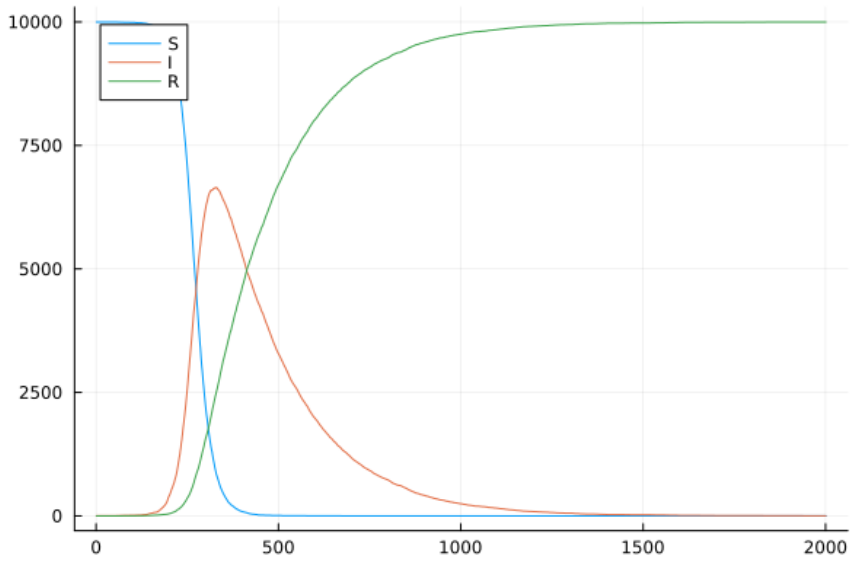

Figure 4.1: SIR plot

JumpProcess.jl While it is feasible to construct such straightforward examples using the DenseGillespieAlgorithm, this is not the typical application. For relatively simple models, the JumpProcess.jl package offers greater flexibility and facilitates the implementation process.

4.3.2 2. Continuous trait space

In this example, we consider an individual-based model of adaptive dynamics, wherein the trait space is a subset of the real line. It is therefore impossible to list all the types and interaction rates between them, as there are uncountably many. It is thus necessary to implement the rates and interactions in a dynamic manner.

In the context of adaptive dynamics models, individuals are characterised by a specific trait, which in this case is a real number. The mortality and fertility rates of individuals are contingent upon this trait. Moreover, competition among individuals is contingent upon the trait in question. Furthermore, at birth, with a probability of μ, the offspring undergoes a mutation and displays a distinct trait in comparison to its parents.

For further insight into the subject of adaptive dynamics models, we would direct the reader to the lecture notes by Anton Bovier.

We present a specific case study of an adaptive dynamics model, originally proposed by Dieckmann and Doebeli [57]. Here the trait space is $\mathcal{X} = [-1, 1] \subset \mathbb{R}$. The birth rate is given by $b(x) = exp(-x^2/2\sigma_b^2)$ for some $\sigma_b > 0$. The death rate is constant $d(x) = d$ and the competion between individuals depends only on theri distance by $c(x, y) = exp(-(x - y)^2/2\sigma_c^2)$ for some $\sigma_c > 0$. Moreover the mutation kernel, that chooses the new trait of an offspring at birth is a Gaussian law with mean 0 and variance 0.1 conditioned to $[-1, 1]$.

The next step is to begin the implementation of this model, starting with the rates function.

```
using Distributions

#birth rate
b(x, g) = exp(-x^2 / (2g^2))
#death rate
d(x,d) = d
#competiton kernel
c(x, y, g, K) = inv(K) * exp(-(x - y)^2 / (2g^2))
#mutation kernel
mutation(x) = rand(truncated(Normal(x, 0.1), -1, 1))
```

Given that we anticipate a relatively limited number of distinct traits to be present at any given time, but a considerable number of representatives for any given trait that we elect to implement this model with, we have opted to utilise dictionaries. Each trait is a key within the dictionary, with the value being a triple consisting of the size of the subpopulation and its intrinsic birth and death rate. By saving the birth and death rate, the need for repeated recalculation of the same rate in each step is avoided; instead, the rate is simply read from the dictionary. To illustrate,starting in a monomorphic equilibrium at the boudary $x_0 = -1$, the initial population state would be as follows.

```
x0 = -1.0

n0 = Dict(
        x0 => [
                (b(x0, g_b) - d(x0,d)) / c(x0,x0,g_c, K),
                b(x0, g_b),
                d(x0,d)
                ]
        )
```

In this manner, the rate values for each trait are stored in a cache once they are incorporated into the population. A similar approach is employed for the competition rates between individuals, with a dedicated dictionary being established to accommodate the various competition rates. In order to establish the competition dictionary, it is necessary to define the following function.

```
#Generate a cach dictionary for all competition rates between
#individuals from the population state ps
function generatecompdict(ps,competition)
        IndividualType = keytype(ps)
        Individuals = collect(keys(ps))
        #generate empty dictionary
        C = Dict{
                Tuple{IndividualType,IndividualType},
                Real
        }()
        #populate dictionary
```

```
        for x in keys(ps), y in keys(ps)
                C[(x,y)] = competition(x,y)
        end
        return C
end
```

The cached values will be passed to the functions via the parameter variable. Additionally, the birth, death, mutation and competition functions, with their fixed parameter values, will be stored there. Furthermore, it is necessary to adjust all model parameters, including the variances of the Gaussian birth and competition rates, the population size, and the time frame.

```
t = 0:1000

par = (
        birth = x -> b(x, 0.9),
        death = x -> d(x,0.0),
        competition = (x, y) -> c(x, y, 0.8, 1000),
        mutate = mutation,
        mu = 0.00015,
        K = 1000,
        compdict = generatecompdict(n0,(x, y) -> c(x, y, 0.8, 1000)),
        historylength = length(t)
)
```

History length As the population history will be stored in the dictionary, it is necessary to inform the algorithm of the duration of the simulation. To this end, the field "historylength" must be added to the parameter variable.

The next step is to define the rates function. In this case, the rates are also provided as a dictionary. Each subpopulation has two rates: a birth rate and a death rate. These are calculated from the cache and written to the dictionary.

```
#define rates function
function rates!(rates::Dict,ps::Dict,par)
        #iterate through current population
        for (x,vx) in ps
                #size of subpopulation
                nx = vx[1]
                #check if rates are already cached, if not do so
                !haskey(rates,x) && (
                        rates[x] = valtype(rates)(undef,2)
                        )
                #birthrate n_x * b(x)
                rates[x][1] = nx*vx[2]
                #deathrate n_x * (d(x) + sum c(x,y) n_y)
```

```
                rates[x][2] = nx* vx[3]
                for (traittuple,c) in par.compdict
                        t1,t2 = traittuple
                        t1 == x && (
                                rates[x][2] += nx * ps[t2][1] * c
                                )
                end
        end
end
```

The process of adding a new trait to the population at birth is rendered challenging by the presence of extensive caching, particularly in relation to competition rates. Consequently, a preliminary function is first devised to facilitate the addition of new traits to the population, prior to the implementation of the `birth!` and `death!` functions.

```
#add a new trait to current population
function addnewtrait!(ps,rates,par,trait)
        #add to population state
        ps[trait] = [par.diff,par.birth(trait),par.death(trait)]
        #set competition
        for other_trait in keys(ps)
                par.compdict[(trait,other_trait)] =
                        par.competition(trait,other_trait)
                par.compdict[(other_trait,trait)] =
                        par.competition(other_trait,trait)
        end
end
```

The `birth!` and `death!` functions can now be defined with relative ease and combined into a single `execute!` function.

```
function birth!(ps, rates, par, trait)
        #Birth with or without mutation
        if par.mu > 0.0 && rand() <= par.mu
                #mutate to new type/species and add to species
                new_trait = par.mutate(trait)
                #setup the size of the new type
                if haskey(ps,new_trait)
                        ps[new_trait][1] += par.diff
                else
                        addnewtrait!(ps,rates,par,new_trait)
                end
        else
                ps[trait][1] += par.diff
        end
        nothing
end
```

```
function death!(ps,trait,pr)
        ps[trait][1] -= par.diff
end

function execute!(i,trait,ps,rates,pr)
        if i==1
                birth!(ps, rates, pr, trait)
        elseif i==2
                death!(ps,trait,pr)
        else
                error("Index Error: Unknown event #$i")
        end
end
```

To initiate the simulation, it is merely necessary to establish an empty rates dictionary and population history, and then to execute the `run_gillespie!` function.

```
#empty population history
hist = Dict(x=>zeros(eltype(valtype(n0)),length(t)) for x in keys(n0))
#empty rates dictionary (gets populated in first iteration)
initrates = Dict{keytype(n0),Vector{Real}}()

#execute the simulation
run_gillespie!(
        t,
        n0,
        par,
        execute!,
        rates!,
        initrates,
        hist
)
```

To observe the findings, the size of the subpopulations is plotted over time, with the different traits represented by varying colours.

```
using Plots

#function to determine the color of the trait
function c(x)
        #find the biggest and smallest key in the population history
        min = min(keys(hist)...)
        max = max(keys(hist)...)
        #if there ever has been only one trait return 1
        #otherwise a color inbetween
        if min == max
```

```
                return 1
        else
                return floor(
                        Integer,(
                                (x-min)/(max-min)) * length(
                                        cgrad(:thermal)
                                )-1
                        ) + 1
        end
end

#setup plot
p=plot(legend=false)

for (x,his_x) in history
        plot!(p,time,his_x,color=cgrad(:thermal).colors.colors[c(x)])
end

p
```

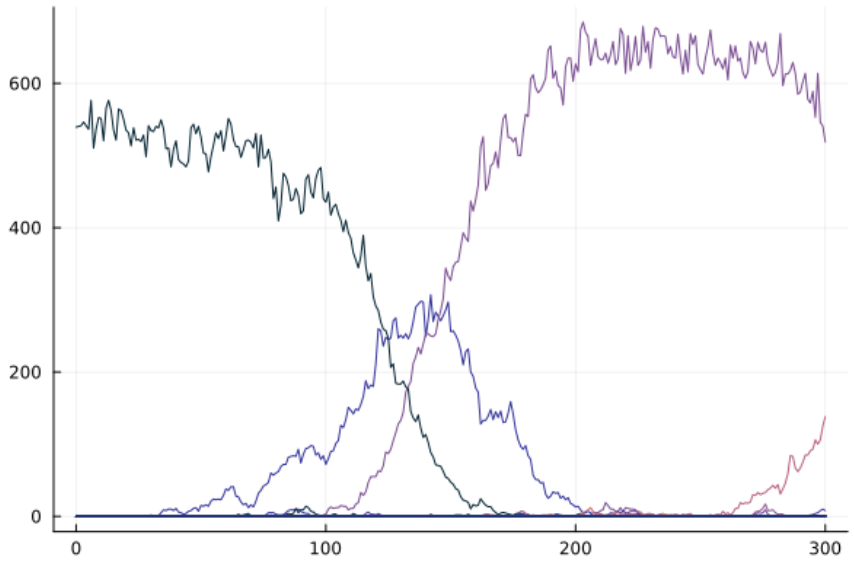

Figure 4.2: Simulation results with a mutation rate of $1/K$ where K is the carrying capacity.

Mutation rate In this scenario, the runtime of the algorithm is highly dependent on the mutation rate. An increase in the mutation rate results in a greater number of different traits. This implementation with dictionaries is most suited for a small number of traits being alive at the same time. However, if the mutation rate is increased to levels of frequent mutation, it is recommended that dictionaries are not used, but instead vectors should be employed for saving the data. The following example demonstrates this technique.

Population size Nevertheless, increasing the population size in this scenario does not significantly prolong the runtime of the algorithm. This is an advantage of using dictionaries and caching the competition. However, this approach is only effective when the mutation rate is scaled with the population size (as demonstrated in the above example).

Empty cache It should be noted that the algorithm performs regular checks for subpopulations in the dictionary with a population size of zero. In the event that such subpopulations are identified, they are removed in order to prevent an excessive expansion of the dictionary. This process is carried out by the `DenseGillespieAlgorithm.dropzeros!` function.

Switch-off caching It is possible to implement the same dynamics without the caching of competition rates. In this case, the only necessary modification is to alter the for-loop in the `rates!` function, which iterates over all pairs of tuples to

```
for t1 in keys(ps)
        rates[x][2] += nx * ps[t1][1] * par.competition(x,t1)
end
```

and delete the for-loop over all other traits in the `addnewtrait!` function. In instances where the number of distinct traits is considerable, this approach is advised.

4.3.3 3. High-dimensional model

The final example we will present is the most complex. We implement a model to analyse the dynamics of complete recessive lethal diseases. Each disease is triggered by the mutation of a gene and is expressed only in a homozygous state. Therefore, the traitspace for this model is $\mathcal{X} = \{0,1\}^{2\times N}$ where N is the number of genes. A detailed description and results of numerous simulations with this exact framework can be found here [136].

Individuals expressing a disease are excluded from the mating process. At birth, each individual randomly selects a fit partner from the population. Following the process of recombination, whereby the diploid genetic information is reduced to a haploid zygote incorporating crossover events, the gametes of the two parents fuse to form a new offspring. New mutations emerge at a constant rate. It is assumed that the intrinsic death rate and the competition among all individuals are identical. Only the birth rate is reduced to zero for infected individuals.

Given that there are 2^{2N} potential configurations with interactions between them, it is not feasible to enumerate them all prior to the start of the simulation. Furthermore, it is of no particular interest to ascertain the precise genetic configuration of the entire population. Typically, one is only concerned with summary statistics, such as the mutation burden (the average number of mutations per individual) and the prevalence (the fraction of individuals affected by a disease). It is therefore only these statistics that will be retained for subsequent analysis. However, for the propagation of the population dynamics, it is essential to have access to the exact configurations. To be more precise, the total birth and death rates can

be calculated via the summary statistics, which we utilise. However, in order to employ an offspring, the configurations are required.

In order to enhance the efficacy of the algorithm, it is essential to make extensive use of the parameter variable, which is passed to all relevant functions. The intrinsic configuration is stored therein in an optimal way, while the population state encompasses only the summary statistics and the total population size, as these are the necessary components for computing the event rates.

The initial stage of the process entails the establishment of all model-specific parameters, including the constant birth, death, and competition rates; the expected number of mutations at birth, denoted by μ; the number of recessive genes; the initial population size; and the recombination rate.

```
using SparseArrays
using Random

t = 0:1000

par = (
        birth = 1.0,
        death = 0.9,
        competition = 0.1 / 1000,
        mu = 0.1,
        Nloci = 100,
        K = 1000,
        recombination = 0.01
)

n0 = Dict(
        "PopSize" => par.K,
        "Ill" => 0,
        "ML" => 0
)
```

The only two possible events are birth and death, the `rates!` function can be expressed as follows:

```
function rates!(rates, ps, par)
        #linear birth for all propagable individuals
        rates[1] = par.birth * (ps["PopSize"]-ps["Ill"])
        #uniform logistic death
        rates[2] = ps["PopSize"] * par.death
                + ps["PopSize"] * (ps["PopSize"] - 1) * par.competition
        nothing
end
```

The definition of the function that executes the birth is a more challenging undertaking, given that it involves three mechanisms: mating, recombination and mutation. Nevertheless, prior to an explication of the implementation of the `execute!` function, it is necessary to describe the means by which the population configuration is saved internally. Since the size of any given trait is relatively large ($2N$ bytes), and since we anticipate a significant number of different traits, but a limited population size, we have chosen to construct a vector of traits that is as large as the expected population size, with additional space for fluctuations. This vector will store all the traits. Furthermore, a dictionary of indices is maintained, which points to the indices of traits in the vector. The dictionary differentiates between traits that are either alive and healthy, or alive but ill (thus expressing the disease and unable to reproduce), or that are not part of the current population. The aforementioned free traits can then be modified if new offsprings are born, eliminating the necessity of initiating a new $2\times N$ matrix of Bool's each time. This method allows for the saving of a considerable amount of memory. The production of new traits is dependent upon the absence of free indices. Upon the death of a fey individual, the index is released into the group of free indices, where it may be reborn as a new trait at a future point in time. In order to initiate the trait vector, which encompasses all individual genetic configurations, we have implemented a function that takes the initial population state as an input and draws a possible trait configuration from it.

```
#empty genetic configuration
emptytraits(Nloci,T=Bool) = [spzeros(T,Nloci),spzeros(T,Nloci)]

#produce trait collection from population state
function inittraits(par,n0)
        #Setup healthy genetic information
        locs = 1:par.Nloci
        #Generate healty population with some buffer for fluctuations
        traits = [
                emptytraits(par.Nloci)
                        for _ in 1:round(Int,par.K + sqrt(par.K))
                ]
        #add two mutations to completely healthy individuals
        #to get the required number of ill individuals
        for i in 1:n0["Ill"]
                l = rand(locs)
                traits[i][1][l] = 1
                traits[i][2][l] = 1
        end
        individuals = 1:n0["PopSize"]
        #add the remaining mutaions to the population
        #to get the required mutation load
        for i in n0["Ill"]+1:n0["ML"]-2*n0["Ill"]
                #choose random individual and location
                ind = rand(individuals)
                l = rand(locs)
```

```
                #recoose random individual and location
                #if the individual has already a mutation
                #at that locationo or at the homologe gene
                while traits[ind][1][l]+traits[ind][2][l] !== 0
                        ind = rand(individuals)
                        l = rand(locs)
                end
                traits[ind][rand(par.choosecopyfrom)][l] = 1
        end
        return traits
end
```

Subsequently, both the traits and the corresponding index dictionary are incorporated into the parameter variable. Additionally, other necessary elements are included, allowing for their reuse rather than generation on each occasion. These include a vector of random numbers for mate selection, the mutation distribution, the distribution of mutation locations and a unit range for gene segment selection.

```
par = (par...,
        rndm = Vector{Int}(undef,2),
        MutationsPerBirth = Poisson(par.mu),
        MutationLocation = 1:par.Nloci,
        traits = inittraits(par,n0),
        indices = Dict(
                "healthy" => collect(n0["Ill"]+1:n0["PopSize"]),
                "ill" => collect(1:n0["Ill"]),
                "free" => collect(
                        n0["PopSize"]+1:round(Int,par.K + sqrt(par.K))
                        )
        ),
        historylength = length(t),
        choosecopyfrom = 1:2,
)
```

The initial step on implementing the `birth!` function is to implement a function that establishes the crossover breakpoints for recombination at random, in accordance with the specified recombination rate. The function initially draws the number of crossover breakpoints from a *Poisson* distribution, and subsequently selects the position at random from among all $N-1$ positions. The resulting vector of 'UnitRange' segments is then returned.

```
#output for recombination rate 1
fullreccuts(par) = [i:i for i in 1:par.Nloci]
#output for recombination rate 0
noreccuts(par) = [1:par.Nloci]

function reccuts(par)
        if par.recombination == 1
```

```
                return fullreccuts(par)
        elseif par.recombination == 0
                return noreccuts(par)
        else
                #draw number of chromosome cuts
                ncuts = rand(Poisson(par.recombination*par.Nloci))
                #equals full recombination
                ncuts >= par.Nloci - 1 && return fullreccuts(par)
                #equals no recombination
                iszero(ncuts) && return noreccuts(par)
                #otherwise produce individual segments at random
                cutsat = sort!(
                        sample(1:par.Nloci-1,ncuts,replace=false)
                        )
                ccuts = [1:cutsat[1]]
                for i in 2:length(cutsat)
                        push!(ccuts,cutsat[i-1]+1:cutsat[i])
                end
                push!(ccuts,cutsat[end]+1:par.Nloci)
                return ccuts
        end
end
```

Subsequently, once two parents have been identified for mating, the process of the offspring generation is defined. This encompasses recombination, mating and mutation. The new genetic configuration of the offspring is stored at a designated index.

```
function offspring!(offspring_index, par, n_mut)
        #randomly recombine the parental genetic information
        #first for one then for the other parent
        for i in par.choosecopyfrom # =1:2
                #randomly choose one copy for each
                #chromosome/gene block
                ccuts = reccuts(par)
                choosecopy = rand(par.choosecopyfrom,length(ccuts))
                for (r,chromosome) in enumerate(ccuts)
                        view(
                                par.traits[offspring_index][i],
                                chromosome
                                ) .=
                        view(
                                par.traits[par.rndm[i]][choosecopy[r]],
                                chromosome
                                )
                end
        end
```

```
        #add n_mut mutations to random positions mutation
        #if there are no mutations to add skip the mutation process
        if n_mut > 0
                for _ in 1:n_mut
                        par.traits[
                                offspring_index
                        ][
                                rand(par.choosecopyfrom)
                        ][
                                rand(par.MutationLocation)
                                ] = 1
                end
        end
        nothing
end
```

The final step before integrating all components into a unified system is to implement a function that updates the population state following the generation of the offspring's genetic configuration.It is therefore necessary to ascertain whether the offspring in question exhibits a mutation in a homogeneous state, which would render it unsuitable for reproduction. Furthermore, the mutation burden of the offspring must be calculated.

```
#check if configuration has homogeneous mutation
ispropagable(a::Vector,Nloci) = ispropagable(a)
function ispropagable(a::Vector)
        for (i,p) in enumerate(a[1])
                isone(p) && isone(a[2][i]) && return false
        end
        return true
end
#calculate mutation load
mutationload(a::Vector) = sum(sum(svec) for svec in a)
#modify population state at birth of offspring
function updateps_birth!(ps,par,offspring_index)
        if ispropagable(par.traits[offspring_index],par.Nloci)
                push!(par.indices["healthy"],offspring_index)
        else
                ps["Ill"] += 1
                push!(par.indices["ill"],offspring_index)
        end
        ps["PopSize"] += 1
        ps["ML"] += mutationload(par.traits[offspring_index])
end
```

The aforementioned processes are unified in the `birth!` function, which initially identifies potential parents for mating, subsequently generates offspring, and finally updates the population state.

```
#execute the addition of an individual
function birth!(ps, par)
        #choose two genetic configurations to mate
        rand!(par.rndm,par.indices["healthy"])
        #clean up parental configurations
        for i in par.choosecopyfrom, j in par.choosecopyfrom
                dropzeros!(par.traits[par.rndm[i]][j])
        end
        #select free index for offspring
        if isempty(par.indices["free"])
                offspring_index = length(par.traits) + 1
                push!(par.traits,emptytraits(par.Nloci))
        else
                offspring_index = pop!(par.indices["free"])
        end
        #generate offsprings genetic configuration
        offspring!(
                offspring_index,
                par,
                rand(par.MutationsPerBirth)
                )
        #add the individual to the current population state dictionary
        updateps_birth!(ps,par,offspring_index)
end
```

The removal of an individual at death is a relatively straightforward process. It merely entails freeing the index and updating the population state in accordance with the relevant changes.

```
#modify population state at death of an individual
function updateps_death!(ps,par,fey_index)
        ps["PopSize"] -= 1
        ps["ML"] -= mutationload(par.traits[fey_index])
end
#execute the removal of an individual
function death!(ps, par)
        #choose fey
        if rand()<=ps["Ill"]/ps["PopSize"]
                fey_index = popat!(
                        par.indices["ill"],
                        rand(1:ps["Ill"])
                        )
                ps["Ill"] -= 1
        else
                fey_index = popat!(
                        par.indices["healthy"],
```

```
                        rand(1:(ps["PopSize"]-ps["Ill"]))
                        )
        end
        #add fey to gravejard
        push!(par.indices["free"],fey_index)
        #update population state
        updateps_death!(ps,par,fey_index)
end
```

As in the preceding cases, it is necessary to collate both the `birth!` and `death!` functions into a single `execute!` function.

```
function execute!(i,ps,par)
        if i == 1
                birth!(ps,par)
        elseif i == 2
                death!(ps,par)
        else
                error("Unknown event index: $i")
        end
end
```

Finally, the simulation can be executed following the initialisation of both a blank rates vector and a blank population history.

```
#setup empty rates vector
initrates = Vector{typeof(par.birth)}(undef,2)

#setup empty population history
hist = Dict(x=>zeros(valtype(n0),length(t)) for x in keys(n0))

run_gillespie!(
        t,n0,
        par,
        execute!,
        rates!,
        initrates,
        hist
)
```

In order to facilitate the analysis of the data, we present a straightforward graphical representation.

```
using Plots
#plot the prevalence
plot(t,hist["Ill"] ./ hist["PopSize"],color=:orange,label="")
#plot the population size (without axis ticks)
plot!(twinx(),hist["PopSize"],color=:gray,yticks=false,label="")
```

```
#plot the mutation burden
plot!(twinx(), hist["ML"] ./ hist["PopSize"], color=:red, label="")
```

Population extinction In the event of the population becoming extinct, the aforementioned method will result in an error, as the division of zero will occur. To circumvent this issue, it is recommended to first invoke the following function on the data set.

```
replace_NaN(v) = map(x -> isnan(x) ? zero(x) : x, v)
```

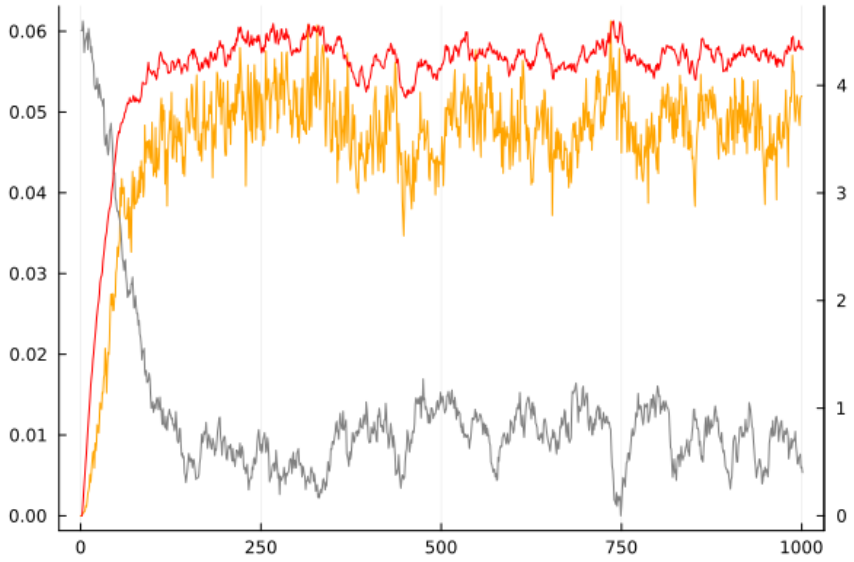

Figure 4.3: Simulation result showing the mutation load, prevalence and population size.

The grey line in Figure 4.3 represents the population size, which is not represented by any axis. On the left y-axis, the prevalence is represented by the yellow line, while the mutation burden is shown by the red line on the right y-axis.

4.3.3.1 Custom `statistic!` function

Thus far, the only function employed for the purpose of saving the population history was the built-in `DenseGillespieAlgorithm.saveonestep!` function, which in this case saves the mutation burden, prevalence, and population size over time. However, given the intricate population structure of the model, it may be beneficial to consider saving additional statistics that extend beyond mere numbers over time. For this example, we are interested in saving the allele frequencies of the mutated allele per position over time. This necessitates the definition of a custom `statistic!` function.

The initial step is to incorporate an additional function call into the existing functions `updateps_death!` and `updateps_birth!` of the form `updatestats_death!(ps, par, fey_index)` and `updatestats_birth!(ps, par, offspring_index)`, respectively. This

allows us to modify the new statistic that we wish to utilise at each event, rather than re-calculating it from the current population state after every full time step, which is the usual process.

Empty functions In the event that there is no intention to update the statistical data at each stage, it is nonetheless recommended that the function call `updatestats_event!` be retained in order to ensure the flexibility and reusability of the code. In the absence of a required function, the implementation of a generic function of the form

```
function updatestats_death! end
function updatestats_birth! end
```

is sufficient.

In order to enhance the flexibility of the system, a custom data type has been defined to accommodate the various statistical elements associated with the population history. This approach facilitates the incorporation of new statistics, should the need arise.

```
#type to hold population history
struct PopHist
        #mutation burden, prevalence and population size
        mlp :: Dict
        #allele frequencies per postion
        loadpos :: Array
end
```

Unmutable struct It is important to note that the variable type of the population history has been defined as unmutable, which may appear counterintuitive at first glance. However, upon closer examination, it becomes evident that the elements within the struct are only generated once and then populated with data. Meanwhile, the container itself (array, dict, etc.) remains unchanged. This allows for the use of a faster and lighter unmutable object. Conversely, if there is a need to modify the fields within the struct, it would be necessary to define it as a mutable struct.

We choose to save the allele frequencies of the mutated allele at each position as a $T \times N$ matrix, where T is the total length of the simulation. Each column of the matrix represents the allele frequencies for a single time step. In fact, we will save the precise number of mutations per gene, leaving the division by the population size to be performed subsequently, once the simulation has been completed. In order to utilise the enhanced performance afforded by the addition of the `updatestats_event!` function, it is necessary to create a temporary storage location for the current allele frequencies prior to their final saving to the storage medium for subsequent analysis. Once again, the parameter variable that is passed to every significant function is employed for this purpose.

```
#add blank current allele frequencies to parameter variable
par = (
        par..., 
```

```
        cafs = zeros(Int,par.Nloci)
)
```

Given the type configuration that is added or removed from the population, the adjustment of the current allele frequencies is a relatively straightforward process.

```
#add or remove one individual from allele frequency vector
function update_allelefreqs!(af,ind,i)
        af .+= i .* ind[1]
        af .+= i .* ind[2]
end
```

Furthermore, the corresponding functions for birth and death can be developed upon this function.

```
updatestats_death!(ps,par,index) =
        update_allelefreqs!(par.cafs,par.traits[index],-1)
updatestats_birth!(ps,par,index) =
        update_allelefreqs!(par.cafs,par.traits[index],+1)
```

The additional statistics have now been incorporated into the system, and the next stage is to save the data at each time step within the specified time horizon into the `PopHist` type. This process is carried out by the following function. Additionally, the function responsible for saving the supplementary statistical data is merged with the one that stores the mutation burden, prevalence, and population size, which were previously saved in the basic example. This allows the creation of the custom `statistic!` function.

```
function saveafs!(allelefreqs,index,ps,par)
        view(allelefreqs,index,:) .= par.cafs
end

function statistic!(pophist::PopHist,index,ps,par)
        #save standard statistic
        DenseGillespieAlgorithm.saveonestep!(pophist.mlp,index,ps,par)
        #save additional statistic
        saveafs!(pophist.allelefreqs,index,ps,par)
end
```

As previously described, the final stage of the process is to set up the initial rates, the initial population history, and then to execute the `run_gillespie!` function together with the newly defined `statistic!` function as a keyword argument.

```
#setup empty rates vector
initrates = Vector{typeof(par.birth)}(undef,2)

#setup empty population history
hist = PopHist(
        Dict(x=>zeros(valtype(n0),length(t)) for x in keys(n0)),
```

```
        zeros(Integer,(length(t),par.Nloci,))
)

run_gillespie!(
        t,n0,
        par,
        execute!,
        rates!,
        initrates,
        hist,
        statistic! =statistic!
)
```

To analyse the data, one possible approach would be to construct a small GIF that generates plots of the allele frequencies at each position over time.

```
using Plots

#calculate frequencies from absolute numbers of mutations
afs = hist.allelefreqs ./ hist.mlp["PopSize"]
#maximal frequencie for axis limit
ymax = maximum(afs)
#create animation
anim = @animate for i in 0:100
        bar(view(afs,i+1,:),ylim=(0,ymax),label="")
end every 10
#show animation
gif(anim)
```

Time intervals In certain instances, the requisite statistic may require a considerable amount of memory space or a significant amount of time to calculate. In such cases, it may be more efficient to save and calculate the statistic not in every time step, but rather only after larger intervals. This can be achieved in two ways. First, the time horizon provided to the algorithm can be adjusted to a coarser resolution, for instance, `0:10:1000` instead of `0:1000`, resulting in a step size of 10 rather than 1. It should be noted that in such instances, the events continue to occur at the (potentially very small) event rates, but the saving mechanism is executed at each full time step. In this scenario, however, all the statistics that are generated are saved exclusively at the larger time steps. Second, in the event that a specific statistic is particularly resource-intensive, it is possible to implement an `if` condition within the `statistic!` function that will then save the statistic only if the time index meets the specified condition.

Snapshots It is a source of considerable frustration to have invested a significant amount of time in running a comprehensive simulation only to realise, upon completion, that an alternative statistic might also warrant examination. Consequently, it was beneficial on

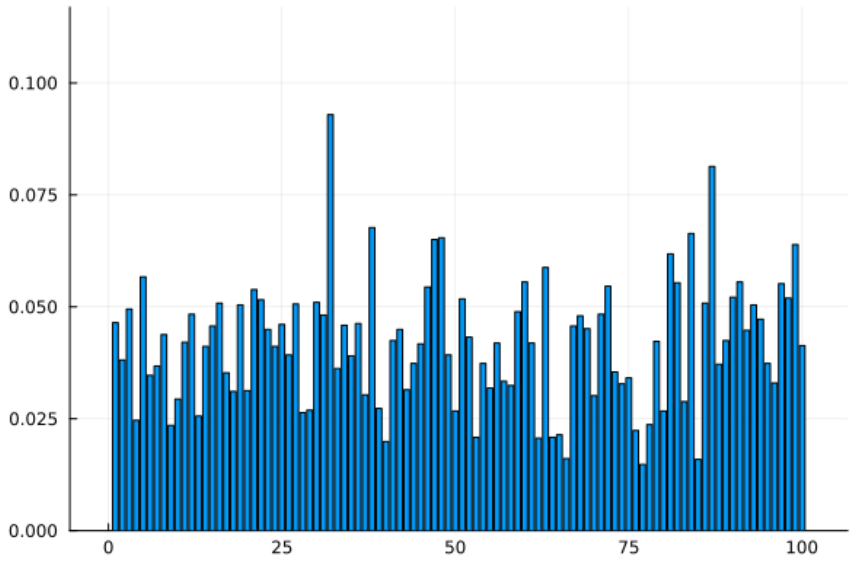

Figure 4.4: As a print medium is static, we're showing a snapshot of the GIF at generation 100. You'll find the moving image in the online version of the documentation.

occasion to also take "snapshots" at every couple of generations. Therefore, a random sample of the population was selected and all the information for that subpopulation was stored. As the population size was reduced by taking a sample from the population and the time horizon was reduced by taking these snapshots on a coarse time grid, the amount of memory required remained within acceptable limits.

4.4 Performance tips

One of the key benefits of the Gillespie algorithm is its ability to trace a single, precise stochastic trajectory. Nevertheless, in order to achieve this for each individual event, the rates must be calculated and re-calculated whenever there is a change in the population configuration. This makes the algorithm computationally demanding. There are numerous modifications that can be made in order to enhance performance, such as τ-leaping[82]. However, in this section, our objective is to maintain the precision of the stochastic simulation and to identify potential bottlenecks and strategies for optimising the performance of the simulation in its current form.

4.4.1 Julia performance tips and benchmarking

The Julia documentation contains a comprehensive and invaluable array of performance tips. Should you encounter any unexpected delays in your simulations, we advise you to consult these tips. It is likely that you will find a solution to the problem by following the guidance provided. Additionally, we recommend that you review the workflow tips and

the style guide. These resources assist in enhancing the efficiency, readability, and overall quality of your code. In particular, we recommend these chapters of the documentation for researchers who are new to Julia. To assess the efficacy of the functions within your model, Julia offers an exemplary benchmarking package called BenchmarkingTools.jl. This enables the comparison of different versions of your code, facilitating the selection of the most optimal version.

4.4.2 Natural bottleneck in Gillespies Algorithm

It should be noted that both the `rates!` and `execute!` functions will be called a considerable number of times during the course of the simulation. To illustrate, consider a species with a reproduction rate of 1 within a population of 1 000 individuals over a time horizon of 1 000 generations. In this scenario, the `rates!` and `execute!` functions will be called approximately one million times solely for reproduction events. Consequently, it is crucial to implement these functions with optimal performance. Every millisecond gained, every byte saved can significantly impact the runtime and efficiency of the program.

4.4.3 Reuse memory space

Avoid to recreate containers such as arrays or dictionaries, particularly within the `execute!` and `rates!` function. Reuse of these containers is preferable to the allocation of new memory space with each iteration. The accumulation of data over the course of a simulation can lead to a reduction in performance.

4.4.4 Recalculate vs. update

In certain scenarios, it is preferable to recalculate data from the current population state rather than update the data after every event, depending on the model and model parameter in question. For other models the opposite is may be true. In cases where there is uncertainty about the optimal implementation strategy, it is recommended to implement both versions and evaluate their performance using the BenchmarkingTools.jl package. This is the case for the event rates, which should be recalculated or updated depending on the circumstances and for the additional statistics that may be saved. However, it should be noted that the `rates!` function is usually called much more frequently than the `statistics!` function. The statistics are saved at regular intervals, as specified by the time horizon input to the algorithm, whereas the rates function is called after each event.

4.4.5 Keep calm

The Gillespie Algorithm is a computationally intensive algorithm due to its exact nature. Following the implementation of complex networks on a high-dimensional trait space, the execution may require a significant amount of time. It is recommended to test the algorithm on a smaller scale, benchmark and adapt functions regularly, and only to be concerned if

the time required to produce the simulation is unexpectedly long or if a disproportionate amount of memory is allocated. Otherwise, grab a coffee and check out the progress meter to see how things are going with the simulation.

4.5 Public API

Puplic documentation of all internal functions.

4.5.1 Detailed API

- `DenseGillespieAlgorithm.chooseevent` - Method

```
chooseevent(
        rates::Vector{Float64},
        total_rate::Float64
)
```

  Choose from the vector of total rates at random one of the indices of the vector according to their rates. The value 0 is returned if the total rates are positive, but too smale to let the evolution continue.

- `DenseGillespieAlgorithm.chooseevent` - Method

```
chooseevent(
        rates::Dict,
        total_rate::Float64
)
```

  Choose from the dictionary of total rates at random one of the keys of the dictionary according to their values. The value 0 is returned if the total rates are positive, but too smale to let the evolution continue.

- `DenseGillespieAlgorithm.dropzeros!` - Method

```
dropzeros!(ps)
```

  Do nothing for non-dictionary inputs.

- `DenseGillespieAlgorithm.dropzeros!` - Method

```
dropzeros!(ps::Dict{Any,Vector})
```

  Eliminates all key value pairs for which the the firts entry of the vector of the value is zero.

- `DenseGillespieAlgorithm.historylength` - Method

```
historylength(population_hisotry,par)
```

Return the simulation time based on the length of the population history. If the population history is neither a Vector nor a Matrix it is assumed that the Parameter has a field called historylength that is then returned.

- `DenseGillespieAlgorithm.mainiteration!` - Method

```
mainiteration!(
pop_hist,
rates,
n0,
ct,
time,
par,
ex!::F1,
r!::F2,
stat!::F3,
hstart
)
```

Main iteration of the GillespieAlgorithm for complex models.

- `DenseGillespieAlgorithm.nexteventandtime` - Method

```
nexteventandtime(rates::Vector{Float64})
```

Sample a exponential distributed random variable to determine the time for the next event and calls `choose_event`. The return value is a tuple consiting of the envent index returned by `choose_event` and the time to the next event.

- `DenseGillespieAlgorithm.nexteventandtime` - Method

```
nexteventandtime(rates::Dict)
```

Sample a exponential distributed random variable to determine the time for the next event and calls `choose_event`. The return value is a triple consisting of the event index and trait returned by `choose_event` and the time to the next event.

- `DenseGillespieAlgorithm.onestep!` - Method

```
onestep!(
        x_0,
        rates,
        t_0,
        t_end,
        par,
        ex!::F1,
        r!::F2
)
```

Execute one step of the evolution by modifying `x0` and rates and returning the current time `t0`.

- `Dense.GillespieAlgorithm.run_gillespie!` - Method

```
run_gillespie!(
        time,
        n_0,
        par,
        execute!,
        rates!,
        initrates,
        population_history[,
        hstart=0,
        statistic!
])
```

Run a exact stochastic simulation, return and fill the `population_history`.

Arguments

- `time::AbstracVector`: time interval for the simulation
- `n0`: initial population state
- `par`: additional parameter (gets passed to ‘execute¡ and ‘rates¡)
- `execute!`: execute function
- `rates!`: rates function
- `initrates`: initial rates
- `population_history`: empty population history
- `hstart=0`: time shift for parameter change *(opitonal)*
- `statistic!`: additional statistic function *(optional)*

Extended help

- Note that `n0,initrates,population_history` all three get modified during the simulation.
- The algorithm expects the `execute!` function to have the following signature

```
execute!(i::Number, n0, par)
```

  where the `i` is the event that gets executed and the population state `n0` gets modified accordingly. The only exception is when the `initrates` are given as a dictionary. In that case the signature is `execute!(i,trait,n0,initrates,par)`, where `trait` is the key that is modified.
- The algorithm expects the `rates!` function to have the following signature

```
rates!(initrates,n0,par)
```

where the rates get modified according to the current population state given in `n0`.

- The algorithm expects the `statistic!` function to have the following signature

```
statistic!(population_hist,t,n0,par)
```

where the population history gets modified at position `t` with the current population state `n0`.

- Note that the `population_history` needs to be accessible via index from 1 to `length(time)`, or if `hstart` is given from `1+hstart` to `length(time)+hstart`. Unless a specified `statistic!` function is given.
- Note that the initial population state `n0` must match the `population_history` in the sense that `population_history :: Vector{typeof(n0)}`. Unless a specified `statistic!` function is given.
- The parameter variable `par` is passed through all functions (`execute!,rates!,statistics!`), thereby affording the user additional flexibility.

- `DenseGillespieAlgorithm.saveonestep!` - Method

```
saveonestep!(pop_hist,index,ps,par)
```

Save one step of the simulation. Generic method if no explicit statistic! function is given.

- `DenseGillespieAlgorithm.stop!` - Method

```
stop!(pop_hist,index,n0,par,stat!)
```

Fill the remaining population history with the (statistic of) the current population state if the evolution came to a halt.

- `DenseGillespieAlgorithm.sumsumdict` - Method

```
sumsumdict(D::Dict{String,Vector})
```

Calculate the sum of the sums of the vectors that are the values of a dictionary.

Bibliography

[1] Eugenische Argumentation im Beschluss des Bundesverfassungsgerichts zum Inzestverbot: Stellungnahme der Deutschen Gesellschaft für Humangenetik (GfH). *Medizinische Genetik*, 20(2):239–239, 2008.

[2] Stübing v. Germany. *European Court of Human Rights.*, Application no. 43547/08, 2012.

[3] *Incest Prohibition. Opinion.* German Ethics Council, 2016. https://www.ethikrat.org/fileadmin/Publikationen/Stellungnahmen/englisch/opinion-incest-prohibition.pdf.

[4] M. Abouelhoda, T. Sobahy, M. El-Kalioby, N. Patel, H. Shamseldin, D. Monies, N. Al-Tassan, K. Ramzan, F. Imtiaz, R. Shaheen, and F. S. Alkuraya. Clinical genomics can facilitate countrywide estimation of autosomal recessive disease burden. *Genetics in Medicine*, 18(12):1244–1249, 2016.

[5] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*. New York, Garland Science, xxv + 1552 pp., 4 edition, 2002.

[6] G. E. Allen and C. Dytham. An efficient method for stochastic simulation of biological populations in continuous time. *Biosystems*, 98(1):37–42, 2009.

[7] C. E. G. Amorim, Z. Gao, Z. Baker, J. F. Diesel, Y. B. Simons, I. S. Haque, J. Pickrell, and M. Przeworski. The population genetics of human disease: The case of recessive, lethal mutations. *PLOS Genetics*, 13(9):e1006915, 2017.

[8] S. E. Antonarakis. Carrier screening for recessive disorders. *Nature Reviews Genetics*, 20(9):549–561, 2019.

[9] J. Audiffren and E. Pardoux. Muller's ratchet clicks in finite time. *Stochastic Processes and their Applications*, 123(6):2370–2397, 2013.

[10] F. J. Ayala. Evolution. *https://www.britannica.com/science/evolution-scientific-theory*, 2024.

[11] M. Baar, A. Bovier, and N. Champagnat. From stochastic, individual-based models to the canonical equation of adaptive dynamics in one step. *The Annals of Applied Probability*, 27(2):1093 – 1170, 2017.

[12] D. J. Balick, R. Do, C. A. Cassa, D. Reich, and S. R. Sunyaev. Dominance of Deleterious Alleles Controls the Response to a Population Bottleneck. *PLOS Genetics*, 11(8):e1005436, 2015.

[13] M. A. Ballinger and M. A. F. Noor. Are Lethal Alleles Too Abundant in Humans? *Trends in Genetics*, 34(2):87–89, 2018.

[14] F. Baumdicker, G. Bisschop, D. Goldstein, G. Gower, and A. P. Ragsdale, et. al. Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*, 220(3):iyab229, 12 2021.

[15] G. Bell. *The Masterpiece of Nature: The Evolution and Genetics of Sexuality*. Croom Helm, vi + 638 pp., 1982.

[16] H. Bernstein, F. A. Hopf, and R. E. Michod. The molecular basis of the evolution of sex. *Advanced Genetics*, 24:323–370, 1987.

[17] J. Bezanson, A. Edelman, and S. Karpinski. Why we created Julia. *https://web.archive.org/web/20200502144010/https://julialang.org/blog/2012/02/why-we-created-julia/*, 2012.

[18] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. Julia: A fresh approach to numerical computing. *SIAM Review*, 59(1):65–98, 2017.

[19] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. The Julia Programming Language. *https://www.julialang.org*, 2024.

[20] S. Biglari, A. Biglari, and S. Mazloomzadeh. The Frequency of Consanguinity and Its Related Factors in Parents of Children with Genetic Disorders. *Journal of Advances in Medical and Biomedical Research*, 30(143):501–506, 2022.

[21] S. Billiard, V. Castric, and V. Llaurens. The integrative biology of genetic dominance. *Biological reviews of the Cambridge Philosophical Society*, 96(6):2925–2942, 2021.

[22] A. H. Bittles and M. L. Black. Evolution in health and medicine Sackler colloquium: Consanguinity, human evolution, and complex diseases. *PNAS*, 107 Suppl 1(Suppl 1):1779–1786, 2010.

[23] D. Boffelli and D. I. K. Martin. Epigenetic inheritance: a contributor to species differentiation? *DNA and Cell Biology*, 31(Suppl 1):S11–S16, 2012.

[24] A. Bovier and F. den Hollander. *Metastability: A Potential-Theoretic Approach*. Die Grundlehren der mathematischen Wissenschaften. Springer International Publishing, xxi + 581 pp., 2015.

[25] A. Bovier and A. Kraut. Stochastic individual based models: From adaptive dynamics to modelling of cancer therapies. *https://wt.iam.uni-bonn.de/bovier/lecture-notes*, ix + 148 pp., 2019. Lecture Notes.

[26] A. Bovier, R. Neukirch, and L. Coquille. The recovery of a recessive allele in a mendelian diploid model. *Journal of Mathematical Biology*, 77(4):971–1033, 2018.

[27] G. E. Box. Robustness in the strategy of scientific model building. In *Robustness in statistics*, pages 201–236. Elsevier, 1979.

[28] K. M. Boycott, J. S. Parboosingh, B. N. Chodirker, R. B. Lowry, D. R. McLeod, J. Morris, C. R. Greenberg, A. E. Chudley, F. P. Bernier, J. Midgley, L. B. Moller, and A. M. Innes. Clinical genetics and the Hutterite population: a review of Mendelian disorders. *American Journal of Medical Genetics*, 146A(8):1088–1098, 2008.

[29] Y. Brandvain and S. I. Wright. The Limits of Natural Selection in a Nonequilibrium World. *Trends in Genetics*, 32(4):201–210, 2016.

[30] A. Bryant. MATLAB, R, and Julia: Languages for data analysis. *https://web.archive.org/web/20140426110631/https://strata.oreilly.com/2012/10/matlab-r-julia-languages-for-data-analysis.html*, 2012.

[31] R. Bürger. *The Mathematical Theory of Selection, Recombination, and Mutation.* Wiley Series in Mathematical and Computational Biology. John Wiley and Sons, Ltd., Chichester, xii + 424 pp., 2000.

[32] BVerfG. Beschluss des Zweiten Senats vom 26. Februar 2008. *https://www.bverfg.de/e/rs20080226_2bvr039207.html*, 2 BvR 392/07,:Rn. 1–128, 2008.

[33] A. Caballero. Developments in the prediction of effective population size. *Heredity*, 73(6):657–679, 1994.

[34] C. Cannings. The latent roots of certain Markov chains arising in genetics: A new approach, I. Haploid Models. *Advances in Applied Probability*, 6(2):260–290, 1974.

[35] C. Cannings. The Latent Roots of Certain Markov Chains Arising in Genetics: A New Approach, II. Further Haploid Models. *Advances in Applied Probability*, 7(2):264–282, 1975.

[36] Y. Cao, H. Li, and L. Petzold. Efficient formulation of the stochastic simulation algorithm for chemically reacting systems. *Journal of Chemical Physics*, 121(9):4059 – 4067, 2004.

[37] R. Chakraborty and A. Chakravarti. On consanguineous marriages and the genetic load. *Human Genetics*, 36(1):47–54, 1977.

[38] N. Champagnat. A microscopic interpretation for adaptive dynamics trait substitution sequence models. *Stochastic Processes and their Applications*, 116(8):1127–1160, 2006.

[39] N. Champagnat, R. Ferrière, and G. Ben Arous. The canonical equation of adaptive dynamics: a mathematical view. *Selection*, 2:73–83, 2001.

[40] N. Champagnat and S. Méléard. Polymorphic evolution sequence and evolutionary branching. *Probability Theory and Related Fields*, 151(1):45–94, 2011.

[41] B. Charlesworth. Model for evolution of Y chromosomes and dosage compensation. *PNAS*, 75(11):5618–5622, 1978.

[42] B. Charlesworth. The evolution of chromosomal sex determination and dosage compensation. *Current Biology*, 6(2):149–162, 1996.

[43] B. Charlesworth. The effects of deleterious mutations on evolution at linked sites. *Genetics*, 190(1):5–22, 2012.

[44] B. Charlesworth. Why we are not dead one hundred times over. *Evolution*, 67(11):3354–3361, 2013.

[45] B. Charlesworth and D. Charlesworth. Rapid fixation of deleterious alleles can be caused by Muller's ratchet. *Genetics Research*, 70(1):63–73, 1997.

[46] B. Charlesworth and D. Charlesworth. Some evolutionary consequences of deleterious mutations. *Genetica*, 102-103(1-6):3–19, 1998.

[47] M. Cieslak and P. Prusinkiewicz. Gillespie-Lindenmayer systems for stochastic simulation of morphogenesis. *in silico Plants*, 1(1):diz009, 2019.

[48] T. H. Clutton-Brock. Mammalian mating systems. *Proceedings of the Royal Society B: Biological Sciences*, 236(1285):339–372, 1989.

[49] P. Collet, S. Méléard, and J. A. J. Metz. A rigorous model study of the adaptive dynamics of Mendelian diploids. *Journal of Mathematical Biology*, 67:569–607, 2013.

[50] C. E. Correns. G. Mendel's Regel über das Verhalten der Nachkommenschaft der Rassenbastarde. *Berichte der Deutschen Botanischen Gesellschaft*, 18(4):158–168, 1900.

[51] J. F. Crow. Advantages of sexual reproduction. *Developmental Genetics*, 15(3):205–213, 1994.

[52] J. F. Crow and M. Kimura. *An Introduction To Population Genetics Theory*. Harper and Row, iv + 609 pp., 1970.

[53] S. Danisch and J. Krumbiegel. Makie.jl: Flexible high-performance data visualization for Julia. *Journal of Open Source Software*, 6(65):3349, 2021.

[54] C. Darwin. *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. John Murray, London, xiv + 502 pp., 1859.

[55] K. J. Dawson. The dynamics of infinitesimally rare alleles, applied to the evolution of mutation rates and the expression of deleterious mutations. *Theoretical Population Biology*, 55(1):1–22, 1999.

[56] H. de Vries. Das Spaltungsgesetz der Bastarde. *Berichte der Deutschen Botanischen Gesellschaft*, 18(3):83–90, 1900.

[57] U. Dieckmann and M. Doebeli. On the origin of species by sympatric speciation. *Nature*, 400:354–357, 1999.

[58] U. Dieckmann and R. Law. The dynamical theory of coevolution: a derivation from stochastic ecological processes. *Journal of Mathematical Biology*, 34(5):579–612, 1996.

[59] M. Doebeli. Quantitative Genetics and Population Dynamics. *Evolution*, 50(2):532–546, 1996.

[60] J. W. Drake. A constant rate of spontaneous mutation in DNA-based microbes. *PNAS*, 88(16):7160–7164, 1991.

[61] S. R. Eichten, R. J. Schmitz, and N. M. Springer. Epigenetics: Beyond Chromatin Modifications and Complex Genetic Regulation. *Plant Physiology*, 165(3):933–947, 2014.

[62] S. Engen, T. H. Ringsby, B.-E. Saether, R. Lande, H. Jensen, M. Lillegärd, and H. Ellegren. Effective size of fluctuating populations with two sexes and overlapping generations. *Evolution*, 61(8):1873–1885, 2007.

[63] A. Etheridge. *Some Mathematical Models from Population Genetics: École d'Été de Probabilités de Saint-Flour XXXIX-2009*. Springer, viii + 119 pp., 2011.

[64] A. Etheridge, P. Pfaffelhuber, and A. Wakolbinger. How often does the ratchet click? Facts, heuristics, asymptotics. https://arxiv.org/abs/0709.2775, 2007.

[65] S. N. Ethier and M. F. Norman. Error estimate for the diffusion approximation of the Wright–Fisher model. *PNAS*, 74(11):5096–5098, 1977.

[66] W. J. Ewens. *Mathematical Population Genetics. I*, volume 27. Springer, New York, xx + 418 pp., second edition, 2004.

[67] J. Felsenstein. The evolutionary advantage of recombination. *Genetics*, 78(2):737–756, 1974.

[68] R. A. Fisher. The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 42:399–433, 1918.

[69] R. A. Fisher. *The genetical theory of natural selection*. Oxford, Clarendon Press, xii + 286 pp., 1930.

[70] R. A. Fisher. The wave of advance of advantageous genes. *Annals of Eugenics*, 7(4):355–369, 1937.

[71] W. H. Fleming and M. Viot. Some Measure-Valued Markov Processes in Population Genetics Theory. *Indiana University Mathematics Journal*, 28(5):817–843, 1979.

[72] N. Fournier and S. Méléard. A microscopic probabilistic description of a locally regulated population and macroscopic approximations. *Annals of Applied Probability*, 14(4):1880–1919, 2004.

[73] F. Foutel-Rodier and A. M. Etheridge. The spatial Muller's ratchet: Surfing of deleterious mutations during range expansion. *Theoretical Population Biology*, 135:19–31, 2020.

[74] R. Frankham, J. D. Ballou, D. A. Briscoe, and K. H. McInnes. *Introduction to Conservation Genetics*. Cambridge University Press, xx + 644 pp., 2002.

[75] R. E. Franklin and R. G. Gosling. Molecular Configuration in Sodium Thymonucleate. *Nature*, 171:740–741, 1953.

[76] Z. Gao, D. Waggoner, M. Stephens, C. Ober, and M. Przeworski. An estimate of the average number of recessive lethal mutations carried by humans. *Genetics*, 199(4):1243–1254, 2015.

[77] M. Ghiselin. *The Economy of Nature and the Evolution of Sex*. University of California Press, 346 pp., 1974.

[78] M. A. Gibson and J. Bruck. Efficient exact stochastic simulation of chemical systems with many species and many channels. *Journal of Physical Chemistry A*, 104(9):1876 – 1889, 2000.

[79] K. J. Gilbert, F. Pouyet, L. Excoffier, and S. Peischl. Transition from Background Selection to Associative Overdominance Promotes Diversity in Regions of Low Recombination. *Current Biology*, 30(1):101–107, 2020.

[80] D. T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4):403 – 434, 1976.

[81] D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, 1977.

[82] D. T. Gillespie. Approximate accelerated stochastic simulation of chemically reacting systems. *Journal of Chemical Physics*, 115(4):1716 – 1733, 2001.

[83] D. T. Gillespie. Stochastic Simulation of Chemical Kinetics. *Annual Review of Physical Chemistry*, 58:35–55, 2007.

[84] D. T. Gillespie, A. Hellander, and L. R. Petzold. Perspective: Stochastic algorithms for chemical kinetics. *Journal of Chemical Physics*, 138(17), 2013.

[85] D. T. Gillespie and M. Mangel. Conditioned averages in chemical kinetics. *The Journal of Chemical Physics*, 75(2):704 – 709, 1981.

[86] R. Giugliani, F. Bender, R. Couto, A. Bochernitsan, and A. C. Brusius-Facchin, et. al. Population medical genetics: translating science to the community. *Genetics and Molecular Biology*, 42(1):312–320, 2019.

[87] S. Glémin. How are deleterious mutations purged? Drift versus nonrandom mating. *Evolution*, 57(12):2678–2687, 2003.

[88] M. R. Goddard, D. Greig, and A. Burt. Outcrossed sex allows a selfish gene to invade yeast populations. *Proceedings of the Royal Society B: Biological Sciences*, 268(1485):2537–2542, 2001.

[89] A. González-Casanova, C. Smadi, and A. Wakolbinger. Quasi-equilibria and click times for a variant of Muller's ratchet. *Electronic Journal of Probability*, 28:1 – 37, 2023.

[90] I. Gordo and B. Charlesworth. The degeneration of asexual haploid populations and the speed of Muller's ratchet. *Genetics*, 154(3):1379–1387, 2000.

[91] J. Goutsias and G. Jenkinson. Markovian dynamics on complex reaction networks. *Physics Reports*, 529(2):199 – 264, 2013.

[92] S. Gravel, B. M. Henn, R. N. Gutenkunst, A. R. Indap, G. T. Marth, A. G. Clark, F. Yu, and R. A. Gibbs, et. al. Demographic history and rare allele sharing among human populations. *PNAS*, 108(29):11983–11988, 2011.

[93] A. Gulani and W. T. *Genetics, Autosomal Recessive*. StatPearls Publishing, 2023.

[94] J. Gunawardena. Models in biology: 'accurate descriptions of our pathetic thinking'. *BMC Biology*, 12(1):29, 2014.

[95] R. N. Gutenkunst, R. D. Hernandez, S. H. Williamson, and C. D. Bustamante. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLOS Genetics*, 5(10):1–11, 2009.

[96] J. Haigh. The accumulation of deleterious genes in a population—Muller's Ratchet. *Theoretical Population Biology*, 14(2):251–267, 1978.

[97] J. B. S. Haldane. A Mathematical Theory of Natural and Artificial Selection, Part V: Selection and Mutation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 23(7):838–844, 1927.

[98] B. C. Haller and P. W. Messer. *SLiM: An Evolutionary Simulation Framework*, 2018. http://benhaller.com/slim/SLiM_Manual.pdf.

[99] B. C. Haller and P. W. Messer. SLiM 3: Forward genetic simulations beyond the Wright-Fisher model. *Molecular Biology and Evolution*, 36(3):632–637, 2019.

[100] H. Hamamy. Consanguineous marriages. *Journal of Community Genetics*, 3(3):185–192, 2012.

[101] G. H. Hardy. Mendelian Proportions in a Mixed Population. *Science*, 28(706):49–50, 1908.

[102] M. Hartfield and P. D. Keightley. Current hypotheses for the evolution of sex and recombination. *Integrative Zoology*, 7(2):192–209, 2012.

[103] B. M. Henn, L. R. Botigué, C. D. Bustamante, A. G. Clark, and S. Gravel. Estimating the mutation load in human genomes. *Nature Reviews Genetics*, 16(6):333–343, 2015.

[104] D. J. Higham. Modeling and Simulating Chemical Reactions. *SIAM Review*, 50(2):347–368, 2008.

[105] J. Hofbauer and K. Sigmund. *Evolutionary Games and Population Dynamics*. Cambridge University Press, xxii + 323 pp., 1998.

[106] H. Hu, K. Kahrizi, L. Musante, Z. Fattahi, R. Herwig, and M. Hosseini, et. al. Genetics of intellectual disability in consanguineous families. *Molecular Psychiatry*, 24(7):1027–1039, 2019.

[107] R. Jamra. Genetics of autosomal recessive intellectual disability. *Journal of Medical Genetics*, 30(3):323–327, 2018.

[108] X. Ji, R. L. Kember, C. D. Brown, and M. Bućan. Increased burden of deleterious variants in essential genes in autism spectrum disorder. *PNAS*, 113(52):15054–15059, 2016.

[109] S. G. Johnson. PyCall.jl: Calling Python functions from the Julia language. `https://github.com/JuliaPy/PyCall.jl`, 2012.

[110] L. B. Jorde and S. P. Wooding. Genetic variation, classification and 'race'. *Nature Genetics*, 36(11 Suppl):S28–S33, 2004.

[111] K. Kahrizi, H. Hu, M. Hosseini, V. M. Kalscheuer, and Z. Fattahi, et. al. Effect of inbreeding on intellectual disability revisited by trio sequencing. *Clinical Genetics*, 95(1):151–159, 2019.

[112] P. J. Keeling and J. D. Palmer. Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics*, 9(8):605–618, 2008.

[113] P. D. Keightley and S. P. Otto. Interference among deleterious mutations favours sex and recombination in finite populations. *Nature*, 443:89–92, 2006.

[114] W. O. Kermack, A. G. McKendrick, and G. T. Walker. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 115(772):700–721, 1927.

[115] M. Kimura. Solution of a process of random genetic drift with a continuous model. *PNAS*, 41(3):144–150, 1955.

[116] M. Kimura. On the evolutionary adjustment of spontaneous mutation rates. *Genetical Research*, 9(1):23–34, 1967.

[117] M. Kimura and T. Maruyama. The mutational load with epistatic gene interactions in fitness. *Genetics*, 54(6):1337–1351, 1966.

[118] J. F. C. Kingman. On the Genealogy of Large Populations. *Journal of Applied Probability*, 19:27–43, 1982.

[119] M. Kircher, D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper, and J. Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46(3):310–315, 2014.

[120] M. Kirkpatrick and P. Jarne. The Effects of a Bottleneck on Inbreeding Depression and the Genetic Load. *The American Naturalist*, 155(2):154–167, 2000.

[121] É. Kisdi and S. A. H. Geritz. Adaptive dynamics in allele space: Evolution of genetic polymorphism by small mutations in a heterogeneous environment. *Evolution*, 53(4):993–1008, 1999.

[122] K. Kochinke, C. Zweier, B. Nijhof, M. Fenckova, P. Cizek, F. Honti, S. Keerthikumar, M. A. W. Oortveld, T. Kleefstra, J. M. Kramer, C. Webber, M. A. Huynen, and A. Schenck. Systematic Phenomics Analysis Deconvolutes Genes Mutated in Intellectual Disability into Biologically Coherent Modules. *American Journal of Human Genetics*, 98(1):149–164, 2016.

[123] A. Kolmogorov, I. Petrovsky, and N. Piskunov. Investigation of the Equation of Diffusion Combined with Increasing of the Substance and Its Application to a Biology Problem. *Bulletin of Moscow State University Series A: Mathematics and Mechanics*, 1:1–25, 1937.

[124] A. S. Kondrashov. Selection against harmful mutations in large sexual and asexual populations. *Genetical Research*, 40(3):325–332, 1982.

[125] A. S. Kondrashov. Classification of hypotheses on the advantage of amphimixis. *Journal of Heredity*, 84(5):372–387, 1993.

[126] A. S. Kondrashov. Modifiers of mutation-selection balance: general approach and the evolution of mutation rates. *Genetical Research*, 66(1):53–69, 1995.

[127] A. S. Kondrashov. *Crumbling Genome: The Impact of Deleterious Mutations on Humans*. Wiley, xv + 281 pp., 2017.

[128] A. S. Kondrashov. Through Sex, Nature Is Telling Us Something Important. *Trends in Genetics*, 34(5):352–361, 2018.

[129] A. Kong, M. L. Frigge, G. Masson, S. Besenbacher, and P. Sulem, et. al. Rate of de novo mutations and the importance of father's age to disease risk. *Nature*, 488:471–475, 2012.

[130] A. Kong, D. F. Gudbjartsson, J. Sainz, G. M. Jonsdottir, and S. A. Gudjonsson, et. al. A high-resolution recombination map of the human genome. *Nature Genetics*, 31(3):241—247, 2002.

[131] P. Krill. New Julia language seeks to be the C for scientists. *https://web.archive.org/web/20140913234252/http://www.infoworld.com/d/application-development/new-julia-language-seeks-be-the-c-scientists-190818*, 2012.

[132] S. M. Krone and C. Neuhauser. Ancestral Processes with Selection. *Theoretical Population Biology*, 51(3):210–237, 1997.

[133] T. Kwong. *Hands-On Design Patterns and Best Practices with Julia*. Packt Publishing, xii + 532 pp., 2020.

[134] L. A. La Rocca, J. Frank, H. B. Bentzen, J. T. Pantel, K. Gerischer, A. Bovier, and P. M. Krawitz. Drop of Prevalence after Population Expansion: A lower prevalence for recessive disorders in a random mating population is a transient phenomenon during and after a growth phase. https://doi.org/10.1101/2021.09.29.462290, 2021.

[135] L. A. La Rocca, J. Frank, H. B. Bentzen, J. T. Pantel, K. Gerischer, A. Bovier, and P. M. Krawitz. Understanding recessive disease risk in multi-ethnic populations with different degrees of consanguinity. *American Journal of Medical Genetics*, 194(3):e63452, 2024.

[136] L. A. La Rocca, K. Gerischer, A. Bovier, and P. M. Krawitz. Refining the drift barrier hypothesis: a role of recessive gene count and an inhomogeneous Muller's ratchet. https://arxiv.org/abs/2406.09094, 2024.

[137] J.-B. Lamarck. *Philosophie Zoologique*. Dentu, Muséum d'Histoire Naturelle, x + 428 pp., 1809.

[138] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, and J. Baldwin, et. al. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[139] B. Lauwens and A. B. Downey. *Think Julia: How to Think Like a Computer Scientist*. O'Reilly Media, Inc., xxi + 295 pp., 2019.

[140] U. Lenz, S. Kluth, E. Baake, and A. Wakolbinger. Looking down in the ancestral selection graph: A probabilistic approach to the common ancestor type distribution. *Theoretical Population Biology*, 103:27–37, 2015.

[141] W. M. Lewis. Interruption of Synthesis as a Cost of Sex in Small Organisms. *The American Naturalist*, 121(6):825–833, 1983.

[142] W. M. Lewis. *The Evolution of Sex and its Consequences: The costs of sex*, volume 4. Birkhäuser Verlag, Basel, xv + 404 pp., 1987.

[143] M. I. Lind and F. Spagopoulou. Evolutionary consequences of epigenetic inheritance. *Heredity*, 121(3):205–209, 2018.

[144] H. Lodish, A. Berk, S. L. Zipursky, P. Matsudaira, D. Baltimore, and J. Darnell. *Molecular Cell Biology*. W.H. Freeman, xxxvi + 1084 pp., 4 edition, 1999.

[145] L. Loewe. Quantifying the genomic decay paradox due to Muller's ratchet in human mitochondrial DNA. *Genetics Research*, 87(2):133–159, 2006.

[146] K. E. Lohmueller, A. R. Indap, S. Schmidt, A. R. Boyko, R. D. Hernandez, M. J. Hubisz, J. J. Sninsky, T. J. White, S. R. Sunyaev, R. Nielsen, A. G. Clark, and C. D. Bustamante. Proportionally more deleterious genetic variation in European than in African populations. *Nature*, 451:994–997, 2008.

[147] A. Lomnicki. Carrying capacity, competition and maintenance of sexuality. *Evolutionary Ecology Research*, 3:603–610, 2001.

[148] A. J. Lotka. Quantitative Studies in Epidemiology. *Nature*, 88:497–498, 1912.

[149] M. Lynch, M. S. Ackerman, J.-F. Gout, H. Long, W. Sung, W. K. Thomas, and P. L. Foster. Genetic drift, selection and the evolution of the mutation rate. *Nature Reviews Genetics*, 17(11):704–714, 2016.

[150] T. R. Malthus. *An Essay on the Principle of Population as it Affects the Future Improvement of Society, with Remarks on the Speculations of Mr. Goodwin, M. Condorcet and Other Writers.* London: J. Johnson in St. Paul's Church-yard, xix + 134 pp., 1798.

[151] J. T. Manning. Diploidy and Muller's ratchet. *Acta Biotheoretica*, 32(4):289–292, 1983.

[152] M. Mariani, É. Pardoux, and A. Velleret. Metastability between the clicks of Muller's ratchet. https://arxiv.org/abs/2007.14715v3, 2020.

[153] H. C. Martin, W. D. Jones, R. McIntyre, G. Sanchez-Andrade, and M. Sanderson, et. al. Quantifying the contribution of recessive coding variation to developmental disorders. *Science*, 362(6419):1161–1164, 2018.

[154] N. Masuda and C. L. Vestergaard. *Gillespie Algorithms for Stochastic Multiagent Dynamics in Populations and Networks.* Elements in the Structure and Dynamics of Complex Networks. Cambridge University Press, xi + 107 pp., 2023.

[155] W. H. Mather, J. Hasty, and L. S. Tsimring. Fast stochastic algorithm for simulating evolutionary population dynamics. *Bioinformatics*, 28(9):1230–1238, 2012.

[156] S. Méléard and V. Bansaye. *Stochastic Models for Structured Populations: Scaling Limits and Long Time Behavior*. Springer International Publishing, x + 107 pp., 2015.

[157] G. Mendel. Versuche über Pflanzen-Hybriden. *Verhandlungen des Naturforschenden Vereines in Brünn*, 4:3–47, 1866.

[158] J. A. J. Metz. Adaptive dynamics. IIASA Interim Report IR-12-052, IIASA, Laxenburg, Austria, 2012.

[159] B. Modell and A. Darr. Genetic counselling and customary consanguineous marriage. *Nature Reviews Genetics*, 3(3):225–229, 2002.

[160] P. A. P. Moran. Random processes in genetics. *Mathematical Proceedings of the Cambridge Philosophical Society*, 54(1):60–71, 1958.

[161] H. J. Muller. Some Genetic Aspects of Sex. *The American Naturalist*, 66(703):118–138, 1932.

[162] H. J. Muller. The relation of recombination to mutational advance. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 1(1):2–9, 1964.

[163] L. Musante and H. H. Ropers. Genetics of recessive cognitive disorders. *Trends in Genetics*, 30(1):32–39, 2014.

[164] T. Nagylaki. *Introduction to Theoretical Population Genetics*. Springer Berlin, Heidelberg, xii + 369 pp., 1 edition, 1992.

[165] V. M. Narasimhan, K. A. Hunt, D. Mason, C. L. Baker, and K. J. Karczewski, et. al. Health and population effects of rare gene knockouts in adult humans with related parents. *Science*, 352(6284):474–477, 2016.

[166] M. Nei. Modification of linkage intensity by natural selection. *Genetics*, 57(3):625–641, 1967.

[167] M. Nei. The frequency distribution of lethal chromosomes in finite populations. *PNAS*, 60(2):517–524, 1968.

[168] R. Neukirch and A. Bovier. Survival of a recessive allele in a Mendelian diploid model. *Journal of Mathematical Biology*, 75(1):145–198, 2017.

[169] H. Ochman, J. G. Lawrence, and E. A. Groisman. Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405:299–304, 2000.

[170] K. O'Sullivan. Access to marriage: consanguinity and affinity prohibitions in national and international context. *Irish Journal of Family Law*, 22(2):8–12, 1 2019.

[171] S. P. Otto. The evolutionary enigma of sex. *Irish Journal of Family Law*, 174(Suppl 1):S1–S14, 2009.

[172] S. P. Otto and N. H. Barton. Selection for recombination in small populations. *Evolution*, 55(10):1921–1931, 2001.

[173] S. P. Otto and T. Lenormand. Resolving the paradox of sex and recombination. *Nature Reviews Genetics*, 3(4):252–261, 2002.

[174] J. Pahle. Biochemical simulations: Stochastic, approximate stochastic and hybrid approaches. *Briefings in Bioinformatics*, 10(1):53 – 64, 2009.

[175] S. Pálsson. The effects of deleterious mutations in cyclically parthenogenetic organisms. *Journal of Theoretical Biology*, 208(2):201–214, 2001.

[176] S. Pálsson and P. Pamilo. The effects of deleterious mutations on linked, neutral variation in small populations. *Genetics*, 153(1):475–483, 1999.

[177] M. Payne, C. A. Rupar, G. M. Siu, and V. M. Siu. Amish, mennonite, and hutterite genetic disorder database. *Paediatrics and Child Health*, 16(3):e23–e24, 2011.

[178] S. Peischl and L. Excoffier. Expansion load: recessive mutations and the role of standing genetic variation. *Molecular Ecology*, 24(9):2084–2094, 2015.

[179] P. Pfaffelhuber, P. R. Staab, and A. Wakolbinger. Muller's ratchet with compensatory mutations. *The Annals of Applied Probability*, 22(5):2108–2132, 2012.

[180] N. Phadnis and J. D. Fry. Widespread correlations between dominance and homozygous effects of mutations: implications for theories of dominance. *Genetics*, 171(1):385–392, 2005.

[181] W. B. Provine. *The origins of theoretical population genetics.* Chicago, University of Chicago Press, v + 240 pp., 1971.

[182] C. Rackauckas and Q. Nie. Differentialequations.jl–a performant and feature-rich ecosystem for solving differential equations in Julia. *Journal of Open Research Software*, 5(1):15, 2017.

[183] R. J. Redfield. *Do Bacteria Have Sex?*, chapter 19, pages 139–144. John Wiley and Sons, Ltd, 2012.

[184] W. R. Rice. Degeneration of a nonrecombining chromosome. *Science*, 263(5144):230–232, 1994.

[185] M. Roser and H. Ritchie. How has world population growth changed over time? *Our World in Data*, 2023. https://ourworldindata.org/population-growth-over-time.

[186] V. Rossi, A. Gandolfi, F. Baraldi, C. Bellavere, and P. Menozzi. Phylogenetic relationships of coexisting Heterocypris (Crustacea, Ostracoda) lineages with different reproductive modes from Lampedusa Island (Italy). *Molecular Phylogenetics and Evolution*, 44(3):1273–1283, 2007.

[187] J. Sang. Drosophila melanogaster: The Fruit Fly. In *Encyclopedia of Genetics*, page 23. Taylor and Francis, 2015.

[188] A. Schmidt, M. Danyel, K. Grundmann, T. Brunet, H. Klinkhammer, and T.-C. Hsieh, et. al. Next-generation phenotyping integrated in a national framework for patients with ultra-rare disorders improves genetic diagnostics and yields new molecular findings. *Nature Genetics*, 56:1644–1653, 2024.

[189] J. Schmidtke and M. C. Cornel. Contentious ethical issues in community genetics: let's talk about them. *Journal of Community Genetics*, 11(1):5–6, 2020.

[190] Y. B. Simons and G. Sella. The impact of recent population history on the deleterious mutation load in humans and close evolutionary relatives. *Current Opinion in Genetics and Development*, 41:150–158, 2016.

[191] S. J. Smale. On the differential equations of species in competition. *Journal of Mathematical Biology*, 3(1):5–7, 1976.

[192] J. Smith. *The Evolution of Sex*. Cambridge University Press, xi + 236 pp., 1978.

[193] W. Stephan, L. Chao, and J. G. Smale. The advance of Muller's ratchet in a haploid asexual population: approximate solutions based on diffusion theory. *Genetics Research*, 61(3):225–231, 1993.

[194] T. Strachan and A. Read. *Human Molecular Genetics*. Garland Science, xxii + 770 pp., 5th edition, 2018.

[195] S. Strome, N. Bhalla, R. Kamakaka, U. Sharma, and W. Sullivan. Clarifying Mendelian vs non-Mendelian inheritance. *Genetics*, 227(3):iyae078, 2024.

[196] W. Sung, M. S. Ackerman, S. F. Miller, T. G. Doak, and M. Lynch. Drift-barrier hypothesis and mutation-rate evolution. *PNAS*, 109(45):18488–18492, 2012.

[197] T. Székely and K. Burrage. Stochastic simulation in systems biology. *Computational and Structural Biotechnology Journal*, 12(20):14–25, 2014.

[198] N. Tagg, C. P. Doncaster, and D. J. Innes. Resource competition between genetically varied and genetically uniform populations of Daphnia pulex (Leydig): does sexual reproduction confer a short-term ecological advantage? *Biological Journal of the Linnean Society*, 85(1):111–123, 2005.

[199] A. Tenesa, P. Navarro, B. J. Hayes, D. L. Duffy, G. M. Clarke, M. E. Goddard, and P. M. Visscher. Recent human effective population size estimated from linkage disequilibrium. *Genome Research*, 17(4):520–526, 2007.

[200] C. C. Traverse and H. Ochman. Conserved rates and patterns of transcription errors across bacterial growth states and lifestyles. *PNAS*, 113(12):3311–3316, 2016.

[201] E. Tschermak. Über künstliche Kreuzung bei Pisum sativum. *Berichte der Deutschen Botanischen Gesellschaft*, 18(6):232–239, 1900.

[202] F. van der Plas, M. Dral, P. Berg, R. Huijzer, and M. Bocheńsk, et. al. Pluto.jl, 2024. https://doi.org/10.5281/zenodo.13329704.

[203] L. van Valen. A new evolutionary law. *Evolutionary Theroy*, 1(1):1–30, 1973.

[204] V. Volterra. Variations and Fluctuations of the Number of Individuals in Animal Species living together. *ICES Journal of Marine Science*, 3(1):3–51, 1928.

[205] R. C. Vrijenhoek. Animal Clones and Diversity: Are natural clones generalists or specialists? *BioScience*, 48(8):617–628, 1998.

[206] J. Wang, E. Santiago, and A. Caballero. Prediction and estimation of effective population size. *Heredity*, 117(4):193–206, 2016.

[207] J. D. Watson and F. H. C. Crick. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, 171:737–738, 1953.

[208] W. Weinberg. Über den Nachweis der Vererbung beim Menschen. *Jahreshefte des Vereins für vaterländische Naturkunde in Württemberg*, 64:368–382, 1908.

[209] S. A. West, C. M. Lively, and A. F. Read. A pluralist approach to sex and recombination. *Journal of Evolutionary Biology*, 12(6):1003–1012, 1999.

[210] M. H. F. Wilkins, A. R. Strokes, and H. R. Wilson. Molecular Structure of Nucleic Acids: Molecular Structure of Deoxypentose Nucleic Acids. *Nature*, 171:738–740, 1953.

[211] D. J. Wilkinson. Stochastic modelling for quantitative description of heterogeneous biological systems. *Nature Reviews Genetics*, 10(2):122–133, 2009.

[212] S. G. Wright. Coefficients of Inbreeding and Relationship. *The American Naturalist*, 56(645):330–338, 1922.

[213] S. G. Wright. Evolution in Mendelian populations. *Genetics*, 16:97–157, 1931.

[214] S. G. Wright. Breeding Structure of Populations in Relation to Speciation. *The American Naturalist*, 74(752):232–248, 1940.

[215] Q. Xiao and V. M. Lauschke. The prevalence, genetic complexity and population-specific founder effects of human autosomal recessive disorders. *NPJ Genomic Medicine*, 6(1):41, 2021.

[216] C. A. Yates and G. Klingbeil. Recycling random numbers in the stochastic simulation algorithm. *Journal of Chemical Physics*, 138(9):094103, 2013.

[217] B. Yuan, K. V. Schulze, N. Assia Batzir, J. Sinson, H. Dai, W. Zhu, F. Bocanegra, C.-T. Fong, J. Holder, J. Nguyen, C. P. Schaaf, Y. Yang, W. Bi, C. Eng, C. Shaw, J. R. Lupski, and P. Liu. Sequencing individual genomes with recurrent genomic disorder deletions: an approach to characterize genes for autosomal recessive rare disease traits. *Genome Medicine*, 14(1):113, 2022.

[218] G. A. Zagatti, S. A. Isaacson, C. Rackauckas, V. Ilin, S. Ng, and S. Bressan. Extending JumpProcesses.jl for fast point process simulation with time-varying intensities. *Proceedings of the JuliaCon Conferences*, 6(58):133, 2024.

[219] M. L. Zeeman. Hopf bifurcations in competitive three-dimensional Lotka–Volterra systems. *Dynamics and Stability of Systems*, 8(3):189–216, 1993.