Technische Universität München
TUM School of Life Sciences

Simulation-based Modeling of Biological Traits and Inference from the Ancestral Recombination Graph

Kevin Korfmann

Vollständiger Abdruck der von der TUM School of Life Sciences der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz:

Prof. Dr. Frank Johannes

Prüfer der Dissertation:

1. Prof. Dr. Aurélien Tellier
2. Hon.-Prof. Dr. Klaus Mayer
3. Prof. Dr. Dirk Metzler

Die Dissertation wurde am 02.10.2023 bei der Technischen Universität München eingereicht und durch die TUM School of Life Sciences am 08.01.2024 angenommen.

List of Figures

List of Tables

Acknowledgements

First and foremost, I'd like to thank my supervisor Prof. Dr. Aurélien Tellier for giving me the opportunity to contribute to the field of population genetics in the form of this doctoral project and for providing the environment in which scientific research can proliferate.

Next, I'd like to thank the International Graduate School of Science and Engineering (IGSSE) for providing the funding, along with countless possibilities for personal growth in the form of network opportunities or allowing me to move abroad.

Furthermore, the thesis wouldn't have been possible without the many co-authors, with whom it was a pleasure to collaborate. Namely, I'd like to thank and in no particular order: Diala Abu Awad, Thibaut Sellinger, Fabian Freund, Matteo Fumagalli, Marie Temple-Boyer, Oscar Gaggiotti and Aurélien Tellier.

In addition, I'd like to say thank you to my mentor Sona John, and the entire IGSSE cohort (GENOMIE_QADOP) for providing such a stimulating framework to work in.

Lastly, I extend my heartfelt gratitude to all those not explicitly mentioned — present and past lab members, friends, and family. The list is extensive, and I refrain from enumerating it to avoid inadvertently omitting anyone.

Publications and Licenses

The following part lists the publications, alongside their respective licences, which have come forward throughout this PhD project.

Chapter 1 was added by KK.

Chapter 2 has been published as:
Kevin Korfmann, Oscar E Gaggiotti, Matteo Fumagalli, Deep Learning in Population Genetics, Genome Biology and Evolution, Volume 15, Issue 2, February 2023, evad008, https://doi.org/10.1093/gbe/evad008

Chapter 3 has been published as:
Kevin Korfmann, Diala Abu Awad and Aurélien Tellier (2023). Weak seed banks influence the signature and detectability of selective sweeps. Journal of Evolutionary Biology, 36, 1282–1294. https://doi.org/10.1111/jeb.14204

The article has been peer-reviewed and is being recommended by PCI and the recommendation is available at https://evolbiol.peercommunityin.org/articles/rec?id=552.

Chapter 4 has been published as:
Kevin Korfmann, Marie Temple-Boyer, Thibaut Sellinger and Aurélien Tellier (2023). Determinants of rapid adaptation in species with large variance in offspring production. Molecular

Ecology, 00, 1– 14. https://doi.org/10.1111/mec.16982

The article of **Chapter 5** was under review, therefore only available on BioRxiv under:
Kevin Korfmann, Thibaut Sellinger, Fabian Freund, Matteo Fumagalli, Aurélien Tellier. Simultaneous Inference of Past Demography and Selection from the Ancestral Recombination Graph under the Beta Coalescent (2023) bioRxiv 2022.09.28.508873;
doi: https://doi.org/10.1101/2022.09.28.508873

A revised version is now published as:
Kevin Korfmann, Thibaut Sellinger, Fabian Freund, Matteo Fumagalli, Aurélien Tellier. Simultaneous Inference of Past Demography and Selection from the Ancestral Recombination Graph under the Beta Coalescent. Peer Community Journal, Volume 4 (2024), article no. e33. doi : https://doi.org/10.24072/pcjournal.397

Chapter 6 was added by KK.

Abstract

Population genetics is a field that imagines evolutionary trajectories to understand the sparse treasure of genetic diversity that has surfaced over time. This is achieved through modeling, either simulation-based or through rigorous mathematical formulation. The later modeling approach acknowledges the foundational work laid down by eminent figures in the discipline, notably Fisher, Haldane, Muller, and Wright. Their pioneering contributions spanned a broad spectrum, from the understanding of gene frequencies and genetic linkage to the intricate roles mutations play in shaping genetic diversity. These early ventures into the genetic underpinnings of populations provided the groundwork for what would later be recognized as the Modern Synthesis. From this area, the intuitive understanding of evolution can thus be modeled by Markovian and individual-based simulations of generation after generation.

Chronologically, however, and in the following years of Kimura's work on diffusion equations, which modeled allele frequencies through time, a pivotal shift in the study of population genetics came with the advent of the Coalescent Theory. This transformative approach revolutionized the analysis of DNA sequences by advocating for a methodology that traced genealogies backward in time. The following thesis stands on the shoulders of both perspectives and countless research endeavors, exploring further, the intuitive, so-called forward-in-time simulation-based strategy by studying process, where contemporary mathematical theory has not been formulated. Specifically, we dedicate Chapter 3 to the fascinating biological phenomenon of seed banking or dormancy. Here, organisms, ranging from bacteria to plants, have evolved mechanisms to store genetic information in a metabolically reduced or halted state. This ensures that the genetic material can be preserved and subsequently revived in more favorable conditions. We, specifically, highlight how seed banking decreases genetic drift, by scaling the ancestral recombination graph, both in terms of the frequency of coalescence and recombination events. There, implying the inception of prominent selective sweep signatures in depths and sharpness.

We continue to paint a more realistic picture of the complete life-cycle, by addressing the aspect of high offspring variance observed in a range of organisms. This phenomenon, where some species produce a large number of offspring with a correspondingly high rate of early mortality, introduces a layer of complexity to population genetics and more specifically acts in contrast to seed banks by increasing genetic drift as a function of the strength of fecundity.

Lastly, in the modern era, computational advancements have surged to the forefront of many scientific disciplines, and population genetics is no exception. The thesis underscores the

transformative potential of deep learning in the field by providing both an overview of the field in Chapter 2 and demonstrating inference capabilities in Chapter 4 of the above-mentioned highly stochastic environment enabled by multiple merger coalescence model, thereby directly benefiting from the extensions of early coalescent theory models, which now lay the basis for the necessary large datasets required to capture enough of the variation for making predictions regarding demographic and selective inferences.
In essence, this work provides a panoramic view of simulation-based modeling in population genetics. And alongside, it weaves together historical perspectives, mathematical intricacies, and cutting-edge computational methodologies to offer a comprehensive understanding of the genetic tapestry that underpins populations and their evolutionary journeys.

Zusammenfassung

Die Populationsgenetik ist ein Bereich, in dem evolutionäre Trajektorien erstellt werden. Ihr Ziel ist es, den spärlichen Schatz genetischer Vielfalt zu verstehen, der im Laufe der Zeit entstanden ist. Dies wird durch Modellierung erreicht, die entweder auf der Grundlage von Simulationen oder durch strenge mathematische Formulierung erfolgt. Der letztere Modellierungsansatz würdigt die grundlegende Arbeit, die von bedeutenden Persönlichkeiten der Disziplin, insbesondere Fisher, Haldane, Muller und Wright, geleistet wurde. Ihre bahnbrechenden Beiträge umfassten ein breites Spektrum, das vom Verständnis der Genhäufigkeit und der genetischen Verknüpfung bis hin zur komplizierten Rolle von Mutationen bei der Gestaltung der genetischen Vielfalt reichte. Diese frühen Erkenntnisse, die genetischen Grundlagen von Populationen darstellen, legten den Grundstein für das, was später als 'Moderne Synthese' bezeichnet wurde. Aus diesem Bereich stammt das intuitive Verständnis der Evolution, die durch Markovianische und individualbasierte Simulationen von Generation zu Generation modelliert werden kann.

Chronologisch gesehen kam es jedoch in den Folgejahren von Kimuras Arbeit an den Diffusionsgleichungen, die die Allelhäufigkeiten im Laufe der Zeit modellierten, mit dem Aufkommen der Koaleszenztheorie zu einem entscheidenden Wandel in der Untersuchung der Populationsgenetik. Dieser transformative Ansatz revolutionierte die Analyse von DNA-Sequenzen, indem er für eine Methodik plädierte, die Genealogien in der Zeit zurückverfolgte. Die folgende Arbeit steht auf den Schultern beider Perspektiven und zahlloser Forschungsbemühungen, indem sie die intuitive, so genannte zeitlich vorwärts gerichtete simulationsbasierte Strategie weiter erforscht und Prozesse untersucht, für die keine zeitgenössische mathematische Theorie formuliert wurde. Insbesondere widmen wir uns in Kapitel 3 dem faszinierenden biologischen Phänomen der Dormanz. Hier haben Organismen, von Bakterien bis hin zu Pflanzen, Mechanismen entwickelt, um genetische Informationen in einem metabolisch reduzierten oder angehaltenen Zustand zu speichern. Dadurch wird sichergestellt, dass das genetische Material erhalten bleibt und später unter günstigeren Bedingungen wiederbelebt werden kann. Wir zeigen insbesondere auf, wie Dormanz den genetischen Drift verringert, indem es den Ahnenrekominationsgraphen sowohl in Bezug auf die Häufigkeit der Koaleszenz- als auch der Rekombinationsereignisse skaliert. Dies impliziert das Auftreten auffälliger selektiver Sweep-Signaturen in Tiefe und Schärfe.

Wir fahren fort, ein realistischeres Bild des gesamten Lebenszyklus zu zeichnen, indem wir uns mit dem Aspekt der hohen Nachkommenschaftsvarianz befassen, die bei einer Reihe von

Organismen beobachtet wird. Dieses Phänomen, bei dem einige Arten eine große Anzahl von Nachkommen mit einer entsprechend hohen frühen Sterblichkeitsrate produzieren, führt eine neue Ebene der Komplexität in die Populationsgenetik ein und wirkt insbesondere im Gegensatz zur Dormanz, indem es den genetischen Drift in Abhängigkeit von der Anzahl der Nachkommen pro Individuum erhöht.
Schließlich sind in der modernen Ära rechnergestützte Fortschritte in vielen wissenschaftlichen Disziplinen in den Vordergrund gerückt, und die Populationsgenetik bildet hier keine Ausnahme. Die Dissertation unterstreicht das transformative Potenzial von neuronalen Netzwerken in diesem Bereich, indem sie sowohl einen Überblick über das Feld in Kapitel 2 bietet als auch in Kapitel 4 die Inferenzfähigkeiten des oben erwähnten hochstochastischen Umfelds demonstriert, das ein Koaleszenzmodell mit mehreren simultanen Koaleszenzen ermöglicht, und so direkt von den Erweiterungen der frühen Koaleszenztheoriemodelle profitiert, die nun die Grundlage für die notwendigen großen Datensätze bilden, die erforderlich sind, um genügend Variation für Vorhersagen hinsichtlich demografischer und selektiver Inferenzen zu erfassen.
Im Wesentlichen bietet diese Arbeit einen Panoramablick auf die simulationsbasierte Modellierung in der Populationsgenetik. Daneben werden historische Perspektiven, mathematische Formulierungen und hochmoderne Berechnungsmethoden miteinander verwoben, um ein umfassendes Verständnis der genetischen Diversität zu vermitteln, die Populationen und deren evolutionäre Entwicklung untersteht.

Contents

1 Introduction

1.1 Motivation

The nature of simulation-based modeling is constrained by the number or dimensions of model-able interacting or non-interacting layers which aim to explain some phenomena. In consequence, the space of possible evolutionary histories is far larger than can be appreciated at first glance. Further, when comparing a variety of modelable processes, the arising simulation is a product of each process contributing a stochastic amount of signal. The result is complex and difficult to interpret or to model mathematically *a priori* especially if the number of dimensions increases. Yet, the variation, that we observe in nature represents an often one-time experiment of all the hidden stochastic processes that have occurred over millions of years, of which only traces have been captured by means of sequencing or historically-electrophoretic technologies. Thus, an open question, when facing any dataset is to what kind of degree a particular process has shaped the observed variation, how many processes are involved, and how their contributions shaped the outcome. This is further complicated by difficulties of the non-identifiability by indifferentiable processes. Despite the complexity, it was proven successful to come up with expectations and theories about the nature of aspects of the observed variation through the mathematical formulation and identification of some of the dimensions throughout the last century. To highlight some historical insights into the field as well as early and mid-century controversies or open questions, we start by summarizing the history in an overview before moving on to the promise of simulation-based modeling of multiple dimensions in the form of biological traits and their interacting genomic processes like recombination or external forces in the form of positive selection and moving to inference by non-model-based deep learning (DL) methodology, as well as giving an extended introduction to the current state of DL inference approaches in population genetics.

1.2 Early Historical Introduction to Population Genetics

Much to the benefit of population genetics the field has been rigorously conceptualized from its inception (Sarkar and Cohen, 1992). Roughly one hundred years ago the retrospectively named founders Fisher, Haldane, Muller, and Wright had the foresight to derive the fundamental concepts of *i.e.* gene frequencies (Fisher, 1919), genetic linkage and drift (Haldane,

1990), dominance effects (Fisher, 1922; Charlesworth, 2022), the role of mutations (Muller, 1927, 1932), genetic load (Muller, 1950; Haldane, 1937) and much more, despite unavailable and yet to derive knowledge of the precise underlying biological machinery, including the discovery of DNA sequence structure decades later (Watson and Crick, 1953). Starting from the rediscovered Mendelian ratios, quantitative observations, and the emerging acceptance of linearly arranged genes (Sturtevant, 1913), Fisher laid down the foundations of population genetics in his historic paper *On the dominance ratio*, followed by the development of statistical methods and their application in series of threes papers in 1922 (Fisher, 1922). The first key ideas of Fisher's foundational paper conceptualized the first description of selectively maintained variation at a heterozygous site, known as *balancing selection*, followed by the study of the survivability of selective mutations and their chance of loss in large populations and stochastic frequency changes in finite populations. While Fisher continued to extend his study towards dominance and adjacent topics culminating in his influential book on *The Genetical Theory of Natural Selection* (Fisher, 1930) Wright likewise introduced fundamental aspects of the field. In 1922, he introduced the inbreeding coefficient, known as a measure of the probability that two alleles are identical by descent while highlighting implications on the fitness as a result of increased homozygosity and thus reduced genetic variation (Wright, 1922). Later, in his lengthy landmark publication *Evolution in Mendelian Populations* (Wright, 1931) Wright introduced most-notably among many other topics the concept of random fluctuation of allele frequencies or *genetic drift*, which was directly impacted by the effective population size, likewise has been in introduced in the same publication. Other contributions in later years include the notation of the adaptive landscape or population structure, setting up the framework Wrights' *Shifting Balance Theory* (Wright, 1931, 1932). The essence of this postulation can be summarized in an attempt to explain how populations caught in sub-optimal areas in the adaptive (epistatic) fitness landscape can still transition to higher fitness peaks essentially through migration leading to one of the earlier controversies of the field due to an unresolved disagreement with Fisher. Staying for a while longer in the pre-50s area of population genetics, Haldanes and Muller's work remains to be acknowledged and is no less important. Continuing with the last mathematician of the four, Haldane investigated the effect of selection starting from a *simple Mendelian population* defining concepts like fitness, gene frequency, and selection coefficients to show that the rate of selection is capable of explaining genetic variation (Haldane, 1990). Later, Haldane went on to discuss the above-mentioned concept of *genetic load*, which measures the cost of deleterious alleles in a population. Yet, one of his earliest contributions evolved around the formulation of genetic linkage, which later have been extended by Maynard Smith with regard to selection. The last of the early contributors, being the only geneticist, among the four above-named pioneers, came with the most fundamental concepts, largely taken for granted by today, with early work centering on the conceptualization of a gene, and intrinsically linked concept of genetic inheritance, description of the evolutionary value of recombination, alongside the

later famous idea of "Muller's ratchet", further emphasizing the value of sexual reproduction as a mechanism to counteract the accumulation of deleterious mutations. Unfortunately, this summary represents only a small introduction of the highly significant contributions of the first half of the 20th century and many ideas can not be named or addressed. However, the shared contributions above ultimately culminated in the formulation of the *Modern Synthesis* (Huxley, 1942), interweaving the predominate fields of Mendelian Genetics with a Darwinian view of evolution. This first theory, though often criticized by today's standards, offered the broader field of evolutionary biology the opportunity of a rigid physics-oriented framework centered around the principles of genetic drift, gene flow, mutational pressure, and selective forces. Holding onto a unifying idea at the center of evolution also enabled the possibility for conscientious scientific investigation and lively discussion and *controversies* of the precise underlying mechanisms (Crow, 2008). One such discussion tried to explain the underlying performance increase of hybrids attributing it either to a dominance phenomenon, a kind of rescue event of deleterious recessive alleles or to overdominance, in which heterozygotes inherently have higher fitness. This discussion as well as the 1955 introduced conversation of the "classical vs. balance hypothesis" led primarily by disagreements by Dobzhansky and Muller, were examples of the experimentally-driven investigation.

Following this era, Kimura shifted the conversation towards the investigation of genetic variation by means of studying diffusion equations, which led to many theoretical expectations in population genetics and most notably led to the formulation of the "neutral theory" and to the description of the rate of evolution using a molecular clock metaphor. As Hahn, however, points out the misinterpretation of the at the time controversial theory, does not mean that the majority of mutations, that arise are associated with a selection coefficient equal to zero, but that alternative alleles in the population have the same selective potential if they happened to be substituted by genetic drift (Hahn, 2018). From today's perspective, the theory is often regarded as a null model for molecular evolution and remains to be accepted as realistic for non-coding or non-functional locations in genomes. Following Kimura's work has been the introduction of a fundamental shift of perspective in the modeling of evolution. Kimuras continued the tradition of letting the evolutionary process unfold itself forward in time, tracing the loss or fixation of an allele. However, *Coalescent Theory* turned this point of view upside-down, and starting from the present traced events of ancestral coalescence of a pool of samples. This mathematical shortcut or approximation leads to many simplifications for the field of large-scale data generation, essential for modern inference methods. A summary of the theory is presented below.

1.3 Coalescent Theory and the State of Evolutionary Simulations

The introduction of *Coalescent Theory* can be regarded as a scientific breakthrough in the field due to the accompanying perspective change of looking at a sample of DNA sequences and letting the genealogy emerge backward in-time. It, thus, provides the possibility of approximating the intuitive but time-consuming way of a continuous and forward in-time evolving process, like it has been historically introduced through the early models described by the *founders* and extended by Kimura's work on diffusion models. Kingman described in his 1982 paper the process of samples which *coalesce* at different times in the past at a given rate ($\frac{1}{2N_e}$), notability the distribution of coalescence events depends on the effective population time (N_e), thus providing a theoretical framework which can be used for parameter estimation, *i.e.* population size changes through time. Extension of the coalescence process quickly followed, generating the possibility of a fast way of generating data, which mirrors real observed variation in accordance with the studied process.

Most notably extensions include the addition of recombination events (Hudson, 1983; Griffiths and Marjoram, 1996), effectively describing the opposite of a coalescence event or bifurcation of a lineage traced backward in time, leading to a graph structure known as the *ancestral recombination graph* (ARG) onto whose branches mutations fall at a given rate (μ) measured per generations per site, accompanied by the recombination rate (r) per generation per site. In this context, the effective population size (N_e), reflecting the number of breeding individuals in an idealized population, scales both μ and r by $4N_e$ to get an estimation of the population-comparable genetic variability in a population due to mutation ($\theta = 4N_e\mu$) and recombination ($\rho = 4N_e r$), respectively.

Another straightforward extension of the original Kingman coalescence model is the generalization of allowing multiple coalescence events as opposed to strictly binary occurrences. The family of models known as multiple merger models due to the mentioned multiple coalescence appearance (Tellier and Lemaire, 2014) resulted in a need for modeling the ecological more realistic offspring distribution of species with sweepstake reproduction capabilities, such as virus, fungi, plant, mollusk or fish species. Further complications of the coalescent deal with the simulation of selection (Krone and Neuhauser, 1997; Neuhauser and Krone, 1997), yet these extensions rely on the mathematical formulation of simultaneous processes, *i.e.* demographic changes under selection with non-WF offspring distribution and recombination, often remains difficult to derive and implement.

Despite the early-on perceived promise of efficient ancestral simulations, the exact coalescent (with recombination) only became simulated-able for large genome sequences and sample sizes under low memory and importantly time constraints after the introduction of *sparse trees* and *coalescent records* in the form the *succinct tree sequence* data structure

by Kelleher's reimplementation of Hudson's algorithm (Kelleher et al., 2016). Up till then, the predominant understanding happened to be that the state-space of ARGs (Griffiths and Marjoram, 1996) is too large to be simulated, which is why much effort was put into developing the approximation of the underlying ancestral process by modeling recombination events along the sequence (Wiuf and Hein, 1999). This viewpoint was picked up later by providing the algorithmic formulation, being the sequentially Markov coalescent (SMC) (McVean and Cardin, 2005), which was improved a year later in the form of the SMC' (Marjoram and Wall, 2006). The class of SMC algorithms is known to give a close approximation of the coalescent, but suffer to capture long-range linkage correlations. Yet, their intrinsic Markovian property gave rise to powerful HMM inference algorithms, discussed below. However, coalescent simulations, as elegant as they are, persist not only to be useful only for scenarios of limited complexity or as a starting point for hybrid coalescent and forward simulation approaches or for verification of theoretical expectations of neutral base models, but also for DL required data generation.

In case the studied model does not require estimations from the underlying genealogy, (markovian) individual-based simulations of each consecutive generation of a given evolutionary process can be sufficient to create data sets. This more historic approach has become more popular again with the introduction of SLiM, a programmable evolutionary simulation framework (Messer, 2013). This *modus operandi*, which has been named *forward in-time simulations* above is suitable for modeling complicated interaction of simultaneous processes as each of them can be tightly linked to the precise and known biological assumptions. Later versions have added the possibility of adding a *tree recording feature*, which added support for keeping track of ARGs as well if required (Haller et al., 2019b). Further, the programmatic aspect makes it intuitive to study statistical distributions for the offspring generations, which deviate from the standard Wright-Fisher assumption of having approximately a Poisson number of offspring with a rate parameter of one for mean and variance. This kind of model has been termed the Cannings model and is described in detail in Chapter 4 for the study of large offspring variance distributions interacting with various kinds of selective processes (Cannings, 1974).

With the availability of coalescent and computational frameworks or data structures, the study into signatures of ecological traits is enabled, in particular, if the coalescent formulation of base models exists. One such trait is summarized under the term dormancy, which effectively describes a sort of inactive or resting stage. Because this trait is fundamental and easily abstracted, it can be found ubiquitously in resting cells (*i.e.* cancer cells or bacteria) to more complex biological structures (*i.e.* spores or seeds). A short introduction of the trait, also otherwise known as seed banking, highlighting the life-cycle adaptation applicability part follows below as a precursor for Chapter 3.

1.4 Seed Banking

Seed banking or dormancy is the biological phenomenon of storing genetic information in a metabolically -reduced or halted state and structure in order for the system to resuscitate at a later generation. Depending on the time horizon spanning only a few (weak version, suitable *i.e.* for plants) up to thousands (strong model, suitable *i.e.* for bacteria or virus) of generations, two seed bank models have been proposed. Despite the straightforward extension of the neutral coalescent, in the form of *i.e.* demographic inference (see below), the consequences and complexity arising from a general model, which in its purest form simply describes a stochastically paused state of a biological object, implying wide-reaching interactions within the field of health (*i.e.* Malaria infections of *Plasmodium vivax* or dormant cancer cells) or in biography, in the form of dispersal or conservation efforts and breeding or biotechnological potential due to their storage of genetic diversity. Both mentioned seed bank models have their own coalescent formulation, which extends the Kingman coalescent model by this biological trait adding a dimension of complexity to the data, thus increasing the space of potential simulated-able or observable information or concretely altering branch length distributions of the underlying coalescent tree or ARG through a rescaling by $\frac{1}{b^2}$ (with b being the germination rate) in the case of the weak model or more complex and multiple coalescence-like signals in the strong version. Moreover, interactions regarding recombination have been formulated, which consequentially lead to informational signal alterations when adding selection modifiers to the sequence as the linkage is affected by the scaling of the underlying genealogy. It has previously been shown for the weak seed bank model, that recombination is scaled by a factor of $\frac{1}{b}$ (assuming no recombination during the dormant state), implying a faster linkage decay with a simultaneous decrease of genetic drift due to the storage effect of a allele, which only can be considered to be lost when it also leaves the seed bank, despite potentially not having any copies in the parent generation. A detailed introduction including references to derivations of seed bank diffusion approximations under selection can be found in Chapter 3. While seed banking scales the ARG and decreases genetic drift, having a lengthening effect on the effective life span, variations of the underlying offspring distributions can have an opposing effect and are introduced next.

1.5 Offspring Variance

Traditionally under Wright-Fisher conditions, the variation of offspring is considered to be low and often modeled through binomial sampling or approximated with a Poisson distribution. Yet, exchanging the probability function leads to the possibility of modeling populations other than humans, and thus explores a range of organisms with relevance to disease management or agricultural importance. The respective species can be biologically characterized as having stochastically large offspring numbers per individual and associated high infant mortality.

Ecologically, the class of species is thus known as choosing a type-III survivorship strategy. From a coalescence modeling perspective, the multiple merger coalescent models have been derived to model these organisms (see above). In Chapter 4, we explore the consequences of adaptation, when modeling the tailored life cycle in interaction with multiple positive selection models and the importance of distinguishing between somatic mutations (mutation emerging in offspring) or heritable mutations (mutation emerging in parents) into account. This, yet, adds another dimension of complexity increasing the space of observable genealogies, here increasing genetic drift.
In the next part we look at the topic of demographic changes and provide a short overview as it will be relevant for Chapter 4.

1.6 Demographic Inference and Beyond

A fundamental parameter of population genetics is the size of a population, often idealized as the effective population size (N_e), which defines the amount of genetic drift acting on a population. The size further leads to baseline understandings of the genetic diversity and by extension adaptive potential of populations by giving insights into the expected number of mutations or recombination events on the population level. On top of that, it characterizes the putative exchange between populations and is a prerequisite for understanding the level of gene flow events. Mean estimates of the effective population size can be approximated for constant population sizes from polymorphism data: $N_e = \frac{\theta_\pi}{4\mu}$. However, the coalescent framework directly enabled the inference of N_e through time as it modifies the effective rate of coalescence leading to longer or shorter branch length distributions and the respective accumulation of mutations. Having a more fine-scaled understanding of population size changes, *i.e.* population bottlenecks or expansion, not only leads to insights regarding the evolutionary history but also can be essential for the calibration of genome-wide scans of selection events. Therefore, a substantial literature exists consisting of methodological advances and their respective applications. On the methods side theoretical advances based on the sequential Markovian Coalescent theory should be especially highlighted as they make use of the underlying derived theoretical expectations of the coalescent. The first inference method PSMC, compared the number of segregating sites between two sequences as a proxy of how far the coalescence time is believed to have occurred in the past and assumed a Markovian Poisson process of recombination events along the genome. This realization spawned a whole series of SMC methods integrating more samples at a time of inferring population genetically relevant parameters like recombination rates along the genome or ecological traits of which Sellinger and colleagues provided a study concerning their limitations and convergence properties (Sellinger et al., 2021b). The underlying method for SMC methods is a Hidden Markov Model whose observed variables are the polymorphisms and latent variable the coalescent times.

For some models of interest, the underlying coalescent model is difficult to derive, but simulation methods can be more easily implemented (see above). For this kind of setup, approximate Bayesian computation (ABC) approaches filled a methodological gap and provided a valuable tool for countless studies (Sunnåker et al., 2013; Robinson et al., 2014; Boitard et al., 2016; Johri et al., 2022b). The objective of the ABC strategy can be summarised by its aim of optimizing a posterior distribution of given parameters of interest by iteratively exploring the state-space of possible simulation-enabled realization. Yet, in recent years shortcomings of the time-consuming exploration step and reliance on summary statistics have been attempted to be overcome through DL methods, which contrary to ABC is gradient-based optimization method unrestricted by the type of input data. It therefore evolved to become a powerful inference framework which like ABC has proven to be not only helpful for complex inference tasks but also provides sophisticated generative abilities. A detailed review of the state of the art of DL in population genetics is provided in Chapter 2.

2 Methods for Deep Learning Population Genomics Analyses

The following chapter has been published as:
Kevin Korfmann, Oscar E Gaggiotti, Matteo Fumagalli, Deep Learning in Population Genetics, Genome Biology and Evolution, Volume 15, Issue 2, February 2023, evad008, https://doi.org/10.1093/gbe/evad008

KK contributed an equal amount as his co-authors to the manuscript and revisions. KK additionally designed and implemented the methodological part '2.5 A novel application: detecting balancing selection from temporal data.'

2.1 Abstract

Population genetics is transitioning into a data-driven discipline thanks to the availability of large-scale genomic data and the need to study increasingly complex evolutionary scenarios. With likelihood and Bayesian approaches becoming either intractable or computationally unfeasible, machine learning, and in particular deep learning, algorithms are emerging as popular techniques for population genetic inferences. These approaches rely on algorithms that learn non-linear relationships between the input data and the model parameters being estimated through representation learning from training data sets. Deep learning algorithms currently employed in the field comprise discriminative and generative models with fully connected, convolutional, or recurrent layers. Additionally, a wide range of powerful simulators to generate training data under complex scenarios are now available. The application of deep learning to empirical data sets mostly replicates previous findings of demography reconstruction and signals of natural selection in model organisms. To showcase the feasibility of deep learning to tackle new challenges, we designed a branched architecture to detect signals of recent balancing selection from temporal haplotypic data, which exhibited good predictive performance on simulated data. Investigations on the interpretability of neural networks, their robustness to uncertain training data, and creative representation of population genetic data, will provide further opportunities for technological advancements in the field.

Significance Deep learning, a powerful class of supervised machine learning, is emerging as a promising inferential framework in evolutionary genomics. In this review, we introduce all deep learning algorithms currently used in population genetic studies, highlighting their strengths, limitations, and empirical applications. We provide perspectives on their interpretability and usage in face of data uncertainty, whilst suggesting new directions and guidelines for making the field accessible and inclusive.

2.2 From model-based to data-driven discipline

Population genetics arose in the early 20^{th} century as a conceptual framework aimed at unifying two opposing views of evolution (Provine, 2020). As such, it developed a rich body of theory that became a vast treasure trove of probabilistic models to develop sophisticated statistical methods when molecular data became available. This body of theory has continued to grow in complexity in order to consider more realistic evolutionary and genetic scenarios as well as more efficient computational algorithms. Therefore, the field of population genetics has been dominated by model-based statistical approaches. One could even say that many population geneticists would agree to the proposition of slightly modifying George E.P. Box's aphorism so as to say that in our field, all models are wrong but ***many*** are useful.

The preeminence of model-based statistical inference may explain the fact that our field has lagged behind other life-science disciplines in the adoption of machine learning methods and, in particular, deep learning approaches. Clearly, the black-box nature of deep learning is an important obstacle to applications in the domain of population genetics, which main objective is to uncover the genetic and evolutionary mechanisms responsible for the diversity of life on our planet. Another deterrent is the apparent difference in foci between the fields of statistics and machine learning. Statistics is focused on inference through the creation and fitting of a probabilistic model while machine learning is focused on prediction using general-purpose algorithms that capture patterns present in complex and large data sets (Bzdok et al., 2018). However, population geneticists are interested in both inference and prediction, as clearly illustrated by the general interest in making inferences about demographic history of species on the one hand and detecting signatures of natural selection or assigning individuals to populations on the other. Nevertheless, most genetic clustering methods and so-called genome scans of selection are based on probabilistic models, in some cases mechanistic (e.g. `Bayescan` (Foll and Gaggiotti, 2008) and `STRUCTURE` (Pritchard et al., 2000)) and in others phenomenological (e.g. `LFMM` (Frichot et al., 2013) and `DAPC` (Jombart et al., 2010)).

The focus on model-based statistical inference in population genetics has been challenged by the massive data sets generated by next generation sequencing technologies (Levy and Myers, 2016). This is particularly the case for maximum likelihood and Bayesian methods, which are implemented using expensive computational methods such as Monte Carlo Markov Chain and Expectation-Maximisation. In principle, the computational cost of calculating the

likelihood function of very complex models, can be overcome using Approximate Bayesian Computation (ABC), which relies on the use of summary statistics to capture the information present in raw population genetic data (Bertorelle et al., 2010). In ABC, the posterior distribution of the parameter(s) to be estimated is approximated without the calculation of a likelihood function. Instead, a model fit is obtained by the collection of simulated summary statistics matching the observed values (Beaumont et al., 2002). ABC has been widely and successfully used for population genetic inferences (Lopes and Beaumont, 2010). However, capturing enough information requires large numbers of summary statistics which lead to a "curse of dimensionality" because, as the number of summary statistics increases, the error in the approximation increases (Prangle, 2015). This problem has led to an increasing interest in machine learning approaches (Schrider and Kern, 2018). The underlying rationale here is that analysing genomic data with machine learning methods can uncover signatures of evolutionary and genetic processes in a model agnostic way and in doing so teach us something new about nature (Schrider and Kern, 2018). But a major motivation for the shift is the practical reality that population genetics has been transitioning from a theory-driven discipline into a data-driven field with vast amounts of genomes and metadata at hand in the past few years. For instance, in human population genetics, scientists have access to high-quality whole-genome sequencing data from more than 150,000 individuals from the UK Biobank (Halldorsson et al., 2022), and more than 3,000 individuals distributed worldwide (Byrska-Bishop et al., 2022), or to hundreds of genomic data from ancient samples (`https://reich.hms.harvard.edu/datasets`).

In this review we will focus on a particular subset of supervised machine learning algorithms, namely deep neural networks. Although such methods can be considered as the epitome of a black box, we will argue that new advances in this field are providing the tools we need to uncover the mechanisms underlying the complex patterns present in population genomic data. Moreover, deep learning can be implemented to analyse raw genetic data as well as summary statistics. Additionally, it has been used to carry out statistical inference about the demographic history of populations as well as to carry out selection scans and assign individuals to geographic locations. Applications to demographic history inference embrace the model-based tradition of population genetics in that the training set (see Glossary) is usually generated through simulations of specific evolutionary scenarios. Applications to genome scan methods on the other hand, rely on new techniques for evaluating the importance of features, in this case loci, in predicting an outcome such as a phenotype or an environmental factor that may exert a selective pressure.

We will first provide a definition of supervised machine learning and its applications in population genetics. We will then focus our attention on various deep learning algorithms currently used in the field, with a discussion on efforts to "open the black box" of said algorithms. We will finally discuss ongoing challenges of deep learning applications in population genetics, and highlight future research directions.

2.3 Machine learning in population genetics

Machine learning, a subset of artificial intelligence, refers to a class of operations using data to perform inferential tasks without explicit mathematical models. To do so, machine learning algorithms identify informative patterns which can be then used to predict unknown outcomes. Typically, the performance of machine learning algorithms increases with the amount of available data. Machine learning comprises both supervised and unsupervised algorithms. Unsupervised machine learning aims at finding patterns and clusters within the data, and does not have a notion of prediction. On the other hand, supervised machine learning algorithms automatically tune their internal parameters to maximise the prediction accuracy and, as such, require a known data set (called training set) to learn the relationship between input and output.

To train a supervised machine learning algorithm, the available data sets are typically divided into training, validation, and testing sets, with the latter two sets used to evaluate the performance during and after training. In supervised learning, a labelled data set (which explicitly relates any given input to a specific output) is given to the algorithm. The loss (the distance between the predicted and true value) is calculated, and at the next iteration the internal parameters are updated towards decreasing loss (and increasing accuracy). Training a supervised machine learning algorithm is a fine balance between prediction accuracy over the training set and generalisation performance over the testing set.

Machine learning has a rich history in biological sciences and genomics (reviewed in Yue and Wang, 2018; Zou et al., 2019; Greener et al., 2022). Additionally, supervised machine learning methods have been designed and deployed to perform population genetic tasks such as variant calling (Poplin et al., 2018) and the prediction, characterisation, and localisation of signatures of natural selection (Pavlidis et al., 2010; Lin et al., 2011; Ronen et al., 2013; Pybus et al., 2015; Schrider and Kern, 2016; Sugden et al., 2018; Mughal and DeGiorgio, 2019; Koropoulis et al., 2020). An important difference between the variant calling application (which only uses observed data) and those aimed at detecting selection is that the latter implement an innovation first introduced by (Pavlidis et al., 2010) whereby the ML algorithms are trained using synthetic data sets generated via simulations. These applications, therefore, can be considered as being part of likelihood-free simulation-based approaches (Cranmer et al., 2020a), which are commonly employed in population genetics. Currently, most population genetics applications of ML use this strategy but, as we describe below, some recent applications only use observed data to train the algorithms. These applications, however, require the combination of genotypic data with phenotypic, environmental or geographic coordinate data.

As already stated, in this review we will focus on deep learning, a class of machine learning algorithms based on artificial neural networks comprising nodes in multiple layers connecting features (input) and responses (output) (LeCun et al., 2015). Weights between nodes are

optimised during the training to minimise the distances between predictions and ground truth. After training, an ANN can predict the response given any arbitrary new input data. Unlike approaches that use a predefined set of summary statistics as input, deep learning algorithms can effectively learn which features are sufficient for the prediction (LeCun et al., 2015). This is an important aspect as summary statistics are meaningful but human-constructed features. When dealing with different sources of raw data, the design of such features has been a major part of information engineering. A key finding of deep learning was that such features emerged within a well-trained deep network: they are effectively suggested and discovered by a network during training (Krizhevsky et al., 2012). This finding has been repeated in different domains: features can be automatically discovered, and new suggestions made, by the approaches of deep learning. Nodes in an ANN can be arranged in various numbers and layers, making this method as flexible and "deep" as needed.

Deep learning in population genetics is in its infancy, and most of current applications rely on synthetic data sets for training. Nevertheless deep learning represents a notable progress over commonly used simulation-based techniques for several reasons. First, they have the capacity to handle any feature extracted from a data set as input and are less sensitive to poorly crafted summary statistics than ABC (Csilléry et al., 2010). Second, neural networks are universal approximators of any complex function provided that they include a sufficiently large number of “neurons”, non-linear units (Hornik et al., 1989). Nevertheless, careful monitoring of networks’ training and a posteriori diagnostic analyses are required to ensure that predictions are robust.

Whilst overviews of machine learning applications for population and molecular genetics are provided elsewhere (Schrider and Kern, 2018; Fountain-Jones et al., 2021; Kumar et al., 2022), here we aim at providing an update on the latest advances in deep learning algorithms and how they have been exploited to address questions in population genetics. Additionally, we focus our attention on deep neural networks, in all their supervised forms, rather than including other commonly used algorithms such as support vector machine (Pavlidis et al., 2010), random forests (Schrider and Kern, 2016; Vizzari et al., 2020), gradient forests (Laruson et al., 2022), and hierarchical boosting (Pybus et al., 2015). Finally, we restrict our review on applications in population genomics while acknowledging that similar algorithms herein described are used in other related disciplines like genomics (Yue and Wang, 2018), phylogenetics (Suvorov et al., 2020; Azouri et al., 2021; Blischak et al., 2021), phylogeography (Fonseca et al., 2021; Perez et al., 2022), and epidemiology (Voznica et al., 2021).

Glossary

- Accuracy: proportion of correct predictions made by a model
- Activation function: operation that each neuron performs

- Attribute: name of a variable describing an observation
- Bias term: a term attached to neurons allowing the model to represent patterns that do not pass through the origin
- Backpropagation: gradient descent-based learning algorithm for calculating derivatives through the network starting from the last layer
- Confusion Matrix: table that summarises the prediction performance by providing false and true positive/negative rates
- Embedding: learned low-dimensional continuous vector representation of a concept (e.g., a word, sentence, genotype matrix or graph)
- Epoch: the number of times the algorithm sees the data set
- Feature: input variable used in making predictions
- Hyperparameters: higher level properties of a model controlling the training process (e.g., learning rate, number of epochs) and that need to be tuned, in principle before the ML model is trained
- Instance: a data point or sample in a data set (observation)
- Learning rate: magnitude at which an algorithm updates its parameters
- Loss: (also called cost) measurement of distance between predictions and ground truth; its function is minimised during training
- Normalisation: scaling technique used when input features have different ranges
- Regularisation: an additional penalty to the loss function for better generalisation
- Testing set: portion of the data set that it is not used for training, but rather to evaluate the performance a neural network
- Training set: portion of the data set that it is used to optimise parameters of a neural network
- Tuning or hyperparameter optimisation: process of finding the hyperparameter values that maximise the performance of the model
- Validation set: portion of the data set that it is used for monitoring the training of a neural network

2.4 Deep learning algorithms

We now introduce, describe and discuss four common families of architectures for deep learning algorithms used in population genetics: fully connected neural networks, convolutional neural networks, recurrent neural networks, and generative models. For each type of algorithm, we illustrate their main applications in the field and the novel findings generated by their deployments. Note that these general algorithms have a long history spanning many decades and numerous original contributions which we cannot properly credit in our review because of space. Thus, we refer readers interested in historical developments to previous publications (Schmidhuber, 2014).

2.4.1 Fully connected neural networks

Fully connected neural networks (FCNNs) are suitable for generic prediction problems when there are no special relations among the input data features. They can be viewed as a generalisation of linear regression. In fact, standard regression is nested in the general neural network framework in the sense that a linear regression fits a hyperplane to the data, while a neural network fits a space of hyperplanes in a transformed space (Qin et al., 2022). This becomes clear by comparing the formulation for the simplest multivariate linear regression model with the equation representing the operations taking place in a single node of a hidden layer of a FCNN,

linear regression: $y_i(\mathbf{x}, \mathbf{w}) = b + \sum_{i=1}^{I} w_i x_i$

FCNN: $s(\mathbf{x}, \mathbf{w}) = f(b + \sum_{i=1}^{I} w_i x_i)$,

where b is the bias (not to be confounded with statistical bias), $\mathbf{w} = \{w_i\}$ is a vector of weights, $\mathbf{x} = \{x_i\}$ is a vector of input features (explanatory variables), and f is a nonlinear activation function. In a FCNN with a single hidden layer, there will be a number J of hidden nodes, each carrying out a similar operation using a different vector of weights, all of which can be represented by a matrix $\mathbf{W} = \{w_{ij}\}$, $i = 1, 2..., I$, $j = 1, 2, ...J$. A very simple example of a FCNN with one hidden layer and only two nodes is presented in Figure 2.1.

In the linear regression case a dependent variable is computed by calculating the dot-product of a set of input data points with a set of parameters. This output variable is then used in the context of a maximum-likelihood or least square approach to optimise the set of learnable parameters. FCNNs extend this idea by computing a matrix-product of the weight matrix with the input data points, which is then transformed with a non-linear activation function. The activation function is applied element-wise and the result is called an embedding. Instead of using the maximum likelihood or least square approaches for optimisation, FCNNs are optimised using the multivariate version of the gradient-descent algorithm, which iteratively adapts the parameters across the network layers (back-propagation algorithm (Linnainmaa, 1976; LeCun et al., 1989)) based on a task-specific loss-function and learning rate. A fundamental property of FCNNs is expressed by the Universal Approximation Theorem,

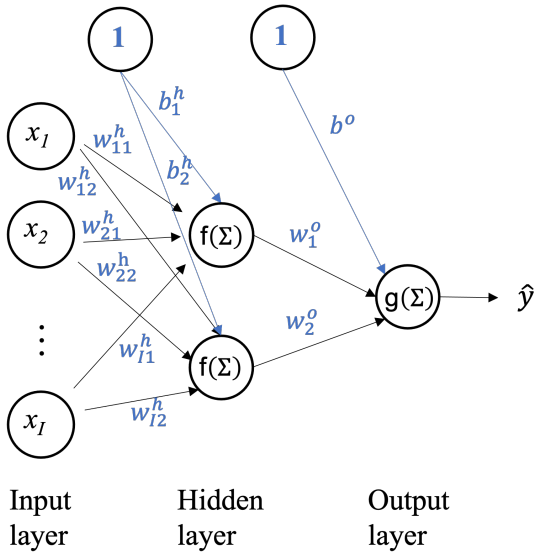


Figure 2.1 A simple FCNN consisting of a single hidden layer with only two nodes. f and g represent different activation functions used respectively in the hidden layer and the output layer and h and o superscripts are used to identify parameters associated with these layers; all other parameters are defined in the text.

which states that a neural network with a single hidden-layer can approximate any continuous function to any desired precision. Precision can be increased by increasing the number of hidden neurons or the number of hidden layers. It is this property that enables the use of neural networks as a viable alternative to common model-based statistical methods.

In an early application of deep learning methods to population genetics, FCNNs are used to simultaneously infer natural selection and population bottlenecks (Sheehan and Song, 2016). This approach was inspired by ABC methods and therefore used summary statistics to extract the information present in the raw data, which was then fed to a fixed-size linear input layer of the network. To discriminate between demographic and natural selection effects, Sheehan and Song trained the FCNN using simulated datasets generated under various models assuming different bottleneck times and selection models (Sheehan and Song, 2016) . The software `evoNet`, which implemented said FCNN, was applied to almost 200 genomes of *Drosophila melanogaster* from Africa to jointly infer the demography history and loci under selection. One interesting analysis in the study is the evaluation of the most informative summary statistics, either by permutation or perturbation. Notably, summary statistics derived from the site frequency spectrum, linkage disequilibrium (LD), number and location of single nucleotide polymorphisms (SNPs), and identity-by-state tracts are among the most important features for the inference of population size changes and type of selection.

Another example of a FCNN application in population genetics that uses simulated data to train the algorithm is provided by the work of Burger and colleagues on the estimation on mutation rates (Burger et al., 2022b). They show that a simple neural network is able to recapitulate estimators of mutation rate for intermediate recombination rates. As a novel

methodological advance, their implementation features an adaptive reweighting of the loss function based on model-based estimators of the mutation rate. By doing so, with sufficient and appropriate training set, only a single hidden layer is required to achieve the same performance of model-based estimators. The method was able to recover variation in mutation rates from synthetic human population genetic data under a realistic recombination map.

There are also recent population genetics applications of FCNNs that implement the standard approach of training algorithms using observed instead of simulated data. A good example is `Locater`, which assigns individual genotypes to their geographic origin (Battey et al., 2020). Interestingly, this method implements a regression approach that is capable of assigning correlated genetic samples to similar geographic space. Uncertainty in the estimates due to drift is taken into account by running predictions in windows across the genome. Simulations indicate that `Locater` has an accuracy comparable to that of other state-of-the-art competing algorithms but with shorter run-times. Its application to an empirical population genetic data set of *Anopheles* mosquitoes, *Plasmodium falciparum*, and human populations, provides results that are in general concordant with current knowledge.

Another example that only uses observed data to train the FCNN is `DeepGenomeScan`(Qin et al., 2022). However, this method's objective departs from the prevalent use of neural networks, *i.e.* prediction and pattern recognition. Its aim is to develop a statistical framework to carry out genome scans or GWAS, much in the same way that PCA and redundancy analysis have been used to develop equivalent approaches ((Capblancq et al., 2018; Luu et al., 2017). Specifically, `DeepGenomeScan` implements a FCNNs that uses genotypes to predict individuals' traits (e.g., geographic coordinates or phenotype), and constructs a feature importance measure based on the weights of the trained network. Furthermore, *p*-values for variable importance are obtained through bootstrapping of the input. As opposed to other methods that can only detect linear associations, `DeepGenomeScan` is able to detect non-linear ones thanks to the non-linear approximation property of FCNNs. Its application to a genomic data set of human samples of European ancestry identified novel targets of natural selection which showed significant geographical variation.

Finally, we note that FCNNs have also been used in the context of ABC frameworks. Early studies used neural networks to construct the posterior distribution of parameters from the collection of accepted values (Blum and François, 2010), as implemented in the `abc` package (Csilléry et al., 2012). More recently, Mondal and colleagues coupled an ABC framework, using the site frequency spectrum (SFS) as summary statistic, with a four-layer FCNN to infer the demographic history of human Eurasian populations (Mondal et al., 2019). Their implementation includes an *ad hoc* noise injection algorithm to partly take into the account any bias associated with a simulated training set. A similar study by Villanea and Schraiber used the joint SFS between Europeans and Neanderthal genomes to fit a demographic model using a 3-layer FCNN (Villanea and Schraiber, 2019). Both studies inferred multiple gene flow events between archaic and anatomically modern humans.

Summary statistics and genotype matrices are not the only way in which population genomic data can be described and used as input to deep learning algorithms. It is also possible to represent samples of sequences as images and, in the next section, we discuss an architecture that is being increasingly applied to such data.

2.4.2 Convolutional neural networks

Convolutional neural networks (CNNs) are specifically designed to analyse data that has a grid-like structure, such as images (LeCun et al., 2004; Krizhevsky et al., 2012). Whilst in theory FCNNs could be used to make predictions from images, the number of features (i.e., pixels) they contain would require networks with a very large number of parameters, which would render them very slow and computationally expensive. Similarly to FCNNs, CNNs are comprised of a set of learnable parameters (LeCun et al., 1989; Lecun and Bengio, 1995). However, as opposed to FCNNs, in which hidden layers are all of the same type (layers of neurons carrying out similar operations), CNNs architecture consists of consecutive sets of convolutional and pooling layers, followed by a fully connected set of layers (similar to a FCNN, Figure 2.2). The first convolutional layer takes the input image and carries out a convolution using a kernel (also known as filter; a matrix of learnable parameters) to generate a feature map that is then fed to the pooling layer. This layer uses a filter to reduce the size of the feature map and to help dissociate a particular feature from its position in the input image. This first set of operations will capture coarse grained features; adding additional convolutional and pooling layers helps capture more fine-grained features (O'Shea and Nash, 2015). The final step of the convolutional layers (flatten step) converts the feature map into a vector that is fed to the fully connected layers that will carry out the image classification step. The number of kernels, their dimensions, and initialisation are all hyperparameters of the model.

CNNs can be regarded as a regularised version of FCNNs with a focus on localised spatial signatures. In fact, a fundamental property of CNNs is the space-invariance of the learned features in the data set, which means that they can identify a pattern regardless of its spatial location in the image. Note, however, that identification of feature realisations like rotations or scaling requires either appropriate samples or perturbations of the input (Goodfellow et al., 2016).

First applications of CNNs in population genetics relied on "image" data sets in the form of stacked summary statistics. The method implemented in software `diploS/HIC` aimed at classifying genomic windows into neutral regions or under soft or hard selective sweeps from unphased genotypes (Kern and Schrider, 2018). It did so by applying convolutional operations on a feature vector of normalised summary statistics calculated in windows surrounding the target location. The architecture consisted of three branches of two dimensional convolutional layers with different filter sizes, followed by max pooling, flattening and two fully connected

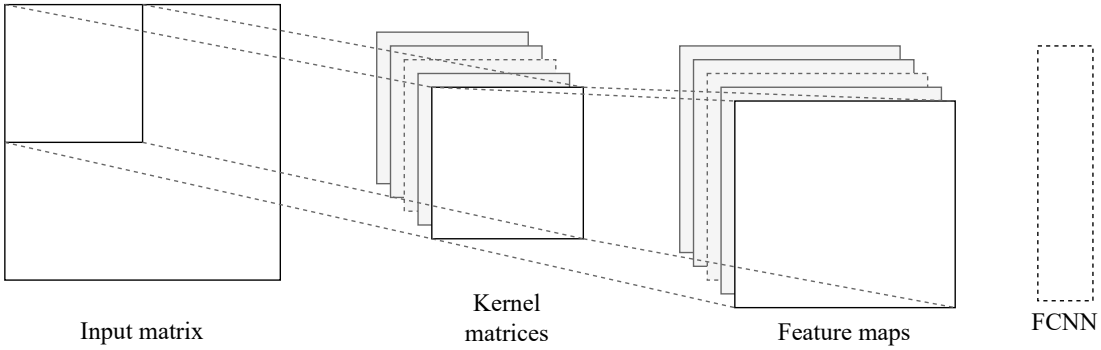


Figure 2.2 A simple CNN illustration consisting of the input matrix (*i.e.* genotype matrix), a user-specified number of kernels (or filters) and the resulting feature maps, followed by a FCNN.

layers. Extensive simulations of tested scenarios were produced to train the CNN. The authors showed that CNNs outperformed competing ML algorithms previously used for this classification task (Schrider and Kern, 2016), possibly because CNNs retain the spatial relationships of summary statistics. Notably, with moderate sample size, `diploS/HIC` appears to be robust to model misspecification as it retains accuracy when predictions for a population growth demography were obtained from CNNs trained on constant size population simulations. As an application of `diploS/HIC`, the authors replicated previous findings of selective sweep in the *Anopheles gambiae* genome. A later extension of this method led to `partialS/HIC` which uses CNNs on a larger feature vector of summary statistics for a finer classification of selective events, including partial sweeps and linked selection (Xue et al., 2020). Finally, an additional application of CNNs based on summary statistics to test against different modes of selective sweeps has been recently proposed (Caldas et al., 2022). This study uses varying window sizes to accommodate the calculation of summary statistics at different genomic extents within the target loci. They also introduced a hybrid simulation strategy to pair the flexibility of forward-in-time simulations with the efficiency of coalescent ones.

An approach that fully exploits the potential of CNNs is to replace summary statistics as input with full information on sequence alignments, with convolutional layers automatically extracting informative features. Input data can consist of either genotype or haplotype sequences. In the simplest form, input data is a binary matrix, with rows and columns corresponding to individuals and alleles at each SNP, respectively. Under this representation, and in opposition to the structured nature of "classic" images, the ordering of individuals (*i.e.* random samples from a population) in an unstructured population is arbitrary and carries no information (Chan et al., 2018); *i.e.* genetic data are exchangeable. However, standard CNNs rely on spatial information and, therefore, the ordering of the data can affect its accuracy. To avoid this problem, individuals need to be sorted in a "biologically meaningful" way. For example, Flagel and collaborators sort chromosomes by genetic

similarity (Flagel et al., 2018). Additionally, they represent the information on genomic positions of SNPs as a separate branch in the architecture. Interestingly, the inclusion of monomorphic sites in windows of fixed length seems to yield good accuracy for predicting natural selection, as shown in a separate study (Nguembang Fadja et al., 2021). Notably, several applications of the proposed method are illustrated, with CNN achieving equal if not better better performance than state-of-the-art methods to detect gene flow and selective sweeps, estimate recombination rates, and infer demographic parameters (Flagel et al., 2018). Therefore, these findings demonstrated the capability of CNNs to infer population genetic parameters, even in cases where a theoretical framework is not available.

To address the exchangeability issue, Chan and coworkers proposed an exchangeable neural network (Chan et al., 2018). This architecture consists of convolutional layers with 1-dimension kernels with a subsequent permutation-invariant function to allow for the network to be insensitive to the order of individuals. Although they employed the mean operation as permutation-invariant function, other functions are possible, including a fully connected layer. Another important contribution of this study is the adoption of a "simulation-on-the-fly" approach: training data is continuously generated by simulations to avoid the network to see the same data twice and therefore to reduce overfitting. This is a valuable consideration since, when reliable simulators are available (as in the case of population genetics), we have access to theoretically infinite training data, the latter being constrained by computing time only. The implemented software `defiNETti` was applied to illustrate the accuracy of exchangeable neural networks to predict recombination hotspots in human data.

Further solutions to tackle the issue of exchangeable genetic data have been explored by Torada *et al.* in the software `ImaGene` (Torada et al., 2019). Specifically, the authors showed how ordering haplotypes and SNPs by frequency leads to accurate predictions of positive selection. Whilst sorting SNPs implied a loss of information on LD patterns, this approach makes training faster with minimal decay in accuracy, as the number of learnable parameters is drastically reduced as the final fully connected layer is not required. However, double-sorting makes the method less appropriate for a general-purpose methodology. Additionally, by training and testing `ImaGene` with simulations conditioned on different demographic models, the authors quantified the drop in accuracy when CNNs are affected by model misspecification during training. Finally, a multiclass classification approach was proposed as an alternative method to approximate the posterior distribution of the selection coefficient, a continuous parameter typically hard to estimate.

In another landmark study, Sanchez and colleagues provide a comprehensive framework for building deep neural networks taking into account several nuances of the input data, such as the variable number of SNPs, their correlation, and the exchangeability of individuals (Sanchez et al., 2021a). These challenges were tackled by proposing an architecture, called `SPIDNA` (Sequence Position Informed Deep Neural Architecture), which consisted of stacks of multiple blocks of convolutional, pooling, and fully connected layers. In addition to deploy

their method to reconstruct changes in effective population size of cattle breed populations, the authors compared the accuracy of several deep neural networks against ABC, including hybrid approaches. Notably, results suggest that integrating deep learning with ABC marginally improves performance, and possibly explainability. Further investigations from the same authors demonstrated a more prominent increased performance using deep neural networks (Sanchez, 2022). These studies depart from previous attempts to adapt existing architectures, and instead they suggest to build novel architectures tailored to the specifics of population genetic data.

In a later study, Gower and colleagues (Gower et al., 2021) aimed to identify signatures of adaptive archaic introgression in the human genome without relying on statistics that capture the frequency of putatively introgressed haplotypes. The authors developed a deep learning method based on CNNs, `genomatnn`, to jointly infer archaic admixture and positive selection. `genomatnn` is trained from a matrix consisting of concatenated genotype alignments encompassing donor (archaic humans) and recipient (modern humans) populations. Matrix entries represent counts of minor alleles in an individual haplotype within a given genomic window. Thus, this approach is applicable to low-quality sequencing data where genotype calling can be bypassed by the statistical estimation of allele frequencies (Kim et al., 2011). Additionally, the authors proposed a framework to visually inspect the input features that are more informative for the prediction by means of saliency maps (Simonyan et al., 2013). Intriguingly, the latter indicated that the network focus most of its attention on Neanderthal and European haplotypes when exposed with data from an adaptive introgression, in line with the expected pairing of donor and recipient populations.

`DeepSweep` is another application of CNNs to detect selective sweeps from "haplotypic" images, as defined by the authors (Deelder et al., 2021). This method selects the longest common haplotype among neighbouring SNPs, and sort all remaining haplotypes based on their distance to it. This sorted alignment of haplotype differences is then fed into a series of convolutional layers. The aim of of the original study was to detect signatures of positive selection in malaria parasites, namely *Plasmodium falciparum* and *Plasmodium vivax*. Interestingly, the algorithm was then trained using real data from regions covering SNPs previously associated with drug resistance, and the validation was performed using a leave-one-out approach. Possibly as a result of both the data processing and training strategies, when deployed on whole-genome data, `DeepSweep` predicted selection targets to be known drug-resistance genes and largely overlapping with predictions using haplotype-based summary statistics. One advantage of this training strategy is that it enables an assessment of which data points are informative during training.

A comparison between the performance of FCNN and CNN to detect natural selection, specifically balancing selection, is presented by Isildak and colleagues (Isildak et al., 2021a) in the software `BaSe`. While both architectures exhibit high classification accuracy to distinguish between neutrality and selection, CNN outperformed FCNN to predict the type of balancing

selection, a task that proved too challenging when relying solely on summary statistics as input. Authors used forward-in-time simulations and conditioned the target variants to a predefined range of final allele frequency. To counterbalance the increased computational time associated with this simulation scheme, a data augmentation to artificially enlarge the training data was adopted.

In recent years, the generation of sequencing data from ancient or historical samples, as well as from capture-recapture and evolve-and-resequence experiments, has allowed for a direct observation of how genetic diversity and allele frequencies change under natural or controlled conditions over time. To detect positive selection with time-series data, Whitehouse and Schrider proposed to stack either allele frequency or haplotype data over sampling times to be fed as input to 1-dimensional CNNs (Whitehouse and Schrider, 2022). Their method was implemented in the software `Timesweeper`, and evaluated under various sampling conditions. Results show overall good accuracy levels for predicting selection, localising the target variant, and distinguishing between selection from *de novo* mutation and from standing variation. Interestingly, using haplotype instead of allele frequency data yields a lower performance, possibly due to the difficulty in properly sorting the input data in a biologically-meaningful way. `Timesweeper` was deployed to time-series pooled-sequencing data from *Drosophila simulans*, and it was able to replicate previously detected sweep signatures with better resolution.

CNNs have quickly become the main deep learning algorithm in population genetic studies thanks to their ability to automatically extract important features from raw genotype data, and their flexibility in accommodating different models to be tested. As a result, novel applications of such algorithms in population genetics are frequently proposed and introduced (Smith et al., 2022). In machine learning, natural language processing (NLP) represents a branch of algorithms that aims at "understanding" words in a text, meaning that they can, for instance, perform speech recognition, text generation, or sentiment analysis (i.e., associating an output label to each word or sentence). As DNA sequences are easily representable as a series of letters or motifs, in the next section, we will introduce NLP applications that are emerging in population genetics.

2.4.3 Recurrent neural networks

Recurrent neural networks (RNNs) are algorithms derived from FCNNs but designed specifically for sequential data as they introduce a mechanism that influence current predictions based on previous outcomes (Elman, 1990; Minsky, 1967; Rumelhart and McClelland, 1987). In fact, RNNs are comprised of connected nodes that form a cycle, with the output of some nodes feeding back to other (or same) nodes. Therefore, simple RNNs can be considered as for-loops iterating along the sequential data, where at each position the current input and the previous output are combined to form the next output (or hidden state). Multiple RNN

layers can be stacked on top of each other to increase the capacity of the network and extract more features from the data. One of the limitations of RNNs is the limited capacity to learn long-range dependencies. Architectures such as Long Short-Term Memory (LSTM) and Gated-Recurrent Units (GRUs) networks circumvent this problem by adding the concept of cell state which is propagated along the sequence in the case of LSTMs, and GRUs enabling the filtering of passing information of long-range information through a Gating mechanism alone (Cho et al., 2014) whilst maintaining similar performance to LSTMs (Hochreiter and Schmidhuber, 1997).

Recurrent layers have been used by Adrion and colleagues (Adrion et al., 2020b) to estimate recombination maps for *D. melanogastor*. The proposed software `ReLERNN` provides a comprehensive modular workflow on how to generalise the method for different model species of interest, including instructions for phased, unphased and pooled-sequencing data. However, caution should be made when estimating recombination rates from genotype alignments using machine learning under certain conditions of low variability (Johnson and Wilke, 2022). Hejase and coworkers proposed a method to detect natural selection by extracting features from estimated genealogical trees (Hejase et al., 2021). They used counts of remaining lineages along a discrete log-transformation of the time dimension. The sequential nature of the trees along the sequence was used to set up an LSTM, which recognises the lack of remaining lineages, *i.e.* zeros in the distant past or upper part of the feature matrix. This approach, implemented in the software `SIA`, gains the possibility to obtain an easily interpretable model at the cost of using an ancestral recombination graph (ARG)-inference method such as `Relate` (Speidel et al., 2019).

Inspired by the sequential nature of the Sequential Markov Chain (SMC) methodology, Khomutov *et al.* proposed a RNN method to estimate times to the most recent common ancestor from simulated data (Khomutov et al., 2021). Interestingly, this method achieved good results after coupling it with a CNN. Their approach is setup as a coalescent event classification strategy, thus creating a probability distribution of the TMRCA coalescent time at any given sequence position. Finally, neural net compression algorithms have been developed (Wang et al., 2018; Silva et al., 2020) making use of recurrent layers for the emphasis of long-range inter-dependencies and convolution layers. These approaches appear useful as the cost of sequencing dramatically decreases and becomes increasingly negligible compared to storage costs.

RNNs, in all their forms, have becoming increasingly popular in population genetics thanks to their ability to incorporate sequential data. Whilst training recurrent layers tend to be more challenging, coupling them with convolutional layers appear to be a suitable solution to overcome such issue whilst incorporating novel information. In the next section, we will explore how CNNs can be embedded in a more general family of machine learning algorithms called generative models.

2.4.4 Generative models

Generative models aim at capturing, and therefore approximating, the probability distribution between data and labels. By their nature, generative models are able to "generate" novel data points according to the captured probability distribution. Fitting a Gaussian mixture model and sampling from the distribution can be interpreted as a generative process, although it is insufficient to capture complex phenomena in high-dimensional spaces. In fact, even if sampling procedures can yield impressive results, *i.e.* for ARG inference (Mahmoudi et al., 2022b), they often remain model-based, and are fundamentally limited by their run-time. For these reasons, deep generative models have become a subject of increased attention, especially for their capability of generating new samples even if the true underlying distribution is unknown. The following section focuses on three among the most popular non-model-based and high-parameter generative methods that have been explored in population genetics: autoencoders (Rumelhart and McClelland, 1987), variational autoencoders (Kingma and Welling, 2014), and generative adversarial networks (Goodfellow et al., 2014).

Autoencoders and variational autoencoders

Similar to Principal Component Analysis (PCA), autoencoders aim to solve a compression problem by step-wise reducing the input parameters into a smaller set of hidden parameters, analogues of the principal components. The number of hidden parameters, known as the latent space, is dependent on the network architecture. In a simple form, compression is achieved by a FCNN, called the encoder, with a decreasing number of learnable parameters in each layer. A second expanding network, called the decoder, rebuilds the original data from said latent space by minimising a suitable loss function. An important part of the autoencoders is the regularisation step, usually introduced as part of the loss function, which is necessary for learning a meaningful latent space by avoiding memorisation.

Variational autoencoders (VAEs) differ from autoencoders as they introduce a generative operation by compressing the data into a latent space distribution, instead of a point representation. Furthermore, the latent space directly offers the possibility to probe the network for any kind of structure as input data, which the encoder has been forced to compress, by plotting the low-dimensional latent variables against each other. Thanks to the non-linearity of neural networks, VAEs outperform classic methods, *i.e.* PCA, for visual data representation (Battey et al., 2021).

VAEs have been implemented by Battey and colleagues in the software `popvae` (Battey et al., 2021). By applying it to genomic data sets, they recovered geographic similarities among human populations, and tested for robustness in the presence of genomic inversions in *Anopheles* mosquitoes. Additionally, low values of population genetic differentiation, as measured by F_{ST} (Holsinger and Weir, 2009), are more likely to be detected by VAEs. Lastly,

whilst the generative property of VAEs has difficulties in detecting more complex relations, like long-range LD signatures, it can produce data with similar SFS patterns.

Other authors proposed a different VAE, named `HaploNet` (Meisner and Albrechtsen) to infer population structure and ancestry proportions. `HaploNet` was shown to be able to infer parameters from very large genomic data sets, such as the UK Biobank and the 1000 Genomes Project. Likewise, others have proposed a multi-headed autoencoder, called `Neural ADMIXTURE` (Mantes et al., 2021), which was evaluated on the Simons Genome Diversity Project and the Human Genome Diversity Project, achieving similar results. Finally, Lopez-Cortes and coworkers combined an autoencoder with common clustering methods, such as hierarchical clustering and K-Means (López-Cortés et al., 2020). They sought to assign maize lines into subpopulations, and achieved marginally better results than by using a Bayesian clustering method.

Generative adversarial networks

Generative adversarial networks (GANs) provide a framework capable of estimating high-dimensional probability distributions by solving a min-max optimisation problem between two opposing networks (Goodfellow et al., 2014). The aim of this architecture is thus to approximate the underlying data generation process (*i.e.* evolutionary process) of a study object of interest (*i.e.* genotype matrix). The model is capable then to *sample* new instances of the study object.

The first part of the architecture, called the generator network, only has access to the random distribution as a prior for constructing the target object, while the second network, called the discriminator has access to a real object (*i.e.* genotype matrix) and the generated object. The loss function from GANs illustrates the objectives of both networks: $L = E_x[log(D(x))] + E_z[log(1 - D(G(z)))]$. The first part $E_x[log(D(x))]$ representing the expected value of real samples x to be classified correctly by the discriminator ($D(x)$) and the second part $E_z[log(1 - D(G(z)))]$ stands for the expected value of generated data ($G(z)$, z being the latent initialisation) to be classified as fake by the discriminator ($1 - D(G(z))$). Thus, the discriminator aims to maximise the loss function, while the generator tries to minimise it. The parameters of both networks are updated alternately. Optimisation can be particularly challenging as neither network should be under-performing nor outperforming the other network too quickly. For instance, when both networks are not training *synchronously*, many values of the random initialisation distribution can collapse into few target estimations, leading to decreased diversity of generated samples of the generator, a phenomenon known as the 'Helvetica scenario' or 'mode collapse' (Arjovsky and Bottou, 2017). The discriminator would become trapped in a rejection space, and eventually end in a local minimum (Che et al., 2016). Another issue focuses around the fleeting convergence property during training, meaning the generator network becomes too good at misleading the discriminator, in which

case the discriminator could only guess the correct class, resulting in poor gradients for both networks overtime.

In the first application of GANs in population genetics, Wang and colleagues (Wang et al., 2021) integrated the coalescent simulator `msprime` (Baumdicker et al., 2021) with a parameter sampling algorithm (called simulated annealing) as the generator, with a CNN as discriminator. The objective was to infer optimal parameters of the simulations that generated realistic data sets. In this study, authors sought to estimate demographic parameters and recombination rate by evaluating both real and simulated data using summary statistics in a likelihood-free approach, similarly to ABC. In fact, authors compared their method, implemented in the software `pg-gan`, to an SFS-based ABC and achieved a similar performance. However, it is still unclear whether ABC or GANs yield a better performance in terms of the number and accuracy of parameters (here demographic changes), number of necessary simulations, and run-time for population genetic applications.

Beyond inferring parameters, the generative property of GANs has been explored in the form of other generative models such as Restricted-Bolthman-Machines (RBMs, (Teh and Hinton, 2000; Smolensky, 1986)). Yelmen *et al.* used RBMs to recreate a population structure data set as genotype matrices extrected from 1000 Genomes Project data set (Yelmen et al., 2021a). The authors successfully demonstrated the ability of RBMs to reconstruct multimodal distributions by reporting various distance measures (such as Wasserstein distance) and by visual inspection via dimensionality reduction. However, this initial attempt is not capable of recovering rare variant patterns, but advanced architectures designed to deal with mode collapse may solve this issue (Ghosh et al., 2017). Despite current limitations, GANs appear to be a promising deep learning framework to infer complex population genetic parameters in face of an uncertain or unknown demographic model (Booker et al., 2022).

2.5 Available resources

2.5.1 Simulators

The application of deep learning methods has been empowered by decades of research into mathematical models of evolution and development of simulators built to recreate the the hidden stochasticity of unseen evolutionary processes. In the context of deep learning, most of the applications in population genetics rely on training algorithms via synthetic data generated by such simulators. Broadly speaking, simulators can be categorised as forward-in-time and backward-in-time approaches. The latter category refers to coalescent simulators which, due to their rigorous underlying models, are extremely efficient as they only keep track of sampled genomes.. Forward-in-time simulation tend to be more intuitive in their development, and are often used for complex selective processes which cannot be described by coalescent models..

The following section is dedicated to name a few popular simulation tools, which can be used to generate data set to train neural networks.

`SLiM` (Messer, 2013), provides a whole programming language Eidos (Haller, 2016) designed to build forward simulation code for a vast range of evolutionary processes. Therefore, it has been used to train deep learning algorithms that aimed at inferring complex models. Interestingly, current developments on spatial simulators, such as `slendr` (Petr et al., 2022), leverage `SLiM`'s capabilities to generate synthetic genetic data variable in time and space. Likewise `SLiM`'s extensions to simulate bacterial populations (Cury et al., 2022) allow for studies of non-model organisms to generate synthetic data sets which could be used in a deep learning framework. Another forward-in-time simulator that has been used in deep learning is `SFS_code` (Hernandez and Uricchio, 2015).

Among coalescent simulators, `msprime` (Baumdicker et al., 2021) is the preferred choice among practitioners due to its carefully designed code base, efficient tree sequence data structure (Kelleher et al., 2018), fast run-time, available choice of coalescent models (Adrion et al., 2020a), easy programmatic access as well as active maintenance. It should be noted that tree sequences are not inherently limited to coalescent simulations, but have also been integrated into forward-in-time simulators such as `SLiM` (Haller et al., 2019b), `fwpyy` (Thornton, 2014) or `sleepy` (Korfmann et al., 2022a). Lastly, `ms` (Hudson, 2002), `msms` (Ewing and Hermisson, 2010), `fastsimcoal2` (Excoffier et al., 2021), and `discoal` (Kern and Schrider, 2016) are coalescent tools that have been applied to train deep neutral networks for population genetic inferences.

2.5.2 Software

Most of the studies herein mentioned provide their implementations, often as user-friendly software, of deep learning algorithms for population genetic analyses. In Table 2.1, we summarise these implementations by the programming language and required (or preferred) simulator (if any) used, and by the input data required (Table 2.1). We further categorise implementations based on their underlying type of neural network. Whilst general-purpose software for simulation-based inferences are available (Tejero-Cantero et al., 2020), here we focus only on implementations specific to population genetic analysis.

From this collection, we note that recent implementations often rely on `python` packages such as `keras` and `tensorflow` which allow for easy building of layers, efficient optimisation of networks, and intuitive monitoring of training performance. Implementations based on `pytorch` (another popular `python` package) allow for more flexibility in constructing complex architectures and investigating internal nodes. These `python` packages are supported by a strong and active community of developers and users, which ensures constant debugging and development.

We also note that forward-in-time simulators are becoming increasingly popular for training deep neural networks despite their significant computational cost, although the adoption of tree-sequence data and 'simulation-on-the-fly' techniques can reduce such burden. Despite the plethora of implementations, each one appears to be suitable to perform specific tasks. At the moment of writing, only `DNADNA` (Sanchez et al., 2022a) is the sole software providing a general framework to both generate simulations and build and training arbitrary networks.

Reference	Language/library	Simulator	Input
evoNet[1] (Sheehan and Song, 2016)	Java	msms	summary statistics
DeepGenomeScan[2] (Qin et al., 2022)	R/keras	not trained by simulations	genotype, phenotype and sampling locations
Locater[3] (Battey et al., 2020)	python/keras	not trained by simulations	phenotype and sampling locations
ML_in_pop_gen[4] (Burger et al., 2022b)	python/keras	msprime	SFS
ABC_DL[5] (Mondal et al., 2019)	Java/Encog and R/abc	fastSimcoal2	SFS
diploS/HIC[6] (Kern and Schrider, 2018)	python/keras and scikit-learn	discoal	summary statistics
partialS/HIC[7] (Xue et al., 2020)	python/keras and scikit-learn	discoal	summary statistics
drosophila-sweeps|[8] (Caldas et al., 2022)	python/pytorch	SLiM/msprime	summary statistics
defiNETti[9] (Chan et al., 2018)	python/tensorflow	msprime	genotype data
pop_gen_cnn[10] (Flagel et al., 2018)	python/keras	ms discoal	genotype data
ImaGene[11] (Torada et al., 2019)	python/keras	msms	haplotype data
dlpopsize[12] (Sanchez et al., 2021a)	python/pytorch	msprime	haplotype data
BaSe[13] (Isildak et al., 2021a)	python/keras	SLiM	haplotype data
genomatnn[14] (Gower et al., 2021)	python/tensorflow	SLiM	genotype data
DeepSweep[15] (Deelder et al., 2021)	python/keras	SFS_code	haplotype data
Timesweeper|[16] (Whitehouse and Schrider, 2022)	python/keras	SLiM	haplotype or allele frequency time-series data
disperseNN[17] (Smith et al., 2022)	python/keras	SLiM or msprime	genotype or tree sequence data and sampling locations
ReLERNN[18] (Adrion et al., 2020b)	python/tensorflow	msprime	genotype data
SIA[19] (Hejase et al., 2021)	python/keras	SLiM or discoal	local trees
DNADNA20 (Sanchez et al., 2022a)	python/pytorch	msprime	haplotype data

Table2.1 List of available software and implementations of deep learning methods (not considering generative models) for population genetic inferences. Software is gratefully supplied at their respective repositories: [1]:https://sourceforge.net/projects/evonet, [2]:https://xinghuq.github.io/DeepGenomeScan, [3]:https://github.com/kr-colab/locator, [4]:https://github.com/fbaumdicker/ML_in_pop_gen, [5]:https://github.com/oscarlao/ABC_DL, [6]:https://github.com/kr-colab/diploSHIC, [7]:https://github.com/xanderxue/partialSHIC, [8]:https://github.com/ianvcaldas/drosophila-sweeps, [9]:https://github.com/popgenmethods/defiNETti, [10]:https://github.com/flag0010/pop_gen_cnn, [11]:https://github.com/mfumagalli/ImaGene, [12]:https://gitlab.inria.fr/ml_genetics/public/dlpopsize, [13]:https://github.com/ulasisik/balancing-selection, [14]:https://github.com/grahamgower/genomatnn, [15]:https://github.com/WDee/Deepsweep, [16]:https://github.com/SchriderLab/timeSeriesSweeps, [17]:https://github.com/kr-colab/disperseNN, [18]:https://github.com/kr-colab/ReLERNN, [19]:https://github.com/CshlSiepelLab/arg-selection, [20]:https://mlgenetics.gitlab.io/dnadna

2.6 A novel application: detecting short-term balancing selection from temporal data

We now wish to illustrate the feasibility and accessibility of deep learning algorithms to perform population genetics predictive tasks which are typically unachievable using classic approaches. To this aim, by using some of the architectures and techniques described above, we seek to develop a novel algorithm to detect signals of recent balancing selection from temporal genomic data.

Balancing selection is a process that generates and maintains genetic diversity within populations (Charlesworth, 2006) whose signals are typically detected by investigating patterns of genetic diversity, allele frequency, and shared polymorphisms between species and populations (Key et al., 2014). Long-term balancing selection has been proved to be a major determinant of important phenotypes, including in humans (Soni et al., 2022). However, recent and fleeting balancing selection leaves cryptic genomic traces which are hard to detect and greatly confounded by neutral evolutionary processes (Sellis et al., 2011). Therefore, currently employed methods are either unsuitable or underpowered to detect short-term balancing selection (Fijarczyk and Babik, 2015).

Information from temporal genetic variation, either from evolve-resequence or ancient DNA (aDNA) experiments, is particularly suitable to identify when and at to what extent natural selection acted (Dehasque et al., 2020). Previous attempts to use deep learning to infer balancing selection from contemporary genomes (Isildak et al., 2021a) and positive selection from temporal data (Whitehouse and Schrider, 2022) suggest that training an algorithm that uses the haplotype information from both contemporary and aDNA data has high potential to characterise signals of recent adaptation (and thus recent balancing selection).

To illustrate the ability of deep learning to detect signals of recent balancing selection, we simulated a scenario inspired by available data in human population genetics. We simulated 2000 50 kbp loci under either neutrality or overdominance (*i.e.* heterozygote advantage, a form of balancing selection) at the centre of the locus, conditioned to a demographic model of European populations (Jouganous et al., 2017). We performed forward-in-time simulations using `SLiM` (Haller and Messer, 2019), similarly to a previous study (Isildak et al., 2021a). We imposed selection on a de novo mutation starting 10k years ago, with selection coefficients of 0.25% and 0.5%. We sampled 40 present-day haplotypes, and 10 ancient haplotypes at four different time points (8k, 4k, 2k, 1k years ago, mirroring a plausible human aDNA data collection).

We trained a deep neural network to distinguish between neutrality and selection. Using `pytorch`, we built a network comprising two branches. One branch receives present-day haplotypes and performs a series of convolutional and pooling layers with permutation-invariant functions. The other branch processes stacked ancient haplotypes at different sampling points, and both branches performing residual convolutions. The two branches

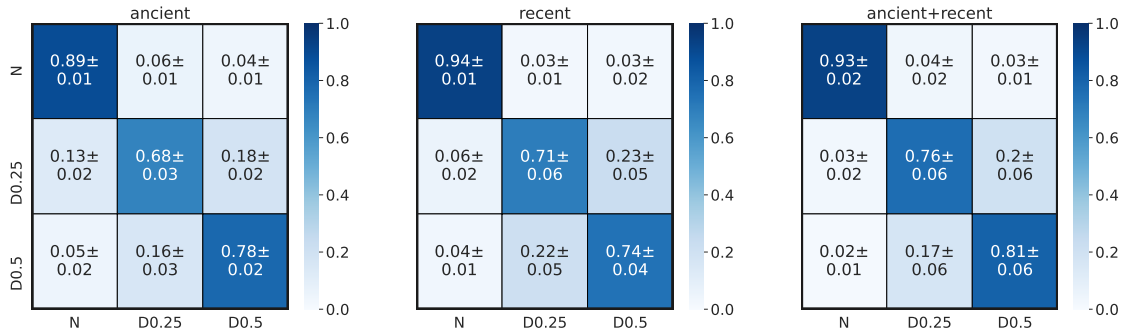


Figure 2.3 Confusion matrices to classify neutrality (N), weak (D0.25), or moderate (D0.5) overdominance with a deep learning algorithm using only ancient, present-day, or both types of samples. True and predicted classes are on the x-axis and y-axis, respectively.

are merged with a dense fully layer that performs a ternary classification. We used 64 filters with 3x3 kernel size and 1x1 padding size after sorting haplotypes by frequency (Torada et al., 2019). We performed 10 separate training operations to obtain confidence intervals in accuracy values. We report results in the form of confusion matrices, a typical representation to summarise the predictive performance at testing. To further showcase the accessibility of deep learning, we made the full implementation and scripts are available at `https://github.com/kevinkorfmann/temporal-balancing-selection`.

Results show that, despite the small training set used, the network has high accuracy to infer recent balancing selection under this tested scenario (Figure 2.3). Notably, we observe a significant decrease in accuracy for distinguishing between weak and moderate selection when silencing the time-series branch, suggesting an important contribution of ancient samples in the prediction. In this illustrative example, we do not attempt to take into account the uncertainty given by degraded and low-coverage aDNA data and population structure across time points, among other confounding factors. Nevertheless, these results demonstrate that building and training novel deep learning algorithms is accessible and generates powerful predictions to address current questions in population genetics.

2.7 Interpretable machine learning

As already mentioned in the introduction, population genetics and evolution in general are aimed at uncovering the mechanisms responsible for the diversity of life in our planet. Thus, the black-box nature of deep learning methods represent an important obstacle for their application in these research fields. However, very recent advances in "interpretable machine learning" algorithms (Linardatos et al., 2021) are providing the tools needed to overcome this hurdle.

But what exactly do we mean by interpretability? There is no general consensus on what the word 'interpretability" means (Fan et al., 2020; Doshi-Velez and Kim, 2017) and

discussions of this concept in the artificial intelligence literature tend to be rather abstract and sometimes highly technical. In the context of machine learning, a common definition is "the ability to explain or present in understandable terms to a human" (Doshi-Velez and Kim, 2017). This abstract definition has been translated into a myriad of different operational definitions based on a wide range of criteria. In fact, several taxonomies for interpretability of neural networks have been proposed and the number of published articles on interpretability has been increasing exponentially since 2000 (Fan et al., 2020). Therefore, here we will restrict ourselves to distinguishing between global and local interpretability and explaining the relevance of these two concepts for population genomics studies. Also, we note that we will not consider very recent efforts aimed at designing inherently interpretable deep neural networks (e.g. (Chen et al., 2020)) and instead focus on post-hoc interpretation methods, *i.e.* algorithms that can be used to interpret an already trained network.

Global interpretability aims at explaining the overall behaviour of a model (Ancona et al., 2019b), which in turn can inform us about the system being studied. In principle, this goal can be achieved by analysing the hyperparameters (which control the learning process and the values taken by the parameters; e.g. learning rate, activation function, number of hidden layers, number of neurons per hidden layer) or parameters (weights and biases) of a deep neural network. However, the information provided by hyperparameters tend to be limited to model complexity, for example in terms of the number of nodes and hidden layers retained after tuning and fitting or the type of activation function. On the other hand, the values taken by parameters (weights and biases) after fitting can provide more meaningful biological information; in particular, they help identify the features that contributed the most to the predictive power of the algorithm. For example, Sheehan and Song (Sheehan and Song, 2016) (see FCNN section above) use random permutation of each summary statistic (feature) and identify as most informative for the detection of population size changes those statistics that, when randomly permuted, lead to the sharpest decrease in accuracy. Another approach is based on feature importance (Olden and Jackson, 2002), which was used by another study (Qin et al., 2022) to identify as outlier loci those that contributed the most to the power of a FCNN to predict an individual's phenotype or geographic origin. Feature importance is based on the idea that the magnitude of connection weights between neurons connecting input and output nodes measure the extent to which each feature contributes to the network's predictive power. The architecture used for these two examples was a FCNN. A different approach is necessary in the case of CNNs. For example, in the case of a CNN that classify images into different categories, a common approach is to use saliency maps, which measure the support that different groups of pixels in an image provides for a particular class (Mohamed et al., 2022). This is implemented by feeding the CNN an image of a particular class and using visualisation techniques to generate heatmaps overlayed on the original image; the image elements that are being used by the CNN to identify the class are highlighted in red. A

population genetics application of this approach is presented by Gower and colleagues, who used a CNN algorithm to detect adaptive introgression (Gower et al., 2021).

Local interpretability aims at understanding the reasons for a specific decision concerning a particular instance. Note that the ability of a particular feature to predict an attribute (e.g. phenotype) for a particular instance (data point), may depend on the values taken by the other features. This is particularly relevant in population genomics applications as the effect that a particular locus variant has on the phenotype of an individual may depend on the variants found at other loci (*i.e.* the genetic background; (Chandler et al., 2013)). A very promising technique to address this important issue is the Shapley value approach (Strumbelj and Kononenko, 2010). Shapley values were first introduced in cooperative game theory (Shapley, 1953) to calculate the contribution of individual players to the outcome of a game. In the context of deep learning, each feature represent a player, different combinations of features (feature subsets) represent a coalition, and the set comprising all features represents the "grand coalition of players". The objective is to explain how values of a feature for a particular instance contribute to the difference between the prediction of a machine learning algorithm with the feature included and the expected prediction when the feature value is ignored (Strumbelj and Kononenko, 2010). Thus, the Shapley value of a feature can be interpreted as the average marginal contribution of the feature to all possible feature subsets that can be formed without it (c.f., Ancona et al., 2019a). An important advantage of the approach is that it is the only explanation method that takes into account all the potential dependencies and interactions between feature values (c.f., Strumbelj and Kononenko, 2010). In principle this requires the evaluation of all 2^N feature subsets (coalitions), were N is the number of features in the full set (grand coalition). Obviously, this is only possible when the number of features is small to moderate (some few dozens). Thus, several algorithms have been proposed for approximating Shapley values and a unified approach proposed by (Lundberg and Lee, 2017) has been implemented in both `python` (KernelShap and DeepShap) and `R` (shapr). However, they are limited to deep neural networks with moderate number of features. Nevertheless, very recent developments have led to new approaches, DASP (Ancona et al., 2019a) and G-DeepShap (Chen et al., 2022), that may scale up to population genomics datasets. For the moment there are no applications of Shapley values to population genomics studies; there is only an application in population genetics but in the context of random forests (Kittlein et al., 2022).

Much work remains to be done in order to incorporate the latest advances in interpretable machine learning to population genomics. Interpretability can lead to important breakthroughs by uncovering complex genomic signatures left by the non-linear interactions among many genetic and evolutionary processes. Although population genetics theory has already provided a deep understanding of the genomic signatures left by complex demographic history and selective processes, the 'agnostic' nature of deep learning has the potential to uncover 'hidden' genomic signatures that traditional model-based statistical methods are unable to

detect. In doing so, they may generate new hypotheses for explaining observed genomic patterns that could then be tested.

2.8 Dealing with uncertainty

Whilst, as described so far, deep learning has led to novel applications in population genetics, the intrinsic challenges associated with uncertain DNA sequencing data, simulated training data sets, and an incomplete statistical framework are limiting factors to fully exploit the power of such technique.

As previously described, data given as input to deep learning algorithms in population genetics typically consist of alignments of genotypes, inferred haplotypes, or summary statistics. Genotype calling, phasing, and calculation of summary statistics are associated with statistical uncertainty (Nielsen et al., 2011), especially when performed from low-coverage sequencing (*i.e.* from museum specimen, ancient samples, or generally non-model species) (Lou et al., 2021). Sequencing data uncertainty could be tackled by providing estimates of summary statistics from genotype likelihoods as input. Additional approaches based on filtering masks to take into account data errors and missingness have been proposed in the literature (Adrion et al., 2020b). Finally, generating sequencing data-like simulations (Escalona et al., 2016; Cury et al., 2022) for training could be a valuable solution to accommodate all nuances of the experimental data, at the expense of increasing computational resources needed. Other sequencing technologies may provide data of different nature (e.g., sample allele frequencies from pooled-sequencing experiments (Anand et al., 2016)), and therefore appropriate considerations should be made in terms of additional statistical uncertainty associated with such output. Approaches based on using trees or local ancestry tracts as input (Hamid et al., 2022) may be more prone to input data uncertainty.

One of the main concerns about current applications of deep learning in population genetics is the use of synthetic data for training neutral networks. For instance, the detection of signals of natural selection typically requires the knowledge of the underlying demography model to generate a null distribution under neutrality (Nielsen, 2005). If the baseline demographic model is ill-defined, inference of natural selection is expected to be biased (Johri et al., 2022c). Whilst such issue is shared with other popular inferential frameworks, such as ABC (Bertorelle et al., 2010), the use of simulations in this context appears to be more problematic given the 'black-box' nature of neural networks. Solutions to address the uncertainty of simulations explored in the literature include testing a network trained on misspecified models (e.g., Flagel et al., 2018; Torada et al., 2019; Adrion et al., 2020b), and deploying it on known cases of selection and neutrality (Isildak et al., 2021a) to quantify false positive and false negative rates. Although post-inference diagnostic analyses are required to ensure robustness of results, as per best-practice in machine learning (Lones, 2021; Whalen et al., 2022), the ever increasing curated list of demographic models (Adrion et al., 2020a) will facilitate the

use of synthetic data for training networks. Likewise, these resources will facilitate the establishment of gold-standard data sets to benchmark newly proposed architectures. Finally, efforts towards the adoption of transfer learning and domain adaptation techniques should further reduce any bias associated with uncertain training data sets.

Most applications described herein aim at classifying data into discrete labels or providing point-estimates of parameters of interests. Statistical uncertainty should be quantified by characterising probability distributions of both the model uncertainty (epistemic or reduceable part) and the inherent stochastic uncertainty of data generating process (aleatoric or irreduce-able uncertainty) (Sanchez et al., 2022b; Hüllermeier and Waegeman, 2021). Solutions to this problem include the prediction of mean and standard deviation (Chan et al., 2018) or confidence intervals alongside point estimates, and the quantification of any errors associated with the training phase (Smith et al., 2022). Thus, we encourage practitioners for the upcoming publications to consider modifying their models to account for uncertainty in a principled manner.

From regular convolutions to graph convolutions

Genotype matrices have been the starting point for doing any kind of population genetics analysis, either by calculating summary statistics (e.g., site frequency spectra), model-based probabilistic optimisation algorithms (e.g., SMC), or Bayesian sampling techniques (e.g., ABC) and non-model based function approximations (e.g., deep learning). Yet, recent trends emphasise a need to combine the power of deep learning approaches with a model-based constraint. A promising idea is to format the input data (genotype matrix) in order for model assumptions to be encoded directly in the data for subsequent training and inference. In the most general case, this model-based formatting can be considered as a representation of the ARG, for which few methods have been developed (Rasmussen et al., 2014b; Speidel et al., 2019; Kelleher et al., 2019; Mahmoudi et al., 2022b). Decoupling the ARG or genealogy construction and inference of evolutionary parameters of interest would create the opportunity to increase collaborations with mathematical modellers, by incorporating more complex coalescent models or biological processes like introgression, structured populations, or species-specific life history traits. Additionally, it may no longer be necessary to try to interpret the inner workings of a CNN trained on (sparse) genotype matrices (which likely rebuilds parts of the ARG through complex aggregation of genotype density patterns). Any type of model-based properties could be questioned through modification of the ARG. An essential step has been developed by Korfmann *et al.*, providing not only a new ARG-parameter inference method based on graph neural networks (GNN) but also a SMC method applied to a particular coalescent model, known for long-range LD interdependencies (Korfmann et al., 2022b). This approach offers the unique opportunity to test for mathematical model-based blind spots in an inherently-Markovian constrained SMC method using GNNs.

2.9 Conclusions

This review illustrates the great diversity of deep learning architectures that have been used in population genetics applications. Currently, the prevailing type of applications involve the training of algorithms with simulated data but there is an increasing number of studies that use a more standard approach where training is carried out using observed data. Thus, we can identify two strands of methods, one that is closely associated with likelihood-free, simulation-based approaches that consider explicit evolutionary models and another one that conforms to a purely data-driven, model-free approach. In both cases, however, deep learning is used as an inferential tool (as opposed to a predictive or pattern recognition approach). However, as the popularity of deep learning increases among population geneticists, we expect that further deep learning algorithms, including the latest diffusion models (Ramesh et al., 2022), will be adapted to solve predictive tasks. Intriguingly, novel applications may go beyond classic inferential tasks and include other aims, such as efficient data compression or generation of synthetic experimental data sets. Likewise, solutions for making neural networks a 'transparent-box', such as neural additive models (Novakovsky et al., 2022) and symbolic metamodelling (Alaa and van der Schaar, 2019), will facilitate the adoption of deep learning among empiricists.

More research is needed in the domain of 'interpretable' machine learning so as to gain an understanding of how deep learning algorithms make their decisions. This in turn would enable population geneticists to uncover novel genomic signatures associated with non-linear processes that current theory has not yet suggested including non-linear interactions among many genetic, ecological, and evolutionary processes. Importantly, further developments in local interpretability (see above) can help us to identify epistatic interactions and gain a better understanding of how genetic background influences the phenotypic effect of mutations.

One key aspect to make deep learning a popular framework in population genetics, is to ensure reproducible analyses and avoid repeating training of highly parameterised networks from scratch. In this context, recent efforts to provide users with documented workflows (Whitehouse and Schrider, 2022) and pre-trained networks (Hamid et al., 2022) will both reduce carbon footprint (Grealey et al., 2022) and facilitate the application of deep learning to a wider range of data sets, allowing users to modify the network's parameters according to the specific requirements of the biological system under examination.

Finally, we urge the community to make the field as inclusive as possible. Whilst open-source software release is common practice among machine learning practitioners, access to appropriate computing resources is still a limiting factor for many researchers. Initiatives to provide GPUs (*i.e.* graphics processing unit) and cloud computing credits to academics in need represent a valuable step towards making deep learning in population genetics accessible and inclusive to a wide range of scientists. Likewise, we encourage the establishment of training opportunities in machine learning for early-career population geneticists. Impor-

tantly, such events should happen either online or in hybrid format, with resources provided in multiple languages to ensure that text or verbal comprehension is not a barrier to learning. Consortia and local networks, properly funded by the wealthiest countries, appear to be a natural solution to fulfil this need. If all these conditions are met, deep learning will soon be established as part of the common toolkit among population geneticists globally.

Acknowledgements

We are grateful to all members of the EvoGenomics.AI consortium (`www.evogenomics.ai`) for helpful discussions. We also wish to thank Ulas Isildak for assistance in using `SLiM`. Two anonymous reviewers provided insightful comments that improved the manuscript.

Funding

KK is supported by a grant from the Deutsche Forschungsgemeinschaft (DFG) through the TUM International Graduate School of Science and Engineering (IGSSE), GSC 81, within the project GENOMIE QADOP. We acknowledge the support of Imperial College London - TUM Partnership award.

Data Availability Statement

No new data were generated in support of this research. An implementation of the neural network illustrated in this review is available at `https://github.com/kevinkorfmann/temporal-balancing-selection`.

3 Weak Seedbanks Influence the Signature and Detectability of Selective Sweeps

The following chapter has been published as:
Kevin Korfmann, Diala Abu Awad and Aurélien Tellier (2023). Weak seed banks influence the signature and detectability of selective sweeps. Journal of Evolutionary Biology, 36, 1282–1294. https://doi.org/10.1111/jeb.14204

KK designed and implemented the simulator, ran the analysis, added the figures, contributed to writing manuscript and worked on the revisions.

3.1 Abstract

Seed banking (or dormancy) is a widespread bet-hedging strategy, generating a form of population overlap, which decreases the magnitude of genetic drift. The methodological complexity of integrating this trait implies it is ignored when developing tools to detect selective sweeps. But, as dormancy lengthens the ancestral recombination graph (ARG), increasing times to fixation, it can change the genomic signatures of selection. To detect genes under positive selection in seed banking species it is important to 1) determine whether the efficacy of selection is affected, and 2) predict the patterns of nucleotide diversity at and around positively selected alleles. We present the first tree sequence-based simulation program integrating a weak seed bank to examine the dynamics and genomic footprints of beneficial alleles in a finite population. We find that seed banking does not affect the probability of fixation and confirm expectations of increased times to fixation. We also confirm earlier findings that, for strong selection, the times to fixation are not scaled by the inbreeding effective population size in the presence of seed banks, but are shorter than would be expected. As seed banking increases the effective recombination rate, footprints of sweeps appear narrower around the selected sites and due to the scaling of the ARG are detectable for longer periods of time. The developed simulation tool can be used to predict the footprints

of selection and draw statistical inference of past evolutionary events in plants, invertebrates, or fungi with seed banks.

3.2 Introduction

Seed banking is an ecological bet-hedging strategy, by which seeds or eggs lay in a dormant state of reduced metabolism until conditions are more favourable to hatch or germinate and complete the life-cycle. This life-history trait acts therefore as a buffer in uncertain environments (Cohen, 1966; Templeton and Levin, 1979) and has evolved several times independently in prokaryotes, fungi, plants, and invertebrates (Evans and Dennehy, 2005; Nara, 2009; Willis et al., 2014; Tellier, 2019; Lennon et al., 2021b). Because several generations of seeds are simultaneously maintained, seed banks act as a temporal storage of genetic information (Evans and Dennehy, 2005), decreasing the effect of genetic drift and lengthening the time to fixation of neutral and selected alleles (Templeton and Levin, 1979; Hairston Jr and De Stasio Jr, 1988). Seed banks are therefore expected to play an important role in determining the adaptive potential of a species (Tellier, 2019). In bacteria (Shoemaker and Lennon, 2018; Lennon et al., 2021b), invertebrates (Evans and Dennehy, 2005) or plants (Willis et al., 2014; Tellier, 2019), dormancy determines the neutral and selective diversity of populations by affecting the effective population size and buffering population size changes (Nunney and Ritland, 2002), affecting mutation rates (Levin, 1990; Whittle, 2006; Dann et al., 2017), genetic structure (Vitalis et al., 2004), rates of population extinction/recolonization (Brown and Kodric-Brown, 1977; Manna et al., 2017) and the efficacy of positive (Hairston Jr and De Stasio Jr, 1988; Koopmann et al., 2017; Heinrich et al., 2018; Shoemaker and Lennon, 2018) and balancing selection (Tellier and Brown, 2009; Verin and Tellier, 2018).

Seed banking, or dormancy, introduces a time delay between the changes in the active population and changes in the dormant population which considerably increases the time to reach the common ancestor of a sample of genes from the active population (Kaj et al., 2001; Blath et al., 2015, 2016, 2020). We note that two models of seed banks are proposed, namely the weak and strong dormancy models. These make different assumptions regarding the scale of the importance of dormancy relative to the evolutionary history of the species. On the one hand, the strong version is conceptualized after a modified two-island model with coalescence events occurring only in the active population as opposed to the dormant population (seed bank) with migration (dormancy and resuscitation) between the two (Blath et al., 2015, 2016, 2019; Shoemaker and Lennon, 2018). Strong seed bank applies more specifically to organisms, such as bacteria or viruses, which exhibit very quick multiplication cycles and can stay dormant for times on the order of the population size (thousands to millions of generations, Blath et al., 2015, 2020; Lennon et al., 2021b). On the other hand, the weak seed bank model assumes that dormancy occurs only over a few generations (tens to hundreds), thus

seemingly negligible when compared to the order of magnitude of the population size (Kaj et al., 2001; Tellier et al., 2011; Živković and Tellier, 2012; Sellinger et al., 2019), making it applicable to plant, fungi or invertebrate (*e.g. Daphnia sp.*) species or when the seed bank is experimentally imposed (as it is in practice difficult to generate the strong seed bank) (Shoemaker et al., 2022). We focus here on a pseudo-diploid version of the weak seed bank model in order to provide novel insights into the population genomic analysis of species which undergo sexual reproduction. The applicability of our results, as well as the differences and similarities between the strong and weak seed bank models, are highlighted in the Discussion.

The weak seed bank model can be formulated forward-in-time as an extension of the classic Wright-Fisher model for a population of size N haploid individuals. The constraint of choosing the parents of offspring at generation t only from the previous generation $(t-1)$ is lifted, and replaced with the option of choosing parents from previous generations $(t-2, t-3, \ldots$ up to a predetermined boundary $t-m)$(Nunney and Ritland, 2002). The equivalent backward-in-time model extends the classic Kingman coalescent and assumes an urn model in which lineages are thrown back-in-time into a sliding window of size m generations, representing the past populations of size N (Kaj et al., 2001). Coalescence events occur when two lineages randomly choose the same parent in the past. The germination probability of a seed of age i is b_i, which is equivalent to the probability of one offspring choosing a parent i generations ago. The weak dormancy model is shown to converge to a standard Kingman coalescent with a scaled coalescence rate of $1/\beta^2$, in which $\beta = \frac{\sum_{i=1}^{m} b_i}{\sum_{i=1}^{m} i b_i}$ is the inverse of the mean time seeds spend in the seed bank, and m is the maximum time seeds can be dormant (Kaj et al., 2001). The intuition in a coalescence framework (Kaj et al., 2001) is that for two lineages to find a common ancestor, *i.e.* to coalesce, they need to choose the same parent in the active population, each the probability β to do so, as only active lineages can coalesce. Thus the probability that two lineages are simultaneously in the active population is a β^2 scaling of the coalescence rate. The germination function was previously simplified by assuming that the distribution of the germination rate follows a truncated geometric function with rate b, so that $b = \beta$ when m is large enough (Tellier et al., 2011; Živković and Tellier, 2012; Sellinger et al., 2019, see methods). A geometric germination function is also assumed in the forward-in-time diffusion model analysed in Koopmann et al., 2017; Heinrich et al., 2018 and Blath et al., 2020.

Seed banking influences neutral and selective processes via its influence on the rate of genetic drift. In a nutshell, a seed bank delays the time to fixation of a neutral allele and increases the inbreeding effective population size (from now on referred to only as the "effective population size") by a factor $1/b^2$. The effective population size under a weak seed bank is defined as $N_e = \frac{N_{cs}}{b^2}$ where N_{cs} is the census size of the active population (Nunney and Ritland, 2002; Tellier et al., 2011; Živković and Tellier, 2012). Mutation under an infinite site model can occur in seeds with probability μ_d and μ_a in the active population, so that we can

define θ the population mutation rate under the weak seed bank model: $\theta = \frac{4N_{cs}(b\mu_a+(1-b)\mu_d)}{b^2}$ (Tellier et al., 2011). If mutations occur in the dormant population at the same rate as in the active population, we define $\mu_d = \mu_a = \mu$ yielding $\theta = \frac{4N_{cs}\mu}{b^2}$, while if the dormant state does not mutate, $\mu_d = 0$ and $\mu_a = \mu$, yielding $\theta = \frac{4N_{cs}\mu}{b}$. Empirical evidence (Levin, 1990; Whittle, 2006; Dann et al., 2017) and molecular biology experiments have shown that even under reduced metabolism DNA integrity has to be protected (Waterworth et al., 2016), and suggest that mutations occur in the dormant population (for simplicity at the same rate as in the active population, see model in Sellinger et al., 2019). Furthermore, recombination and the rate of crossing-over is also affected by seed banking. However, only the non-dormant lineages are affected by recombination in the backward-in-time model so that the population recombination rate is $\rho = 4N_e rb = \frac{4N_{cs}r}{b}$. The recombination rate r needs to be multiplied by the probability of germination b as only active individuals can recombine (Živković and Tellier, 2018; Sellinger et al., 2019). The ratio of the population mutation rate and the recombination rate defines the amount of nucleotide diversity in the genome as well as the amount of linkage disequilibrium, a property which has been used to develop a sequential Markovian coalescent (SMC) approach to jointly estimate past demographic history and the germination rate (Sellinger et al., 2019, 2021b).

While there is now a thorough understanding of how neutral diversity is affected by seed banking, the dynamics of alleles under selection have not been fully explored. Koopmann et al., 2017 developed a diffusion model of infinite (deterministic) seed bank model with positive selection and show that the time to fixation is not multiplied by $1/b^2$ (as for neutral alleles) but by a higher factor (between $1/b^2$ and $1/b$). The interpretation is as follows: while the time to fixation of an advantageous allele is lengthened compared to a model without dormancy, the efficacy of selection should be altered compared to a neutral allele (the effect of genetic drift). Namely, the Site Frequency Spectrum (SFS) of independently selected alleles shows an increased deviation from neutrality with a decreasing value of b. By relaxing the deterministic seed bank assumption, Heinrich et al., 2018 find that: 1) a finite small seed bank decreases the efficacy of selection, and 2) selection on fecundity (production of offspring/seeds) yields a different selection efficiency compared to selection on viability (seed viability), as can be seen from their estimated Site-Frequency Spectrum (SFS) of independent alleles under selection. Furthermore, based on the effect of seed banks on θ and ρ and on selection, verbal predictions on the genomic signatures of selection have been put forth (Živković and Tellier, 2018).

These theoretical and conceptual approaches, while paving the way for studying selection under seed banks, did not consider the following argument. If the time to fixation of an advantageous allele increases due to the seed bank, it can be expected that 1) drift has more time to drive this allele to extinction, and 2) the signatures of selective sweeps can be erased

by new mutations appearing in the vicinity of the selected alleles. These effects would counter-act predictions that selection is more efficient under a stronger seed bank compared to genetic drift, as well as, that selective sweeps are more easily observable under stronger seed bank (Koopmann et al., 2017; Živković and Tellier, 2018). In order to resolve this paradox, we develop and make available the first simulation method for the weak seed bank model, which allows users to generate full genome data under neutrality and selection. We first present the simulation model, which we use to follow the frequencies of an adaptive allele in a population with seed banking. We aim to provide insights into the characteristics of selective sweeps, including the time and probability of fixation, as well as recommendations for their detection in species exhibiting seed banks.

3.3 Methods

Forward-in-time individual-based simulations are implemented in C++. Genealogies are stored and manipulated with the tree sequence toolkit (tskit, Kelleher et al., 2018), which allows for a general approach to handling arbitrary evolutionary models and an efficient workflow through well-documented functions.

3.3.1 Model

The model represents a single, panmictic population of N hermaphroditic pseudo-diploid adults, which will henceforth be referred to as diploids for brevity. Population size is fixed and generations are discrete, so that in the absence of dormancy and selection, the population follows a classic Wright-Fisher model. In this case, at the beginning of each generation, a new individual is produced by sampling two parents from the previous generation. Once sampled, each parent contributes a (recombined) gamete to generate the new individual. Each parent is sampled with probability $\frac{1}{N}$ (multinomial sampling), leading to two vectors $\mathbf{X}_{parent1}$ and $\mathbf{X}_{parent2}$, containing the indices of the respective parents:

$$\mathbf{X}_{parent1} = (X_1^1, X_2^1, ..., X_N^1) \sim Mult(N, \tfrac{1}{N}) \text{ with } \{X_i^1 \in \mathbb{N} : X_i^1 \leq N\}$$

$$\mathbf{X}_{parent2} = (X_1^2, X_2^2, ..., X_N^2) \sim Mult(N, \tfrac{1}{N}) \text{ with } \{X_i^2 \in \mathbb{N} : X_i^2 \leq N\}$$

Dormancy adds a layer of complexity, by introducing seeds that can germinate after being dormant for many generations. This relaxes the implicit Wright-Fisher assumption, as parents are no longer only sampled from the previous generation, but also from dormant individuals produced up to m generations in the past. The probability of being sampled from generation k depends on the probability of germination, which is a function of the age of the dormant individual. As for the classical Wright-Fisher model, there are $2N$ possible parents. The parents are sampled using a probability vector $\mathbf{Y}^{norm}$ written as:

$$\mathbf{Y} = (Y_1, Y_2, Y_k, ..., Y_m) \text{ with } Y_k = b(1-b)^{k-1} \text{ and } \{Y_k \in \mathbb{R} : Y_k > 0\}$$
$$\text{from which we obtain: } \mathbf{Y}^{norm} = \frac{\mathbf{Y}}{\sum_{j=1}^{m} Y_j}$$

From the expression above, the probability of being sampled follows a truncated geometric distribution parameterized with germination rate b and then normalized. The generation G of each parent is randomly sampled using a multinomial sampling with the probability vector $\mathbf{Y}^{norm}$.

$$\mathbf{G}_{parent1} = (G_1^1, G_2^1, ..., G_N^1) \sim Mult(N, \mathbf{Y}^{norm}) \text{ with } \{G_i^1 \in \mathbb{N} : G_i^1 \leq N\}$$

$$\mathbf{G}_{parent2} = (G_1^2, G_2^2, ..., G_N^2) \sim Mult(N, \mathbf{Y}^{norm}) \text{ with } \{G_i^2 \in \mathbb{N} : G_i^2 \leq N\}$$

Once the age of each of the $2N$ parents has been determined, random individuals from the corresponding age groups are sampled (the same individual can be sampled more than once) and one recombined gamete from each of these $2N$ individuals is generated. These gametes are then randomly combined to form N new diploid individuals which constitute the current active population. Thus, the forward simulation process models two haploid dormant individuals (with different ages) which become active at the current generation and join to form a diploid individual (Figure 3.1). This pseudo-diploid model formulation is implicitly equivalent to haploid gametes being resuscitated from the dormant state and fusing to create a diploid individual capable of reproduction. The probability of coalescence (p_{coal}) is therefore expected to follow haploid expectations $p_{coal} = (\frac{1}{2N}) \times b^2$). The number of recombination events is sampled from a Poisson distribution with parameter r (for example 1×10^{-8} per bp per generation). At the end of this process, new mutations can be introduced (only necessary for sweep detection tools). Generally neutral mutations are not simulated and statistics are computed using branch lengths. We assume here that mutations are also introduced at every generation in dormant individuals at the same rate (following Sellinger et al., 2019), even if they are not explicitly simulated. Recombination breakpoints are uniformly distributed across the genome with each coalescent tree being delineated by two recombination breakpoints.

To model selection signatures within a neutral genomic background, we consider non-neutral bi-allelic loci, placed at predefined and fixed genomic positions, with beneficial mutations arising after the burn-in period. A locus under selection has a dominance h and selection coefficient s, respectively. The expressions for the fitness of heterozygote and homozygote individuals with the beneficial mutation are thus $1 + hs$ and $1 + s$, respectively. Fitness affects the probability that an individual's gametes can leave the dormant state and contribute to reproduction. The choice of the germinating generation when sampling the parents is unaffected by their fitness values, but the sampling of individuals within a given generation is determined by the fitness. In other words, selection acts on fecundity, as the fitness of an allele determines the number of offspring produced and not survival (viability selection). A selection coefficient of 0 would lead to multinomial Wright-Fisher sampling,

which can be used to track neutral mutations over time. This two-step process of first choosing the generation followed by the individual is presented in Figure 3.1.

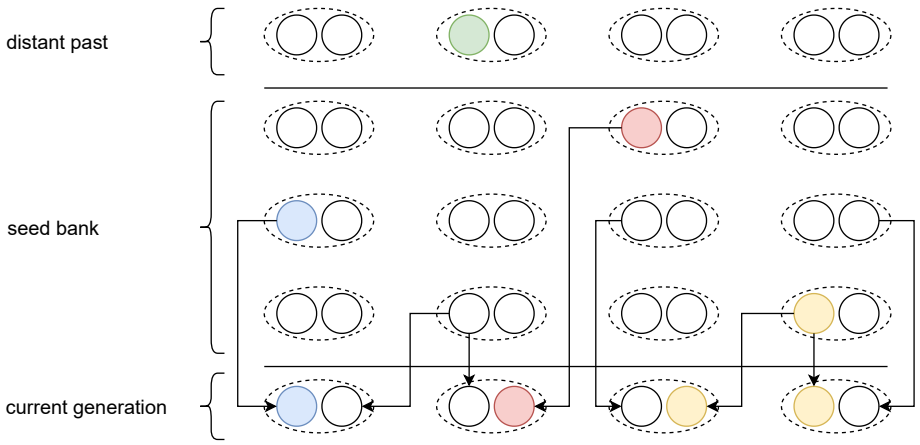


Figure 3.1 Schematic representation of our pseudo-diploid weak dormancy seed bank model by a forward-in-time two step process in the spirit of Kaj et al., 2001 for haploid dormant seeds. The arrows originating from the parent or seed generation represent the geometric sampling process of the current generation, and the sampling of the individual within the given generation of the past based on the respective fitness value.

From a technical perspective, individuals can be tracked in the tskit-provided table data structures, if the *tree_sequence_recording* feature is enabled. This feature is not required when computing statistics on allele frequency dynamics only (*i.e.* to compute fixation times or probabilities). The tables used in this simulation are as follows: 1) a node table representing a set of genomes, 2) an edge-table defining parent-offspring relationships between node pairs over a genomic interval, 3) a site table to store the ancestral states of positions in the genome, and 4) a mutation table defining state changes at particular sites. The last two tables are only used to introduce the mutation under selection. If neutral mutations are required for down-stream analysis, they are simulated after this step. The simulation code works with the aforementioned tables through tskit functions, e.g. the addition of information to a table after sampling a particular individual or through the removal of parents who do not have offspring in the current generation in a recurrent simplification process. This clean-up process is a requirement to reduce RAM-usage during the simulation, because keeping track of every individual ever simulated to build the genealogy quickly becomes infeasible. However, a noticeable difference to the classic use of the tskit function is that in our case individuals which have not produced offspring in the past, but are still within the dormancy upper-bound defined range of m generations, need to be protected from the simplification process, which is achieved by marking them as *sample nodes* during the simulation. Indeed, forward-in-time, a parent can give offspring many generations later (maximum m) through germinating seeds. As previously stated, the simulation process can be run independently of tskit, but the latter is required when planning to analyze the genealogy.

3.3.2 Simulations

Except when indicated otherwise, the population size is generally set to $N = 500$ individuals or $2N = 1,000$ haploid genomes. We specifically change population size when testing whether sweep signatures can be explained by simple size scaling. In this case we use $N = 2000$ individuals with a germination rate of $b = 1$, corresponding to $N = 245$ for $b = 0.35$ (Figure 3.14). Our focal seed bank setup is that of a population of $N = 500$ individuals with a germination rate $b = 0.35$ and dominance coefficient $h = 0.5$.

The genome sequence length is set to 100,000 bp, 1MB or 10 MB. Simulations start with a burn in or calibration phase of 50,000 generations for $b = 1$, and 200,000 generations for $b = 0.5$ (see Figure 3.6, Table 3.1 for the calibration method used to define the of generations needed for a given recombination rate), to make sure full coalescence has occurred and a most-recent common ancestor is present. We consider that after this initial phase, the population is at an equilibrium state in terms of neutral diversity, including within the seed bank. After this phase, one selectively advantageous mutation is introduced at the predefined site. To study sweep signatures as well as the time it takes for sweep signatures to recover, simulations are run for several generations after fixation of the beneficial allele (up to 16,000 generations after fixation).

Neutral diversity is calculated based on the branch length, meaning that explicitly simulating mutations is not required. To check whether the strength of a sweep behaves in accordance to expectations *i.e.* lower recombination rates result in wider sweeps, recombination rates ranging from 5×10^{-8} to $r = 10^{-7}$ are tested for all parameter sets. Simulations are run for the germination rate b ranging from 0.25 up to 1 (with $b = 1$ meaning no dormancy). The upper-bound number of generations m which is the maximum time that seeds can remain dormant (*i.e.* seeds older than m are removed from the population) is set at 30 generations. Beneficial mutations have a selective coefficient $N_e^{b=1}s$ ranging from 0.1 to 100 and dominance h takes values 0.1, 0.5 and 1.1, representing recessive, co-dominant and overdominant beneficial mutations.

3.3.3 Statistics and sweep detection

We first calculate several statistics relative to the forward-in-time change of the frequency of an advantageous allele in the population, such as the mean time to fixation and the probability of fixation, using 1,000 simulations per parameter configuration. Each simulation run consists of the recurrent introduction over time of an allele (mutant at frequency $1/2N$) which is either lost or fixed. When an allele is lost and the simulation is conditioned on fixation a new simulation starts from a neutral genetic diversity background (see below for more details). An allele is considered to be fixed if its number of copies is 2N for m consecutive generations. For each simulation run we store 1) the time it takes for the last introduced allele to reach fixation (time between allele introduction until fixation), and 2) the number of alleles which

were introduced until one has reached fixation (yielding the probability of fixation of an allele per simulation run). The resulting times to fixation and fixation probabilities are calculated as the averages over the 1,000 simulation runs.

We also compute statistics on the underlying coalescent tree and ancestral recombination graph (ARG) such as time to the most recent common ancestor, linkage disequilibrium (r^2, Hill and Robertson, 1968), as well as Tajima's π and D (Tajima, 1983; Nei and Li, 1979; Tajima, 1989) over windows of size 5,000 (giving 200 windows for a sequence length of 1 MB). This allows us to analyse the effects of seed-dormancy on the amount of linkage disequilibrium and nucleotide diversity along the genome, as well as the footprint of a selective sweep on these quantities. *Tskit* functions are used for diversity and linkage disequilibrium calculations. Nucleotide diversity (π) is calculated based on the branch length. Sweeps are detected using Omega and SweeD statistic, the first one quantifies the degree to which LD is elevated on both sides of the selective sweeps, as implemented and applied with OmegaPlus (Alachiotis et al., 2012), while SweeD (Pavlidis et al., 2013) uses changes in SFS across windows to detect sweeps. A difficult issue in detecting selective sweeps is choosing the correct window size to perform the computations. It is documented that the optimal window size depends on the recombination rate and thus the observed amount of linkage disequilibrium (Alachiotis et al., 2012; Alachiotis and Pavlidis, 2016). We use two different setups with different window sizes: –minwin 2000 –maxwin 50000 and –minwin 1000 –maxwin 25000 . The window sizes refer to the minimum and maximum region used to calculate LD values between mutations. Importantly the –minwin parameter determines the sensitivity, meaning the degree to which false positives or false negatives (high –minwin values) are detected, while the –maxwin parameter determines run-time and memory requirements. A detailed graphical description can be found in the online OmegaPlus manual. In theory the larger window size is more appropriate for the model without dormancy ($b = 1$), and the narrower window size for the model with dormancy ($b < 1$). For both cases, we set –grid 1000 –length 10 MB. SweeD is only tested using a –grid 1000 parameter. The statistic is computed for a sample size of 100 over 400 simulations for each sweep signature at multiple generations after fixation (sweep recovery scenerios).

3.3.4 Code description and availability

Source code of the simulator and demonstration of the analysis can be found at `https://gitlab.lrz.de/kevin.korfmann/sleepy` and `https://gitlab.lrz.de/kevin.korfmann/sleepy-analysis`. A convenient feature of the simulator is the option to choose between switching the tree sequence recording on or off depending on the question, *i.e.* if analysing fixation time and probability of fixation it is unnecessary to record the tree sequence (or use a calibration phase). To analyse the sweep signatures, the simulation pro-

cess has been divided into two phases to alleviate the large run-times of forward simulations. During the first phase, a tree sequence will be generated under neutrality and stored to disk. And in the second phase the neutral tree sequence is loaded and a parameter of interest is tested until fixation or loss. Additionally, if the simulation is conditioned on fixation, then the simulation can start again from the beginning of the second phase that will have been run for tree sequence calibration, saving the time.

```
import tskit
from utils import sleepy

sleepy(N=500, b=1.0, m=30, output_directory="./b1/")
ts_no_selection = tskit.load("./b1/run_0.trees")

sleepy(N=500, b=1.0, m=30, s=0.01, output_directory="./b1_s001/")
ts_selection = tskit.load("./b1_s001/run_0.trees")
```

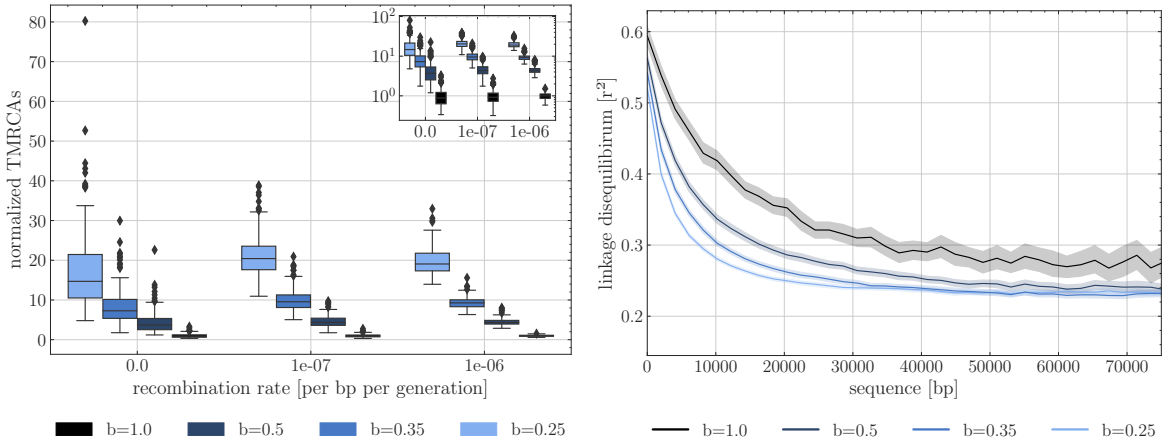
Listing 1 Simplified, demonstrative Python code example for a simulation with and without selection. Tree sequence results are stored in a specified output directory and are loaded via *tskit* function for further processing or analysis of e.g. linkage disequilibrium or nucleotide diversity along the genome. A more detailed version with more parameters can be found in the example notebook at https://gitlab.lrz.de/kevin.korfmann/sleepy-analysis.

Simulations rely on regular simplification intervals for efficiency of the genealogy recording, yet the weak dormancy model requires keeping up to m generations in memory even for past individuals (seeds) which do not have offspring in the current generation. To make sure that this assumption is realized in the code, up to m generations are technically defined as leaf nodes, thus hiding them from the regular memory clean-up process. Furthermore, the presence or absence of an allele with an associated selection coefficient needs to be retrievable, even under the influence of recombination, for all individuals for up to m generations in order to determine the fitness of the potential parents. Therefore, recombination and selective alleles are tracked additionally outside of the *tskit* table data structure, allowing the running of the the simulation without the tree sequence. Both of these model requirements, namely maintaining individuals which do not have offspring in the current generation (but potentially could have due to stochastic resuscitation of a seed) as well as the knowledge about the precise state of that given individual in the past, are reasons to choose our own implementation over SLiM (Haller and Messer, 2019).

3.4 Results

3.4.1 Neutral coalescence

We first verify that our simulator accurately produces the expected coalescent tree in a population with a seed bank with germination parameter b and population size N. To do



(a) **(b)**

Figure 3.2 (a) Time to the most recent common ancestor (TMRCA) as a function of the germination rate b and scaled by results under $b = 1$. For each germination rate, three recombination rates per site are presented ($r = 0$, $r = 10^{-7}$ and $r = 10^{-6}$. Boxes describe the 25th (Q1) to 75th percentile (Q3), with the lower whisker representing Q1-1.5×(Q3-Q1) outlier threshold and the upper whisker is calculated analogously. The mean is plotted between Q3 and Q1. Each boxplot represents the distribution of 200 TMRCA values over 200 sequences of 0.1 Mb. Per sequence the oldest TMRCA is retained. (b) Monotonous decrease of linkage disequilibrium as a function of distance between pairs of SNPs, setting $r = 10^{-7}$ per generation per bp, sequence length to 10^5 bp. While population size is 500, linkage decay was calculated by subsetting 200 individuals, purely to constrain the computational burden. In total 200 replicates were used for TMRCA and LD calculations. Shaded areas represent the 95 % confidence interval.

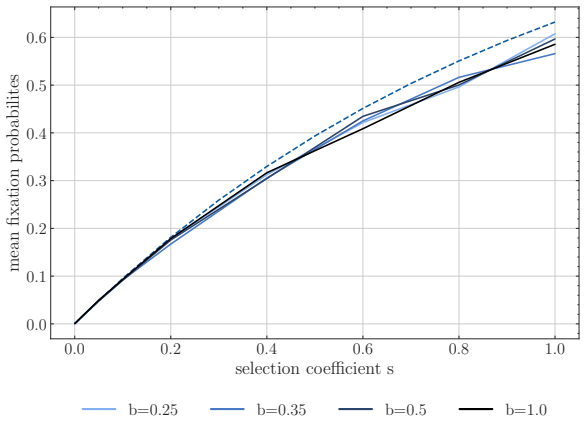
so, we first compute the time to the most recent common ancestor (TMRCA) of a coalescent tree for a sample size $n = 500$. We find that the coalescent trees are scaled by a factor $\frac{1}{b^2}$ independently of the chosen recombination rate (Figure 3.2a). The variance of the TMRCA decreases with increasing recombination due to lower linkage disequilibrium among adjacent loci, as expected under the classic Kingman coalescent with recombination (Hudson, 1983). Moreover, we also find that decreasing the value of b (*i.e.* maintaining the dormant population for longer) decreases linkage disequilibrium (Figure 3.2b). This is a direct consequence of the scaling of the recombination rate by $\frac{1}{b}$, because any active individual can undergo recombination (and can be picked as a parent with a probability b backwards in time). Therefore, we observe here two simultaneous effects of seed banks on the ARG: 1) the length of the coalescent tree and the time between coalescence events is increased by a factor $\frac{1}{b^2}$ meaning an increase in nucleotide diversity (under a given mutation parameter μ), and 2) a given lineage has a probability br to undergo an event of recombination backward in time. In other words, even if the recombination rate r is slowed down by a factor b (because only active individuals recombine), since the coalescent tree is lengthened by a factor $\frac{1}{b^2}$ there are on average $\frac{1}{b}$ more recombination events per chromosome. This property of the ARG was used in Sellinger et al., 2019 to estimate the germination parameter using the sequential Markovian coalescent approximation along the genome.

3.4.2 Allele fixation under positive selection

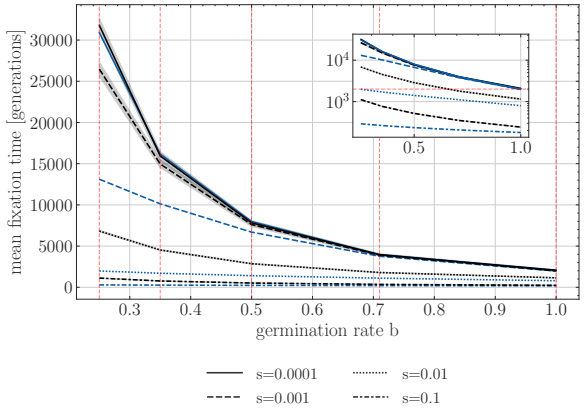
We examine the trajectory of allele frequency of neutral and beneficial mutations, by computing the probabilities and times to fixation over 1000 simulations. As expected for the case without dormancy ($b = 1$), the probability of fixation of a beneficial allele increases with the strength of selection (Figure 3.3a). We note, that the mean fixation probability is unaffected by the seed bank, as when N_e is large enough and the coefficient of selection s is not too strong, the probability of fixation of a beneficial mutation depends only on hs (Barrett et al., 2006). As expected from the neutral case, the time to fixation with dormancy becomes longer with smaller values of b (Figure 3b). When selection is weak the time to fixation is close to the expectation for neutral mutations (Figure 3.3b, $b = 1$: $4N = 2000$ generations and $b = 0.25$: $4N \times \frac{1}{b^2} = 32,000$ generations). However, increasing s changes the scaling of the time to fixation. Dormancy significantly increases the times to fixation, beyond that expected by N_e. This can be seen by comparing the expectations for the times to fixation for the rescaled effective population size without dormancy (blue lines in Figure 3.3b) to those obtained from our simulations (black lines). In order to understand this observation, we examine the time an allele under selection remains at given frequencies in the active population. The trajectory of an allele undergoing selection can be separated into three phases: two that are qualified as "stochastic", when the allele is at a very low or very high frequency, and one "deterministic", during which the frequency of the allele increases exponentially (see Kim and Stephan, 2002). As shown in Figures 3.7-3.9, we find that the proportion of time spent at very low and very high frequencies increases with increasing selection and decreasing b (it is unaffected by b when selection is weak *i.e.* $s = 0.0001$). This observation, along with generally shorter relative times spent in the deterministic phase (Figure 3.9) with increasing b, imply that the seed bank contributes to increasing the duration of the stochastic phases, slowing down the selection process.

3.4.3 Footprints of selective sweep

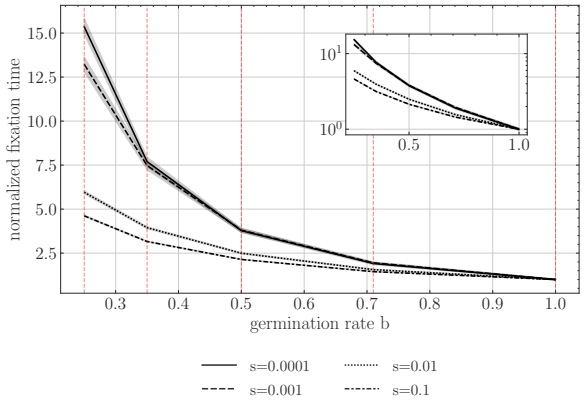
Now that we have a clearer indication of the dynamics of allele fixation, we use our new simulation tool to investigate the genomic diversity and signatures of selective sweeps at and near the locus under positive selection by simulating long portions of the genome (Figure 3.4). In accordance with the results from Figures 3.2a and 3.2b and the effects of the seed bank in maintaining genetic diversity, smaller germination rates lead to higher neutral genetic diversity due to the lengthening of the coalescent trees (e.g. Figure 3.4a measured as Tajima's π). Moreover, comparing the width of the selective sweeps valley of polymorphism in presence and absence of dormancy, we conclude that stronger dormancy generates narrower selective sweeps around sites under positive selection which have reached fixation (Figures 3.4b, 3.4d, 3.10-3.12, and 3.14-3.15). In other words, there is a narrower genomic region of hitch-hiking effect around the site under selection (Maynard Smith and Haigh, 1974b).



(a)



(b)



(c)

Figure 3.3 (a) Simulated estimates of the probability of fixation of an advantageous allele with different coefficients of selection s under absence of seed bank $b = 1$ (black solid line) and various seed bank strength $b = 0.5, 0.35, 0.25$ (blue lines) along with the theoretical expectations for a neutral allele (dashed). (b) Time to fixation for different selection coefficients. Y-axis is the time in generations, and X-axis is the germination rate b. (c) Normalized time to fixation with respect to the number of generations for $b = 1$ for each selection coefficient version of b). In b) and c) black lines represent time to fixation under seed bank. The blue lines indicate the time to fixation in a population without dormancy but with an effective population size scaled by $\frac{1}{b^2}$ and the respective scaled effective selection coefficient $N_e^b s$. For example, for $s = 0.001$, we quantify the fixation time of alleles under $N_e^{b=1.0} s = 1$, $N_e^{b=0.71} s = 1.98$, $N_e^{b=0.5} s = 4$, $N_e^{b=0.35} s = 8.2$, and $N_e^{b=0.25} s = 16$ (indicated by the red vertical dashed lines). Population size is 500 diploids, $h = 0.5$, 1,000 replicates are used for each parameter combination, and shaded areas represent the 95% confidence interval.

This is due to the re-scaling of the recombination rate as a consequence of dormancy. We note that with lower germination rates the depth of the sweeps increases in absolute diversity terms (Figure 3.4a) but not in relative diversity (Figure 3.4b), when scaling by $\frac{1}{b^2}$. However, we observe that nucleotide diversity close to the site under selection is not zero (Figure 3.4a) because of the longer times to fixation of a positive mutation and longer time for drift and new mutations to occur at neutral alleles close to the selected site. The results in Figure 3.4 reflect the manifold effect of dormancy on neutral and selected diversity as well as the recombination rate (Figures 3.2b and 3.3c). Furthermore, as recombination and selection are scaled by different functions of the germination rate, the results in Figure 3.4 cannot be produced by scaling by the expected effective population size in the absence of dormancy (Figure 3.14), since that would likewise scale the recombination rate by $\frac{1}{b^2}$, when it should be only be scaled by $\frac{1}{b}$. Scaling only by the effective population size, leads to narrower sweeps for $b = 1$ (Figure 3.14). Additionally, seed bank diversity appears to decrease visibility of the sweep when mutations are overdominant ($d = 1.1$ with $b = 0.35$, Figures 3.11-3.12) due to the increased time over which recombination can act to reduce linkage within the region (Figure 3.15b). We finally point out that while the signatures of sweeps appear smooth in Figures 3.4 and 3.15, it is because these are averaged footprints over 400 repetitions. Each simulation shows variance in both nucleotide diversity and the sweep signature, both of which condition the detectability of the sweep against the genomic background.

3.4.4 Detectability of selective sweeps

Based on the previous results, we hypothesize that, compared to the absence of seed banking, the detectability of selective sweeps in a species with seed bank is affected 1) in the genome space, that is the ability to detect the site under selection, and 2) in time, that is the ability to detect a sweep after the fixation of the beneficial allele. First, as the footprints of selective sweeps are sharper and narrower in the genome under a stronger seed bank, we expect that the detection of these sweeps likely requires adapting the different parameters of sweep detection tools, namely the window size to compute sweep statistics. Second, in a population without dormancy, the time for which the detection of a selective sweep signature is possible is approximately 0.1N generations (Kim and Stephan, 2002). We hypothesize that as the mutation rate and genetic drift are scaled by $1/b^2$, the time it takes a sweep to recover after it has reached the state of fixation is slowed down. The time window for which a sweep could still be detected would then be potentially longer than $0.1N$ generations.
In Figure 3.5 we show the results obtained using OmegaPlus and SweeD, both tools for detecting selective sweeps (Alachiotis et al., 2012; Pavlidis et al., 2013). As noted above, individual simulations show significant variation in nucleotide diversity and LD, which is not captured by the mean diversity over several runs plotted in the figures above. As the detection of sweeps is performed against the genomic background of each individual simulation, these

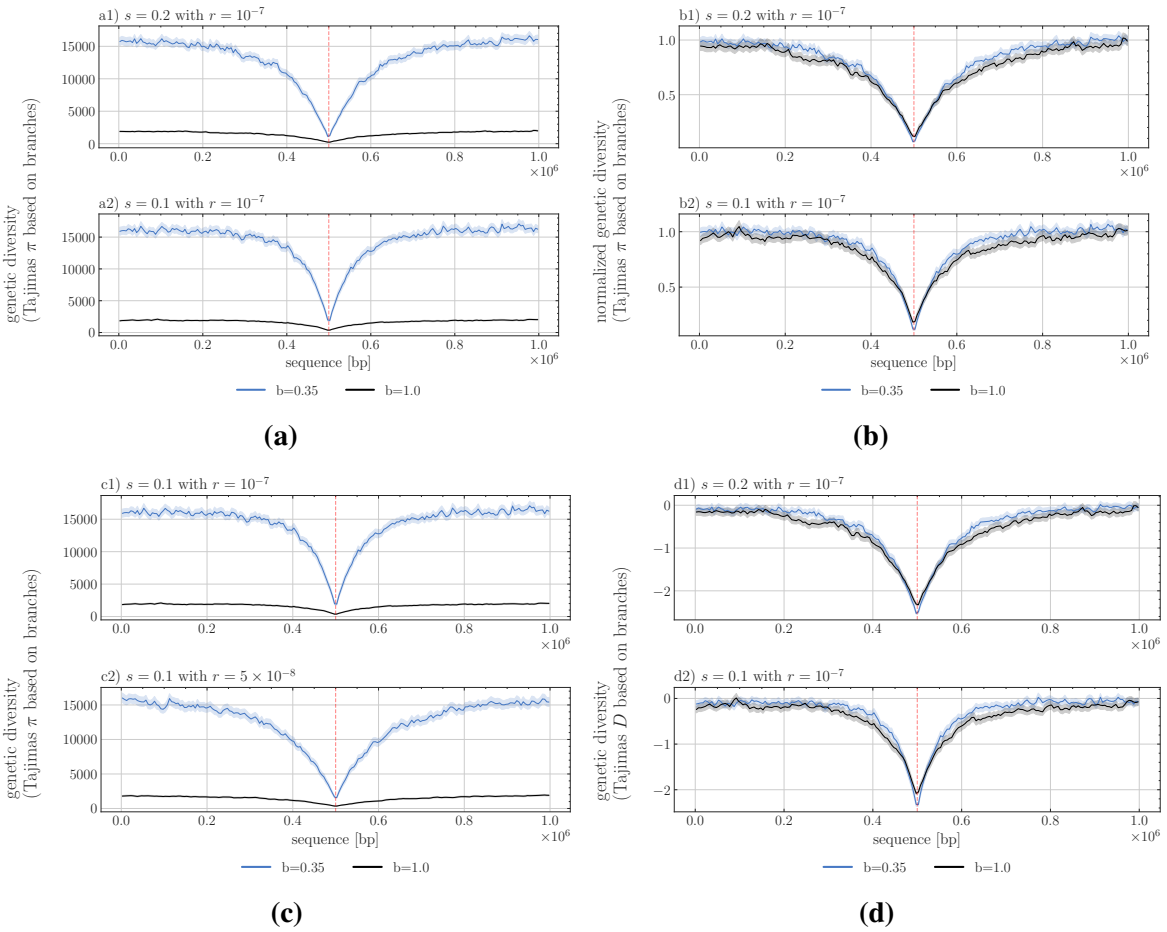


Figure 3.4 Signature of selective sweeps as measured by nucleotide diversity (Tajimas π in a, b, c) and Tajimas D (in d) over 1Mb sequence length (X-axis), the selected site being located in the middle of the segment. The statistics are computed per windows of size 5,000 bp and averaged over 200 repetitions, the shaded area representing the 95% confidence interval. The black line indicates the value without a seed bank ($b = 1$) and the blue line with dormancy ($b = 0.35$). a) π assuming two selection coefficients $N_e^{b=1}s = 200$ (a1) and $N_e^{b=1}s = 100$ (a2) with $h = 0.5$. (b) Normalized π as divided by the average neutral branch diversity, namely approx. 2000 for $b = 1$ and approx. 16000 for $b = 0.35$ (see (a) or (c) between sequence range of 0 to 0.2×10^6 or from 0.8×10^6 to 1×10^6). (c) π assuming two recombination rates $r = 10^{-7}$ per bp per generation (c1) and $r = 5 \times 10^{-8}$ per bp per generation (c2).

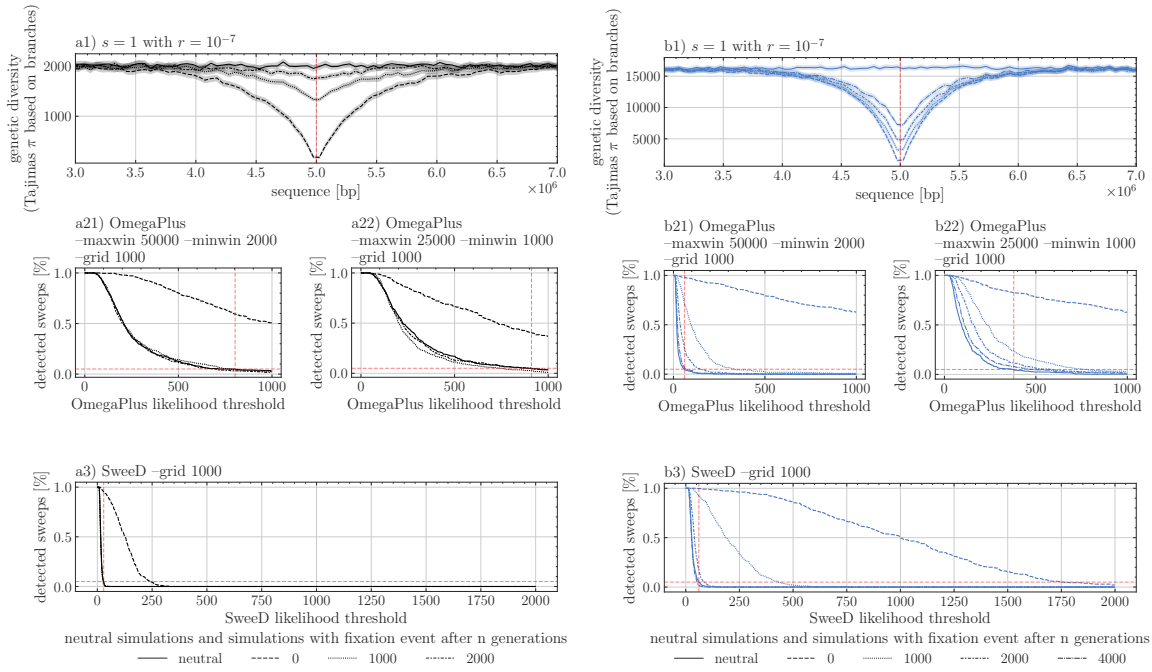
variations in nucleotide diversity and LD generate confounding effects and define the rates of false positives expected from the detection test.

Following the classic procedure to detect sweeps, we use neutral simulations to define different thresholds for detection, for which we obtain a false positive rate of less than 0.05. We find that when using the same large detection window "–minwin 2000 –maxwin 50000" for $b = 1$ and $b = 0.35$ (Figures 3.5 a21 and 3.5 b21), sweep detection almost completely fails for $b = 1$, unless the fixation has just occurred, meaning that no generation has passed since the fixation event. For $b = 0.35$ sweeps are detectable up to >2000 generations after fixation. Decreasing the window size is generally associated with a loss of sensitivity, increasing the rate of true and false positives. This is true for $b = 1$ (see neutral threshold line in Figure 3.5 b21 and b22), indicating a decrease from roughly 60 % detected sweeps to 40 % (after 400 repetitions). However, the detectability of older sweeps (>2,000 generations) is increased for $b = 0.35$ (Figure 3.5 b22). Results using SweeD support this increased detectability, also when using the SFS statistics, showing the possibility of locating sweeps approximately up to 2,000 generations after fixation (Figure 3.5 a3 and b3)

We note that there is a much sharper decrease in the rate of detection of false positive sweeps (neutral simulation line in Figure 3.5) under seed bank compared to the absence of a seed bank, likely being a direct consequence of the increased linkage decay around the site. Lastly, the possibility to locate sweeps multiple generations after the fixation event emphasizes the slower recovery of nucleotide diversity post-fixation in combination with the already established narrowness of the signature in the presence of a seed bank for a given population size N ($b = 0.35$, Figure 3.10).

3.5 Discussion

We investigate the neutral and selective genome-wide characteristics of a weak seed bank model by means of a newly developed simulator. We first characterize the emergent behavior of an adaptive allele under a weak seed bank model, and simulate the times to and probabilities of fixation, considering different strengths of selection and recombination. In populations without seed banks, a neutral mutation is expected to fix after a time of $2N_e$ generations and $\approx 2N_e s$ if the allele is under weak selection (Kimura, 1962). Though both processes are re-scaled by the weak dormancy model (Koopmann et al., 2017), the time to fixation of a neutral mutation can be obtained by rescaling N_e appropriately ($N_e = \frac{N}{b^2}$ in the case of a seed bank, with b the germination rate). This remains true under weak selection, however under strong selection the time to fixation is significantly decreased and cannot be explained by the change in N_e alone. In accordance with existing theory, the probability of fixation is unaffected by the seed bank (since it depends only on sh, see for example Barrett et al., 2006), implying that the main effect of seed banks is on the dynamics of allelic frequencies, but not on the outcome of selection at a single locus. Combining this observation and the effect



(a) **(b)**

Figure 3.5 Selective sweep detection depending on the threshold of OmegaPlus or SweeD statistics on a 10MB sequence with a strong selective mutation of $N_e^{b=1}s = 1,000$ located in the middle of the sequence. Two germination rates apply: a1) $b = 1$ and b1) $b = 0.35$, with the signature of sweep being shown at various time points after the fixation event (0, 1000, 2000 and 4000 generations). Results for two window sizes "–minwin 2000 –maxwin 50000" (a21,b21) and "–minwin 1000 –maxwin 25000" (a22,b22) for analysis with OmegaPlus and SweeD (a3 and b3) using a grid size of 1,000. The percentage of detected sweeps is indicated for a given user-defined threshold value on the X-axis. Vertical dashed lines indicate the 5% sweep detection based on neutral simulations, setting up the false positive rate. Recombination rate is $r = 1 \times 10^{-7}$ per bp per generation for all sweep simulations, and 400 replicates for each parameter.

of seed banks on increasing the effective recombination rate, we suggest that the signatures of sweeps may be slightly easier to detect in the presence of seed banking as shown by the sharpness and depth of the nucleotide diversity pattern (the so-called valley of polymorphism due to genetic hitch-hiking, Maynard Smith and Haigh, 1974b; Kim and Stephan, 2002) against the genomic background.

3.5.1 Dynamics of alleles under positive selection

Our results regarding the time to fixation of advantageous alleles are in line with previous works in showing that a weak seed bank delays the time to fixation (Hairston Jr and De Stasio Jr, 1988; Koopmann et al., 2017; Heinrich et al., 2018; Shoemaker and Lennon, 2018). However, a novelty here is that we refine these results in showing that the time to fixation of a weakly ($s < 0.01$) and a strongly ($s \geq 0.01$) positively selected allele differ under seed bank: the selection on weak alleles is delayed by a factor $\frac{1}{b^2}$ while for strong selection, the time to fixation is delayed by more than would be expected for a population without a seed bank but the same effective population size(see Figure 3.3b,3.3c, and Koopmann et al. 2017 for an analytical approach with an infinite deterministic seed bank). We show that this delay can be explained by an increase in the time spent in the stochastic phases of allele fixation (at below 10% and above 90% in the active population). In other words, dormancy delays the action of selection under the weak seed bank model (due to the dormant population acting as a buffer slowing down allele frequency change). In the initial phase of selection when the advantageous allele is at a very low frequency in the (active) population, and before reaching the exponential phase, the allele frequency increases almost deterministicly (Kim and Stephan, 2002). This delay in the initial selection phase is visible in Figure 4a in Shoemaker and Lennon, 2018. Our results are valid for the weak seed bank model (as studied in Figure 4a in Shoemaker and Lennon, 2018, and Koopmann et al., 2017) and we find that there exists a unique phase of selection encompassing the time until all individuals (in the active and dormant population) have fixed the advantageous allele. Strong seed bank models behave differently with respect to time to fixation of alleles under selection (Shoemaker and Lennon, 2018), showing two distinct phases: a first rapid phase of selection in the active population, followed by a second long delay until there is fixation in the dormant population. We are not aware of any results regarding the effect of strong seed banking on the probability of allele fixation. Our results thus mitigate the previous claim that (weak) seed banks may amplify selection, making it relatively more efficient with regards to the effects of genetic drift, while it does not alter the probability of fixation of an advantageous allele. Longer times to fixation should promote genetic diversity, but as the probability of fixation at a single locus is unchanged by the seed bank, dormancy does not necessarily enhance the adaptive potential (by positive selection) of a population.

3.5.2 Signals of selective sweeps

The precise signature of a positive selective sweep is dependent on a variety of factors, *i.e.* age of the observation after fixation, degree of linkage due to recombination, and its detectability depends on the specified window size to compute polymorphism statistics. However, in the case of sweeps under seed bank, two effects are at play and change the classic expectations based on the hitch-hiking model without generation overlap. First, as the effective population size under seed bank increases with smaller values of b, an excess of new mutations is expected to occur after fixation around the site under selection compared to the absence of seed bank. As these new mutations are singleton SNPs, we suggest that the signature of selective sweeps observed in the site-frequency spectrum (U-shaped SFS) should be detectable under seed bank (Maynard Smith and Haigh, 1974b; Kim and Stephan, 2002). Additionally, this effect was also detectable by the other sweep detection methods based on the SFS (SweeD, Pavlidis et al., 2013), finding sweeps older than 2000 generations (for N=500).
Second, the signature of sweeps also depends on the distribution of linkage disequilibrium (LD) around the site under selection (Alachiotis et al., 2012; Bisschop et al., 2021b), which is affected by the seed bank (Figure 3.4). Theoretically, it has been shown that patterns of LD both on either side and across the selected site generally provide good predictive power to detect the allele under selection. We use this property when using OmegaPlus, which relies on LD patterns across sites. Further past demography should be accounted to correct for false positives, due for example to bottlenecks (see review in Stephan, 2019c). We speculate that a high effective recombination rate around the site under selection, as a consequence of the seed bank, maybe an advantage when detecting sweeps. This allows the avoidance of confounding effects due to the SFS shape, which is sensitive to demographic history. We also highlight that the narrower shape of the selective sweep under stronger seed bank, and the smaller number of loci contained in the window, reduce the number of false positives.
As mentioned above, a crucial parameter to detect sweeps is the window length to compute the statistics that the various methods rely on. The optimal window size depends on the neutral background diversity around the site of interest, which is a consequence not only of the rate of recombination but also the scaled rate of neutral mutations. We choose a constant mutation rate over time, and make the assumption of mutations being introduced during the dormant phase at this constant rate (see equations in introduction). This simplifying assumption is partially supported by empirical evidence (Levin, 1990; Whittle, 2006; Dann et al., 2017), and has so far been made in the wider field of inference models, notably in the ecological sequential Markovian coalescent method (eSMC, Sellinger et al., 2019). While assuming mutation in the dormant population favors the inference of footprints of selection by simply adding additional data, which subsequently increases the likelihood to observe recombination events, it remains unclear if this assumption is justified for all species with a dormant phase and/or if mutations occur at a different rate depending on the age of the dormant population.

More research on the rate of mutation and stability of DNA during dormant phases is needed in plant (*e.g.*Waterworth et al., 2016), fungi and invertebrate species. Nevertheless, even if this mutation rate in seeds is relatively low, our results of a stronger signal of selection under seed banking than in populations without seed banking are still valid. In contrast to the weak seed bank model, it is possible to test for the existence of mutations during the dormant stage under a strong seed bank model as assumed in prokaryotes, because of the much longer dormant phase compared to the coalescence times (Blath et al., 2020).
Finally, as for all sweep models, we show that selective events that are too far back in the past cannot be detected under seed banks. Nonetheless, we show that when there is a seed bank, older sweeps can be detected with increasing accuracy. The presence of a long persistent seed bank could therefore be convenient when studying older adaptation events in plants, fungi and invertebrates that have some form of dormancy. This prediction also agrees with the previous observation that the footprint of older demographic events is stored in the seed bank (predicted in Živković and Tellier, 2012, observed theoretically in Sellinger et al., 2019, and empirically observed in *Daphnia* in Möst et al., 2015). Our results open avenues for further testing the correlation between past demographic events and selective events for species that present this life-history strategy. However, current methods estimating the age of selective sweeps (Tournebize et al., 2019; Bisschop et al., 2021b) would need to use an *ad hoc* simulator (*e.g.* such as the one we present here) to generate neutral and selected simulations under seed banking.

3.5.3 Strengths and limitations of the simulation method

The simulation program developed and used in this work, written in C++, is centered on the use of *tskit*. The toolkit allows for the efficient storage of genealogies through time, by removing lineages that have effectively gone extinct in the current population, thus simplifying the genealogy at regular intervals during the program run-time. Despite all our efforts to streamline the process, forward simulations are inherently limited, because each generation has to be produced sequentially. Thus, while being more flexible and intuitively easier to understand than their coalescent counterparts, forward simulations sacrifice computational efficiency in terms of memory and speed. While simulating hundreds or thousands of individuals is possible (also storing their genealogies in a reasonable amount of time), this limitation becomes exaggerated when adding genomic phenomena such as recombination, and even more so when considering ecological characteristics such as seed banking. The latter scales the process of finding the most recent common ancestor by an inverse factor of b^2. As this leads to an increase in run-time of the order of $O(1/b^2)$, we kept the population size at 500 (hermaphroditic) diploid individuals. Furthermore, the output format of the simulations are tree sequences, which enables downstream processing and data analysis without the elaborate design of highly specific code. We believe that our code is the first to allow simulations of

long stretches of DNA under the seed bank model including recombination and selection. In a previous study, we developed a modified version of the neutral coalescent simulator *scrm* (Staab et al., 2015) which includes a seed bank with recombination (Sellinger et al., 2019). Our current simulator can be used to study the effect and signatures of selection along the genome under dormancy for non-model species with reasonably small population sizes. For a strict application of our model to diploid plants, future work would need to consider the constraint of having only N individual diploid parents to choose from. We expect this to likely yield slightly shorter coalescent times than in our pseudo-diploid model (based on the haploid Kaj et al., 2001), while our insights should still be valid.

3.5.4 Towards more complete scenarios of selection

We here explore a scenario in which a single beneficial allele is introduced. The much longer times to fixation in the presence of seed banks suggest that such a scenario may be unlikely. Indeed, it is probable that several alleles under selection, potentially affecting the same biological processes, are maintained simultaneously in populations for longer periods of time. We can therefore surmise that under seed banking, polygenic selective processes and/or competing selective sweeps, often associated with complex phenotypes and adaptation to changing environmental conditions in space and time, should be common.

From the point of view of genomic signatures of selection, the overall effectiveness of selection at a locus coupled with increased effective recombination with seed banking generate narrower selective sweeps, hence less genetic hitch-hiking throughout the genome. While we show that these effects can be advantageous to detect selective sweeps, we speculate that this might not be the case for balancing selection. If seed banks do promote balancing selection (Tellier and Brown, 2009), the expected genomic footprints would be likely narrowly located around the site under selection, and the excess of nucleotide diversity would not be significantly different from the rest of the genome. The presence of seed banking would therefore obscure the signatures of balancing selection. Concomitantly, the Hill-Robertson-Effect and background selection are expected to be weaker under longer seed banks. These predictions could ultimately define the relationship between linkage disequilibrium, the efficacy of selection and observed nucleotide diversity in species with seed banks compared to species without it (Tellier, 2019, Živković and Tellier, 2018).

Acknowledgements

The authors gratefully acknowledge the computational and data resources provided by the Leibniz Supercomputing Centre (www.lrz.de). KK is supported by a grant from the Deutsche Forschungsgemeinschaft (DFG) through the TUM International Graduate School of Science and Engineering (IGSSE), GSC 81, within the project GENOMIE QADOP. AT receives fund-

ing from the Deutsche Forschungsgemeinschaft (DFG) grant TE809/1-4, project 254587930. DAA was a Humboldt Post-Doctoral fellow. A preprint version of this article has been peer-reviewed and recommended by PCIEvolBiol (https://doi.org/10.24072/pci.evolbiol.100552).

Conflict of interest disclosure

The authors declare that they have no financial conflict of interest with the content of this article.

3.6 Supplementary Information

The following part contains the supplementary information, including left-out tables and figures to further describe statistics and signatures of the weak dormancy model.

3.7 Absolute TMRCAs for different germination and recombination rates

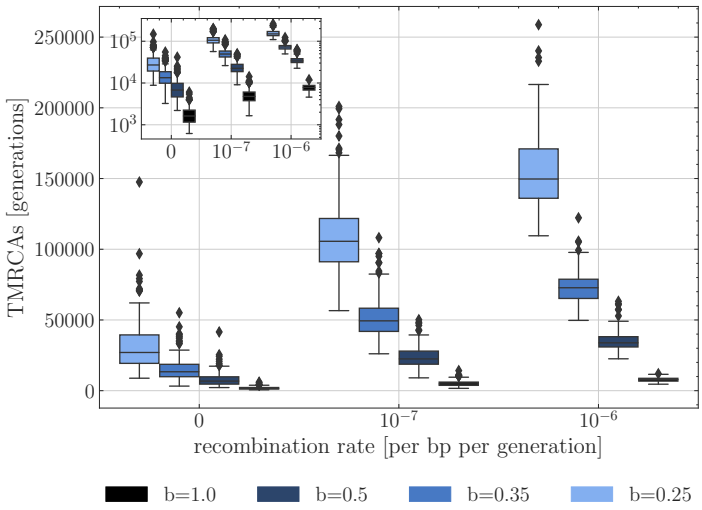
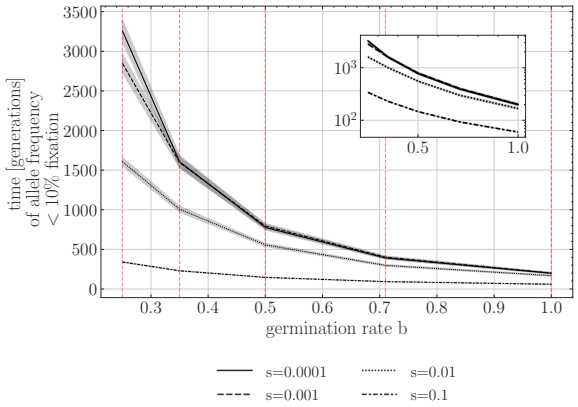


Figure 3.6 Absolute time to the most recent common ancestor (TMRCA) as a function of the germination rate b. For each germination rate, three recombination rates per site are presented ($r = 0$, $r = 10^{-7}$ and $r = 10^{-6}$. Boxes describe the 25th (Q1) to 75th percentile (Q3), with the lower whisker representing Q1-1.5×(Q3-Q1) outlier threshold and the upper whisker is calculated analogously. The mean is plotted between Q3 and Q1. Each boxplot represents the distribution of 200 TMRCA values over 200 sequences of 0.1 Mb. Per sequence the oldest TMRCA is retained.

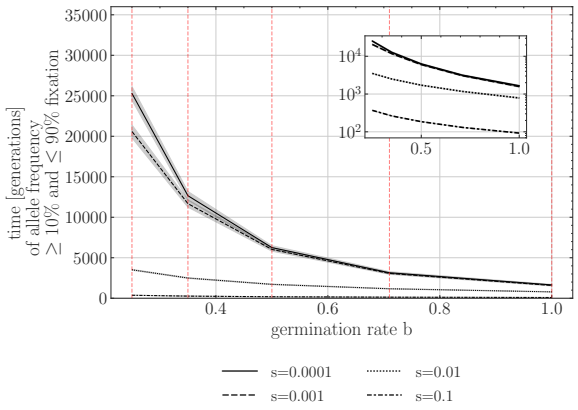
Table 3.1 Proposed number of generations to simulate before adding selective mutation (for 2N=1000). Also, the number of generations simulated to estimate TMRCA.

germination rate (b)	recombination rate (r)	calibration generations
1	0	40000
0.5	0	80000
0.35	0	160000
0.25	0	320000
1	10^{-8}	48000
0.5	10^{-8}	96000
0.35	10^{-8}	192000
0.25	10^{-8}	384000
1	10^{-7}	56000
0.5	10^{-7}	112000
0.35	10^{-7}	224000
0.25	10^{-7}	448000
1	10^{-6}	64000
0.5	10^{-6}	128000
0.35	10^{-6}	256000
0.25	10^{-6}	512000

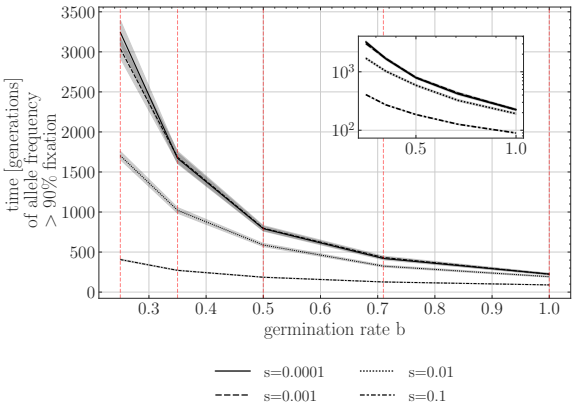
3.8 Fixation time phase contribution



(a)

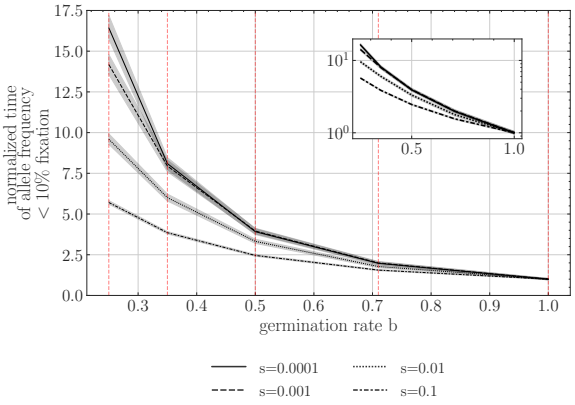


(b)

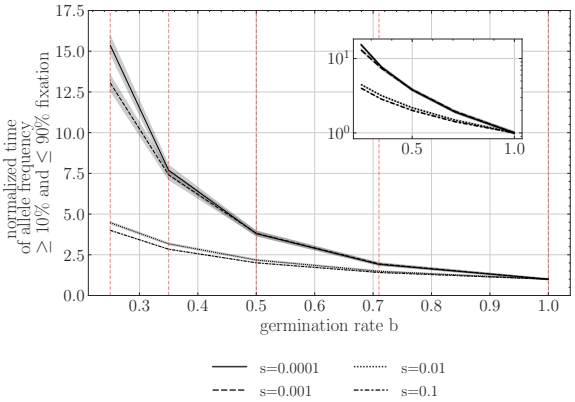


(c)

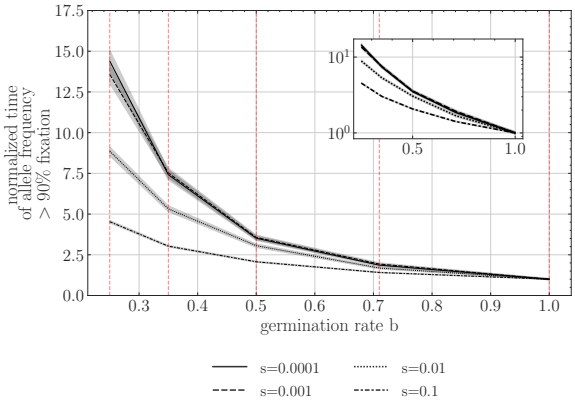
Figure 3.7 Time to fixation for different selection coefficients. Y-axis is the unnormalized time in generations, and X-axis is the germination rate b. a) Time allele spends below 10% frequecy in population and b) time allele spends above 10 % and below 90% frequency c) above 90% in population. A dormancy effective population size coefficient $N_e^b s$ can be calculated with each intersection of vertical dashed lines with fixation times by scaling with b^2, e.g. for $N_e^{b=1.0}s = 1$: $N_e^{b=0.71}s = 2.0$, $N_e^{b=0.5}s = 4$, $N_e^{b=0.35}s = 8.2$, $N_e^{b=0.25}s = 16$. In total 1000 replicates were used for each parameter configurations



(a)

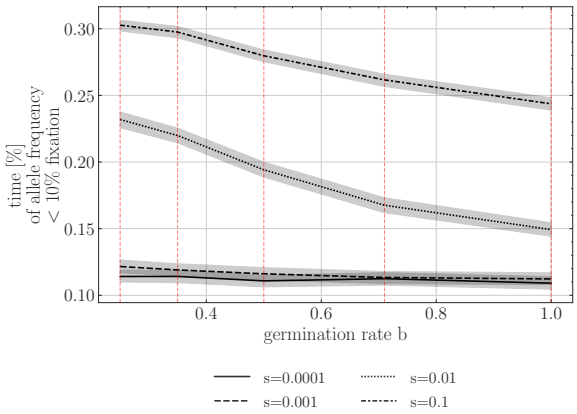


(b)

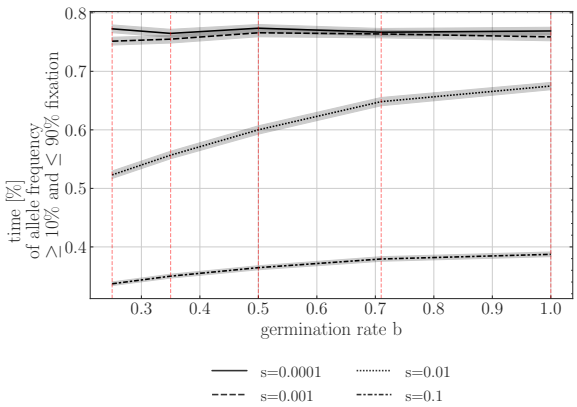


(c)

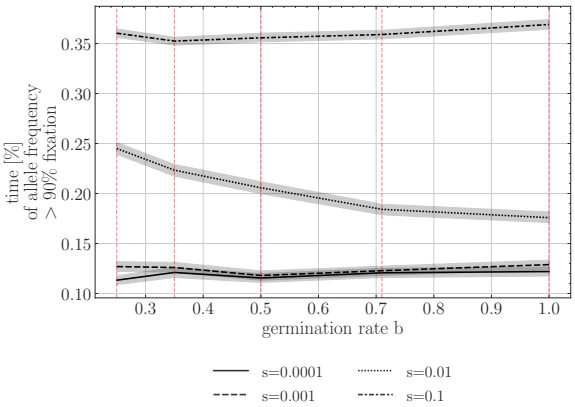
Figure 3.8 Time to fixation for different selection coefficients. Y-axis is the time normalized by the $b = 1$ estimate for each respective selection coefficient, and X-axis is the germination rate b. a) Time allele spends below 10% frequency , b) time allele spends above or equal to 10 % and below or equal to 90% frequency and c) above 90% frequency in population. A dormancy effective population size coefficient $N_e^b s$ can be calculated with each intersection of vertical dashed lines with fixation times by scaling with b^2, e.g. for $N_e^{b=1.0} s = 1$: $N_e^{b=0.71} s = 2.0$, $N_e^{b=0.5} s = 4$, $N_e^{b=0.35} s = 8.2$, $N_e^{b=0.25} s = 16$. In total 1000 replicates were used for each parameter configurations



(a)



(b)



(c)

Figure 3.9 Time contribution to fixation for different selection coefficients in percent of the phases (a) below 10 % allele frequency and (b) above 10 % and below 90% allele frequency and (c) above 90% allele frequency. In total 1000 replicates were used for each parameter configurations.

3.9 Sweep recovery signatures after fixation

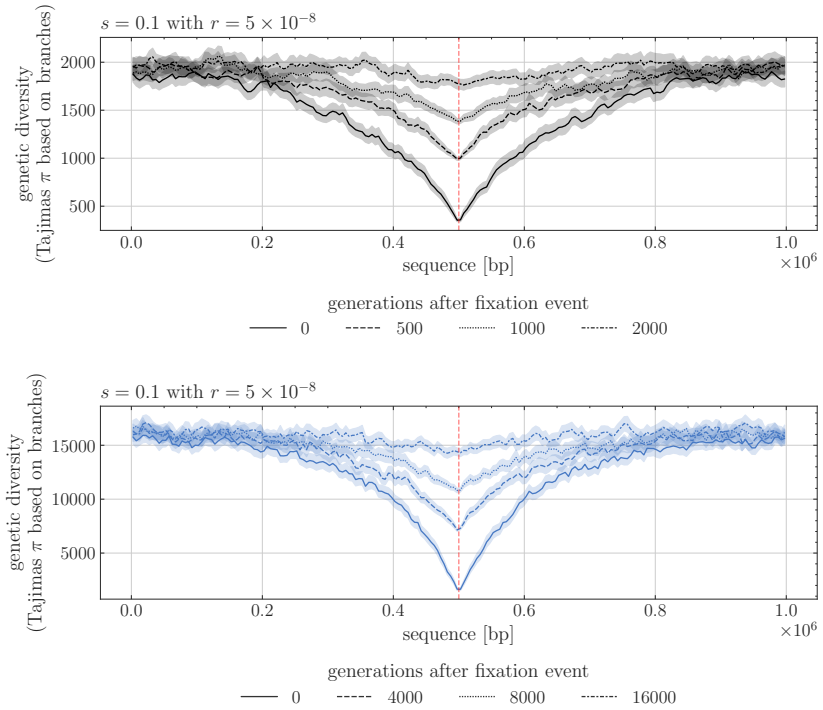


Figure 3.10 Nucleotide diversity (Tajima's π, Y-axis) over sequence length (X-axis) for windows of size 5000 mapping length and averaged over 400 repetitions. Comparison between germination rate a) $b = 1$ and b) $b = 0.35$ for different sweep recovery times. A selection coefficient of $N_e^{b=1}s = 100.0$ and recombination rate $r = 5 \times 10^{-8}$ per generations per bp was set for all simulations.

3.10 Effect of different dominance coefficients

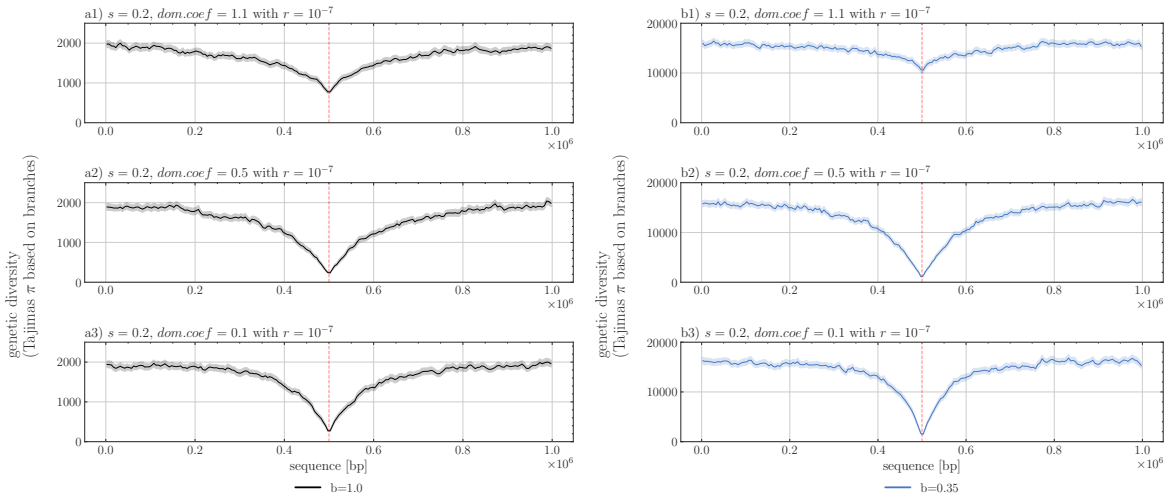


Figure 3.11 Nucleotide diversity (Tajima's π, Y-axis) over sequence length (X-axis) for windows of size 5000 and averaged over 200 repetitions. Comparison between germination rate a) $b = 1$ and b) $b = 0.35$ for different dominance coefficients a1, b1) $h = 1.1$, a2,b2) $h = 0.5$, a3, b3) $h = 0.1$, respectively. A selection coefficient of $N_e^{b=1} s = 200$ and recombination rate $r = 5 \times 10^{-7}$ per bp per generation was set for all simulations.

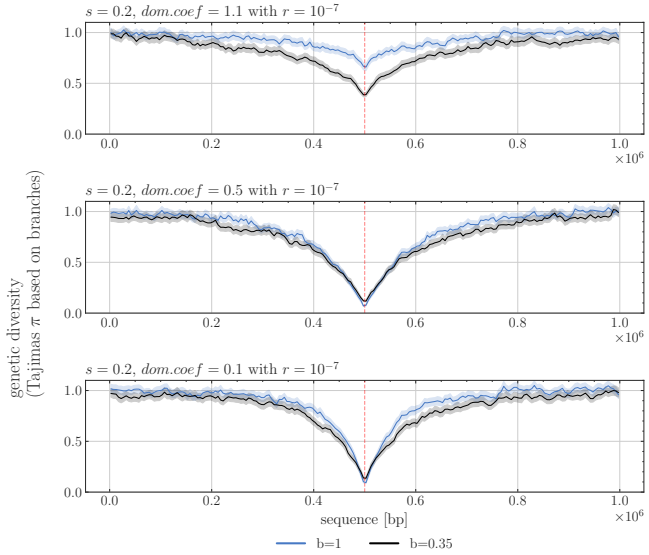
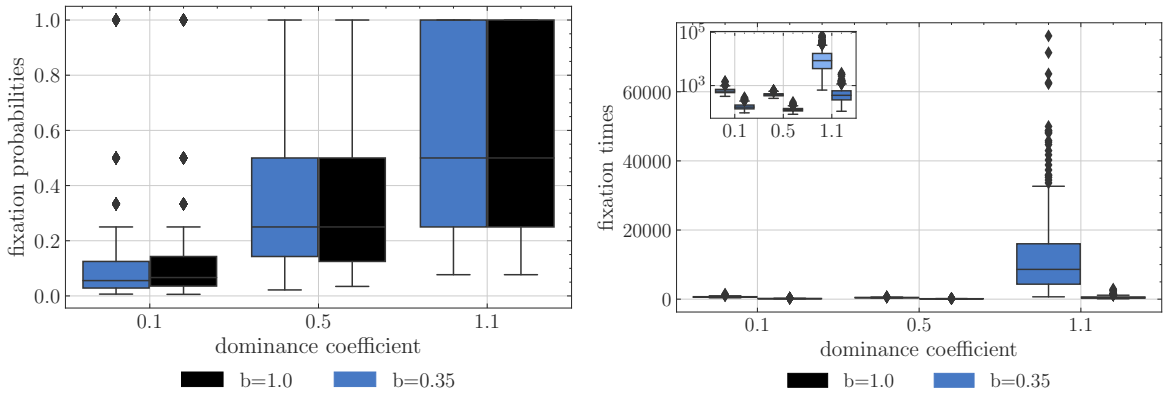


Figure 3.12 Normalized nucleotide diversity (Tajima's π, Y-axis) over sequence length (X-axis) for windows of size 5000 and averaged over 200 repetitions. Comparison between germination rate $b = 1$ and $b = 0.35$ for different dominance coefficients $h = 1.1$, $h = 0.5$, $h = 0.1$, respectively. A selection coefficient of $N_e^{b=1} s = 200$ and recombination rate $r = 5 \times 10^{-7}$ per bp per generation was set for all simulations.



(a) (b)

Figure 3.13 Fixation probability (a) and time (b) for different dominance coefficients and two different germination rates, namely $b = 1$ and $b = 0.35$ for 400 replicates. Selection coefficient was set to 0.2, corresponding to $N_e^{b=1}s = 200$ and $N_e^{b=0.35}s = 1632.7$. In total 1000 replicates were used for each parameter configurations and simulations were conditioned on fixation.

3.11 Scaling population size by $\frac{1}{b^2}$

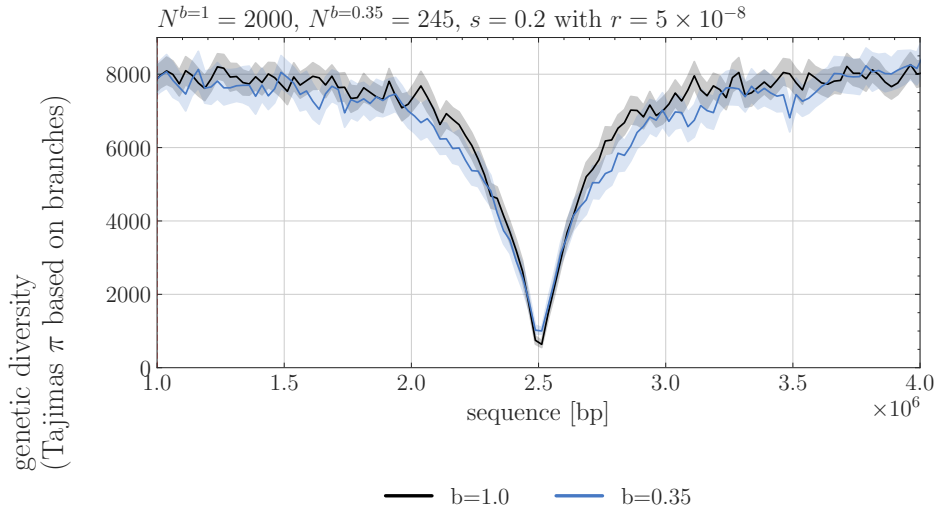
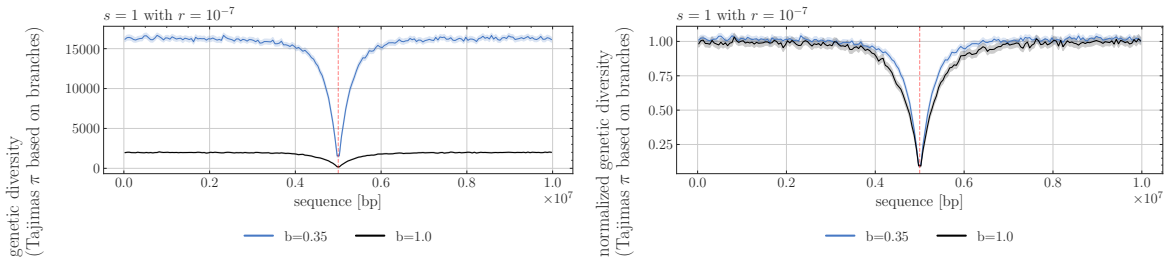


Figure 3.14 Tajima's π, Y-axis) over sequence length of 5 mb (X-axis) for windows of size 25000 and averaged over 150 repetitions. Comparison between germination rate $b = 1$ and $b = 0.35$ under a selection coefficient of $s = 0.2$, corresponding to $N_e^{b=1}s = 400$ without a seed bank (black) and to $N_e^{b=0.35}s = 400$ (blue), assuming population sizes of 2000 and 245 diploid individuals, for no seed bank and seed bank, respectively. A recombination rate of $r = 5 \times 10^{-8}$ per bp per generation was set for all simulations.

3.12 Narrow sweep signature of

a large sequence lengths



(a) **(b)**

Figure 3.15 Nucleotide diversity (a) Tajimas π and (b) normalized diversity Y-axis) over sequence length of 10 mb (X-axis) for windows of size 50000 and averaged over 400 repetitions, the shaded area represents a 95% confidence interval. Comparison between germination rate $b = 1$ (black) and $b = 0.35$ (blue) for selection coefficients $s = 1$ corresponding to $N_e^{b=1}s = 1000$ and $N_e^{b=0.35}s = 8163.3$ with a dominance coefficient of $h = 0.5$ and recombination rate of $r = 10^{-7}$ per bp and generation.

4 Determination of Rapid Adaptation in Species with Large Offspring Variance

The following chapter has been published as:
Kevin Korfmann, Marie Temple-Boyer, Thibaut Sellinger and Aurélien Tellier (2023). Determinants of rapid adaptation in species with large variance in offspring production. Molecular Ecology, 00, 1– 14. https://doi.org/10.1111/mec.16982

KK re-implemented and extended the simulator in Julia, based on an original Pythonic implementation of M. Temple-Boyer. All simulations and figures were contributed by KK, while additionally working on the manuscript and revisions.

4.1 Abstract

The speed of population adaptation to changing biotic and abiotic environments is determined by the interaction between genetic drift, positive selection and linkage effects. Many marine species (fish, crustaceans), invertebrates and pathogens of humans and crops, exhibit sweepstakes reproduction characterized by the production of a very large amount of offspring (fecundity phase) from which only a small fraction may survive to the next generation (viability phase). Using stochastic simulations, we investigate whether the occurrence of sweepstakes reproduction affects the efficiency of a positively selected unlinked locus and thus the speed of adaptation since fecundity and/or viability have distinguishable consequences on mutation rate, probability and fixation time of advantageous alleles. We observe that the mean number of mutations at the next generation is always function of the population size, but the variance increases with stronger sweepstakes reproduction when mutations occur in the parents. On the one hand, stronger sweepstakes reproduction magnifies the effect of genetic drift thus increasing the probability of fixation of neutral allele, and decreasing that of selected alleles. On the other hand, the time to fixation of advantageous (as well as neutral) alleles is shortened by stronger sweepstakes reproduction. Importantly, fecundity and viability selection exhibit different probabilities and times to fixation of advantageous alleles

under intermediate and weak sweepstakes reproduction. Finally, alleles under both strong fecundity and viability selection display a synergistic efficiency of selection. We conclude that measuring and modelling accurately fecundity and/or viability selection is crucial to predict the adaptive potential of species with sweepstakes reproduction.

4.2 Introduction

Starting from the seminal work of Darwin, evolution and adaptation of species/populations to their environment through a particular phenotype or trait have been traditionally assumed to be occurring at a slow pace and be beyond the direct observation of evolutionary biologists. However, numerous counter-examples recently challenge this view, and demonstrate that adaptation to environmental change can be fast, that is occurring over few generations (Zhou et al., 2019). To name but a few, rapid adaptation is observed in natural settings during colonization of novel habitats (Losos and Ricklefs, 2009; Hu et al., 2019), but also in response to human activities: in plants (Anderson et al., 2012) and fish (Crotti et al., 2021) responding to anthropogenic changes and destruction of habitats, in insects colonizing new urban habitats (Diamond et al., 2022), in bacteria resisting antibiotics (Barbosa et al., 2021), in fungi resistant to fungicides (Fisher et al., 2022), or in crop pathogens overcoming new plant resistance (Persoons et al., 2017). We define here rapid adaptation as occurring over few tens (up to hundred) generations thus including the effect of generation time.

With advances in sequencing technology, it becomes feasible to dissect the genetic architecture of the trait underpinning rapid adaptation, namely to decipher whether one or few genes with major effects or many loci with small effects are involved. In the former case, so-called selective sweeps occur at the given genes by selection of an advantageous allele (Maynard Smith and Haigh, 1974a; Stephan, 2019b), while in the latter case, there is so-called polygenic selection driven by simultaneous changes in allele frequencies across many loci (Barghi et al., 2020; Jain and Stephan, 2017). Arguably, these both types of selection processes and models represent two extreme in a continuum of possible genetic architectures (number of genes involved, epistatic and pleiotropic interactions) and distribution of selection coefficients across loci (Barghi et al., 2020; Jain and Stephan, 2017; Stephan, 2016). Based on the current climate and global change accompanied by loss of biodiversity, as well as the importance of adaptive mechanisms for medicine and agriculture, there is a tremendous interest in measuring the speed of adaptation of species and uncovering the underlying genetic (coding genes, gene expression change, gene duplication, genome rearrangement, transposable elements,..) or epigenetic (RNA silencing, methylation,. . .) mechanisms. However, we suggest here that for evolutionary genomics of rapid adaptation to move into a predictive science, that is to be able to assess the adaptive potential of species and predict how and when adaptation may occur, it is important to account for the variability in life cycles and

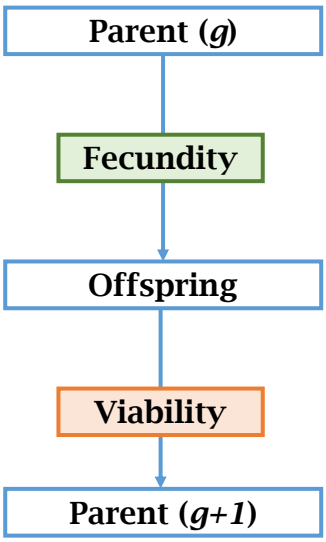
specificity of the life history traits of each species. Indeed, a potential pitfall lies in the relative inadequacy of classic population genetics theory to describe the diversity of life cycles and life history traits found in nature (Ellegren and Galtier, 2016). To date, analyses and predictions of adaptation (rapid or not) are mostly biased by the use of mathematical models with simplifying assumptions (*e.g.* assuming a Wright-Fisher model) which are violated in many species of bacteria, viruses, fish, invertebrates, fungi or plants which exhibit clonal reproduction, large variance in offspring production (sweepstakes reproduction) and dormancy/quiescence (Ellegren and Galtier, 2016; Tellier and Lemaire, 2014; Tellier, 2019; Lennon et al., 2021a; Sabin et al., 2022; Eldon, 2020a).

In this study we chose to focus on adaptation provided by one locus with a significant positive fitness effect. As a result this allele may spread and become fixed in the population, generating a so-called selective sweep at the locus under selection (Maynard Smith and Haigh, 1974a; Stephan, 2016, 2019b). The speed of rapid adaptation depends on three distinct processes (Charlesworth and Charlesworth, 2010; Charlesworth, 2020): 1) the probability that a given advantageous allele appears by mutation at this locus, 2) the probability of fixation of this allele, and 3) the time to fixation of that advantageous allele. First, an advantageous allele needs to appear by random mutation. This process is quantified by the population mutation rate $\theta = 4N_e\mu$, where N_e is the inbreeding effective population size and μ the mutation rate (per site or per locus). In general, the time for new mutations to appear is only small enough to play an important role in rapid adaptation when both the mutation rate and population size are high such as in viruses, bacteria or fungi (especially crop pathogens, Stam and McDonald, 2018). Conversely, in most animal or plant species, mutation rates are too small to promote new mutations over few generations, and it suffices to analyze rapid adaptation based on standing variation (Eldon and Stephan, 2018, 2023. Second, a new advantageous allele (mutant) has a probability to reach fixation (P_{fix}) or to be lost, as genetic drift may counter-act the effect of positive selection. To assess the efficiency of selection, the probability of fixation of an advantageous allele should be compared to the probability of fixation of a neutral allele. Third, if the advantageous allele reaches fixation, it can do so more or less rapidly (measured in generations). This is termed as the time to fixation (T_{fix}), and to measure the effectiveness of selection it ought to be compared to that of a neutral allele. While we focus, for simplicity, on rapid adaptation due to positive selection at one locus, note that our predictions are affected in the genomic context by the effect of linked selection at neighbouring sites.

As mentioned above, population genetics theory is built on a mathematical framework which models these three processes and is historically based on the so-called Wright-Fisher (WF) model of population evolution (see description in textbooks such as Charlesworth and Charlesworth, 2010). The Wright-Fisher model was rapidly extended to continuous-time

diffusion and coalescent models (*e.g.* Charlesworth and Charlesworth, 2010). In its simplest version, the WF neutral model of evolution assumes a simplified life-cycle in which each of the N haploid parents at generation g produce more than enough offspring (an infinite number), in order for N of them to constitute the next reproducing generation (g+1). There is no overlap between generations, and offspring choose their parent at random from generation g following a binomial sampling. An emerging property resulting from this random sampling scheme is that parents exhibit a distribution of offspring number which is well approximated by a Poisson distribution with mean (and variance) equals to one (Charlesworth and Charlesworth, 2010). In other words, the variance in offspring production is small as most parents produce zero, one or two offspring which become the next reproducing generation (g+1).

To explicit the limitations of the WF model, we draw in Figure 4.1 a simplified life-cycle starting from parents at generation g producing offspring regrouping various developmental stages: plants produce seeds germinating into seedlings, fish but also invertebrates such as nematodes, crustaceans or insects produce eggs developing into juveniles, fungi produce various forms of spores, and bacteria can produce dormant/quiescent spores. The production of offspring constitutes the fecundity phase. These produced offspring hatch, germinate and/or grow to form potential parents which are mature for reproduction, and constitute the next generation (g+1). The survival of the produced offspring is termed the viability phase of the cycle. At each phase, life history traits determine the rate of genetic drift, as well as mutational and selective processes, which can potentially result in different expectations from the Wright-Fisher model for θ, P_{fix}, and T_{fix}. First, some species present a very large fecundity meaning that parents can produce a number of offspring/developmental stages much larger than the population size of adults (number of offsprings >> number of parents N) which can generate a large variance in offspring production (and/or survival) between parents (*e.g.* Árnason et al., 2023; Arnason and Halldorsdottir, 2015; Sabin et al., 2022; Menardo et al., 2020; Hedgecock and Pudovkin, 2011; Eldon and Wakeley, 2006; Eldon et al., 2015; Eldon and Stephan, 2023 and reviewed in Tellier and Lemaire, 2014; Eldon, 2020a) resulting in a neutral sweepstakes reproduction life-cycle. The neutral sweepstakes reproduction life-cycle can be qualified as a specific case of "boom-and-bust" population dynamics within a generation.



Processes		
Genetic drift	Mutation	Selection
Variance in offspring production among parents	Mutation in parent and/or mutation in offspring	Fecundity selection = mean production of of offspring per parent
Life span of offspring (dormancy, quiescence)	Mutation during the life span of offspring	Viability selection = mean viability of offspring

Figure 4.1 Schematic view of species life-cycle. The cycle has two phases: fecundity and production of developmental stages (green), and viability (orange) with the growth and development to become a reproducing parent at generation g+1. Each phase has neutral, mutational and selective processes associated.

Furthermore, under the explicit life-cycle of Figure 4.1, defining the mutational process can be important as new mutations can occur in 1) the parental germ lines and are inherited by all produced offspring, or 2) in the offspring. Additionally, fecundity selection occurs if a parent with an advantageous allele produces on average a larger number of offspring than other genotypes (Figure 4.1). Second, once the offspring are produced, their viability and life span can exhibit large variance within and between population/species. Dormancy is a common life history trait of plants (seeds), invertebrates (eggs) or fungi and bacteria (also called quiescence or covert infections for human diseases) which increases the life span of the offspring, generating overlap of generations (Charlesworth et al., 1994; Charlesworth and Charlesworth, 2010; Kaj et al., 2001; Lennon et al., 2021a; Tellier, 2019). Defining the mutational process is also here non-trivial as mutations may or may not occur in the dormant stage (Tellier, 2019; Lennon et al., 2021a). Viability selection occurs when the survival of offspring carrying an advantageous allele increases compared to other genotypes (Fig. 4.1). When adding mutations in the population model, the fecundity and viability phases are not distinguished nor distinguishable under the assumptions of the WF model. As a result, for an advantageous allele, the probability of fixation and the time to fixation are identical for fecundity and viability selection (He et al., 2017).

Recent studies have chiefly investigated the effect of sweepstakes reproduction on the polymorphism of neutral alleles with the aim to infer the strength of the skew in offspring

production (Schweinsberg, 2003; Eldon and Wakeley, 2006; Eldon et al., 2015; Koskela, 2018; Koskela and Berenguer, 2019; Freund, 2020; Vendrami et al., 2021; Árnason et al., 2023; Sackman et al., 2019; Irwin et al., 2016; Korfmann et al., 2022b), while the influence of sweepstakes reproduction on selected alleles is not yet fully understood. It is shown in (Der et al., 2011) that selection should act more deterministically under sweepstakes reproduction than under the WF model though under a peculiar model of reproduction (the Dirac model). Further, the time to fixation of an advantageous allele is much faster than expected under the WF model under a sweepstakes reproduction model with viability selection (Eldon and Stephan, 2018). A complementary analysis of the effect of sweepstakes reproduction with viability selection including the effect of allele dominance and random demographic changes is concomitantly available to the present study (Eldon and Stephan, 2023). These studies suggest that strong sweepstakes reproduction may, on the one hand, increase the effectiveness of positive selection and speed up adaptation by shortening the time to fixation of an advantageous allele (T_{fix} conditioned on fixation). On the other hand, when observing all allele trajectories (Supplementary Results in Eldon and Stephan, 2018, and Eldon and Stephan, 2023), there is a high chance that advantageous alleles get lost under sweepstakes reproduction. This means that the efficiency of viability selection, that is the probability of fixation (P_{fix}), may be decreased under sweepstakes reproduction. These studies mostly consider constant population size (but see Eldon and Stephan, 2023) and constant selection in time, while varying population size and fluctuating selection are shown to be also key determinants of the fate of advantageous alleles. Indeed, such non-constant conditions affect the outcome of the interaction between genetic drift and selection under the WF model (Devi and Jain, 2020; Kaushik and Jain, 2021).

Our study is unique in disentangling the effect of viability and fecundity effect on genetic drift (sweepstakes reproduction under boom and bust cycle) and selection. We first investigate the effect of sweepstakes reproduction on the three components of speed of adaptation θ, P_{fix}, and T_{fix} by comparing neutral and advantageous alleles and considering fecundity and/or viability selection under constant and varying population size as well as under fluctuating selection. Second we also investigate the effect of joint viability and fecundity selection on P_{fix}, and T_{fix}. We rely on the Cannings Model (Cannings, 1974) which allows us to model the life cycle depicted in Figure 4.1. We use simple analytical derivations to provide intuitions on the main results, and then use stochastic forward-in-time simulations, to assess the effectiveness and efficiency of positive selection under a wide range of biologically plausible life cycles and scenarios. We conclude on the importance of taking into account life-cycles, including neutral sweepstakes, when predicting the potential of species for rapid adaptation in medicine or agriculture.

4.3 Materials and Methods

4.3.1 Model Description

With the idea to focus on sweepstakes reproduction in invertebrates, fungi, viruses or bacteria, we consider a population of haploid individuals of constant size N evolving under a Cannings model (Cannings, 1974). We furthermore assume in our population the existence a bi-allelic locus (with alleles A or a). We begin by setting up the definitions of the neutral Cannings model following the model of Schweinsberg (Schweinsberg, 2003) in which allele A and a have an identical fitness value.

Definition 1 (Neutral Cannings model of reproduction) *Let $(X_i)_{1\leq i\leq N}$ be identical independent $\mathbb{N}$-valued random variables such that $\mathbb{E}(X_1) > 1$.*
For each $1 \leq i \leq N$, the number of offspring of the i^{th} parental individual is X_i. The N surviving offspring are then drawn without replacement amongst the $X_1 + ... + X_N$ offspring ($\mathbb{N}$ is the set of natural numbers).

The hypothesis $\mathbb{E}(X_1) > 1$ guarantees there will be more than N offspring generated at each generations assuring at least N surviving offspring when N goes to infinity as required in mathematical models (Schweinsberg, 2003; Eldon and Stephan, 2018, 2023). As we wish to investigate sweepstakes reproduction, we assume throughout this manuscript that the number of offspring produced by each parents to be independent and identically distributed as in Schweinsberg, 2003. Hence the distribution of the number of produced offspring is shaped by a unique parameter α (dropping the scaling constant C). More precisely, the distribution of the number of produced offspring by a parent i (*i.e.* X_i), given the parameter α, is:

$$\begin{aligned} P(X_i = 0) &= 0 \\ P(X_i \geq k) &= \frac{1}{k^\alpha} \text{with } k \geq 1 \end{aligned} \tag{4.1}$$

Let us define now a Cannings model with selection with the allele of type A exhibiting a selection advantage with coefficient s. Under fecundity selection individuals with allele A produce on average more offspring than a individuals.

Definition 2 (Reproduction with fecundity selection) *Let $(X_i)_{1\leq i\leq N}$ be identical independent $\mathbb{N}$-valued random variables such that $\mathbb{E}(X_i) > 1$. Let s be the selection coefficient of the type A. Let $(Y_i)_{1\leq i\leq N}$ be identical independent $\mathbb{N}$-valued random variable such that*

$\mathbb{E}(Y_i) > 1$ *and* $\mathbb{E}(Y_i) = (1+s)\mathbb{E}(X_i)$.
For each $1 \leq i \leq N$*, the number of offspring of the* i^{th} *parental individual is* X_i *if the* i^{th} *individual has the type a, and* Y_i *otherwise.*
The N surviving offspring are then drawn without replacement amongst all produced offspring.

We first assume that parental individuals with allele A produces on average $1+s$ more (or less is if $s < 0$) offspring than parents with allele a. Yet each produced offspring have the same probability to be drawn for the next generation (viability). We define an advantageous allele for fecundity, as an allele that increases on average the number of produced offspring by a factor s ($s > 0$).

In other words the expected number of produced offspring by parent with the allele A is $(1+s)$ times of the number of offspring produced by parents with allele a. Additionally, there is no limit to the total number of offspring produced. This definition, referred to as the model F1 for fecundity selection, is our fecundity selection default model. Our F1 fecundity model is implemented by applying the *inverse transform sampling* method (see Appendix and equation 4.1), multiplying the number of offspring produced by the selection component $(1+s)$ and flooring the resulting number of type A offspring for each generation.

However, we also suggest that alleles might affect the fecundity in a different way than in the model F1 (*i.e.* scaling the average of the offspring distribution for individuals carrying allele A). We wish to investigate an alternative fecundity selection model F2, in which the distribution of produced offspring by parents with allele A is possibly more skewed (towards high values) than that of parents with allele a. This effect takes place in addition to producing on average $(1+s)$ times the expected number of offspring produced by parents with allele a. Such model might be a more realistic model for selection in species already displaying skewed offspring distribution. In this alternative fecundity selection scenario (F2), the advantageous (A) allele increases on average the number of produced offspring by a factor s as well compared to individuals with allele a (as in F1). However, individuals with the allele a produce offspring with parameter of the Cannings model α_a whereas the individuals with the allele A produce their offspring with a modified α_A (where $\alpha_A < \alpha_a$). Hence in our fecundity F2 model, α_A is obtained by the following formula:

$$\alpha_A = \alpha_a/(1+s) \tag{4.2}$$

Definition 3 (Reproduction with viability selection) *Let* $(X_i)_{1\leq i\leq N}$ *be identical independent* $\mathbb{N}$*-valued random variables such that* $\mathbb{E}(X_i) > 1$*. Let* $s \geq 0$ *be the selection coefficient of the type A.*
For each $1 \leq i \leq N$*, the number of offspring of the* i^{th} *individual is* X_i*.*

The N surviving offspring are the draw according to the Wallenius non central hypergeometric distribution where the weight of an offspring is $w = 1 + s$ *if it has the type A, and* $w = 1$ *otherwise.*

Here, viability selection increases the probability of offspring with allele A to be drawn to constitute the next generation $g+1$ (similar to Eldon and Stephan, 2018). In the Supplementary Material, we present a cursory analysis of our discrete time Cannings model with and without selection. We provide some useful general expressions for the expectation and variance of the number of offspring (of allele A) at a given generation. These analytical results generate the insight that fecundity and viability selection do differ in their mean and variance of the number of offspring produced under our sweepstakes reproduction model. However, this formalism does not allow to compute further analytical results because the precise distribution of offspring would need to be specified. We therefore use simulations to generate quantitative results on the comparison of neutral and selection models.

4.3.2 Stochastic simulations

Simulating offspring production

The neutral and selection Cannings models are implemented in $Julia$. As defined above (equation 4.1), under neutrality we assume the number of offspring produced by each parents to be independent and identically distributed as defined in (Schweinsberg, 2003). The main parameter α determines the distribution of offspring produced (Schweinsberg, 2003), with α close to one meaning a large variance in offspring production with some individuals producing extremely large number of offspring (on the order of N). When α is close to two, there is a small variance (of one) in offspring production between parents. Note that as described in (Eldon and Stephan, 2018, 2023), due to scaling of large offspring production under the Cannings model, a value of $\alpha = 2$ does not generate exactly a WF model. In order to obtain the WF model, we specify in our Cannings model a Poisson distribution of offspring (with parameter λ). Further details of the implementation are found in the Supplementary Material. Implementations and scripts to run simulations can be found at **https://github.com/kevinkorfmann/CanningsSimulator**.

Simulating selection and demography

Each α parameter setting is coupled with viability selection and/or fecundity selection model under the influence of weak and strong selection coefficient, namely $s = [0, 0.01, 0.1]$ ($s = 0$ being the neutral case). We use population sizes N of 500, 1000, 5000, so that the effective selection coefficients are $Ns = [[0, 0, 0,], [5, 10, 50], [50, 100, 500]]$, respectively.

We follow previous work to model time dependent sinus functions with varying amplitudes and periods for the population size (Devi and Jain, 2020) or selection coefficient (Kaushik and Jain, 2021) (see Supplementary Material, Figs. SI1 and SI2). As environmental conditions change, so can the fitness provided by alleles. To model this fluctuating effect, we now assume population size to be constant, but the selective advantage provided by the allele A to be changing through time. To simplify our approach (as in Kaushik and Jain, 2021), we assume the average value of the selection coefficient to be 0 over a period of fluctuating selection, but the amplitude of the selective coefficient s ranges up to 0.01 or 0.1 (*i.e.* maximum fitness provided by the allele A).

We test the effects on the probability of fixation and time to fixation regarding the initial introduction phase of the allele (initially advantageous or deleterious) and the speed at which the selection coefficient changes through time (slow period ~ 1,000 generations or fast period ~ 100 generations, see Supplementary Material and Figure 4.8).

We analyze the allele fixation probabilities and fixation times under different scenarios: constant population size, fluctuating population size, and fluctuating selection. The obtained results are based on 5×10^5 simulations to compute the probability of fixation, while the time to fixation simulations are conditioned on fixation. Thus, simulations are run until 5,000 fixation events have occurred. Simulations start with time independent constant population size and selection coefficient. Finally, to produce simulations under the WF model, we use a Poisson distribution with $\lambda = 1.2$ (based on empirical simulation). We perform simulations demonstrating that we recover the known expectations for probability of fixation and time to fixation of neutral and advantageous alleles under the WF model (Table 4.1). Furthermore, the probabilities and times to fixation are found to be identical for viability and fecundity selection (Table 4.1) as expected under the WF model (He et al., 2017).

4.3.3 Mutation Process

Mutations may occur during the life cycle of the parent (in the cells of the germ line) and can be heritable, so that all offspring produced by this parent share these mutations. Mutations can occur during the gamete production and are thus offspring specific, with independence between mutations. Hence the total number of occurring mutations M in the produced offspring at generation g can be decomposed as:

$$M = L \times (\mu_p \times N_p + \mu_o \times N_o) \tag{4.3}$$

Where L is the sequence length (or number of loci), μ_p is the mutation rate per generation per site/locus during the life cycle of the parent, N_p the number of parents (in our case,

$N_p = N$), μ_o the mutation rate per generation per site/locus during gamete production, and N_o the number of produced offspring.

We simulate two different models. In the first model, parents mutate at rate 0.01 (*i.e.* $\mu_p = 0.01$ and $\mu_o = 0$), while in the second model the offspring mutate (and not the parents) individually and independently under the same mutation rate (*i.e.* $\mu_p = 0$ and $\mu_o = 0.01$). We then simulate mutations under the two models and compare the observe diversity, that is the number of A alleles produced in one generation. Each replicate consists of a population of size N=10,000 (starting with all individuals of type a allele) and only one generation is simulated. The number of offspring produced is simulated under the distribution described above (equation 4.1) for different values of the α parameter. We then randomly sample among the produced offspring, N_p surviving individuals to assess the diversity at the next generation (g+1).

4.4 Results

4.4.1 Distribution of new mutations

We first focus on the distribution of new alleles obtained by mutations under a parental or offspring mutational process and as function of the α parameter. As expected from the equation 4.3, the average number of new alleles are very similar for the parental and offspring mutational model and all values of α, with a mean of 100 (as $N_p \times \mu_p = N_o \times \mu_o = 10000 \times 0.01 = 100$, Figure 4.2). However, if mutations occur in the parents the variance in the number of new mutations across simulation replicates is higher than if mutations occur in offspring. Furthermore, the variance across replicates in the number of new alleles increases with diminishing α when mutations occur in parents, while α has a negligible effect on the variance of number of new alleles when mutations occur in the offspring (Figure 4.2). These first results indicate that in a population of fixed size N, species with mutation in parents or in offspring would exhibit on average similar speed of adaptation based on the average rate of appearance of novel advantageous mutations, but species with mutations in offspring are more likely to consistently produce this number at each generation.

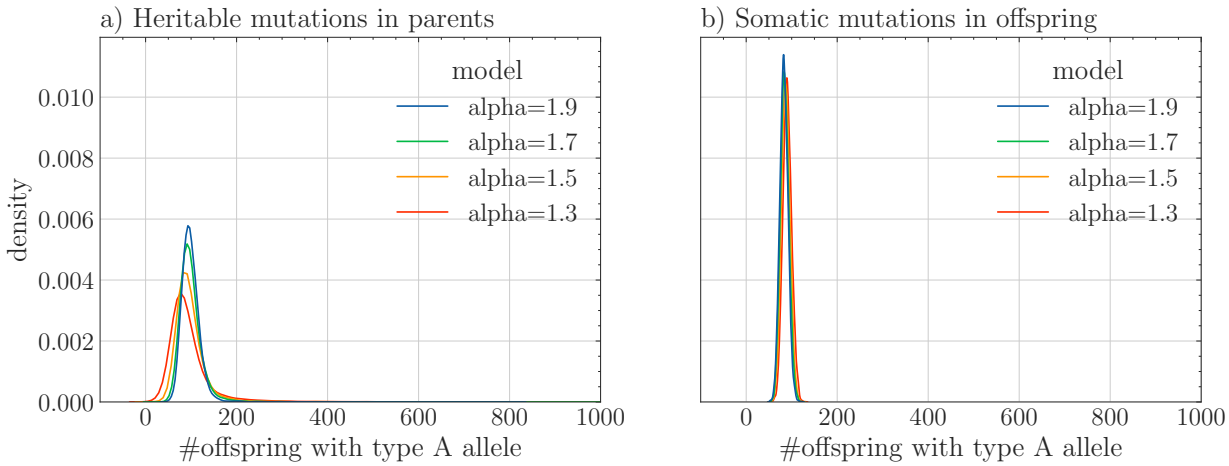


Figure 4.2 Density distribution of new mutants (individuals with type A allele) after one generation for two different mutation models. Density distribution of individuals with type A allele after one generation under our Cannings model of offspring distribution with α measures of 1.3, 1.5, 1.7, 1.9 for two different mutation models. In a) parents mutate, and the mutation is heritable for all offspring of a given parent, in b) mutations occur in each offspring individually. Population size is $N_p = N_o = 10^4$ and mutation rate $\mu_p = \mu_o = 0.01$. The density estimation is the result of 5×10^3 repetitions per α value.

4.4.2 Constant population size and constant selection

Viability and F1 fecundity selection

We first estimate the probability for alleles to reach fixation for different α values under constant population size (Figure 4.3 with $N = 1,000$, $N = 500$ in Figure 4.9 and $N = 5,000$ in Figure 4.10). The fixation probability of a neutral allele, *i.e.* $s = 0$ in Figure 4.3a, diminishes from 0.001 to 0.0002 with increasing α values. Hence, neutral stronger sweepstakes reproduction increases the probability of a neutral allele to become fixed. As expected from classic theoretical results under the WF model, the probability of fixation of a neutral allele diminishes with increasing population size (compare Figure 4.3a, d to Figure 4.9 and 4.9). Furthermore, viability and fecundity have a similar effect on the probability of fixation (advantageous alleles for viability in Figure 4.3a, or for fecundity, Figure 4.3d). As could be expected, under weaker selection ($s = 0.01$), the probability of fixation of the advantageous allele is similar to that under neutrality, while stronger selection ($s = 0.1$) increases the fixation probability (Figure 4.3a, d). Note that in contrast to the neutral allele, stronger sweepstakes reproduction decreases the probability of fixation of an advantageous allele (from ca. 0.005 for $\alpha = 1.1$ to 0.02 for $\alpha = 2$). This means that that sweepstakes reproduction is equivalent to stronger genetic drift which can counter-act the action of selection, thus decreasing the probability of advantageous allele to be fixed by up to a factor four. Under strong sweepstakes reproduction (low α values), viability and fecundity advantageous alleles exhibit ultimately similar probabilities of fixation as neutral alleles (Figure 4.3c, f). To better assess the dissimilarity between viability and fecundity selection, we compute the difference of fixation

probability between allele providing a fecundity and viability advantage (Figure 4.3c). When an allele provides no or only a small advantage, the difference is equal to zero for all α values and population sizes (see Figure 4.9 and 4.10). However under strong selection for α smaller than 1.7, fecundity selection yields higher probabilities of fixation than viability selection, a trend which reverses for higher α values (Figure 4.3c).

We provide estimates of the fixation time (in generation) of a neutral or advantageous allele (see Figure 4.3b, e). As expected from the classic theory under the WF, the time to fixation decreases with the strength of selection and with population size (Figure 4.9b,e and 4.10b,e). As reported before (Eldon and Stephan, 2018), the time to fixation of advantageous alleles increases with increasing α values (the effect being more apparent for large population size $N = 5,000$ in Figure 4.10b,e). Yet under neutrality ($s = 0$ in Figure 4.3e), the time to fixation (and its variance) increases exponentially with α, while under strong selection the time to fixation increases almost linearly with α. Overall, neutral alleles tend to reach fixation as fast as advantageous alleles under strong sweepstakes reproduction (low α values), *i.e.* confirming the stronger effect of genetic drift with diminishing α (note that the neutral case is not shown in Eldon and Stephan, 2018). We assess the difference in time to fixation between fecundity or viability selection (Figure 4.3f) and observe, as expected, no difference under neutrality. Under selection, the time to fixation of an allele providing fecundity advantage is shorter than an allele providing viability advantage, while this difference tends to zero when α tends to one.

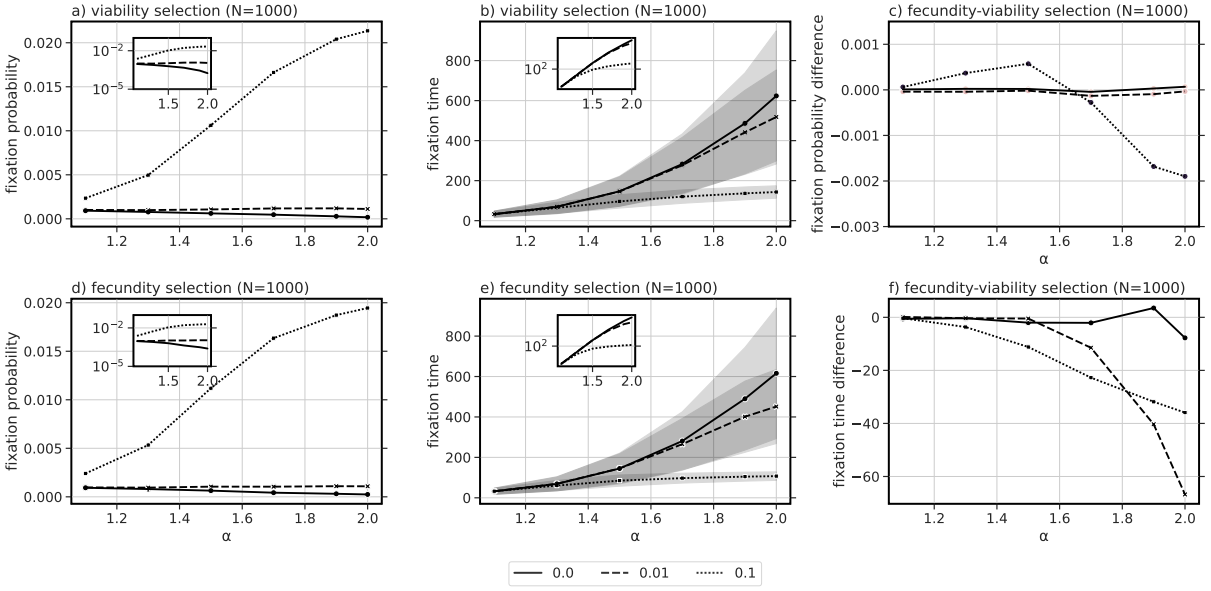


Figure 4.3 Allele fixation probability and average time to fixation of alleles under constant population size. Fixation probability (a, d) and average fixation times (b, e) under different sweepstakes strength (α ranging from 1.1 to 2.0) for (a, b) viability selection and (d, e) fecundity selection (F1). Three selection coefficients are shown ($s = 0; 0.01; 0.1$) and $N = 1,000$. Probabilities are obtained from 5×10^5 simulations, while fixation times are estimated based on 5×10^3 simulations conditioned on fixation. Shaded areas correspond to 95% confidence intervals. Panels (c) and (f) present the difference between the probability of fixation (c) and time to fixation (f) for fecundity compared to viability selection.

Joint fecundity and viability selection

We now generalize the previous results by considering that the advantageous allele affects simultaneously fecundity (F1) with weak $s = 0.01$ or strong $s = 0.1$ coefficient and viability with weak $s = 0.01$ or strong $s = 0.1$ effect. As above, for an allele with strong effect for both types of selection (*i.e.* $s = 0.1$), the probability of allele fixation increases with increasing α. However, the probability of fixation under joint strong selection (blue line in Figure 4.4a) can become greater than the additive effect of viability and fecundity as measured by summing up the probability for strong viability and strong fecundity obtained separately (solid black line in Figure 4.4a). This synergistic effect is also observed for the time to fixation when comparing the time to fixation under joint strong selection with the sum of each strong selection measured independently (blue line versus solid black line in Figure 4.4b). When both selection coefficients are smaller ($s = 0.01$) or sweepstakes reproduction is strong (small α), this synergistic effect diminishes so the probability of fixation and time to fixation are close to the weak combined effect (see Figure 4.4). These results hold for all tested population sizes (Figure 4.11 and Figure 4.12). We therefore demonstrate that as we

decouple fecundity and viability selection in our model of offspring production, a synergistic effect appears for simultaneous strong fecundity and strong viability selection.

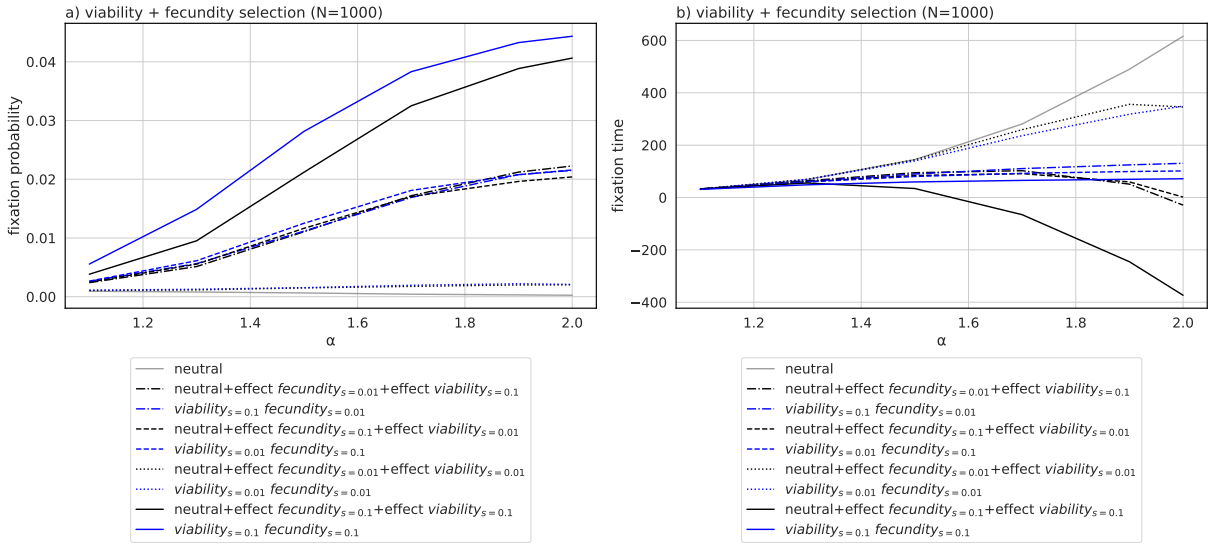


Figure 4.4 Allele fixation probability and average fixation time under constant population size and additive selection models. a) Probability of fixation under simultaneous viability and fecundity selection (F1, blue) in comparison to neutral fixation probability with added effect (net contribution) of fecundity (F1) and viability selection, when simulated individually (black). b) Average time to fixation of alleles under simultaneous selection models (blue) in comparison to neutral estimates of time to fixation summed up with the effect (net contribution) of each selection type (black). Probabilities are obtained from 5×10^5 simulations, while fixation times are estimated based on 5×10^3 simulations conditioned on fixation.

Fecundity selection (F2) model

We then estimate the probability (Figure 4.5a) and time to fixation (Figure 4.5b) of neutral and advantageous alleles under the fecundity selection F2 model and compare with previous results under F1 fecundity selection. The probability of fixation under model F2 is at least five times greater than that under model F1 (Figure 4.5a), and the the time of fixation is at least twice as fast than under F1 (Figure 4.5b).

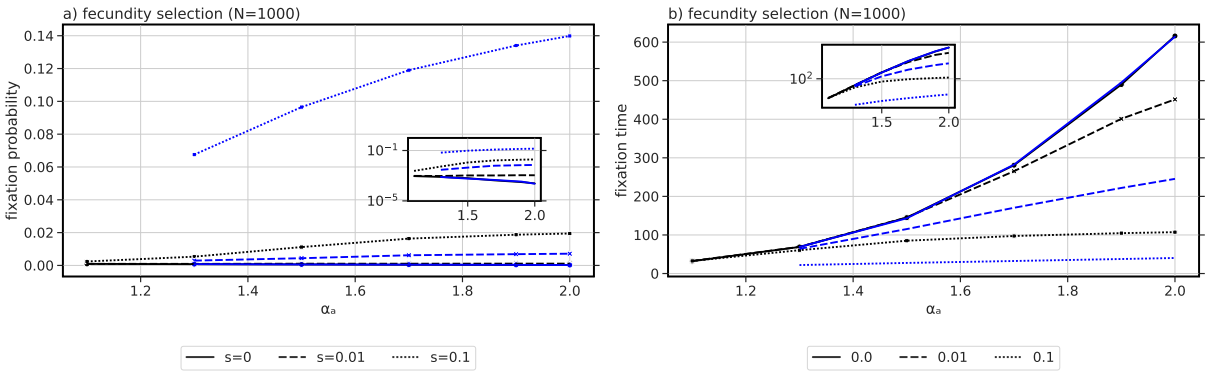


Figure 4.5 Allele fixation probability and average fixation time under constant population size and two constant selection fecundity models. a) Probability of fixation and b) fixation time for two positive selection fecundity models. In black fecundity is modelled by increasing the offspring number (F1) , and in blue the fecundity F2 selection. Probabilities are obtained from 5×10^5 simulations, and fixation times are estimated based on 5×10^3 simulations.

4.4.3 Fluctuating conditions

Fluctuating population sizes

Under fluctuating population sizes (Figure 4.7), the probability of fixation of an allele under fecundity F1 or under viability selection is similar as that under constant population size (Figure 4.13 and 4.14). When comparing two different demographic scenarios, namely with different amplitude and period of oscillations, the probability of fixation only slightly vary. Similarly, the times to fixation of an advantageous allele with fecundity or viability selection under fluctuating population sizes, are slightly shorter and with higher variance than those obtained under constant population size (Figure 4.15 and 4.16). This confirms the intuition and previous theoretical results under the WF model that varying population size around the mean N does increase genetic drift and decrease the time to fixation of alleles. Overall, the difference in time to fixation and in fixation probability between the two selection models (fecundity F1 and viability) also follow those under constant populations size (Figure 4.17, 4.18, 4.19 and 4.20).

Fluctuating selection

Under a slow variation of s through time, the probability of fixation of the A allele depends strongly on whether the initial fitness of that allele is beneficial or deleterious (Figure 4.6). Comparing these two situations, when the A allele is initially advantageous, the fixation probability is higher (less chance for the allele to be lost) and time to fixation is smaller (Figure 4.21, 4.22, 4.23). Furthermore, we observe no noticeable difference between the probability of fixation under viability or fecundity selection. The probability of fixation is increasing with α (for all population sizes Figure 4.24 and 4.25) when the allele is initially

highly beneficial, otherwise in other cases, the probability of fixation is decreasing with α when we consider slow fluctuating selection (Figure 4.6a,c). A new result emerges when varying the speed of the fluctuating fitness (Figure 4.6b, d). When the speed of fitness variation increases (fast fluctuations), the probability of fixation becomes non-monotonic and is maximized at intermediate values of α (at high amplitude of fitness).

For small amplitude of fitness variation, the results are similar to slow variation of fitness (Figure 4.6). If the allele is initially deleterious (disadvantageous, light curves Figure 4.6a-d), the probability to loose this allele is higher when the amplitude of fitness variation increases. Similar results are obtained if the speed of fitness variation is increased (Figure 4.6). This demonstrates that there is an interaction between the speed and amplitude of fitness fluctuation on one hand, and the occurrence of sweepstakes reproduction events on the other hand. When selection is fast and strong enough, it can counter-act the stronger genetic drift generated by sweepstakes reproduction (intermediate values of α), increasing thus the probability of fixation for advantageous alleles.

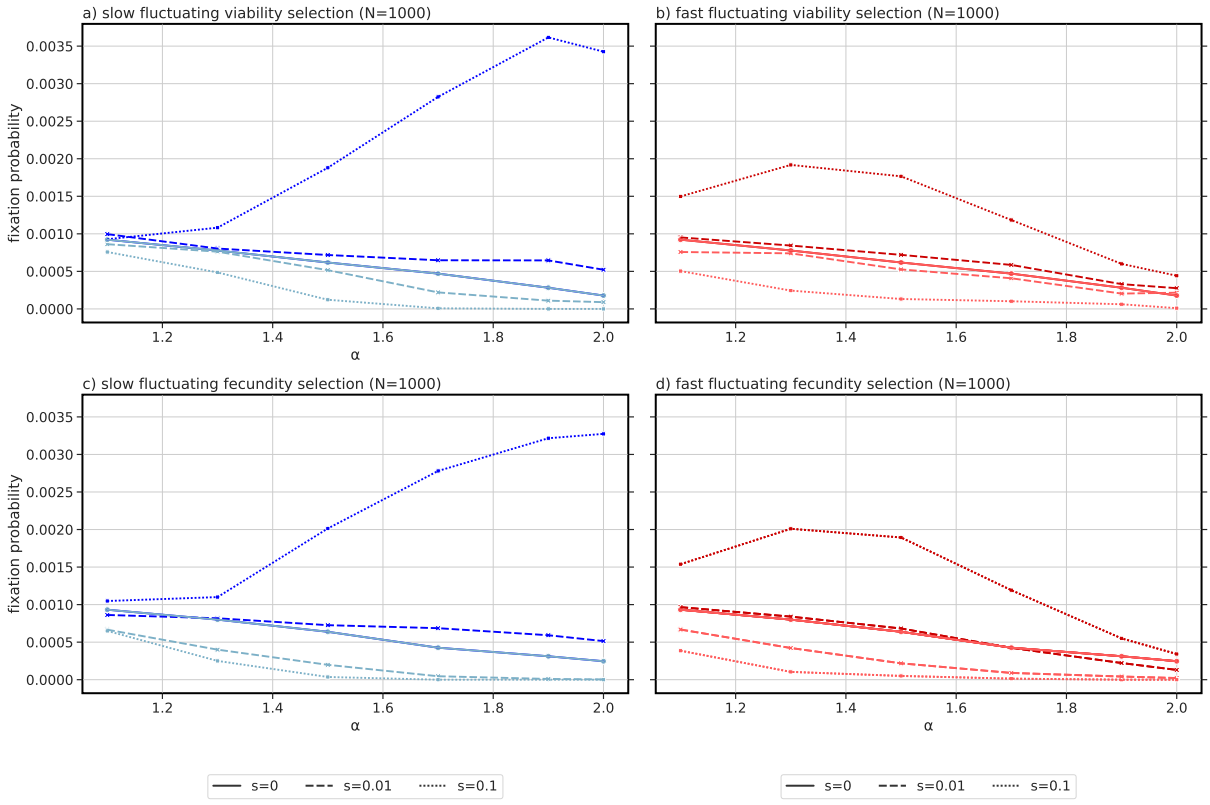


Figure 4.6 Allele fixation probability under constant population size and fluctuating selection coefficients. Fixation probability under different sweepstakes strength (α ranging from 1.1 to 2.0) for (a) slow fluctuating viability selection with initial positive selection s (dark blue) or starting with a negative selection coefficient (light blue). b) fast fluctuating viability selection starting with allele A being advantageous (dark red) and allele A starting being deleterious (light red). The same color code is used for slow (c) and fast (d) fluctuating fecundity selection (F1). Estimates were obtained from 5×10^5 simulations.

4.5 Discussion

Throughout this manuscript, we investigate the effect of sweepstakes reproduction on the three components of speed of adaptation θ, P_{fix}, and T_{fix}. Our results are in line with previous results under viability selection regarding P_{fix} becoming higher, and T_{fix} becoming shorter, when decreasing the parameter α (Eldon and Stephan, 2018, 2023). In the following, we highlight few far reaching implications of our results. We focus on pathogens of crops and plant species where sweepstakes reproduction may occur (Tellier and Lemaire, 2014) to complement the abundant literature on already investigated biological systems such as fish and marine organisms, or bacteria and viruses parasites of humans (Vendrami et al., 2021; Árnason et al., 2023; Sackman et al., 2019; Irwin et al., 2016; Menardo et al., 2020; Sabin et al., 2022; Matuszewski et al., 2017; Morales-Arce et al., 2020). On a technical note of caution, our model assumes that the potential maximum number of produced offspring is unbounded (as in Schweinsberg, 2003). As we intend to keep our model generally applicable to many species, from fish to fungi, we refrain to fix a species specific boundary, although we are aware that this may be biologically unrealistic. Investigating the effect of different boundary values for the number of offspring produced is beyond the scope of this study and will be part of future work focusing on given species of interest.

4.5.1 Mutational process under sweepstakes reproduction

We first analyze the number of new mutations produced either in the parental germ lines giving rise to the gametes or in the offspring. Species with a short life span and with strong sweepstakes reproduction, such as fungi (*e.g.* of crop pathogens), viruses, bacteria, invertebrates, and vertebrates (fish) presenting a typical type-III survivorship ecological strategy (*i.e.* life span is exponentially distributed), do benefit from mutations occurring in offspring (eggs, spores, particles). Indeed, the number of new alleles is proportional to the number of produced offspring (which is affected by α and can possibly be very large) and not to the number of parents, but the probability that individuals with new mutation survive the viability phase is very small. Nevertheless, in such sweepstakes reproducing species with mutations in offspring, a more consistent number of new (advantageous) mutations are observed than when mutations occur in the parents (centered around the expected mean $N \times \mu$). This small variance in number of new mutations would reduce the variance for the waiting time for a new advantageous mutation to appear (the random variable θ), thereby likely speeding up rapid adaptation compared to the case where mutations occur in the parents. When mutations occur in offspring, even under large boom-and-bust cycle of population size, that is the production of large amount of offspring which can mutate, the number of new mutants observed at the next generation is determined by the number of offspring surviving (N) and does not depend on the parameter α. Hence, for parasites, the effective population

size depends on the number of spores/infectious particles which successfully land on, transmit to and infect a host (N). If we define the total number of offspring (spores, infectious particles) produced during the boom cycle of fecundity as the census size (N_{cs}), we obtain the following relationships: the inbreeding effective population size (N_e) is smaller than the number of surviving offspring (N), which is itself smaller than N_{cs}. These discrepancies between N_e, N and N_{cs} become larger as α becomes smaller, meaning that the sweepstakes reproduction model shows increasing discrepancy with the WF model of population evolution in which $N = N_e = N_{cs}$ (Waples, 2005; Hare et al., 2011; Eldon, 2020a and references therein). In other words, sweepstakes reproduction as well as dormancy in seed banks, disentangle the process of genetic drift and mutation as usually considered in the classic expression of the population mutation rate $\theta = 4N_e\mu$ (see Figure 4.1) possibly explaining the so-called Lewontin's paradox in some species (Tellier, 2019; Charlesworth and Jensen, 2022).

We consequently draw two conclusions. First, the high stochastic processes of fecundity and viability under sweepstakes reproduction does generate different, but likely more realistic, predictions regarding the number of new mutations compared to average simple computations based on census size (N_{cs}) observations in the field. Indeed, the census size of the boom fecundity part of the life cycle is routinely assessed in plant pathology by measuring the spore production of fungal lesions (see computations in Stam and McDonald, 2018), while measuring the viability part of the infection cycle is more difficult but essential to define the value of the population size N (likely overestimated in Stam and McDonald, 2018). Conversely, the estimation of inbreeding effective population size (N_e) using polymorphism sequence data is likely biased and likely underestimates (N), meaning that predictions from N_e do not fully assess the full adaptive potential of crop pathogens (see for example McDonald and Linde, 2002). We suggest as a way forward to use full genome polymorphism data to estimate simultaneously N and α (Korfmann et al., 2022b). Second, species with parents exhibiting long live life span (trees, mammals) and in which mutations occurs in parental germ lines (generating the gametes) may be expected to show more variable amount of new (advantageous) mutations each generation. Our results show that this number of new mutations is reduced under strong sweepstakes reproduction (α close to one). Nonetheless, as sweepstakes reproduction and a boom-and-bust life cycle is not common to species with type-I and type-II survivorship ecological strategy (*e.g.* mammals), we may expect our results to be relevant to explain the variance in number of new mutations (and variance in waiting time for new advantageous mutations) in species such as long lived plant species which produce large amount of seeds, few of which surviving to the next generation.

4.5.2 Neutrality versus selection under sweepstakes reproduction

We study the probability of fixation of an allele providing fecundity and/or viability advantage. We first attempt to derive fixation probabilities analytically under a general reproduction

model. Yet, those probabilities could not be obtained without the exact definition of offspring distributions. Hence, we use simulations. We quantify the effect of sweepstakes reproduction on increasing, respectively decreasing, the fixation probabilities of neutral and, respectively, selected alleles. Under strong sweepstakes reproduction (α tends to one), the probability of fixation and the time to fixation of advantageous alleles tend to the probability of fixation of neutral alleles (namely a probability of 0.001 for $N = 1000$ when $\alpha = 1.1$ similar to $1/N$ under the WF model), while these quantities are not notably affected by population size or population size variations.

We draw two further conclusions. First, as the probability of fixation of neutral allele increases under stronger sweepstakes reproduction, we suggest that genomic divergence (substitution rate) is not any longer a sole function of the mutation rate (μ) as assumed under the classic population genetics neutral theory (Charlesworth and Charlesworth, 2010). Phylogenetic methods and dating of past events would need to account for the higher substitution rate in species with low α. We speculate that this effect of neutral sweepstakes may explain the large variance in substitution rates observed across bacteria species (Cui et al., 2013; Gibson and Eyre-Walker, 2019), especially as pervasive and strong recurrent selection generates sweepstakes in bacteria (Neher and Hallatschek, 2013; Menardo et al., 2020). As a follow up, we speculate that classic methods to detect the action of positive selection based on divergence versus polymorphism analyses at synonymous and non-synonymous sites (the McDonald-Kreitman test, McDonald and Kreitman, 1991); and its derivatives (Stoletzki and Eyre-Walker, 2011) could be biased and could show low statistical power in species with sweepstakes. This is because neutral (synonymous) and selected (non-synonymous) allele exhibit similar probability and time to fixation. Second, we suggest to revise the previous claim that sweepstakes reproduction speed up rapid adaptation (Eldon and Stephan, 2018, but a slightly different model generates a significant decrease of the probability of fixation; Eldon and Stephan, 2023), because while we agree that time to fixation is faster, the probability of fixation of advantageous allele is significantly smaller for small values of α (see also the Supplementary Table A11 in Eldon and Stephan, 2018). Furthermore, the detection of selective sweeps is likely complicated by the fact that strong sweepstakes reproduction generates fixation (sweep) of neutral alleles as often as of positively selected alleles. It is therefore expected that genome scan for positive selection would overestimate the number of true selective events and show a high rate of false positives depending on the recombination rate in the genome, a small rate increasing the effect of linkage disequilibrium and pervasive sweepstakes signatures (possibly partially explaining the results in cod fish; Árnason et al., 2023; discussed in Eldon and Stephan, 2023).

4.5.3 Fecundity and viability selection under sweepstakes reproduction

Our results suggest an interaction between the type of selection, fecundity and viability, and the strength of the sweepstakes reproduction. At intermediate sweepstakes reproduction fecundity selection is more efficient than viability, and the reverse is true at low sweepstakes reproduction. Note that as expected when our model is adjusted to fit a WF model, we find the probability and time to fixation of advantageous alleles to be the same for both types of selection (He et al., 2017). These results hold under varying population size. Population size variation only slightly shortens the time to fixation of beneficial alleles, likely because in contrast to (Devi and Jain, 2020), sweepstakes reproduction is the strongest determinant of the effect of genetic drift compared to the weaker effect of population size change with a random bottlenecks model (see also Eldon and Stephan, 2023). Furthermore, when an allele jointly affects fecundity and viability, the probability of fixation and time to fixation of weakly advantageous alleles are largely additive as opposed to multiplicative, but a synergistic effect is observed for strong selection coefficients. This demonstrates the non-additive interaction between the fecundity and viability phases of the life-cycle for species with boom-and-bust dynamics, possibly explaining the high adaptive potential and rapid adaptation of crop or animal pathogens in response to drug/fungicide treatments (Barbosa et al., 2021; Fisher et al., 2022) or to new crop resistant varieties (Persoons et al., 2017). Additionally, we observe that the probability of fixation of a beneficial allele affecting fecundity is elevated if it jointly increases the mean and skewness of the offspring distribution (model F2) compared to only increasing the average fitness (model F1). Therefore, the effect and magnitude of selection can only be appreciated and measured in the light of the offspring distribution and life cycle. We speculate that our F2 selection model may be typical of species such as asexual bacteria (Neher and Hallatschek, 2013) or asexual fungi (pathogen of crops), possible exhibiting some kind of "winner takes all" dynamics whereby one clone would invade the population extremely rapidly due to selection accelerated by neutral sweepstakes reproduction.

Finally, we also investigate the effect of fluctuating fecundity or viability selection. As stronger sweepstakes reproduction corresponds to a higher occurrence of strong genetic drift events, the efficiency of selection depends on the period of selection variation. When the period of selection change is smaller (and with large amplitudes) than the occurrence of sweepstakes reproductive events, selection overruns drift. In other cases, selected alleles do not differ in their probability and time to fixation from the neutral alleles. As previously acknowledged (Kaushik and Jain, 2021), the initial phase of the cycle of selection is crucial to determine the fate of an allele: if the selection coefficient is initially positive, the allele can become fixed, if it is negative, the allele is lost from the population. In this respect, there is no difference between fecundity or viability selection, and the strength of sweepstakes reproduction does not influence this behaviour. As far as we are aware, our results may

be the first to demonstrate the different effect of fecundity and viability selection on fitness under sweepstakes reproduction (Figure 4.1), and thus further work is needed to explore the possible selection models and their biological implications as well as how to test the predictions in empirical data. We highlight especially the cases where fecundity and viability selection occur simultaneously, or that of fluctuating selection, as testing and measuring each selection phase and varying the selection coefficient in time maybe be doable empirically under controlled laboratory conditions.

4.5.4 Conclusion

As indicated in the introduction, we advocate to remain cautious regarding the applicability of our results for the analysis of genome-wide polymorphism data (see for example Johri et al., 2021, 2022a). Indeed, our model considers only one single locus under positive selection. The speed of rapid adaptation in species with sweepstakes, and the signatures of selective sweeps, are realistically affected by linked (positive and negative) selection at neighbouring sites with various distribution of fitness effects (*e.g.* Hill and Robertson, 1966; Eyre-Walker and Keightley, 2007; Johri et al., 2022a). Nevertheless, life history traits, such as sweepstakes reproduction, dormancy and clonality are ubiquitous in many virus, bacteria, fungi, invertebrates and plant species, and do impact the reproductive mechanism and consequently the relative importance of the various forces shaping genome evolution. We disentangle here the fecundity and viability phases of the life cycle (Figure 4.1) under sweepstakes reproduction. We note that similar observations have been made for seed dormancy disentangling fecundity and viability selection in dormant seeds (Heinrich et al., 2018). However, in nature, organisms may exhibit several peculiar life history traits (dormancy, sweepstakes reproduction) such that there is a large variance in produced offspring, some of which remain dormant for more than one generation before they become activated and can reproduce. Hence, it may be important to consider more complex but realistic models of evolution for species with peculiar life cycles to assess the limits of current population genetics theory, and better predict the evolutionary potential of such species possibly important for medicine or agriculture.

Acknowledgments

KK is supported by a grant from the Deutsche Forschungsgemeinschaft (DFG) through the TUM International Graduate School of Science and Engineering (IGSSE), Graduate School (GSC) 81, within the project GENOMIE QADOP. TS is supported by the Austrian Science Fund (project no. TAI 151-B) to Anja Hörger. AT is supported by grant TE809/3 (project 274542535) from the Deutsche Forschungsgemeinschaft (DFG).

Data accessibility statement

The simulated data that support the findings of this study are openly available at:
https://github.com/kevinkorfmann/CanningsSimulator

4.6 Supplementary Material

The following part contains the supplementary material, including the theoretical setup, implementation details of the offspring distribution, tables for comparison to theoretical expectations and supplementary figures.

4.7 Theoretical set-up: Genetic drift and allele frequencies

We describe here the diffusion of the allele A in our population of haploid individuals of constant size N under the different Cannings models: the neutral Cannings model and the models with fecundity selection or viability selection.

4.7.1 The neutral case

We note n_t the number of individual of type A, w_t the number of offspring produced by individual of type A and M_t the total number of offspring produced at the generation t. In addition we write X_i the number of offspring produced by individual i and for $k \geq 1$ we define $\overline{X}_k = X_1 + ... + X_k$. Without loss of generality, we assume that the individuals of type A are indexed from 1 to n_t and the individual of type a from $n_t + 1$ to N. Hence $\overline{X}_{n_t} = w_t$ and $\overline{X}_N = M_t$.

Theorem 1 (Neutral case) *Under the neutral Cannings model n_t is a martingale and we have the following:*

$$\mathbb{E}(n_{t+1}|n_t = n) = n$$

$$Var(n_{t+1}|n_t = n, w_t = w, M_t = M) = N\frac{w}{M}\frac{(M-w)}{M}\frac{(M-N)}{(M-1)}$$

In order to prove Theorem 1 we need the following lemma

Lemma 1 *Let $(X_i)_{i\geq 1}$ be identical independent $\mathbb{N}$-valued random variable. Let $k \geq 1$ and let note $\overline{X}_k = X_1 + ... + X_k$. We have the following result:*

$$\frac{m}{k} = \mathbb{E}(X_1|\overline{X}_k = m)$$

[Proof of Lemma 1] We first compute the expectation of the sum and the square of the sum conditionally to the value of the sum.

We hence have:

$$\begin{aligned} m &= \mathbb{E}(\overline{X}_k|\overline{X}_k = m) \\ &= \sum_{i=1}^{k} \mathbb{E}(X_i|\overline{X}_k = m) \text{ by linearity of the expectation} \\ &= k\mathbb{E}(X_1|\overline{X}_k = m) \text{ because the } X_i \text{ are exchangeable} \end{aligned}$$

[Proof of Theorem 1]

The number of surviving offspring of type A in the next generation n_{t+1} correspond to a drawing without replacement of N offspring amongst M_t in which w_t have the type A. Hence the law of n_{t+1} knowing that $w_t = w$ and $M_t = M$ is an hypergeometric $\mathcal{H}(w, M, N)$. We therefore have :

$$\mathbb{E}(n_{t+1}|n_t = n, M_t = M, w_t = w) = \frac{Nw}{M} \tag{4.6}$$

Since for any random variable X, Y, Z we have $\mathbb{E}(X|Y = y) = \sum_z \mathbb{P}(Z = z|Y = y)\mathbb{E}(X|Y = y, Z = z)$ and using 4.6 we have:

$$\begin{aligned} \mathbb{E}(n_{t+1}|n_t = n, M_t = M) &= \sum_w \mathbb{P}(w_t = w|n_t = n, M_t = M)\mathbb{E}(n_{t+1}|n_t = n, M_t = M, w_t = w) \\ &= \frac{N}{M}\mathbb{E}(w_t|n_t = n, M_t = M) \end{aligned}$$

Knowing that $n_t = n$ we have then $w_t = \sum_{i=1}^{n} X_i$ hence by linearity of the expectation and because the $(X_i)_{1\leq i\leq N}$ have all the same law and are exchangeable, we have $\mathbb{E}(w_t|n_t = n, M_t = M) = n\mathbb{E}(X_1|n_t = n)$ which gives:

$$\mathbb{E}(n_{t+1}|n_t = n, M_t = M) = \frac{N}{M}n\mathbb{E}(X_1|M_t = M)$$

Because $M_t = \overline{X}_N$ and of Lemma 1 we have $\mathbb{E}(X_1|\overline{X}_N = M) = \frac{M}{N}$ and $\mathbb{E}(n_{t+1}|n_t = n, M_t = M) = n$ which leads to $\mathbb{E}(n_{t+1}|n_t = n) = n$. Hence n_t is a martingale.

As mentioned above, the number of surviving offspring of type A in the next generation n_{t+1} is an hypergeometric $\mathcal{H}(w, M, N)$ knowing that $w_t = w$ and $M_t = M$. Hence:

$$\mathrm{Var}(n_{t+1}|n_t = n, w_t = w, M_t = M) = N\frac{w}{M}\frac{(M-w)}{M}\frac{(M-N)}{(M-1)}$$

An explicit formula for the variance is not obtainable without the explicit definition of the offspring distribution. At that stage, we leave to future work more involving analytical derivations, and use stochastic simulations thereafter.

4.7.2 Fecundity Selection

We note X_t the number of offspring produced by individual of type A and Y_t the number of offspring produced by individual of type a at the generation t. In addition we write Y_i the number of offspring produced by individual i of type a such that $E(Y_i) = \nu_y$ with $\nu_y > 1$, and $\mathrm{Var}(Y_i) = \sigma_y^2$ with $\sigma_y > 0$. We write X_i the number of offspring produced by individual i of type A. Individuals of type A display a fecundity advantage of s such that $E(X_i) = \nu_x = (1+s)\nu_y = (1+s)E(Y_i)$, with $s > 0$, and $\mathrm{Var}(X_i) = \sigma_x^2$,with $\sigma_x > 0$. We further assume all individuals to produce offspring independently from one another.

The number of surviving offspring of type A in the next generation, written n_{t+1}, correspond to a drawing without replacement of N offspring amongst $Y_t + X_t$ in which X_t have the type A. Hence the law of n_{t+1} knowing that $Y_t = Y$ and $X_t = X$ is an hypergeometric $\mathcal{H}(X, X+Y, N)$. We therefore have :

$$\mathbb{E}(n_{t+1}|n_t = n, X_t = X, Y_t = Y) = N\frac{X}{(X+Y)}$$

$$\mathrm{Var}(n_{t+1}|n_t = n, X_t = X, Y_t = Y) = N\frac{X}{(X+Y)}\frac{Y}{(X+Y)}\frac{(X+Y-N)}{(X+Y-1)}$$

However, because X_t and Y_t are independent we can write :

$$\mathbb{E}(n_{t+1}|n_t = n) = \sum_X \sum_Y \mathbb{P}(Y_t = Y|n_t = n)\mathbb{P}(X_t = X|n_t = n)N\frac{X}{(X+Y)}$$

$$\mathrm{Var}(n_{t+1}|n_t = n) = \sum_X \sum_Y \mathbb{P}(Y_t = Y|n_t = n)\mathbb{P}(X_t = X|n_t = n)N\frac{X}{(X+Y)}\frac{Y}{(X+Y)}\frac{(X+Y-N)}{(X+Y-1)}$$

As above, an explicit formula is not obtainable without a proper definition the offspring production's distribution. However, in practice the population size N is assumed to be large. We can hence assume the number of individuals of type a and A (respectively $N-n$ and n) to be large as well. Because, $X = X_1 + ... + X_n$ and $Y = Y_1 + ... + Y_{N-n}$, the central limit theorem applies so that:

$$Y \approx \mathcal{N}((N-n)\nu_y, (N-n)\sigma_y^2)$$
$$X \approx \mathcal{N}(n(1+s)\nu_y, n\sigma_x^2)$$

In this case, we can write :

$$\mathbb{E}(n_{t+1}|n_t = n) \approx \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} N\frac{x}{(x+y)}\frac{e^{-\frac{(x-(1+s)n\nu_y)^2}{2(n\sigma_x)^2}}}{n\sigma_x\sqrt{2\pi}}\frac{e^{-\frac{(y-(N-n)\nu_y)^2}{2((N-n)\sigma_y)^2}}}{(N-n)\sigma_y\sqrt{2\pi}}d_x d_y$$

$$\mathrm{Var}(n_{t+1}|n_t = n) \approx \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} N\frac{x}{(x+y)}\frac{y}{(x+y)}\frac{(x+y-N)}{(x+y-1)}\frac{e^{-\frac{(x-(1+s)n\nu_y)^2}{2(n\sigma_x)^2}}}{n\sigma_x\sqrt{2\pi}}\frac{e^{-\frac{(y-(N-n)\nu_y)^2}{2((N-n)\sigma_y)^2}}}{(N-n)\sigma_y\sqrt{2\pi}}d_x d_y$$

4.7.3 Viability Selection

We note X_t the number of offspring produced by individual of type A and Y_t the number of offspring produced by individual of type a at the generation t. We write X_i the number of offspring produced by individuals i. We here assume all X_i to be independent and identically distributed. However, individuals of type A display a viability advantage (*i.e.* more likely to survive) of $s > 0$ such that the N surviving offspring are the draw according to the Wallenius non central hypergeometric distribution where the weight of an offspring is $1 + s$ if it has the type A, and 1 otherwise.

Hence the number of surviving offspring of type A in the next generation, written n_{t+1}, correspond to a drawing without replacement of N offspring amongst $Y_t + X_t$ in which X_t have the type A. Hence the law of n_{t+1} knowing that $Y_t = Y$ and $X_t = X$ is a Wallenius non central hypergeometric distribution $wnchypg(N, X, Y, (1+s))$. Unfortunately there are no explicit formulae for the variance or the expectation of a Wallenius non central hypergeometric distribution, but these can be computed numerically and by stochastic simulations.

4.8 Simulating offsprings

We build a forward simulator model as described in the main text. At each generation parents produce offspring and the distribution is obtained as follows. The number X_i of offspring being obtained through inverse transform sampling by:

$$X_i = e^{\frac{1}{\alpha} \times \ln \frac{1-p_0}{u}} \tag{4.12}$$

With u being a random number ($u \sim U(0,1)$) and p_0 the probability for an individual to not produce any offspring. All X_i are summed, and floored to discretize the number of sampled alleles. The parameter α ranging from $\alpha = 2$ to $\alpha = 1$ affects the probability of sweepstake reproductive events to occur. When $\alpha = 2$ the reproductive mechanism is similar to the one in a Wright-Fisher. This way the number of type A allele individuals are sampled, while the number of type a allele are calculated by subtracting the number of type A from the total population size.

Likewise, modeling the offspring of a Wright-Fisher model in a Cannings formulation is achieved by sampling a Poisson number of offspring with rate parameter λ.

$$X_i = Poisson(\lambda) \tag{4.13}$$

Further implementation details can be found at:
https://github.com/kevinkorfmann/CanningsSimulator/blob/main/src/CanningsSimulator.jl.

4.9 Simulated scenarios

We investigate three possible scenarios: constant population size, varying population size and fluctuating selection. The fluctuating population size is set to vary in regular cycles with a given period and amplitude, to yield an harmonic mean of size 500, 1,000 and 5,000 as defined in Figure 4.7. These values of the harmonic mean are the same as the values used for the constant population size scenario. The cycles of population size variation can either start with a population increase or population decrease as modelled by the shift in the sinus function by a factor π (Figure 4.7). Fluctuating selection is modelled to obtain an average selection coefficient of zero across a full period. We choose two different amplitudes for s (0.01 and 0.1) and two periods (fast with 100, and slow with 1,000 generations) as shown in Figure 4.8. Slow fluctuating selection is defined as $s \times sin(0.006283x)$ and fast fluctuating selection as $s \times sin(0.06283x)$. Importantly, the initial part of the cycle can yield that the A allele is advantageous, or deleterious when shifting the sin function by π (Figure 4.8).

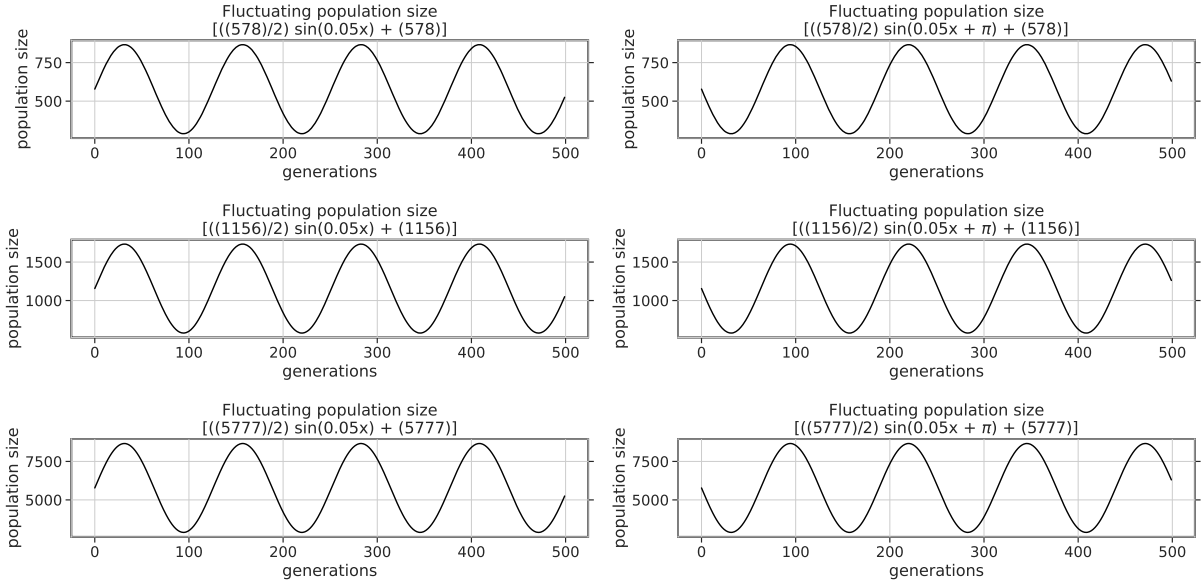


Figure 4.7 Fluctuating population size model. The curves indicate the population size variation with harmonic means of 500, 1,000 and 5,000 and the same periods.

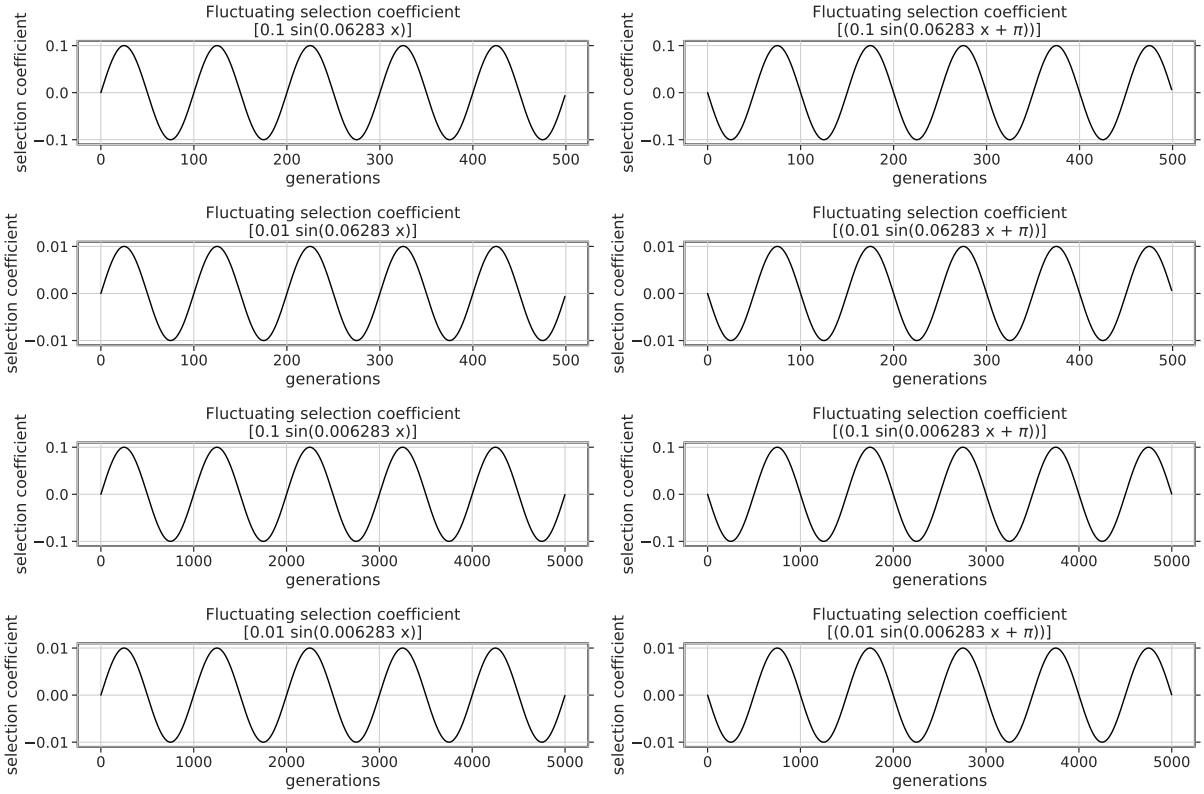


Figure 4.8 Fluctuating selection coefficient model. The curves indicate the change in selection coefficient with mean $s = 0$, two amplitudes (0.01 and 0.1) and two periods (fast and slow).

4.10 Supplementary Table

Table 4.1 Average fixation probability and time for Cannings model with Poisson offspring distribution with rate parameter $\lambda = 1.2$. Probabilities are obtained from 5×10^5 simulations, while fixation times are estimated based on 5×10^3 simulations conditioned on fixation. In bold the neutral cases for which the expectation is $1/N$ for the probability of fixation, and $2N$ for the time to fixation.

Selection type	Selection coefficient	Population size N	Fixation probability	Fixation time
fecundity	**0.00**	**500**	**0.00197**	**999.15**
fecundity	0.01	500	0.00685	587.38
fecundity	0.10	500	0.06851	108.64
fecundity	**0.00**	**1,000**	**0.00097**	**2000.45**
fecundity	0.01	1,000	0.00689	740.88
fecundity	0.10	1,000	0.06826	123.08
fecundity	**0.00**	**5,000**	**0.00023**	**9976.43**
fecundity	0.01	5,000	0.00674	1070.92
fecundity	0.10	5,000	0.06871	156.69
viability	**0.00**	**500**	**0.00201**	**995.14**
viability	0.01	500	0.00736	862.7
viability	0.10	500	0.06169	239.36
viability	**0.00**	**1,000**	**0.00095**	**2005.19**
viability	0.01	1,000	0.00705	1328.78
viability	0.10	1,000	0.06274	280.35
viability	**0.00**	**5,000**	**0.0002**	**9973.58**
viability	0.01	5,000	0.00702	2303.67
viability	0.10	5,000	0.06276	373.39

4.11 Supplementary Figures

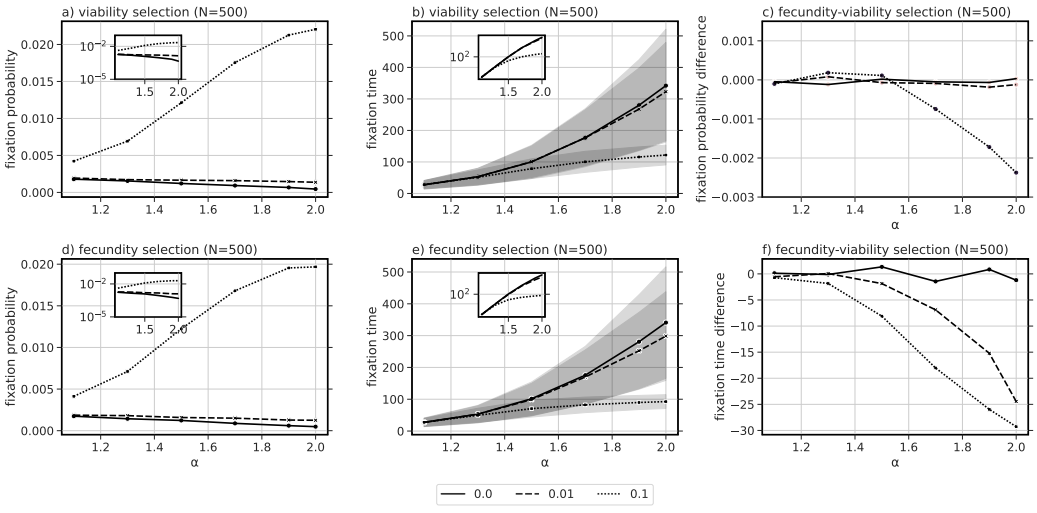


Figure 4.9 Allele fixation probability and average time to fixation of alleles under constant population size. Fixation probability (a, d) and average fixation times (b, e) under different sweepstake strength (α ranging from 1.1 to 2.0) for (a, b) viability selection and (d, e) fecundity selection (F1). Three selection coefficients are shown ($s = 0; 0.01; 0.1$) and $N = 500$. Probabilities are obtained from 5×10^5 simulations, while fixation times are estimated based on 5×10^3 simulations conditioned on fixation. Shaded areas correspond to 95% confidence intervals. Panels (c) and (f) present the difference between the probability of fixation (c) and time to fixation (f) for fecundity compared to viability selection.

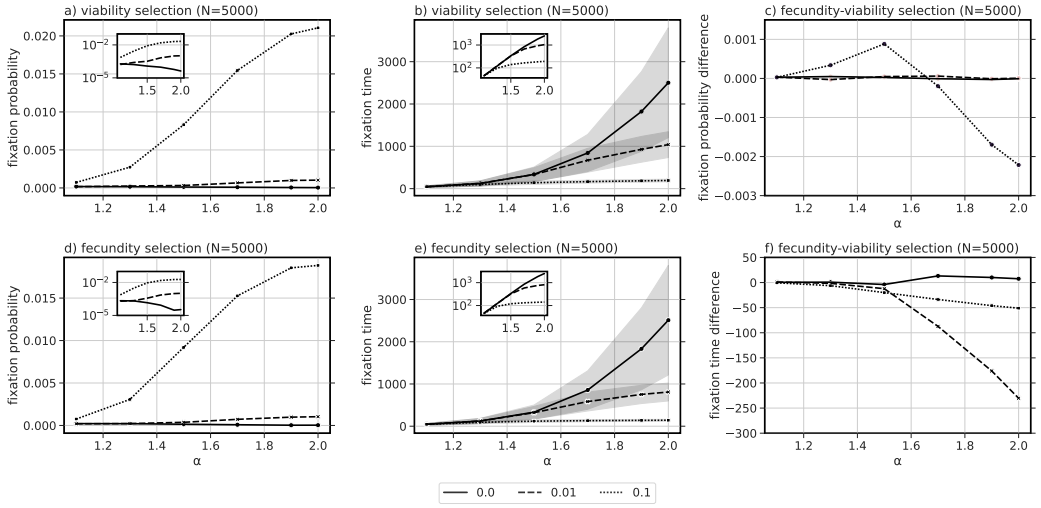


Figure 4.10 Allele fixation probability and average time to fixation of alleles under constant population size. Fixation probability (a, d) and average fixation times (b, e) under different sweepstake strength (α ranging from 1.1 to 2.0) for (a, b) viability selection and (d, e) fecundity selection (F1). Three selection coefficients are shown ($s = 0; 0.01; 0.1$) and $N = 5,000$. Probabilities are obtained from 5×10^5 simulations, while fixation times are estimated based on 5×10^3 simulations conditioned on fixation. Shaded areas correspond to 95% confidence intervals. Panels (c) and (f) present the difference between the probability of fixation (c) and time to fixation (f) for fecundity compared to viability selection.

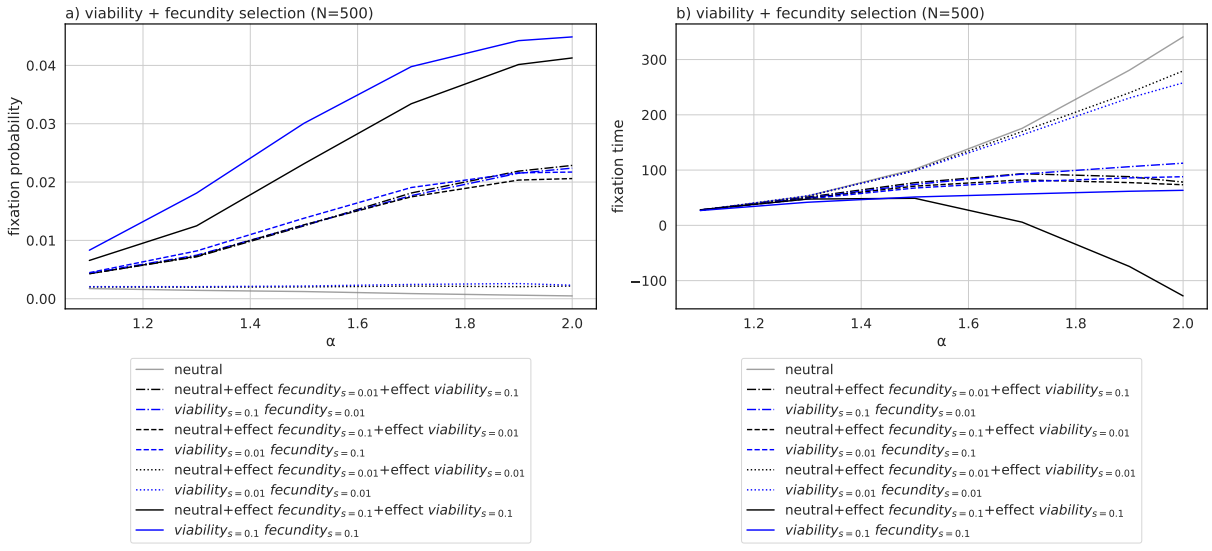


Figure 4.11 Allele fixation probability and average fixation time under the joint selection model and N=500. a) Probability of fixation under joint viability and fecundity selection (F1, blue) in comparison to neutral fixation probability with added effect of fecundity (F1) and viability selection, when simulated individually (black). b) Average time to fixation of alleles under simultaneous selection models (blue) in comparison to neutral estimates of time to fixation summed up with the effect of each selection type (black). Probabilities are obtained from 5×10^5 simulations, while fixation times are estimated based on 5×10^3 simulations conditioned on fixation.

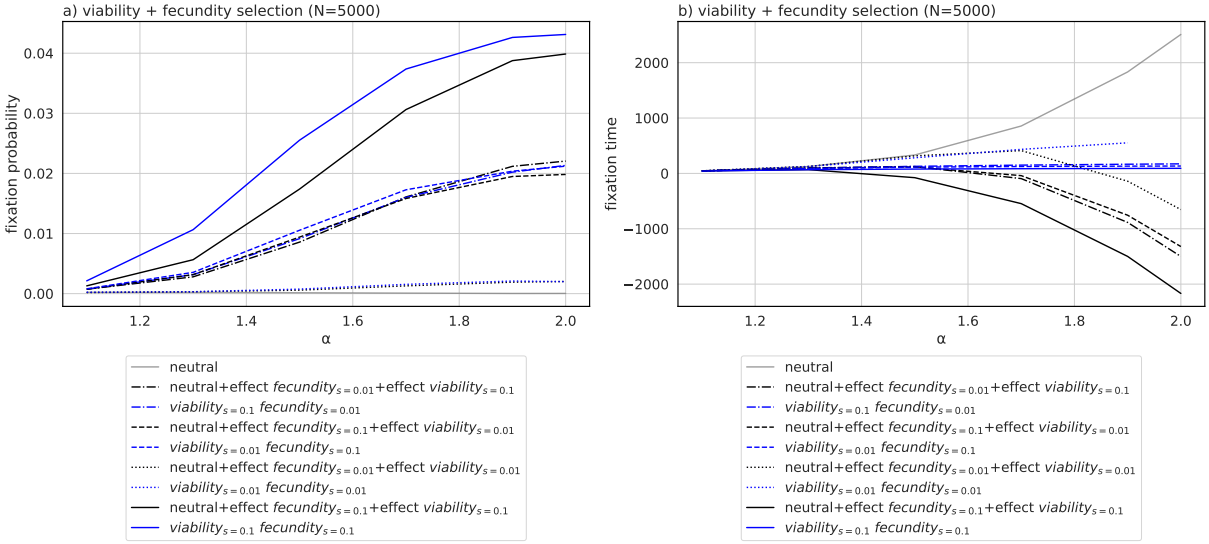


Figure 4.12 Allele fixation probability and average fixation time under the joint selection model and N=5,000. a) Probability of fixation under joint viability and fecundity selection (F1, blue) in comparison to neutral fixation probability with added effect of fecundity (F1) and viability selection, when simulated individually (black). b) Average time to fixation of alleles under simultaneous selection models (blue) in comparison to neutral estimates of time to fixation summed up with the effect of each selection type (black). Probabilities are obtained from 5×10^5 simulations, while fixation times are estimated based on 5×10^3 simulations conditioned on fixation.

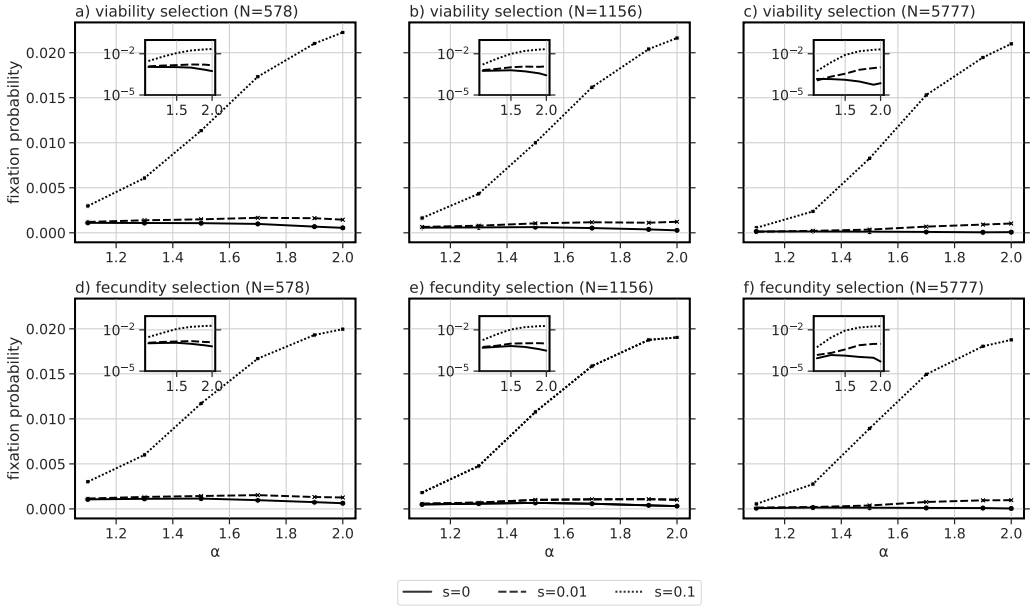


Figure 4.13 Allele fixation probability under variable population size with initial increase in size. Forward in-time simulations of a Cannings model with Schweinsberg offspring probability of different sweepstake strength parameter α range from 1.1 to 2.0 under viability (a-c) and fecundity(d-f) selection (F1) with selection coefficients ranging from the neutral case to strong positive selection. Population size at time t p_t is defined as $p_t = \frac{N}{2} * sin(0.05 \times t) + N$ with $N = 578$, $N = 1,156$ and $N = 5,777$ individuals leading to harmonic mean of $N^{HM} = 500$ (a, d), $N^{HM} = 1,000$ (b, e), $N^{HM} = 5,000$ (c, f). Each combination of α, selection coefficient and population size parameter was simulated 5×10^5 times, while counting the number of fixation events to estimate the fixation probability.

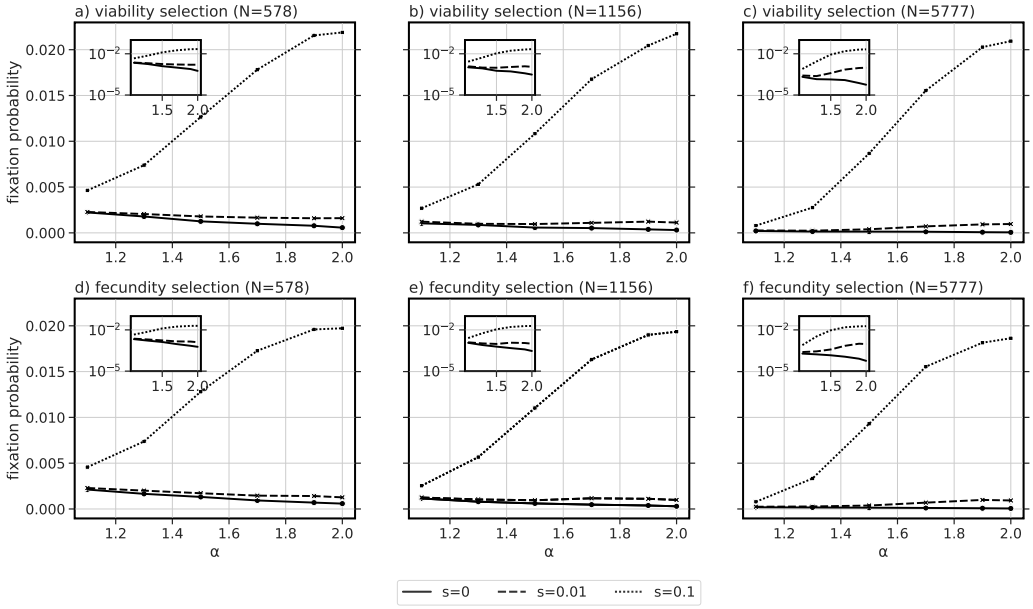


Figure 4.14 Allele fixation probability under variable population size with initial decrease in size. Forward in-time simulations of a Cannings model with Schweinsberg offspring probability of different sweepstake strength parameter α range from 1.1 to 2.0 under viability (a-c) and fecundity (d-f) selection (F1) with selection coefficients ranging from the neutral case to strong positive selection. Population size at time t p_t is defined as $p_t = \frac{N}{2} * sin(0.05 \times t + \pi) + N$ with $N = 578$, $N = 1,156$ and $N = 5,777$ individuals leading to harmonic mean of $N^{HM} = 500$ (a, d), $N^{HM} = 1,000$ (b, e), $N^{HM} = 5,000$ (c, f). Each combination of α, selection coefficient and population size parameter was simulated 5×10^5 times, while counting the number of fixation events to estimate the fixation probability.

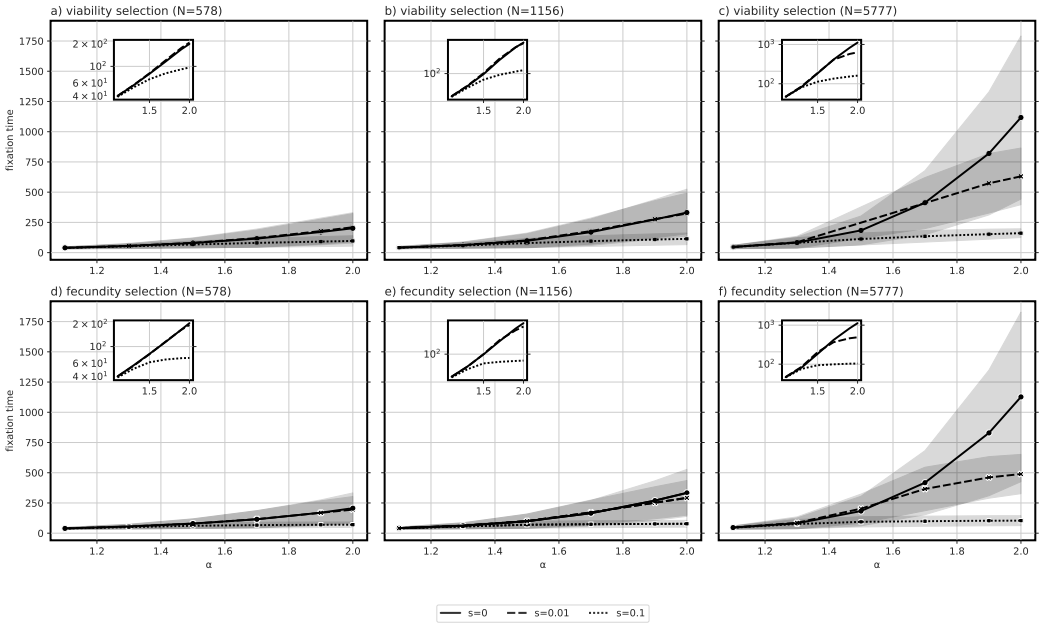


Figure 4.15 Allele fixation time under variable population size with initial increase in size. Forward in-time simulations of a Cannings model with Schweinsberg offspring probability of different sweepstake strength parameter α range from 1.1 to 2.0 under viability (a-c) and fecundity (d-f) selection (F1) with selection coefficients ranging from the neutral case to strong positive selection. Population size at time t p_t is defined as $p_t = \frac{N}{2} * sin(0.05 \times t) + N$ with $N = 578$, $N = 1,156$ and $N = 5,777$ individuals leading to harmonic mean of $N^{HM} = 500$ (a, d), $N^{HM} = 1,000$ (b, e), $N^{HM} = 5,000$ (c, f). Each combination of α, selection coefficient and population size parameter was simulated until 5,000 alleles reached fixation. Mean and 95% confidence interval of fixation time are presented.

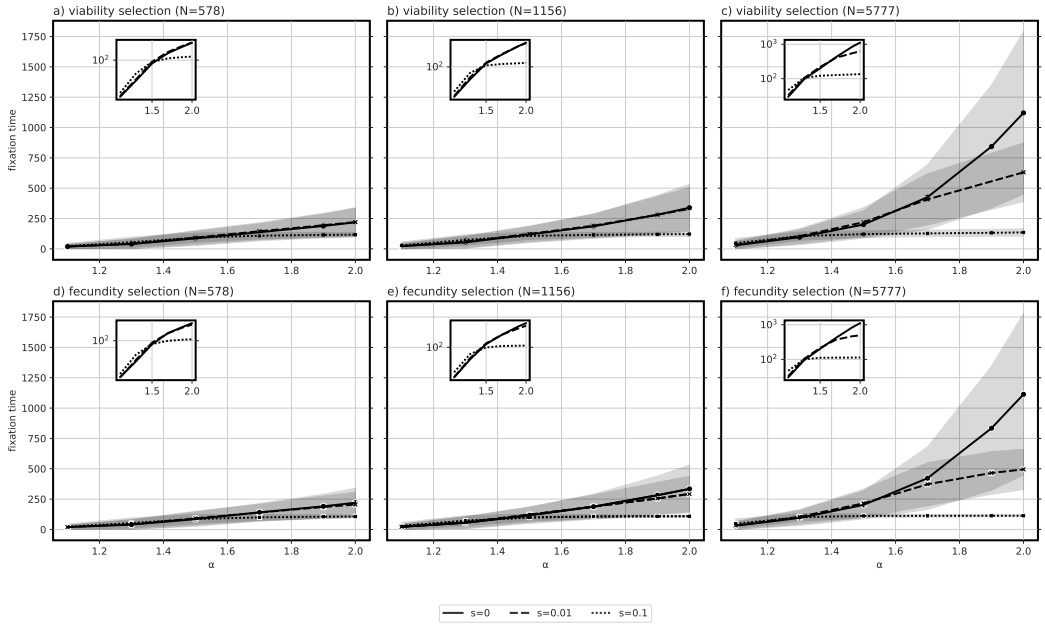


Figure 4.16 Allele fixation time under variable population size with initial decrease in size. Forward in-time simulations of a Cannings model with Schweinsberg offspring probability of different sweepstake strength parameter α range from 1.1 to 2.0 under viability (a-c) and fecundity (d-f) selection (F1) with selection coefficients ranging from the neutral case to strong positive selection. Population size at time t p_t is defined as $p_t = \frac{N}{2} * sin(0.05 \times t + \pi) + N$ with $N = 578$, $N = 1,156$ and $N = 5,777$ individuals leading to harmonic mean of $N^{HM} = 500$ (a, d), $N^{HM} = 1,000$ (b, e), $N^{HM} = 5,000$ (c, f). Each combination of α, selection coefficient and population size parameter was simulated until 5,000 alleles reached fixation. Mean and 95% confidence interval of fixation time are presented.

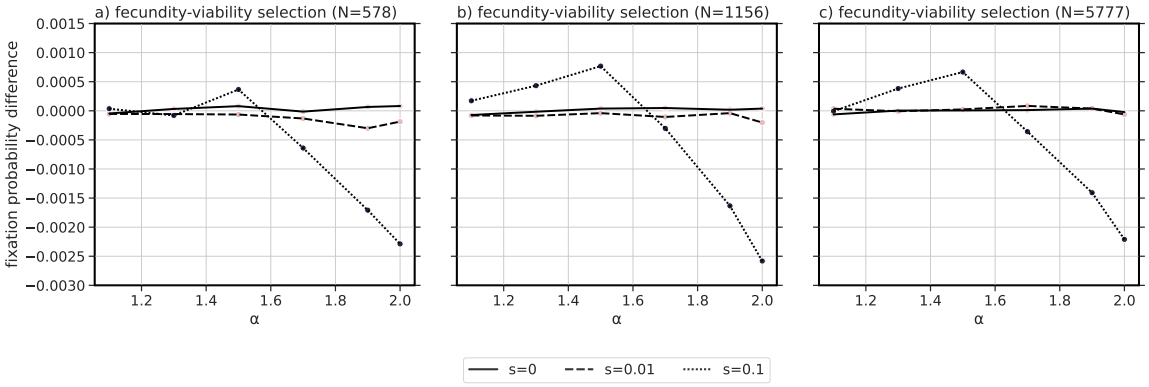


Figure 4.17 Difference in allele fixation probability between fecundity and viability selection under variable population size with initial size increase. Forward in-time simulations of a Cannings model with Schweinsberg offspring probability difference (fecundity (F1)- viability selection)of different sweepstake strength parameter α range from 1.1 to 2.0 with selection coefficients ranging from the neutral case to strong positive selection. Population size at time t p_t is defined as $p_t = \frac{N}{2} * sin(0.05 \times t) + N$ with $N = 578$, $N = 1,156$ and $N = 5,777$ individuals leading to harmonic mean of $N^{HM} = 500$ (a), $N^{HM} = 1,000$ (b), and $N^{HM} = 5,000$ (c). Each combination of α, selection coefficient and population size parameter was simulated 5×10^5 times, while counting the number of fixation events to estimate the fixation probability.

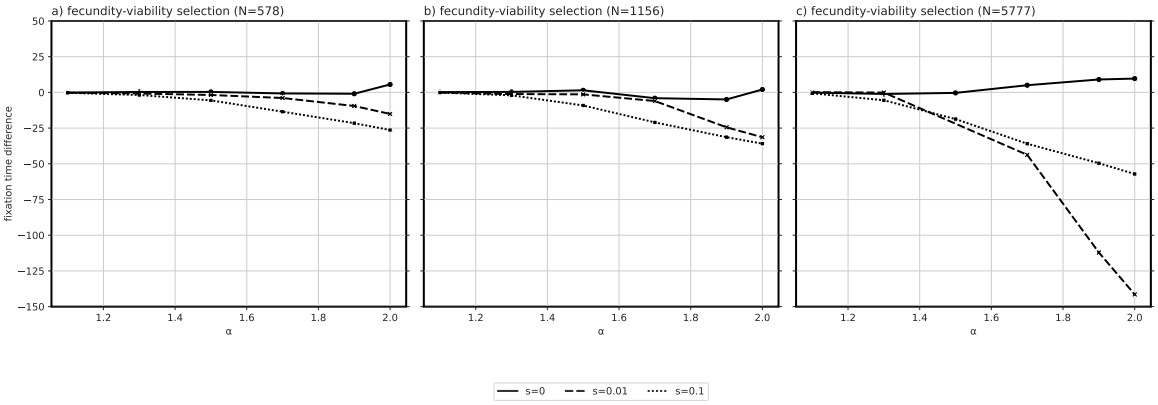


Figure 4.18 Difference in time to fixation between fecundity and viability selection under under variable population size with initial size increase. Forward in-time simulations of a Cannings model with Schweinsberg offspring probability difference (fecundity (F1) - viability selection)of different sweepstake strength parameter α range from 1.1 to 2.0 with selection coefficients ranging from the neutral case to strong positive selection. Population size at time t p_t is defined as $p_t = \frac{N}{2} * sin(0.05 \times t) + N$ with $N = 578$, $N = 1,156$ and $N = 5,777$ individuals leading to harmonic mean of $N^{HM} = 500$ (a), $N^{HM} = 1,000$ (a), and $N^{HM} = 5,000$ (c). Each combination of α, selection coefficient and population size parameter was simulated until 5,000 alleles reached fixation.

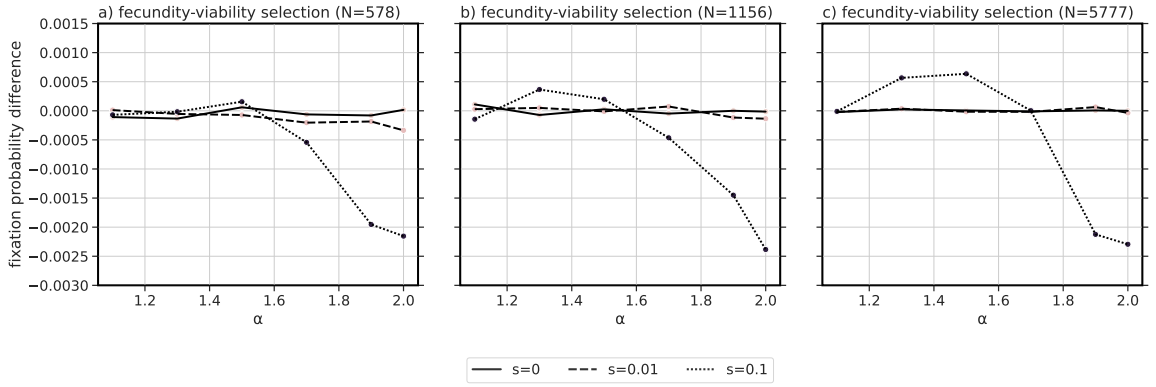


Figure 4.19 Difference in allele fixation probability between fecundity and viability selection under variable population size with initial size decrease. Forward in-time simulations of a Cannings model with Schweinsberg offspring probability difference (fecundity (F1)- viability selection) of different sweepstake strength parameter α range from 1.1 to 2.0 with selection coefficients ranging from the neutral case to strong positive selection. Population size at time t p_t is defined as $p_t = \frac{N}{2} * sin(0.05 \times t + \pi) + N$ with $N = 578$, $N = 1,156$ and $N = 5,777$ individuals leading to harmonic mean of $N^{HM} = 500$ (a), $N^{HM} = 1,000$ (b), and $N^{HM} = 5,000$ (c). Each combination of α, selection coefficient and population size parameter was simulated 5×10^5 times, while counting the number of fixation events to estimate the fixation probability.

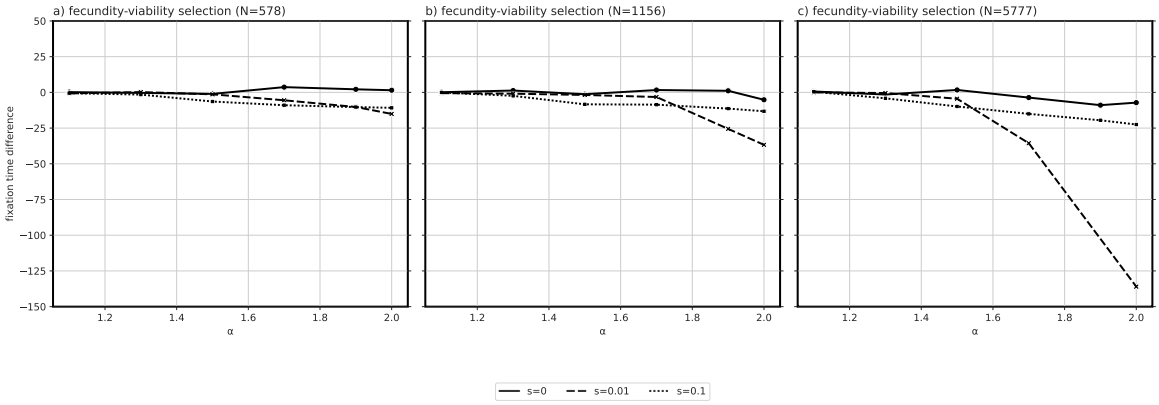


Figure 4.20 Difference in time to fixation between fecundity and viability selection under under variable population size with initial size decrease. Forward in-time simulations of a Cannings model with Schweinsberg offspring time difference (fecundity (F1) - viability selection) of different sweepstake strength parameter α range from 1.1 to 2.0 with selection coefficients ranging from the neutral case to strong positive selection. Population size at time t p_t is defined as $p_t = \frac{N}{2} * sin(0.05 \times t + \pi) + N$ with $N = 578$, $N = 1,156$ and $N = 5,777$ individuals leading to harmonic mean of $N^{HM} = 500$ (a), $N^{HM} = 1,000$ (b), and $N^{HM} = 5,000$ (c). Each combination of α, selection coefficient and population size parameter was simulated until 5,000 alleles reached fixation.

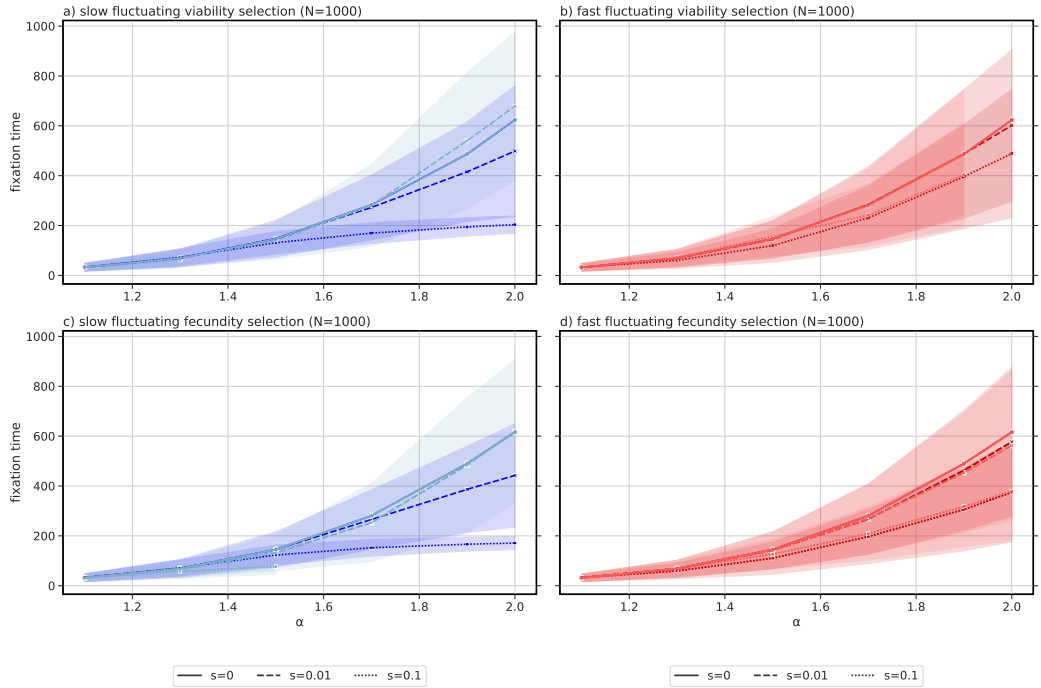


Figure 4.21 Time to fixation under constant population size and fluctuating selection coefficients with N=1,000. Fixation times under different sweepstake strength (α ranging from 1.1 to 2.0) for (a) slow fluctuating viability selection starting with a positive value of s (dark blue) and starting with a negative value of s (light blue) and (b) fast fluctuating viability selection starting with a positive (dark red) and negative (light red) value of s. The same color code is used for slow (c) and fast (d) fluctuating fecundity selection (F1). Each combination of α, selection coefficient and population size parameter was simulated until 5,000 alleles reached fixation. Mean and 95% confidence interval of fixation time are presented.

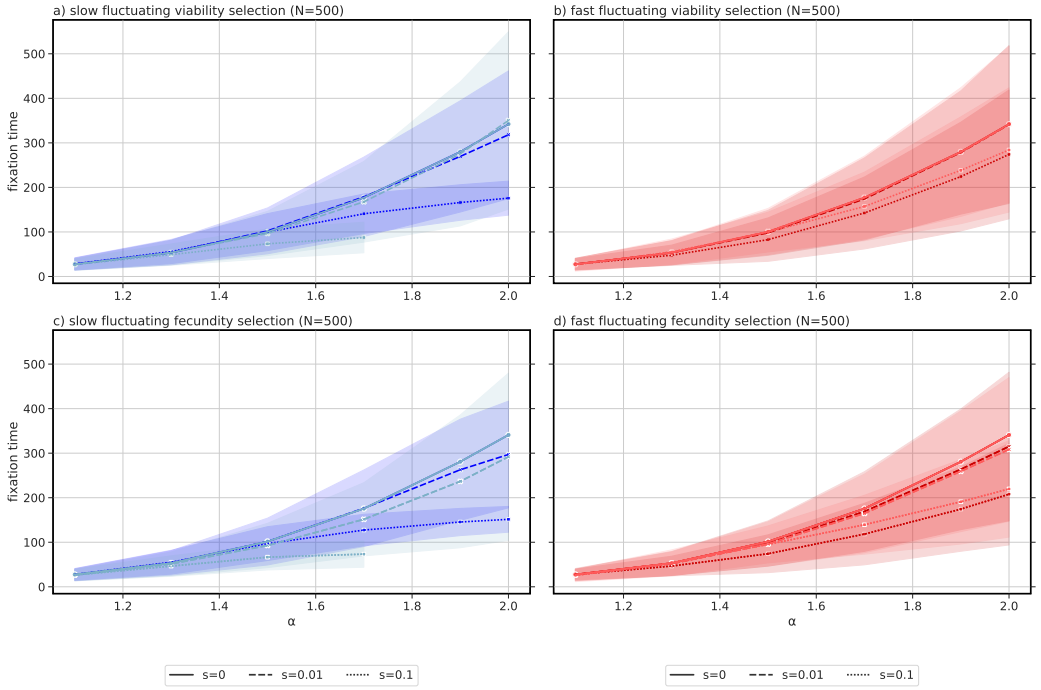


Figure 4.22 Time to fixation under constant population size and fluctuating selection coefficients with N=500. Fixation times under different sweepstake strength (α ranging from 1.1 to 2.0) for (a) slow fluctuating viability selection starting with a positive value of s (dark blue) and starting with a negative value of s (light blue) and (b) fast fluctuating viability selection starting with a positive (dark red) and negative (light red) value of s. The same color code is used for slow (c) and fast (d) fluctuating fecundity selection (F1). Each combination of α, selection coefficient and population size parameter was simulated until 5,000 alleles reached fixation. Mean and 95% confidence interval of fixation time are presented.

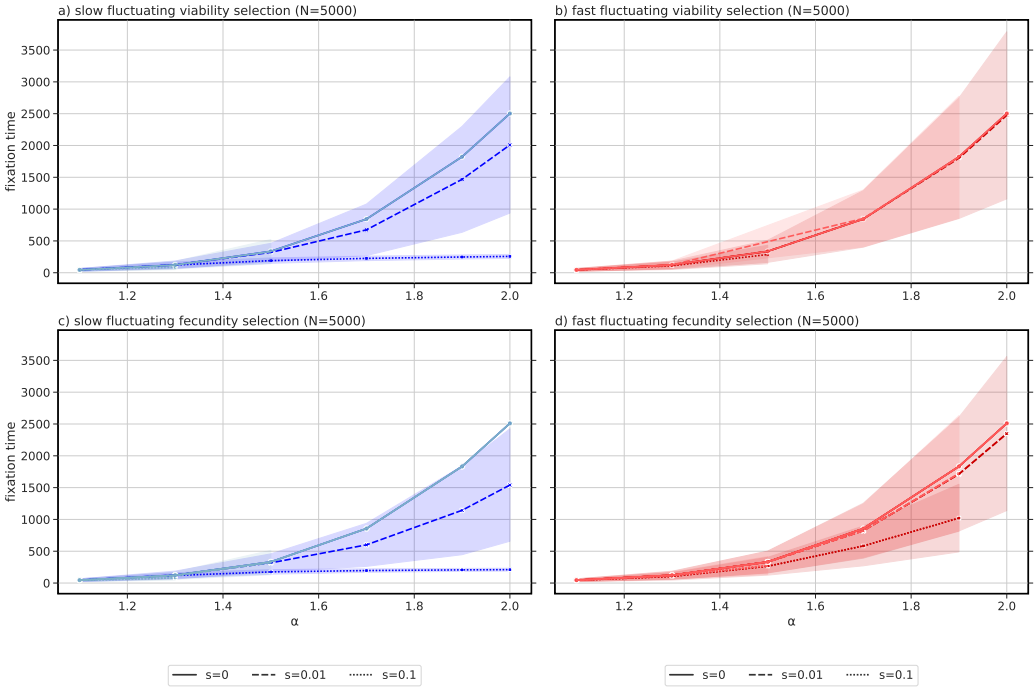


Figure 4.23 Time to fixation under constant population size and fluctuating selection coefficients with N=5,000. Fixation times under different sweepstake strength (α ranging from 1.1 to 2.0) for (a) slow fluctuating viability selection starting with a positive value of s (dark blue) and starting with a negative value of s (light blue) and (b) fast fluctuating viability selection starting with a positive (dark red) and negative (light red) value of s. The same color code is used for slow (c) and fast (d) fluctuating fecundity selection (F1). Each combination of α, selection coefficient and population size parameter was simulated until 5,000 alleles reached fixation. Mean and 95% confidence interval of fixation time are presented.

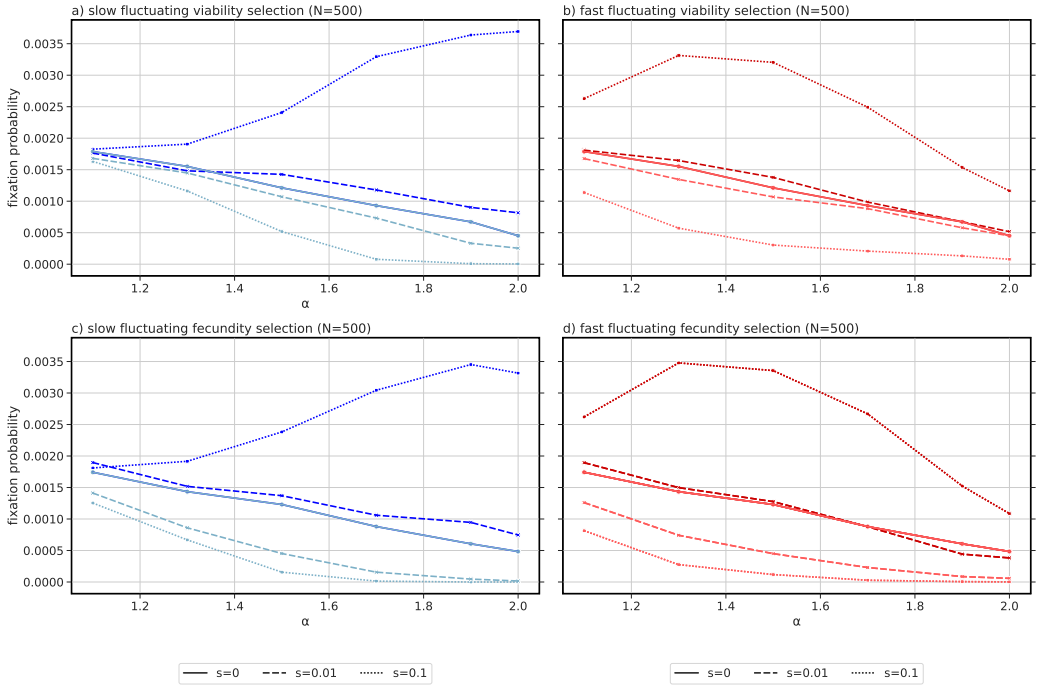


Figure 4.24 Allele fixation probability under constant population size and fluctuating selection coefficients with N=500. Fixation probability under different sweepstake strength (α ranging from 1.1 to 2.0) for (a) slow fluctuating viability selection starting with a positive value of s (dark blue) and starting with a negative value of s (light blue) and (b) fast fluctuating viability selection starting with a positive (dark red) and negative (light red) value of s. The same color code is used for slow (c) and fast (d) fluctuating fecundity selection (F1). Estimates were obtained from 5×10^5 simulations.

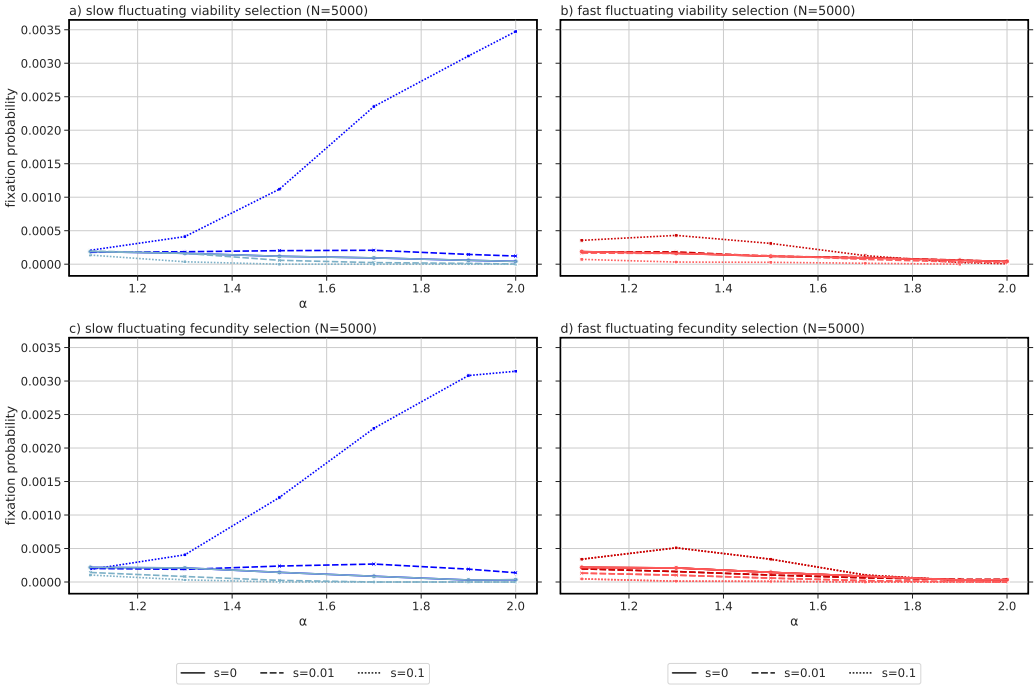


Figure 4.25 Allele fixation probability under constant population size and fluctuating selection coefficients with N=5,000. Fixation probability under different sweepstake strength (α ranging from 1.1 to 2.0) for (a) slow fluctuating viability selection starting with a positive period (dark blue) and starting with a negative period (light blue) and (b) fast fluctuating viability selection starting with a positive (dark red) and negative (light red) period. The same color code is used for slow (c) and fast (d) fluctuating fecundity selection (F1). Estimates were obtained from 5×10^5 simulations.

5 Simultaneous Inference of Demography and Selection under the Beta coalescent from the Ancestral Recombination Graph

The following chapter was under review and on BioRxiv as:
Kevin Korfmann, Thibaut Sellinger, Fabian Freund, Matteo Fumagalli, Aurélien Tellier. Simultaneous Inference of Past Demography and Selection from the Ancestral Recombination Graph under the Beta Coalescent (2023) bioRxiv 2022.09.28.508873;
doi: https://doi.org/10.1101/2022.09.28.508873

KK designed and implemented GNN$coal$, while Thibaut Sellinger (TS) contributed SMβC as shared first-co-author. KK contributed equally to the manuscript, however the original work on the β-Coalescent has been started by TS. Any perceived overlap between the β-Coalescent Chapter of the PhD thesis of TS and our preprint can be attributed to his original work.

A revised version is now published as:
Kevin Korfmann, Thibaut Sellinger, Fabian Freund, Matteo Fumagalli, Aurélien Tellier. Simultaneous Inference of Past Demography and Selection from the Ancestral Recombination Graph under the Beta Coalescent. Peer Community Journal, Volume 4 (2024), article no. e33. doi : https://doi.org/10.24072/pcjournal.397

5.1 Abstract

The reproductive mechanism of a species is a key driver of genome evolution. The standard Wright-Fisher model for the reproduction of individuals in a population assumes that each individual produces a number of offspring negligible compared to the total population size. Yet many species of plants, invertebrates, prokaryotes or fish exhibit neutrally skewed offspring distribution or strong selection events yielding few individuals to produce a number of offspring of up to the same magnitude as the population size. As a result, the genealogy of a

sample is characterized by multiple individuals (more than two) coalescing simultaneously to the same common ancestor. The current methods developed to detect such multiple merger events do not account for complex demographic scenarios or recombination, and require large sample sizes. We tackle these limitations by developing two novel and different approaches to infer multiple merger events from sequence data or the ancestral recombination graph (ARG): a sequentially Markovian coalescent (SMβC) and a graph neural network (GNN$coal$). We first give proof of the accuracy of our methods to estimate the multiple merger parameter and past demographic history using simulated data under the β-coalescent model. Secondly, we show that our approaches can also recover the effect of positive selective sweeps along the genome. Finally, we are able to distinguish skewed offspring distribution from selection while simultaneously inferring the past variation of population size. Our findings stress the aptitude of neural networks to leverage information from the ARG for inference but also the urgent need for more accurate ARG inference approaches.

5.2 Introduction

With the availability of genomes of increasing quality for many species across the tree of life, population genetics models and statistical methods have been developed to recover the past history of a population/species from whole genome sequence data from several individuals (Sheehan and Song, 2016; Li and Durbin, 2011; Schiffels and Durbin, 2014; Speidel et al., 2019; Sellinger et al., 2020; Barroso et al., 2019; Barroso and Dutheil, 2021; Stephan, 2019a; Johri et al., 2022a, 2020). Indeed, the inference of the past demographic history of a species, *i.e.* population expansion, contraction, or bottlenecks, extinction/colonisation, is not only interesting in its own right, but also essential to calibrate genome-wide scans to detect genes under (*e.g.* positive or balancing) selection (Stephan, 2019a; Johri et al., 2021). A common feature of inference methods that make full use of whole genome sequences is the underlying assumption of a Kingman coalescent process (Kingman, 1982) to describe the genealogy distribution of a sample. The Kingman coalescent process and its properties stem from using the traditional forward-in-time Wright-Fisher (WF) model to describe the reproduction mechanism of a population. Besides non-overlapping generations, a key assumption of the neutral WF model is that an individual offspring chooses randomly (*i.e.* uniformly) its parents from the previous generation. More precisely, each chromosome chooses a parental chromosome from the previous generation. Therefore, an essential factor is how many offspring parents can produce. In the Wright-Fisher (WF) model, because of binomial sampling, the distribution of the number of offspring for each parent closely follows a Poisson distribution where both the average and variance are one. This implies that parents will most likely have zero, one, or two offspring individuals, but it is improbable that one parent would have many offspring individuals (*i.e.* on the order of the population size, under the Wright-Fisher haploid model the probability for a parent to have 10 or more offspring is $\approx 10^{-8}$). The

assumption of small variance in offspring distribution between individual parents is realistic for species with low juvenile mortality (so-called type I and II survivorship in ecology), such as mammals.

As genome sequence data become available for a wide variety of species with different biological traits and/or life cycles, the applicability of the Kingman coalescent relying on the WF model can be questioned (Steinruecken et al., 2013; Arnason and Halldorsdottir, 2015; Árnason et al., 2023; Niwa et al., 2016; Kato et al., 2017; Morales-Arce et al., 2020; Tellier and Lemaire, 2014; Menardo et al., 2020; Freund et al., 2022). Indeed, for some species, such as fish, with high fecundity and high juveniles mortality (type III survivorship), it is expected that the variance in reproduction between parents can be much larger than under the Poisson distribution (Tellier and Lemaire, 2014). The phenomenon is commonly referred to as sweepstake reproduction Hedgecock and Pudovkin, 2011; Arnason and Halldorsdottir, 2015. Several neutral mechanisms, including pronounced seed banking Blath et al., 2020, heightened fecundity with an imbalanced offspring distribution Hedgecock and Pudovkin, 2011; Eldon and Wakeley, 2006, intense and repeated bottlenecks Birkner et al., 2008; Casanova et al., 2020, and potent selective forces (specifically positive selection) Durrett and Schweinsberg, 2005; Brunet et al., 2006, 2007; Harris and Jensen, 2020; Árnason et al., 2023, are theoretically identified as factors diverging from the conventional WF model. This divergence is such that the associated genealogies don't align with the Kingman coalescent process. In these scenarios, an alternate set of processes emerges to characterize the genealogy distribution. This set allows for the fusion of multiple individuals and/or the concurrent occurrence of several notable coalescence events Sagitov, 2003; Mohle and Sagitov, 2001; Donnelly and Kurtz, 1999; Sagitov, 1999; Pitman, 1999; Bolthausen and Sznitman, 1998. This group of genealogical processes is widely recognized as the Multiple Merger Coalescent (MMC). MMC models are more biologically appropriate than the Kingman coalescent to study many species of fish (Eldon et al., 2015; Arnason and Halldorsdottir, 2015; Árnason et al., 2023; Hedgecock and Pudovkin, 2011), invertebrates (insects, crustaceans, etc.), viruses (Matuszewski et al., 2017), bacteria (Menardo et al., 2020; Neher and Hallatschek, 2013), plants and their pathogens (Tellier and Lemaire, 2014). While we would like to assess which population model best describes the species genealogy, field experiments to quantify the underlying reproduction mechanism of a species can be costly and time consuming at best, or intractable at worst. Therefore, an alternative solution is to use inference methods based on genome data to identify which model best describes the genealogy of a given species/population.

In this study we use the so-called β-coalescent, a specific class of MMC models. Unlike under the WF model, under MMC models the ploidy level strongly affects the distribution of genealogies (Birkner et al., 2013). For simplicity, in this study we focus on haploid

organisms. The demonstrations shows that when the likelihood of a parent producing k or more descendants aligns with $k^{-\alpha}$, and $1 < \alpha < 2$, the genealogy can be represented using a Λ-coalescent (Schweinsberg, 2003). The latter is a general class of coalescent process describing how and how fast ancestral lineages merge (Pitman, 1999; Sagitov, 1999). When using the Beta(2 α,α) distribution as a probability measure for the Λ-coalescent, the transition rates (*i.e.* coalescent rate) can be analytically obtained leading to the β-coalescent, a specific MMC model.When α approaches 2, the coalescent process transitions to resemble a Kingman coalescent, adjusted by a scaling constant. (Birkner et al., 2013; Koskela, 2018; Koskela and Berenguer, 2019). If α tends to one, the model tends to a Bolthausen-Sznitman coalescent process (*i.e.* dominated by strong multiple merger events) (Bolthausen and Sznitman, 1998). The β-coalescent has the property that the observed polarized Site Frequency Spectrum (SFS) of a sample of single nucleotide polymorphisms (SNPs) exhibits a characteristic U-shape with an excess of rare and high frequency variants (compared to the Kingman coalescent) (Sargsyan and Wakeley, 2008). Current methods to draw inference under MMC models leverage information from the summary statistics extracted from full genome data such as Site Frequency Spectrum (SFS, or derived summary statistics) (Koskela and Berenguer, 2019; Harris and Jensen, 2020; Sackman et al., 2019), minor allele frequency (Rice et al., 2018) or copy number alteration (Kato et al., 2017). It is shown that the SFS is robust to the effect of recombination (Koskela and Berenguer, 2019; Rice et al., 2018) and its shape allows to discriminate between simple demographic models (population expansion or contraction) under the Kingman coalescent and MMC models with constant population size (Koskela and Berenguer, 2019; Koskela, 2018; Eldon et al., 2015). However, methods relying on genome-wide SFS have two main disadvantages. First, in absence of strong prior knowledge, they can suffer from non-identifiability (Johri et al., 2022a) as several complex neutral demographic and/or selective models under the Kingman or MMC models can generate similar SFS distributions. Second, as they summarize the collection of underlying genealogies, they require high sample sizes (>50) to produce trustworthy results (Koskela and Berenguer, 2019; Koskela, 2018; Eldon et al., 2015), relying on experimental designs which are prohibitive for the study of non-model species. To tackle these limitations, we develop two methods that integrate recombination events along the genome in order to leverage more information from full genome data, thus requiring fewer samples.

In species undergoing sexual reproduction, recombination events break the genealogy of a sample at different position of the genome (*i.e.* the genealogy of a sample varies along the genome), leading to what is called the Ancestral Recombination Graph (ARG) (Hudson, 1983; Birkner et al., 2013). Because all the genealogical information is contained in the ARG, in this study we aim at the interpretation of the ARGs to recover model parameters in presence of multiple merger events. With the development of the sequentially Markovian coalescent theory (McVean and Cardin, 2005; Marjoram and Wall, 2006; Wiuf and Hein, 1999), it

becomes tractable to integrate linkage disequilibrium over chromosomes in inferences based on the Kingman coalescent (Li and Durbin, 2011). Hence, we first develop an SMC approach based on the β-coalescent named the Sequentially Markovian β Coalescent (SMβC). The β-coalescent has the additional property that, under recombination, long range dependency can be generated between coalescent trees along the genome if multiple-merger events happen in a single generation (Birkner et al., 2013). In other words, coalescent trees which are located at different places in the genome, and expected to be unlinked from one another (Nelson et al., 2020), would show non-zero correlation in their topology and coalescent times. This is because coalescent trees from different genomic regions may all be affected by the same MMC event (merger event of multiple lineages in the past) which then leaves traces in the genome at several loci (Birkner et al., 2008). To overcome the theoretically predicted non-Markovian property of the distribution of genealogies along the genome under the β-coalescent with recombination (Birkner et al., 2013), we develop a second method based on deep learning (DL) trained from efficient coalescent simulations (Baumdicker et al., 2022). In evolutionary genomics, DL approaches trained by simulations are shown to be powerful inference tools (Sheehan and Song, 2016; Korfmann et al., 2023). Previous work demonstrated that DL approach can help overcome problems mathematically insolvable or computationally intractable in the field of population genetics (Sheehan and Song, 2016; Battey et al., 2020; Wang et al., 2021; Yelmen et al., 2021a; Flagel et al., 2018; Chen et al., 2018; Qin et al., 2022; Burger et al., 2022a; Isildak et al., 2021b). The novelty of our neural network relies on its structure (Graph Neural Network, GNN) and its training algorithm based on the ARG of a sample, or its tree sequence representation (Kelleher et al., 2018). GNNs are an emerging category of DL algorithm (Bronstein et al., 2017; Xu et al., 2019; Cao et al., 2020; Zhou et al.) that benefit by using irregular domain data (*i.e.* graphs). GNNs are designed for the prediction of node features (Kipf and Welling, 2016; Yang et al., 2016), edge features (link prediction) (Zhang and Chen, 2018; Schlichtkrull et al., 2017), or additional properties of entire graphs (Ying et al.; Lee et al., 2018). Therefore, GNNs represent a new tool to address the large dimensionality of ARGs, while simultaneously leveraging information from the genealogy (namely topology and age of coalescent events) as a substantial improvement over convolutions of genotype matrices, as currently done in the field (Sanchez et al., 2021b).

We first quantify the bias of previous SMC methods (MSMC and MSMC2 (Schiffels and Durbin, 2014; Wang et al., 2020)) when performing inference of past population size variation under the β-coalescent. We then describe our two methods, SMβC and GNN$coal$, and demonstrate their statistical power as well as their respective limitations. From simulated tree-sequence (*i.e.* ARG) and sequence (*i.e.* SNPs) data, we assess the accuracy of both approaches to recover the past variation of population size and the α parameter of the Beta-distribution. This parameter indicates how frequent and strong multiple merger events occur.

We demonstrate that our approaches can infer the evolutionary mechanism responsible for multiple merger events and distinguish local selection events from genome-wide effects of multiple mergers. We highlight the limits of the Markovian property of SMC to describe data generated under the β-coalescent. Finally, we show that both our approaches can model and identify the presence of selection along the genome while simultaneously accounting for non-constant population size, recombination, and skewed offspring distribution. Thus our methods represents a major and necessary leap forward in the field of population genetic inferences.

5.3 Materials and Methods

In our study we first assume the true ARG to be known. Hence, the ARG of the sample is given as input to our methods to estimate recover model parameters of interest (*e.g.* the α parameter and/or the past variation of population size). We then show the applicability of our methods by using as input simulated sequence data (*i.e.* SNPs) and/or ARG inferred using ARGweaver (Rasmussen et al., 2014a) from simulated sequence data.

5.3.1 SMC-based method

In this study, we use different SMC-based algorithms: two previously published, MSMC and MSMC2 (Schiffels and Durbin, 2014; Wang et al., 2020), and the new SMβC. In the latter, the software backbone stems from our previous eSMC (Sellinger et al., 2020, 2021a) whilst the theoretical framework originates from the MSMC algorithm (Schiffels and Durbin, 2014) (see Supplementary Text 5.9). All approaches can use the ARG or sequence data as input. Giving ARG as input for MSMC an MSMC2 is enabled by a re-implementation included in the R package eSMC2 (Sellinger et al., 2021a). The MSMC2 algorithm focuses on the coalescence time between two haploid samples along the genome. In the event of recombination, there is a break in the current genealogy and the coalescence time consequently takes a new value. A detailed description of the algorithm can be found in (et al, 2016; Wang et al., 2020). The MSMC algorithm simultaneously analyses multiple sequences (up to 10) and follows the distribution of the first coalescence event in a sample of size $n > 2$ along the sequence based on the Kingman coalescent (Kingman, 1982). A detailed description of MSMC can be found in (Schiffels and Durbin, 2014).

Our new approach, SMβC, is a theoretical extension of the MSMC algorithm, simultaneously analyzing multiple haploid sequences and focusing on the first coalescence event of a sample size 3 or 4. The SMβC follows the distribution of the first coalescence event of a sample along sequences assuming a β-coalescent process. Therefore, our SMβC allows for more than two ancestral lineages to join the first coalescence event, or new lineages to join

an already existing binary (or triple) coalescent event. Hence, the SMβC extends the MSMC theoretical framework by adding hidden states at which more than two lineages coalesce. Currently, the SMβC has been derived to analyze for up to 4 sequences simultaneously (due to computational load and mathematical complexity). The emission matrix is similar to the one of MSMC. As in the MSMC software, the population size is assumed piece-wise constant in time and we discretize time in 40 bins throughout this study. A detailed description of SMβC can be found in Supplementary Text 5.9. To test and validate the theoretical accuracy of our approach, we first study its best case convergence (introduced in Sellinger et al., 2021a) which corresponds to the model's performance when the true genealogy is given as input, *i.e.* as if the hidden states are known. Additionally, we also validate the practical accuracy of the SMβC on simulated sequence data taking the same input as the MSMC software (Schiffels and Durbin, 2014), or using the inferred ARGs by ARGweaver (Rasmussen et al., 2014a). All SMC approaches used in this manuscript are found in the R package eSMC2 (`https://github.com/TPPSellinger/eSMC2`).

5.3.2 GNN$coal$ method

Inspired by results obtained from inferences based on tree sequence data (Gattepaille et al., 2016; Sellinger et al., 2021a), we develop a graph neural network (GNN) taking tree sequence data as input. Our GNN is designed to infer population size along with the α parameter of the Beta distribution describing the distribution of offspring production. In practice, the ARG is reshaped into a sequence of genealogies (more precisely a sequence of undirected graphs), and then given as input to the GNN (similar to what is described above for the SMβC). In our analyses, we fixed the batch size to 500. This value represents the number coalescence trees being processed before updating parameters of the neural network. As batch size is fixed to 500, only simulations displaying at least 500 recombination events are considered for the training data sets. If more than 500 recombination events occur along the sequence, the ARG is truncated and the GNN will only take as input the first 500 genealogies and remove the rest. Thanks to the GNN architecture, the algorithm can account for the topology of the genealogy. Hence, the GNN leverages information from coalescence time and branch lengths but also from the topology of the ARG. This operation is known as a graph convolution. By doing so, the GNN is capable of learning from local features of the ARG and extract information from its complex structure. To learn from global genealogy patterns (which SMC-based methods cannot do), an additional pooling strategy is implemented as part of the network (Ying et al.). To do so, the ARG is broken into smaller ARGs (*i.e.* subgraphs) during the forward-pass step. To illustrate the GNN strategy, we visualize the compression-like process, from the coalescent trees (1) being processed by GNN$coal$ (2,3) to the inferred variable of interest (4, 5) in Figure 5.1.

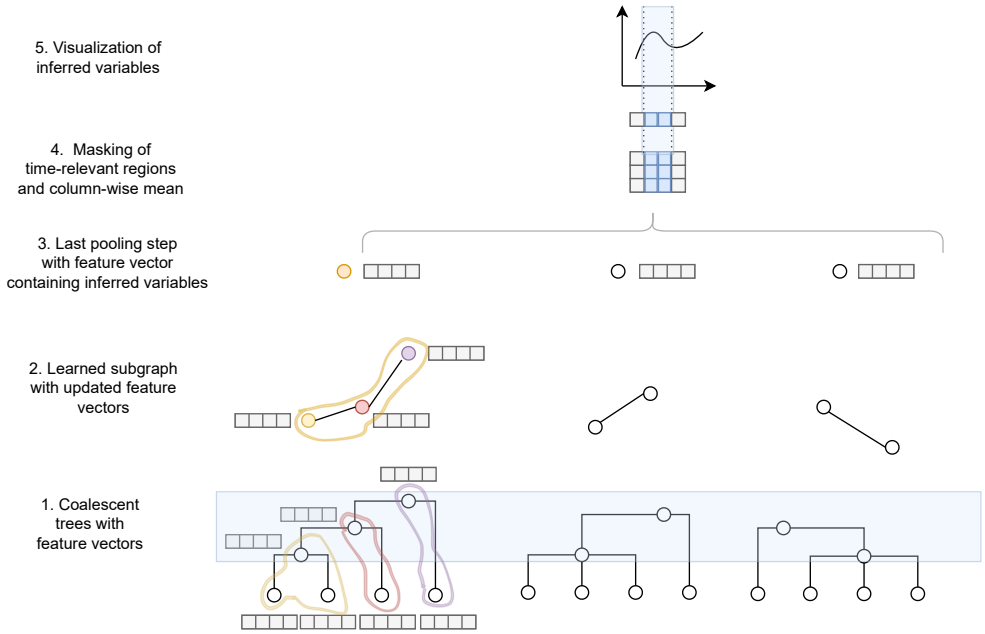


Figure 5.1 Schematic representation of GNN*coal* processing an ARG Hierarchical pooling compression of a sequence of coalescent trees into the inferred variable of interest (*i.e.* demographic changes) using graph convolutions. Each coalescent ancestor or leaf node is initialized by a feature vector (light grey boxes) (1). Sub-graphs are generated by a pooling network with updated feature vectors and a final compression step is performed until ideally one node per graph remains (2-3). Lastly, the column-wise mean is taken after applying a time mask (blue - based on number of coalescent events), so that single feature vector remains (4-5). Detailed description of the graph convolution, feature vector initialization, pooling methodology, coalescent time mask construction, and dataset generation can be found in Supplementary Text 5.10.

To infer parameters from our neural network, we need to define an objective function to be optimized. We use a masked root-mean-squared error (RMSE) loss function as objective function which is computed for each inputted ARG (*i.e.* minimizing the average square difference between predicted and true parameter value). In practice, time is discretized (as for the SMβC) and time windows are defined. The true α value and true demography at 60 predefined time points are given as input to the GNN to compute the loss function. The GNN captures the stochastic complexity arising from the underlying demographic scenario and model parameters. Furthermore, our algorithm naturally defines an appropriate time window to have sufficient observation at each time point. A more detailed description of the GNN*coal* can be found in Supplementary Text 5.10. The code of the model architecture is implemented in *Pytorch* (Paszke et al., 2017) using the extension *Pytorch Geometric* (Fey and Lenssen). The model is available with the simulated training dataset at `https://github.com/kevinkorfmann/GNNcoal` and `https://github.com/kevinkorfmann/GNNcoal-analysis`.

5.3.3 ARGweaver

As the ARG is not known in practice, it needs to be inferred from sequence data. ARGweaver displays the best performance at recovering the ARG from whole genome polymorphism data at the sample sizes employed in this study (*i.e.* $\ll 50$) (Rasmussen et al., 2014a; Brandt et al., 2022). Briefly, ARGweaver samples the ARG of n chromosomes/scaffolds conditional on the ARG of $n-1$ chromosomes/scaffolds. To this aim, ARGweaver relies on hidden Markov models while assuming a sequentially Markov coalescent process and a discretization of time, similarly to the SMC-based methods previously described. For a more detail description of the algorithm, we refer the reader to the supplementary material of (Rasmussen et al., 2014a).

5.3.4 Simulation of data

Validation dataset for both methods The ARG is given as input to the DL approach and the SMβC (see (Sellinger et al., 2021a)). We use msprime (Baumdicker et al., 2022) to simulate the ARG of a sample (individuals are assumed to be haploid) under the β-coalescent based on (Schweinsberg, 2003; Birkner et al., 2013) or under the Kingman coalescent (under neutrality or selection). We simulate 10 sequences of 100 Mbp under five different demographic scenarios: 1) Constant population size; 2) Bottleneck with sudden decrease of the population size by a factor 10 followed by a sudden increase of population by a factor 10; 3) Expansion with sudden increase of the population size by a factor 10, 4) Contraction with sudden decrease of the population size by a factor 10; and 5) "Saw-tooth" with successive exponential decreases and increases of population size through time, resulting in continuous population size variation (as shown in (Terhorst et al., 2017; Schiffels and Durbin, 2014; Sellinger et al., 2021a)). We simulate data under different α values (*i.e.* parameters of the β-distribution) including values of 1.9 (almost no multiple merger events), 1.7, 1.5, and 1.3 (frequent and strong multiple merger events). Mutation and recombination rate (respectively μ and r) are set to 10^{-8} per generation per bp in order to obtain the best compromise between realistic values and number of SNPs. When specified, some specific scenarios assume recombination and mutation rate set to produce sufficient data or to avoid violation of the finite site hypothesis. All python scripts used to simulate data sets are available at `https://github.com/kevinkorfmann/GNNcoal-analysis`.

Additionally, to generate sequence data, we simulate 10 sequences of 10 Mbp under the five different demographic scenarios described above and for the same α values. For each scenario, 10 replicates are simulated. In order to obtain sufficient SNPs for inference, we simulate sequence data with mutation and recombination rate (respectively μ and r) of 10^{-8} per generation per bp when α is set to 1.9 and 1.7, 10^{-7} per generation per bp when α is set to 1.5, and 10^{-6} per generation per bp when α is set to 1.3.

Training dataset for the GNN$coal$

In our study we train two GNNs, one to infer past variation of population size through time along with α, and one for model selection. The training dataset used for both GNNs is described below.

Training dataset for the GNN inferring α and demography

We generate an extensive number of ARGs to train our GNN. The ARGs are simulated under many demographic scenarios and α values. The model parameters are updated in supervised manner. The loss function is calculated for each batch with respect to how much the machine-learning estimates differ from to the true parameters used for simulation. The simulations strategy to recover past demographic history is based on the strategy described and used in (Boitard et al., 2016; Sanchez et al., 2021b). The idea of this approach is to generate a representative set of demographic scenarios over which the network generalizes to consequently infer similar demographic changes after training. More details on the training strategy can be found in Supplementary Text 5.10.

To improve the outputted demographic history, we introduce a smoothing of the demography allowing to infer continuous variation of population size through time. We do so by interpolating I time points cubically, and choosing w (set to 60) uniformly spaced new time points of the interpolation in log space. All time points more recent than ten generations in the past are discarded, since inference is too imprecise in the very recent present under our models. An example of this process can be seen in Supplementary Text 5.10.

Training dataset to disentangling coalescent and selection signatures

Beyond parameter inference, deep learning approaches can also be used for clustering. Hence, we train a GNN to disentangle between different scenarios and models. In total, we define eight classes, namely K (S0) (Kingman, no selection), K (WS) (Kingman, weak selection), K (MS) (Kingman, medium selection), K (SS) (Kingman, strong selection) and four different β-coalescent classes (]2.0-1.75],]1.75-1.5],]1.5-1-1.25],]1.25-1.01]) without selection. The three different selection regimes are defined, corresponding to $Ne \times s$ in: [0.1, 0.01[for SS, [0.01, 0.001[for MS, [0.001, 0.0001[for WS and [0] for absence of selection. Demography is kept constant and set to 10^5 individuals and sequence length is set to 10^5 bp. The simulation is discarded if it resulted in less than 2,000 obtained trees and is rerun with twice the sequence length until the tree number required is satisfied. This procedure avoids simulating large genome segments of which only a small fraction of trees is used for the given scenario during training and inference. The selection site is introduced in the centre of the respective sequence, so that 249 trees left and 250 right of the middle tree under selection

form a training sample, using 500 trees for each sample. One hundred replicates are generated for each training sample. The complete training dataset consists of 1,000 parameter sets, 500 for the Kingman cases and 500 for the β-coalescent cases, with approximately 125 parameter sets per class. The model itself is trained on one epoch (number of time the data is analyzed), and the evaluation performed afterwards on 1,000 randomly generated parameter sets, with one replicate per parameter set. The same architecture used for demography estimation is employed with additional linear layers to reduce the number of output dimensions from 60 to 8. The loss function is set to a Cross-Entropy-Loss for the network to be trainable for categorical labels. Otherwise all architecture and training parameters is the same as described above and detailed in Supplementary Text 5.10.

5.4 Results

5.4.1 Inference bias under the wrongly assumed Kingman coalescent

We first study the effect of assuming a Kingman coalescent when the underlying true model is a β-coalescent (*i.e.* in presence of multiple merger events) by applying MSMC and MSMC2 to our simulated data. The inference results from MSMC and MSMC2 when the population undergoes a sawtooth demographic scenario are displayed in Figure 5.2. For $\alpha > 1.5$ the shape of the past demography is fairly well recovered. Decreasing the parameter α of the β-coalescent (*i.e.* higher probability of multiple merger events occurring) increases the variance of inferences and flattens the demography. Yet, both methods fail to correctly infer the correct population size, due to the scaling discrepancy between the Kingman and β-coalescent. Hence, we perform the same analysis and correct for the scaling effect of the MMC versus a Kingman coalescent in order to better capture the specific effects of assuming binary mergers only. The results are displayed in Figure 5.11. For $\alpha > 1.5$ the demography is accurately recovered providing we know the true value of α to adjust the y-axis (population size) scale. However, for smaller α values the observed variance is extremely high and a flattened past variation of population size is observed.

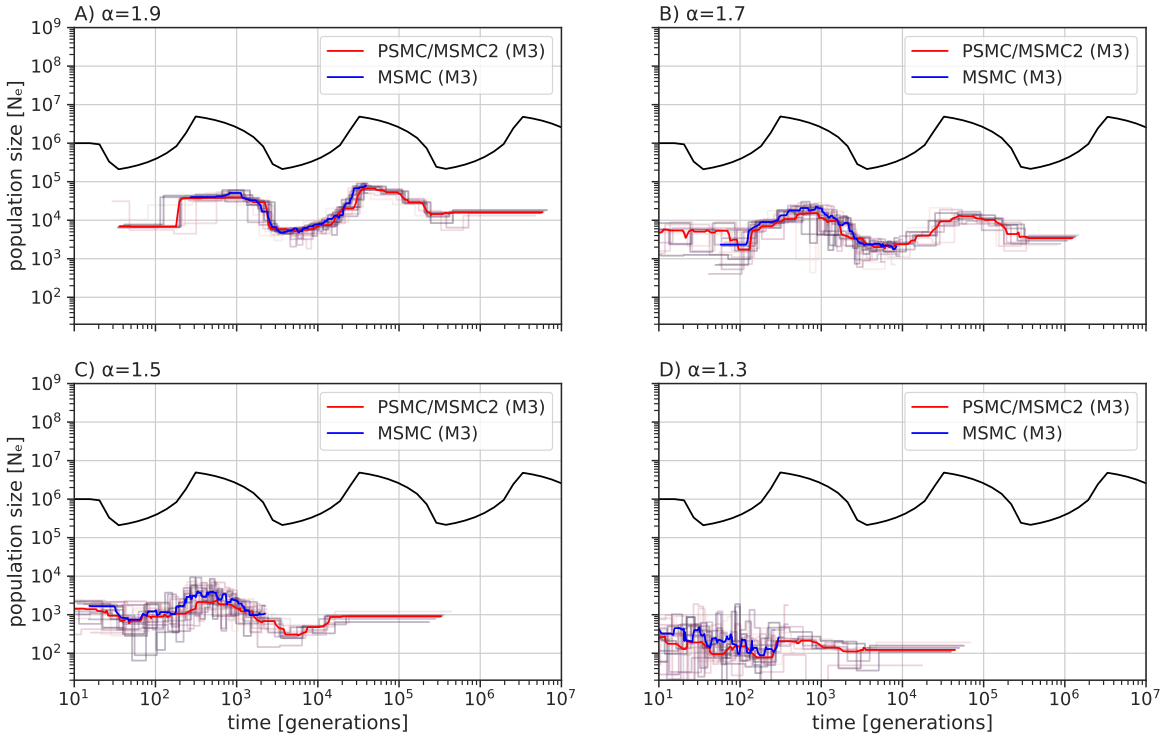


Figure 5.2 Performance of MSMC and MSMC2 under a β-coalescent. Averaged estimated demographic history by MSMC (blue) and MSMC2 (red) based on 10 sequences of 100 Mb with $\mu = r = 10^{-8}$ per generation per bp over ten repetitions (while analyzing simultaneously 3 sequences, noted by M=3). Each repetition result is represented in light red (PSMC/MSMC2) or in light blue (MSMC). Population undergoes a sawtooth demographic scenario (black) for A) $\alpha = 1.9$, B) $\alpha = 1.7$, C) $\alpha = 1.5$, and D) $\alpha = 1.3$.

5.4.2 The limit of the Markovian hypothesis

As SMC approaches rely on the hypothesis of Markovian change in genealogy along the genome, we study the effect of α on the linkage disequilibrium of pairs of SNPs (r^2, (Rogers and Huff; Miles et al.)) in data simulated under the Kingman Coalescent or the β-coalescent (with $\alpha = 1.5$ and $\alpha = 1.3$) and constant population size (Figure 5.3). Linkage monotonously decreases with distance under the Kingman coalescent. Under the β-coalescent a similar distribution is observed but with higher amount of LD. We find a higher variance in LD for smaller α values. More precisely, this increased variance results in the occurrence of high spikes of linkage disequilibrium along the genome (*e.g.* Figure 5.3 B). This stochastic increase of linkage along the genome demonstrates that the Markovian hypothesis used to model genealogies along the genome is violated under the β-coalescent due to the long range effect of strong multiple merger events (Birkner et al., 2013).

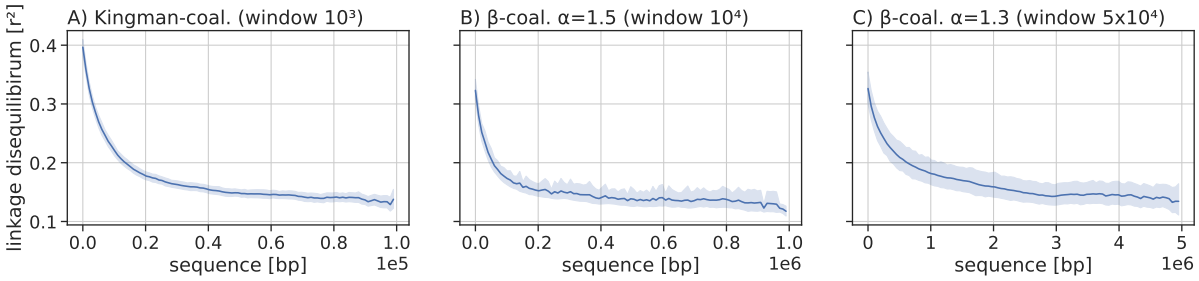


Figure 5.3 Linkage disequilibrium under a Kingman and β-coalescent. Pairwise linkage disequilibrium between SNPs (r^2) under a Kingman and β-coalescent with $\alpha = 1.5$ and $\alpha = 1.3$ using 50 sequences of length A) 0.1 Mb, B) 1Mb, and C) 5Mb. The population size is constant at $N = 10^4$ for the Kingman model and $N = 10^6$ for the β-coalescent, with $\mu = 2 \times 10^{-7}$ and $r = 2 \times 10^{-8}$ per generation per bp. For each LD analysis, the linkage disequilibrium is calculated by averaging it over a window of size 10^3, 10^4 and 5×10^4 bp respectively in A), B) and C).

We further investigate the effect of multiple merger events on linkage disequilibrium. To do so, we first assume an SMC framework (*e.g.* MSMC2 or eSMC) to predict the transition matrix (*i.e.* matrix containing the probabilities for the coalescent time to change to another value along the genome) and looking at the absolute difference between the observed transition events. Under the Kingman coalescent, the distribution of coalescent times between two positions in a sample of size two ($n = 2$) is well approximated by the SMC as shown in Figure 5.12 (*i.e.* absence of structured difference between observed and predicted). However, under the β-coalescent (with $\alpha = 1.3$) we observe significant and structured differences between observed and predicted at times points where multiple merger events occur (Figure 5.13). In practice multiple merger events do not occur at each time point (as they remain rare events), unveiling a discrepancy between the expectation from the SMC (*i.e.* approximating the distribution of genealogies along the genome by a Markov chain) and the simulated data. This discrepancy does not stem from the simulator, because it correctly generates ARG under the β-coalescent model (Birkner et al., 2013; Baumdicker et al., 2022), but from the limits of the SMC approximation to model events with long range effects on the ARG (Figure 5.13).

5.4.3 Inferring α and past demography on ARG

To test if our two approaches (GNN*coal* and SMβC) can recover the past variation of population size and the α parameter, we run both methods on simulated tree sequences under different α values and demographic scenarios. Figure 5.4 displays results for data simulated under a sawtooth past demography and for α ranging from 1.9, 1.7, 1.5 to 1.3. In all cases, the DL approach exhibits high accuracy and low variance to infer the variation of population size. For high α values (>1.5), the shape of population size variation is well recovered by SMβC. However, for smaller values, an extremely high variance is observed which demonstrates the

limits of SMC inferences. On average, both approaches seem to recover fairly well the true α value (Figure 5.4 and Table 5.3), especially GNN*coal* which performs well by displaying high accuracy and lower standard deviation. We note that the variance in the estimation of α increases with diminishing α value. Moreover, increasing the number of simultaneously analyzed sequences by SMβC does not seem to improve the inferred α value (Table 5.3). These conclusions are also valid for the results in Figures 5.14-5.17 and Table 5.3 based on inference under four additional demographic scenarios: constant population size, bottleneck, sudden increase and sudden decrease of population size.

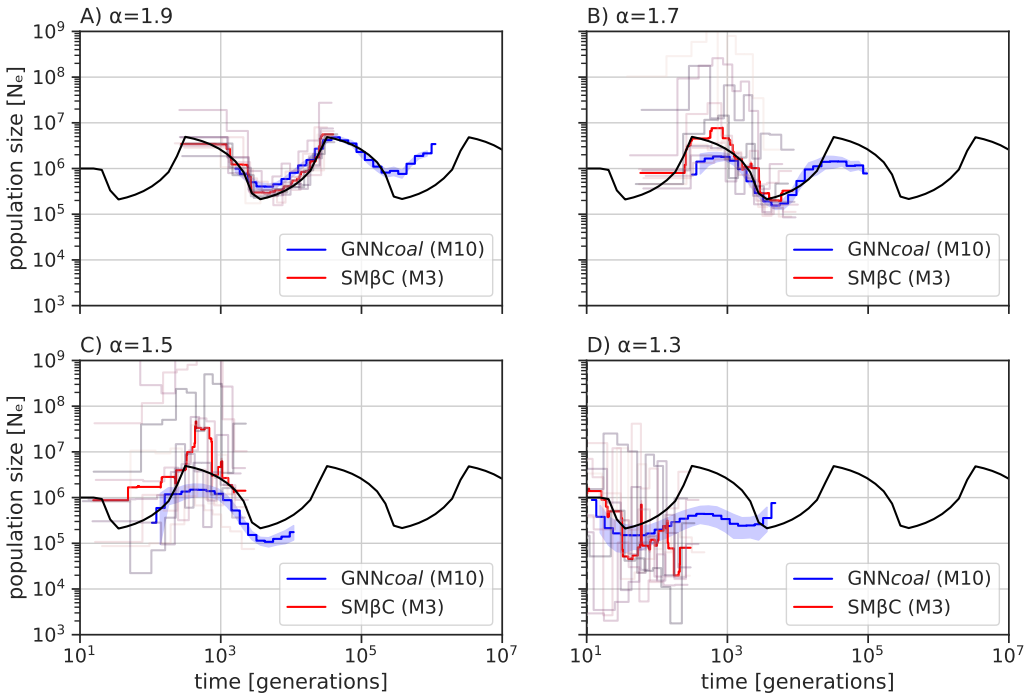


Figure 5.4 Best-case convergence estimations of SMβC and GNN*coal* under a β-coalescent. Estimations of past demographic history by SMβC in red (median) and by GNN*coal* in blue (mean and 95% confidence interval, CI95; while analyzing simultaneously 3 or 10 sequences, noted by M=3 or M=10) when population undergoes a sawtooth demographic scenario (black) under A) $\alpha = 1.9$, B) $\alpha = 1.7$, C) $\alpha = 1.5$ and D) $\alpha = 1.3$. SMβC runs on 10 sequences and 100 Mb, GNN*coal* runs on 10 sequences and 500 trees, and $\mu = r = 10^{-8}$ per generation per bp.

Because when α diminishes, the effective population size decreases as well (*i.e.* the population size calculated from SNPs assuming a Wright-Fisher model), the number of recombination events plummets for small values of $\alpha < 1.5$. To demonstrate the theoretical accuracy of the SMβC, *i.e.* the convergence to the correct values, we run SMβC on data simulated with mutation and recombination rate fifty times higher under similar scenarios as in Figure 5.4, namely to increase the amount of data (SNPs and number of independent coalescent trees by recombination). Results of SMβC for α values of 1.7, 1.5 and 1.3 are

displayed on Table 5.4, where it appears that that SMβC can recover α with higher accuracy when given more data (*i.e.* recombination events).

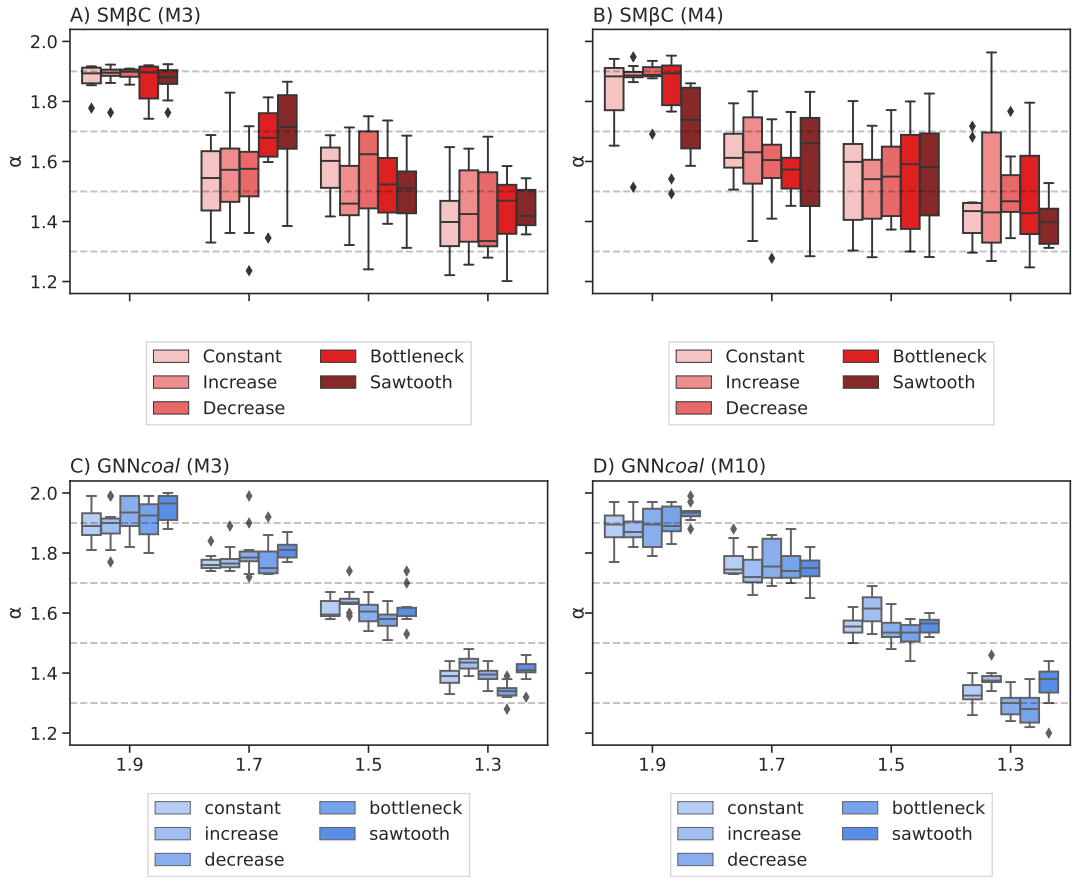


Figure 5.5 Estimated values of α by SMβC and GNN*coal* over ten repetitions using 10 sequences of 100 Mb with $\mu = r = 10^{-8}$ per generation per bp under a β-coalescent process (with different α parameter). The analysis are run on five different demographic scenarios (Constant population size, Bottleneck, Sudden increase, Sudden decrease and a Sawtooth demography) using a sample size $n = 3$ for A) and C), $n = 4$ for B), and $n = 10$ for D). Grey dashed lines indicate the true α values.

Although 10 sequences are given to the SMβC in the previous analyses, the method can only analyze three or four simultaneously. However, the GNN*coal* approach can simultaneously analyze 10 sequences, that is the whole simulated ARG. As we observe that GNN*coal* has a higher performance than SMβC we wish to test whether the GNN*coal* better leverages information from the ARG or whether it benefits from simultaneously analyzing bigger sample size. Thus, we run GNN*coal* on the same dataset, but downsampling the coalescent trees to a sample size three. Results are displayed in Figure 5.14 to Figure 5.17. Results with sample size three of the GNN*coal* approach are similar to results with sample size 10, demonstrating that the GNNs can better leverage information from the ARG in presence of multiple merger events (Figure 5.18).

Additionally, we test if both approaches can recover a Kingman coalescent from the ARG when data are simulated under the Kingman coalescent, namely both approach should recover $\alpha = 2$. To do so, we simulate the same five demographic scenarios as above under a Kingman coalescent and infer the α parameter along with the past variation of population size. Estimations of α values are provided in Table 5.1 and are systematically higher than 1.85, suggesting mostly binary mergers (*i.e* an underlying Kingman coalescent process). The associated inferred demographies are shown in Figures S9-S13. Both approaches correctly infer the past demographic shape up to the scaling discrepancy between the Beta and the Kingman coalescent mentioned above. Furthermore, we notice that the scaling effect only affects the y-axis for the SMβC but affect both axes for GNN$coal$.

As GNN$coal$ was not trained on data simulated under the Kingman coalescent (especially with such high population size), some events fall beyond the scope of the GNN (due to the scaling discrepancy between the Beta and Kingman coalescence). Hence, we run our GNN$coal$ on data simulated under the Kingman coalescent but with smaller population size (scaled down by a factor 100) to assure that all events fall within the scope of the GNN. Values of α inferred by the GNN$coal$ and the SMβC under the five demographic scenarios are available in Table 5.5. The associated inference of population size are plotted in Figure 5.19-5.23. Both approaches recover high α values (*i.e.*>1.85) suggesting a genealogy with almost exclusively binary mergers. In addition, both approaches accurately recover the shape of the past variation of population size up to a scaling constant (but only on the population size y-axis).

5.4.4 Inferring α and past demography from simulated sequence data

We first investigate results for both GNNcoal and SMβC when the ARG is reconstructed with ARGweaver (Rasmussen et al., 2014a), which is currently the best ARG inference approach for sample size smaller than 20 (Brandt et al., 2022). Demographic inference results by both approaches are displayed in Figure 5.11.1, and α inference results in Table 5.6. GNN$coal$ does not manage to recover the shape of the demographic history from the inferred ARGs and fails to recover α by highly overestimating it. The α inferences of SMβC (when giving the inferred ARG as input) are better than GNN$coal$ and hints toward the presence or absence of multiple merger events although variance is high. SMβC recovers the shape of the past variation of population size for $\alpha > 1.3$ but displays extremely high variance for $\alpha = 1.3$.
Second, we run SMβC on simulated sequence data and found that α is well recovered on average (Table 5.2) and results are similar to what is obtained when the true ARG is given. Concerning inference of past variation of population size, the shape of the past variation of population size is well recovered under the sawtooth demographic scenario for $\alpha > 1.3$

(Figure 5.25). In the other four scenarios, the shape of the demography is recovered in recent times, but underestimated in the past (Figure 5.26). Furthermore, as found above from inputted ARG, the variance generally increases with diminishing α.

5.4.5 Inferring MMC and accounting for selection

As specific reproductive mechanisms and selection can lead to the occurrence of multiple merger-like events, we train our neural network on data simulated under the β-coalescent, and under the Kingman coalescent in presence or absence of selection. We then use the trained GNN$coal$ to cluster and determine if multiple merger events originate from skewed offspring distribution or positive selection, or if the data follows a neutral Kingman coalescent process. The classification results are displayed in Figure 5.6 in the form of a confusion matrix, that is the percentage of times the GNN$coal$ correctly assigns the true model shown on the diagonal (evaluated on a test dataset of size 1,000 of known ARGs). Our approach can accurately select the model except in two cases. Our approach shows limited power to distinguish strong from average selection under the Kingman coalescent, as well as to distinguish the β-coalescent with a small amount of multiple merger events (*i.e.* α>1.75) from the β-coalescent case with 1.75>α>1.5.

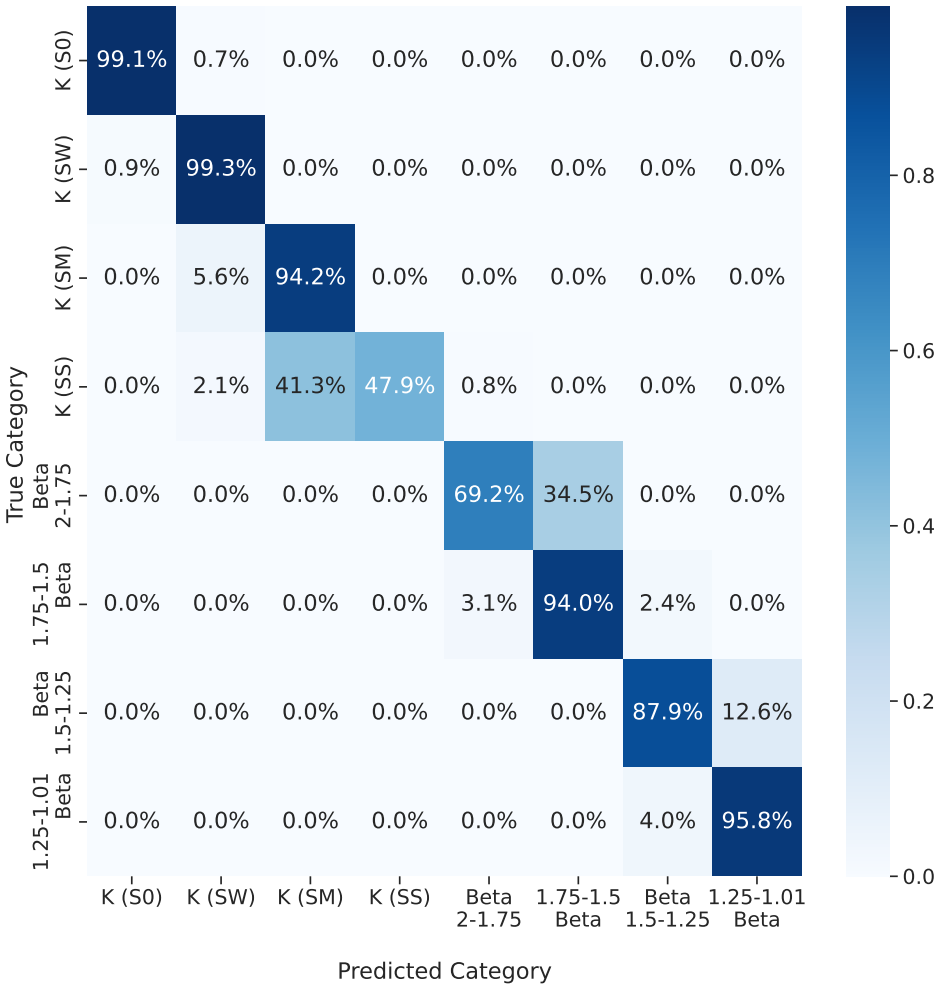


Figure 5.6 Confusion matrix for Kingman and β-coalescent classification model under varying selection coefficients. Evaluation of classification accuracy for Kingman (K) and β-coalescent (B) for no selection (S0), weak selection (SW), medium selection (SM) and strong selection (SS) using a 1,000 repetition validation dataset. Population size was kept constant at $N = 10^5$ individuals, using a sample size $n = 10$ and $r = 10^{-8}$ per bp per generation.

To assess the effect of selection, we infer the α along the genome with both approaches from data simulated with strong positive selection or neutrality under a Kingman coalescent with population size being constant through time. SMβC infers α on windows of 10kbp along the genome, and GNN*coal* infers α every 20 trees along the genome. Results for the GNN*coal* approach and for the SMβC are displayed in Figure 5.7. Both approaches recover smaller α value around the locus under strong selection. However under neutrality or weak selection, inferred α values remain high (>1.6) along the genome.

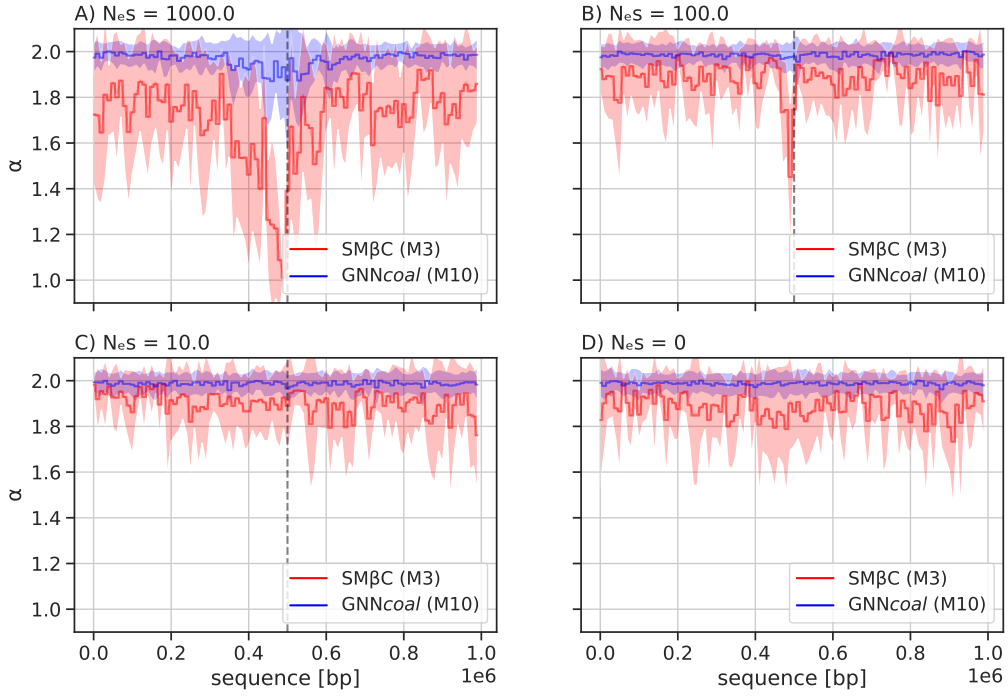


Figure 5.7 Averaged estimations by GNN*coal* and SMβC under selection Estimations of α along the genome by the GNN*coal* approach and the SMβC when population undergoes as strong positive selective sweep event (at position 0.5 Mb) under different strengths of selection: A) $s = 0.01$, B)$s = 0.001$, C) $s = 0.0001$, and D) $s = 0$ meaning neutrality (mean and standard deviation for both methods). The population size is constant and set to $N = 10^5$ with $\mu = r = 10^{-8}$ per generation per bp. We hence have in A) $N_e \times s = 1000$,B) $N_e \times s = 100$, C) $N_e \times s = 10$ and D) $N_e \times s = 0$. SMβC uses 20 sequences of 1Mb (red) and GNN*coal* uses 10 sequences through down-sampling the sample nodes (blue)

Similarly, we run both approaches on data simulated under the β-coalescent (assuming neutrality) and we infer the α value along the genome. Inferred α values by both approaches are plotted in Figure 5.27. GNN*coal* is able to recover the α value along the genome (also with overestimation due to tree sparsity), but the SMβC systematically underestimate these α values. However, unlike in presence of positive selection at a given locus, the inferred α values are found in all cases to be fairly similar along the genome.

We finally simulate data under a strong selective sweeps or under neutrality and under a sawtooth demographic scenario. Under neutrality, our approach recovers high α value along the genome and can accurately recover the past variation of population size (only up to a scaling constant for GNN*coal*, since it was only trained on the β-coalescent) (Figure 5.8). Similarly, when the simulated data contains strong selection, a small α value is recovered at the locus under selection and the past variation of population size is accurately recovered, albeit with a small underestimation of population size in recent times for SMβC (Figure 5.8).

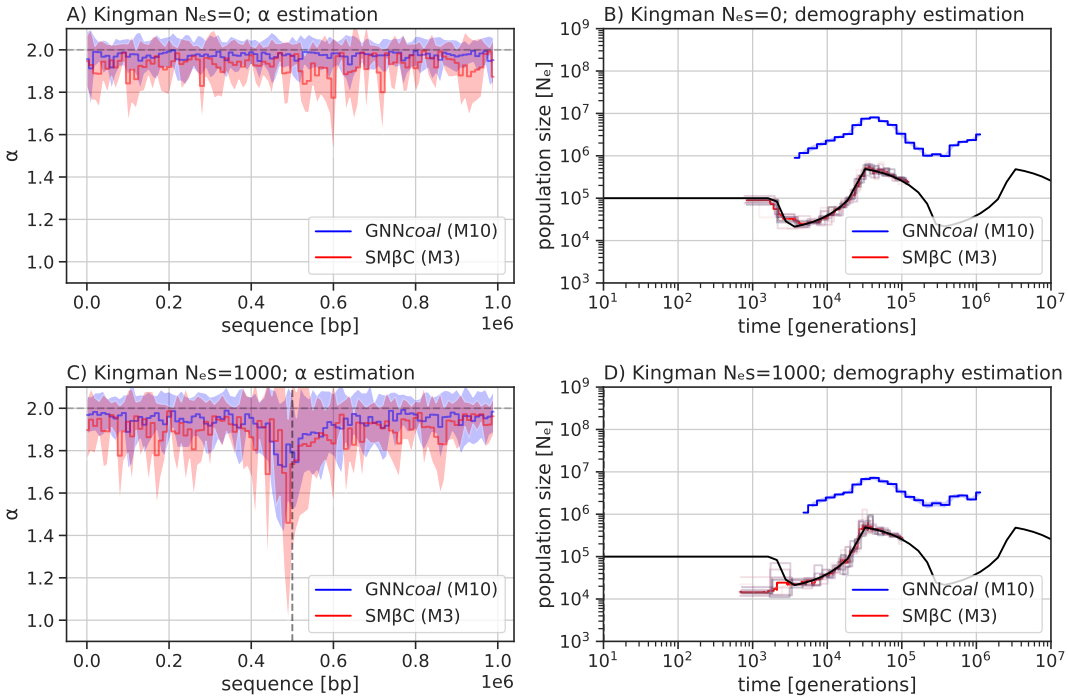


Figure 5.8 Simultaneous estimations of α along the sequence under demographic change by GNN*coal* and SMβC. Simultaneous estimation of α along the genome under a partial sawtooth scenario: A) and B) in the absence of selection (mean and standard deviation for both methods), and C) and D) presence of selection with $N_e S = 1,000$ (mean and CI95 for GNN*coal* and median for SMβC). SMβC uses 20 sequences of 1Mb (red) and GNN*coal* uses 10 sequences through downsampling the sample nodes (blue), and $\mu = r = 10^8$ per generation per bp.

5.5 Discussion

With the rise of SMC approaches (Li and Durbin, 2011), most current methods leverage information from whole genome sequences by simultaneously reconstructing a portion of the ARG in order to infer past demographic history (Li and Durbin, 2011; Schiffels and Durbin, 2014; Terhorst et al., 2017; Upadhya and Steinrücken, 2021), migration rates (Kim et al., 2020; Wang et al., 2020), variation in recombination and mutation along the genome (Barroso et al., 2019; Barroso and Dutheil, 2021), as well as ecological life history traits such as selfing or seed banking (Sellinger et al., 2020; Struett et al., 2022). However in previous studies authors discussed uncoupling both steps, namely reconstructing the ARG and then inferring parameters from its distribution (Sellinger et al., 2021a; Gattepaille et al., 2016; Rasmussen et al., 2014a). Indeed, recent efforts have been made to improve approaches recovering the ARG (Speidel et al., 2019; Kelleher et al., 2019; Hubisz and Siepel, 2020; Rasmussen et al., 2014a; Mahmoudi et al., 2022a; Brandt et al., 2022), as well as its interpretation (Gattepaille et al., 2013; Sellinger et al., 2021a). Our results on data simulated under the β-coalescent clearly show the strong effect of multiple merger events on the topology and branch length of the ARG. We find that the more multiple merger events occur, the more information

concerning the past demography is lost. Both methods, whether given sequence data, the true or inferred ARG, can recover the α parameter and the variation of past population size for α values high enough (*i.e.* $\alpha > 1.3$), however for lower values of α a larger amount of data is necessary for any inference (which becomes nearly impossible when α tends to one). Both approaches can also recover the Kingman coalescent (*i.e.* α>1.8). We find the GNN$coal$ approach outperforms the SMβC approach in almost all cases when given the true ARG, and one can use GNN$coal$ to disentangle (classify) between models of β-coalescent and Kingman models with selection.

Overall, our results provide a qualitative increase in the developments of inference methods for models with multiple merger events, a key step to understand the underlying reproduction mechanism of a species. We first develop a new SMC method which takes simultaneously three or four sequences to draw inference of the coalescent tree and transitions along the genome. While fairly accurate, SMβC is outperformed by GNN$coal$ when given true ARGs as input. As we directly compare our theoretical SMC to the GNN based on the same input data (coalescent trees), we are ideally placed to dissect the mechanisms underlying the power of the GNN$coal$ method. We identify four main reasons for the difference in accuracy between both methods in this specific case. First, the SMβC approach suffers from the limit of the sequential Markovian coalescent hypothesis along the genome when dealing with strong multiple merger events (Birkner et al., 2013; Casanova et al., 2020). Second, most current SMC approach rely on a discretization of the coalescent times into discrete hidden states, meaning that simultaneous mergers of three lineages may not be easily distinguished from two consecutive binary mergers occurring over a short period of time. Third, the SMβC relies on a complex hidden Markov model and due to computational and mathematical tractability, it cannot leverage information on a whole genealogy (as MSMC, SMβC only focuses on the first coalescent event) and cannot simultaneously analyze large sample size. Furthermore, the SMβC approach leverages information from the distribution of genealogies along the genome. It means that in the near absence of recombination events, both approaches cannot utilize any information from the genealogy itself. However, the GNN$coal$ approach can overcome this by increasing sample size. Fourth, the SMβC is based on a coalescent model where α is constant in time. Yet multiple merger events do not appear regularly across the genealogical timescale, but occur at few random time points. Hence, the SMC approach suffers from a strong identifiability problem between the variation of population size and the α parameter (for low α values). As for example, if during one hidden state one strong multiple merger event occurs, multiple merger events are seldom observed and SMβC may rather assume a small population size at this time point (hidden state). This may explain the high variance of inferred population sizes under the β-coalescent.

By contrast, the GNN*coal* approach makes use of the whole ARG, and can easily scale to big sample size (>10), although it recovers α with high accuracy with sample size 3 only. Our interpretation is thus that the GNN*coal* is able of simultaneously leveraging information from topology and the age of coalescent events (nodes) across several genealogies (here 500). The GNN*coal* approach ultimately leverages information from observing recurrent occurrences of the same multiple merger events at different locations on the genome, while being aware of true multiple merger events from rapid successive binary mergers. We believe that our results pave the way towards the interpretability of GNN and DL methods applied to population genetics.

When applying both approaches to simulated sequence data (and not to true ARGs), both approaches behave differently. The GNN*coal* is not capable to accurately infer model parameter whether it is past variation of population size or α. In contrast, we find the SMβC to perform better than the GNN*coal* when dealing with sequence data (and not true ARG). SMβC is capable of recovering α and the shape of the demographic scenario in recent times irrespective of whether sequence data or ARG inferred by ARGweaver is given as input. This is most likely due to the fact that the statistic used by the SMβC (*i.e.* first coalescent event in discrete time) is more coarse than the statistic used by the GNN*coal* (*i.e.* the exact ARG). We therefore speculate that the theoretical framework of the SMβC, although being in theory less accurate than GNN*coal*, is more robust and suited for application to sequence data. More specifically the issue being faced by the GNN*coal* is known as out-of-distribution inference (Hüllermeier and Waegeman, 2021), which requires the network to generalize over an untrained data distribution. More precisely, this happens because GNN*coal* is not trained using ARG inferred by ARGweaver. Building a training data set for GNN*coal* to overcome this issue is currently impractical due to the inference speed of ARGweaver (very slow). However, future work will aim on increasing robustness of GNN inferences (*e.g.* by adding uncertainty or multiple models during the training process). Improving the performance of GNN*coal* on sequence data requires a more suitable ARG inference methods (*i.e.* efficient and accurate), which can be used on a broad spectrum of data sets (*i.e.* under a more general class of model to avoid bias from potential hypothesis violations of the chosen ARG inference approach).

Among the major forces driving the genome evolution there are the past demographic history, the reproductive mechanisms and natural selection (Johri et al., 2022a). Hence, in the second part of this manuscript we focus on integrating selection in both approaches. Currently, no method (especially if relying only on SFS information) can account for the presence of selection, linkage, non-constant population size and multiple merger events (Johri et al., 2022a) although recent theoretical framework might render this possible in the future (Alberti et al., 2021). As a first step to fill this gap, we demonstrate that the GNN*coal*

approach can be used for model selection (classification), reducing the number of potential hypotheses. Determining which evolutionary forces are driving the genome evolution is key, as only under the appropriate neutral population model can results of past demography and selection scans be correctly interpreted (Johri et al., 2022a, 2021). The high accuracy of the GNN*coal* approach in model selection is promising, especially as other methods based on the SFS alone (Koskela and Berenguer, 2019; Kato et al., 2017) have limits in presence of complex demographic scenarios which the GNN can possibly overcome, as it is easier to scale the GNN to estimate more parameters. We follow a thread of previous work (Sackman et al., 2019; Hejase et al., 2022; Bisschop et al., 2021a), by integrating and recovering selection, multiple merger and population size variation by simply allowing each (fixed) region in the genome to have its own α parameter. In presence of strong selection, we find lower α value around the selected loci and high α value in neutral neighbouring regions. In presence of weak selection, no effect on the estimated α value is observed, demonstrating that weak selection can be modeled by binary merger (no multiple merger) and only has a local effect on the branch length (*i.e.* by shortening it). Hence, our results point out that strong selection can indeed be seen/modeled as a local multiple merger event (see (Durrett and Schweinsberg, 2005; Bisschop et al., 2021a; Sackman et al., 2019)). In theory, both approaches could be able to infer the global α parameter linked to the reproductive mechanism, the local α parameter resulting from selection jointly with the variation of population size. However the absence of a simulator capable of simulating data with selection and non-constant population size under a β-coalescent model prevents us from delivering such proofs. We show strong evidence that under neutrality our approaches can recover a constant (and correct) α along the genome as well as the past variation of the population size. We further predict, that while selective processes favor coding regions, local variations in α as a consequence of sweepstake events should be indifferent to coding or non-coding regions. Hence, we suggest that current sequence simulators (Baumdicker et al., 2022; Haller et al., 2019a) could be extended to include the aforementioned factors and *de facto* facilitate the development of machine learning approaches.

Our study is unique in developing a new state of the art SMC approach to demonstrate that computational and mathematical problems can be overcome by DL (here GNN) approaches. The GNN*coal* approach is, in principle, not limited to the β-coalescent, and should work for other multiple merger models (*e.g.* Dirac coalescents (Eldon and Wakeley, 2006)) with the appropriate training. Furthermore, our SMβC approach is the first step to build a full genome method with an underlying model accounting for positive selection. In the future, further implementations may be added for a more realistic approach. The α parameter should be varying along the genome (as a hidden state), as the recombination rate in the iSMC (Barroso et al., 2019). This would allow to account for the local effect of strong and weak selection (Alberti et al., 2021). The effect of the α parameter could be also changing through time to

better model the non uniform occurrence of multiple merger events through time. Although it is mathematically correct to have α as a constant in time, it is erroneous in practice (5.12). We speculate that those additional features will allow to accurately model and infer multiple merger events, variation of population size, and selection at each position on the genome. We believe the DL approach could also be improved to recover more complex scenarios, providing in depth development on the structure of the graph neural networks, for example, by accounting for more features. At last, further investigation are required to make progress in the interpretability of the GNN methods, namely which statistics and convolution of statistics are used by GNN$coal$ to infer which parameters.

Currently, the applications of our methods are constrained by the assumptions of haploidy. Throughout the study we have assumed haploid data, but the GNN$coal$ and SMβC approach can easily be extended to diploidy (Birkner et al., 2013). Our results on inferred ARGs stress the need for efforts to be made on improving ARG inference (Brandt et al., 2022). Thanks to the SMC we are close to model the ARG allowing to infer demographic history, selection and specific reproductive mechanism. Moreover, the comparison of DL approaches with model driven *ad hoc* SMC methods may have the potential to help us solve issues (*e.g.* simultaneously inferring and accounting for recombination, variation of population size, different type of selection, population structure and the variation of the mutation and recombination rate along the genome) which have puzzled theoreticians and statisticians since the dawn of population genetics (Johri et al., 2022a).

On a final note, as environmental considerations hit us all, we suggest that decreasing the computer and power resources needed to perform DL/ GNN analyses should be attempted (Sapoval et al., 2022). Based on our study, we suggest that population genetics DL methods could be built as a two step process: 1) inferring ARGs, and 2) inferring demography and selection based on the ARGs. We speculate that general training sets based on ARGs could be build and be widely applicable for inference across many species with different life cycles and life history traits, while the inference of ARGs could be undertaken by complementary DL or Hidden Markov methods.

5.6 Tables

scenario	True α	α:SMβC,M=3	α:SMβC,M=4	α : GNN, M=3	α : GNN, M=10
Constant	2	1.97 (0.005)	1.97 (0.008)	1.99 (0.002)	1.99 (0.003)
Sawtooth	2	1.94 (0.017)	1.87 (0.019)	1.99 (0.002)	1.99 (0.003)
Bottleneck	2	1.97 (0.01)	1.97 (0.009)	1.99 (0.003)	1.99 (0.004)
Decrease	2	1.97 (0.007)	1.97 (0.008)	1.99 (0.003)	1.99 (0.004)
Increase	2	1.97 (0.007)	1.97 (0.008)	1.99 (0.004)	1.99 (0.002)

Table 5.1 Average estimated values of α by SMβC and GNN$coal$ over ten repetitions under the Kingman coalescent using 10 haploid sequences of 10 Mb and $\mu = r = 10^{-8}$ per generation per bp. The standard deviation is indicated in brackets.

scenario	True α	α^*:SMβC,M=3
Constant	1.9	1.86 (0.16)
Bottleneck	1.9	1.89 (0.09)
Increase	1.9	1.93 (0.07)
Decrease	1.9	1.96 (0.04)
Sawtooth	1.9	1.76 (0.17)
Constant	1.7	1.82 (0.10)
Bottleneck	1.7	1.64 (0.23)
Increase	1.7	1.82 (0.10)
Decrease	1.7	1.89 (0.13)
Sawtooth	1.7	1.71 (0.27)
Constant	1.5	1.52 (0.30)
Bottleneck	1.5	1.64 (0.33)
Increase	1.5	1.57 (0.24)
Decrease	1.5	1.60 (0.18)
Sawtooth	1.5	1.66 (0.14)
Constant	1.3	1.31 (0.20)
Bottleneck	1.3	1.2 (0.17)
Increase	1.3	1.24 (0.13)
Decrease	1.3	1.57 (0.11)
Sawtooth	1.3	1.37 (0.16)

Table 5.2 Average estimated α values by SMβC on simulated sequence data over ten repetitions using 10 sequences of 10 Mb with recombination and mutation rate set to 1×10^{-8} for α 1.9 and 1.7, 1×10^{-7} for α 1.5 and 1×10^{-6} for α 1.3 per generation per bp under a Beta coalescent process. The analysis are run on five different demographic scenarios (Constant population size, Bottleneck, Sudden increase, Sudden decrease and a Sawtooth demography).

5.7 Data availability

Code used to generate the simulated data for analysis, training and validation alongside (trained) deep learning models can be found at `https://github.com/kevinkorfmann/GNNcoal` and `https://github.com/kevinkorfmann/GNNcoal-analysis`. Code for SMC approaches used in this manuscript are available in the R package eSMC2 `https://github.com/TPPSellinger/eSMC2`.

5.8 Acknowledgments

This work was supported by the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A532B, 031A533A, 031A533B, 031A534A, 031A535A, 031A537A, 031A537B, 031A537C, 031A537D, 031A538A). KK is supported by a grant from the Deutsche Forschungsgemeinschaft (DFG) through the TUM International Graduate School of Science and Engineering (IGSSE), GSC 81, within the project GENOMIE QADOP. TS is supported by the Austrian Science Fund (project no. TAI 151-B). AT acknowledges funding from the DFG grant TE809/1-4 (project 254587930) and TE809/7-1 (project 317616126). FF and AT acknowledge funding from the DFG Priority Program SPP1590 on "Probabilistic Structures in Evolution". MF and AT acknowledge the support from the Imperial College - TUM Partnership award.

5.9 Supplementary Material: The Sequentially Markovian β Coalescent

5.9.1 SMβC

The Sequentially Markovian β Coalescent is a Hidden Markov Model based on the Multiple Sequentially Markovian Coalescent (MSMC) where Multiple Merger event are allowed to occur under the β coalescent.

To define our Hidden Markov Model (HMM) we need to define :

- Hidden States
- The signal (observed data)
- A Transition matrix (Probability of passing from one state to another)
- An Emission matrix (Probability of observing the signal conditional to the hidden state)
- An Initial probability (Probability of hidden states at the first position of the sequence)

Notations and Assumptions

We here define the different notations used and their meaning:

- r : recombination rate per nucleotide
- μ : Mutation rate per nucleotide
- u : recombination time, follows a continuous uniform distribution between 0 and first coalescent time.

- ξ_t : Scaled population size at time t ($N_t = \xi_t N_0$)
- $\chi_t = \xi_t^{\beta-1}$
- M : Number of analyzed haploid sequences (or haploid individuals)
- β : The multiple merger parameter

The model's assumptions are :

- $(\chi_t)_{t\geq 0}$ is piece-wise constant (intervals are specified in the following)

We first define the transition rates of the Beta n-coalescent (Kersting et al., 2014). The rate of transition from a state with b lineages to $b-n+1$ lineages, *i.e.* a merger of n lineages is

$$\lambda_{b,\beta,b-n+1} = \frac{B(n-\beta, b-n+\beta)}{\Gamma(2-\beta)\Gamma(\beta)}. \tag{5.1}$$

$$\Lambda_{b,\beta,b-n+1} = \frac{\binom{b}{n} B(n-\beta, b-n+\beta)}{\Gamma(2-\beta)\Gamma(\beta)}. \tag{5.2}$$

Thus, the total rate is

$$\lambda_{b,\beta} = \sum_{k=2}^{b} \frac{\binom{b}{k} B(k-\beta, b-k+\beta)}{\Gamma(2-\beta)\Gamma(\beta)} \tag{5.3}$$

Waiting times are exponentially distributed in the coalescent for population size constant in time. For time-varying population sizes, we define the time-changed Λ-n-coalescent as the (rescaled) genealogy limit from a Wright-Fisher type Cannings model with skewed offspring distributions as introduce in (Schweinsberg, 2003), which leads to a time-change waiting time for coalescence events: If a waiting time has rate λ in the standard Beta n-coalescent (started at some time t_0), it has a waiting time density of

$$f(t) = \frac{\lambda}{\chi(t)} e^{-\int_{t_0}^{t} \frac{\lambda}{\chi(s)} ds}, \tag{5.4}$$

which follows as described in (Freund, 2020).

Hidden States

Our hidden states at one position are the first coalescent time $t > 0$ at that position and which individuals $i := (i_1, \ldots, i_n)$ coalesce in the corresponding coalescence. A transition from coalescent time s to time t or a change in the index i can only occur when a recombination happens.

Observations

The observation signal is the comparison of the M analyzed sequences. Thus the signal is a series of number indicating the allelic state of the sequences at each position. For M=3, under

the infinite site model hypothesis, only 4 different state can be observed along the sequence. All sequences are the same at this position (indicated by a 0), or one of the three sequences is different from the two other (indicated by i, if the sequence i is different from the two other).

Transition Matrix

Five transitions are possible, we transition from (s, j) to (t, i). Here we assume that t and s are in interval time α and γ. At indices i and j, n and m individuals coalesce. In addition, recombination occurs with probability :

$$P(rec|s) = 1 - e^{-rMs} \tag{5.5}$$

We assume that only a single recombination event can happen before a merger. A recombination event of one of M lineages splits one ancestral lineage in two (backwards in time). The additional lineage is not yet described by the coalescent without the recombination event, we call this free. It can merge with any of the remaining M lineages, but also with the second parental ancestral lineage (*i.e.* the second split lineage from the recombination event). The transition probabilities/rates conditional on the (known) behaviour of the other lineages are as described in Dhersin et al., 2013, Sect. 5: Conditional on the mergers of the M other lineages, a binary merger of the "freed" lineage appears with rate $M\lambda_{M+1,\beta,M}$ and it joins an existing merger of m lineages at some time t with probability $1 - \frac{\lambda_{M+1,\beta,M+1-m+1}}{\lambda_{M,\beta,M-m+1}} = \frac{\lambda_{M+1,\beta,M+1-m}}{\lambda_{M,\beta,M-m+1}}$, where the second equation is due to the consistency of rates in Λ-n-coalescents. In the following, we derive conditional probabilities and/or conditional densities for certain events.

t<s

For this to happen, a recombination must occur before time t. The new number of individuals first coalescing is now $n = 2$, and the recombination event needs to affect one of these two individuals $i = \{i_1, i_2\}$ (which happens with probability $2/M$), splitting one lineage in two. Then, we just multiply the density of a binary merger of the free lineage with any of the other M lineages in the time-changed coalescent, which is Eq. (5.4) with rate $M\lambda_{M+1,\beta,M}$ and the probability that the first merger is indeed merging i, which is $\frac{1}{M}$ (we pick the second lineage of i at random from M lineages):

$$f(t, i|s, j, u) = \frac{2\lambda_{M+1,\beta,M}}{M\chi_t} e^{-\int_u^t \frac{M\lambda_{M+1,\beta,M}}{\chi_v} dv} \tag{5.6}$$

t=s

Case 1: a non coalescing individual joins the coalescent event $j = \{j_1, \ldots, j_n\}$. For this to happen, the recombination event must occur before time s in a non coalescing branch, which happens with probability $\frac{M-n}{M}$. Then, the newly split second ancestral lineage of i need to not coalesce in a binary collision until time s, which equals $\exp(-\int_u^s \frac{M\lambda_{M+1,\beta,M}}{\chi_v} dv)$ (by

integrating Eq. (5.4) with rate $M\lambda_{M+1,\beta,M}$). Finally, it then needs to join in the coalescent event j, which happens with probability $\frac{\lambda_{M+1,\beta,M+1-m}}{\lambda_{M,\beta,M-m+1}}$. This shows that

$$P(s,i|s,j,u) = \frac{(M-n)\lambda_{M+1,\beta,M+1-m}}{M\lambda_{M,\beta,M-m+1}} e^{-\int_u^s \frac{M\lambda_{M+1,\beta,M}}{\chi_v} dv} \tag{5.7}$$

for $i = j \cup \{i\}$.

Case 2: Recombination occurs in a coalescing individual $i \in j = \{j_1, \ldots, j_n\}$ of a multiple merger event with $n > 2$ (happens with probability $\frac{1}{n}$). The new lineage then coalesces higher in time, *i.e.* it does neither coalesce in a binary merger before s (as in case 1) nor in the collision at s (which is the complementary event from case 1). As above, this leads to

$$P(s,i|s,j,u) = \frac{\lambda_{M+1,\beta,M+1-m+1}}{n\lambda_{M,\beta,M-m+1}} e^{-\int_u^s \frac{M\lambda_{M+1,\beta,M}}{\chi_v} dv} \tag{5.8}$$

for $i = j \setminus \{i\}$

Case 3: Nothing changes. This happens if a) there is no recombination event (so $u > s$), b) the lineage split makes a binary merger between the two lineages resulting from the split ("self-coalesce") before s, c) recombination splits a lineage merged at the coalescence event at s, but that the second ancestral lineage from the split joins the merger.
a) happens with probability 1 if $u > s$, b) has conditional density as in Eq. 5.6 without the factor $2/M$, integrating over $[u, s]$ yields the probability $(1 - (1/M)\exp(-\int_u^t \frac{M\lambda_{M+1,\beta,M}}{\chi_v} dv)$ and c) follows as case 2, only we need the recombination event on a lineage already participating in the merger at time s (so just replacing $M - n$ with n in Eq. (5.8)).

t>s

For this to happen, a recombination must occur before time s and break a coalescent event of only two individuals (j_1, j_2) (w. probability $2/M$). Assume without restriction j_1 was affected by recombination. For (the new ancestral lineage of) j_1 to coalesce at time t, it must not coalesce until time s, then not coalesce in the former coalescent event and then the next coalescence event happens at time t. The next coalescence event can take any form and does not need to merge j_1. Additionally, we just need to keep track of the $M - 1$ non-recombining lineages and j_1, since we are not conditioning on the behaviour after s and thus both self-coalescence and the coalescence of the second split lineage (not j_1) can be ignored. Thus, to compute the conditional rate for merging into i, we first compute the probability that the new ancestral lineage representing j_1 in the new DNA segment does not coalesce until time s, given by Eq. (5.8) with $n = 2$, and multiply this by the conditional density for merging into any i of M lineages afterwards. Thus, this is just Eq. (5.4) with rate $\lambda_{M,\beta}$ multiplied with $\frac{\lambda_{M,\beta,M-n+1}}{\lambda_{M,\beta}}$. This leads to

$$f(t,i|s,j,u) = \frac{2\lambda_{M+1,\beta,M}}{M\lambda_{M,\beta,M-1}} e^{-\int_u^s \frac{M\lambda_{M+1,\beta,M}}{\chi_v} dv} \frac{\lambda_{M,\beta,M-n+1}}{\chi_t} e^{-\int_s^t \frac{\lambda_{M,\beta}}{\chi_v} dv} \tag{5.9}$$

for $i = \{i_1, \ldots, i_n\}$.

Full transition probability

$$p(t, i|s, j, u) =$$

$$\begin{cases} (1 - e^{-Mrs}) \frac{2\lambda_{M+1,\beta,M}}{\chi_t M} e^{-\int_u^t \frac{M\lambda_{M+1,\beta,M}}{\chi_v} dv} & u < t < s \\ e^{-Mrs} + (1 - e^{-Mrs})(\int_u^s \frac{e^{\int_u^k -\frac{M\lambda_{M+1,\beta,M}}{\chi_v} dv}}{\chi_k} dk + \frac{(M-n)\lambda_{(n+1),\beta,2}}{M\lambda_{(n+1),\beta,2}+\lambda_{(n+1),\beta,1}} e^{\int_u^t -\frac{M\lambda_{M+1,\beta,M}}{\chi_v} dv}) & t = s, m = n \\ (1 - e^{-Mrs}) \frac{(M-n)\lambda_{(n+1),\beta,1} e^{-\int_u^s \frac{M\lambda_{M+1,\beta,M}}{\chi_v} dv}}{M(\lambda_{(n+1),\beta,2}+\lambda_{(n+1),\beta,1})} & t = s, m = n+1 \\ (1 - e^{-Mrs}) \frac{1}{n} \frac{\lambda_{(n+1),\beta,2} e^{-\int_u^s \frac{M\lambda_{M+1,\beta,M}}{\chi_v} dv}}{s(\lambda_{(n+1),\beta,2}+\lambda_{(n+1),\beta,1})} & t = s, m+1 = n \\ (1 - e^{-Mrs}) \frac{\lambda_{M,\beta,(M-m+1)}}{\binom{M}{m}\chi_\alpha} e^{-\int_s^t \frac{\lambda_{M,\beta}}{\chi_v} dv} e^{-\int_u^s \frac{M\lambda_{M+1,\beta,M}}{\chi_v} dv} \frac{2\lambda_{(n+1),\beta,2}}{M(\lambda_{(n+1),\beta,2}+\lambda_{(n+1),\beta,1})} & t > s, i = l, j = k \end{cases} \tag{5.10}$$

As explained before, the state space must be finite. We therefore discretized time in k intervals. At one point the hidden state is α if $t \in [T_\alpha, T_{\alpha+1}]$, where $\alpha \in [0, (n-1)]$. We define T_α :

$$T_\alpha = \frac{-\ln(1 - \frac{\alpha}{n})}{\lambda_{M,\beta}} \tag{5.11}$$

We therefore have:

$$p(\alpha, i|s, j) = \int_{T_\alpha}^{T_{\alpha+1}} p(t, i|s, j) dt \tag{5.12}$$

Note: Because time is discretized, if the first coalescent time is bigger than $T_{(n-1)}$, then all individual coalesce.

Initial Probability

We use the equilibrium probability as initial probability while assuming m individual coalesce. The equilibrium probability is given by :

$$\begin{aligned} q_o(\alpha, i) &= \int_{T_\alpha}^{T_{\alpha+1}} \frac{\lambda_{M,\beta,(M-m+1)}}{\chi_\alpha \binom{M}{m}} e^{-\int_0^t \frac{\lambda_{M,\beta}}{\chi_v} dv} dt \\ &= \frac{\lambda_{M,\beta,(M-m+1)} e^{\sum_{\eta=0}^{\alpha-1} \frac{\lambda_{M,\beta}}{\chi_\eta} \Delta_\eta}}{\binom{M}{m} \lambda_{M,\beta}} (1 - e^{-\Delta_\alpha \frac{\lambda_{M,\beta}}{\chi_t}}) \end{aligned} \tag{5.13}$$

Where :

$$\Delta_\gamma = T_{\gamma+1} - T_\gamma \tag{5.14}$$

Calculation of $t_{\gamma,j}$ Assuming n individual coalesces.

$$t_{\gamma,j} = E[\text{Coalescent time}|\gamma, j] = \frac{E[\text{Coalescent time} \cap \gamma, j]}{P(\gamma, j)} = \frac{\int_{T_\gamma}^{T_{\gamma+1}} t\lambda_{M,\beta,(M-n+1)} e^{-\int_0^t \frac{\lambda_{M,\beta}}{\chi_v} dv}}{\binom{M}{n} q_0(\gamma, j)} dt$$
$$= \frac{T_\gamma - T_{\gamma+1} e^{-\Delta_\gamma \frac{\lambda_{M,\beta}}{\chi_\gamma}}}{(1 - e^{-\Delta_\gamma \frac{\lambda_{M,\beta}}{\chi_\gamma}})} + \frac{\chi_\gamma}{\lambda_{M,\beta}} \tag{5.15}$$

Where :

$$\Delta_\gamma = T_{\gamma+1} - T_\gamma \tag{5.16}$$

We note that $t_{\gamma,j}$ is independent of j, thus $t_{\gamma,j} = t_\gamma$.

Calculation of $p(\alpha, i|\gamma, j)$

$\alpha < \gamma$

We here calculate the transition probability from the state γ to a time t in the time interval α.

$$P(t, i|t_\gamma, j) = \frac{P_\gamma}{t_\gamma} \int_0^t \frac{2\lambda_{M+1,\beta,M}}{\chi_v M} e^{-\int_u^t \frac{M\lambda_{M+1,\beta,M}}{\chi_v} dv} du$$
$$= \frac{P_\gamma}{t_\gamma} \left(\sum_{\eta=0}^{\alpha-1} \int_{T_\eta}^{T_{\eta+1}} \frac{2\lambda_{M+1,\beta,M}}{\chi_v M} e^{-\int_u^t \frac{M\lambda_{M+1,\beta,M}}{\chi_v} dv} du + \int_{T_\alpha}^t \frac{2\lambda_{M+1,\beta,M}}{\chi_v M} e^{-\int_u^t \frac{M\lambda_{M+1,\beta,M}}{\chi_v} dv} du\right)$$
$$= \frac{P_\gamma}{t_\gamma} \left(\sum_{\eta=0}^{\alpha-1} \int_{T_\eta}^{T_{\eta+1}} \frac{2\lambda_{M+1,\beta,M}}{\chi_\alpha M} e^{-\int_{T_{\eta+1}}^{T_\alpha} \frac{M\lambda_{M+1,\beta,M}}{\chi_v} dv} e^{-\int_{T_\alpha}^t \frac{M\lambda_{M+1,\beta,M}}{\chi_v} dv} e^{-\int_u^{T_{\eta+1}} \frac{M\lambda_{M+1,\beta,M}}{\chi_v} dv} du + \right.$$
$$\left. \int_{T_\alpha}^t \frac{2\lambda_{M+1,\beta,M}}{\chi_\alpha M} e^{-(t-u)\frac{M\lambda_{M+1,\beta,M}}{\chi_\alpha}} dv du\right)$$
$$= \frac{P_\gamma}{t_\gamma} \frac{2\lambda_{M+1,\beta,M}}{\chi_\alpha M} \left(\sum_{\eta=0}^{\alpha-1} \int_{T_\eta}^{T_{\eta+1}} e^{-\sum_{\zeta=\eta+1}^{\alpha-1} \Delta_\zeta \frac{M\lambda_{M+1,\beta,M}}{\chi_\zeta}} e^{-(t-T_\alpha)\frac{M\lambda_{M+1,\beta,M}}{\chi_\alpha}} e^{-(T_{\eta+1}-u)\frac{M\lambda_{M+1,\beta,M}}{\chi_\eta}} du \right. \tag{5.17}$$
$$\left. + \frac{(1 - e^{-(t-T_\alpha)\frac{M\lambda_{M+1,\beta,M}}{\chi_\alpha}})}{\frac{M\lambda_{M+1,\beta,M}}{\chi_\alpha}}\right)$$
$$= \frac{P_\gamma}{t_\gamma} \frac{2\lambda_{M+1,\beta,M}}{\chi_\alpha M} \left(\sum_{\eta=0}^{\alpha-1} e^{-\sum_{\zeta=\eta+1}^{\alpha-1} \Delta_\zeta \frac{M\lambda_{M+1,\beta,M}}{\chi_\zeta}} e^{-(t-T_\alpha)\frac{M\lambda_{M+1,\beta,M}}{\chi_\alpha}} \frac{(1 - e^{-\Delta_\eta \frac{M\lambda_{M+1,\beta,M}}{\chi_\eta}})}{\frac{M\lambda_{M+1,\beta,M}}{\chi_\eta}} \right.$$
$$\left. + \frac{(1 - e^{-(t-T_\alpha)\frac{M\lambda_{M+1,\beta,M}}{\chi_\alpha}})}{\frac{M\lambda_{M+1,\beta,M}}{\chi_\alpha}}\right)$$

We then have to integrate of the time interval alpha to have the transition probability from the state γ to the state α.

$$
\begin{aligned}
P(\alpha, i|\gamma, j) = \int_{T_\alpha}^{T_{\alpha+1}} \frac{P_\gamma}{t_\gamma} \frac{2\lambda_{M+1,\beta,M}}{\chi_\alpha M} \Big(\sum_{\eta=0}^{\alpha-1} e^{-\sum_{\zeta=\eta+1}^{\alpha-1} \Delta_\zeta \frac{M\lambda_{M+1,\beta,M}}{\chi_\zeta}} e^{-(t-T_\alpha)\frac{M\lambda_{M+1,\beta,M}}{\chi_\alpha}} \frac{(1-e^{-\Delta_\eta \frac{M\lambda_{M+1,\beta,M}}{\chi_\eta}})}{\frac{M\lambda_{M+1,\beta,M}}{\chi_\eta}} \\
+ \frac{(1-e^{-(t-T_\alpha)\frac{M\lambda_{M+1,\beta,M}}{\chi_\alpha}})}{\frac{M\lambda_{M+1,\beta,M}}{\chi_\alpha}} \Big) dt \\
= \frac{P_\gamma}{t_\gamma} \frac{2\lambda_{M+1,\beta,M}}{\chi_\alpha M} \Big(\sum_{\eta=0}^{\alpha-1} e^{-\sum_{\zeta=\eta+1}^{\alpha-1} \Delta_\zeta \frac{M\lambda_{M+1,\beta,M}}{\chi_\zeta}} \frac{(1-e^{-\Delta_\alpha \frac{M\lambda_{M+1,\beta,M}}{\chi_\alpha}})}{\frac{M\lambda_{M+1,\beta,M}}{\chi_\alpha}} \frac{(1-e^{-\Delta_\eta \frac{M\lambda_{M+1,\beta,M}}{\chi_\eta}})}{\frac{M\lambda_{M+1,\beta,M}}{\chi_\eta}} \\
+ \frac{(\Delta_\alpha - \frac{(1-e^{-(\Delta_\alpha \frac{M\lambda_{M+1,\beta,M}}{\chi_\alpha})})}{\frac{M\lambda_{M+1,\beta,M}}{\chi_\alpha}})}{\frac{M\lambda_{M+1,\beta,M}}{\chi_\alpha}} \Big) \\
= \frac{P_\gamma}{t_\gamma} \frac{2}{M^2} \Big(\sum_{\eta=0}^{\alpha-1} e^{-\sum_{\zeta=\eta+1}^{\alpha-1} \Delta_\zeta \frac{M\lambda_{M+1,\beta,M}}{\chi_\zeta}} (1-e^{-\Delta_\alpha \frac{M\lambda_{M+1,\beta,M}}{\chi_\alpha}}) \frac{(1-e^{-\Delta_\eta \frac{M\lambda_{M+1,\beta,M}}{\chi_\eta}})}{\frac{M\lambda_{M+1,\beta,M}}{\chi_\eta}} \\
+ (\Delta_\alpha - \frac{(1-e^{-(\Delta_\alpha \frac{M\lambda_{M+1,\beta,M}}{\chi_\alpha})})}{\frac{M\lambda_{M+1,\beta,M}}{\chi_\alpha}}) \Big)
\end{aligned}
\tag{5.18}
$$

Where the recombination probability if defined as:

$$
P_\gamma = (1 - e^{-Mrt_\gamma}) \tag{5.19}
$$

$\gamma < \alpha$ We here calculate the transition probability from the state γ to a time t in the time interval α.

$$
\begin{aligned}
P(t, i|t_\gamma, j) = \int_0^{t_\gamma} \frac{P_\gamma}{t_\gamma} \frac{\lambda_{M,\beta,(M-m+1)}}{\binom{M}{m}\chi_\alpha} e^{-\int_{t_\gamma}^{t} \frac{\lambda_{M,\beta}}{\chi_\nu} d\nu} e^{-\int_u^{t_\gamma} \frac{M\lambda_{M+1,\beta,M}}{\chi_\nu} d\nu} \frac{2\lambda_{(n+1),\beta,2}}{M(\lambda_{(n+1),\beta,2} + \lambda_{(n+1),\beta,1})} du \\
= \frac{P_\gamma}{t_\gamma} \frac{\lambda_{M,\beta,(M-m+1)}}{\binom{M}{m}\chi_\alpha} \Big(\sum_{\eta=1}^{\gamma-1} \int_{T_\eta}^{T_{\eta+1}} e^{-\int_{t_\gamma}^{t} \frac{\lambda_{M,\beta}}{\chi_\nu} d\nu} e^{-\int_{T_{\eta+1}}^{T_\gamma} \frac{M\lambda_{M+1,\beta,M}}{\chi_\nu} d\nu} e^{-\int_{T_\gamma}^{t_\gamma} \frac{M\lambda_{M+1,\beta,M}}{\chi_\nu} d\nu} \\
\times e^{-\int_u^{T_{\eta+1}} \frac{M\lambda_{M+1,\beta,M}}{\chi_\nu} d\nu} \frac{2\lambda_{(n+1),\beta,2}}{M(\lambda_{(n+1),\beta,2} + \lambda_{(n+1),\beta,1})} du \\
+ \int_{T_\gamma}^{t_\gamma} e^{-\int_{t_\gamma}^{t} \frac{\lambda_{M,\beta}}{\chi_\nu} d\nu} e^{-\int_u^{t_\gamma} \frac{M\lambda_{M+1,\beta,M}}{\chi_\nu} d\nu} \frac{2\lambda_{(n+1),\beta,2}}{M(\lambda_{(n+1),\beta,2} + \lambda_{(n+1),\beta,1})} du \Big) \\
= \frac{P_\gamma}{t_\gamma} \frac{2\lambda_{(n+1),\beta,2} e^{-\int_{t_\gamma}^{t} \frac{\lambda_{M,\beta}}{\chi_\nu} d\nu}}{M(\lambda_{(n+1),\beta,2} + \lambda_{(n+1),\beta,1})} \frac{\lambda_{M,\beta,(M-m+1)}}{\binom{M}{m}\chi_\alpha} \Big(\sum_{\eta=1}^{\gamma-1} e^{-\sum_{\zeta=\eta+1}^{\gamma-1} \Delta_\zeta \frac{M\lambda_{M+1,\beta,M}}{\chi_\zeta}} \\
\times e^{-(t_\gamma - T_\gamma)\frac{M\lambda_{M+1,\beta,M}}{\chi_\gamma}} \frac{(1-e^{-\Delta_\eta \frac{M\lambda_{M+1,\beta,M}}{\chi_\eta}})}{\frac{M\lambda_{M+1,\beta,M}}{\chi_\eta}} \\
+ \frac{(1-e^{-(t_\gamma - T_\gamma)\frac{M\lambda_{M+1,\beta,M}}{\chi_\gamma}})}{\frac{M\lambda_{M+1,\beta,M}}{\chi_\gamma}} \Big)
\end{aligned}
\tag{5.20}
$$

We then have to integrate of the time interval alpha to have the transition probability from the state γ to the state α.

$$
\begin{aligned}
P(\alpha, i|\gamma, j) &= \int_{T_\alpha}^{T_{\alpha+1}} \frac{P_\gamma}{t_\gamma} \frac{2\lambda_{(n+1),\beta,2} e^{-\int_{t_\gamma}^{t} \frac{\lambda_{M,\beta}}{\chi_\nu} d\nu}}{M(\lambda_{(n+1),\beta,2} + \lambda_{(n+1),\beta,1})} \frac{\lambda_{M,\beta,(M-m+1)}}{\binom{M}{m}\chi_\alpha} \\
&\Big(\sum_{\eta=1}^{\gamma-1} e^{-\sum_{\zeta=\eta+1}^{\gamma-1} \Delta_\zeta \frac{M\lambda_{M+1,\beta,M}}{\chi_\zeta}} e^{-(t_\gamma - T_\gamma)\frac{M\lambda_{M+1,\beta,M}}{\chi_\gamma}} \frac{(1 - e^{-\Delta_\eta \frac{M\lambda_{M+1,\beta,M}}{\chi_\eta}})}{\frac{M\lambda_{M+1,\beta,M}}{\chi_\eta}} + \frac{(1 - e^{-(t_\gamma - T_\gamma)\frac{M\lambda_{M+1,\beta,M}}{\chi_\gamma}})}{\frac{M\lambda_{M+1,\beta,M}}{\chi_\gamma}}\Big) dt \\
&= \int_{T_\alpha}^{T_{\alpha+1}} \frac{P_\gamma}{t_\gamma} \frac{2\lambda_{(n+1),\beta,2} e^{-\int_{t_\gamma}^{T_\alpha} \frac{\lambda_{M,\beta}}{\chi_\nu} d\nu} e^{-\int_{T_\alpha}^{t} \frac{\lambda_{M,\beta}}{\chi_\nu} d\nu}}{M(\lambda_{(n+1),\beta,2} + \lambda_{(n+1),\beta,1})} \frac{\lambda_{M,\beta,(M-m+1)}}{\binom{M}{m}\chi_\alpha} \\
&\Big(\sum_{\eta=1}^{\gamma-1} e^{-\sum_{\zeta=\eta+1}^{\gamma-1} \Delta_\zeta \frac{M\lambda_{M+1,\beta,M}}{\chi_\zeta}} e^{-(t_\gamma - T_\gamma)\frac{M\lambda_{M+1,\beta,M}}{\chi_\gamma}} \frac{(1 - e^{-\Delta_\eta \frac{M\lambda_{M+1,\beta,M}}{\chi_\eta}})}{\frac{M\lambda_{M+1,\beta,M}}{\chi_\eta}} + \frac{(1 - e^{-(t_\gamma - T_\gamma)\frac{M\lambda_{M+1,\beta,M}}{\chi_\gamma}})}{\frac{M\lambda_{M+1,\beta,M}}{\chi_\gamma}}\Big) dt \\
&= \frac{P_\gamma}{t_\gamma} \frac{2\lambda_{(n+1),\beta,2} e^{-\int_{t_\gamma}^{T_\alpha} \frac{\lambda_{M,\beta}}{\chi_\nu} d\nu} (1 - e^{-\Delta_\alpha \frac{\lambda_{M,\beta}}{\chi_\alpha}})}{M(\lambda_{(n+1),\beta,2} + \lambda_{(n+1),\beta,1})} \frac{\lambda_{M,\beta,(M-m+1)}}{\binom{M}{m}\lambda_{M,\beta}} \\
&\Big(\sum_{\eta=1}^{\gamma-1} e^{-\sum_{\zeta=\eta+1}^{\gamma-1} \Delta_\zeta \frac{M\lambda_{M+1,\beta,M}}{\chi_\zeta}} e^{-(t_\gamma - T_\gamma)\frac{M\lambda_{M+1,\beta,M}}{\chi_\gamma}} \frac{(1 - e^{-\Delta_\eta \frac{M\lambda_{M+1,\beta,M}}{\chi_\eta}})}{\frac{M\lambda_{M+1,\beta,M}}{\chi_\eta}} + \frac{(1 - e^{-(t_\gamma - T_\gamma)\frac{M\lambda_{M+1,\beta,M}}{\chi_\gamma}})}{\frac{M\lambda_{M+1,\beta,M}}{\chi_\gamma}}\Big)
\end{aligned}
\tag{5.21}
$$

$\gamma = \alpha, m = n + 1$

For a multiple merger event to happen, there are three possibilities. A non coalescing branch join the coalescent event, or it coalesces in the same hidden state (before or after the coalescent event).

$$
\begin{aligned}
P(\gamma, i|\gamma, j) =& \frac{P_\gamma}{t_\gamma} \int_0^{t_\gamma} \frac{(M-n)\lambda_{(n+1),\beta,1} e^{-\int_u^{t_\gamma} \frac{M\lambda_{M+1,\beta,M}}{\chi_\nu} d\nu}}{M(\lambda_{(n+1),\beta,2} + \lambda_{(n+1),\beta,1})} du + Pc_1 + Pc_2 \\
=& \frac{P_\gamma}{t_\gamma} \frac{(M-n)\lambda_{(n+1),\beta,1}}{M(\lambda_{(n+1),\beta,2} + \lambda_{(n+1),\beta,1})} \Big(\sum_{\eta=1}^{\gamma-1} \int_{T_\eta}^{T_{\eta+1}} e^{-\int_u^{T_{\eta+1}} \frac{M\lambda_{M+1,\beta,M}}{\chi_\nu} d\nu} e^{-\int_{T_{\eta+1}}^{t_\gamma} \frac{M\lambda_{M+1,\beta,M}}{\chi_\nu} d\nu} du \\
&+ \int_{T_\gamma}^{t_\gamma} e^{-\int_u^{t_\gamma} \frac{M\lambda_{M+1,\beta,M}}{\chi_\nu} d\nu} du\Big) + Pc_1 + Pc_2 \\
=& \frac{P_\gamma}{t_\gamma} \frac{(M-n)\lambda_{(n+1),\beta,1}}{M(\lambda_{(n+1),\beta,2} + \lambda_{(n+1),\beta,1})} \Big(\sum_{\eta=1}^{\gamma-1} \frac{(1 - e^{-\Delta_\eta \frac{M\lambda_{M+1,\beta,M}}{\chi_\eta}})}{\frac{M\lambda_{M+1,\beta,M}}{\chi_\eta}} e^{-\int_{T_{\eta+1}}^{t_\gamma} \frac{M\lambda_{M+1,\beta,M}}{\chi_\nu} d\nu} \\
&+ \frac{(1 - e^{-(t_\gamma - T_\gamma)\frac{M\lambda_{M+1,\beta,M}}{\chi_\gamma}})}{\frac{M\lambda_{M+1,\beta,M}}{\chi_\gamma}}\Big) + Pc_1 + Pc_2 \\
=& \frac{P_\gamma (M-n)\lambda_{(n+1),\beta,1}}{t_\gamma M(\lambda_{(n+1),\beta,2} + \lambda_{(n+1),\beta,1})} \Big(\sum_{\eta=1}^{\gamma-1} \frac{(1 - e^{-\Delta_\eta \frac{M\lambda_{M+1,\beta,M}}{\chi_\eta}})}{\frac{M\lambda_{M+1,\beta,M}}{\chi_\eta}} e^{-\sum_{\zeta=\eta+1}^{\gamma-1} \frac{\Delta_\zeta M\lambda_{M+1,\beta,M}}{\chi_\zeta}} e^{-(t_\gamma - T_\gamma)\frac{M\lambda_{M+1,\beta,M}}{\chi_\gamma}} \\
&+ \frac{(1 - e^{-(t_\gamma - T_\gamma)\frac{M\lambda_{M+1,\beta,M}}{\chi_\gamma}})}{\frac{M\lambda_{M+1,\beta,M}}{\chi_\gamma}}\Big) + Pc_1 + Pc_2
\end{aligned}
\tag{5.22}
$$

Pc_1 is the probability that a recombination happens before the first coalescent event in the non coalescing branch, it then coalesce before the current first coalescent event but in the same hidden state (resulting in a multiple merger coalescent because of the discretized time)

Pc_2 is the probability that a recombination happens before $T_{\gamma+1}$ in the non coalescing branch, it then coalesce after the current first coalescent event but in the same hidden state (resulting in a multiple merger coalescent because of the discretized time)

$$
\begin{aligned}
Pc_1 = \int_{T_\gamma}^{t_\gamma} \frac{P_\gamma}{t_\gamma} \frac{\lambda_{2,\beta}}{\chi_\gamma M} \Big(\sum_{\eta=0}^{\gamma-1} e^{-\sum_{\zeta=\eta+1}^{\gamma-1}\Delta_\zeta \frac{M\lambda_{M+1,\beta,M}}{\chi_\zeta}} e^{-(t-T_\gamma)\frac{M\lambda_{M+1,\beta,M}}{\chi_\gamma}} \frac{(1-e^{-\Delta_\eta \frac{M\lambda_{M+1,\beta,M}}{\chi_\eta}})}{\frac{M\lambda_{M+1,\beta,M}}{\chi_\eta}} \\
+ \frac{(1-e^{-(t-T_\gamma)\frac{M\lambda_{M+1,\beta,M}}{\chi_\gamma}})}{\frac{M\lambda_{M+1,\beta,M}}{\chi_\gamma}}\Big) dt \\
= \frac{P_\gamma}{t_\gamma} \frac{\lambda_{2,\beta}}{\chi_\gamma M} \Big(\sum_{\eta=0}^{\gamma-1} e^{-\sum_{\zeta=\eta+1}^{\gamma-1}\Delta_\zeta \frac{M\lambda_{M+1,\beta,M}}{\chi_\zeta}} \frac{(1-e^{-(t_\gamma-T_\gamma)\frac{M\lambda_{M+1,\beta,M}}{\chi_\gamma}})}{\frac{M\lambda_{M+1,\beta,M}}{\chi_\gamma}} \frac{(1-e^{-\Delta_\eta \frac{M\lambda_{M+1,\beta,M}}{\chi_\eta}})}{\frac{M\lambda_{M+1,\beta,M}}{\chi_\eta}} \\
+ \frac{((t_\gamma-T_\gamma) - \frac{(1-e^{-(t_\gamma-T_\gamma)\frac{M\lambda_{M+1,\beta,M}}{\chi_\gamma}})}{\frac{M\lambda_{M+1,\beta,M}}{\chi_\gamma}})}{\frac{M\lambda_{M+1,\beta,M}}{\chi_\gamma}}\Big) \\
= \frac{P_\gamma}{t_\gamma} \frac{1}{M^2} \Big(\sum_{\eta=0}^{\gamma-1} e^{-\sum_{\zeta=\eta+1}^{\gamma-1}\Delta_\zeta \frac{M\lambda_{M+1,\beta,M}}{\chi_\zeta}} (1-e^{-(t_\gamma-T_\gamma)\frac{M\lambda_{M+1,\beta,M}}{\chi_\gamma}}) \frac{(1-e^{-\Delta_\eta \frac{M\lambda_{M+1,\beta,M}}{\chi_\eta}})}{\frac{M\lambda_{M+1,\beta,M}}{\chi_\eta}} \\
+ ((t_\gamma-T_\gamma) - \frac{(1-e^{-(t_\gamma-T_\gamma)\frac{M\lambda_{M+1,\beta,M}}{\chi_\gamma}})}{\frac{M\lambda_{M+1,\beta,M}}{\chi_\gamma}})\Big)
\end{aligned}
\tag{5.23}
$$

$$
\begin{aligned}
Pc_2 = \int_{t_\gamma}^{T_{\gamma+1}} \frac{P_\gamma}{t_\gamma} \frac{\lambda_{(n+1),\beta,2} e^{-\int_{t_\gamma}^{t} \frac{\lambda_{2,\beta}}{\chi_\gamma} dv}}{M(\lambda_{(n+1),\beta,2}+\lambda_{(n+1),\beta,1})} \frac{\lambda_{2,\beta}}{\chi_\gamma} \\
\Big(\sum_{\eta=1}^{\gamma-1} e^{-\sum_{\zeta=\eta+1}^{\gamma-1}\Delta_\zeta \frac{M\lambda_{M+1,\beta,M}}{\chi_\zeta}} e^{-(t_\gamma-T_\gamma)\frac{M\lambda_{M+1,\beta,M}}{\chi_\gamma}} \frac{(1-e^{-\Delta_\eta \frac{M\lambda_{M+1,\beta,M}}{\chi_\eta}})}{\frac{M\lambda_{M+1,\beta,M}}{\chi_\eta}} + \frac{(1-e^{-\Delta_\gamma \frac{M\lambda_{M+1,\beta,M}}{\chi_\gamma}})}{\frac{M\lambda_{M+1,\beta,M}}{\chi_\gamma}}\Big) dt \\
= \frac{P_\gamma}{t_\gamma} \frac{\lambda_{(n+1),\beta,2}(1-e^{-(T_{\gamma+1}-t_\gamma)\frac{\lambda_{2,\beta}}{\chi_\gamma}})}{M(\lambda_{(n+1),\beta,2}+\lambda_{(n+1),\beta,1})} \\
\Big(\sum_{\eta=1}^{\gamma-1} e^{-\sum_{\zeta=\eta+1}^{\gamma-1}\Delta_\zeta \frac{M\lambda_{M+1,\beta,M}}{\chi_\zeta}} e^{-(t_\gamma-T_\gamma)\frac{M\lambda_{M+1,\beta,M}}{\chi_\gamma}} \frac{(1-e^{-\Delta_\eta \frac{M\lambda_{M+1,\beta,M}}{\chi_\eta}})}{\frac{M\lambda_{M+1,\beta,M}}{\chi_\eta}} + \frac{(1-e^{-\Delta_\gamma \frac{M\lambda_{M+1,\beta,M}}{\chi_\gamma}})}{\frac{M\lambda_{M+1,\beta,M}}{\chi_\gamma}}\Big)
\end{aligned}
\tag{5.24}
$$

$\gamma = \alpha, m = n-1$

$$P(\gamma, i|\gamma, j) = \frac{P_\gamma}{t_\gamma}\int_0^{t_\gamma}\frac{\lambda_{(n+1),\beta,2}e^{-\int_u^{t_\gamma}\frac{M\lambda_{M+1,\beta,M}}{\chi_v}dv}}{n(\lambda_{(n+1),\beta,2}+\lambda_{(n+1),\beta,1})}du$$
$$= \frac{P_\gamma}{t_\gamma}\frac{\lambda_{(n+1),\beta,2}}{n(\lambda_{(n+1),\beta,2}+\lambda_{(n+1),\beta,1})}(\sum_{\eta=1}^{\gamma-1}\int_{T_\eta}^{T_{\eta+1}}e^{-\int_u^{T_{\eta+1}}\frac{M\lambda_{M+1,\beta,M}}{\chi_v}dv}e^{-\int_{T_{\eta+1}}^{t_\gamma}\frac{M\lambda_{M+1,\beta,M}}{\chi_v}dv}du$$
$$+\int_{T_\gamma}^{t_\gamma}e^{-\int_u^{t_\gamma}\frac{M\lambda_{M+1,\beta,M}}{\chi_v}dv}du)$$
$$= \frac{P_\gamma}{t_\gamma}\frac{\lambda_{(n+1),\beta,2}}{n(\lambda_{(n+1),\beta,2}+\lambda_{(n+1),\beta,1})}(\sum_{\eta=1}^{\gamma-1}\frac{(1-e^{-\Delta_\eta\frac{M\lambda_{M+1,\beta,M}}{\chi_\eta}})}{\frac{M\lambda_{M+1,\beta,M}}{\chi_\eta}}e^{-\int_{T_{\eta+1}}^{t_\gamma}\frac{M\lambda_{M+1,\beta,M}}{\chi_v}dv}+\frac{(1-e^{-(t_\gamma-T_\gamma)\frac{M\lambda_{M+1,\beta,M}}{\chi_\gamma}})}{\frac{M\lambda_{M+1,\beta,M}}{\chi_\gamma}})$$
$$= \frac{P_\gamma}{t_\gamma}\frac{\lambda_{(n+1),\beta,2}}{n(\lambda_{(n+1),\beta,2}+\lambda_{(n+1),\beta,1})}(\sum_{\eta=1}^{\gamma-1}\frac{(1-e^{-\Delta_\eta\frac{M\lambda_{M+1,\beta,M}}{\chi_\eta}})}{\frac{M\lambda_{M+1,\beta,M}}{\chi_\eta}}e^{-\sum_{\zeta=\eta+1}^{\gamma-1}\Delta_\zeta\frac{M\lambda_{M+1,\beta,M}}{\chi_\zeta}}e^{-(t_\gamma-T_\gamma)\frac{M\lambda_{M+1,\beta,M}}{\chi_\gamma}}$$
$$+\frac{(1-e^{-(t_\gamma-T_\gamma)\frac{M\lambda_{M+1,\beta,M}}{\chi_\gamma}})}{\frac{M\lambda_{M+1,\beta,M}}{\chi_\gamma}})$$
(5.25)

$\gamma = \alpha, m = n$

$$P(\gamma, j|\gamma, j) = 1 - \sum_{\alpha\neq\gamma, i\neq j}P(\alpha, i|\gamma, j) \tag{5.26}$$

Emission Matrix

To compute the emission probabilities, we need the external branch length at each position of the genome (noted as Ts). Ts is estimated with an additional HMM as described in (Schiffels and Durbin, 2014).

M=3

One can only observe 4 possibilities. Observation 0, no mutation within the 3 individual. Observation 1 (2,3), individual 1 (2,3) is different from individual 2 and 3 (1 and 3 , 1 and 2).

if the first coalescent event involves 2 individuals we have :

$$P(0|\gamma) = e^{-\mu(Ts)} \tag{5.27}$$

$i \in 1, 2, 3$. Mutation occurred and did not occurred in the first coalescent event.

$$P(i|\gamma, i) = (1 - e^{-\mu(Ts-2t_\gamma)}) \tag{5.28}$$

$i \in 1, 2, 3; \bar{i} \neq i$. Mutation occurred and is in the first coalescent event.

$$P(i|\gamma, \bar{i}) = (1 - e^{-\mu(2t_\gamma)}) \tag{5.29}$$

if the first coalescent event involves 3 individuals we have :

$$P(0|\gamma) = e^{-\mu(3t_\gamma)} \tag{5.30}$$

$i \in 1, 2, 3$

$$P(i|\gamma) = 1 - e^{-\mu(3t_\gamma)} \tag{5.31}$$

M=4

One can only observe 8 possibilities as we only focus on SNPs. Observation 0, no mutation within the 4 individual. Observation 1 (2,3,4), individual 1 (2,3,4) is different from all other individual. Observation 5 to 7 , two individual are different from the other two.

if the first coalescent event involves 4 individuals:

$$P(0|\gamma) = e^{-\mu(4t_\gamma)} \tag{5.32}$$

$i \in 1, 2, 3, 4$

$$P(i|\gamma) = 1 - e^{-\mu(4t_\gamma)} \tag{5.33}$$

if the first coalescent event involves 3 individuals:

$$P(0|\gamma) = e^{-\mu(Ts)} \tag{5.34}$$

$i \in 1, 2, 3, 4$

$$P(i|\gamma, \bar{i}) = (1 - e^{-\mu(t_\gamma)}) \tag{5.35}$$

$i \in 1, 2, 3, 4$

$$P(i|\gamma, i) = (1 - e^{-\mu(Ts-3t_\gamma)}) \tag{5.36}$$

if the first coalescent event involves 2 individuals:

No mutation occurred.

$$P(0|\gamma) = e^{-\mu(Ts)} \tag{5.37}$$

$i \in 1, 2, 3, 4$ and mutation is within on one of the individual that coalesce.

$$P(i|\gamma) = (1 - e^{-\mu(t_\gamma)}) \tag{5.38}$$

$i \in 1, 2, 3, 4$ and mutation is not on one of the individual that coalesce.

$$P(i|\gamma) = (1 - e^{-\mu(\frac{(Ts-4t_\gamma)}{2}+t_\gamma)}) \tag{5.39}$$

$i \in 5, 6, 7$ and the two individual coalescing are identical.

$$P(i|\gamma) = e^{-\mu(Ts)} \tag{5.40}$$

$i \in 5, 6, 7$ and the two individual coalescing are different,then two mutation must occur, which has probability 0.

$$P(i|\gamma) = 0 \tag{5.41}$$

5.10 Supplementary Material: Description of the Graph Neural Network Approach (GNN$coal$)

The following part contains a detailed description of the graph neural net approach.

5.10.1 Brief Introduction to neural network

We wish to define a neural network that has some learnable parameters (*i.e.* weights) to predict parameters of interest. In this study, we focus on the past variation of population size and on the α parameter of the Beta distribution shaping the offspring distribution (Schweinsberg, 2003). To define a neural network, we first generate a training dataset, *i.e.* a dataset containing information to learn from (the ancestral recombination graph) and predictive variables of interest (past demography and α parameter). We then process inputs through the neural network and compute the loss, which evaluates the distance between predictions and ground truth. Afterwards, we propagate gradients back into the network's parameters and update its weights. We now describe further details of this procedure.

5.10.2 Datasets

Set 1: Variable demography and MMC

To build our training data set, we rely on the efficient coalescent and sequence simulator *msprime* (Baumdicker et al., 2022; Kelleher et al., 2018). We simulated 10 haploid genomes for a total of 2×10^4 demographic scenarios with 100 replicates each. We followed the procedure outlined in Figure 5.9 (see below) to define our scenarios. To ensure that each simulation contains sufficient data points, we imposed that 95% of the 100 replicates had to contain at least 500 trees. This procedure resulted in a training data set containing between 9.5×10^8 and 1×10^9 coalescent trees. Each simulated data point was stored as a tree sequence and converted into the graph format (*i.e. Pytorch Geometric* data object for the forward pass. This procedure drastically decreased the storage space requirements at the marginal cost of training speed. Furthermore, we trained our network over a range of the multiple merger parameter α from 1.01 to 1.99.

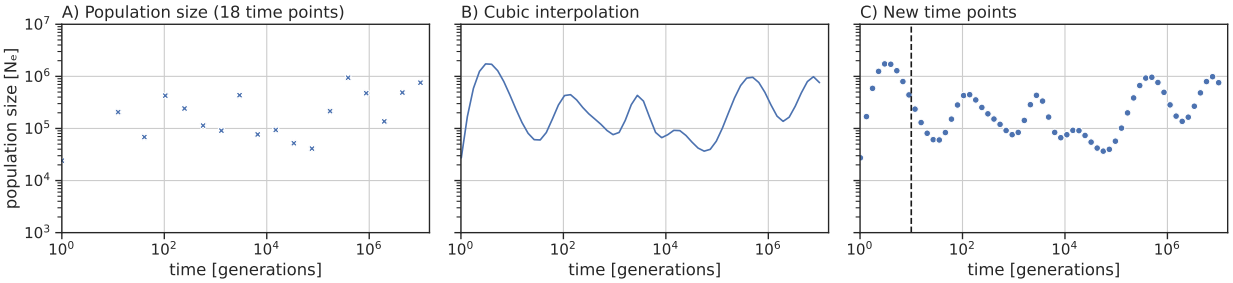


Figure 5.9 Demography sampling process Population size (y-axis) changes through time in generations (x-axis). Demography sampling process according to (Boitard et al., 2016; Sanchez et al., 2021b) (A) and cubic interpolation (B), and re-sampling for 60 uniformly spaced points. The sampling process is repeated if any population size is outside the 10^4 and 10^7 population size window.

Set 2: MMC versus selection under constant population size

To discern between multiple merger and selection, we simulated 1000 replicates under constant population size ($N_e = 10^5$) by randomly choosing between Kingman and β-coalescent models. Additionally, we randomly chose one of the selection regimes when a Kingman coalescent scenario was drawn. For each scenario, we simulated 100 replicates with 99% of each replicate containing at least 2,000 trees (*i.e.* a batch size of 2000). This procedure resulted in a training data set of more than 1.98×10^8 coalescent trees. An additional validation dataset was generated consisting of 1,000 simulations under the same scenarios. A script for replicating these simulations is available at `https://github.com/kevinkorfmann/GNNcoal-analysis`.

5.10.3 Training

The demography inference model was trained for two epochs, an epoch being defined by iterating one time over the training dataset. *i.e.* the data is processed twice, once by the untrained neural network and a second time after the first epoch. The model to infer the α parameter was trained on 10% of the dataset used for demography inference as we found the procedure to be sufficient for accurate estimates. We set the batch size to 500 (*i.e.* number of trees processed before updating the neural network parameters). Jupyter notebooks to perform the training are accessible at `https://github.com/kevinkorfmann/GNNcoal-analysis`.

5.10.4 Neural network

The neural network is implemented in *PyTorch* (Paszke et al., 2017) using the extension *Pytorch Geometric* for the graph convolution operation and differential hierarchical pooling strategy (Fey and Lenssen; Kipf and Welling, 2016; Ying et al.). For training and inference, each coalescent tree was interpreted as an undirected graph, with each ancestral node or leaf having edges to their parent or children, and vice versa. Each node i contains a feature vector (a feature vector being the predicted values by the neural network) $\mathbf{x_i}$ of size 60 (number of desired estimated parameters), which are horizontally stacked into a feature vector matrix (*i.e.* each node/coalescent event has feature vector) $\mathbf{X}$ and initialized with ones on the diagonal and zeros everywhere else. Each feature vector are unique.

As mentioned above, we chose a feature vector of size 60, which corresponds to the final output dimension of the demography network. A convolution of an individual feature vector $\mathbf{x_i}$ of node i is computed following (Kipf and Welling, 2016):

$$\mathbf{x}'_i = \mathbf{\Theta}^\top \sum_{j \in \mathcal{N}(v) \cup \{i\}} \frac{e_{j,i}}{\sqrt{\hat{d}_j \hat{d}_i}} \mathbf{x}_j \tag{5.42}$$

With $\hat{d}_i = 1 + \sum_{j \in \mathcal{N}(i)} e_{j,i}$ and $e_{j,i}$ being the pivot edge from j to i. The node features $\mathbf{x_i}$ are transformed by a learnable weight matrix Θ, normalized according to the degrees of the respective nodes, and summed over the neighboring nodes. Alongside this node-wise formulation, the same computation can be written as matrix-operations. The latter will be the notation used to explain the following pooling strategy:

$$\mathbf{X}' = \hat{\mathbf{D}}^{-1/2} \hat{\mathbf{A}} \hat{\mathbf{D}}^{-1/2} \mathbf{X} \mathbf{\Theta}, \tag{5.43}$$

Where $\hat{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ represents the adjacency matrix and $\hat{D}_{ii} = \sum_{j=0} \hat{A}_{ij}$, represent the diagonal degree matrix. Here, the edge information enters the convolution as part of the values of the adjacency matrix. Both representations are valid descriptions of the same convolution (Fey and Lenssen).

We defined our GNN as a series of three consecutive graph convolutions, with input, hidden and output dimensions of size 60, each convolution separated by a one-dimensional batch normalization step. Increasing the number of convolutions above a certain value will lead to a *smoothing* of the feature vectors, as the information of more distant neighbours will be taken into account for each node (Kipf and Welling, 2016). The output of each convolution is then concatenated and passed through-linear activation and ReLu-function ($Relu(z) = max(0, z)$). The complete implementation is available at `https://github.com/kevinkorfmann/GNNcoal`.

Because each node has a feature vector (*i.e.* each coalescent event of a coalescent tree lead to a specific prediction), the feature vector matrix (*i.e.* the collection of estimated demography parameters from a coalescent tree) needs to be processed to obtain a consensus predicted demography. This step is called the pooling strategy. A detailed description of the pooling strategy used in our study is formulated in (Ying et al.) and short description can be found below.

We compress the number of nodes hierarchically in a step-wise manner, from the l state to $l + 1$ state. Close nodes in the original graph are summarized by a supernode in a new smaller graph, ideally having only one supernode at the end with one feature vector, containing our variables of interest. Thus, this downsizing can be formulated as decreasing the original adjacency matrix $\mathbf{A}$ of a given genealogy to a smaller subgraph $\mathbf{A}^{l+1}$ ($\mathbf{A} \in \mathbb{R}^{n \times n} \rightarrow \mathbf{A}^{l+1} \in \mathbb{R}^{m \times m}$ with n, m being the number of nodes in each graph and $\mathbf{Z} \in \mathbb{R}^{n \times d} \rightarrow \mathbf{Z}^{l+1} \in \mathbb{R}^{m \times d}$ with condition of $m < n$; here d being the length of the feature vector). To achieve this, we cluster a set of nodes in the original genealogy into a smaller subset of supernodes. A cluster assignment matrix $\mathbf{S} \in \mathbb{R}^{n^l \times n^{l+1}}$ (with n^l is the number of clusters in the current step and n^{l+1} the number of clusters in the next step). In the first iteration, the number of clusters n^l is equal to the number of nodes, and it is $learned$ using an GNN_{pool}. This means that the criterion to compress individual nodes is dependent on the overall objective of network, which is to

reduce the loss function with respect to predicting parameters of interest. The specific computation using two GNNs ((GNN_{pool} and GNN_{embed})) is explained in the following:

1. We compute the feature vectors $\mathbf{Z^l}$ using GNN_{embed} at step l (equation (5.44)).

$$\mathbf{Z}^l = GNN_{l,embed}(\mathbf{A}^l, \mathbf{X}^l) \tag{5.44}$$

2. Next we compute the cluster assignment matrix, by computing "feature vectors" and applying the row-wise softmax function ($\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{d} e^{z_j}} \quad for\ i = 1, 2, \ldots, d$) as normalization step. This results is a probability distribution of each node to be part of the next cluster.

$$\mathbf{S}^l = softmax(GNN_{l,pool}(\mathbf{A}^l, \mathbf{X}^l)) \tag{5.45}$$

3. Lastly, the new adjacency matrix $\mathbf{A}^{l+1}$ and corresponding feature vector matrix $\mathbf{X}^{l+1}$ are calculated ((5.46) and (5.47)). Notably, in equation $\mathbf{A^l}$ is multiplied twice with $\mathbf{S}$ to ensure that $\mathbf{A}^{l+1}$ is square (*i.e.* the number of columns is equal to the number of rows).

$$\mathbf{X}^{l+1} = \mathbf{S}^{l^T}\mathbf{Z}^l \rightarrow \mathbb{R}^{n^{l+1} \times d} \tag{5.46}$$

$$\mathbf{A}^{l+1} = \mathbf{S}^{l^T}\mathbf{A^l}\mathbf{S}^l \rightarrow \mathbb{R}^{n^{l+1} \times n^{l+1}} \tag{5.47}$$

In our approach, we used three consecutive pooling iterations $l = 1, ..., 3$. At each iteration, a graph is compressed down to 30% of the remaining nodes. In total, six GNNs (three embedding GNNs and three pooling GNNs) contribute to the number of learnable parameters, followed by two linear layers with interspaced ReLU-activation to compress the output into the desired demographic-time window dimension. The demography model implementation can be found at `https://github.com/kevinkorfmann/GNNcoal`.

We now focus on the α inference and the classification. The model for inferring the α parameter adds three linear layers to the demography network model, while the classification model adds two linear layers. To illustrate the simplicity to estimate α and classify models, we show the code for reducing the number of dimensions below.

α inference model:

```
class AlphaInferenceModel(nn.Module):

    def __init__(self, DemographyNet, time_window=60):
        super().__init__()
        self.l1 = nn.Linear(time_window, time_window//2)
        self.l2 = nn.Linear(time_window//2, time_window//4)
        self.l3 = nn.Linear(time_window//4, 1)
        self.DemographyNet = DemographyNet

    def forward(self, batch):
        x = self.DemographyNet(batch)
        return self.l3(F.relu(self.l2(F.relu(self.l1(x)))))
```

Classification model:

```
class ClassificationModel(nn.Module):

    def __init__(self, DemographyNet, num_classes, time_window=60):
        super().__init__()
        self.l1 = nn.Linear(time_window, time_window//2)
        self.l2 = nn.Linear(time_window//2, num_classes)
        self.DemographyNet = DemographyNet

    def forward(self, batch):
        x = self.DemographyNet(batch)
        return self.l2(F.relu(self.l1(x)))
```

5.10.5 Time window

The output of the neural network is a fixed-size vector containing the values of inferred population size. However, not all trees contain coalescent events falling into these pre-specified time-bins. As a consequence, the population size can only be inferred for simulations where trees contain coalescence events in those specific bins. A multi-step heuristic has been developed to obtain an estimate of the inference window for each simulation, respectively. The procedure is summarized as:

```
    1. Simulate 100 repetitions for a demographic parameter set.
    2. Count number of coalescent events for first 500 trees for each simulation.
    3. Smoothen coalescent event vectors with sliding window*.
    4. Only retain the largest consecutive coalescent time-window per repetition.
    5. Create Boolean mask by retaining the time-windows with at least X coalescent events
    along repetition axis.

    *Sliding window: If left and right of pivot element up to Y coalescent events
    are present, retain the pivot element, move one step to the right and repeat.
```

This heuristically constructed mask is used to zero out all inferences outside of the coalescent events before computing the RMSE-loss function and back-propagating the weights (avoiding under- and over-fitting). In practice, a single row of the mask, is multiplied into a matrix of dimensions equal to the number of trees times the number of windows (here: 500x60). With this procedure, each tree receives the same masked loss based on all repetitions (to balance the training data set). Coalescent events and masks for constant demography of size 10^6 are provided in Figure 5.10 below.

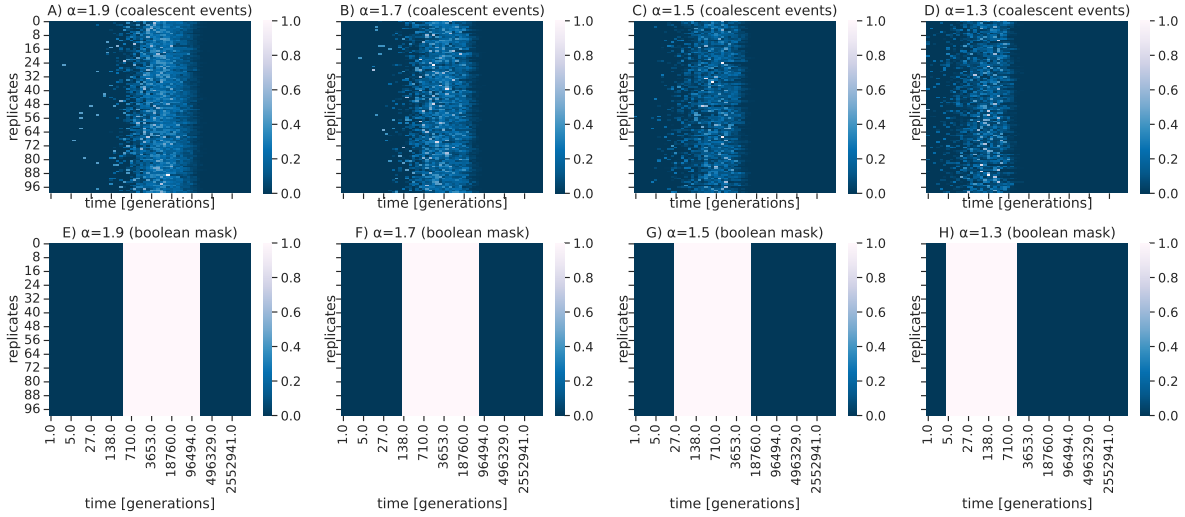


Figure 5.10 Coalescent events and boolean masks A)-D) The number of coalescent events for 100 repetitions (y-axis) across the first 500 trees of a simulation for 60 time windows (x-axis). E)-H): Boolean mask after applying a heuristic to determine a suitable time-window for demography inference.

During the inference, when only one sample is available, the masking procedure was approximated by calculating the mean of the log-scaled node times of the first 500 trees and by taking two standard deviations at both sides from the mean to form a similar time-window even if only one repetition is available (see below).

```
def alternative_coalescent_mask(ts, population_time, x_times_std=2, n_trees=500):

    trees = ts.aslist()[0:n_trees]
    nodes_n_trees = []
    for tree in trees:
        nodes_n_trees += list(tree.nodes())

    node_times = [ts.get_time(node.id) for node in ts.nodes() if node.id >= ts.num_samples and
        node.id in nodes_n_trees]

    log_node_times = np.log(node_times)
    mean = log_node_times.mean()
    std = log_node_times.std()
    lowerbound = np.exp(mean-x_times_std*std)
    upperbound = np.exp(mean+x_times_std*std)
    mask1 = population_time > lowerbound
    mask2 = population_time < upperbound
    mask = np.logical_and(mask1, mask2)
    return mask
```

5.11 Supplementary Figures and Tables: Simultaneous Inference of Past Demography and Selection from the Ancestral Recombination Graph under the Beta Coalescent

The following part contains all additional supplementary figures and tables, which are not part of the main text.

5.11.1 Supplementary Figures

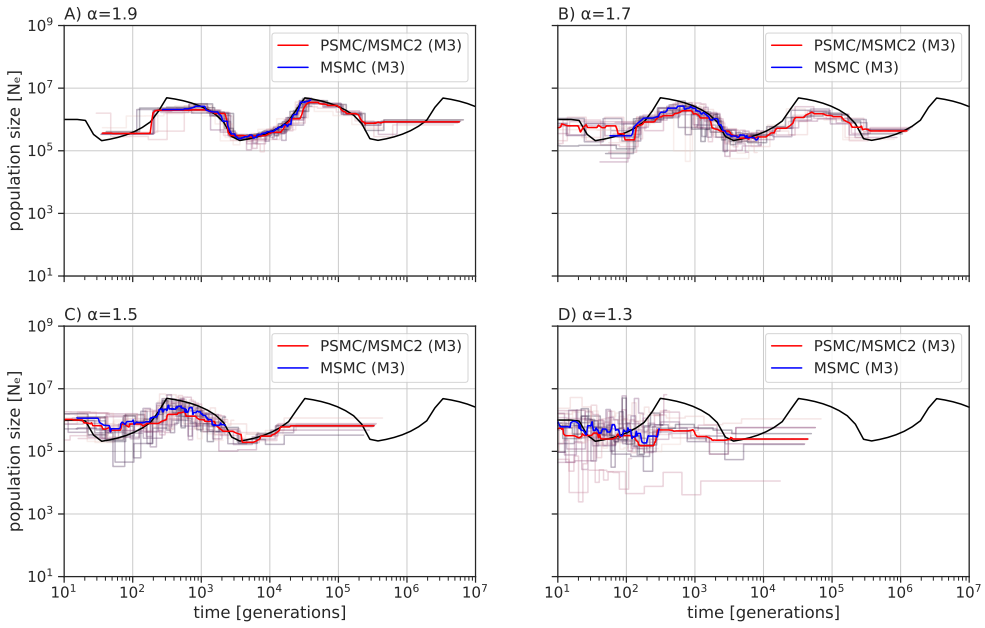


Figure 5.11 Estimated demographic history by MSMC (blue) and MSMC2 (red) from the true ARG using 10 sequences of 100 Mb when population undergoes a sawtooth demographic scenario (black) for different α values: A) 1.9, B) 1.7, C) 1.5,and D) 1.3. The estimated population size is corrected in order to mask the scaling difference between the Kingman coalescent and the Beta coalescent (*i.e.* $Ne = (((mu_{estimated}/mu_{real})/scale)^{(}1/(alpha-1)))$, where $m = 1+(1/((2^{(}alpha-1))*(alpha-1)))$, $scale = (m^{alpha})/(alpha*beta(2-alpha, alpha))$ and $mu_{estimated} = \frac{\theta}{((2*sum(1/(1:(nb_{ind}-1))))*L)}$). The recombination and mutation rate are set to 1×10^{-8} per generation per bp.

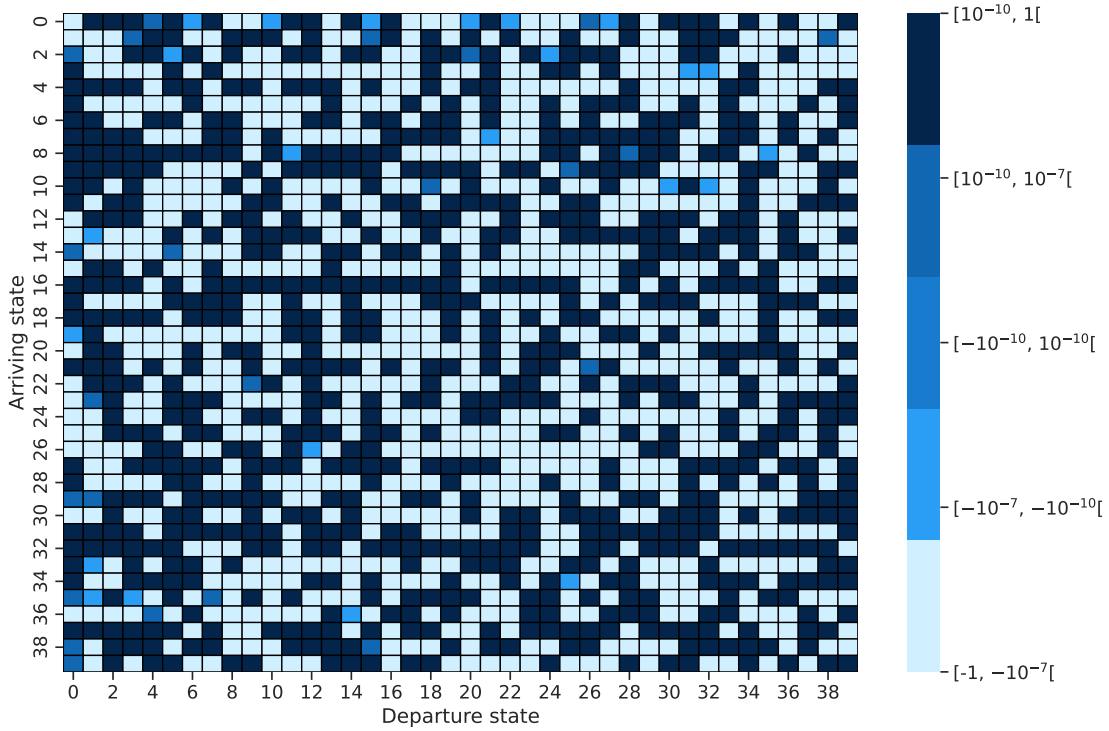


Figure 5.12 Difference between the observed transition matrix and the theoretical prediction under the eSMC2 under a constant population size under the Kingman Coalescent

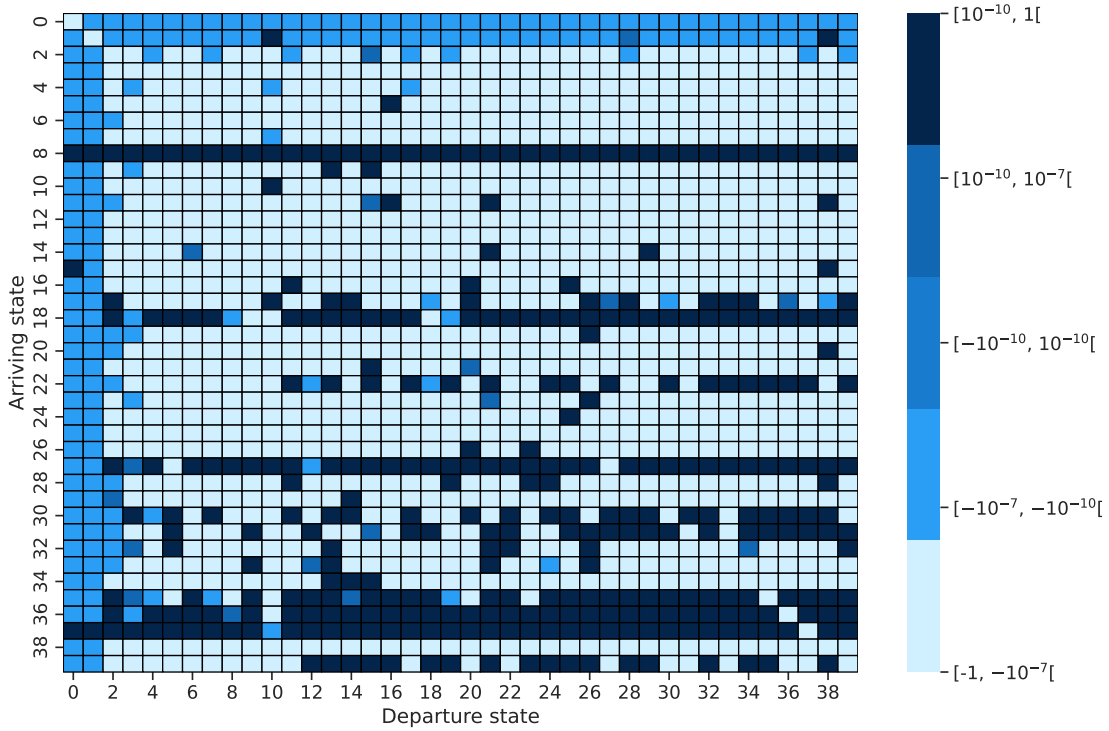


Figure 5.13 Difference between the observed transition matrix and the theoretical prediction under the eSMC2 under a constant population size under the Beta Coalescent with $\alpha = 1.3$

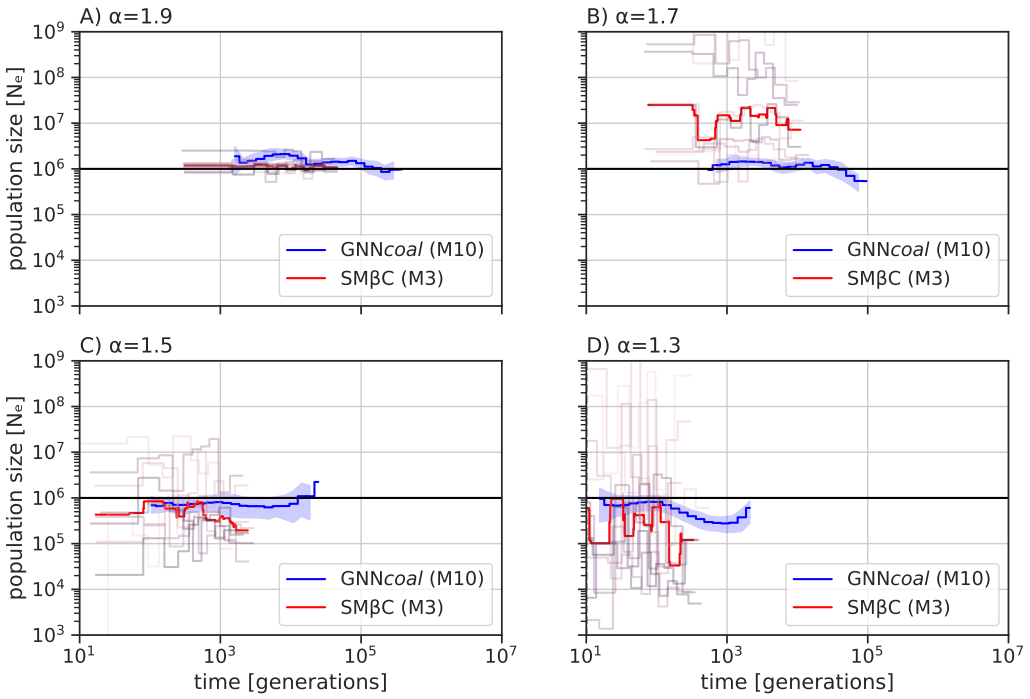


Figure 5.14 Best-case convergence estimations of SMβC and GNN$coal$ under a Beta coalescent. Estimations of past demographic history by SMβC using 10 sequences and 100 Mb in red (median) and by GNN using 10 sequences and 500 trees in blue (mean and CI95) when population undergoes a "constant" demographic scenario (black) under 4 different α values 1.9,1.7,1.5 and 1.3, respectively in A),B),C) and D). The recombination and mutation rate are set to 1×10^{-8} per generation per bp.

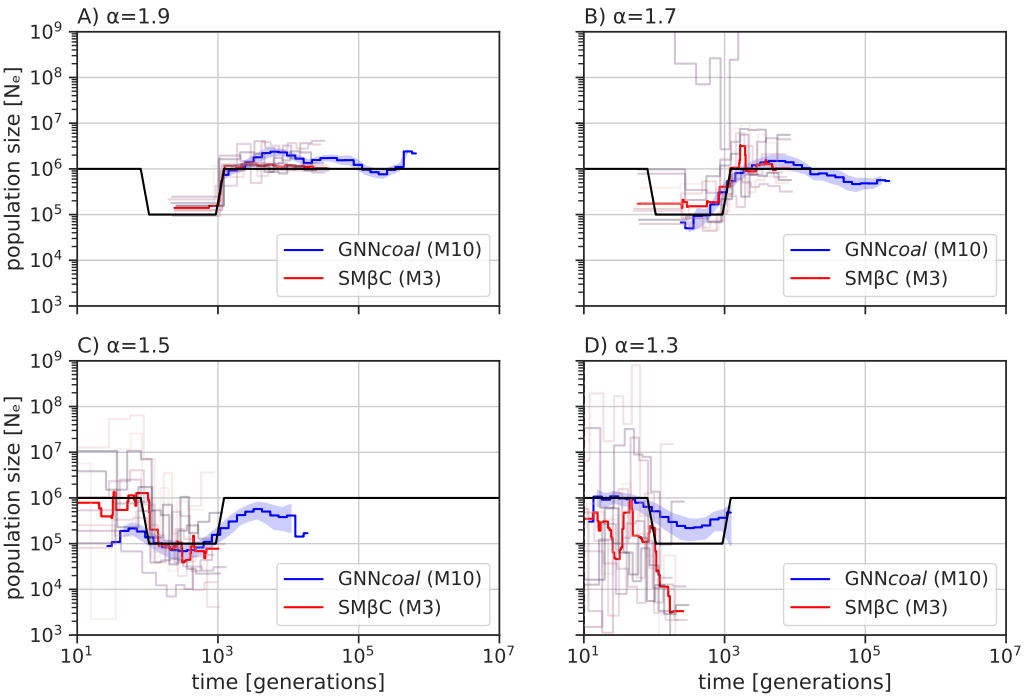


Figure 5.15 Best-case convergence estimations of SMβC and GNN*coal* under a Beta coalescent. Estimations of past demographic history by SMβC using 10 sequences and 100 Mb in red (median) and by GNN using 10 sequences and 500 trees in blue (mean and CI95) when population undergoes a "bottleneck" demographic scenario (black) under 4 different α values A) 1.9, B) 1.7, c) 1.5 and D) 1.3. The recombination and mutation rate are set to 1×10^{-8} per generation per bp.

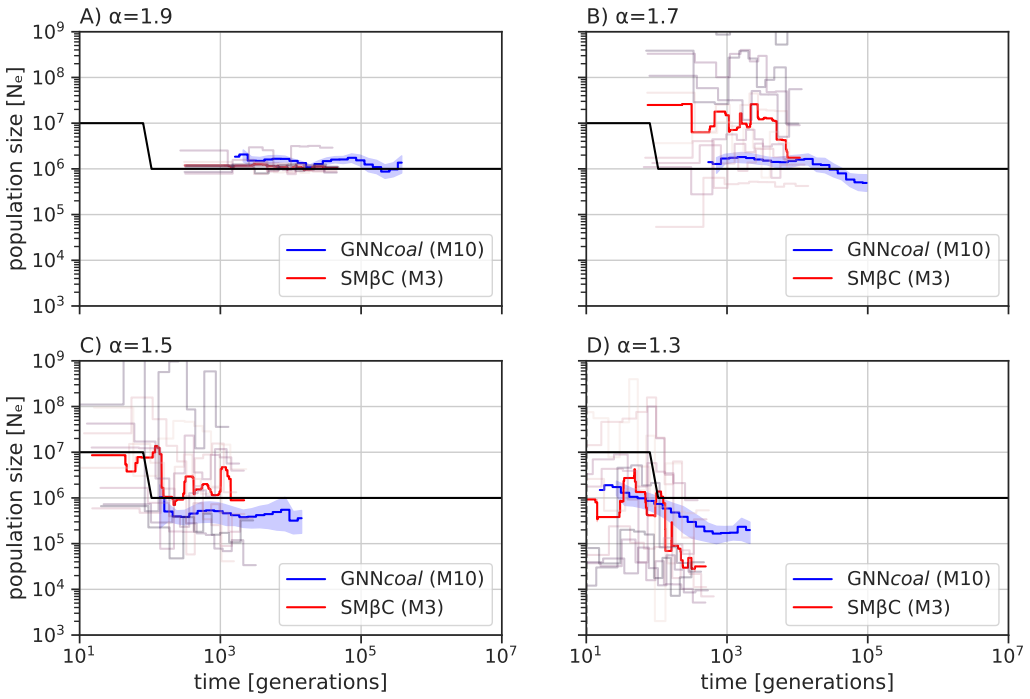


Figure 5.16 Best-case convergence estimations of SMβC and GNN$coal$ under a Beta coalescent. Estimations of past demographic history by SMβC using 10 sequences and 100 Mb in red (median) and by GNN using 10 sequences and 500 trees in blue (mean and CI95) when population undergoes a "increase" demographic scenario (black) under 4 different α values A) 1.9, B) 1.7, c) 1.5 and D) 1.3. The recombination and mutation rate are set to 1×10^{-8} per generation per bp.

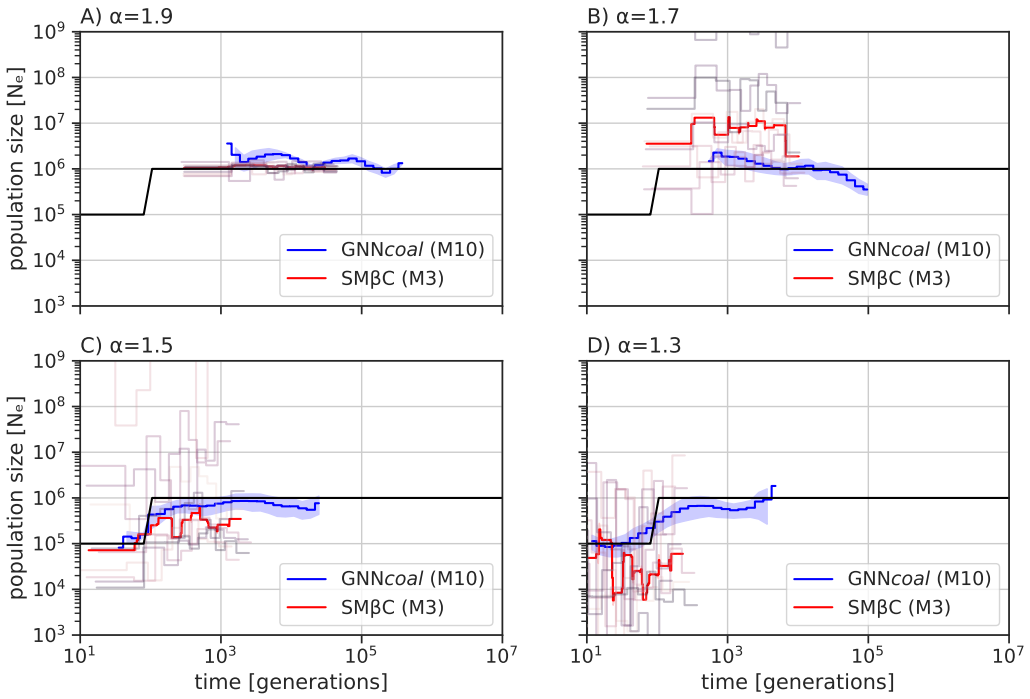


Figure 5.17 Best-case convergence estimations of SMβC and GNN*coal* under a Beta coalescent. Estimations of past demographic history by SMβC using 10 sequences of 100 Mb in red (median) and by GNN using 10 sequences and 500 trees in blue (mean and CI95) when population undergoes a "decrease" demographic scenario (black) under 4 different α values A) 1.9, B) 1.7, c) 1.5 and D) 1.3. The recombination and mutation rate are set to 1×10^{-8} per generation per bp.

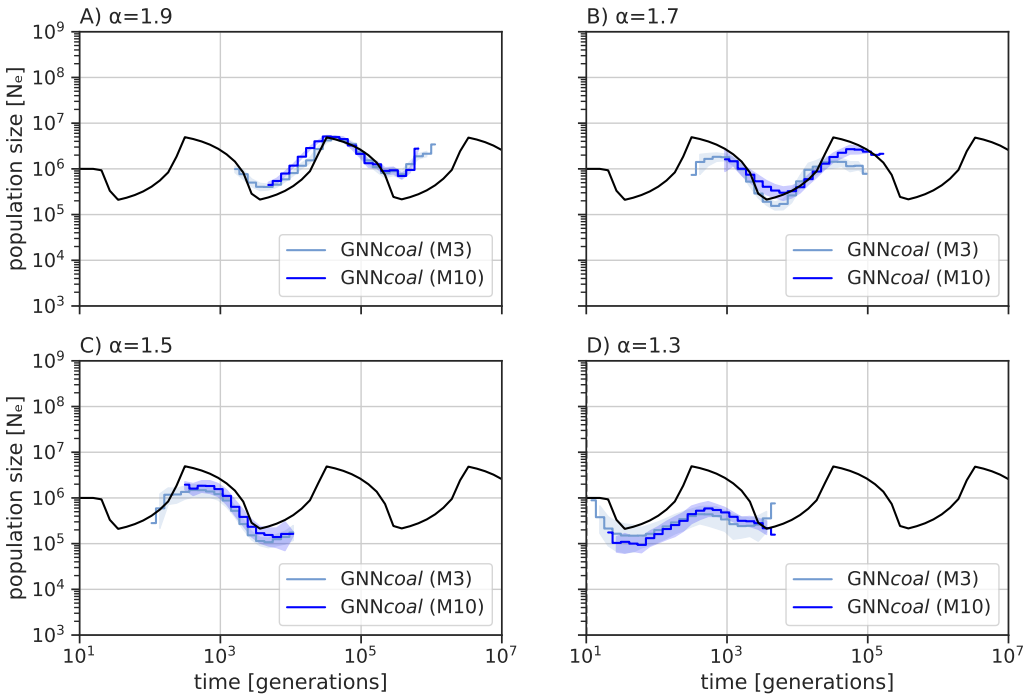


Figure 5.18 Best-case convergence estimations of two GNNs under a Beta coalescent Estimations of past demographic history by GNN*coal* using 10 sequences in blue and by a GNN*coal* using 3 sequences in ice blue when population undergoes a "Sawtooth" demographic scenario (black) under 4 different α values A) 1.9, B) 1.7, c) 1.5 and D) 1.3 (mean and CI95). The recombination and mutation rate are set to 1×10^{-8} per generation per bp.

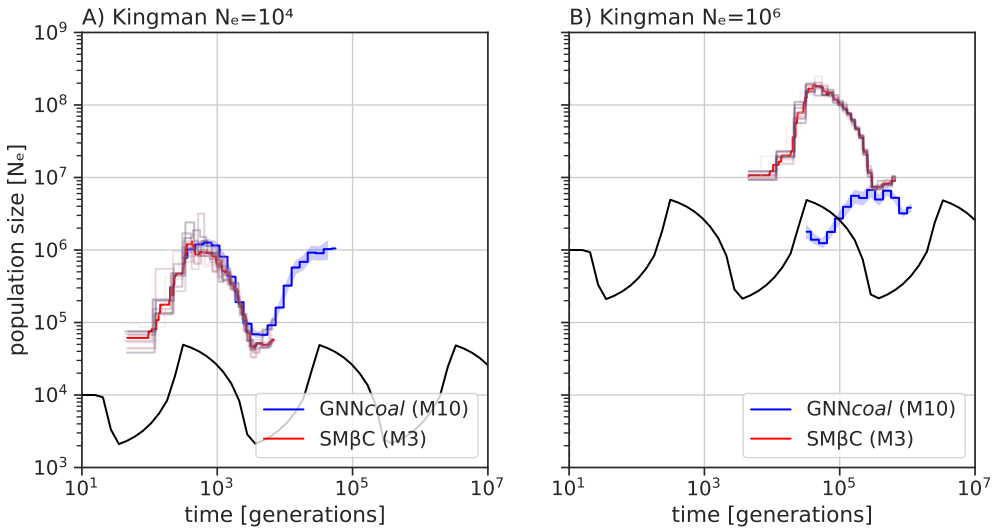


Figure 5.19 Best-case convergence estimations of SMβC and GNN*coal* under a Kingman coalescent. Estimations of past demographic history by SMβC using 10 sequences and 10 Mb in red (median) and by GNN using 10 sequences and 500 trees in blue (mean and CI95) when population undergoes a "sawtooth" demographic scenario (black) under 2 different population sizes A) $N_e = 10^4$ and B) $N_e = 10^6$. The recombination and mutation rate are set to 1×10^{-8} per generation per bp.

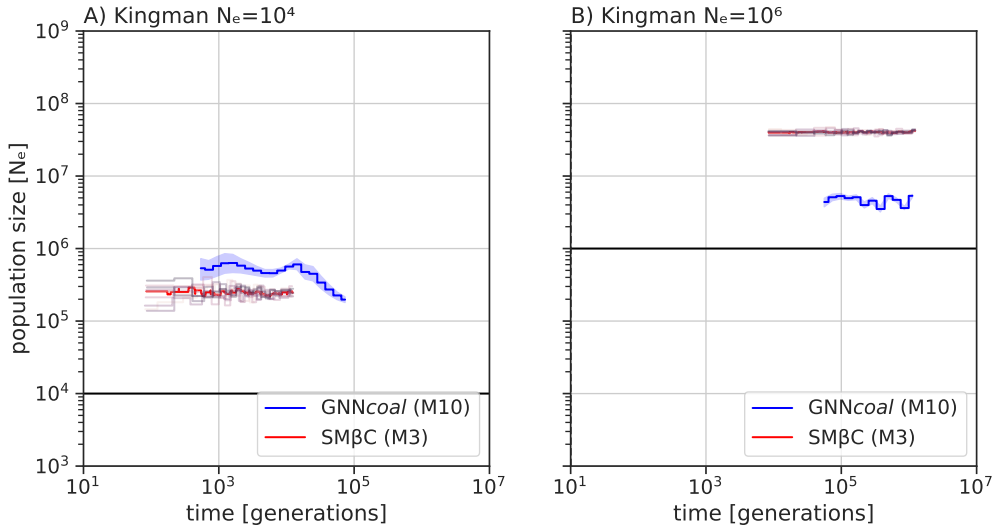


Figure 5.20 Best-case convergence estimations of SMβC and GNN*coal* under a Kingman coalescent. Estimations of past demographic history by SMβC using 10 sequences and 10 Mb in red (median) and by GNN using 10 sequences and 500 trees in blue (mean and CI95) when population undergoes a "constant" demographic scenario (black) under 2 different population sizes A) $N_e = 10^4$ and B) $N_e = 10^6$. The recombination and mutation rate are set to 1×10^{-8} per generation per bp.

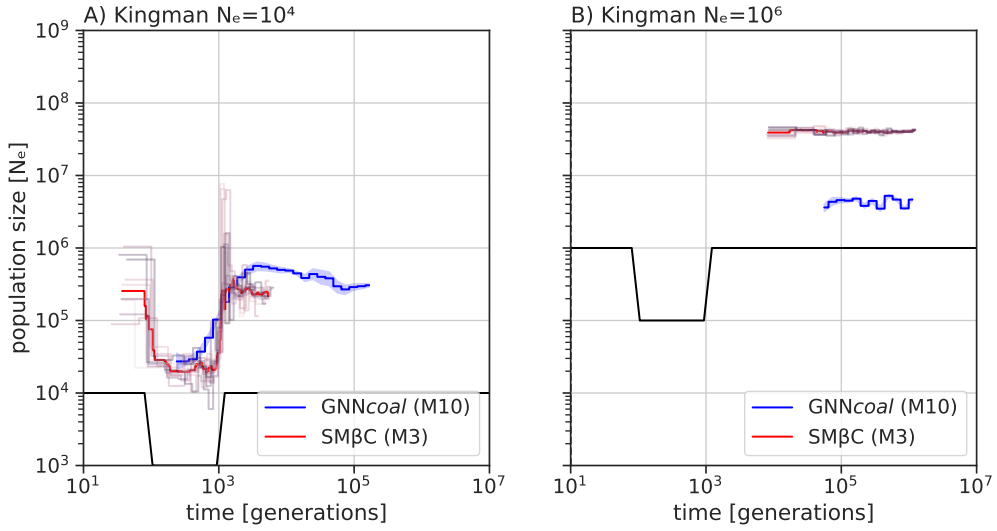


Figure 5.21 Best-case convergence estimations of SMβC and GNN*coal* under a Kingman coalescent. Estimations of past demographic history by SMβC using 10 sequences and 10 Mb in red (median) and by GNN using 10 sequences and 500 trees in blue (mean and CI95) when population undergoes a "bottleneck" demographic scenario (black) under 2 different population sizes A) $N_e = 10^4$ and B) $N_e = 10^6$. The recombination and mutation rate are set to 1×10^{-8} per generation per bp.

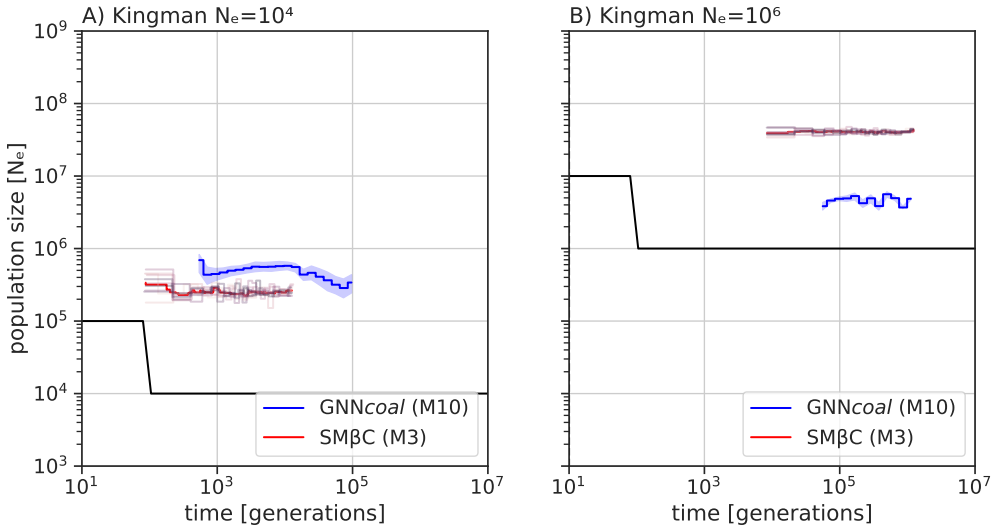


Figure 5.22 Best-case convergence estimations of SMβC and GNN*coal* under a Kingman coalescent. Estimations of past demographic history by SMβC using 10 sequences and 10 Mb in red (median) and by GNN using 10 sequences and 500 trees in blue (mean and CI95) when population undergoes a "increase" demographic scenario (black) under 2 different population sizes A) $N_e = 10^4$ and B) $N_e = 10^6$. The recombination and mutation rate are set to 1×10^{-8} per generation per bp.

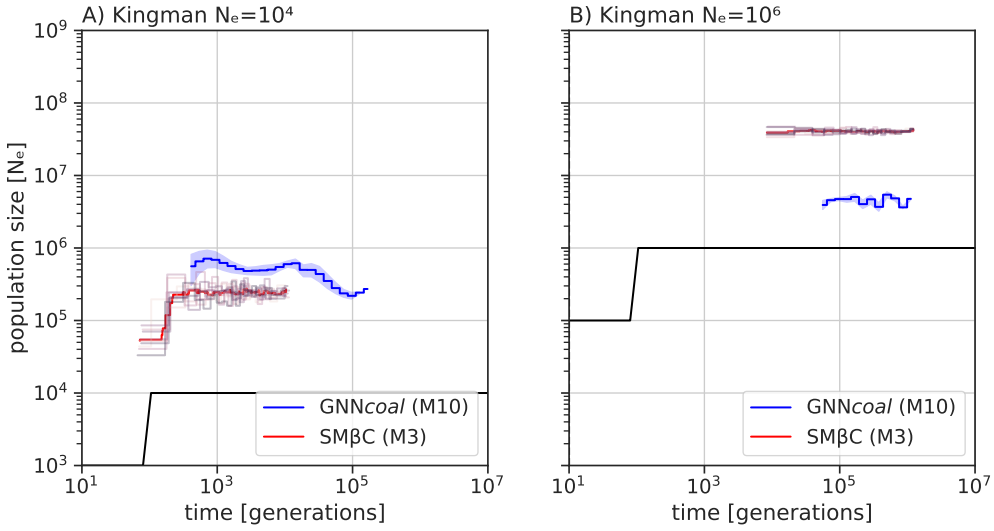


Figure 5.23 Best-case convergence estimations of SMβC and GNN*coal* under a Kingman coalescent. Estimations of past demographic history by SMβC using 10 sequences and 10 Mb in red (median) and by GNN using 10 sequences and 500 trees in blue (mean and CI95) when population undergoes a "decrease" demographic scenario (black) under 2 different population sizes A) $N_e = 10^4$ and B) $N_e = 10^6$. The recombination and mutation rate are set to 1×10^{-8} per generation per bp.

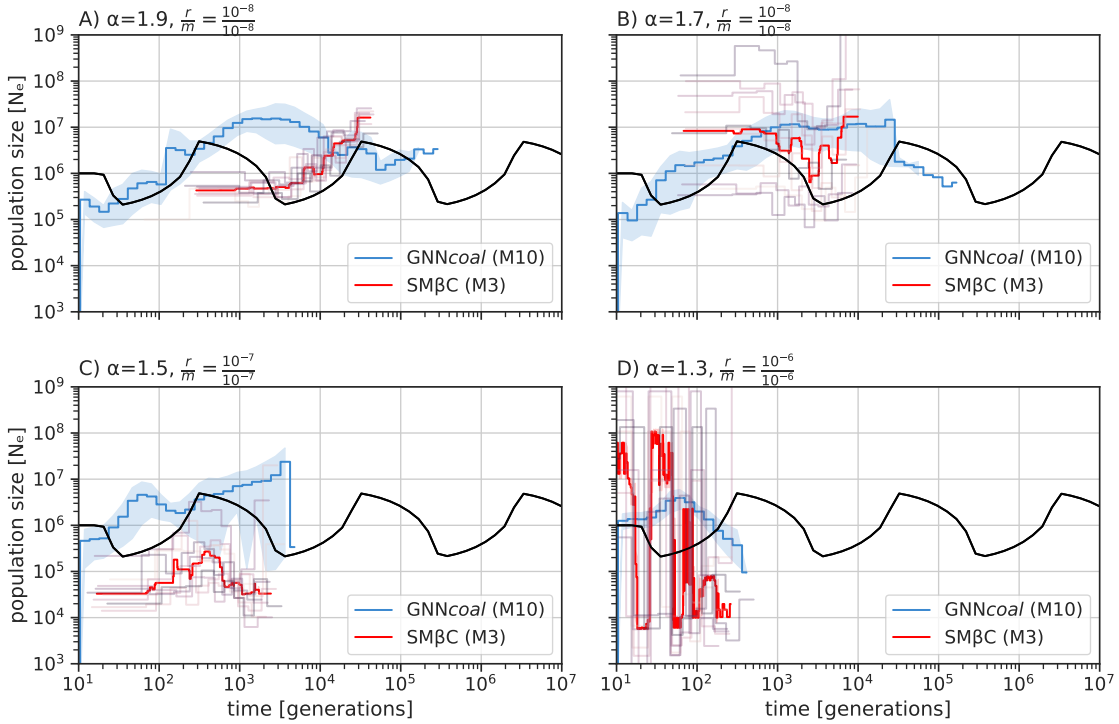


Figure 5.24 Demographic inference estimations of SMβC and GNN*coal* using ARGweaver output under a sawtooth scenario.. Estimations of the ARG is first performed by ARGweaver using 10 sequences of10 Mb. Estimations of past demographic history is then performed on the inferred ARG by SMβC in red (median) and by GNN using 10 sequences and 500 trees in blue (mean and CI95) when population undergoes a "sawtooth" demographic scenario (black). The recombination and mutation rate per generation per bp are set to 1×10^{-8} for α equal to 1.9 and 1.7 , 1×10^{-7} for α equal to 1.5 and set to 1×10^{-6} for α equal to 1.3.

Figure 5.25 Demographic inference estimations of SMβC on simulated data under a sawtooth demographic scenario.Estimated population size by SMβC from simulated sequence data under a sawtooth demographic scenario (black) and the Beta coalescent. Estimations of past demographic history by SMβC using 10 sequences of 10 Mb under different α values (1.9 in red, 1.7 in orange, 1.5 in green and 1.3 in blue. The recombination and mutation rate per generation per bp are set to 1×10^{-8} for α equal to 1.9 and 1.7 , 1×10^{-7} for α equal to 1.5 and set to 1×10^{-6} for α equal to 1.3 .

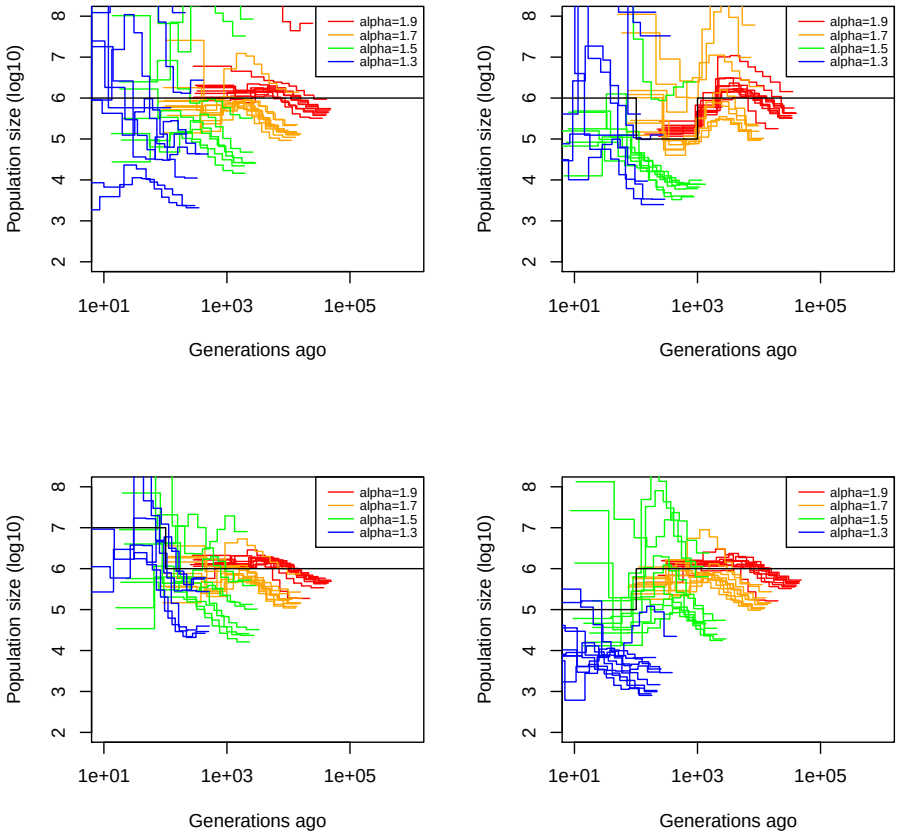


Figure 5.26 Demographic inference estimations of SMβC on simulated data under 4 difference demographic scenarios. Estimated population size by SMβC from simulated sequence data under 4 demographic scenarios (black) (Constant population size in A, Bottleneck in B, Sudden increase of population size in C, and sudden decrease in D). Sequences are simulated under the Beta coalescent. Estimations of past demographic history by SMβC using 10 sequences of 10 Mb under different α values (1.9 in red, 1.7 in orange, 1.5 in green and 1.3 in blue. The recombination and mutation rate per generation per bp are set to 1×10^{-8} for α equal to 1.9 and 1.7 , 1×10^{-7} for α equal to 1.5 and set to 1×10^{-6} for α equal to 1.3 .

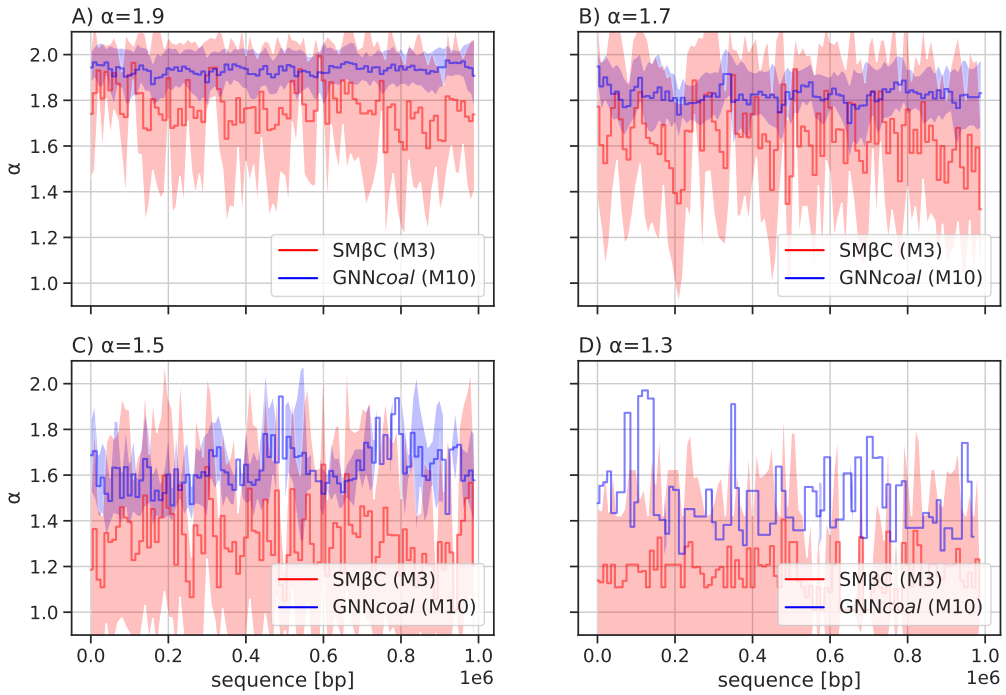


Figure 5.27 Averaged estimations of α by the GNN$coal$ approach and the SMβC along the sequence Estimations of α by SMβC using 20 sequences 1 Mb in red and blue by GNN$coal$ using 10 sequences and in blue under 4 different α values 1.9,1.7,1.5 and 1.3 for a constant demography of size 10^6 in A),B),C) and D) (mean and standard deviation for both methods). The GNN$coal$ used at most 20 coalescent trees per batch along the sequence. Both methods used a window size of 10^3 bp. The recombination and mutation rate are set to 1×10^{-8} per generation per bp.

5.11.2 Supplementary Tables

scenario	True α	α^*:SMβC,M=3	α^*:SMβC,M=4	α^* : GNN, M=3	α^* : GNN, M=10
Constant	1.9	1.88 (0.04)	1.84 (0.1)	1.90 (0.03)	1.88 (0.03)
Bottleneck	1.9	1.86 (0.07)	1.82 (0.16)	1.91 (0.03)	1.91 (0.02)
Increase	1.9	1.88 (0.05)	1.86 (0.12)	1.91 (0.03)	1.88 (0.02)
Decrease	1.9	1.89 (0.02)	1.88 (0.07)	1.93 (0.03)	1.88 (0.04)
Sawtooth	1.9	1.84 (0.06)	1.74 (0.11)	1.96 (0.02)	1.94 (0.02)
Constant	1.7	1.53 (0.12)	1.63 (0.09)	1.78 (0.03)	1.77 (0.03)
Bottleneck	1.7	1.67 (0.14)	1.57 (0.09)	1.78 (0.04)	1.76 (0.03)
Increase	1.7	1.57 (0.14)	1.63 (0.17)	1.77 (0.01)	1.74 (0.03)
Decrease	1.7	1.54 (0.15)	1.57 (0.14)	1.80 (0.04)	1.78 (0.04)
Sawtooth	1.7	1.67 (0.14)	1.60 (0.19)	1.81 (0.02)	1.74 (0.03)
Constant	1.5	1.58 (0.10)	1.56 (0.14)	1.61 (0.02)	1.56 (0.02)
Bottleneck	1.5	1.53 (0.11)	1.55 (0.19)	1.58 (0.02)	1.53 (0.03)
Increase	1.5	1.50 (0.12)	1.52 (0.14)	1.64 (0.05)	1.61 (0.03)
Decrease	1.5	1.57 (0.17)	1.55 (0.14)	1.60 (0.03)	1.55 (0.03)
Sawtooth	1.5	1.47 (0.11)	1.55 (0.19)	1.61 (0.04)	1.56 (0.02)
Constant	1.3	1.40 (0.13)	1.45 (0.14)	1.39 (0.03)	1.33 (0.03)
Bottleneck	1.3	1.43 (0.13)	1.48 (0.20)	1.34 (0.02)	1.28 (0.04)
Increase	1.3	1.45 (0.14)	1.51 (0.24)	1.42 (0.02)	1.38 (0.02)
Decrease	1.3	1.44 (0.16)	1.50 (0.12)	1.4 (0.02)	1.30 (0.03)
Sawtooth	1.3	1.39 (0.09)	1.39 (0.07)	1.41 (0.03)	1.36 (0.05)

Table 5.3 Average estimated values of α by SMβC and the GNN*coal* approach over ten repetitions using the true ARG of 10 sequences of 100 Mb with recombination and mutation rate set to 1×10^{-8} per generation per bp under a Beta coalescent process (with different α parameter). The analysis were run on five different demographic scenarios (Constant population size, Bottleneck, Sudden increase, Sudden decrease and a Sawtooth demography). The standard deviation is indicated in brackets. These results complement the demographic estimates from Figures 5.4 and 5.5, as well as 5.14 to 5.17

scenario	True α	α^*:SMβC,M=3	α^*:SMβC,M=4
Constant	1.7	1.64 (0.12)	1.56 (0.11)
Bottleneck	1.7	1.59 (0.15)	1.55 (0.12)
Increase	1.7	1.65 (0.10)	1.61 (0.15)
Decrease	1.7	1.71 (0.10)	1.67 (0.12)
Sawtooth	1.7	1.60 (0.18)	1.51 (0.20)
Constant	1.5	1.46 (0.11)	1.51 (0.16)
Bottleneck	1.5	1.50 (0.14)	1.45 (0.16)
Increase	1.5	1.56 (0.13)	1.47 (0.15)
Decrease	1.5	1.51 (0.11)	1.47 (0.12)
Sawtooth	1.5	1.40 (0.13)	1.44 (0.17)
Constant	1.3	1.38 (0.15)	1.37 (0.08)
Bottleneck	1.3	1.34 (0.10)	1.34 (0.06)
Increase	1.3	1.34 (0.08)	1.46 (0.17)
Decrease	1.3	1.41 (0.15)	1.44 (0.09)
Sawtooth	1.3	1.35 (0.07)	1.34 (0.05)

Table 5.4 Average estimated values of α by SMβC over ten repetitions using simulated sequence data of 10 sequences of 100 Mb with recombination and mutation rate set to 5×10^{-8} for $\alpha = 1.7$ per generation per bp and 5×10^{-7} for α equal to 1.5 and 1.3 under a Beta coalescent process (with different α values). The coefficient of variation is indicated in brackets..

scenario	True α	α^*:SMβC,M=3	α^*:SMβC,M=4	α^* : GNN, M=3	α^* : GNN, M=10
Constant	2	1.98 (0.023)	1.97 (0.023)	1.99 (0.002)	1.99 (0.005)
Sawtooth	2	1.96 (0.027)	1.87 (0.064)	1.98 (0.003)	1.99 (0.005)
Bottleneck	2	1.96 (0.047)	1.96 (0.04)	1.99 (0.002)	1.99 (0.003)
Decrease	2	1.98 (0.022)	1.99 (0.012)	1.98 (0.009)	1.99 (0.005)
Increase	2	1.97 (0.020)	1.97 (0.025)	1.99 (0.002)	1.99 (0.004)

Table 5.5 Average estimated values of α by SMβC and the GNN$coal$ approach over ten repetitions using the true ARG of 10 sequences of 100 Mb for SMβC and at 500 trees for the GNN$coal$ with recombination and mutation rate set to 1×10^{-8} per generation per bp under a Beta coalescent process (with different α parameter) with $N_e = 10^4$ at generation 0. The standard deviation is indicated in brackets. These results complement the estimations of demography from Figure 5.19 to Figure 5.23 .

scenario	True α	α^*:SMβC,M=3	α^*:SMβC,M=4	α^* : GNN, M=3	α^* : GNN, M=10
Sawtooth	1.9	1.77 (0.05)	1.86 (0.05)	1.55 (0.08)	2.05 (0.01)
Sawtooth	1.7	1.56 (0.13)	1.79 (0.1)	1.7 (0.11)	2.00 (0.06)
Sawtooth	1.5	1.66 (0.11)	1.92 (0.07)	1.61 (0.07)	1.96 (0.04)
Sawtooth	1.3	1.41 (0.13)	1.53 (0.20)	1.30 (0.06)	1.78 (0.08)

Table 5.6 Average estimated values of α by SMβC and the GNN$coal$ approach using ARG inferred by ARGweaver over ten repetitions using 10 simulated sequences of 10 Mb with recombination and mutation rate set to 1×10^{-8} per generation per bp under a Beta coalescent process (with different α parameter). The analysis were run under Sawtooth demographic scenario. The standard deviation is indicated in brackets. These results complement the demographic estimates from Figure 5.11.1

6 General Discussion

Before moving on to the placement of our work into the broader research context, highlighting the current limitations, and providing references to future work, we start by giving a summary of the main results of each individual Chapter starting with the Introduction.

6.1 Summary

We opened the thesis by highlighting the vast stochastic complexity underlying the evolutionary processes, and by providing a short historical overview of the theoretical groundwork, which gave rise to the field of *population genetics*. We then went on to call attention to perspective shift and the importance of *Coalescent Theory*, both with regards to providing a framework for understanding evolution as well as for bestowing the means for large-scale simulations. Afterward, we went on to give an in-depth overview of DL methods and their promise to conquer some of the natural complexity in the field through the gradient-based approximation and deduction through thousands of *trainable* parameters of the networks. This point of view can be regarded as a kind of top-down approach, where *understanding* the genomic signatures are outsourced to the DL algorithm of choice. The other bottom-up direction is described through modeling a process of interest directly either mathematically or through simulations. Especially in the later case of stochastic simulations extending the model is often more straightforward since through a detailed and precise understanding of the biology, *i.e.* life-cycle, simulation code can often be extended intuitively. The resulting signatures are then captured and described in the form of popular summary statistics or measurements that trace individual alleles. In Chapters 2 and 3 we used this simulation-based modeling approach to push our understanding of the biological traits of dormancy and variations of the offspring distribution both under the influence of positive selection. Our studies of the *weak dormancy* model of Kaj and colleagues confirmed previously known scaling expectations of the Coalescent accompanied by the expected decrease of genetic drift proportional to the inverse quadratic of the germination rate and a respective linear scaling of the recombination rate due to the assumed inability of recombination gametes in the dormant stage (Kaj et al., 2001; Koopmann et al., 2017; Živković and Tellier, 2018). We went on to show that the probability of fixation of a positively segregating allele is unaffected by the germination rate and therefore strength of the seed bank. However, the time scales equally to the scaling of the coalescence events, namely also quadratically, of which the stochastic phases, where the allele is either at low or very high frequencies, contribute a disproportional amount of time. Lastly, we showed the implications for the expected hard sweep signatures and their inherent detectability. As announced through increased linkage decay in one of the introductory measurements of the chapter, the sweep appears narrower as well as deeper due to the high diversity gradient in seed banks, which was shown to improve the detection possibility.

In Chapter 3 we extend our perspective and conceptualize a picture of the life-cycle in which both dormancy as an effective measure of the lifespan extension or general viability as well as the fecundity phase can be embedded. In that concept, it becomes important to be aware of the underlying biological implications of certain model choices. For example, the placement of mutations can fall on the parents and thus be heritable or onto the individual offspring themselves, creating a new level of diversity between each descendent of the same parent. Here we observed that the expected number of mutations is always tied to the population size, but the variance increases under strong sweepstakes if mutations occur in the parent generation. Going back to the life-cycle concept, we directly can compare the antagonistic relationship between dormancy and offspring distribution changes. While the former is observed to have a stabilizing effect on drift (decreasing the randomness of the system and maintaining variation) the opposite is the case for species with sweepstake events. Genetic drift is increased, the stronger the frequency of the occurrences of the sweepstakes becomes, leading to the increase in fixation probability of neutral alleles and a decrease for selective alleles. In contrast to the increase in fixation time for the dormancy trait, sweepstakes have the opposite effect of decreasing the time and thus making it possible to push an allele to fixation sometimes multiple orders of magnitude faster than under low variance offspring distributions. Fluctuation conditions for selective alleles seemed to rely mostly on the initial phase, while population size changes only had a negligible effect on fixation times or probabilities. Selection on the fecundity was shown to be more important with regard to fixation probability on the high sweepstake spectrum, while our viability selection model became more important during the low sweepstake spectrum. And lastly, both selection models showed a synergistic non-additive effect during a strong selection regime. The insights from Chapter 3 extended our understanding of adaptation capability through the ways in which stochasticity is increased and painted a more complete picture of the importance of life-cycle modeling.

In the next part, Chapter 4, we switched from simulations to inference placing the β-Coalescent as a backward in-time model, which describes large offspring variance, at the center of our investigations. Here, we developed the SMβC method, alongside a DL-based approach, with the aim of increasing our understanding of necessary model assumptions for the estimation of parameters when being confronted with real data. We started by providing fine-scaled demographic inferences using our methods and went on to estimate the strength of the sweepstakes globally and locally along the genome by capturing the α parameter of the β-Coalescent, as a proxy for the scale of the strength of multiple multiple merger events. In doing so, we differentiated the capabilities of both methods and their respective distinctness, *i.e.* the DL method is flexible and fast to be trained on different sample sizes and can more easily capture larger parts of the ARG, while SMβC is only formulated up to a sample size of four. But the most striking difference is the implicit Markovian assumption of the SMC, a property which especially doesn't hold under multiple merger conditions and thus requires better modeling approaches to capture long-range correlations. Lastly, we highlight the similarities of selection and neutrally occurring sweepstake events and provide a proof of concept, which allows us to look for variations of local changes in α as an estimator for selection events, specifically but not limited to species with rather low offspring variances, as the power is expected to decrease when mixing both processes because an issue of non-identifiability could arise.

We now will continue by placing our methods into the broader research context by highlighting adjacent studies immediately to be relevant to our results.

6.2 Contemporary and adjacent research landscape

Integrating life history traits into models remains to be an open frontier, necessary to overcome for accurate non-human model inferences. The behavior of species altered by the expression of the trait has proven to shape the underlying genomic architecture and continues to be a source of complexity which requires to be carefully accounted for when planning to understand and predict complex signatures on the genetic level and beyond.

With regards to dormancy, the complexity of the trait arises mainly due to its simple abstractable formulation of having a population that is active and a population that is dormant with some migration in between. Lennon and colleagues highlight the generalizable properties of this trait and give a detailed overview with relevance ranging from something as small as transcriptional regulation of genes (Dworkin and Losick, 2005), to cells including cancer (Phan and Croucher, 2020), to plants and entire ecosystems (Lamont et al., 2020; Lennon et al., 2021b). However, since this thesis only looked at one of such dormancy models (weak dormancy by Kaj et al., 2001), it is relevant to mention the strong counterpart (Blath et al., 2015, 2016), whose interaction with selection remains unexplored, but interestingly is predicted to have the capacity of generating long-range genomic signals, due to the model's capability of generating *star-shaped* genealogies (Cordero et al., 2022). Further models introduce or study the notion of metapopulations and their mathematical properties in combination with dormancy (Louvet, 2022) or highlight the necessity to account for traits with regards to demographical inference (Živković and Tellier, 2012; Sellinger et al., 2020).

Switching for a moment away from the lifespan to the offspring generation phase, whose relevance and pervasiveness have been thoroughly justified by the review article of Eldon, 2020b, increasing attention has been crystallized into fitting multiple merger coalescence processes to real data. On the one hand sweepstakes offer an explanation of the prevalent U-shape SFS found across many species (Freund et al., 2023), and on the other hand it becomes more and more necessary to be able to distinguish between neutral sweepstake processes or selectively driven processes in the genome, both of which yield similar signatures of decreased genetic diversity and therefore run into issues of model differentiation (Korfmann et al., 2022b). Having an estimate of the strength of the underlying sweepstake capability of a species may prove as an important initial step to the assessment of putative regions under selections, similar to how demographic estimates are important to be excluded *a priori* before rushing to conclusions about non-neutral evolutionary pressures shaping the genome (Korfmann et al., 2022b; Miró Pina et al., 2023).

Following up on selection and going beyond diversity or SFS-based statistics for selective sweep detection, we emphasize, that more complicated approaches for selection detection may be necessary, including a range of summary statistics or machine learning-based approaches to cope with biases of synonymous and non-synonymous loci, that may occur under frequent sweepstakes or *vice versa* establish a way of quantifying the emerging bias to learn something about the strength of suggested sweepstakes. Last by not least we highlight again the importance of phylogenetic dating in the light of sweepstake organisms as our findings indicate that solely relying on substitution rate may be biased, but needs to be assessed, *i.e.* through simulations, to understand the scope of the arising error when neglecting offspring variations (Zuckerkandl and Pauling, 1965; dos Reis et al., 2016).

After shining a spotlight on a by no means exhaustive list of adjacent articles, we continue by naming limits of both evolutionary simulations and inferences as deployed throughout the projects.

6.3 Limitations of simulations and inferences

Since the stochasticity of nature is endless and computational resources are limited, it is always an open question of how much model complexity is required to capture most of the variation of a piece of genome sequence. Additionally, multiple processes lead to similar signatures, either locally or globally or both, and may require multiple data layers to disentangle them. Therefore, it becomes necessary to investigate the full range of possible scenarios explaining the given observation. Johri et al., 2022a argued to follow a flow-chart-like approach for population genetic inference, which described an ABC strategy for the disentanglement of a set of hypotheses. However, the models analyzed were relatively simple comparing neutral models to sweeps and bottlenecks and when uncertainty may increase, could benefit state-of-the-art DL approaches, which ideally incorporate the advantages of ABC, *i.e.* giving distributional rather than point-estimates and therefore an insight into parameter uncertainty. Though DL alone certainly will not easily solve to overwhelming complexity it may serve here as a valuable tool, when applied carefully to capture a large part of the variance without relying on random ABC-like exploration of the parameter space (Cranmer et al., 2020b).

Another yet fundamental limitation is the speed of simulation. When considering an individual-based approach, exploring variations of multiple replicates is limited by the Markovian dependency of simulating each generation sequentially. The Coalescent elegantly approximates this process, but it is less accessible due to its reliance on the rigorous mathematical description required, which becomes quickly complex when adding a few fundamental processes, in the form of recombination, demography, migration, selection, or variations of the offspring distribution. Therefore, a popular approach has taken shape in combining both perspectives, simulating only part of the more complex and difficult-to-model process forward in time and attaching a neutral coalescence part to the initial simulations in a post-processing step (Kelleher et al., 2018; Haller et al., 2019a).

A new generative kind of paradigm is beginning to emerge which could address issues of simulation speed through learning the underlying distribution of the simulations with generative networks and (after training) producing new samples following the same approximated data distribution. This approach is similar to older naive versions of resampling an original dataset, with the difference that the DL algorithm would be expected to create more realistic samples by learning any kind of interactions between the samples and making estimations of the underlying parameter space. Interesting results that validate this approach have come forward in the form of generating new samples from real genomic datasets, *i.e.* to preserve privacy in medical contexts, and could show that summary statistics remain preserved in newly generated data (Yelmen et al., 2021b, 2023; Booker et al., 2023; Yelmen and Jay, 2023).

Another limitation manifesting itself in the field of population genetics is the lack of benchmarks when developing inference algorithms. The usual procedure involves training networks on some unique custom dataset, and evaluating it separately on a unique dataset created by the same group, that developed the inference algorithm. This *modus operandi* is susceptible to unconscious bias during the creation of the datasets and may lead to the neglect of biological or technical sources of variation. Therefore, evaluating datasets on reoccurring tasks of for example the inference or introduced loss of power through selection may be best assessed by external and carefully created simulated datasets as provided by the *stdpopsim* consortium. However, it is important to highlight that the main objective should always be the scientific knowledge discovery through our created methods,

rather than marginal performance increase over previous methods. Nevertheless, having a standardized way of method evaluation is needed in the field to be able to quantify methodological progress.
After mentioning the limiting factors, we now continue onward to the description of future research directions.

6.4 Future work

In addition to focusing on novel or adjacent research directions, we continue by highlighting some of the more straightforward extensions of the individual Chapters.
Starting with our effective lifespan-extending trait of dormancy. A natural extension that remains to be explored either mathematically or by simulations is the degree to which the weak and strong models overlap in their respective signatures or expressed differently where the edge cases of both models could lead to issues of model identification. If in fact, the weak dormancy model is either part of the strong version or if a new model can be formulated that incorporates both models are still open and unanswered questions. Likewise, how positive selection affects the strong version still requires to be thought of, possibly in concert with other types of selection. Simultaneously the questions if and when mutations occur in the seed bank could be explored again in light of new and emerging experimental evidence. One likely model choice may be through letting mutations occur at two different rates.
Continuing with the simulation or modeling part, concurrently building dormancy and variations of offspring into the same model likewise can be worthy of exploration. As a baseline model for comparison, it may be further interesting to compare not only to Wright-Fisher-Model but also the Moran model, because of its inherent age structure and known expectations.
Making the models more realistic will also become more important, of which the most immediate step is to assess linkage effects between sites, and by extension having a detailed idea of the underlying recombination map is necessary.
On the inference side, two routes remain to be explored. One of each is to address questions of possible neural network interpretability enabled through the usage of graphs as input over genotype matrices and to figure out how additions of *attention* modules could create the possibility of scientific knowledge discovery while adding methodological improvements (Vaswani et al., 2017). The second route aims to add the possibility of robust data application, ideally through an end-to-end learnable framework. Here inferring the ARG itself would be outsourced to a generative network, rather than using any of the existing non-DL ARG algorithms, continuously improving the ARG or ARG-like generative output, while inferring any parameter of interest by a second parameter-estimation network, like GNN$coal$ or future more efficient iterations.
Even though this was all but a brief outline of possible directions, it may serve as a reference point, when revisiting the individual Chapters.

6.5 Conclusion

While we showed the extended promises and powers of simulation-based modeling for answering open questions of population genetics, fitting our insights to real biological data will become non-trivial. However, through the ongoing computational progress both on the software side, *i.e.* in the form of

libraries like *tskit* (Kelleher et al., 2018) or *PyTorch* (Paszke et al., 2017) or ever-increasing hardware capabilities, research questions may be addressed in an ever-increasing speed and the stochasticity of nature and its current accompanied post-modern *sentiment* (Koonin, 2009; Stoltzfus, 2017) of insurmountably will become a feeling of the past.

Bibliography

J. R. Adrion, C. B. Cole, N. Dukler, J. G. Galloway, A. L. Gladstein, G. Gower, C. C. Kyriazis, A. P. Ragsdale, G. Tsambos, F. Baumdicker, J. Carlson, R. A. Cartwright, A. Durvasula, I. Gronau, B. Y. Kim, P. McKenzie, P. W. Messer, E. Noskova, D. Ortega-Del Vecchyo, F. Racimo, T. J. Struck, S. Gravel, R. N. Gutenkunst, K. E. Lohmueller, P. L. Ralph, D. R. Schrider, A. Siepel, J. Kelleher, and A. D. Kern. A community-maintained standard library of population genetic models. *eLife*, 9:e54967, jun 2020a. ISSN 2050-084X. doi: 10.7554/eLife.54967. URL https://doi.org/10.7554/eLife.54967.

J. R. Adrion, J. G. Galloway, and A. D. Kern. Predicting the Landscape of Recombination Using Deep Learning. *Molecular Biology and Evolution*, 37(6):1790–1808, 02 2020b. ISSN 0737-4038. doi: 10.1093/molbev/msaa038. URL https://doi.org/10.1093/molbev/msaa038.

A. M. Alaa and M. van der Schaar. Demystifying black-box models with symbolic metamodels. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/567b8f5f423af15818a068235807edc0-Paper.pdf.

N. Alachiotis and P. Pavlidis. Scalable linkage-disequilibrium-based selective sweep detection: a performance guide. *GigaScience*, 5:7, 2016. ISSN 2047-217X. doi: 10.1186/s13742-016-0114-9.

N. Alachiotis, A. Stamatakis, and P. Pavlidis. OmegaPlus: a scalable tool for rapid detection of selective sweeps in whole-genome datasets. *Bioinformatics*, 28(17):2274–2275, Sept. 2012. ISSN 1460-2059, 1367-4803. doi: 10.1093/bioinformatics/bts419. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts419.

F. Alberti, C. Herrmann, and E. Baake. Selection, recombination, and the ancestral initiation graph. *THEORETICAL POPULATION BIOLOGY*, 142:46–56, DEC 2021. ISSN 0040-5809. doi: 10.1016/j.tpb.2021.08.001.

S. Anand, E. Mangano, N. Barizzone, R. Bordoni, M. Sorosina, F. Clarelli, L. Corrado, F. Martinelli Boneschi, S. D'Alfonso, and G. De Bellis. Next generation sequencing of pooled samples: Guideline for variants' filtering. *Scientific Reports*, 6(1):33735, Sep 2016. ISSN 2045-2322. doi: 10.1038/srep33735. URL https://doi.org/10.1038/srep33735.

M. Ancona, C. Oztireli, and M. Gross. Explaining deep neural networks with a polynomial time algorithm for shapley values approximation. In K. Chaudhuri and R. Salakhutdinov, editors, *INTERNATIONAL CONFERENCE ON MACHINE LEARNING, VOL 97*, volume 97 of *Proceedings of*

Machine Learning Research, 2019a. 36th International Conference on Machine Learning (ICML), Long Beach, CA, JUN 09-15, 2019.

M. Ancona, C. Öztireli, and M. H. Gross. Explaining deep neural networks with a polynomial time algorithm for shapley values approximation. *CoRR*, abs/1903.10992, 2019b. URL http://arxiv.org/abs/1903.10992.

J. T. Anderson, A. M. Panetta, and T. Mitchell-Olds. Evolutionary and ecological responses to anthropogenic climate change: update on anthropogenic climate change. *Plant physiology*, 160(4): 1728–1740, 2012.

M. Arjovsky and L. Bottou. Towards principled methods for training generative adversarial networks, 2017. URL https://arxiv.org/abs/1701.04862.

E. Arnason and K. Halldorsdottir. Nucleotide variation and balancing selection at the Ckma gene in Atlantic cod: analysis with multiple merger coalescent models. *PEERJ*, 3, FEB 24 2015. ISSN 2167-8359. doi: {10.7717/peerj.786}.

E. Árnason, J. Koskela, K. Halldórsdóttir, and B. Eldon. Sweepstakes reproductive success via pervasive and recurrent selective sweeps. *Elife*, 12:e80781, 2023.

D. Azouri, S. Abadi, Y. Mansour, I. Mayrose, and T. Pupko. Harnessing machine learning to guide phylogenetic-tree search algorithms. *Nature communications*, 12(1):1983–1983, Mar 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-22073-8. URL https://pubmed.ncbi.nlm.nih.gov/33790270. 33790270[pmid].

C. Barbosa, N. Mahrt, J. Bunk, M. Graßer, P. Rosenstiel, G. Jansen, and H. Schulenburg. The genomic basis of rapid adaptation to antibiotic combination therapy in pseudomonas aeruginosa. *Molecular biology and evolution*, 38(2):449–464, 2021.

N. Barghi, J. Hermisson, and C. Schlötterer. Polygenic adaptation: a unifying framework to understand positive selection. *Nature Reviews Genetics*, 21(12):769–781, 2020.

R. D. H. Barrett, L. K. M'Gonigle, and S. P. Otto. The Distribution of Beneficial Mutant Effects Under Strong Selection. *Genetics*, 174(4):2071–2079, 12 2006. ISSN 1943-2631. doi: 10.1534/genetics.106.062406. URL https://doi.org/10.1534/genetics.106.062406.

G. V. Barroso and J. Y. Dutheil. Mutation rate variation shapes genome-wide diversity in *Drosophila melanogaster*. preprint, Evolutionary Biology, Sept. 2021. URL http://biorxiv.org/lookup/doi/10.1101/2021.09.16.460667.

G. V. Barroso, N. Puzovic, and J. Y. Dutheil. Inference of recombination maps from a single pair of genomes and its application to ancient samples. *PLOS Genetics*, 15(11), NOV 2019. ISSN 1553-7404. doi: {10.1371/journal.pgen.1008449;10.1371/journal.pgen.1008449.r001;10.1371/journal.pgen.1008449.r002;10.1371/journal.pgen.1008449.r003;10.1371/journal.pgen.1008449.r004}.

C. Battey, P. L. Ralph, and A. D. Kern. Predicting geographic location from genetic variation with deep neural networks. *eLife*, 9:e54507, June 2020. ISSN 2050-084X. doi: 10.7554/eLife.54507. URL https://doi.org/10.7554/eLife.54507.

C. J. Battey, G. C. Coffing, and A. D. Kern. Visualizing population structure with variational autoencoders. 11(1):jkaa036, 2021. ISSN 2160-1836. doi: 10.1093/g3journal/jkaa036. URL https://academic.oup.com/g3journal/article/doi/10.1093/g3journal/jkaa036/6105578.

F. Baumdicker, G. Bisschop, D. Goldstein, G. Gower, A. P. Ragsdale, G. Tsambos, S. Zhu, B. Eldon, E. C. Ellerman, J. G. Galloway, A. L. Gladstein, G. Gorjanc, B. Guo, B. Jeffery, W. W. Kretzschumar, K. Lohse, M. Matschiner, D. Nelson, N. S. Pope, C. D. Quinto-Cortés, M. F. Rodrigues, K. Saunack, T. Sellinger, K. Thornton, H. van Kemenade, A. W. Wohns, Y. Wong, S. Gravel, A. D. Kern, J. Koskela, P. L. Ralph, and J. Kelleher. Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*, 220(3), 12 2021. ISSN 1943-2631. doi: 10.1093/genetics/iyab229. URL https://doi.org/10.1093/genetics/iyab229. iyab229.

F. Baumdicker, G. Bisschop, D. Goldstein, G. Gower, A. P. Ragsdale, G. Tsambos, S. Zhu, B. Eldon, E. C. Ellerman, J. G. Galloway, A. L. Gladstein, G. Gorjanc, B. Guo, B. Jeffery, W. W. Kretzschumar, K. Lohse, M. Matschiner, D. Nelson, N. S. Pope, C. D. Quinto-Cortes, M. F. Rodrigues, K. Saunack, T. Sellinger, K. Thornton, H. van Kemenade, A. W. Wohns, Y. Wong, S. Gravel, A. D. Kern, J. Koskela, P. L. Ralph, and J. Kelleher. Efficient ancestry and mutation simulation with msprime 1.0. *GENETICS*, 220(3), MAR 3 2022. ISSN 0016-6731. doi: 10.1093/genetics/iyab229.

M. A. Beaumont, W. Zhang, and D. J. Balding. Approximate Bayesian Computation in Population Genetics. *Genetics*, 162(4):2025–2035, 12 2002. ISSN 1943-2631. doi: 10.1093/genetics/162.4.2025. URL https://doi.org/10.1093/genetics/162.4.2025.

G. Bertorelle, A. Benazzo, and S. Mona. Abc as a flexible framework to estimate demography over space and time: some cons, many pros. *Molecular Ecology*, 19(13):2609–2625, 2010. doi: https://doi.org/10.1111/j.1365-294X.2010.04690.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-294X.2010.04690.x.

M. Birkner, J. Blath, M. Moehle, M. Steinruecken, and J. Tams. A modified lookdown construction for the Xi-Fleming-Viot process with mutation and populations with recurrent bottlenecks. *arXiv:0808.0412*, 2008. URL https://arxiv.org/abs/0808.0412.

M. Birkner, J. Blath, and B. Eldon. An Ancestral Recombination Graph for Diploid Populations with Skewed Offspring Distribution. *Genetics*, 193(1):255–290, JAN 2013. ISSN 0016-6731. doi: {10.1534/genetics.112.144329}.

G. Bisschop, K. Lohse, and D. Setter. Sweeps in time: leveraging the joint distribution of branch lengths. *GENETICS*, 219(2), OCT 2021a. ISSN 0016-6731. doi: 10.1093/genetics/iyab119.

G. Bisschop, K. Lohse, and D. Setter. Sweeps in time: leveraging the joint distribution of branch lengths. *Genetics*, 219(2):iyab119, Oct. 2021b. ISSN 1943-2631. doi: 10.1093/genetics/iyab119.

J. Blath, A. González Casanova, B. Eldon, N. Kurt, and M. Wilke-Berenguer. Genetic Variability Under the Seedbank Coalescent. *Genetics*, 200(3):921–934, July 2015. ISSN 1943-2631. doi: 10.1534/genetics.115.176818.

J. Blath, A. G. Casanova, N. Kurt, and M. Wilke-Berenguer. A NEW COALESCENT FOR SEED-BANK MODELS. *The Annals of Applied Probability*, 26(2):857–891, 2016. ISSN 1050-5164. URL https://www.jstor.org/stable/43859616.

J. Blath, E. Buzzoni, A. González Casanova, and M. Wilke-Berenguer. Structural properties of the seed bank and the two island diffusion. *Journal of Mathematical Biology*, 79(1):369–392, July 2019. ISSN 1432-1416. doi: 10.1007/s00285-019-01360-5. URL https://doi.org/10.1007/s00285-019-01360-5.

J. Blath, E. Buzzoni, J. Koskela, and M. Wilke Berenguer. Statistical tools for seed bank detection. *Theoretical Population Biology*, 132:1–15, Apr. 2020. ISSN 00405809. doi: 10.1016/j.tpb.2020.01.001. URL https://linkinghub.elsevier.com/retrieve/pii/S0040580920300010.

J. Blath, A. Gonzalez Casanova, N. Kurt, and M. Wilke-Berenguer. The seed bank coalescent with simultaneous switching. *Electronic Journal of Probability*, 25, 2020. ISSN 1083-6489. doi: {10.1214/19-EJP401}.

P. D. Blischak, M. S. Barker, and R. N. Gutenkunst. Chromosome-scale inference of hybrid speciation and admixture with convolutional neural networks. *Molecular Ecology Resources*, 21(8):2676–2688, 2021. doi: https://doi.org/10.1111/1755-0998.13355. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.13355.

M. G. B. Blum and O. François. Non-linear regression models for approximate bayesian computation. *Statistics and Computing*, 20(1):63–73, Jan 2010. ISSN 1573-1375. doi: 10.1007/s11222-009-9116-0. URL https://doi.org/10.1007/s11222-009-9116-0.

S. Boitard, W. Rodríguez, F. Jay, S. Mona, and F. Austerlitz. Inferring Population Size History from Large Samples of Genome-Wide Molecular Data - An Approximate Bayesian Computation Approach. *PLOS Genetics*, 12(3):e1005877, Mar. 2016. ISSN 1553-7404. doi: 10.1371/journal.pgen.1005877. URL https://dx.plos.org/10.1371/journal.pgen.1005877.

E. Bolthausen and A.-S. Sznitman. On ruelle's probability cascades and an abstract cavity method. *Communications in mathematical physics*, 197(2):247–276, 1998.

W. W. Booker, D. D. Ray, and D. R. Schrider. This population doesn't exist: learning the distribution of evolutionary histories with generative adversarial networks. *bioRxiv*, 2022. doi: 10.1101/2022.09.17.508145. URL https://www.biorxiv.org/content/early/2022/09/17/2022.09.17.508145.

W. W. Booker, D. D. Ray, and D. R. Schrider. This population does not exist: learning the distribution of evolutionary histories with generative adversarial networks. *GENETICS*, 224(2):iyad063, May 2023. ISSN 1943-2631. doi: 10.1093/genetics/iyad063. URL https://academic.oup.com/genetics/article/doi/10.1093/genetics/iyad063/7126666.

D. Y. C. Brandt, X. Wei, Y. Deng, A. H. Vaughn, and R. Nielsen. Evaluation of methods for estimating coalescence times using ancestral recombination graphs. *GENETICS*, 221(1), MAY 5 2022. ISSN 0016-6731. doi: 10.1093/genetics/iyac044.

M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst. Geometric deep learning: Going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, jul 2017. doi: 10.1109/msp.2017.2693418. URL https://doi.org/10.1109%2Fmsp.2017.2693418.

J. H. Brown and A. Kodric-Brown. Turnover Rates in Insular Biogeography: Effect of Immigration on Extinction. *Ecology*, 58(2):445–449, 1977. ISSN 0012-9658. doi: 10.2307/1935620. URL https://www.jstor.org/stable/1935620.

E. Brunet, B. Derrida, A. H. Mueller, and S. Munier. Noisy traveling waves: Effect of selection on genealogies. *Europhysics Letters*, 76(1):1–7, OCT 2006. ISSN 0295-5075. doi: {10.1209/epl/i2006-10224-4}.

E. Brunet, B. Derrida, A. H. Mueller, and S. Munier. Effect of selection on ancestry: An exactly soluble case and its phenomenological generalization. *Physical Review E*, 76(4, 1), OCT 2007. ISSN 1539-3755. doi: {10.1103/PhysRevE.76.041104}.

K. E. Burger, P. Pfaffelhuber, and F. Baumdicker. Neural networks for self-adjusting mutation rate estimation when the recombination rate is unknown. *PLOS Computational Biology*, 18(8):1–17, 08 2022a. doi: 10.1371/journal.pcbi.1010407. URL https://doi.org/10.1371/journal.pcbi.1010407.

K. E. Burger, P. Pfaffelhuber, and F. Baumdicker. Neural networks for self-adjusting mutation rate estimation when the recombination rate is unknown. *bioRxiv*, 2022b. doi: 10.1101/2021.09.02.457550. URL https://www.biorxiv.org/content/early/2022/05/17/2021.09.02.457550.

M. Byrska-Bishop, U. S. Evani, X. Zhao, A. O. Basile, H. J. Abel, A. A. Regier, A. Corvelo, W. E. Clarke, R. Musunuri, K. Nagulapalli, S. Fairley, A. Runnels, L. Winterkorn, E. Lowy, E. E. Eichler, J. O. Korbel, C. Lee, T. Marschall, S. E. Devine, W. T. Harvey, W. Zhou, R. E. Mills, T. Rausch, S. Kumar, C. Alkan, F. Hormozdiari, Z. Chong, Y. Chen, X. Yang, J. Lin, M. B. Gerstein, Y. Kai, Q. Zhu, F. Yilmaz, C. Xiao, P. Flicek, S. Germer, H. Brand, I. M. Hall, M. E. Talkowski, G. Narzisi, and M. C. Zody. High-coverage whole-genome sequencing of the expanded 1000 genomes project cohort including 602 trios. *Cell*, 185(18):3426–3440.e19, Sep 2022. ISSN 0092-8674. doi: 10.1016/j.cell.2022.08.004. URL https://doi.org/10.1016/j.cell.2022.08.004.

D. Bzdok, N. Altman, and M. Krzywinski. Statistics versus machine learning. *Nature Methods*, 15(4):233–234, Apr 2018. ISSN 1548-7105. doi: 10.1038/nmeth.4642. URL https://doi.org/10.1038/nmeth.4642.

I. V. Caldas, A. G. Clark, and P. W. Messer. Inference of selective sweep parameters through supervised learning. *bioRxiv*, 2022. doi: 10.1101/2022.07.19.500702. URL https://www.biorxiv.org/content/early/2022/07/20/2022.07.19.500702.

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.

W. Cao, Z. Yan, Z. He, and Z. He. A comprehensive survey on geometric deep learning. *IEEE Access*, 8:35929–35949, 2020. doi: 10.1109/ACCESS.2020.2975067.

T. Capblancq, K. Luu, M. G. B. Blum, and E. Bazin. Evaluation of redundancy analysis to identify signatures of local adaptation. *MOLECULAR ECOLOGY RESOURCES*, 18(6):1223–1233, NOV 2018. ISSN 1755-098X. doi: 10.1111/1755-0998.12906.

A. G. Casanova, V. M. Pina, and A. Siri-Jégousse. The Symmetric Coalescent and Wright-Fisher models with bottlenecks. *arXiv:1903.05642 [math]*, Sept. 2020. URL http://arxiv.org/abs/1903.05642. arXiv: 1903.05642.

J. Chan, V. Perrone, J. P. Spence, P. A. Jenkins, S. Mathieson, and Y. S. Song. A likelihood-free inference framework for population genetic data using exchangeable neural networks. *Advances in neural information processing systems*, 31:8594–8605, Dec 2018. ISSN 1049-5258. URL https://pubmed.ncbi.nlm.nih.gov/33244210. 33244210[pmid].

C. H. Chandler, S. Chari, and I. Dworkin. Does your gene need a background check? how genetic background impacts the analysis of mutations, genes, and evolution. *TRENDS IN GENETICS*, 29 (6):358–366, JUN 2013. ISSN 0168-9525. doi: 10.1016/j.tig.2013.01.009.

B. Charlesworth. How long does it take to fix a favorable mutation, and why should we care? *The American Naturalist*, 195(5):753–771, 2020.

B. Charlesworth. Fisher's historic 1922 paper *On the dominance ratio*. *Genetics*, 220(3):iyac006, Mar. 2022. ISSN 1943-2631. doi: 10.1093/genetics/iyac006. URL https://academic.oup.com/genetics/article/doi/10.1093/genetics/iyac006/6541947.

B. Charlesworth and D. Charlesworth. Elements of evolutionary genetics. 2010.

B. Charlesworth and J. D. Jensen. How can we resolve lewontin's paradox? *Genome biology and evolution*, 14(7):evac096, 2022.

B. Charlesworth et al. *Evolution in age-structured populations*, volume 2. Cambridge University Press Cambridge, 1994.

D. Charlesworth. Balancing selection and its effects on sequences in nearby genome regions. *PLOS Genetics*, 2(4):1–6, 04 2006. doi: 10.1371/journal.pgen.0020064. URL https://doi.org/10.1371/journal.pgen.0020064.

T. Che, Y. Li, A. P. Jacob, Y. Bengio, and W. Li. Mode regularized generative adversarial networks, 2016. URL https://arxiv.org/abs/1612.02136.

H. Chen, S. M. Lundberg, and S.-I. Lee. Explaining a series of models by propagating shapley values. *NATURE COMMUNICATIONS*, 13(1), AUG 3 2022. doi: 10.1038/s41467-022-31384-3.

J. Chen, P. Ni, X. Li, J. Han, I. Jakovlic, C. Zhang, and S. Zhao. Population size may shape the accumulation of functional mutations following domestication. *BMC Evolutionary Biology*, 18, JAN 19 2018. ISSN 1471-2148. doi: {10.1186/s12862-018-1120-6}.

Z. Chen, Y. Bei, and C. Rudin. Concept whitening for interpretable image recognition. *NATURE MACHINE INTELLIGENCE*, 2(12), DEC 2020. doi: 10.1038/s42256-020-00265-z.

K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *CoRR*, abs/1409.1259, 2014. URL http://arxiv.org/abs/1409.1259.

D. Cohen. Optimizing reproduction in a randomly varying environment. *Journal of Theoretical Biology*, 12(1):119–129, Sept. 1966. ISSN 0022-5193. doi: 10.1016/0022-5193(66)90188-3. URL https://www.sciencedirect.com/science/article/pii/0022519366901883.

F. Cordero, A. G. Casanova, J. Schweinsberg, and M. Wilke-Berenguer. -coalescents arising in a population with dormancy. *Electronic Journal of Probability*, 27(none):1–34, Jan. 2022. ISSN 1083-6489, 1083-6489. doi: 10.1214/22-EJP739. URL https://projecteuclid.org/journals/electronic-journal-of-probability/volume-27/issue-none/%ce%9b-coalescents-arising-in-a-population-with-dormancy/10.1214/22-EJP739.full.

K. Cranmer, J. Brehmer, and G. Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020a. doi: 10.1073/pnas.1912789117. URL https://www.pnas.org/doi/abs/10.1073/pnas.1912789117.

K. Cranmer, J. Brehmer, and G. Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, Dec. 2020b. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1912789117. URL https://pnas.org/doi/full/10.1073/pnas.1912789117.

M. Crotti, E. Yohannes, I. J. Winfield, A. A. Lyle, C. E. Adams, and K. R. Elmer. Rapid adaptation through genomic and epigenomic responses following translocations in an endangered salmonid. *Evolutionary Applications*, 14(10):2470–2489, 2021. doi: https://doi.org/10.1111/eva.13267. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/eva.13267.

J. F. Crow. Mid-Century Controversies in Population Genetics. *Annual Review of Genetics*, 42(1):1–16, Dec. 2008. ISSN 0066-4197, 1545-2948. doi: 10.1146/annurev.genet.42.110807.091612. URL https://www.annualreviews.org/doi/10.1146/annurev.genet.42.110807.091612.

K. Csilléry, M. G. Blum, O. E. Gaggiotti, and O. François. Approximate bayesian computation (abc) in practice. *Trends in Ecology & Evolution*, 25(7):410–418, 2010. ISSN 0169-5347. doi: https://doi.org/10.1016/j.tree.2010.04.001. URL https://www.sciencedirect.com/science/article/pii/S0169534710000662.

K. Csilléry, O. François, and M. G. B. Blum. abc: an r package for approximate bayesian computation (abc). *Methods in Ecology and Evolution*, 3(3):475–479, 2012. doi: https://doi.org/10.1111/j.2041-210X.2011.00179.x. URL https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/j.2041-210X.2011.00179.x.

Y. Cui, C. Yu, Y. Yan, D. Li, Y. Li, T. Jombart, L. A. Weinert, Z. Wang, Z. Guo, L. Xu, et al. Historical variations in mutation rate in an epidemic pathogen, yersinia pestis. *Proceedings of the National Academy of Sciences*, 110(2):577–582, 2013.

J. Cury, B. C. Haller, G. Achaz, and F. Jay. Simulation of bacterial populations with SLiM. *Peer Community Journal*, 2:e7, 2022. doi: 10.24072/pcjournal.72. URL https://peercommunityjournal.org/articles/10.24072/pcjournal.72/.

M. Dann, S. Bellot, S. Schepella, H. Schaefer, and A. Tellier. Mutation rates in seeds and seed-banking influence substitution rates across the angiosperm phylogeny. Technical report, bioRxiv, June 2017. URL https://www.biorxiv.org/content/10.1101/156398v1. Type: article.

W. Deelder, E. D. Benavente, J. Phelan, E. Manko, S. Campino, L. Palla, and T. G. Clark. Using deep learning to identify recent positive selection in malaria parasite sequence data. *Malaria Journal*, 20(1):270, Jun 2021. ISSN 1475-2875. doi: 10.1186/s12936-021-03788-x. URL https://doi.org/10.1186/s12936-021-03788-x.

M. Dehasque, M. C. Ávila Arcos, D. Díez-del Molino, M. Fumagalli, K. Guschanski, E. D. Lorenzen, A.-S. Malaspinas, T. Marques-Bonet, M. D. Martin, G. G. R. Murray, A. S. T. Papadopulos, N. O. Therkildsen, D. Wegmann, L. Dalén, and A. D. Foote. Inference of natural selection from ancient dna. *Evolution Letters*, 4(2):94–108, 2020. doi: https://doi.org/10.1002/evl3.165. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/evl3.165.

R. Der, C. L. Epstein, and J. B. Plotkin. Generalized population models and the nature of genetic drift. *Theoretical population biology*, 80(2):80–99, 2011.

A. Devi and K. Jain. The Impact of Dominance on Adaptation in Changing Environments. *Genetics*, 216(1):227–240, 09 2020. ISSN 1943-2631. doi: 10.1534/genetics.120.303519. URL https://doi.org/10.1534/genetics.120.303519.

J.-S. Dhersin, F. Freund, A. Siri-Jegousse, and L. Yuan. On the length of an external branch in the Beta-coalescent. *Stochastic Processes and their Applications*, 123(5):1691–1715, MAY 2013. ISSN 0304-4149. doi: {10.1016/j.spa.2012.12.010}.

S. E. Diamond, E. G. Prileson, and R. A. Martin. Adaptation to urban environments. *Current Opinion in Insect Science*, 51:100893, 2022. ISSN 2214-5745. doi: https://doi.org/10.1016/j.cois.2022.100893. URL https://www.sciencedirect.com/science/article/pii/S2214574522000281.

P. Donnelly and T. Kurtz. Particle representations for measure-valued population models. *Annals of Probability*, 27(1):166–205, JAN 1999. ISSN 0091-1798.

M. dos Reis, P. C. J. Donoghue, and Z. Yang. Bayesian molecular clock dating of species divergences in the genomics era. *Nature Reviews Genetics*, 17(2):71–80, Feb. 2016. ISSN 1471-0064. doi: 10.1038/nrg.2015.8. URL https://www.nature.com/articles/nrg.2015.8.

F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning, 2017. URL https://arxiv.org/abs/1702.08608.

R. Durrett and J. Schweinsberg. A coalescent model for the effect of advantageous mutations on the genealogy of a population. *Stochastic Processes and their Applications*, 115(10):1628–1657, OCT 2005. ISSN 0304-4149. doi: {10.1016/j.spa.2005.04.009}.

J. Dworkin and R. Losick. Developmental Commitment in a Bacterium. *Cell*, 121(3):401–409, May 2005. ISSN 00928674. doi: 10.1016/j.cell.2005.02.032. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867405001996.

B. Eldon. Evolutionary genomics of high fecundity. *Annual Review of Genetics*, 54:213–236, 2020a.

B. Eldon. Evolutionary Genomics of High Fecundity. *Annual Review of Genetics*, 54(1):213–236, Nov. 2020b. ISSN 0066-4197, 1545-2948. doi: 10.1146/annurev-genet-021920-095932. URL https://www.annualreviews.org/doi/10.1146/annurev-genet-021920-095932.

B. Eldon and W. Stephan. Evolution of highly fecund haploid populations. *Theoretical population biology*, 119:48–56, 2018.

B. Eldon and W. Stephan. Sweepstakes reproduction facilitates rapid adaptation in highly fecund populations. *Molecular Ecology*, This issue, 2023.

B. Eldon and J. Wakeley. Coalescent processes when the distribution of offspring number among individuals is highly skewed. *Genetics*, 172(4):2621–2633, APR 2006. ISSN 0016-6731. doi: {10.1534/genetics.105.052175}.

B. Eldon, M. Birkner, J. Blath, and F. Freund. Can the Site-Frequency Spectrum Distinguish Exponential Population Growth from Multiple-Merger Coalescents? *Genetics*, 199(3):841+, MAR 2015. ISSN 0016-6731. doi: {10.1534/genetics.114.173807}.

H. Ellegren and N. Galtier. Determinants of genetic diversity. *Nature Reviews Genetics*, 17(7): 422–433, JUL 2016. ISSN 1471-0056. doi: {10.1038/nrg.2016.58}.

J. L. Elman. Finding Structure in Time. *Cognitive Science*, 14(2):179–211, Mar. 1990. ISSN 03640213. doi: 10.1207/s15516709cog1402_1. URL http://doi.wiley.com/10.1207/s15516709cog1402_1.

M. Escalona, S. Rocha, and D. Posada. A comparison of tools for the simulation of genomic next-generation sequencing data. *Nature Reviews Genetics*, 17(8):459–469, Aug 2016. ISSN 1471-0064. doi: 10.1038/nrg.2016.57. URL https://doi.org/10.1038/nrg.2016.57.

M. et al. A genomic history of Aboriginal Australia. *Nature*, 538(7624):207+, OCT 13 2016. ISSN 0028-0836. doi: {10.1038/nature18299}.

M. E. K. Evans and J. J. Dennehy. Germ banking: bet-hedging and variable release from egg and seed dormancy. *The Quarterly Review of Biology*, 80(4):431–451, Dec. 2005. ISSN 0033-5770. doi: 10.1086/498282.

G. Ewing and J. Hermisson. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics*, 26(16): 2064–2065, Aug. 2010. ISSN 1460-2059, 1367-4803. doi: 10.1093/bioinformatics/btq322. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btq322.

L. Excoffier, N. Marchi, D. A. Marques, R. Matthey-Doret, A. Gouy, and V. C. Sousa. fastsimcoal2: demographic inference under complex evolutionary scenarios. *Bioinformatics*, 37(24):4882–4885, 06 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab468. URL https://doi.org/10.1093/bioinformatics/btab468.

A. Eyre-Walker and P. D. Keightley. The distribution of fitness effects of new mutations. *Nature Reviews Genetics*, 8(8):610–618, Aug. 2007. ISSN 1471-0056, 1471-0064. doi: 10.1038/nrg2146. URL http://www.nature.com/articles/nrg2146.

F. Fan, J. Xiong, and G. Wang. On interpretability of artificial neural networks. *CoRR*, abs/2001.02522, 2020. URL http://arxiv.org/abs/2001.02522.

M. Fey and J. E. Lenssen. Fast graph representation learning with PyTorch geometric. URL http://arxiv.org/abs/1903.02428.

A. Fijarczyk and W. Babik. Detecting balancing selection in genomes: limits and prospects. *Molecular Ecology*, 24(14):3529–3545, 2015. doi: https://doi.org/10.1111/mec.13226. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.13226.

M. C. Fisher, A. Alastruey-Izquierdo, J. Berman, T. Bicanic, E. M. Bignell, P. Bowyer, M. Bromley, R. Brüggemann, G. Garber, O. A. Cornely, et al. Tackling the emerging threat of antifungal resistance to human health. *Nature Reviews Microbiology*, pages 1–15, 2022.

R. A. Fisher. XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh*, 52(2):399–433, 1919. ISSN 0080-4568, 2053-5945. doi: 10.1017/S0080456800012163. URL https://www.cambridge.org/core/product/identifier/S0080456800012163/type/journal_article.

R. A. Fisher. Xxi.—on the dominance ratio. *Proceedings of the Royal Society of Edinburgh*, 42: 321–341, 1922. doi: 10.1017/S0370164600023993.

R. A. Fisher. *The genetical theory of natural selection*. Clarendon Press, Oxford, 1930. doi: 10.5962/bhl.title.27468. URL https://www.biodiversitylibrary.org/bibliography/27468.

L. Flagel, Y. Brandvain, and D. R. Schrider. The Unreasonable Effectiveness of Convolutional Neural Networks in Population Genetic Inference. *Molecular Biology and Evolution*, 36(2):220–238, 12 2018. ISSN 0737-4038. doi: 10.1093/molbev/msy224. URL https://doi.org/10.1093/molbev/msy224.

M. Foll and O. Gaggiotti. A Genome-Scan Method to Identify Selected Loci Appropriate for Both Dominant and Codominant Markers: A Bayesian Perspective. *Genetics*, 180(2):977–993, 10 2008. ISSN 1943-2631. doi: 10.1534/genetics.108.092221. URL https://doi.org/10.1534/genetics.108.092221.

E. M. Fonseca, G. R. Colli, F. P. Werneck, and B. C. Carstens. Phylogeographic model selection using convolutional neural networks. *Molecular Ecology Resources*, 21(8):2661–2675, 2021. doi: https://doi.org/10.1111/1755-0998.13427. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.13427.

N. M. Fountain-Jones, M. L. Smith, and F. Austerlitz. Machine learning in molecular ecology. *Molecular Ecology Resources*, 21(8):2589–2597, 2021. doi: https://doi.org/10.1111/1755-0998.13532. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.13532.

F. Freund. Cannings models, population size changes and multiple-merger coalescents. *Journal of mathematical biology*, 80(5):1497–1521, 2020.

F. Freund, E. Kerdoncuff, S. Matuszewski, M. Lapierre, M. Hildebrandt, J. D. Jensen, L. Ferretti, A. Lambert, T. B. Sackton, and G. Achaz. Interpreting the pervasive observation of U-shaped Site Frequency Spectra. preprint, Evolutionary Biology, Apr. 2022. URL http://biorxiv.org/lookup/doi/10.1101/2022.04.12.488084.

F. Freund, E. Kerdoncuff, S. Matuszewski, M. Lapierre, M. Hildebrandt, J. D. Jensen, L. Ferretti, A. Lambert, T. B. Sackton, and G. Achaz. Interpreting the pervasive observation of U-shaped Site Frequency Spectra. *PLOS Genetics*, 19(3):e1010677, Mar. 2023. ISSN 1553-7404. doi: 10.1371/journal.pgen.1010677. URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1010677.

E. Frichot, S. D. Schoville, G. Bouchard, and O. François. Testing for Associations between Loci and Environmental Gradients Using Latent Factor Mixed Models. *Molecular Biology and Evolution*, 30(7):1687–1699, 03 2013. ISSN 0737-4038. doi: 10.1093/molbev/mst063. URL https://doi.org/10.1093/molbev/mst063.

L. Gattepaille, T. Guenther, and M. Jakobsson. Inferring Past Effective Population Size from Distributions of Coalescent Times. *Molecular Biology and Evolution*, 204(3):1191+, NOV 2016. ISSN 0016-6731. doi: {10.1534/genetics.115.185058}.

L. M. Gattepaille, M. Jakobsson, and M. G. B. Blum. Inferring population size changes with sequence and SNP data: lessons from human bottlenecks. *Heredity*, 110(5):409–419, MAY 2013. ISSN 0018-067X. doi: {10.1038/hdy.2012.120}.

A. Ghosh, V. Kulharia, V. Namboodiri, P. H. S. Torr, and P. K. Dokania. Multi-agent diverse generative adversarial networks, 2017. URL https://arxiv.org/abs/1704.02906.

B. Gibson and A. Eyre-Walker. Investigating evolutionary rate variation in bacteria. *Journal of molecular evolution*, 87(9-10):317–326, 2019.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Networks. *arXiv:1406.2661 [cs, stat]*, June 2014. URL http://arxiv.org/abs/1406.2661. arXiv: 1406.2661.

G. Gower, P. I. Picazo, M. Fumagalli, and F. Racimo. Detecting adaptive introgression in human evolution using convolutional neural networks. *eLife*, 10:e64669, may 2021. ISSN 2050-084X. doi: 10.7554/eLife.64669. URL https://doi.org/10.7554/eLife.64669.

J. Grealey, L. Lannelongue, W.-Y. Saw, J. Marten, G. Méric, S. Ruiz-Carmona, and M. Inouye. The Carbon Footprint of Bioinformatics. *Molecular Biology and Evolution*, 39(3), 02 2022. ISSN 1537-1719. doi: 10.1093/molbev/msac034. URL https://doi.org/10.1093/molbev/msac034. msac034.

J. G. Greener, S. M. Kandathil, L. Moffat, and D. T. Jones. A guide to machine learning for biologists. *Nature Reviews Molecular Cell Biology*, 23(1):40–55, Jan 2022. ISSN 1471-0080. doi: 10.1038/s41580-021-00407-0. URL https://doi.org/10.1038/s41580-021-00407-0.

R. Griffiths and P. Marjoram. Ancestral Inference from Samples of DNA Sequences with Recombination. *Journal of Computational Biology*, 3(4):479–502, Jan. 1996. ISSN 1066-5277, 1557-8666. doi: 10.1089/cmb.1996.3.479. URL http://www.liebertpub.com/doi/10.1089/cmb.1996.3.479.

M. Hahn. *Molecular Population Genetics*. Oxford University Press, Oxford, New York, Aug. 2018. ISBN 9780878939657.

N. G. Hairston Jr and B. T. De Stasio Jr. Rate of evolution slowed by a dormant propagule pool. *Nature*, 336(6196):239–242, Nov. 1988. ISSN 1476-4687. doi: 10.1038/336239a0. URL https://www.nature.com/articles/336239a0.

J. Haldane. The effect of variation of fitness. *The American Naturalist*, 71(735):337–349, 1937.

J. B. S. Haldane. A mathematical theory of natural and artificial selection—I. *Bulletin of Mathematical Biology*, 52(1):209–240, Jan. 1990. ISSN 1522-9602. doi: 10.1007/BF02459574. URL https://doi.org/10.1007/BF02459574.

B. V. Halldorsson, H. P. Eggertsson, K. H. S. Moore, H. Hauswedell, O. Eiriksson, M. O. Ulfarsson, G. Palsson, M. T. Hardarson, A. Oddsson, B. O. Jensson, S. Kristmundsdottir, B. D. Sigurpalsdottir, O. A. Stefansson, D. Beyter, G. Holley, V. Tragante, A. Gylfason, P. I. Olason, F. Zink, M. Asgeirsdottir, S. T. Sverrisson, B. Sigurdsson, S. A. Gudjonsson, G. T. Sigurdsson, G. H. Halldorsson, G. Sveinbjornsson, K. Norland, U. Styrkarsdottir, D. N. Magnusdottir, S. Snorradottir, K. Kristinsson, E. Sobech, H. Jonsson, A. J. Geirsson, I. Olafsson, P. Jonsson, O. B. Pedersen, C. Erikstrup, S. Brunak, S. R. Ostrowski, S. Andersen, K. Banasik, K. Burgdorf, M. Didriksen, K. M. Dinh, D. Gudbjartsson, T. F. Hansen, H. Hjalgrim, G. Jemec, P. Jennum, P. I. Johansson, M. A. H. Larsen, S. Mikkelsen, K. R. Nielsen, M. Nyegaard, S. Sækmose, E. Sørensen, U. Thorsteinsdottir, M. T. Brun, H. Ullum, T. Werge, G. Thorleifsson, F. Jonsson, P. Melsted, I. Jonsdottir, T. Rafnar, H. Holm, H. Stefansson, J. Saemundsdottir, D. F. Gudbjartsson, O. T. Magnusson, G. Masson, A. Helgason, H. Jonsson, P. Sulem, K. Stefansson, and D. G. Consortium. The sequences of 150,119 genomes in the uk biobank. *Nature*, 607(7920):732–740, Jul 2022. ISSN 1476-4687. doi: 10.1038/s41586-022-04965-x. URL https://doi.org/10.1038/s41586-022-04965-x.

B. C. Haller. Eidos: A simple scripting language - ben haller, 2016. URL http://benhaller.com/slim/Eidos_Manual.pdf.

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, Mar. 2019. ISSN 1537-1719. doi: 10.1093/molbev/msy228.

B. C. Haller, J. Galloway, J. Kelleher, P. W. Messer, and P. L. Ralph. Tree-sequence recording in slim opens new horizons for forward-time simulation of whole genomes. *MOLECULAR ECOLOGY RESOURCES*, 19(2):552–566, MAR 2019a. ISSN 1755-098X. doi: 10.1111/1755-0998.12968.

B. C. Haller, J. Galloway, J. Kelleher, P. W. Messer, and P. L. Ralph. Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes. *Molecular Ecology Resources*, 19(2):552–566, Mar. 2019b. ISSN 1755-098X, 1755-0998. doi: 10.1111/1755-0998.12968. URL https://onlinelibrary.wiley.com/doi/10.1111/1755-0998.12968.

I. Hamid, K. L. Korunes, D. R. Schrider, and A. Goldberg. Localizing post-admixture adaptive variants with object detection on ancestry-painted chromosomes. *bioRxiv*, 2022. doi: 10.1101/2022.09.04.506532. URL https://www.biorxiv.org/content/early/2022/09/05/2022.09.04.506532.

M. P. Hare, L. Nunney, M. K. Schwartz, D. E. Ruzzante, M. Burford, R. S. Waples, K. Ruegg, and F. Palstra. Understanding and estimating effective population size for practical application in marine species management. *Conservation Biology*, 25(3):438–449, 2011.

R. B. Harris and J. D. Jensen. Considering genomic scans for selection as coalescent model choice. *GENOME BIOLOGY AND EVOLUTION*, 12(6):871–877, JUN 2020. ISSN 1759-6653. doi: 10.1093/gbe/evaa093.

Z. He, M. Beaumont, and F. Yu. Effects of the ordering of natural selection and population regulation mechanisms on wright-fisher models. *G3: Genes, Genomes, Genetics*, 7(7):2095–2106, 2017.

D. Hedgecock and A. I. Pudovkin. Sweepstakes reproductive success in highly fecund marine fish and shellfish: a review and Commentary. *Bulletin of Marine Science*, 87(4):971–1002, OCT 2011. ISSN 0007-4977. doi: {10.5343/bms.2010.1051}.

L. Heinrich, J. Müller, A. Tellier, and D. Živković. Effects of population- and seed bank size fluctuations on neutral evolution and efficacy of natural selection. *Theoretical Population Biology*, 123:45–69, Sept. 2018. ISSN 0040-5809. doi: 10.1016/j.tpb.2018.05.003. URL https://www.sciencedirect.com/science/article/pii/S0040580917301715.

L. Heinrich, J. Mueller, A. Tellier, and D. Zivkovic. Effects of population- and seed bank size fluctuations on neutral evolution and efficacy of natural selection. *Theoretical Population Biology*, 123:45–69, SEP 2018. ISSN 0040-5809. doi: {10.1016/j.tpb.2018.05.003}.

H. A. Hejase, Z. Mo, L. Campagna, and A. Siepel. A Deep-Learning Approach for Inference of Selective Sweeps from the Ancestral Recombination Graph. *Molecular Biology and Evolution*, 39 (1), 11 2021. ISSN 1537-1719. doi: 10.1093/molbev/msab332. URL https://doi.org/10.1093/molbev/msab332. msab332.

H. A. Hejase, Z. Mo, L. Campagna, and A. Siepel. A deep-learning approach for inference of selective sweeps from the ancestral recombination graph. *MOLECULAR BIOLOGY AND EVOLUTION*, 39 (1), JAN 7 2022. ISSN 0737-4038. doi: 10.1093/molbev/msab332.

R. D. Hernandez and L. H. Uricchio. SFS_code: More Efficient and Flexible Forward Simulations. preprint, Bioinformatics, Aug. 2015. URL http://biorxiv.org/lookup/doi/10.1101/025064.

W. G. Hill and A. Robertson. The effect of linkage on limits to artificial selection. *Genetics Research*, 8(3):269–294, 1966. doi: 10.1017/S0016672300010156.

W. G. Hill and A. Robertson. Linkage disequilibrium in finite populations. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik*, 38(6):226–231, June 1968. ISSN 0040-5752. doi: 10.1007/BF01245622.

S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, Nov. 1997. ISSN 0899-7667, 1530-888X. doi: 10.1162/neco.1997.9.8.1735. URL https://direct.mit.edu/neco/article/9/8/1735-1780/6109.

K. E. Holsinger and B. S. Weir. Genetics in geographically structured populations: defining, estimating and interpreting F(ST). *Nature Reviews. Genetics*, 10(9):639–650, Sept. 2009. ISSN 1471-0064. doi: 10.1038/nrg2611.

K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *NEURAL NETWORKS*, 2(5):359–366, 1989. ISSN 0893-6080. doi: 10.1016/0893-6080(89)90020-8.

J. Hu, A. M. Askary, T. J. Thurman, D. A. Spiller, T. M. Palmer, R. M. Pringle, and R. D. Barrett. The epigenetic signature of colonizing new environments in anolis lizards. *Molecular biology and evolution*, 36(10):2165–2170, 2019.

M. Hubisz and A. Siepel. Inference of ancestral recombination graphs using argweaver. In J. Dutheil, editor, *STATISTICAL POPULATION GENOMICS*, volume 2090 of *Methods in Molecular Biology*, pages 231–266. 2020. ISBN 978-1-0716-0199-0; 978-1-0716-0198-3. doi: 10.1007/978-1-0716-0199-0_10.

R. Hudson. Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, 23(2):183–201, 1983. ISSN 0040-5809. doi: {10.1016/0040-5809(83)90013-8}.

R. R. Hudson. Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, 23(2):183–201, Apr. 1983. ISSN 00405809. doi: 10.1016/0040-5809(83)90013-8. URL https://linkinghub.elsevier.com/retrieve/pii/0040580983900138.

R. R. Hudson. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338, Feb. 2002. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/18.2.337. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/18.2.337.

J. Huxley. *Evolution: The Modern Synthesis*. Allen & Unwin, 1942. URL https://books.google.de/books?id=wVAQzQEACAAJ.

E. Hüllermeier and W. Waegeman. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110(3):457–506, Mar. 2021. ISSN 0885-6125, 1573-0565. doi: 10.1007/s10994-021-05946-3. URL http://link.springer.com/10.1007/s10994-021-05946-3.

K. K. Irwin, S. Laurent, S. Matuszewski, S. Vuilleumier, L. Ormond, H. Shim, C. Bank, and J. D. Jensen. On the importance of skewed offspring distributions and background selection in virus population genetics. *Heredity*, 117(6):393–399, 2016.

U. Isildak, A. Stella, and M. Fumagalli. Distinguishing between recent balancing selection and incomplete sweep using deep neural networks. *Molecular Ecology Resources*, 21(8):2706–2718, 2021a. doi: https://doi.org/10.1111/1755-0998.13379. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.13379.

U. Isildak, A. Stella, and M. Fumagalli. Distinguishing between recent balancing selection and incomplete sweep using deep neural networks. *Molecular Ecology Resources*, 21(8):2706–2718, Nov. 2021b. ISSN 1755-098X, 1755-0998. doi: 10.1111/1755-0998.13379. URL https://onlinelibrary.wiley.com/doi/10.1111/1755-0998.13379.

K. Jain and W. Stephan. Modes of rapid polygenic adaptation. *Molecular biology and evolution*, 34 (12):3169–3175, 2017.

M. M. Johnson and C. O. Wilke. Recombination rate inference via deep learning is limited by sequence diversity. *bioRxiv*, 2022. doi: 10.1101/2022.07.01.498489. URL https://www.biorxiv.org/content/early/2022/07/02/2022.07.01.498489.

P. Johri, B. Charlesworth, and J. D. Jensen. Toward an evolutionarily appropriate null model: Jointly inferring demography and purifying selection. *GENETICS*, 215(1):173–192, MAY 2020. ISSN 0016-6731. doi: 10.1534/genetics.119.303002.

P. Johri, K. Riall, H. Becher, L. Excoffier, B. Charlesworth, and J. D. Jensen. The impact of purifying and background selection on the inference of population history: Problems and prospects. *MOLECULAR BIOLOGY AND EVOLUTION*, 38(7):2986–3003, JUL 2021. ISSN 0737-4038. doi: 10.1093/molbev/msab050.

P. Johri, C. F. Aquadro, M. Beaumont, B. Charlesworth, L. Excoffier, A. Eyre-Walker, P. D. Keightley, M. Lynch, G. McVean, B. A. Payseur, S. P. Pfeifer, W. Stephan, and J. D. Jensen. Recommendations for improving statistical inference in population genomics. *PLOS Biology*, 20(5):e3001669, May 2022a. ISSN 1545-7885. doi: 10.1371/journal.pbio.3001669. URL https://dx.plos.org/10.1371/journal.pbio.3001669.

P. Johri, C. F. Aquadro, M. Beaumont, B. Charlesworth, L. Excoffier, A. Eyre-Walker, P. D. Keightley, M. Lynch, G. McVean, B. A. Payseur, S. P. Pfeifer, W. Stephan, and J. D. Jensen. Recommendations for improving statistical inference in population genomics. *PLOS Biology*, 20(5):e3001669, May 2022b. ISSN 1545-7885. doi: 10.1371/journal.pbio.3001669. URL https://dx.plos.org/10.1371/journal.pbio.3001669.

P. Johri, A. Eyre-Walker, R. N. Gutenkunst, K. E. Lohmueller, and J. D. Jensen. On the prospect of achieving accurate joint estimation of selection with population history. *Genome Biology and Evolution*, 14(7), 06 2022c. ISSN 1759-6653. doi: 10.1093/gbe/evac088. URL https://doi.org/10.1093/gbe/evac088. evac088.

T. Jombart, S. Devillard, and F. Balloux. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics*, 11(1):94, Oct 2010. ISSN 1471-2156. doi: 10.1186/1471-2156-11-94. URL https://doi.org/10.1186/1471-2156-11-94.

J. Jouganous, W. Long, A. P. Ragsdale, and S. Gravel. Inferring the Joint Demographic History of Multiple Populations: Beyond the Diffusion Approximation. *Genetics*, 206(3):1549–1567, 07 2017. ISSN 1943-2631. doi: 10.1534/genetics.117.200493. URL https://doi.org/10.1534/genetics.117.200493.

I. Kaj, S. M. Krone, and M. Lascoux. Coalescent theory for seed bank models. *Journal of Applied Probability*, 38:285–300, 2001.

I. Kaj, S. Krone, and M. Lascoux. Coalescent theory for seed bank models. *Journal of Applied Probability*, 38(2):285–300, JUN 2001. ISSN 0021-9002. doi: {10.1017/S0021900200019860}.

M. Kato, D. A. Vasco, R. Sugino, D. Narushima, and A. Krasnitz. Sweepstake evolution revealed by population-genetic analysis of copy-number alterations in single genomes of breast cancer. *Royal Society of Open Science*, 4(9), SEP 2017. ISSN 2054-5703. doi: {10.1098/rsos.171060}.

S. Kaushik and K. Jain. Time to fixation in changing environments. *Genetics*, 219(3):iyab148, 2021.

J. Kelleher, A. M. Etheridge, and G. McVean. Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLOS Computational Biology*, 12(5):e1004842, May 2016. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004842. URL https://dx.plos.org/10.1371/journal.pcbi.1004842.

J. Kelleher, K. R. Thornton, J. Ashander, and P. L. Ralph. Efficient pedigree recording for fast population genetics simulation. *PLOS Computational Biology*, 14(11):e1006581, Nov. 2018. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1006581. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006581.

J. Kelleher, Y. Wong, A. W. Wohns, C. Fadil, P. K. Albers, and G. McVean. Inferring whole-genome histories in large population datasets (vol 51, pg 1330, 2019). *Nature Genetics*, 51(11):1660, NOV 2019. ISSN 1061-4036. doi: {10.1038/s41588-019-0523-7}.

A. D. Kern and D. R. Schrider. Discoal: flexible coalescent simulations with selection. *Bioinformatics*, 32(24):3839–3841, Dec. 2016. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btw556. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btw556.

A. D. Kern and D. R. Schrider. diploS/HIC: An Updated Approach to Classifying Selective Sweeps. *G3 Genes|Genomes|Genetics*, 8(6):1959–1970, 06 2018. ISSN 2160-1836. doi: 10.1534/g3.118.200262. URL https://doi.org/10.1534/g3.118.200262.

G. Kersting, J. Schweinsberg, and A. Wakolbinger. The evolving beta coalescent. *Electronic Journal of Probability*, 19:1–27, 2014.

F. M. Key, J. C. Teixeira, C. de Filippo, and A. M. Andrés. Advantageous diversity maintained by balancing selection in humans. *Current Opinion in Genetics & Development*, 29: 45–51, 2014. ISSN 0959-437X. doi: https://doi.org/10.1016/j.gde.2014.08.001. URL https://www.sciencedirect.com/science/article/pii/S0959437X14000823. Genetics of human evolution.

E. Khomutov, K. Arzymatov, and V. Shchur. Deep learning based methods for estimating distribution of coalescence rates from genome-wide data. *Journal of Physics: Conference Series*, 1740(1): 012031, Jan. 2021. ISSN 1742-6588, 1742-6596. doi: 10.1088/1742-6596/1740/1/012031. URL https://iopscience.iop.org/article/10.1088/1742-6596/1740/1/012031.

S. Y. Kim, K. E. Lohmueller, A. Albrechtsen, Y. Li, T. Korneliussen, G. Tian, N. Grarup, T. Jiang, G. Andersen, D. Witte, T. Jorgensen, T. Hansen, O. Pedersen, J. Wang, and R. Nielsen. Estimation of allele frequency and association mapping using next-generation sequencing data. *BMC Bioinformatics*, 12(1):231, Jun 2011. ISSN 1471-2105. doi: 10.1186/1471-2105-12-231. URL https://doi.org/10.1186/1471-2105-12-231.

Y. Kim and W. Stephan. Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics*, 160(2):765–777, Feb. 2002. ISSN 0016-6731. doi: 10.1093/genetics/160.2.765.

Y. Kim, F. Koehler, A. Moitra, E. Mossel, and G. Ramnarayan. How Many Subpopulations Is Too Many? Exponential Lower Bounds for Inferring Population Histories. *Journal of Computational Biology*, 27(4):613–625, APR 1 2020. ISSN 1066-5277. doi: {10.1089/cmb.2019.0318}.

M. Kimura. On the Probability of Fixation of Mutant Genes in a Population. *Genetics*, 47(6):713–719, June 1962. ISSN 0016-6731. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1210364/.

D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes, May 2014. URL http://arxiv.org/abs/1312.6114. arXiv:1312.6114 [cs, stat].

J. Kingman. The Coalescent . *Stochastic Processes and their Applications*, 13, 1982.

T. N. Kipf and M. Welling. Semi-Supervised Classification with Graph Convolutional Networks. 2016. doi: 10.48550/ARXIV.1609.02907. URL https://arxiv.org/abs/1609.02907.

M. J. Kittlein, M. S. Mora, F. J. Mapelli, A. Austrich, and O. E. Gaggiotti. Deep learning and satellite imagery predict genetic diversity and differentiation. *METHODS IN ECOLOGY AND EVOLUTION*, 13(3):711–721, MAR 2022. ISSN 2041-210X. doi: 10.1111/2041-210X.13775.

E. V. Koonin. Towards a postmodern synthesis of evolutionary biology. *Cell cycle (Georgetown, Tex.)*, 8(6):799–800, Mar. 2009. ISSN 1538-4101. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3410441/.

B. Koopmann, J. Müller, A. Tellier, and D. Živković. Fisher–Wright model with deterministic seed bank and selection. *Theoretical Population Biology*, 114:29–39, Apr. 2017. ISSN 0040-5809. doi: 10.1016/j.tpb.2016.11.005. URL https://www.sciencedirect.com/science/article/pii/S0040580916301009.

K. Korfmann, D. Abu Awad, and A. Tellier. Weak seed banks influence the signature and detectability of selective sweeps. preprint, Evolutionary Biology, Apr. 2022a. URL http://biorxiv.org/lookup/doi/10.1101/2022.04.26.489499.

K. Korfmann, T. Sellinger, F. Freund, M. Fumagalli, and A. Tellier. Simultaneous inference of past demography and selection from the ancestral recombination graph under the beta coalescent. *bioRxiv*, 2022b. doi: 10.1101/2022.09.28.508873. URL https://www.biorxiv.org/content/early/2022/09/30/2022.09.28.508873.

K. Korfmann, O. E. Gaggiotti, and M. Fumagalli. Deep Learning in Population Genetics. *Genome Biology and Evolution*, 15(2):evad008, Feb. 2023. ISSN 1759-6653. doi: 10.1093/gbe/evad008. URL https://academic.oup.com/gbe/article/doi/10.1093/gbe/evad008/6997869.

A. Koropoulis, N. Alachiotis, and P. Pavlidis. *Detecting Positive Selection in Populations Using Genetic Data*, pages 87–123. Springer US, New York, NY, 2020. ISBN 978-1-0716-0199-0. doi: 10.1007/978-1-0716-0199-0_5. URL https://doi.org/10.1007/978-1-0716-0199-0_5.

J. Koskela. Multi-locus data distinguishes between population growth and multiple merger coalescents. *STATISTICAL APPLICATIONS IN GENETICS AND MOLECULAR BIOLOGY*, 17(3), JUN 2018. ISSN 2194-6302. doi: {10.1515/sagmb-2017-0011}.

J. Koskela and M. W. Berenguer. Robust model selection between population growth and multiple merger coalescents. *Mathematical Biosciences*, 311:1–12, MAY 2019. ISSN 0025-5564. doi: {10.1016/j.mbs.2019.03.004}.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

S. M. Krone and C. Neuhauser. Ancestral Processes with Selection. *Theoretical Population Biology*, 51(3):210–237, June 1997. ISSN 00405809. doi: 10.1006/tpbi.1997.1299. URL https://linkinghub.elsevier.com/retrieve/pii/S0040580997912995.

H. Kumar, M. Panigrahi, A. Panwar, D. Rajawat, S. S. Nayak, K. Saravanan, K. Kaisa, S. Parida, B. Bhushan, and T. Dutt. Machine-learning prospects for detecting selection signatures using

population genomics data. *Journal of Computational Biology*, 0(0):null, 2022. doi: 10.1089/cmb.2021.0447. URL https://doi.org/10.1089/cmb.2021.0447. PMID: 35639362.

B. B. Lamont, J. G. Pausas, T. He, E. T. F. Witkowski, and M. E. Hanley. Fire as a Selective Agent for both Serotiny and Nonserotiny Over Space and Time. *Critical Reviews in Plant Sciences*, 39(2): 140–172, Mar. 2020. ISSN 0735-2689, 1549-7836. doi: 10.1080/07352689.2020.1768465. URL https://www.tandfonline.com/doi/full/10.1080/07352689.2020.1768465.

A. J. Laruson, M. C. Fitzpatrick, S. R. Keller, B. C. Haller, and K. E. Lotterhos. Seeing the forest for the trees: Assessing genetic offset predictions from gradient forest. *Evolutionary Applications*, 15 (3):403–416, 2022. doi: https://doi.org/10.1111/eva.13354. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/eva.13354.

Y. Lecun and Y. Bengio. *Convolutional networks for images, speech, and time-series*. MIT Press, 1995.

Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel. Handwritten digit recognition with a back-propagation network. In D. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 2. Morgan-Kaufmann, 1989. URL https://proceedings.neurips.cc/paper/1989/file/53c3bce66e43be4f209556518c2fcb54-Paper.pdf.

Y. LeCun, F. J. Huang, and L. Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages II–104 Vol.2, 2004. doi: 10.1109/CVPR.2004.1315150.

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015. ISSN 1476-4687. doi: 10.1038/nature14539. URL https://doi.org/10.1038/nature14539.

J. B. Lee, R. Rossi, and X. Kong. Graph Classification using Structural Attention. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1666–1674, London United Kingdom, July 2018. ACM. ISBN 9781450355520. doi: 10.1145/3219819.3219980. URL https://dl.acm.org/doi/10.1145/3219819.3219980.

J. T. Lennon, F. den Hollander, M. Wilke-Berenguer, and J. Blath. Principles of seed banks and the emergence of complexity from dormancy. *Nature Communications*, 12(1):1–16, 2021a.

J. T. Lennon, F. den Hollander, M. Wilke-Berenguer, and J. Blath. Principles of seed banks and the emergence of complexity from dormancy. *Nature Communications*, 12(1):4807, Aug. 2021b. ISSN 2041-1723. doi: 10.1038/s41467-021-24733-1.

D. A. Levin. The Seed Bank as a Source of Genetic Novelty in Plants. *The American Naturalist*, 135 (4):563–572, 1990. ISSN 0003-0147. URL https://www.jstor.org/stable/2462053.

S. E. Levy and R. M. Myers. Advancements in next-generation sequencing. *Annual Review of Genomics and Human Genetics*, 17(1):95–115, 2016. doi: 10.1146/annurev-genom-083115-022413. URL https://doi.org/10.1146/annurev-genom-083115-022413. PMID: 27362342.

H. Li and R. Durbin. Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357):493–U84, JUL 28 2011. ISSN 0028-0836. doi: {10.1038/nature10231}.

K. Lin, H. Li, C. Schlötterer, and A. Futschik. Distinguishing positive selection from neutral evolution: Boosting the performance of summary statistics. *Genetics*, 187(1):229–244, 2011. ISSN 00166731. doi: 10.1534/genetics.110.122614.

P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis. Explainable ai: A review of machine learning interpretability methods. *ENTROPY*, 23(1), JAN 2021. doi: 10.3390/e23010018.

S. Linnainmaa. Taylor expansion of the accumulated rounding error. *BIT*, 16(2):146–160, June 1976. ISSN 0006-3835, 1572-9125. doi: 10.1007/BF01931367. URL http://link.springer.com/10.1007/BF01931367.

M. A. Lones. How to avoid machine learning pitfalls: a guide for academic researchers, 2021. URL https://arxiv.org/abs/2108.02497.

J. Lopes and M. Beaumont. Abc: A useful bayesian tool for the analysis of population data. *Infection, Genetics and Evolution*, 10(6):825–832, 2010. ISSN 1567-1348. doi: https://doi.org/10.1016/j.meegid.2009.10.010. URL https://www.sciencedirect.com/science/article/pii/S1567134809002251. MEEGID IX.

J. B. Losos and R. E. Ricklefs. Adaptation and diversification on islands. *Nature*, 457(7231):830–836, 2009.

R. N. Lou, A. Jacobs, A. P. Wilder, and N. O. Therkildsen. A beginner's guide to low-coverage whole genome sequencing for population genomics. *Molecular Ecology*, 30(23):5966–5993, 2021. doi: https://doi.org/10.1111/mec.16077. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.16077.

A. Louvet. Extinction threshold and large population limit of a plant metapopulation model with recurrent extinction events and a seed bank component. *Theoretical Population Biology*, 145:22–37, June 2022. ISSN 00405809. doi: 10.1016/j.tpb.2022.02.003. URL https://linkinghub.elsevier.com/retrieve/pii/S0040580922000144.

S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In I. Guyon, U. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 30 (NIPS 2017)*, volume 30 of *Advances in Neural Information Processing Systems*, 2017. 31st Annual Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, DEC 04-09, 2017.

K. Luu, E. Bazin, and M. G. B. Blum. pcadapt: an r package to perform genome scans for selection based on principal component analysis. *MOLECULAR ECOLOGY RESOURCES*, 17(1):67–77, JAN 2017. ISSN 1755-098X. doi: 10.1111/1755-0998.12592.

X. A. López-Cortés, F. Matamala, C. Maldonado, F. Mora-Poblete, and C. A. Scapim. A Deep Learning Approach to Population Structure Inference in Inbred Lines of Maize. *Frontiers in*

Genetics, 11, 2020. ISSN 1664-8021. URL https://www.frontiersin.org/articles/10.3389/fgene.2020.543459.

A. Mahmoudi, J. Koskela, J. Kelleher, Y.-b. Chan, and D. Balding. Bayesian inference of ancestral recombination graphs. *PLOS COMPUTATIONAL BIOLOGY*, 18(3), MAR 2022a. ISSN 1553-734X. doi: 10.1371/journal.pcbi.1009960.

A. Mahmoudi, J. Koskela, J. Kelleher, Y.-b. Chan, and D. Balding. Bayesian inference of ancestral recombination graphs. 18(3):e1009960, 2022b. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1009960. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1009960.

F. Manna, R. Pradel, R. Choquet, H. Fréville, and P.-O. Cheptou. Disentangling the role of seed bank and dispersal in plant metapopulation dynamics using patch occupancy surveys. *Ecology*, 98(10): 2662–2672, Oct. 2017. ISSN 0012-9658. doi: 10.1002/ecy.1960.

A. D. Mantes, D. M. Montserrat, C. D. Bustamante, X. Giró-i Nieto, and A. G. Ioannidis. Neural ADMIXTURE: rapid population clustering with autoencoders. preprint, Genomics, June 2021. URL http://biorxiv.org/lookup/doi/10.1101/2021.06.27.450081.

P. Marjoram and J. Wall. Fast “coalescent” simulation. *BMC Genetics*, 7, MAR 15 2006. ISSN 1471-2156. doi: {10.1186/1471-2156-7-16}.

P. Marjoram and J. D. Wall. Fast "coalescent" simulation. *BMC Genetics*, 7(1):16, Mar. 2006. ISSN 1471-2156. doi: 10.1186/1471-2156-7-16. URL https://doi.org/10.1186/1471-2156-7-16.

S. Matuszewski, M. E. Hildebrandt, G. Achaz, and J. D. Jensen. Coalescent processes with skewed offspring distributions and non-equilibrium demography. *Genetics*, 2017. ISSN 0016-6731. URL http://www.genetics.org/content/early/2017/11/10/genetics.117.300499.

J. Maynard Smith and J. Haigh. The hitch-hiking effect of a favourable gene. *Genetics Research*, 23 (1):23–35, 1974a.

J. Maynard Smith and J. Haigh. The hitch-hiking effect of a favourable gene. *Genetics Research*, 23 (1):23–35, Feb. 1974b. ISSN 1469-5073, 0016-6723. doi: 10.1017/S0016672300014634.

B. A. McDonald and C. Linde. Pathogen population genetics, evolutionary potential, and durable resistance. *Annual review of phytopathology*, 40(1):349–379, 2002.

J. H. McDonald and M. Kreitman. Adaptive protein evolution at the adh locus in drosophila. *Nature*, 351(6328):652–654, 1991.

G. McVean and N. Cardin. Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B-Biological Sciences*, 360(1459):1387–1393, JUL 29 2005. ISSN 0962-8436. doi: {10.1098/rstb.20053.1673}.

G. A. McVean and N. J. Cardin. Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1459):1387–1393, July 2005. ISSN 0962-8436, 1471-2970. doi: 10.1098/rstb.2005.1673. URL https://royalsocietypublishing.org/doi/10.1098/rstb.2005.1673.

J. Meisner and A. Albrechtsen. Haplotype and population structure inference using neural networks in whole-genome sequencing data. page genome;gr.276813.122v2. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.276813.122. URL http://genome.cshlp.org/lookup/doi/10.1101/gr.276813.122.

F. Menardo, S. Gagneux, and F. Freund. Multiple merger genealogies in outbreaks of Mycobacterium tuberculosis. *Molecular Biology and Evolution*, 07 2020. ISSN 0737-4038. doi: 10.1093/molbev/msaa179. URL https://doi.org/10.1093/molbev/msaa179. msaa179.

P. W. Messer. SLiM: Simulating Evolution with Selection and Linkage. *Genetics*, 194(4):1037–1039, Aug. 2013. ISSN 1943-2631. doi: 10.1534/genetics.113.152181. URL https://academic.oup.com/genetics/article/194/4/1037/5935309.

A. Miles, p. i. bot, M. R, P. Ralph, N. Harding, R. Pisupati, S. Rae, and T. Millar. cggh/scikit-allel: v1.3.3. URL https://zenodo.org/record/4759368.

M. L. Minsky. *Computation: Finite and Infinite Machines*. Prentice-Hall Series in Automatic Computation. Prentice-Hall, 1967.

V. Miró Pina, Joly, and A. Siri-Jégousse. Estimating the Lambda measure in multiple-merger coalescents. preprint, Evolutionary Biology, Mar. 2023. URL http://biorxiv.org/lookup/doi/10.1101/2023.03.10.532088.

E. Mohamed, K. Sirlantzis, and G. Howells. A review of visualisation-as-explanation techniques for convolutional neural networks and their evaluation. *DISPLAYS*, 73, JUL 2022. ISSN 0141-9382. doi: 10.1016/j.displa.2022.102239.

M. Mohle and S. Sagitov. A classification of coalescent processes for haploid exchangeable population models. *Annals of Probability*, 29(4):1547–1562, OCT 2001. ISSN 0091-1798.

M. Mondal, J. Bertranpetit, and O. Lao. Approximate bayesian computation with deep learning supports a third archaic introgression in asia and oceania. *Nature Communications*, 10(1):246, Jan 2019. ISSN 2041-1723. doi: 10.1038/s41467-018-08089-7. URL https://doi.org/10.1038/s41467-018-08089-7.

A. Y. Morales-Arce, R. B. Harris, A. C. Stone, and J. D. Jensen. Evaluating the contributions of purifying selection and progeny-skew in dictating within-host Mycobacterium tuberculosis evolution. *Evolution*, 74(5):992–1001, MAY 2020. ISSN 0014-3820. doi: {10.1111/evo.13954}.

M. R. Mughal and M. DeGiorgio. Localizing and classifying adaptive targets with trend filtered regression. *Molecular Biology and Evolution*, 36(2):252–270, 2019. ISSN 15371719. doi: 10.1093/molbev/msy205.

H. J. Muller. Artificial Transmutation of the Gene. *Science*, 66(1699):84–87, July 1927. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.66.1699.84. URL https://www.science.org/doi/10.1126/science.66.1699.84.

H. J. Muller. Some Genetic Aspects of Sex. *The American Naturalist*, 66(703):118–138, 1932. ISSN 0003-0147. URL https://www.jstor.org/stable/2456922.

H. J. Muller. Our load of mutations. *American journal of human genetics*, 2(2):111, 1950.

M. Möst, S. Oexle, S. Marková, D. Aidukaite, L. Baumgartner, H.-B. Stich, M. Wessels, D. Martin-Creuzburg, and P. Spaak. Population genetic dynamics of an invasion reconstructed from the sediment egg bank. *Molecular Ecology*, 24(16):4074–4093, Aug. 2015. ISSN 1365-294X. doi: 10.1111/mec.13298.

K. Nara. Spores of ectomycorrhizal fungi: ecological strategies for germination and dormancy. *New Phytologist*, 181(2):245–248, Jan. 2009. ISSN 0028-646X, 1469-8137. doi: 10.1111/j.1469-8137.2008.02691.x. URL https://onlinelibrary.wiley.com/doi/10.1111/j.1469-8137.2008.02691.x.

R. A. Neher and O. Hallatschek. Genealogies of rapidly adapting populations. *Proceedings of the National Academy of Sciences*, 110(2):437–442, Jan. 2013. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1213113110. URL https://pnas.org/doi/full/10.1073/pnas.1213113110.

M. Nei and W. H. Li. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences*, 76(10):5269–5273, Oct. 1979. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.76.10.5269. URL https://pnas.org/doi/full/10.1073/pnas.76.10.5269.

D. Nelson, J. Kelleher, A. P. Ragsdale, C. Moreau, G. McVean, and S. Gravel. Accounting for long-range correlations in genome-wide simulations of large cohorts. *PLOS Genetics*, 16(5), MAY 2020. ISSN 1553-7404. doi: {10.1371/journal.pgen.1008619;10.1371/journal.pgen.1008619.r001;10.1371/journal.pgen.1008619.r002;10.1371/journal.pgen.1008619.r003;10.1371/journal.pgen.1008619.r004;10.1371/journal.pgen.1008619.r005;10.1371/journal.pgen.1008619.r006}.

C. Neuhauser and S. M. Krone. The Genealogy of Samples in Models With Selection. *Genetics*, 145(2):519–534, Feb. 1997. ISSN 1943-2631. doi: 10.1093/genetics/145.2.519. URL https://academic.oup.com/genetics/article/145/2/519/6018087.

A. Nguembang Fadja, F. Riguzzi, G. Bertorelle, and E. Trucchi. Identification of natural selection in genomic data with deep convolutional neural network. *BioData Mining*, 14(1):51, Dec 2021. ISSN 1756-0381. doi: 10.1186/s13040-021-00280-9. URL https://doi.org/10.1186/s13040-021-00280-9.

R. Nielsen. Molecular signatures of natural selection. *Annual Review of Genetics*, 39(1):197–218, 2005. doi: 10.1146/annurev.genet.39.073003.112420. URL https://doi.org/10.1146/annurev.genet.39.073003.112420. PMID: 16285858.

R. Nielsen, J. S. Paul, A. Albrechtsen, and Y. S. Song. Genotype and snp calling from next-generation sequencing data. *Nature Reviews Genetics*, 12(6):443–451, Jun 2011. ISSN 1471-0064. doi: 10.1038/nrg2986. URL https://doi.org/10.1038/nrg2986.

H.-S. Niwa, K. Nashida, and T. Yanagimoto. Reproductive skew in japanese sardine inferred from dna sequences. *ICES Journal of Marine Science*, 73(9):2181–2189, 2016.

G. Novakovsky, O. Fornes, M. Saraswat, S. Mostafavi, and W. W. Wasserman. Explainn: interpretable and transparent neural networks for genomics. *bioRxiv*, 2022. doi: 10.1101/2022.05.20.492818. URL https://www.biorxiv.org/content/early/2022/05/25/2022.05.20.492818.

L. Nunney and A. E. K. Ritland. The Effective Size of Annual Plant Populations: The Interaction of a Seed Bank with Fluctuating Population Size in Maintaining Genetic Variation. *The American Naturalist*, 160(2):195–204, 2002. ISSN 0003-0147. doi: 10.1086/341017. URL https://www.jstor.org/stable/10.1086/341017.

J. Olden and D. Jackson. Illuminating the black box: a randomization approach for understanding variable contributions in artificial neural networks. *Ecological Modelling*, 154:135–150, 2002.

K. O'Shea and R. Nash. An introduction to convolutional neural networks, 2015. URL https://arxiv.org/abs/1511.08458.

A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in PyTorch. Oct. 2017. URL https://openreview.net/forum?id=BJJsrmfCZ.

P. Pavlidis, J. D. Jensen, and W. Stephan. Searching for footprints of positive selection in whole-genome SNP data from nonequilibrium populations. *Genetics*, 185(3):907–922, 2010. ISSN 00166731. doi: 10.1534/genetics.110.116459.

P. Pavlidis, D. Živković, A. Stamatakis, and N. Alachiotis. SweeD: Likelihood-Based Detection of Selective Sweeps in Thousands of Genomes. *Molecular Biology and Evolution*, 30(9):2224–2234, Sept. 2013. ISSN 0737-4038. doi: 10.1093/molbev/mst112. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3748355/.

M. F. Perez, I. A. S. Bonatelli, M. Romeiro-Brito, F. F. Franco, N. P. Taylor, D. C. Zappi, and E. M. Moraes. Coalescent-based species delimitation meets deep learning: Insights from a highly fragmented cactus system. *Molecular Ecology Resources*, 22(3):1016–1028, 2022. doi: https://doi.org/10.1111/1755-0998.13534. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.13534.

A. Persoons, K. J. Hayden, B. Fabre, P. Frey, S. De Mita, A. Tellier, and F. Halkett. The escalatory red queen: Population extinction and replacement following arms race dynamics in poplar rust. *Molecular Ecology*, 26(7):1902–1918, 2017. doi: https://doi.org/10.1111/mec.13980. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.13980.

M. Petr, B. C. Haller, P. L. Ralph, and F. Racimo. slendr: a framework for spatio-temporal population genomic simulations on geographic landscapes, Mar. 2022. URL https://www.biorxiv.org/content/10.1101/2022.03.20.485041v1.

T. G. Phan and P. I. Croucher. The dormant cancer cell life cycle. *Nature Reviews Cancer*, 20 (7):398–411, July 2020. ISSN 1474-1768. doi: 10.1038/s41568-020-0263-0. URL https://www.nature.com/articles/s41568-020-0263-0.

J. Pitman. Coalescents with multiple collisions. *Annals of Probability*, 27(4):1870–1902, OCT 1999. ISSN 0091-1798. doi: {10.1214/aop/1022677552}.

R. Poplin, P.-C. Chang, D. Alexander, S. Schwartz, T. Colthurst, A. Ku, D. Newburger, J. Dijamco, N. Nguyen, P. T. Afshar, S. S. Gross, L. Dorfman, C. Y. McLean, and M. A. DePristo. A universal snp and small-indel variant caller using deep neural networks. *Nature Biotechnology*, 36(10):983–987, Nov 2018. ISSN 1546-1696. doi: 10.1038/nbt.4235. URL https://doi.org/10.1038/nbt.4235.

D. Prangle. Summary statistics in approximate bayesian computation, 2015.

J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, 155(2):945–959, June 2000. ISSN 1943-2631. doi: 10.1093/genetics/155.2.945. URL https://academic.oup.com/genetics/article/155/2/945/6048111.

W. B. Provine. *The origins of theoretical population genetics*. University of Chicago Press, 2020.

M. Pybus, P. Luisi, G. M. Dall'Olio, M. Uzkudun, H. Laayouni, J. Bertranpetit, and J. Engelken. Hierarchical boosting: a machine-learning framework to detect and classify hard selective sweeps in human populations. *Bioinformatics*, 31(24):3946–3952, 08 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv493. URL https://doi.org/10.1093/bioinformatics/btv493.

X. Qin, C. W. K. Chiang, and O. E. Gaggiotti. Deciphering signatures of natural selection via deep learning. *BRIEFINGS IN BIOINFORMATICS*, 2022. ISSN 1467-5463. doi: 10.1093/bib/bbac354.

A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents, 2022. URL https://arxiv.org/abs/2204.06125.

M. D. Rasmussen, M. J. Hubisz, I. Gronau, and A. Siepel. Genome-wide inference of ancestral recombination graphs. *PLOS GENETICS*, 10(5), MAY 2014a. ISSN 1553-7404. doi: 10.1371/journal.pgen.1004342.

M. D. Rasmussen, M. J. Hubisz, I. Gronau, and A. Siepel. Genome-Wide Inference of Ancestral Recombination Graphs. *PLOS Genetics*, 10(5):e1004342, May 2014b. ISSN 1553-7404. doi: 10.1371/journal.pgen.1004342. URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1004342.

D. P. Rice, J. Novembre, and M. M. Desai. Distinguishing multiple-merger from kingman coalescence using two-site frequency spectra. *bioRxiv*, 2018. doi: 10.1101/461517. URL https://www.biorxiv.org/content/early/2018/11/03/461517.

J. D. Robinson, L. Bunnefeld, J. Hearn, G. N. Stone, and M. J. Hickerson. ABC inference of multi-population divergence with admixture from unphased population genomic data. *Molecular Ecology*, 23(18):4458–4471, Sept. 2014. ISSN 0962-1083. doi: 10.1111/mec.12881. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4285295/.

A. R. Rogers and C. Huff. Linkage disequilibrium between loci with unknown phase. 182(3):839–844. ISSN 0016-6731. doi: 10.1534/genetics.108.093153. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2710162/.

R. Ronen, N. Udpa, E. Halperin, and V. Bafna. Learning natural selection from the site frequency spectrum. *Genetics*, 195(1):181–193, 2013. ISSN 00166731. doi: 10.1534/genetics.113.152587.

D. E. Rumelhart and J. L. McClelland. *Learning Internal Representations by Error Propagation*, pages 318–362. 1987.

S. Sabin, A. Y. Morales-Arce, S. P. Pfeifer, and J. D. Jensen. The impact of frequently neglected model violations on bacterial recombination rate estimation: a case study in mycobacterium canettii and mycobacterium tuberculosis. *G3*, 12(5):jkac055, 2022.

A. M. Sackman, R. B. Harris, and J. D. Jensen. Inferring demography and selection in organisms characterized by skewed offspring distributions. *GENETICS*, 211(3):1019–1028, MAR 2019. ISSN 0016-6731. doi: 10.1534/genetics.118.301684.

S. Sagitov. The general coalescent with asynchronous mergers of ancestral lines. *Journal of Applied Probability*, 36(4):1116–1125, DEC 1999. ISSN 0021-9002. doi: {10.1239/jap/1032374759}.

S. Sagitov. Convergence to the coalescent with simultaneous multiple mergers. *Journal of Applied Probability*, 40(4):839–854, DEC 2003. ISSN 0021-9002. doi: {10.1239/jap/1067436085}.

T. Sanchez. *Reconstructing our past deep learning for population genetics*. Theses, Université Paris-Saclay, Mar. 2022. URL https://theses.hal.science/tel-03701132.

T. Sanchez, J. Cury, G. Charpiat, and F. Jay. Deep learning for population size history inference: Design, comparison and combination with approximate bayesian computation. *Molecular Ecology Resources*, 21(8):2645–2660, 2021a. doi: https://doi.org/10.1111/1755-0998.13224. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.13224.

T. Sanchez, J. Cury, G. Charpiat, and F. Jay. Deep learning for population size history inference: Design, comparison and combination with approximate bayesian computation. 21 (8):2645–2660, 2021b. ISSN 1755-098X, 1755-0998. doi: 10.1111/1755-0998.13224. URL https://onlinelibrary.wiley.com/doi/10.1111/1755-0998.13224.

T. Sanchez, E. M. Bray, P. Jobic, J. Guez, A.-C. Letournel, G. Charpiat, J. Cury, and F. Jay. dnadna: a deep learning framework for population genetics inference. *Bioinformatics*, 11 2022a. ISSN 1367-4803. doi: 10.1093/bioinformatics/btac765. URL https://doi.org/10.1093/bioinformatics/btac765. btac765.

T. Sanchez, B. Caramiaux, P. Thiel, and W. E. Mackay. Deep Learning Uncertainty in Machine Teaching. In *IUI 2022 - 27th Annual Conference on Intelligent User Interfaces*, Helsinki / Virtual, Finland, Mar. 2022b. doi: 10.1145/3490099.3511117. URL https://hal.archives-ouvertes.fr/hal-03579448.

N. Sapoval, A. Aghazadeh, M. G. Nute, D. A. Antunes, A. Balaji, R. Baraniuk, C. J. Barberan, R. Dannenfelser, C. Dun, M. Edrisi, R. A. L. Elworth, B. Kille, A. Kyrillidis, L. Nakhleh, C. R. Wolfe, Z. Yan, V. Yao, and T. J. Treangen. Current progress and open challenges for applying deep learning across the biosciences. *Nature Communications*, 13(1):1728, Dec. 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-29268-7. URL https://www.nature.com/articles/s41467-022-29268-7.

O. Sargsyan and J. Wakeley. A coalescent process with simultaneous multiple mergers for approximating the gene genealogies of many marine organisms. *Theoretical population biology*, 74(1): 104–114, 2008.

S. Sarkar and R. S. Cohen, editors. *The Founders of Evolutionary Genetics*, volume 142 of *Boston Studies in the Philosophy of Science*. Springer Netherlands, Dordrecht, 1992. ISBN 978-0-7923-3392-0 978-94-011-2856-8. doi: 10.1007/978-94-011-2856-8. URL http://link.springer.com/10.1007/978-94-011-2856-8.

S. Schiffels and R. Durbin. Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, 46(8):919–925, AUG 2014. ISSN 1061-4036. doi: {10.1038/ng.3015}.

M. Schlichtkrull, T. N. Kipf, P. Bloem, R. v. d. Berg, I. Titov, and M. Welling. Modeling relational data with graph convolutional networks, 2017. URL https://arxiv.org/abs/1703.06103.

J. Schmidhuber. Deep learning in neural networks: An overview. *CoRR*, abs/1404.7828, 2014. URL http://arxiv.org/abs/1404.7828.

D. R. Schrider and A. D. Kern. S/hic: Robust identification of soft and hard sweeps using machine learning. *PLOS Genetics*, 12(3):1–31, 03 2016. doi: 10.1371/journal.pgen.1005928. URL https://doi.org/10.1371/journal.pgen.1005928.

D. R. Schrider and A. D. Kern. Supervised machine learning for population genetics: A new paradigm. *Trends in Genetics*, 34(4):301–312, 2018. ISSN 0168-9525. doi: 10.1016/j.tig.2017.12.005. URL https://www.sciencedirect.com/science/article/pii/S0168952517302251.

J. Schweinsberg. Coalescent processes obtained from supercritical Galton-Watson processes. *Stochastic Processes and their Applications*, 106(1):107–139, JUL 2003. ISSN 0304-4149. doi: {10.1016/S0304-4149(03)00028-0}.

T. Sellinger, D. Abu Awad, M. Möst, and A. Tellier. Inference of past demography, dormancy and self-fertilization rates from whole genome sequence data. preprint, Evolutionary Biology, July 2019. URL http://biorxiv.org/lookup/doi/10.1101/701185.

T. P. P. Sellinger, D. Abu-Awad, and A. Tellier. Limits and convergence properties of the sequentially markovian coalescent. *MOLECULAR ECOLOGY RESOURCES*, 21(7):2231–2248, OCT 2021a. ISSN 1755-098X. doi: 10.1111/1755-0998.13416.

T. P. P. Sellinger, D. Abu-Awad, and A. Tellier. Limits and convergence properties of the sequentially Markovian coalescent. *Molecular Ecology Resources*, 21(7):2231–2248, Oct. 2021b. ISSN 1755-0998. doi: 10.1111/1755-0998.13416.

T. P. P. Sellinger, D. Abu Awad, M. Moest, and A. Tellier. Inference of past demography, dormancy and self-fertilization rates from whole genome sequence data. *PLOS Genetics*, 16(4), APR 2020. ISSN 1553-7404. doi: {10.1371/journal.pgen.1008698;10.1371/journal.pgen.1008698.r001; 10.1371/journal.pgen.1008698.r002;10.1371/journal.pgen.1008698.r003;10.1371/journal.pgen. 1008698.r004;10.1371/journal.pgen.1008698.r005;10.1371/journal.pgen.1008698.r006}.

D. Sellis, B. J. Callahan, D. A. Petrov, and P. W. Messer. Heterozygote advantage as a natural consequence of adaptation in diploids. *Proceedings of the National Academy of Sciences*, 108(51): 20666–20671, 2011. doi: 10.1073/pnas.1114573108. URL https://www.pnas.org/doi/abs/10.1073/pnas.1114573108.

L. S. Shapley. A value for n-person games. In H. W. Kuhn and A. W. Tucker, editors, *Contributions to the Theory of Games II*, pages 307–317. Princeton University Press, Princeton, 1953.

S. Sheehan and Y. S. Song. Deep learning for population genetic inference. *PLOS Computational Biology*, 12(3):1–28, 03 2016. doi: 10.1371/journal.pcbi.1004845. URL https://doi.org/10.1371/journal.pcbi.1004845.

S. Sheehan and Y. S. Song. Deep Learning for Population Genetic Inference. *PLOS Computational Biology*, 12(3), MAR 2016. ISSN 1553-734X. doi: {10.1371/journal.pcbi.1004845}.

W. R. Shoemaker and J. T. Lennon. Evolution with a seed bank: The population genetic consequences of microbial dormancy. *Evolutionary Applications*, 11(1):60–75, Jan. 2018. ISSN 1752-4571. doi: 10.1111/eva.12557.

W. R. Shoemaker, E. Polezhaeva, K. B. Givens, and J. T. Lennon. Seed banks alter the molecular evolutionary dynamics of Bacillus subtilis. *Genetics*, 221(2), 05 2022. ISSN 1943-2631. doi: 10.1093/genetics/iyac071. URL https://doi.org/10.1093/genetics/iyac071. iyac071.

M. Silva, D. Pratas, and A. J. Pinho. Efficient DNA sequence compression with neural networks. *GigaScience*, 9(11), 11 2020. ISSN 2047-217X. doi: 10.1093/gigascience/giaa119. URL https://doi.org/10.1093/gigascience/giaa119. giaa119.

K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2013. URL https://arxiv.org/abs/1312.6034.

C. C. R. Smith, S. Tittes, P. L. Ralph, and A. D. Kern. Dispersal inference from population genetic variation using a convolutional neural network. *bioRxiv*, 2022. doi: 10.1101/2022.08.25.505329. URL https://www.biorxiv.org/content/early/2022/08/26/2022.08.25.505329.

P. Smolensky. Information processing in dynamical systems: Foundations of harmony theory. In *Parallel distributed processing: Explorations in the microstructure of cognition*, pages 194–281–. MIT Press, Cambridge, MA, 1986.

V. Soni, M. Vos, and A. Eyre-Walker. A new test suggests hundreds of amino acid polymorphisms in humans are subject to balancing selection. *PLOS Biology*, 20(6):1–27, 06 2022. doi: 10.1371/journal.pbio.3001645. URL https://doi.org/10.1371/journal.pbio.3001645.

L. Speidel, M. Forest, S. Shi, and S. R. Myers. A method for genome-wide genealogy estimation for thousands of samples. *Nature Genetics*, 51(9):1321–1329, Sept. 2019. ISSN 1061-4036, 1546-1718. doi: 10.1038/s41588-019-0484-x. URL http://www.nature.com/articles/s41588-019-0484-x.

L. Speidel, M. Forest, S. Shi, and S. R. Myers. A method for genome-wide genealogy estimation for thousands of samples. *Nature Genetics*, 51(9):1321+, SEP 2019. ISSN 1061-4036. doi: {10.1038/s41588-019-0484-x}.

P. R. Staab, S. Zhu, D. Metzler, and G. Lunter. scrm: efficiently simulating long sequences using the approximated coalescent with recombination. *Bioinformatics*, 31(10):1680–1682, May 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu861. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4426833/.

R. Stam and B. A. McDonald. When resistance gene pyramids are not durable—the role of pathogen diversity. *Molecular Plant Pathology*, 19(3):521, 2018.

M. Steinruecken, M. Birkner, and J. Blath. Analysis of DNA sequence variation within marine species using Beta-coalescents. *Theoretical Population Biology*, 87:15–24, AUG 2013. ISSN 0040-5809. doi: {10.1016/j.tpb.2013.01.007}.

W. Stephan. Signatures of positive selection: from selective sweeps at individual loci to subtle allele frequency changes in polygenic adaptation. *Molecular Ecology*, 25(1, SI):79–88, JAN 2016. ISSN 0962-1083. doi: {10.1111/mec.13288}.

W. Stephan. Selective Sweeps. *Genetics*, 211(1):5–13, Jan. 2019a. ISSN 1943-2631. doi: 10.1534/genetics.118.301319. URL https://academic.oup.com/genetics/article/211/1/5/5931139.

W. Stephan. Selective sweeps. *Genetics*, 211(1):5–13, 2019b.

W. Stephan. Selective Sweeps. *Genetics*, 211(1):5–13, Jan. 2019c. ISSN 0016-6731. doi: 10.1534/genetics.118.301319. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6325696/.

N. Stoletzki and A. Eyre-Walker. Estimation of the neutrality index. *Molecular biology and evolution*, 28(1):63–70, 2011.

A. Stoltzfus. Why we don't want another "Synthesis". *Biology Direct*, 12(1):23, Oct. 2017. ISSN 1745-6150. doi: 10.1186/s13062-017-0194-1. URL https://doi.org/10.1186/s13062-017-0194-1.

S. Struett, T. Sellinger, S. Glémin, A. Tellier, and S. Laurent. Inference of evolutionary transitions to self-fertilization using whole-genome sequences. *bioRxiv*, 2022. doi: 10.1101/2022.07.29.502030. URL https://www.biorxiv.org/content/early/2022/08/01/2022.07.29.502030.

E. Strumbelj and I. Kononenko. An efficient explanation of individual classifications using game theory. *JOURNAL OF MACHINE LEARNING RESEARCH*, 11:1–18, JAN 2010. ISSN 1532-4435.

A. H. Sturtevant. The linear arrangement of six sex-linked factors in drosophila, as shown by their mode of association. *Journal of Experimental Zoology*, 14(1):43–59, 1913. doi: https://doi.org/10.1002/jez.1400140104. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jez.1400140104.

L. A. Sugden, E. G. Atkinson, A. P. Fischer, S. Rong, B. M. Henn, and S. Ramachandran. Localization of adaptive variants in human genomes using averaged one-dependence estimation. *Nature Communications*, 9(1), 2018. ISSN 20411723. doi: 10.1038/s41467-018-03100-7. URL http://dx.doi.org/10.1038/s41467-018-03100-7.

M. Sunnåker, A. G. Busetto, E. Numminen, J. Corander, M. Foll, and C. Dessimoz. Approximate Bayesian Computation. *PLOS Computational Biology*, 9(1):e1002803, Jan. 2013. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1002803. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002803.

A. Suvorov, J. Hochuli, and D. R. Schrider. Accurate inference of tree topologies from multiple sequence alignments using deep learning. *Systematic biology*, 69(2):221–233, Mar 2020. ISSN 1076-836X. doi: 10.1093/sysbio/syz060. URL https://pubmed.ncbi.nlm.nih.gov/31504938. 31504938[pmid].

F. Tajima. Evolutionary relationship of DNA sequences in finite populations. *Genetics*, 105(2): 437–460, Oct. 1983. ISSN 0016-6731. doi: 10.1093/genetics/105.2.437.

F. Tajima. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3):585–595, Nov. 1989. ISSN 0016-6731. doi: 10.1093/genetics/123.3.585.

Y. W. Teh and G. E. Hinton. Rate-coded restricted boltzmann machines for face recognition. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2000. URL https://proceedings.neurips.cc/paper/2000/file/c366c2c97d47b02b24c3ecade4c40a01-Paper.pdf.

A. Tejero-Cantero, J. Boelts, M. Deistler, J.-M. Lueckmann, C. Durkan, P. J. Gonçalves, D. S. Greenberg, and J. H. Macke. sbi: A toolkit for simulation-based inference. *Journal of Open Source Software*, 5(52):2505, 2020. doi: 10.21105/joss.02505. URL https://doi.org/10.21105/joss.02505.

A. Tellier. Persistent seed banking as eco-evolutionary determinant of plant nucleotide diversity: novel population genetics insights. *New Phytologist*, 221(2):725–730, JAN 2019. ISSN 0028-646X. doi: {10.1111/nph.15424}.

A. Tellier. Persistent seed banking as eco-evolutionary determinant of plant nucleotide diversity: novel population genetics insights. *New Phytologist*, 221(2):725–730, Jan. 2019. ISSN 0028-646X, 1469-8137. doi: 10.1111/nph.15424. URL https://onlinelibrary.wiley.com/doi/10.1111/nph.15424.

A. Tellier and J. K. M. Brown. The influence of perenniality and seed banks on polymorphism in plant-parasite interactions. *The American Naturalist*, 174(6):769–779, Dec. 2009. ISSN 1537-5323. doi: 10.1086/646603.

A. Tellier and C. Lemaire. Coalescence 2.0: a multiple branching of recent theoretical developments and their applications. *Molecular Ecology*, 23(11):2637–2652, JUN 2014. ISSN 0962-1083. doi: {10.1111/mec.12755}.

A. Tellier and C. Lemaire. Coalescence 2.0: a multiple branching of recent theoretical developments and their applications. *Molecular Ecology*, 23(11):2637–2652, June 2014. ISSN 09621083. doi: 10.1111/mec.12755. URL https://onlinelibrary.wiley.com/doi/10.1111/mec.12755.

A. Tellier, S. J. Y. Laurent, H. Lainer, P. Pavlidis, and W. Stephan. Inference of seed bank parameters in two wild tomato species using ecological and genetic data. *Proceedings of the National Academy of Sciences*, 108(41):17052–17057, Oct. 2011. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1111266108. URL https://pnas.org/doi/full/10.1073/pnas.1111266108.

A. R. Templeton and D. A. Levin. Evolutionary Consequences of Seed Pools. *The American Naturalist*, 114(2):232–249, Aug. 1979. ISSN 0003-0147. doi: 10.1086/283471. URL https://www.journals.uchicago.edu/doi/abs/10.1086/283471.

J. Terhorst, J. A. Kamm, and Y. S. Song. Robust and scalable inference of population history froth hundreds of unphased whole genomes. *Nature Genetics*, 49(2):303–309, FEB 2017. ISSN 1061-4036. doi: {10.1038/ng.3748}.

K. R. Thornton. A C++ Template Library for Efficient Forward-Time Population Genetic Simulation of Large Populations. *Genetics*, 198(1):157–166, Sept. 2014. ISSN 1943-2631. doi: 10.1534/genetics.114.165019. URL https://academic.oup.com/genetics/article/198/1/157/6073418.

L. Torada, L. Lorenzon, A. Beddis, U. Isildak, L. Pattini, S. Mathieson, and M. Fumagalli. Imagene: a convolutional neural network to quantify natural selection from genomic data. *BMC Bioinformatics*, 20(9):337, Nov 2019. ISSN 1471-2105. doi: 10.1186/s12859-019-2927-x. URL https://doi.org/10.1186/s12859-019-2927-x.

R. Tournebize, V. Poncet, M. Jakobsson, Y. Vigouroux, and S. Manel. McSwan: A joint site frequency spectrum method to detect and date selective sweeps across multiple population genomes. *Molecular Ecology Resources*, 19(1):283–295, Jan. 2019. ISSN 1755-0998. doi: 10.1111/1755-0998.12957.

G. Upadhya and M. Steinrücken. Robust Inference of Population Size Histories from Genomic Sequencing Data. preprint, Genetics, May 2021. URL http://biorxiv.org/lookup/doi/10.1101/2021.05.22.445274.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2017.

D. L. Vendrami, L. S. Peck, M. S. Clark, B. Eldon, M. Meredith, and J. I. Hoffman. Sweepstake reproductive success and collective dispersal produce chaotic genetic patchiness in a broadcast spawner. *Science advances*, 7(37):eabj4713, 2021.

M. Verin and A. Tellier. Host-parasite coevolution can promote the evolution of seed banking as a bet-hedging strategy. *Evolution*, 72(7):1362–1372, 2018. ISSN 0014-3820. URL https://www.jstor.org/stable/48575279.

F. A. Villanea and J. G. Schraiber. Multiple episodes of interbreeding between neanderthal and modern humans. *Nature Ecology & Evolution*, 3(1):39–44, Jan 2019. ISSN 2397-334X. doi: 10.1038/s41559-018-0735-8. URL https://doi.org/10.1038/s41559-018-0735-8.

R. Vitalis, S. Glémin, and I. Olivieri. When genes go to sleep: the population genetic consequences of seed dormancy and monocarpic perenniality. *The American Naturalist*, 163(2):295–311, Feb. 2004. ISSN 0003-0147. doi: 10.1086/381041.

M. T. Vizzari, A. Benazzo, G. Barbujani, and S. Ghirotto. A revised model of anatomically modern human expansions out of africa through a machine learning approximate bayesian computation approach. *Genes*, 11(12), 2020. ISSN 2073-4425. doi: 10.3390/genes11121510. URL https://www.mdpi.com/2073-4425/11/12/1510.

J. Voznica, A. Zhukova, V. Boskova, E. Saulnier, F. Lemoine, M. Moslonka-Lefebvre, and O. Gascuel. Deep learning from phylogenies to uncover the transmission dynamics of epidemics. *bioRxiv*, 2021. doi: 10.1101/2021.03.11.435006. URL https://www.biorxiv.org/content/early/2021/03/31/2021.03.11.435006.

K. Wang, I. Mathieson, J. O'Connell, and S. Schiffels. Tracking human population structure through time from whole genome sequences. *PLOS Genetics*, 16(3), MAR 2020. ISSN 1553-7404. doi: {10.1371/journal.pgen.1008552}.

R. Wang, Y. Bai, Y.-S. Chu, Z. Wang, Y. Wang, M. Sun, J. Li, T. Zang, and Y. Wang. Deepdna: a hybrid convolutional and recurrent neural network for compressing human mitochondrial genomes. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 270–274, 2018. doi: 10.1109/BIBM.2018.8621140.

Z. Wang, J. Wang, M. Kourakos, N. Hoang, H. H. Lee, I. Mathieson, and S. Mathieson. Automatic inference of demographic parameters using generative adversarial networks. *Molecular Ecology Resources*, 21(8):2689–2705, 2021. doi: https://doi.org/10.1111/1755-0998.13386. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.13386.

R. S. Waples. Genetic estimates of contemporary effective population size: to what time periods do the estimates apply? *Molecular Ecology*, 14(11):3335–3352, 2005.

W. M. Waterworth, S. Footitt, C. M. Bray, W. E. Finch-Savage, and C. E. West. DNA damage checkpoint kinase ATM regulates germination and maintains genome stability in seeds. *Proceedings of the National Academy of Sciences of the United States of America*, 113(34):9647–9652, Aug. 2016. ISSN 0027-8424. doi: 10.1073/pnas.1608829113. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5003248/.

J. D. Watson and F. H. C. Crick. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, 171(4356):737–738, Apr. 1953. ISSN 0028-0836, 1476-4687. doi: 10.1038/171737a0. URL https://www.nature.com/articles/171737a0.

S. Whalen, J. Schreiber, W. S. Noble, and K. S. Pollard. Navigating the pitfalls of applying machine learning in genomics. *Nature Reviews Genetics*, 23(3):169–181, Mar 2022. ISSN 1471-0064. doi: 10.1038/s41576-021-00434-9. URL https://doi.org/10.1038/s41576-021-00434-9.

L. S. Whitehouse and D. R. Schrider. Timesweeper: Accurately identifying selective sweeps using population genomic time series. *bioRxiv*, 2022. doi: 10.1101/2022.07.06.499052. URL https://www.biorxiv.org/content/early/2022/07/07/2022.07.06.499052.

C.-A. Whittle. The influence of environmental factors, the pollen : ovule ratio and seed bank persistence on molecular evolutionary rates in plants. *Journal of Evolutionary Biology*, 19(1):302–308, Jan. 2006. ISSN 1010-061X. doi: 10.1111/j.1420-9101.2005.00977.x.

C. G. Willis, C. C. Baskin, J. M. Baskin, J. R. Auld, D. L. Venable, J. Cavender-Bares, K. Donohue, R. Rubio de Casas, and NESCent Germination Working Group. The evolution of seed dormancy: environmental cues, evolutionary hubs, and diversification of the seed plants. *The New Phytologist*, 203(1):300–309, July 2014. ISSN 1469-8137. doi: 10.1111/nph.12782.

C. Wiuf and J. Hein. Recombination as a point process along sequences. *Theoretical Population Biology*, 55(3):248–259, JUN 1999. ISSN 0040-5809. doi: {10.1006/tpbi.1998.1403}.

S. Wright. Coefficients of Inbreeding and Relationship. *The American Naturalist*, 56(645):330–338, July 1922. ISSN 0003-0147, 1537-5323. doi: 10.1086/279872. URL https://www.journals.uchicago.edu/doi/10.1086/279872.

S. Wright. EVOLUTION IN MENDELIAN POPULATIONS. *Genetics*, 16(2):97–159, Mar. 1931. ISSN 1943-2631. doi: 10.1093/genetics/16.2.97. URL https://academic.oup.com/genetics/article/16/2/97/6045152.

S. Wright. The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proceedings of the XI International Congress of Genetics*, 8:209–222, 1932.

K. Xu, W. Hu, J. Leskovec, and S. Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=ryGs6iA5Km.

A. T. Xue, D. R. Schrider, A. D. Kern, and A. Consortium. Discovery of Ongoing Selective Sweeps within Anopheles Mosquito Populations Using Deep Learning. *Molecular Biology and Evolution*,

10 2020. ISSN 0737-4038. doi: 10.1093/molbev/msaa259. URL https://doi.org/10.1093/molbev/msaa259. msaa259.

Z. Yang, W. W. Cohen, and R. Salakhutdinov. Revisiting semi-supervised learning with graph embeddings. *CoRR*, abs/1603.08861, 2016. URL http://arxiv.org/abs/1603.08861.

B. Yelmen and F. Jay. An Overview of Deep Generative Models in Functional and Evolutionary Genomics. *Annual Review of Biomedical Data Science*, 6(1):annurev–biodatasci–020722–115651, Aug. 2023. ISSN 2574-3414, 2574-3414. doi: 10.1146/annurev-biodatasci-020722-115651. URL https://www.annualreviews.org/doi/10.1146/annurev-biodatasci-020722-115651.

B. Yelmen, A. Decelle, L. Ongaro, D. Marnetto, C. Tallec, F. Montinaro, C. Furtlehner, L. Pagani, and F. Jay. Creating artificial human genomes using generative neural networks. *PLOS Genetics*, 17 (2):1–22, 02 2021a. doi: 10.1371/journal.pgen.1009303. URL https://doi.org/10.1371/journal.pgen.1009303.

B. Yelmen, A. Decelle, L. Ongaro, D. Marnetto, C. Tallec, F. Montinaro, C. Furtlehner, L. Pagani, and F. Jay. Creating artificial human genomes using generative neural networks. *PLOS Genetics*, 17(2):e1009303, Feb. 2021b. ISSN 1553-7404. doi: 10.1371/journal.pgen.1009303. URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1009303.

B. Yelmen, A. Decelle, L. L. Boulos, A. Szatkownik, C. Furtlehner, G. Charpiat, and F. Jay. Deep convolutional and conditional neural networks for large-scale genomic data generation, Mar. 2023. URL https://www.biorxiv.org/content/10.1101/2023.03.07.530442v1.

R. Ying, J. You, C. Morris, X. Ren, W. L. Hamilton, and J. Leskovec. Hierarchical graph representation learning with differentiable pooling. URL http://arxiv.org/abs/1806.08804.

T. Yue and H. Wang. Deep learning for genomics: A concise overview, 2018.

M. Zhang and Y. Chen. Link prediction based on graph neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/53f0d7c537d99b3824f0f99d62ea2428-Paper.pdf.

J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun. Graph neural networks: A review of methods and applications. 1:57–81. ISSN 2666-6510. doi: 10.1016/j.aiopen.2021.01.001. URL https://www.sciencedirect.com/science/article/pii/S2666651021000012.

M. Zhou, N. J. Sng, C. E. LeFrois, A.-L. Paul, and R. J. Ferl. Epigenomics in an extraterrestrial environment: organ-specific alteration of dna methylation and gene expression elicited by spaceflight in arabidopsis thaliana. *BMC genomics*, 20(1):1–17, 2019.

J. Zou, M. Huss, A. Abid, P. Mohammadi, A. Torkamani, and A. Telenti. A primer on deep learning in genomics. *Nature Genetics*, 51(1):12–18, Jan 2019. ISSN 1546-1718. doi: 10.1038/s41588-018-0295-5. URL https://doi.org/10.1038/s41588-018-0295-5.

E. Zuckerkandl and L. Pauling. In evolving genes and proteins, ed. by v. bryson & hj vogel, 1965.

D. Živković and A. Tellier. Germ banks affect the inference of past demographic events. *Molecular Ecology*, 21(22):5434–5446, Nov. 2012. ISSN 1365-294X. doi: 10.1111/mec.12039.

D. Živković and A. Tellier. All But Sleeping? Consequences of Soil Seed Banks on Neutral and Selective Diversity in Plant Species. In R. J. Morris, editor, *Mathematical Modelling in Plant Biology*, pages 195–212. Springer International Publishing, Cham, 2018. ISBN 9783319990705. doi: 10.1007/978-3-319-99070-5_10. URL https://doi.org/10.1007/978-3-319-99070-5_10.