Technische Universität München

TUM School of Life Sciences

Professur für Populationsgenetik

Integration and inference of biological traits from whole genome sequence polymorphism data

Thibaut Paul Patrick Sellinger

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

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitz: Prof. Donna Ankerst, Ph.D.

Prüfer der Dissertation:

1. Prof. Dr. Aurélien Tellier
2. Prof. Asger Hobolth, Ph.D.
3. Prof. Dr. Johannes Müller

Die Dissertation wurde am 26.10.2020 bei der Technischen Universität München eingereicht und durch die TUM School of Life Sciences am 15.03.2021 angenommen.

Declaration

I declare that all results shown were made by me.

Acknowledgements

I would like to thank everyone. I especially thank Diala Abu Awad, Hanna Maerkle and Gustavo Silva-Arias for their help and advices. It was my privilege to work with Pr. Aurélien Tellier.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Genomes, and especially genetic polymorphisms, are shaped by molecular forces, such as mutation and recombination, but also ecological forces intrinsic to, or independent of, the biology of the species [34]. Polymorphism data therefore contain a plethora of information that goes beyond the physiological functions encoded therein. Recent advances in sequencing technologies enable us to obtain whole genome sequence data for many individuals across several populations, even for non-model species [170, 171, 108, 51]. Unlocking information contained in genomes can provide unprecedented results, unveiling the history of humans [103] or of other existing and/or extinct species [69, 51].

The demographic history of a population (the variation of effective population size over time) results from environmental and demographic changes that existing and/or extinct species have experienced (population expansion, colonization of new habitats, past bottlenecks) [72, 52, 35, 59, 47, 51]. The demographic history can thus be linked to archaeological or climatic data, providing new insights on their consequent genomic signatures [59, 47, 51, 6, 69, 103, 99]. Evidence for migration events have been uncovered [103, 15, 176], as have genomic consequences of human activities on other species [37]. Linking demographic history to climate and environmental data can greatly supports the field of conservation genetics [40, 42, 128]. Such information can help ecologist in detecting effective population size decrease

[178], and thus serve as a guide in maintaining or avoiding the erosion of genetic diversity in endangered populations, and potentially predicting the consequences of climate change on genetic diversity [44]. In addition, studying the demographic histories of different species in relation to one another can unveil latent biological or environmental evolutionary forces [75], unveiling links and changes within entire ecosystems. At last, inferring the demographic history is also necessary in order to define thresholds for selection scan methods. Hence, an accurate demographic inference should yield more reliable selection results [163, 133, 156]. With the increased accuracy of current methods, the availability of very large and diverse data sets and the development of new theoretical frameworks, the demographic history has become a central theme in the field of evolutionary biology [54].

Each species undergoes specific ecological and biological forces. Hence, in this thesis, we will focus on explaining a population's (or species) history in the light of its biology by simultaneously inferring its demographic history and biological traits. Unveiling the demographic and biological history of a population from genomes requires statistical inferences. The latter relies on mathematical models describing the evolutionary processes populations undergo. This field of science is called Population Genetics.

1.2 Introduction to Population Genetics

Sequencing data of genome has revealed the existence of DNA sequence polymorphism among individuals of a population[170]. The frequency and distribution of polymorphism in the sequence results from the population's history. Thus, from a theoretical perspective, once can try to reconstruct the population's history (*i.e.* its complete genealogy) based on this observed sequence polymorphism. From this point of view, the main aspect of the presented work is to interpret and infer the genealogy topology along genomes from sampled sequences. Therefore, the work presented in this thesis greatly relies on the coalescence theory. It is therefor of primary importance to understand the foundation of this theory. We will first present the Wright-Fisher model which models the behavior of a population in the most simple and idealized way. We then briefly introduce the coalescence theory which describes a sample genealogy's distribution in the Wright-Fisher model.

Wright-Fisher Model

We first assume we have a panmictic population (*i.e* every individual can mate with every other individual with the same probability) of haploid individuals in absence of natural selection (*i.e* the genotype of an individual does not affect on its number of offspring). We also assume the population size to be constant in time. We assume the population to evolve per generation (*i.e* we assume time to be discrete and to be measured in number of generations). Assuming each individual has only one gene with two allelic states (A or a) observed in the population, which we assume do not affect on fecundity (neutral hypothesis), and that individuals of the generation t+1 chooses uniformly and independently of others one parent among the individuals of the generation t and inherits its allele. From here we can define the genetic drift, which is the variation of the different allele frequencies from one generation to the other due to random sampling. Genetic drift is a fundamental stochastic process and a major component of evolution. We call the described model/process the Wright-Fisher Model, capturing the stochasticity behind genetic drift. The Wright-Fisher Model is represented in Figure 1.1.

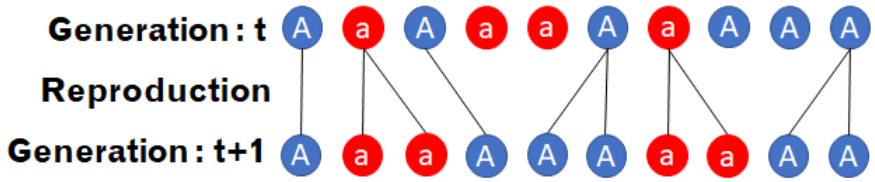

Figure 1.1: **Description of the Wright-Fisher Model.** Representation of a reproduction event in the Wright-Fisher model. Individual with allelic state a are represented in red and those with allelic state A in blue.

The Kingman Coalescent

Assuming we have a population old enough following the Wright-Fisher Model, we now study the time to the most recent common ancestor (TMRCA) of two randomly chosen distinct individuals from the population. We call the "merging" of the two lineages, a coalescence event. The coalescence process is represented in Figure 1.2.

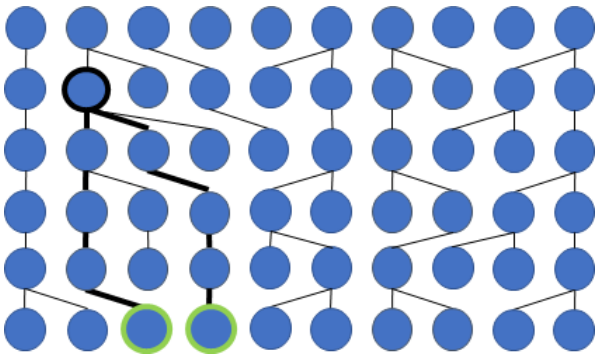

Figure 1.2: **Schematic representation of the coalescence process.** Representation of a Wright-Fisher model with the genealogy of a sample (circled in green) highlighted up to their most recent common ancestor (circled in black)

Then the TMRCA (in unit of twice the population size) tends to an exponential distribution of parameter 1 as the population size tends to infinite. The demonstration can be found in [95]. The formula can easily extended to any sample size. The intuition, is that we search among all pair of sampled individual which will coalesce first (*i.e.* we study the minimum of $\binom{M}{2}$ independent exponential distribution). The density of the waiting time to the most recent common ancestor of any 2 individuals in a sample of size M is thus:

$$P(TMRCA = t) = e^{-\binom{M}{2}t} \tag{1.1}$$

Hence, using the coalescence theory of Kingman, one can model the genealogical process of a sample in a population following the Wright-Fisher model. One can thus reconstruct (under the model's assumptions) the genealogy of a sample with no prior information on the population.

Understanding the coalescence process is of major importance since the history of a population is characterized by its genealogy. In practice, the distribution of genealogies of a sample will describe the population's history since the coalescent rate is in a unit of 2 N (N being the population size). Thus if N changes in time, there will be a variation of the coalescence rate in time, which will affect the distribution of genealogies in time. Thus Variation of population size but also ecological forces or specific species biological traits leading to violation of the Wright-Fisher model's assumptions will affect the sample's genealogy distribution. Hence detecting deviation from expected results can help unveil the population's specific history. Because we start to

understand the evolution of genomes throughout generations (*e.g.* mutation and recombination) and model them, we are able to infer genealogies from sequenced individuals.

1.3 The Sequentially Markovian Coalescent

As mentioned before, genomes can evolve throughout generations (especially during the reproductive events). During reproduction, there can be errors while replicating DNA sequences (mutation rate per generation is noted μ), resulting in what we call mutations (Figure 1.3). If two individuals have a very recent common ancestor, the number of generations (or reproductive events) splitting them is small. Hence, few mutations are expected, and thus less diversity (or the number of differences when comparing their respective DNA sequences) between the two individuals is expected. If their most recent common ancestor is far in the past (*i.e.* many reproductive events split the two individuals) then many mutations between their respective genome sequences are expected.

Figure 1.3: **Schematic representation of a mutation event.** A mutation occurred on the sequence during replication resulting in appearance of a different nucleotide (red) a the mutation position on the replicated sequence

In addition to mutation, recombination events (*i.e.* crossing over) can occur during the meiosis (recombination rate per generation is noted r). Assuming individuals are diploid (*i.e.* they have 2 copies of each chromosome), then the two chromosomes can randomly "exchange" pieces of chromosome (Figure 1.4). Because of recombination, pieces of the same chromosome can thus have a different genealogy.

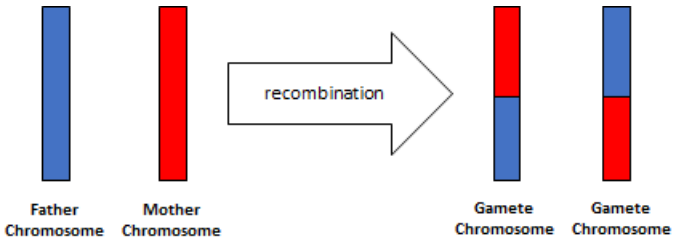

Figure 1.4: **Schematic representation of a recombination event.** The chromosome inherited by the father is represented in blue and the one by the mother in red. Effect of a recombination event on the chromosome is represented.

We classically define ρ as the effective number of recombination and θ as the effective number of mutation per locus (*i.e.* on a defined piece of genome sequence). In the Wright-Fisher model, we have $\rho = 4N_e r$ and $\theta = 4N_e \mu$ (N_e being the effective population size). Thus : $\frac{r}{\mu} = \frac{\rho}{\theta}$.

To model the variation of the genealogy along the chromosome (or sequence) we use the Ancestral Recombination Graph (ARG). A schema of ARG is represented in Figure 1.5. The distribution of the ancestral recombination graph of a sample has been described under a Wright-Fisher model in [84]. However, the model of [84] can become computationally very heavy with increasing sample size or sequence length. This computational load can make inferences or simulations intractable. Therefore, a new process has been introduced to model the ARG. In [180], they model the ARG as an inhomogeneous Poisson process along the sequence. This process has been approximated using a markov chain to model ARG by [115], corrected by [111] to become the Sequentially Markovian Coalescent (SMC). The SMC approximates with high accuracy the ARG assuming the Poisson process modeling recombination event which goes along the sequence as Markovian [179].

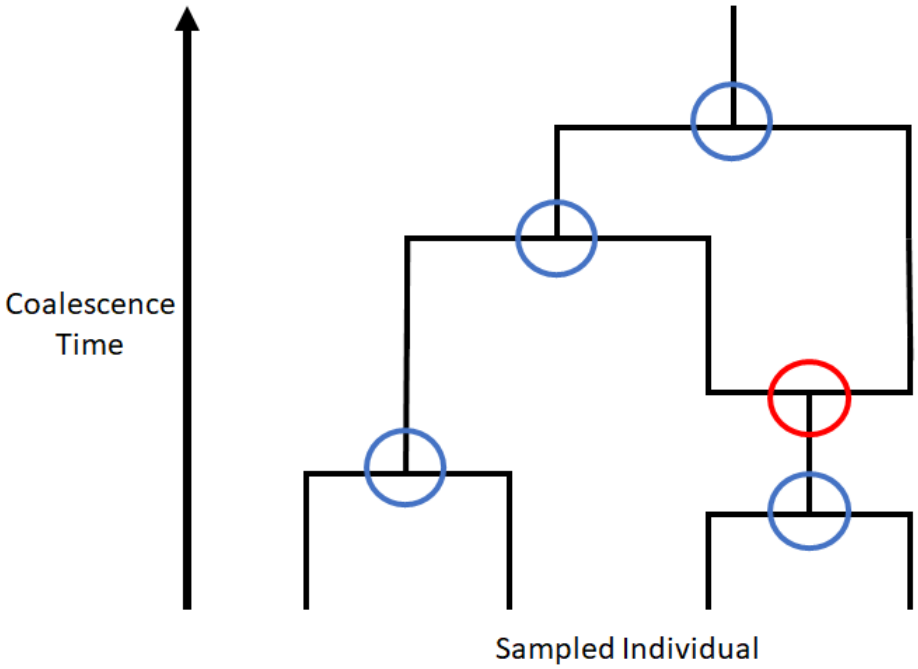

Figure 1.5: **Schematic representation of an ARG.** Representation of a genealogy of 4 sampled individual in presence of recombination. Coalescence event are represented by blue circle. Recombination event are represented by a red circle. Genealogy of the individual undergoing a recombination is now represented by two branches. The sequence on the left of the position of the recombination event has the left genealogy on the ARG, the sequence on the right has the genealogy represented on the right.

Using the SMC, [103] have build a Hidden Markov Model (HMM) named the Pairwise Sequentially Markovian Coalescent (PSMC) to infer parameters shaping the ARG distribution. The insight behind the PSMC is represented in Figure 1.6. The PSMC compares two sequences, if the number of mutations between the two sequences is high, the coalescence time must be far in the past. If no or few mutations are observed, the most recent common ancestor at this position of the sequences must be in very recent time. Hence the PSMC captures the variation segregating sites density along the genomes and uses the SMC to model and infer the variation of genealogy along the sequence (*i.e.* the ARG). From the distribution of genealogies along the se-

quence, the recombination rate and the demographic history can be inferred.

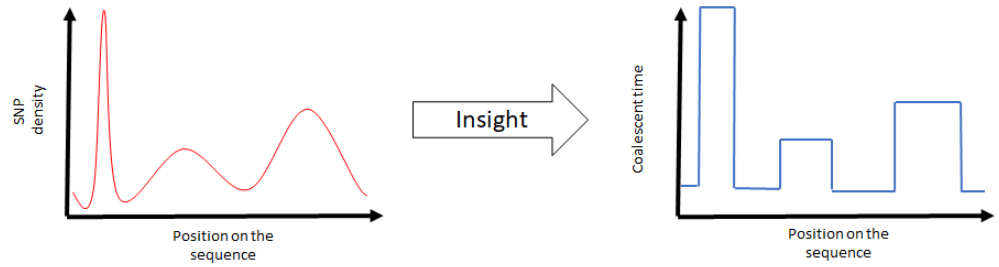

Figure 1.6: **Insight behind the PSMC.** SNPs density along the sequence is represented on the left and the insight behind the coalescence time along the sequence is represented on the right.

1.4 Hidden Markov Models

As the PSMC is a Hidden Markov Model, we will here briefly describe what are HMM and why we here have an HMM. An HMM models an unobserved Markov process which emits a signal (observed data) conditioned on the unobserved Markov process. Here, the exact genealogy of all individuals from a population/species is unknown (when looking at an individual, its complete genealogy does not display). Hence, the genealogy can be considered as a latent (hidden/unobserved) variable. However, we can sequence individual (to obtain their DNA sequences), using the population genetics theory and molecular biology, we know that the distribution of mutation and recombination events are conditioned by the genealogy (more precisely the ARG). Thus, if we take a sample of size two, the genealogy (or coalescence time to their most recent common ancestor) can become a hidden state, and the sequence polymorphism data the observed signal. Under the SMC theory, the sequence of hidden states (*i.e.* the coalescence time) is a Markov process. Thus, we have an HMM. A schematic Hidden Markov Model is represented in Figure 1.7.

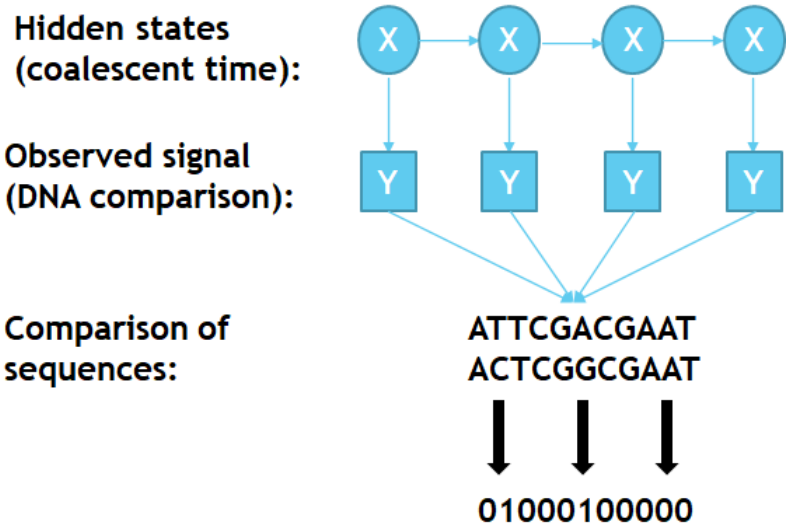

Figure 1.7: **Schematic representation of our HMM.** Hidden state are presented by variation X and the signal by Y. The hidden states conditioned the

1.5 Overview and Authors contribution

Now that the main topic and essential tools used in this thesis have been defined, we will briefly describe the chapters of this thesis.

Chapter 2 of this thesis describes the ecological Sequentially Markovian Coalescent (eSMC), a new implementation of the Pairwise Sequentially Markovian Coalescent [146] where self-fertilization and dormancy have been integrated. In this chapter I demonstrate the capacity of our model to simultaneously infer demographic history and self-fertilization or dormancy. We also demonstrate how self-fertilization and dormancy can strongly affect demographic history estimation if not correctly accounted for. This chapter was partially published in Plos Genetics (https://doi.org/10.1371/journal.pgen.1008698) since the manuscript version is more detailed. Thibaut Paul Patrick Sellinger designed the study, simulated all the data sets and build the data sets for *Arabidopsis thaliana* and *Daphnia pulex*. Thibaut Paul Patrick Sellinger carried out all analyses and implemented the new approach eSMC through an R package (https://github.com/TPPSellinger/eSMC). Further, Thibaut Paul Patrick Sellinger drafted the manuscript which was then read and com-

mented by Pr. Dr. Aurélen Tellier, Dr. Diala Abu Awad and Pr. Dr. Markus Möst.

Since this thesis focuses on inferring biological trait , encountered species might violate commonly made hypotheses in addition to the fact that only poor data sets may be available for non-model species. Hence, chapter 3 of this thesis will focus on the convergence properties of the Sequentially Markovian Coalescent and understanding the consequences of commonly violated hypothesis or data issue. This chapter has been recommended by PCI evolutionary Biology. Thibaut Paul Patrick Sellinger designed the study, simulated all the data sets, carried out all analyses, implemented the new approach eSMC2 through an R package (https://github.com/TPPSellinger/eSMC2). Thibaut Paul Patrick Sellinger drafted the manuscript which was then read and commented by Pr. Dr. Aurélen Tellier and Dr. Diala Abu Awad. (10.24072/pci.evolbiol.100115).

Chapter 4 will focus on breaking down the hypothesis of the Kingman coalescent. Chapter 4 describes the Sequentially Markovian Beta Coalescent (SMβC) (*i.e.* we assume the genealogy distribution follows a Beta Coalescence process). The SMβC thus accounts for potential multiple merger (or collisions) events in the sample's genealogy, allowing for more than two individuals to simultaneously coalesce. Multiple collisions might be more realistic than the Kingman coalescent process (which is limited to binary coalescence event) to model genealogy of species with skewed offspring distribution or populations undergoing strong selection. Thibaut Paul Patrick Sellinger designed the study, simulated all the data sets, carried out all analyses, implemented the new approach SMBC through an R package. The simulator was jointly build by Dr. Jerome Kelleher, Dr. Eldon Bjarki, Dr. Jere Koskela and Thibaut Paul Patrick Sellinger. The theoretical foundation of SMBC was jointly made By Dr. Fabian Freund and Thibaut Paul Patrick Sellinger.

Chapter 5 focuses on simultaneously integrating methylation data (*i.e.* presence or absence of methylated cytosine) and genome polymorphism data (*i.e.* segregating sites) in the Sequentially Markovian coalescent theory. Accounting methylated data in genome analysis increases methods accuracy to unreached heights but also potentially unlock events that do not have enough "SNPs signature" to be recovered with mutations only. Integrating

methylated data makes our approach the state of the art by outperforming any other similar method (which none of them integrate methylation) given the same amount of data. Thibaut Paul Patrick Sellinger designed the study, simulated all the data sets, carried out all analyses, implemented the new approach SMCm through an R package. The simulator was build by Thibaut Paul Patrick Sellinger as well as the theoretical foundation of SMCm under the supervision and advises of Pr. Dr. Aurélien Tellier and Pr. Dr. Frank Johannes.

Chapter 6 will summarize and discuss the obtained results throughout this thesis. A detailed argumentation on why and how species ecology and biology should be accounted to better extract information from whole genome sequence data is given. Finally, a discussion concerning central topics in population genetics but going beyond the scope of this thesis will be presented.

Chapter 2

Inference of demographic history, dormancy and self-fertilization rates

2.1 Motivation

Models and methods have been developed to extract previously unavailable information from whole genome sequence data [168, 103, 146, 152, 45, 156]. Inferences are based on modeling single nucleotide polymorphism (SNPs) along the genome across individuals, the density of which results from the interplay between mutation, time to common ancestors, and recombination. The common denominator in all these methods is their reliance on the per-site ratio recombination (r) and the mutation (μ) rate of the species ($\frac{r}{\mu}$), or, more precisely, on its effective value $\frac{\rho}{\theta}$. We note that so far, applications of these approaches have considered these ratios as interchangeable [146, 168, 103], which is a strong assumption and may be violated in some species that do not fulfill the assumptions of the classic Wright-Fisher diploid model with two sexes (*e.g.* equal sex-ratio, sexual reproduction at each generation and no overlap of generations). In humans and mammals, as $\rho = 4N_e r$ and $\theta = 4N_e \mu$ (N_e being the effective population size), we indeed find $\frac{r}{\mu} = \frac{\rho}{\theta}$. Yet, even in this case, biases arise if $\frac{\rho}{\theta} > 1$ [168]. In such cases, the number of mutations is not sufficient to detect all recombination events. The model is therefore no longer able to correctly estimate the Ancestral Recombination Graph (ARG) of the sample, *i.e* the superposition of coalescence trees at different positions

on the genome to display genealogies of sequences in the presence of recombination. Generally, it becomes necessary to extend existing approaches to account for characteristics and traits of species that can influence ρ or θ, and thus define when these methods can be accurately applied. It is additionally of interest to assess the accuracy of such methods for various values of the ratio $\frac{\rho}{\theta}$.

Current methods rely on the Sequentially Markovian coalescence (SMC) [115, 111] to account for the linear structure of genome sequences. At the first position, a genealogy is built under the neutral coalescence and, in a second step, recombination and linkage disequilibrium are incorporated using a Poisson process [180, 181]. By applying Hidden Markov Models it, therefore, becomes possible to calculate the probabilities of whole genome sequence data and infer the most likely values of the model parameters. These approaches can thus infer 1) the changes in population size (χ_t, where $\chi_t = \frac{N_t}{N_e}$, N_e and N_t being the effective population size and the current population size at time t, respectively) by inferring any variation of the coalescence rate in time, and 2) the ratio of effective recombination over the effective mutation rate $\frac{\rho}{\theta}$. From this ratio, the recombination rate can be estimated if the mutation rate is known (assuming $\frac{\rho}{\theta} = \frac{r}{\mu}$).

The described methods have been almost exclusively built to be applied to hominid data, therefore rely on several assumptions that are violated in many species (and most likely also in hominids): non-overlapping generations, equal sex ratio, sexual reproduction through random mating. Indeed, with the rise of next-generation sequencing technology, these methods are now frequently applied to whole genome sequences of species with characteristics that greatly differ from humans [69, 108, 39, 171]. In many species (*e.g.* plants, invertebrates) life-history strategies, such as mating systems or offspring production, influence the relationship between r and ρ and μ and θ [34]. If these effects are not accounted for, inferences using these methods may be biased and lead to misinterpretation of the results.

Two very common features in plant and invertebrate species are the maintenance of offspring as seed- or egg-banks [14, 63, 7] and self-fertilization [86]. Indeed, as a consequence of environmental fluctuations, species can develop bet-hedging strategies such as seed-banking [166, 62, 101]. This strategy increases the observed diversity [127, 174] and affects the rate of selection and

neutral genomic evolution [77, 165]. Due to the discrepancy between census (N_{cs}) and effective population size (N_e) caused by seed-banks [167, 127], we expect that $\frac{\rho}{\theta} \neq \frac{r}{\mu}$. Seed-banks can therefore strongly bias demographic inference if ignored [188]. Self-fertilization, on the other hand, decreases the effective population size. This reproductive strategy has evolved many times independently and is one of the most common evolutionary transitions observed in flowering plants [4]. The main consequence of this mating system is increased homozygosity, which directly results in a decreased effective recombination rate (ρ) compared to the molecular recombination rate r (since recombination events between two homozygous haplotypes are invisible), as well as a reduction in genetic diversity [5]. Due to their contradictory effects on the effective population size, the simultaneous occurrence of these traits (dormancy and self-fertilization) may in fact be missed, and extensions of inference methods to account for them could not only allow for more accurate inferences of parameters and demographic histories of species with these traits but could also provide a means with which to detect their respective rates.

To account for self-fertilization and seed-banks (or egg-banks) we develop a modified version of PSMC' [146], named ecological Sequentially Markovian Coalescent (eSMC). PSMC' refers to the MSMC using only two haplotypes [146], which is slightly different from the original PSMC [103]. Our model uses the deviation between the ratios $\frac{\rho}{\theta}$ and $\frac{r}{\mu}$ to infer self-fertilization and the existence of seed-banks. However, confounding effects arise when estimating both simultaneously. We first apply eSMC to simulated data to demonstrate its accuracy and then to genome sequence data of a plant, *Arabidopsis thaliana*, and an invertebrate species, *Daphnia pulex*. In these species, self-fertilization and/or seed/egg-banks have been observed or suspected. *A. thaliana* presents a very high self-fertilization rate of 99% [1] and it has been suggested that Scandinavian populations may have evolved seed-banks in Sweden [93] and Norway [107]. *D. pulex* exhibits cyclical parthenogenesis, *i.e.* a cyclical alternation of phases with asexual and sexual reproduction and is known to have dormant eggs produced through sexual reproduction [108, 30]. These resting eggs can potentially build up an egg-bank in the lake sediment as observed in many *Daphnia* species [14, 2]. First, we present a method based on the SMC using polymorphism data to infer the germination and/or self-fertilization rates jointly with the past demographic history.

Second, we study the effect of variable ratios of $\frac{\rho}{\theta}$ (and $\frac{r}{\mu}$) on the accuracy of estimates of past demography. Third, we apply our method to existing datasets from *Arabidopsis thaliana* and *Daphnia pulex* which have well documented high self-fertilization rates and egg-banks, respectively. We find a strong signature of self-fertilization in *Arabidopsis thaliana* and a strong signature of egg-banks in *Daphnia pulex*. We find that self-fertilization has little effect on the inference of the demographic history, whereas neglecting seed-banks can strongly affect the inferred population size.

2.2 Materials and methods

The coalescence with seed-bank and self-fertilization

We model population seed-banks using the same hypotheses described in [88]. Under these assumptions, seed-banking can be accounted for by rescaling the coalescence rate by β^2, where β ($0 \leq \beta \leq 1$) is the germination rate, more precisely the expected germination probability at each generation ($\beta = 1$ implying that there is no seed-bank). The probability that two lineages find a common ancestor in the active population is slowed by a factor $\beta \times \beta$ when looking backward in time. Hence, the expected coalescence times are increased by a factor $\frac{1}{\beta^2}$. Assuming mutations can arise during the dormant stage at the same rate as in the active population, we expect to have $\frac{1}{\beta^2}$ more mutations [127, 166, 77]. As recombination only occurs in the active population and concerns only one lineage backward in time, it is rescaled by β [165]. Because coalescence times are $\frac{1}{\beta^2}$ longer, we obtain (scaled in units of 4N):

$$\rho = \frac{\beta r}{\beta^2} = \frac{r}{\beta} \text{ and } \theta = \frac{\mu}{\beta^2}, \text{ so that } \frac{\rho}{\theta} = \frac{\beta^2 r}{\mu\beta} = \frac{\beta r}{\mu} \tag{2.1}$$

To model self-fertilization, we adopted the island model described in [126], where σ ($0 \leq \sigma \leq 1$) represents the proportion of offspring produced through self-fertilization (if $\sigma = 1$ all individuals are produced through self-fertilization). As a consequence, the coalescence rate is increased by a factor $\frac{2}{2-\sigma}$ [117] and the recombination rate is decreased by a factor $\frac{2-2\sigma}{2-\sigma}$ [126] since recombination events in homozygous individual are invisible. In the case of self-fertilization, we thus find (scaled in units of 4N):

$$\rho = \frac{2(1-\sigma)r}{(2-\sigma)}\frac{(2-\sigma)}{2} = (1-\sigma)r \text{ and } \theta = \frac{\mu(2-\sigma)}{2}, \tag{2.2}$$

$$\text{so that } \frac{\rho}{\theta} = \frac{r2(1-\sigma)}{\mu(2-\sigma)} \tag{2.3}$$

To simultaneously model seed-banking and self-fertilization we assume their effects to be independent and that there is no correlation between dormancy and the rate of self-fertilization. Under this assumption we can simply multiply their effects as in [174], giving the relationship $\frac{\rho}{\theta} = \frac{2(1-\sigma)\beta r}{(2-\sigma)\mu}$. We, therefore, have a confounding effect between self-fertilization and seed-banking when observing the recombination and mutation ratio $\frac{\rho}{\theta}$. Because of their opposing effects on the effective population size (seed dormancy increasing it, and self-fertilization decreasing it), the effects of these traits can be compensated by one another. As consequence, in our model seed-banking is mathematically equivalent to self-fertilization with a higher effective population size.

ecological Sequentially Markovian Coalescent (eSMC)

The eSMC is a Hidden Markov Model along two haplotypes. It is an extension of the PSMC' algorithm [146]. It adds the possibility of taking seed-banks and self-fertilization into account and simultaneously estimating their rates along with the demographic history. As in PSMC', we assume neutrality, an infinite site model, and a piece-wise constant population size. To define our HMM we need to precisely define all the following objects: the signal (observed data), the hidden states (coalescence time), the emission probabilities (probabilities of observing the data conditional to the hidden states), transition probabilities (probabilities of jumping from one hidden state to another) and the probabilities of the initial hidden states. The demonstrations of the results presented here can be found in Appendix 1.2. A detailed list of symbols and parameters used throughout chapter 2 can be found in Table 2.1.

Symbol	Meaning
r	Molecular Recombination rate per nucleotide per generation
μ	Molecular Mutation rate per nucleotide per generation
ρ	Effective recombination rate per nucleotide per generation
θ	Effective mutation rate per nucleotide per generation
β	Germination rate
σ	Self-fertilization rate
χ_t	Population size scaling vector (i.e $N_t = \chi_t N_0$)

Table 2.1: Symbol table

The signal (or observed data) depends on the hidden state and is a chain of 0s and 1s. To construct this signal, as in PSMC', two genome sequences are compared at each position; if at a given position, the two nucleotides are the same on both sequences, this is indicated by a 0, otherwise by a 1. As is necessary in HMM, the number of hidden states (or the coalescence times) must be finite, which is achieved by discretizing time. Therefore, the hidden state at one position is α if the coalescence time between the two haplotypes at that position is between T_α and $T_{\alpha+1}$. Given the model parameters, we know the expected coalescence time (which is $\frac{(2-\sigma)}{2\beta^2}$), and we define T_α as :

$$T_\alpha = \frac{-(2-\sigma)\ln(1-\frac{\alpha}{k})}{2\beta^2} \tag{2.4}$$

Here, k is the number of hidden states and α is an integer value between 0 and $k-1$. σ and β are the self-fertilization and the germination rate, respectively.

The emission probability P is the probability of observing the signal (chain of 0's and 1's) conditional to the hidden states (coalescence time). As in the PSMC' algorithm, we consider an infinite site model. The emission rate is therefore given by:

$$\begin{aligned} P(0|\alpha) &= 1 - e^{-2\mu t_\alpha} \\ P(1|\alpha) &= e^{-2\mu t_\alpha}, \end{aligned} \tag{2.5}$$

Where μ is the mutation rate per base pair and t_α the expected coalescence time in interval α. We find :

$$t_\alpha = \frac{T_\alpha - T_{\alpha+1}e^{-\Delta_\alpha \Lambda_\alpha}}{(1-e^{-\Delta_\alpha \Lambda_\alpha})} + \frac{1}{\Lambda_\alpha} \tag{2.6}$$

With:

$$\Delta_\alpha = T_{\alpha+1} - T_\alpha \text{ , } \Lambda_\alpha = \frac{2\beta^2}{(2-\sigma)\chi_\alpha} \text{ , and } \chi_\alpha = N_\alpha/N_e \tag{2.7}$$

Where Δ_α is the duration (in coalescence time) of interval α, Λ_α is the coalescence rate in the time window α, N is the effective population size and N_α is the population size during the time interval α. Using N and N_α, we can calculate χ_α which represents the variation of population size over time. It is this value that is inferred by the model.

The transition probabilities are the probabilities of going from one hidden state to another. We find:

$$p(\alpha|\gamma) = \begin{cases} \frac{P_\gamma}{2t_\gamma}((\sum_{\eta=1}^{\alpha-1} \frac{(1-e^{-2\Delta_\alpha\Lambda_\alpha})e^{-\int_{T_{\eta+1}}^{T_\alpha} 2\Lambda_v dv}(1-e^{-\Delta_\eta 2\Lambda_\eta})}{2\Lambda_\eta}) \\ \qquad +(\Delta_\alpha - \frac{(1-e^{-\Delta_\alpha 2\Lambda_\alpha})}{2\Lambda_\alpha})) & \alpha < \gamma \\ \frac{P_\gamma}{t_\gamma}(\sum_{\eta=1}^{\gamma-1} e^{-\int_{T_{\eta+1}}^{t_\gamma} 2\Lambda_v dv}\frac{(1-e^{-2\Delta_\eta\Lambda_\eta})}{2\Lambda_\eta} \\ \qquad +\frac{(1-e^{2(T_\gamma - t_\gamma)\Lambda_\gamma})}{2\Lambda_\gamma})e^{-\int_{t_\gamma}^{T_\alpha}\Lambda_v dv}(1-e^{-\Delta_\alpha\Lambda_\alpha}) & \alpha > \gamma \\ 1 - (\sum_{\alpha=0}^{\gamma-1} p(\alpha|\gamma) + \sum_{\alpha=\gamma+1}^{k} p(\alpha|\gamma)) & \alpha = \gamma \end{cases} \tag{2.8}$$

Where P_γ is the recombination probability between two base-pairs:

$$P_\gamma = (1 - e^{-2rt_\gamma \frac{2\beta(1-\sigma)}{(2-\sigma)}}) \tag{2.9}$$

The initial probability corresponds to the first state probability. We assume this probability to be the equilibrium probability $q_o(\alpha)$ (probability of being in state α at the first position). We find:

$$q_o(\alpha) = e^{\sum_{\eta=0}^{\alpha-1} -\Lambda_\eta\Delta_\eta}(1 - e^{-\Delta_\alpha\Lambda_\alpha}) \tag{2.10}$$

Simulated (pseudo-observed) Sequence data

Throughout this chapter we use five different demographic scenarios: 1) constant population size, 2) expansion, 3) bottleneck and recovery, 4) decrease

and 5) "sawtooth" (a succession of expansions and decreases). These scenarios are simulated for different combinations of the self-fertilization rate ($0 \leq \sigma \leq 0.9$) and the germination rate ($0.1 \leq \beta \leq 1$). Different sequence lengths are tested, as are combinations of mutation and recombination rates. To simulate our data, we use a modified version of the coalescence simulation program scrm [159]. This modified version integrates seed-banking (or egg-banking) and self-fertilization. The simulator is available on our GitHub repository (https://github.com/TPPSellinger/escrm). On all the simulated data, four different algorithms are used to estimate demographic history and recombination rate: our algorithm eSMC, which we compare to PSMC', MSMC ,and MSMC2. PSMC', MSMC and MSMC2 are run with default parameter and 1 as initial value for $\frac{\rho}{\theta}$.

Sequence data

We use 12 whole genome sequences (hence all five chromosomes) of European *A. thaliana* from the 1001 genome project [36, 171], six individuals sampled in Sweden (id : 5830, 5836, 5865, 6077, 6085 and 6087) and six from Germany (id : 7231, 7250, 7255, 7337, 7415 and 7419). Each individual is considered haploid because of very high levels of homozygosity [69]. We obtained polymorphism data (that is processed vcf files) from the authors of the study [69]. The mapping to the reference genome and SNP calling was performed based on the pipeline in [69]. The mutation rate is set at 7×10^{-9} per generation per bp [129] and the chromosome-specific recombination rates are 3.4×10^{-8}, 3.6×10^{-8}, 3.5×10^{-8}, 3.8×10^{-8}, 3.6×10^{-8} per generation per bp for chromosome 1 to 5 respectively [144]. We first run the four different algorithms to estimate the demographic history and recombination rate (ignoring self-fertilization and seed-banks for eSMC). Analyses are run per chromosome (represented by the different lines in the figures). We then analyse the data again with eSMC, first accounting only for self-fertilization (β is fixed to 1 and σ is estimated), and then accounting for both self-fertilization and seed-banks using reasonable priors ($0.5 \leq \beta \leq 1$ and $0.5 \leq \sigma \leq 1$).

To infer the demographic history and the dormancy rates of *D. pulex*, we use six whole genome sequences from [108](id: SRR5004865, SRR5004866, SRR5004867, SRR5004868, SRR5004869 and SRR5004872) which are available under the accession SAMN06005639 in the NCBI Sequence Read Archive (SRA). We used the reference genome assembly PA42 v3.0 which is avail-

able at the European Molecular Biology Laboratory (EMBL) nucleotide sequencing database under accession PRJEB14656 [58]. The raw data is first trimmed using bbtools to remove duplicates, trim adapters, remove synthetic artifacts, spike-ins, and perform quality-trimming based on minimum read quality of 40. Then we mapped reads using bwa (default parameters) onto the reference genome [102]. We used Samtools to convert sam to bam files [104] and GATK to remove PCR duplicates and perform local realignment around indels [48]. We used freebayes to call the SNPs and vcftools for post-processing (filtering). The pipeline is available on the GitHub repository: https://github.com/TPPSellinger/Daphnia_pulex_data . Note that the reference genome consists of 1,822 scaffolds (average length of 85,849) and thus to avoid bias in the analyses, we only kept scaffolds above 1 Mb retaining only the 19 largest scaffolds. As the phasing quality could not be guaranteed, we analyze sequence data of each *D. pulex* individual separately. The mutation rate is set at 4.33×10^{-9} per generation per bp [66] and the recombination rate at 8×10^{-8} per event of sexual reproduction per bp [183, 108]. To account for the number of generations before sexual reproduction takes place, the recombination rate is re-scaled by n_p which represents the total number of generations per year. If we consider $n_p = 5$, the recombination rate is scaled by 0.2 [108]. We also test how the number of parthenogenetic generations between sexual reproductive events could affect the quality of the inference, and rescale the recombination rate accordingly. The scenarios we test are: no parthenogenesis, two generations of parthenogenesis, and five generations of parthenogenesis, therefore rescaling the recombination rate by 1, 0.5, and 0.2 respectively. The sequences of each individual are analyzed with PSMC' and eSMC only, as MSMC and MSMC2 require accurate and reliable phasing, which is not the case for these sequences. We then account for egg-banks using eSMC and imposing no priors on β and setting $\sigma = 0$. The multihetsep files for *A. thaliana* and *D. pulex* analyses are available on the GitHub repository at https://github.com/TPPSellinger/Daphnia_pulex_data and at https://github.com/TPPSellinger/Arabidopsis_thaliana_data.

2.3 Results

We first study the theoretical accuracy and properties of our method on sequence data simulated under different scenarios. A demonstration of the model's accuracy in absence of seed bank and self-fertilization can be found

in [151]. We then analyze real sequence data from two European populations of *Arabidopsis thaliana*: one from Tübingen, Germany, where there is no seed-bank, and one from Sweden, where seed-banking is suspected while accounting for self-fertilization. We also analyze data from *Daphnia pulex*, for which egg-banking is known to be a prominent biological feature.

Simulation results

Convergence property with dormancy (seed- or egg-banks). Using eSMC on sequences simulated under the "sawtooth" scenario in the presence of seed-banks (mutation and recombination rates are set to 1.25×10^{-8} per generation per bp, Figure 2.1), we obtain an accurate estimation of the demography (χ_t) and of the germination rates (β). Under four germination rates β with values 1 (no seed-bank), 0.5 (two-year seed-bank), 0.2 (long-lived five-year seed-bank) and 0.1 (long-lived ten-year seed-bank), we respectively estimate an average germination rate of $0.88, 0.55, 0.24$ and 0.13. As seed-banks affect the time window of the estimated demography, more ancient events can be inferred when $\beta < 1$ [188]. In models where seed-banks cannot be accounted for (PSMC', MSMC, MSMC2), census population size is strongly overestimated when $\beta < 1$.

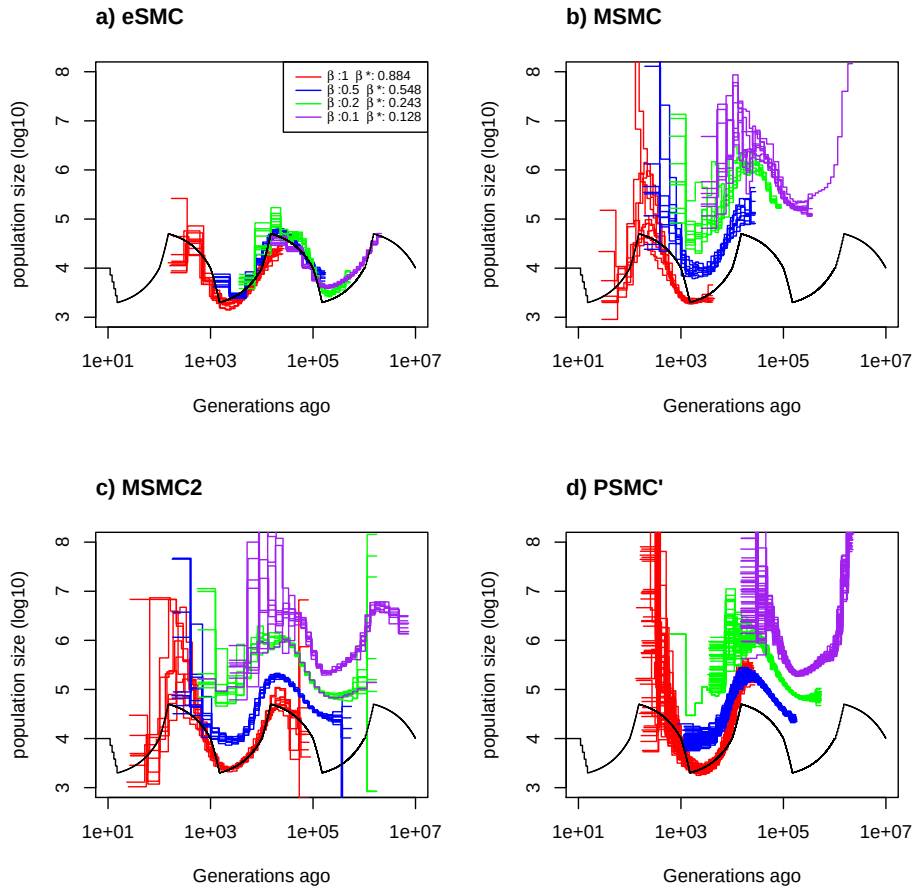

Figure 2.1: **Estimated demographic history with seed banking.**Estimated demographic history using four simulated sequences of 10 Mb and ten replicates under a sawtooth demographic scenario (black). The mutation and recombination rates are set to 1.25×10^{-8} per generation per bp. Therefore $\frac{r}{\mu} = 1$. We simulate under four different germination rates $\beta = 1$ (red), 0.5 (blue), 0.2 (green) and 0.1 (purple), hence we respectively have $\frac{\rho}{\theta} = 1$, $= 0.5$,0.2 and 0.1. The demographic history is estimated using a) eSMC , b) MSMC, c) MSMC2 and d) PSMC'. β^* represents the estimated germination rate by eSMC.

For simpler demographic scenarios (constant population size, bottleneck, expansion, and decrease, see Supplementary Figure A.1) and $\mu = r = 1.25 \times 10^{-8}$ per generation per bp, the germination rate and the demographic histories estimated by eSMC are accurate for all the demographic scenarios

considered.

Convergence property with self-fertilization. Under the "sawtooth" scenario with different rates of self-fertilization σ, with mutation and recombination rates set to 1.25×10^{-8} per generation per bp ($\frac{r}{\mu} = 1$), for four different self-fertilization rates $\sigma = 0$ (no self-fertilization), 0.5 (50% selfing), 0.8 (80% selfing) and 0.9 (90% selfing), we estimate the self-fertilization rate respectively at $0.17, 0.51, 0.76$ and 0.85 (Figure 2.2). eSMC infers a residual rate of self-fertilization (below 0.2). Yet, eSMC accurately estimates demography, while MSMC, MSMC2, and PSMC' exhibit a small bias in the estimation of the demographic history. Neglecting self-fertilization, therefore, seems to be of a smaller consequence than neglecting dormancy (see above), as self-fertilization has a very small impact on the inferred demographic history. Variance in the estimations increases for higher rates of σ. When the mutation rate is set to 1.25×10^{-8} per generation per bp and the recombination rate to 6.25×10^{-8} per generation per nucleotide ($\frac{r}{\mu} = 5$), the self-fertilization rate is overestimated for small values of σ (Supplementary Figure A.2), but well estimated for higher values of σ. The estimation of the demographic history remains accurate, though slightly biased for small values of self-fertilization. The other methods tested (PSMC', MSMC, MSMC2) present stronger biases in the estimated demographic history.

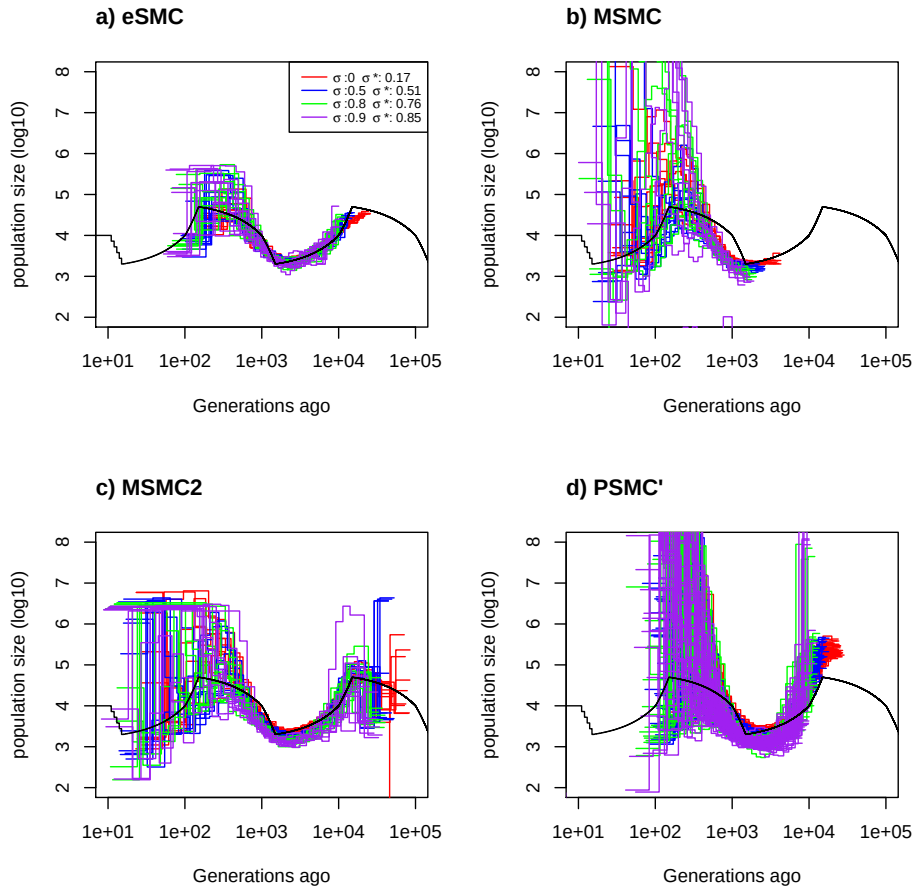

Figure 2.2: **Estimated demographic history with selfing.** Estimated demographic history using four simulated sequences of 10 Mb and ten replicates under a sawtooth demographic scenario (black). The mutation and recombination rates are set to 1.25×10^{-8} per generation per bp, and simulations were run for four different self-fertilization rates ($\sigma = 0$ (red), 0.5 (blue), 0.8 (green) and 0.9 (purple)), and as $\frac{r}{\mu} = 1$, this gives $\frac{\rho}{\theta} = 1$, 0.667,0.333 and 0.182 respectively. The demographic history is estimated using a) eSMC , b) MSMC, c) MSMC2 and d) PSMC'. σ^* represents the self-fertilization rate estimated by eSMC.

In the simpler demographic scenarios tested (Supplementary Figure A.3), the rate of self-fertilization is estimated fairly well, though there is an impact of the considered demographic scenario. However, in absence of self-fertilization, eSMC still infers a residual rate of self-fertilization (below 0.2).

Demographic history remains accurately estimated.

Convergence property with both dormancy and self-fertilization. Here we test different combinations of seed/egg-banks and self-fertilization rates that result in the same ratio $\frac{\rho}{\theta} = 0.15$, with $\frac{r}{\mu} = 1$ (setting $\mu = r = 1.25 \times 10^{-8}$ per generation per bp). Self-fertilization and dormancy have opposing effects on the coalescence rate, and thus cannot be simultaneously estimated from whole-genome data alone (Figure 2.3). These two rates are indeed simultaneously non-identifiable. Without any prior knowledge (blue in Figure 2.3), eSMC can't estimate the correct set of parameters. However, this shortcoming can be corrected to some extent by setting general "ecological" priors for either β or σ (*e.g.* $0 \leq \beta \leq 0.5$ or $0.5 \leq \sigma \leq 1$). In this case, eSMC can infer a demographic history of the correct shape but slightly shifted away from the true values of population size and time. eSMC tends to overestimate the values of β and σ, a consequence of which is the overestimation of the census population size (see Figure 2.3). However, while integrating prior knowledge on both parameters does not solve the non-identifiability issue, it does reduce the inferred range of values. Hence, including priors on both rates reduces the parameter space for which the confounding effect of joint estimation occurs. This is shown in Supplementary Figure A.4 in a plot showing all possible estimations of coupled variables of β and σ for a given parameter set. We also test how recombination can influence the output of these models, notably by taking a higher recombination rate (8.335×10^{-8} per-site per generation), more representative of the high recombination to mutation ratio observed in some species (notably *D. pulex* and *A. thaliana* [108, 144]). This gives $\frac{r}{\mu} = 6.667$ and $\frac{\rho}{\theta} = 1$, parameters for which the variance of the demographic history is smaller and the estimation of self-fertilization and germination parameters remain unchanged (Supplementary Figure A.5). All the possible estimated combinations of β and σ, given different sets of priors for this recombination rate and ratio $\frac{r}{\mu}$ are given in Supplementary Figure A.6 (results similar to those for $\frac{r}{\mu} = 1$ in Supplementary Figure A.4 are observed.)

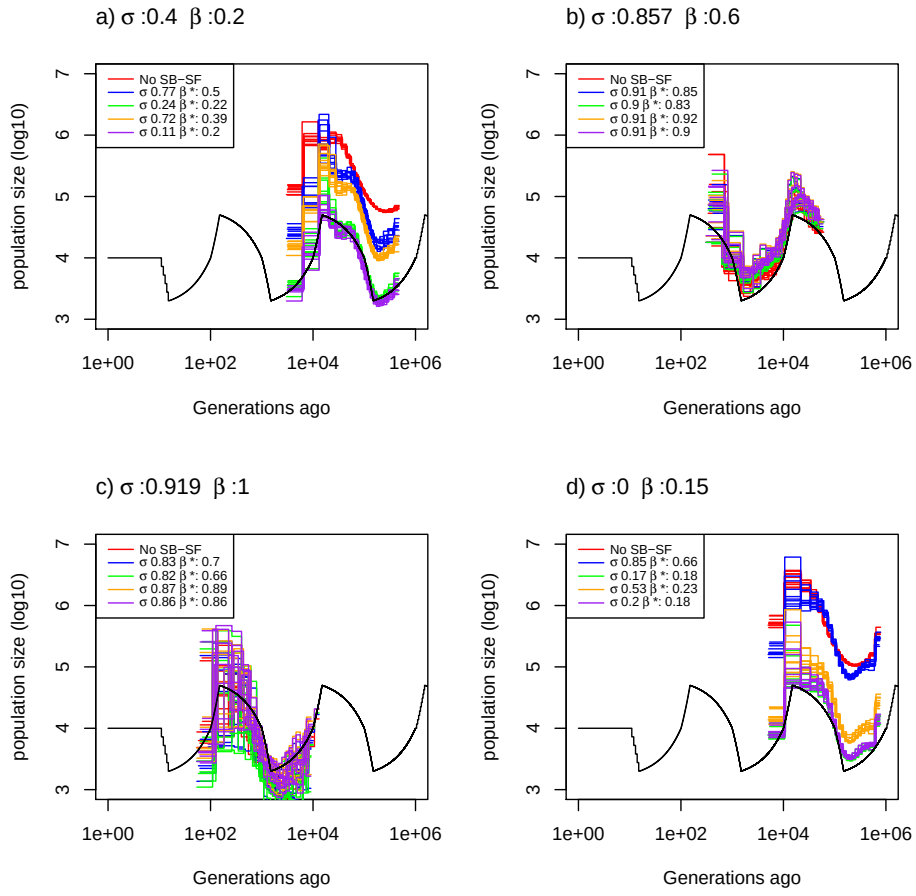

Figure 2.3: **Estimated demographic history with selfing and seed banking.** Demographic history estimated by eSMC for ten replicates using four simulated sequences of 10 Mb under a sawtooth demographic scenario and four different combinations of germination (β) and self-fertilization (σ) rates but resulting in the same $\frac{\rho}{\theta}$. Mutation and recombination rates are set to 1.25×10^{-8} per generation per bp, giving $\frac{r}{\mu} = 1$. The four combinations are : a) $\sigma = 0.4$ and $\beta = 0.25$, b) $\sigma = 0.75$ and $\beta = 0.6$, c) $\sigma = 0.85$ and $\beta = 1$ and d) $\sigma = 0$ and $\beta = 0.15$. Hence, for each scenario $\frac{\rho}{\theta} = 0.15$ For each combination of β and σ, eSMC was launched with five different prior settings: ignoring seed-banks and self-fertilization (red), accounting for seed-banks and self-fertilization but without setting priors (blue), accounting for seed-banks and self-fertilization with a prior set only for the self-fertilization rate (green), only for the germination rate (orange) or for both (purple). σ^* and β^* respectively represent the estimated self-fertilization and germination rate.

Inferring self-fertilization, seed-banks and demography in *Arabidopsis thaliana*

Using 12 individuals whole genome sequence data obtained from two accessions of *A. thaliana* (one from Sweden and the other from Germany), we inferred the demography of each population using eSMC, PSMC', MSMC, and MSMC2 (Supplementary Figure A.7). When ignoring self-fertilization, both populations have a common demographic history, similarly inferred by the different methods, except for MSMC, whose results exhibit a higher variance for the Swedish population. Furthermore, we observe a non-negligible deviation between the recombination rate estimated using these inference methods ($\frac{\rho}{\theta} < 1.2$, Supplementary Figure A.7) and what has been obtained using experimental approaches ($\frac{r}{\mu} = 5$) [144]. MSMC finds lower $\frac{\rho}{\theta}$ than other methods. When accounting only for self-fertilization (hence imposing $\beta = 1$), eSMC estimates a high self-fertilization rate averaged at $\sigma = 0.86$ in the German population and 0.87 in the Swedish one. These rates are not as high as what has been recorded previously [55, 1]. When running analyses per chromosome, we found no significant chromosome effect on these estimations. When simultaneously estimating β, σ, and the population size, we find a slightly lower $\sigma = 0.84$ in the German population and 0.86 in the Swedish one (Figure 2.4). Here, eSMC estimates a germination rate β higher than 0.9 in both populations, implying that there is no long-term seed-bank in either of them.

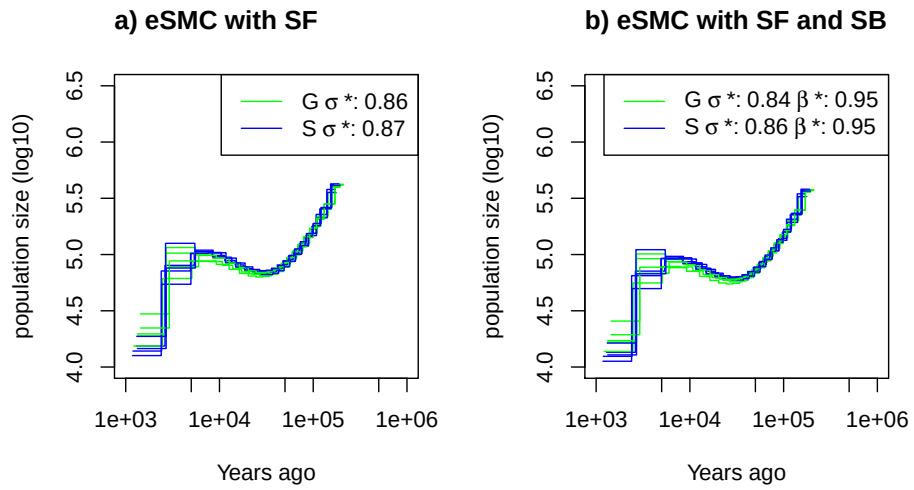

Figure 2.4: **Estimated demographic history of *Arabidopsis thalinana*.** Demographic history of two European (Sweden (S, blue) and German (G, green)) populations of *A. thaliana* estimated using eSMC : a) accounting only for selfing (σ is a variable and $\beta = 1$) and b) accounting simultaneously for selfing and seed-banking (σ bounded between 0.5 and 0.99 and β bounded between 0.5 and 1). Mutation rate is set to 7×10^{-9} per generation per bp and recombination respectively set for chromosome 1 to 5 to $3.4 \times 10^{-8}, 3.6 \times 10^{-8}, 3.5 \times 10^{-8}, 3.8 \times 10^{-8}, 3.6 \times 10^{-8}$) per generation per bp. σ^* and β^* respectively represent the estimated self-fertilization and germination rates.

Inferring egg-banks and demography in *Daphnia pulex*

The inferred demographic history of a single population of *D. pulex* has a similar shape using eSMC and PSMC' (Figure 2.6). The demographic history estimated by PSMC' is shifted vertically compared to eSMC since dormancy is ignored. The effective population size is hence overestimated compared to eSMC. We fix the self-fertilization rate at $\sigma = 0$ because during sexual cycles although *D. pulex* in principal could self-fertilize via intraclonal matings during sexual cycles, these matings are rare and it has been shown that selfing is negligible in this case [108]. *D. pulex* reproduces via cyclical parthenogenesis, i.e. alternating phases of ameiotic parthenogenesis (or more exactly abortive meiosis with no or very little recombination [78, 79]) and sexual reproduction. Hence, the inferred mean generation time before the hatching of dormant eggs produced by sexual reproduction depends on the number of parthenogenetic cycles that occur on average per year since mutations can occur during ameiotic parthenogenesis but recombination is very unlikely. Here one generation is considered to be of one cycle of asexual or sexual reproduction, several generations taking place in a single year. For this specific population, a maximum of five parthenogenetic cycles before sexual reproduction is assumed [108]. It is important to note that the number of parthenogenetic cycles can affect the ratio $\frac{\rho}{\theta}$. Therefore we tested the effect of the value of the average number of parthenogenetic cycles on the estimation of the germination rate. Independently of the number of parthenogenetic cycles, eSMC always detects dormancy, with $\beta < 0.5$. The average dormancy can therefore be bounded between 3 and 18 generations, revealing the existence of at least moderate dormancy in this species. Running analysis per scaffold and not per individual lead to slightly less dormancy and demographic history (Supplementary Figure A.8).

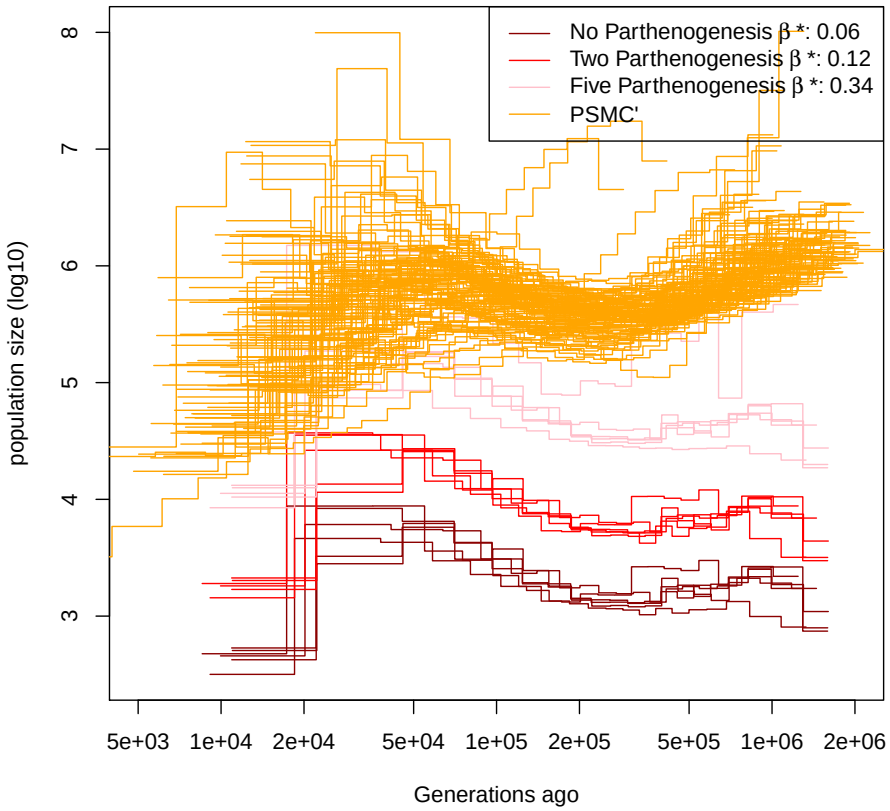

Figure 2.5: **Estimated demographic history of *Daphnia pulex*.** Demographic history estimated by eSMC on six individuals of *D. pulex* accounting for egg-banks (β is a variable and $\sigma = 0$). Different assumptions concerning the number of parthenogenetic cycles before the production of the dormant egg are made: Five cycles (pink), two cycles (red) and no parthenogenesis (dark red). Demographic history estimated by PSMC' are ploted in organge. Mutation and recombination rates are respectively set to 4.33×10^{-9} and $\frac{8\times10^{-8}}{n_p}$ per generation per bp, where n_p is the number of reproductive cycles per year, parthenogenetic and sexual.

Because of the life cycle of *D. pulex* is extremely short ($<$ 1 year) [108, 30], the hypothesis of having mutation accumulating in the seed bank at the same speed as those from reproductive events might be violated. Hence, we now assume that mutations in the dormancy stage accumulate 5 times slower

as during the reproductive event (Supplementary Figure A.9). Using this setting we still find dormancy but also a demographic history shifted in the past compared to Figure 2.6.

2.4 Discussion

The existing statistical inference methods based on full genome polymorphism data estimate the past demographic history under the assumptions of a model that is violated in many species. Here, we develop a method where ecological and life history traits can not only be accounted for but can also be inferred from sequence data along with the past demography. Ecology and life history traits can affect ρ and θ differently and our HMM can detect these differences through the estimation of $\frac{\rho}{\theta}$. However, this implies that some knowledge of the molecular ratio of recombination over mutation ($\frac{r}{\mu}$) is required. We demonstrate the capacity of our method to accurately recover the germination rate (and therefore the presence and strength of dormancy), though simulated results show that the violation of the infinite site assumption can lead to it being slightly underestimated. Similarly, our model can also retrieve the self-fertilization rate and we show that for high rates ($\sigma \geq 0.9$), more data are required to compensate for the variance observed and increase the accuracy of the estimation. Finally, our model cannot disentangle the genomic signatures of self-fertilization and seed-banks due to non-identifiability. The simultaneous estimation should thus be avoided, and only be performed if a priori knowledge on the presence of seed-banks (or a dormant stage) as well as the reproductive mode is available [7, 14].

Throughout the chapter, we have highlighted two main ratios that are of great importance when using inference methods: the ratios $\frac{r}{\mu}$ and $\frac{\rho}{\theta}$, respectively the per-site molecular and effective ratios of recombination over mutation rate. We have used the deviation between $\frac{r}{\mu}$ and $\frac{\rho}{\theta}$ to estimate the self-fertilization and/or germination rate. We also show that the demographic history can contribute to a departure of $\frac{\rho}{\theta}$ from $\frac{r}{\mu}$. Indeed, care must be taken concerning the initial value used for $\frac{\rho}{\theta}$: if the initial value is greater than one, the inferred demographic history will be flattened, regardless of the actual value of $\frac{\rho}{\theta}$. Furthermore, if the true value of $\frac{\rho}{\theta}$ is indeed greater than one, similar biases are expected, such as flattened demographic history. This observation, which is true irrespective of the presence of seed-banks or

self-fertilization, is due to insufficient information to correctly reconstruct the local genealogy, as a high ratio $\frac{\rho}{\theta}$ implies that few SNPs are present between the recombination spots on the genome. We highlight that the importance of $\frac{\rho}{\theta}$ for inference was mentioned in [168], but has largely been ignored in the literature on SMC-based methods, although a ratio of $\frac{\rho}{\theta}$ greater than one significantly alters the accuracy of inference.

When applying eSMC to sequences data of *A. thaliana*, we find evidence of strong self-fertilization with an estimated selfing rate of around 0.87. However, this rate is slightly smaller than what is known empirically for this species, where the current rate of self-fertilization has been estimated at 0.99 [1]. There are three possible explanations for this discrepancy. First, *A. thaliana* most probably evolved from outcrossing to highly self-fertilizing less than 400 thousand years ago [55], whereas our demographic inference dates further in the past. Self-fertilization would therefore have appeared within the time window of the inferred demographic history. As a consequence, our estimate of self-fertilization (constant in time) reflects the average effect of the varying the real self-fertilization rate within the time window. Second, the under-estimation may be due to limits of the self-fertilization model, which accounts only for homologous recombination events. Yet, other types of recombination or chromosomic re-arrangements do occur in genomes. These non-accounted mechanisms could increase the signature of recombination leading to an underestimation of the self-fertilization rate. Third, we infer the self-fertilization rate of a single population in isolation (Germany or Sweden) while the past demography of *A. thaliana* consists of episodes of admixture, migration, and recolonization from glacial refugia [36, 171], all of which are ignored in our model. The resulting complex population structure likely affects our estimates (see discussion in [140]).

It has long been observed that many *Daphnia* species, including *D. pulex*, have resting egg-banks. The sequences analyzed using eSMC agree with this hypothesis, as we find strong evidence of dormancy. The inferred duration of dormancy greatly depends on the number of parthenogenetic generations between sexual reproduction events. Indeed, parthenogenetic cycles increase the number of mutations compared to recombination events. If we take two extreme scenarios for the specific sampled population (no parthenogenesis versus 5 generations of parthenogenesis [108]) we find a duration of dormancy between 3 and 18 generations, slightly less than when performing the

analysis per scaffold. The discrepancy between those results suggest issues in the data or a hypothesis violation (*e.g* mutation or recombination rate are no constant along the genome). However, reducing the mutation rate during the dormant stage (which is a more realistic hypothesis), will shift the demographic history in the past, demonstrating that dormancy can not only shift the y axis but also the x axis. Finding dormancy in *Daphnia* agrees with empirical observations [30, 14], and confirms the major role of egg-banks in maintaining diversity in this species. The sequences used here originate from an ephemeral (*i.e.* non-permanent pond), and populations in such environments are expected to have both higher rates of sexual reproduction as well as longer-lived egg-banks [30]. It would therefore be interesting to test the existence of egg-banks and assess the germination rates in several *Daphnia* species and from different permanent and ephemeral water bodies. Our method presents a way forward to the detection of egg/seed/spore-banks of many invertebrates, plant, and fungal species, as well as their past demographic history using sequence data (as experimental validation of dormancy is difficult to obtain [166, 165]).

Our method represents a small but important step for the integration of ecological traits in whole genome sequence analysis through the ratio $\frac{\rho}{\theta}$. We nevertheless advise caution when using our proposed, or other HMM methods (further advice and recommendations are found in [132]), for the inference of demography, as some assumptions may still be violated. For example, we assume that mutations occur in the seed/egg-bank (a consequence of DNA damage) at the same rate as in the active population. While there is support in plants for this hypothesis [177, 24], we do not know supporting data in *Daphnia*. Note finally, that all results rely on the quality of the sequences used and of the reference genome assembly.

As the conclusion of the chapter, the presented method allows the joint estimation of life-history traits and past demographic history based on full genome data. It is specifically adapted to the many species presenting violations of the classic Wright-Fisher model and can be used to study the evolution of seed/egg-banking as an important bet-hedging strategy with large consequences on the rate of genome evolution [165].

Chapter 3

Limits and Convergence properties of the Sequentially Markovian Coalescent

3.1 Motivation

Recovering the demographic history of a population has become a central theme in evolutionary biology. The demographic history (the variation of effective population size over time) is linked to environmental and demographic changes that existing and/or extinct species have experienced (population expansion, colonization of new habitats, past bottlenecks) [72, 52, 35]. Current statistical tools to estimate the demographic history rely on genomic data [147] and these inferences are often linked to archaeological or climatic data, providing novel insights on the evolutionary history [59, 47, 51, 6, 69, 103, 99]. From these analyses, evidences for migration events have been uncovered [103, 15], as have genomic consequences of human activities on other species [37]. Linking demographic history to climate and environmental data greatly supports the field of conservation genetics [40, 42, 128]. Indeed, using such approaches can help ecologists in detecting effective population size decrease [178], and thus serve as a guide in maintaining or avoiding the erosion of genetic diversity in endangered populations, and potentially predicting the consequences of climate change on genetic diversity [44]. Besides, studying the demographic histories of different species in relation to one another can unveil latent biological or environmental evolutionary forces [75], unveiling

links and changes within entire ecosystems. With the increased accuracy of current methods, the availability of very large and diverse data sets, and the development of new theoretical frameworks, the demographic history has become information that is essential in the field of evolution [54, 36]. However, obtaining unbiased estimations/interpretations of the demographic history remains challenging [9, 22].

The most sophisticated methods to infer demographic history make use of whole genome polymorphism data. Among the state-of-the-art methods, are those based on the theory of the Sequentially Markovian Coalescent (SMC) developed by McVean and Cardin [115] after the work of Wiuf and Hein [180], corrected by Marjoram and Wall [111] and first applied to whole genome sequences by Li and Durbin [103], who introduced the now well-known, Pairwise Sequentially Markovian Coalescent (PSMC) method. PSMC allows demographic inference of populations with unprecedented accuracy, while requiring only one sequenced diploid individual. This method uses the distribution of SNPs along the genome between the two sequences to account for and infer recombination and demographic history of a given population, assuming neutrality and panmixia. Although PSMC was a breakthrough in demographic inference, it has limited power in inferring more recent events. To address this issue, PSMC has been extended to account for multiple sequences (*i.e.* more than two) into the method known as the Multiple Sequentially Markovian Coalescent (MSMC) [146]. By using more sequences, MSMC better infers recent events and also provides the possibility of inferring population splits using the cross-coalescent rate. MSMC, unlike PSMC, is not based on SMC theory [115] but on SMC' theory [111], therefore MSMC applied to only two sequences has been defined as PSMC'. Methods developed after MSMC followed suit, with MSMC2 [46] extending PSMC by incorporating pairwise analysis, increasing efficiency, and the number of sequences that can be inputted (up to a hundred), resulting in more accurate results. SMC++ [168] brings the SMC theory to another level by allowing the use of hundreds of unphased sequences (MSMC requires phased input data) and breaking the piece-wise constant population size hypothesis while accounting for the sample frequency spectrum (SFS). Because SMC++ incorporates the SFS in the estimation of demographic history, it increases accuracy in recent times [168]. SMC++ is currently the state of the art SMC-based method for big data sets (>20 sequences) but seems to be outperformed by PSMC when using smaller data sets [53]. In a similar vein,

the Ascertained Sequentially Markovian Coalescent (ASMC) [131] extends the SMC theory to estimate coalescence times at the locus scale from ascertained SNP array data, something that was made possible by the theory presented by Hobolth and Jensen [81].

More recently, the second generation of SMC-based methods has been developed. New features have been added to the initial SMC theory, extending its application beyond simply inferring past demography [6, 151, 176]. The development of C-PSMC [75] allows the interpretation of estimated demographic history in the light of coevolution between species, making the first link between demographic history estimated by PSMC and evolutionary forces (although biological interpretation remains limited). iSMC [6] extends the PSMC theory to account and infer the variation of the recombination rate along sequences, unlocking recombination map estimations. An impressive advancement is the development of MSMC-IM, which to some extent solves the population structure problem, allowing the accurate and simultaneous inference of the demographic history and population admixture [176]. eSMC [151] incorporates common biological traits (such as self-fertilization and dormancy) and demonstrated the strong effect life-history traits can have on demographic history estimations. Results that could not be explained under the initial SMC hypotheses can now be explained by the potential presence of measurable phenomena.

New methods have been developed since PSMC, that have been either strongly inspired by the SMC [152, 161] or that are completely dissociated from it [156, 8, 140, 89, 109, 87, 155, 175]. Though there are alternative approaches, methods based on the SMC are still considered state of the art and remain widely used [112, 9, 157], notably in human evolution studies [157, 53]. However, each described method has its specificity, being based on different hypotheses to solve a particular problem or shortcomings of existing methodology. Although all these methods allow a new and different interpretation of genomic data, none of these methods guarantees unbiased inference, and their limitations have underlined how crucial and challenging demographic inference is, highlighting the complementarity and usefulness of applying several inference methods on a given dataset.

SMC-based methods display very good fits when using simulated data, especially when using simple single population models based on typical hu-

man data parameters [168, 146, 151, 176]. However, the SMC makes a large number of hypotheses [103, 146] that are often violated in data obtained from natural populations. When inputting data from natural populations, extracting information or correctly interpreting the results can become troublesome [22, 169, 9] and several studies address the consequences of hypothesis violation [74, 22, 140, 114, 148]. They bring to light how strongly population structure or introgression influence demographic history estimation if not correctly accounted for [74, 22]. Furthermore, some SMC-based methods require phased data (such as MSMC [146] and MSMC-IM [176]), and phasing errors can lead to a strong overestimation of population size in recent time [168]. The effect of sequencing coverage has also been tested in Nadachowska et al. [121], showing the importance of high coverage in order to obtain trustworthy results, and yet, SMC methods seem robust to genome quality [53]. Selection, if not accounted for, can result in a bottleneck signature [148], and there is currently no solution to this issue within the SMC theory, though it could be addressed using different theoretical frameworks that are being developed [153, 122]. More problematic, is the ratio of effective recombination over effective mutation rates $\frac{\rho}{\theta}$, which, if it is greater than one, biases estimations [168, 6, 151]. It is also important to keep in mind that there can be deviations between $\frac{\rho}{\theta}$ and the ratio of recombination rate over mutation rate measured experimentally ($\frac{r}{\mu}$), as the former can be greatly influenced by life-history, such as in organisms displaying self-fertilization, parthenogenesis or dormancy, and this can lead to issues when interpreting results (*e.g.* [151]). It is thus necessary to keep in mind that the accuracy of SMC-based methods depends on which of the many underlying hypotheses are prone to be violated by the data sets being used.

In an attempt to complement previous works, we here study the limits and convergence properties of methods based on the Sequentially Markovian Coalescent. We first define the limits of SMC-based methods (*i.e.* how well they perform theoretically), which we will call the best-case convergence. To do this, we use a similar approach to [71, 130, 87], and compare simulation results obtained with the simulated Ancestral Recombination Graph (ARG) as input to results obtained from sequences simulated under the same ARG, to study the convergence properties linked to data sets in the absence of hypothesis violation. We test several scenarios to check whether there are instances, where even without violating the underlying hypotheses of the methodology, the demographic scenarios cannot be retrieved because of the-

oretical limits (and not issues linked with data). We also study the effect of the optimization function (or composite likelihood) and the time window of the analysis on the estimations of different variables. Lastly, we test the effect of commonly violated hypotheses, such as the effect of the variation of recombination and mutation rates along the sequence and between scaffolds, errors in SNP calls, and the presence of transposable elements and link abnormal results to specific hypothesis violations. Through this work, we aim to provide guidelines concerning the interpretation of results when applying this methodology on data sets that may violate the underlying hypotheses of the SMC framework.

3.2 Materials and Methods

In this study we use four different SMC-based methods: MSMC, MSMC2, SMC++ and eSMC. All methods are Hidden Markov Models and use whole genome sequence polymorphism data. The hidden states of these methods are the coalescence times (or genealogies) of the sample. To have a finite number of hidden states, they are grouped into x bins (x being the number of hidden states). The reasons for our model choices are as follows: *i*) MSMC, unlike any other method, focuses on the first coalescence event of a sample of size n, and thus exhibits different convergence properties [146], *ii*) MSMC2 computes coalescence times of all pairwise analysis from a sample of size n, and can deal with a large range of data sets [156], *iii*) SMC++ [168] is the most advanced and efficient SMC method and lastly, *iv*) eSMC [151] is a re-implementation of PSMC' (similar to MSMC2), which will contribute to highlighting the importance of algorithmic translations as it is very flexible in its use and outputs intermediate results necessary for this study.

3.2.1 SMC methods

PSMC', MSMC2 and eSMC

PSMC' and methods that stem from it (such as MSMC2 [46] and eSMC [151]) focus on the coalescence events between only two individuals (or sequences in practice), and, as a result, do not require phased data. The algorithm goes along the sequence and estimates the coalescence time at each position. To do this, it checks whether the two sequences are similar or different at each

position. The presence or absence of a segregating site along the sequence (determined by the population mutation rate θ) is used to infer the hidden state (*i.e.* coalescence time). However, the hidden state is only allowed to change in the event of a recombination, which leads to a break in the current genealogy. Thus, the population recombination rate ρ constrains the inferred changes of hidden states along the sequence (for a detailed description of the algorithm see [146, 176, 151]).

MSMC

MSMC is mathematically and conceptually very similar to the PSMC' method. Unlike other SMC methods, it simultaneously analyses multiple sequences and because of this, MSMC requires the data to be phased. In combination with a second HMM, to estimate the external branch length of the genealogy, it can follow the distribution of the first coalescence event in the sample along the sequences. However, due to computational load, MSMC cannot analyze more than 10 sequences simultaneously (for a detailed description see [146]).

SMC++

SMC++ is slightly more complex than MSMC or PSMC. Though it is conceptually very similar to PSMC', mathematically it is quite different. SMC++ has a different emission matrix compared to previous methods because it calculates the sample frequency spectrum of sample size $n + 2$, conditioned on the coalescence time of two "distinguished" haploids and n "undistinguished" haploids. In addition SMC++ offers features such as a cubic spline to estimate demographic history (*i.e.* not a piece-wise constant population size). The SMC++ algorithm is fully described in [168].

Best-case convergence

Using sequence simulators such as msprime [91] or scrm [159], one can simulate the Ancestral Recombination Graph (ARG) of a sample. Usually, the ARG is given through a sequence of genealogies (*e.g.* a sequence of trees in Newick format). From this ARG, one can find what state of the HMM the sample is in at each position. Hence, one can build the series of states along the genomes, and build the transition matrix. The transition matrix is a square matrix of dimension x (where x is the number of hidden states)

counting all the possible pairwise transitions between the x states (including from a given state to itself). Using the transition matrix built directly from the exact ARG, one can estimate parameters using eSMC or MSMC as if they could correctly infer the hidden states. Hence, estimations using the exact transition matrix represents the upper bound of performance for these methods. We choose to call this upper bound the best-case convergence (since it can never be reached in practice). For this study's purpose, a second version of the R package eSMC [151] was developed. This package enables the building of the transition matrix (for eSMC or MSMC), and can then be used to infer the demographic history. The package is mathematically identical to the previous version but includes extra functions, features, and new outputs necessary for this study. The package and its description can be found at https://github.com/TPPSellinger/eSMC2.

Baum-Welch algorithm

SMC-based methods can use different optimization functions to infer the demographic parameters (*i.e.* likelihood or composite likelihood). The four studied methods use the Baum-Welch algorithm to maximize the likelihood. MSMC2 and SMC++ implement the original Baum-Welch algorithm (which we call the complete Baum-Welch algorithm), whereas eSMC and MSMC compute the expected composite likelihood $Q(\theta|\theta^t)$ based only on the transition matrix (which we call the incomplete Baum-Welch algorithm). The use of the complete Baum-Welch algorithm or the incomplete one can be specified in the eSMC package. The composite likelihood for SMC++ and MSMC2 is given by equations 1 and the composite likelihood for eSMC and MSMC by equation 2:

$$\begin{aligned} Q(\Theta|\Theta^t) = \nu_{\Theta^t} log(P(X_1|\Theta)) + \sum_{X,Y} E(X,Z|\Theta^t) log(P(X|Z,\Theta)) \\ + \sum_{X,Y} E(Y,X|\Theta^t) log(P(Y|X,\Theta)) \end{aligned} \tag{3.1}$$

and :

$$Q(\Theta|\Theta^t) = \sum_{X,Y} E(X,Z|\Theta^t) log(P(X|Z,\Theta)), \tag{3.2}$$

with:

- ν_Θ : The equilibrium probability conditional to the set of parameters Θ.
- $P(X_1|\Theta)$: The probability of the first hidden state conditional to the set of parameters Θ.
- $E(X, Z|\Theta^t)$: The expected number of transitions of X from Z conditional to the observation and set of parameters Θ^t.
- $P(X|Z, \Theta)$: The transition probability from state Z to state X, conditional to the set of parameters Θ.
- $E(Y, X|\Theta^t)$ The expected number of observations of type Y that occurred during state X conditional to observation and set of parameters Θ^t.
- $P(Y|X, \Theta)$: The emission probability conditional to the set of parameters Θ.

Time window

Each tested SMC-based method has its own specific time window for which estimations are made. Note that hidden states are defined as discretized intervals, as consequences of which the boundaries, length, and number of states of the time window do implicitly affect inferences. For example, the original PSMC has a time window wider than PSMC', so that estimations cannot be compared one to one. To measure the effect of choosing different time window parameters, we analyze the same data with four different settings. The first time window is the one used for PSMC' defined in [146]. The second time window is that of MSMC2 [176] (similar to the one of the original PSMC [103]), which we call "big" since it goes further in the past and in more recent time than that of PSMC'. We then define a time window equivalent to the first one (i.e. PSMC') shifted by a factor of five in the past (first time window, *i.e.* hidden states, multiplied by five). The last one is a time window equivalent to the first one (i.e. PSMC') shifted by a factor of five in recent times (i.e. first time window divided by five).

3.2.2 Simulated sequence data

Throughout this chapter, we simulate different demographic scenarios using either the coalescence simulation program scrm [159] or msprime [91]. We use scrm for the best-case convergence as it can output the genealogies in a Newick format (which we use as input). We use scrm, which outputs simulated sequences in the ms format, to simulate data for eSMC, MSMC, MSMC2. We use msprime to simulate data for SMC++ since msprime is more efficient than scrm for big sample sizes [91] and can directly output .vcf files (which is the input format of SMC++).

Absence of hypothesis violation

We simulate five different demographic scenarios: sawtooth (successions of population size exponential expansion and decrease), bottleneck, exponential expansion, exponential decrease, and constant population size. Each of the scenarios with varying population size is tested under four amplitude parameters (*i.e.* by how many folds the population size varies: 2, 5, 10, 50). We infer the best-case convergence under four different sequence lengths (10^7, 10^8, 10^9, and 10^{10} bp) and choose the per site mutation and recombination rates recommended for humans in MSMC's manual, respectively 1.25×10^{-8} and 1×10^{-8}. When analyzing simulated sequence data, we simulate sequences of 100 Mb: two sequences for eSMC and MSMC2, four sequences for MSMC, and twenty sequences for SMC++.

Calculation of the mean square error (MSE)

To measure the accuracy of inferences we calculate the Mean Square Error (MSE). We first divide the time window (in log10 scale) of each analysis into ten thousand points. We then calculate the MSE by comparing the actual population size and the one estimated by the method at each of the ten thousand points. We thus have the following formula:

$$MSE = \frac{\sum_{i=1}^{10^4}(y_i - y_i^*)^2}{10^4} \tag{3.3}$$

Where:

- y_i is the population size at the time point i.
- y_i^* is the estimated population size at the time point i.

All the command lines to simulate data can be found in S1 of the Appendix.

Presence of hypothesis violation

SNP calling: In practice, SNP calling from next generation sequencing can yield different numbers and frequencies of SNPs depending on the chosen parameters for the different steps of analysis (read trimming, quality check, read mapping, and SNP calling) as well as the quality of the reference genome, data coverage and depth of sequencing, species ploidy [135]. Therefore, based on raw sequence data, the stringency of filters can lead to excluding SNPs (false negatives) or including spurious ones (false positives). When dealing with complex genomes or ancient

DNA [154, 20], SNPs can be simultaneously missed and added. We thus simulate four sequences of 100 Mb under a "sawtooth" scenario and then a certain percentage (5,10 and 25 %) of SNPs is randomly added to and/or deleted from the simulated sequences. We then analyze the variation and bias in SNP calling on the accuracy of demographic parameter estimations. As an additional analysis, we test the effect of ascertainment bias on inferences (a prominent issue in microarray SNP studies) by simulating 100 sequences with msprime where only SNPs above a certain (Minor Allele Frequency) MAF threshold (1%,5%, and 10%) are kept, then run SMC methods on a subset of the obtained data.

Changes in mutation and recombination rates along the sequence: Because the recombination rate and the mutation rate can change along the sequence [6], and chromosomes are not always fully assembled in the reference genome (which consists of possibly many scaffolds), we simulate short sequences where the recombination and/or mutation rate randomly change between the different scaffolds around an average value of $1.25 \times 10-8$ per generation per base pair (between $2.5 \times 10-9$and $6.25 \times 10-8$). We simulate 20 scaffolds of size 2 Mb, as this seems representative of the best available assembly for non-model organisms [108, 160]. We then analyze the simulated sequences to study the effect of assuming scaffolds share the same mutation and recombination rates. In addition, we simulate sequences of 40 Mb (assuming genomes are fully assembled) where the recombination rate along the sequence randomly changes every 2 Mbp (up to five-fold) around an average value of $1.25 \times 10-8$ (the mutation rate being fixed at $1.25 \times 10-8$ per generation per bp) to study the effect of the assumption of a constant recombination rate along the sequence.

Transposable elements (TEs): Genomes can contain transposable elements whose dynamics violate the classic infinite site mutational model for SNPs and thus potentially affect the estimation of different parameters. Although methods have been developed to detect [123] and simulate them [96], understanding how their presence/absence influences demographic inferences remains unclear. TEs are usually masked when detected in the reference genome and thus not taken into account in the mapped individuals due to the redundancy of read mapping for TEs. Due to their repetitive nature, it can be difficult to correctly detect and assemble them if using short reads, as well as to assess the presence/absence polymorphism of individuals of a population [64]. In addition, the quality and completeness of the reference genome (*e.g.* using the reference genome of a sister species as the reference genome) can strongly affect the accuracy of detecting, assembling, and masking TEs [137]. To best capture and mimic the effect of TEs unaccounted for in the data, we altered four simulated sequences of length 20 Mb in four different

ways. The first way to simulate the effect of unmapped and unaccounted TEs is to assume they exhibit presence/absence polymorphism, hence creating gaps in the sequence. For each individual, we remove small pieces of the sequence of different lengths (1kb, 10 kb or 100kb), so that up to a certain percentage (5,10,25,50%) of the original simulated sequence is removed, so as to shorten and fragment the whole sequence to be analyzed. The second way is to consider unmasked TEs, done by randomly selecting small pieces of the original simulated sequence (1kb, 10 kb or 100kb), making up to a certain percentage of it (5,10,25,50%), and removing all the SNPs found in those regions (*i.e.* removing mutations from TEs). The removed SNPs are hence structured in many small regions along the genome. Thirdly, we test the consequences of simultaneously having both removed and unmasked TEs in the data set. Lastly, to measure the importance of detecting and masking TEs, we assume all TEs to be present and masked when building the multihetsep file (*i.e.* considering TEs as missing data).

3.3 Results

3.3.1 Best-case convergence

Results of the best-case convergence of eSMC under the sawtooth demographic history are displayed in Figure 1. Increasing the sequence length increases accuracy and reduces variability, leading to better convergence and reducing the mean square error (see Figures 3.1 a-c and Supplementary Table A.1). However, when the amplitude of population size variation is too great (here for 50 fold), the demographic history cannot be retrieved, even when using very large data sets (see Figure 1d). In Supplementary Figure A.12, we show that even when changing the number of hidden states (*i.e.* number of inferred parameters), some scenarios with a very strong variation of population size remain badly inferred.

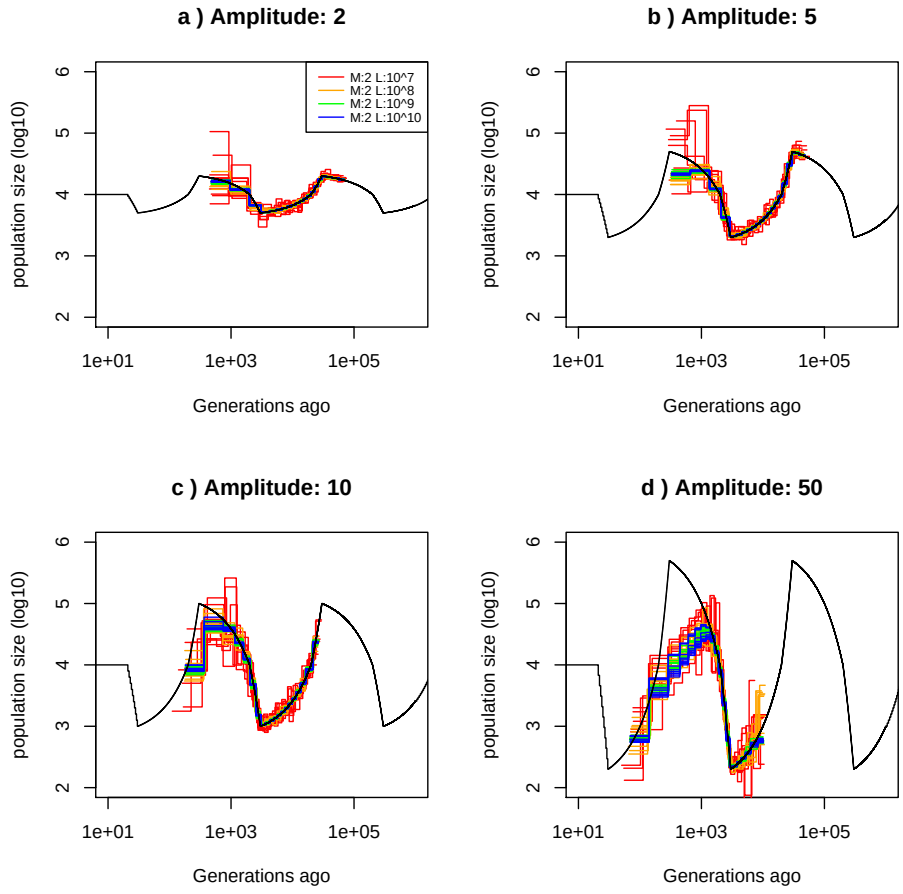

Figure 3.1: **Best-case convergence of eSMC.** Estimated demographic history using simulated genealogy over sequences of 10,100,1000,10000 Mb (respectively in red,orange, green and blue) under a sawtooth scenario (original scenario in black) with 10 replicates for different amplitudes of size change: a) 2-fold, b) 5-fold, c) 10-fold, and d) 50-fold. The recombination rate is set to 1×10^{-8} per generation per bp and the mutation rate to 1.25×10^{-8} per generation per bp.

In Supplementary Figures A.13, we show the best-case convergence of MSMC with four genome sequences and generally find that these analyses present a higher variance than eSMC. However, MSMC shows better fits in recent times and is better able to retrieve population size variation than eSMC (see Supplementary Figure A.13 d). Scenarios with a strong variation of population size (*i.e.* with large amplitudes) still pose a problem (see Supplementary Figure A.14), and no

matter the number of estimated parameters, such scenarios cannot be correctly inferred using MSMC.

To better understand these results, we examine the coefficient of variation calculated from the replicates at each entry of the transition matrix. We can see that increasing the sequence length reduces the coefficient of variation (the ratio of the standard deviation to the mean, hence indicating convergence when equal to 0, see Supplementary Figure A.15). Yet increasing the amplitude of population size variation decreases the number of some hidden state transitions leading to unobserved transitions. Unobserved transitions result from the reduced probability of coalescence events in specific time intervals (*i.e.* hidden states). In these cases, matrices display higher coefficients of variation and can be partially empty (Figure 3.2). This explains the increase of variability of the inferred scenarios, as well as the incapacity of SMC methods to correctly infer the demographic history with strong population size variation in specific time intervals independently of the amount of data available.

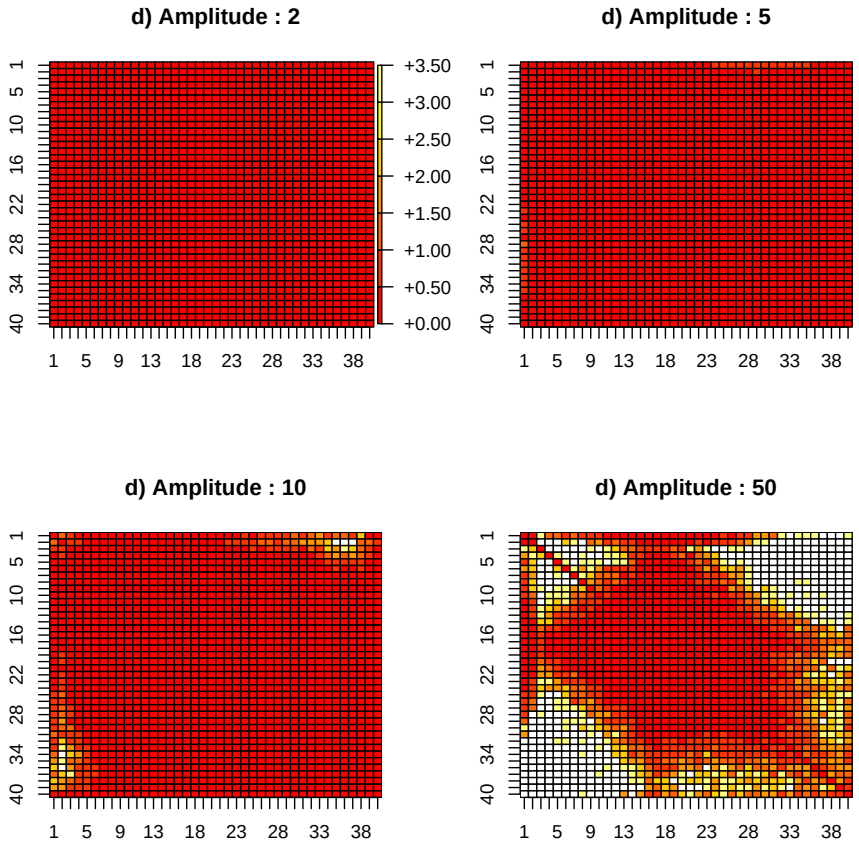

Figure 3.2: **Estimated transition matrix in sharp sawtooth scenario.** Estimated coefficient of variation of the transition matrix using simulated genealogy over sequences of 10000 Mb under a sawtooth scenario of amplitude 2, 5,10 and 50 (respectively in a, b, c and d) each with 10 replicates. Recombination and mutation rates are as in Figure 1. White squares indicate absence of observed transitions (*i.e. no data*).

3.3.2 Simulated sequence results

Scenario effect

In the previous section, we explored the theoretical performance limitations of eSMC and MSMC using trees in Newick format as input. In this section, we evaluate how these methods perform when inputting simulated sequence data using

the same recombination and mutation rates. We first investigate the effect of amplitude of population size variation as in Figure 3.1. Results for the sawtooth scenario are displayed in Figure 3.3, where the different models display a good fit but are not as good as expected from the best-case convergence given the same amount of data (Figure 3.1 (orange line) and Supplementary Table A.1 vs Figure 3.3 (red line) and Supplementary Table A.2). As predicted by Figures 1 and 2, the case with the greatest amplitude of population size variation (Figure 3.1d) is the least well fitted (see Supplementary Table A.2 for the MSE). To study the origin of differences between simulation results and theoretical results, we measure the difference between the transition matrix estimated by eSMC and the one built from the actual genealogy. Results show that hidden states are harder to correctly infer in scenarios with strong population size variation, explaining the high variance (see Supplementary Figure A.16). We demonstrate there that for the same amount of data, the simulation, and thus by extension the real data, shows additional stochastic behavior than the best-case convergence (Figure 3.1).

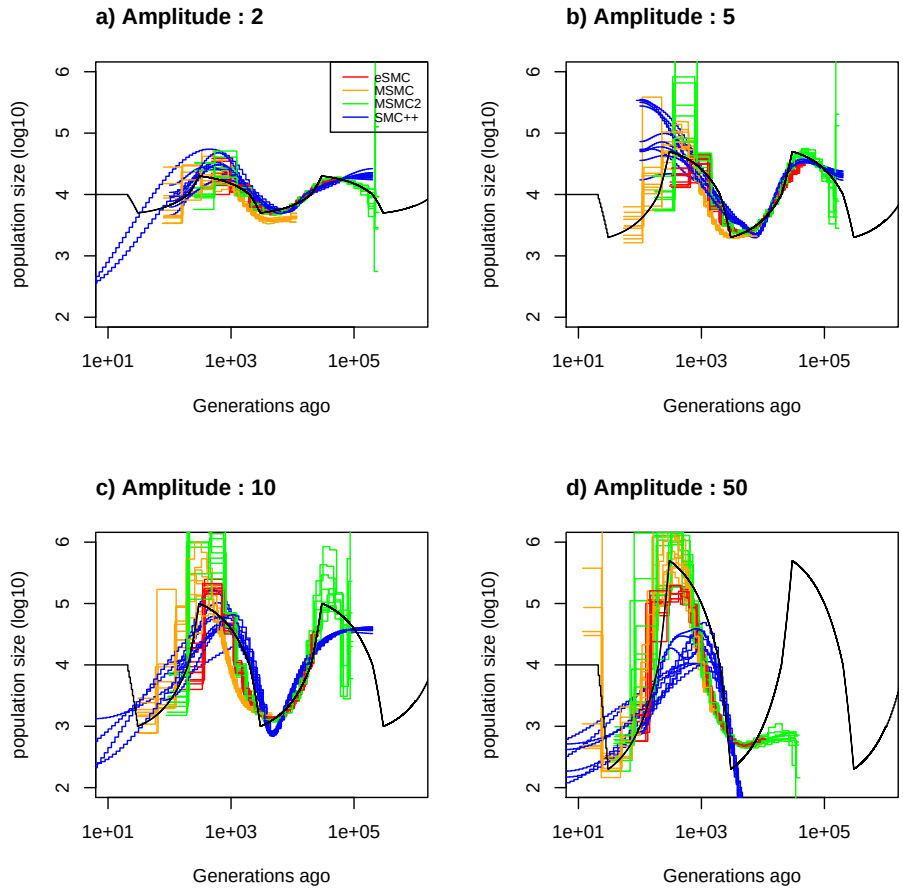

Figure 3.3: **Estimated demography using simulated sequences as input.** Estimated demographic history (black) under a sawtooth scenario with 10 replicates using simulated sequences for different amplitudes of population size change: a) 2, b) 5, c) 10 and d) 50. Two sequences of 100 Mb for eSMC and MSMC2 (respectively in red and green), four sequences of 100 Mb for MSMC (orange) and 20 sequences of 100 Mb for SMC++ (blue) were simulated. Recombination and mutation rates are respectively set to $1 \times 10^{-}8$ and $1.25 \times 10^{-}8$.

Increasing the time window results in an increased variance of the inferences (Supplementary Figure A.17). In addition, shifting the window towards more recent time leads to poor demographic estimations, but shifting it further in the past does not seem to bias it (there are however consequences on estimations of the recombination rates, see Table 1 for more details). Concerning the optimization

function, we find that the complete Baum-Welch algorithm gives similar results to the incomplete one (Table 3.1).

Optimization function	Scenario	real $\frac{\rho}{\theta}$	normal window $\frac{\rho^*}{\theta}$	Big Window $\frac{\rho^*}{\theta}$	Old window $\frac{\rho^*}{\theta}$	Recent window $\frac{\rho^*}{\theta}$
Incomplete Baum-Welch	sawtooth	0.8	0.79 (0.036)	0.72 (0.039)	0.72 (0.042)	0.94 (0.005)
Complete Baum-Welch	sawtooth	0.8	.79 (0.044)	0.72 (0.039)	0.72 (0.042)	1.56 (0.087)
Incomplete Baum-Welch	Constant	0.8	0.86 (0.019)	0.85 (0.020)	0.84 (0.019)	0.98 (0.002)
Complete Baum-Welch	Constant	0.8	0.86 (0.019)	0.85 (0.020)	0.84 (0.019)	1.06 (0.02)

Table 3.1: Average estimated values for the recombination over mutation ratio $\frac{\rho}{\theta}$ over ten repetitions for different sizes of the time window. The coefficient of variation is indicated in brackets. Four sequences of 50 Mb were simulated with a recombination rate set to 1×10^{-8} per generation per bp and a mutation rate to 1.25×10^{-8} per generation per bp.

Effect of the ratio of the recombination over the mutation rate

The ratio of the effective recombination over effective mutation rates ($\frac{\rho}{\theta}$) can influence the ability of SMC-based methods to retrieve the coalescence time between two points along the genome [168]. Intuitively, if recombination occurs at a higher rate compared to mutation, then it renders it more difficult to detect any recombination events that may have taken place before the introduction of a new mutation, and thus bias the estimation of the coalescence time [151, 168]. Under the bottleneck scenario, we find that the lower $\frac{\rho}{\theta}$, the better the fit of the inferred demography by eSMC and SMC++ in the past, but also the higher the variance of the inferences (see Figure 3.4). However, each method displays the worse fit when $\frac{\rho}{\theta} = 10$ (Supplementary Table A.3). SMC++ seems slightly less sensitive to $\frac{\rho}{\theta}$ than other methods.

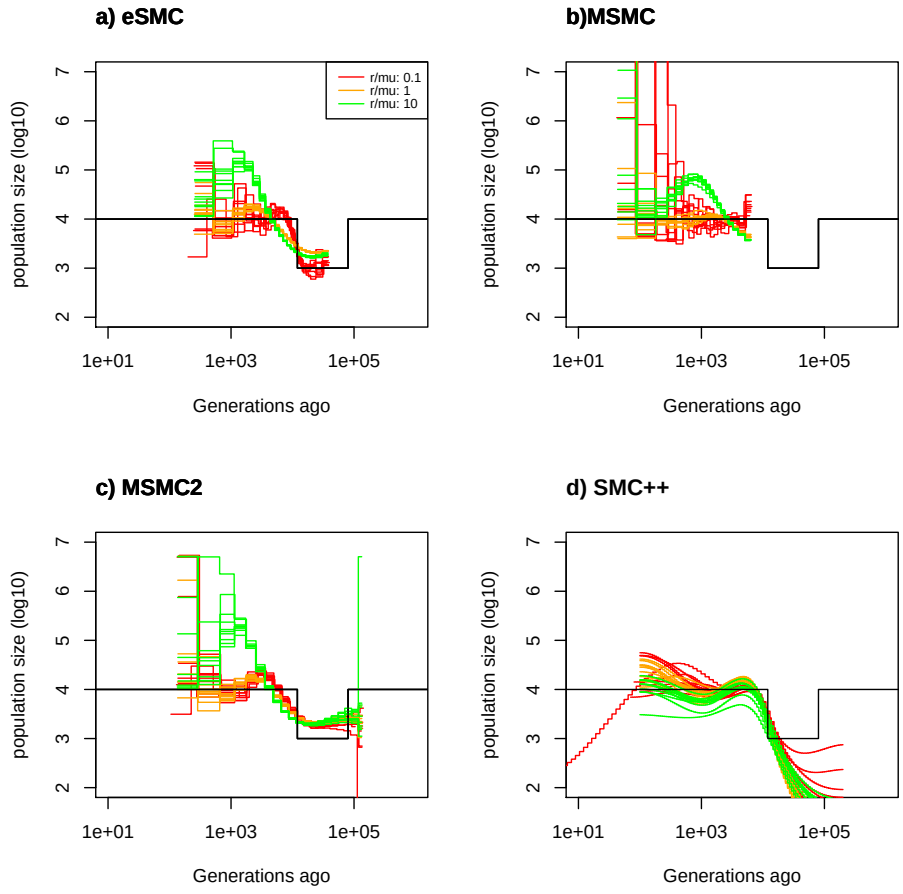

Figure 3.4: **Effect of $\frac{\rho}{\theta}$ on inference of demographic history.** Estimated demographic history under a bottleneck scenario with 10 replicates using simulated sequences. We simulate two sequences of 100 Mb for eSMC and MSMC2 (respectively in a and b), four sequences of 100 Mb for MSMC (c) and twenty sequences of 100 Mb for SMC++ (d). The mutation rate is set to 1.25×10^{-8} per generation per bp and the recombination rates are 1.25×10^{-9},1.25×10^{-8} and 1.25×10^{-7} per generation per bp, giving $\frac{\rho}{\theta} = 0.1$, 1 and 2 and the inferred demographies are in red, orange and green respectively. The demographic history is simulated under a bottleneck scenario of amplitude 10 and is represented in black.

It is, in some instances, possible to compensate for a $\frac{\rho}{\theta}$ ratio that is not ideal by increasing the number of iterations. Indeed, for eSMC, the demographic history is better inferred (Supplementary Figure A.18), although the correct recombina-

tion rate cannot be retrieved (Table 2). MSMC is able to better infer the correct recombination rate than other methods despite $\frac{\rho}{\theta} > 1$ but poorly estimates the demographic history. The past demographic inferences obtained using MSMC2 and SMC++ are not improved when increasing the number of iterations (see Supplementary Figure A.18 and Table 3.2).

method	real $\frac{\rho}{\theta}$	set 1 , $\frac{\rho^*}{\theta}$	set 2 , $\frac{\rho^*}{\theta}$	set 3 , $\frac{\rho^*}{\theta}$	set 4 , $\frac{\rho^*}{\theta}$	set 5 , $\frac{\rho^*}{\theta}$
eSMC	10	1.35 (0.026)	1.76 (0.047)	1.29 (0.027)	1.74 (0.048)	1.80 (0.041)
MSMC	10	2.70 (0.011)	6.58 (0.031)	2.68 (0.011)	6.57 (0.032)	6.62 (0.030)
MSMC2	10	1.27 (0.055)	1.65 (0.13)	1.26 (0.060)	1.75 (0.060)	1.60 (0.29)
SMC++	10	0.56 (0.38)	0.48 (0.38)	1.32 (0.15)	0.21 (0.62)	0.98 (0.24)

Table 3.2: Average estimated values for the recombination over mutation ratio $\frac{\rho}{\theta}$ over ten repetitions. The coefficient of variation is indicated in brackets. For eSMC, MSMC and MSMC2 we have: set 1: 20 hidden states; set 2: 200 iterations; set3: 60 hidden states; set 4: 60 hidden states and 200 iterations and set 5: 20 hidden states and 200 iterations. For SMC++: set 1: 16 knots; set 2: 200 iterations; set 3: 4 knots in green; set 4: regularization penalty set to 3 and set 5: regularization-penalty set to 12.

3.3.3 Simulation results under hypothesis violation

Imperfect SNP calling

We analyze simulated sequences that have been modified by removing and/or adding SNPs using the different SMC methods. We find that, when using MSMC2, eSMC and MSMC, having more than 10% of spurious SNPs (*e.g.* no quality filtering) can lead to a strong overestimation of population size in recent time but that missing SNPs have no effects on inferences in the far past and only mild effects on inferences of recent time for MSMC2 (Figure 5). The mean square error is displayed in Supplementary Table A.4.

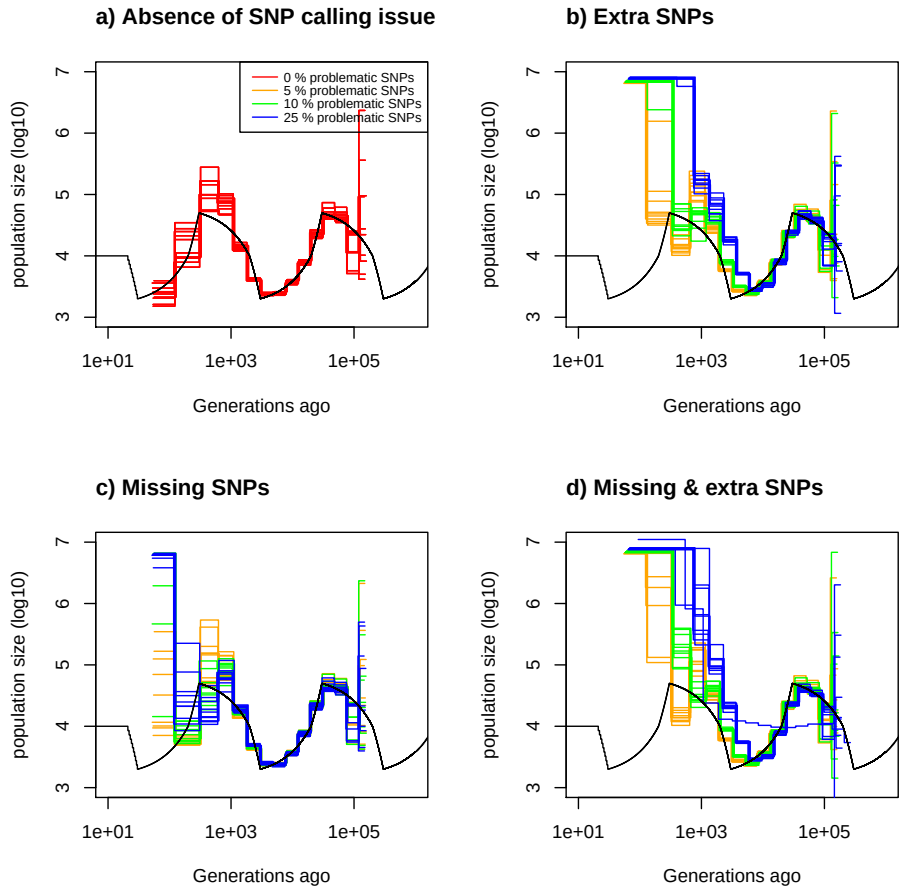

Figure 3.5: **Consequences of SNP calling errors.** Estimated demographic history using MSMC2 under a sawtooth scenario with 10 replicates using four simulated sequences of 100 Mb. Recombination and mutation rates are as in Figure 1 and the simulated demographic history is represented in black. a) Demographic history simulated with ibsence of SNP calling issue (red). b) Demographic history simulated with 5% (orange),10% (green) and 25% (blue) missing SNPs. c) Demographic history simulated with 5% (orange), 10% (green) and 25% (blue) additional SNPs. d) Demographic history simulated with 5% (orange),10% (green) and 25% (blue) of additional and missing SNPs .

Specific scaffold parameters

We simulate sequence data where scaffolds have either been simulated with the same recombination and mutation rates or with different recombination and mutation rates. Data sets are then analyzed assuming scaffolds share or do not share the same recombination and mutation rates. We can see in Figure 3.6 (and Supplementary Table A.5) that when scaffolds all share the same parameter values, estimated demography is accurate both when the analysis assumed shared or differing mutation and recombination rates. However, when scaffolds are simulated with different parameter values, analyzing them under the assumption that they have the same mutation and recombination rates leads to poor estimations. Assuming scaffolds do not share recombination and mutation rates does improve the results somewhat, but the estimations remain less accurate than when scaffolds all share with the same parameter values. If only the recombination rate changes from one scaffold to another, the demographic history is only slightly biased, whereas, if the mutation rate changes from one scaffold to the other, demographic history is poorly estimated.

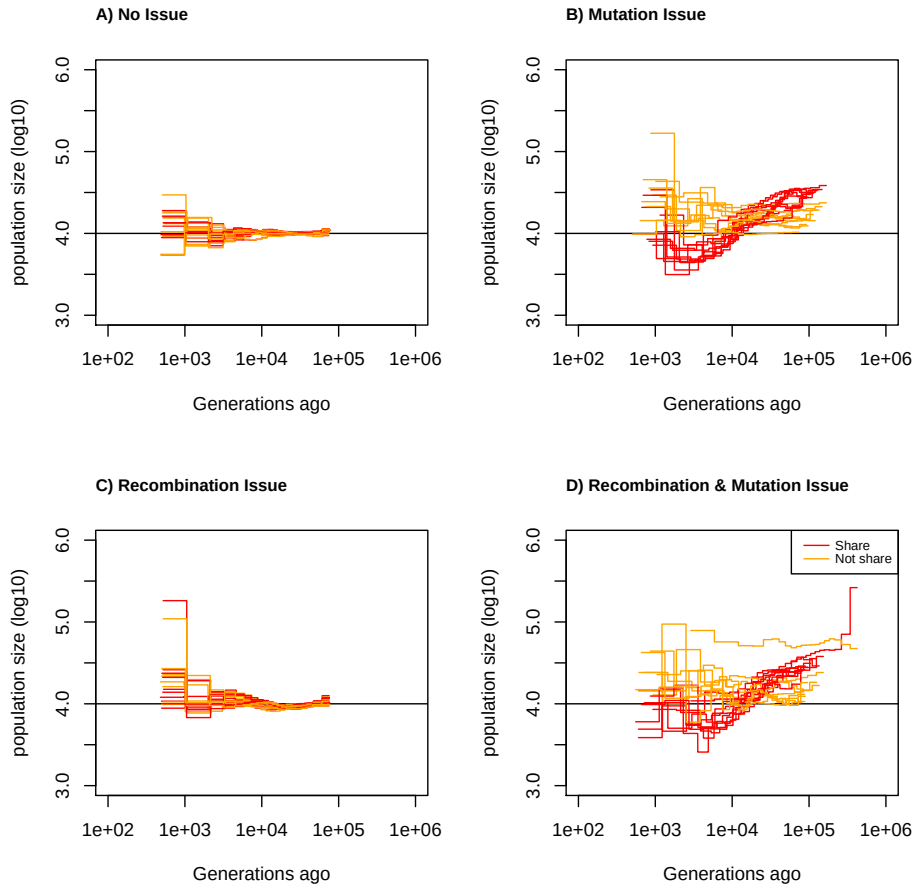

Figure 3.6: **Estimating demographic history using scaffolds sharing or differing in mutation and recombination rates.** Estimated demographic history using eSMC under a sawtooth scenario with 10 replicates using twenty simulated scaffolds of two sequences of 2 Mb assuming scaffolds share (red) or do not share recombination and mutation rates (orange). The simulated demographic history is represented in black. a) Scaffolds share the same parameters, recombination and mutation rates are set at 1.25×10^{-8}, b) Each scaffold is randomly assigned a recombination rate between 2.5×10^{-9} and 6.25×10^{-8} and the mutation rate is 1.25×10^{-8}, c) Each scaffold is randomly assigned a mutation rate between 2.5×10^{-9} and 6.25×10^{-8} and the recombination rate is 1.25×10^{-8} and d) Each scaffold is assigned a random mutation and an independently random recombination rate, both being between 2.5×10^{-9} and 6.25×10^{-8}.

Even if chromosomes are fully assembled, assuming we here have one scaffold of 40 Mb (chromosome fully assembled), there may be variations of the recombination rate along the sequence, however, this seems of little consequence when applying eSMC. As can be seen in Supplementary Figure A.19, the demographic scenario is well inferred, despite an increase in variance and a smooth "wave" shaped demographic history when sequences simulated with varying recombination rates are compared to those with a fixed recombination rate throughout the genome. Overall we see that when the recombination rate is heterogeneous along the genome by a factor 5, it is not untypical to falsely estimate a two-fold variation of Ne even though the true Ne is constant in time.

How transposable elements bias inference

Transposable elements (TEs) are present in most species, and are (if detected) taken into account as missing data by SMC methods [146]). Depending on how TEs affect the data set, we find that methods are more or less sensitive to TEs. If TEs are unmapped/removed from the data set, there does not appear to be any bias in the estimated demographic history when using MSMC2 (see Figure 3.7), but there is an overestimation of $\frac{\rho}{\theta}$ (see Table 3). We find that the higher the proportion of sequences removed, the more $\frac{\rho}{\theta}$ is over-estimated. For a fixed amount of missing/removed data, the smaller the sequences that are removed, the more $\frac{\rho}{\theta}$ is over-estimated (Table 3.3). If TEs are present but unmasked in the data set (and thus not accounted for missing data by the model [146]), we find that this is equivalent to a faulty calling of SNPs, in which SNPs are missing, hence resulting in demographic history estimations by MSMC2 similar to those observed in Figure 3.5a. However, if the size of unmasked TEs increases, different results are obtained (see Supplementary Figures A.20 and A.20). Indeed, in recent times there is a strong underestimation of population size and the model fails to capture the correct demographic history. The longer the TEs are, the stronger the effect on the estimated demographic history. However, when TEs are detected and correctly masked, there is no effect on demographic inferences (Supplementary Figures A.22).

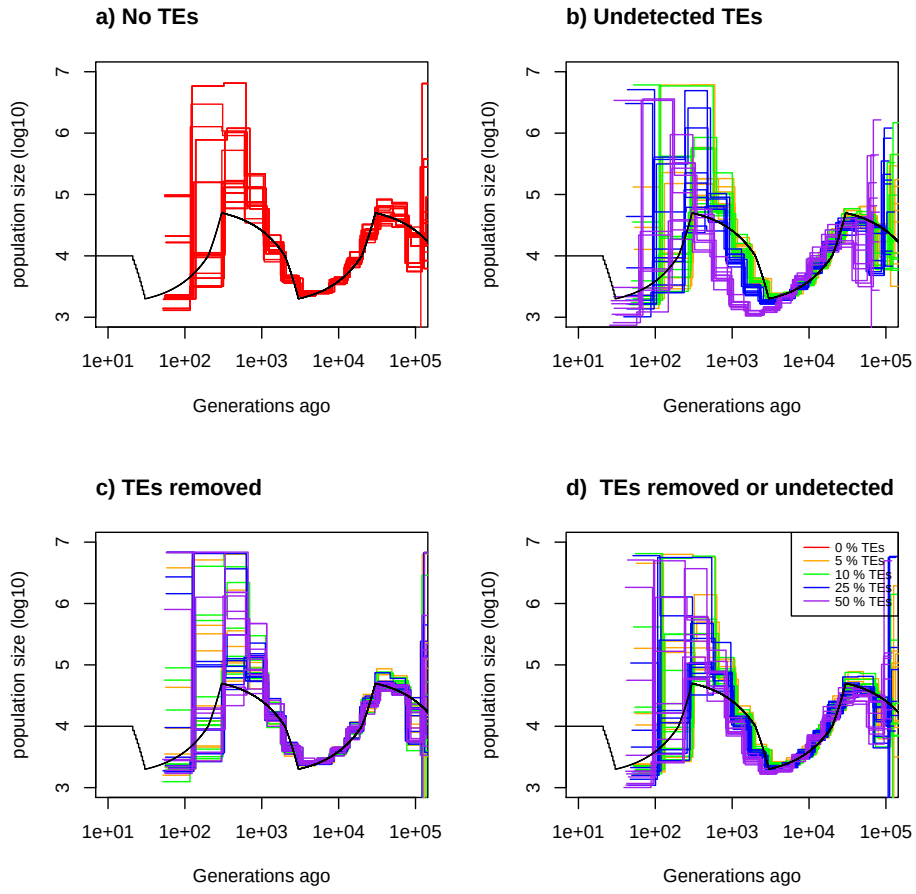

Figure 3.7: **Estimated demography of MSMC2 under a sawtooth scenario with transposable elements.** Estimated demographic history by MSMC2 under a sawtooth scenario with 10 replicates using simulated sequences. 4 sequences of 20 Mb. Recombination rate is set to 1.25×10^{-8} per generation per bp and mutation rate to 1.25×10^{-8} per generation per bp . The simulated demographic history is represented in black. Here tansposable element are of length 1kbp. a) Demographic history simulated with no transposable elements. b) Demographic history simulated where transposable element are removed. c) Demographic history simulated where SNPs on transposable are removed. d) Demographic history simulated where half of transposable are removed and SNPs on the other half are removed. Proportion of transpobable element of the genome is set to 0% (red), 5% (orange), 10% (green), 25 % (blue) and 50 % (purple).

TE length	method	real $\frac{\rho}{\theta}$	$\frac{\rho}{\theta}^*$ and 5% TEs	$\frac{\rho}{\theta}^*$ and 10% TEs	$\frac{\rho}{\theta}^*$ and 25% TEs	$\frac{\rho}{\theta}^*$ and 50% TEs
1 kb	MSMC2	1	0.87 (0.047)	0.88 (0.049)	1.0 (0.036)	1.35 (0.035)
10 kb	MSMC2	1	0.87 (0.064)	0.89 (0.067)	.99 (0.15)	1.13 (0.30)
100 kb	MSMC2	1	0.87 (0.056)	0.88 (0.050)	0.91 (0.079)	0.91 (0.073)

Table 3.3: Average estimated values for the recombination over mutation ratio $\frac{\rho}{\theta}$ by MSMC2 over ten repetitions. The coefficient of variation is indicated in brackets. TEs are of length 1kb, 10kb or 100 kb and are completely removed and the proportion of the genome made up by TEs is 5%,10% ,25% and 50%.

3.4 Discussion

Throughout this chapter, we have outlined the limits of PSMC' and MSMC methodologies, which had, until now, not been clearly defined. We find that, in most cases, if enough genealogies (*i.e.* data) are inputted then the demographic history is accurately estimated, tending to results obtained previously [71, 22], however, we find that the amount of data required for an accurate fit depends on the underlying demographic scenario. The differences with previous works stem from estimations being made using the actual series of coalescence times [71, 22], whereas we use the series of hidden states built from the discretization of time summarized in a simple matrix. We also find that some scenarios are better retrieved when using either MSMC or methods based on PSMC', indicating that there are complementary convergence properties between these methodologies.

We develop a method to indicate if the amount of data is enough to retrieve a specific scenario, notably by calculating the coefficient of variation of the transition matrix using either real or simulated data and therefore offer guidelines to build appropriate data sets (see also Supplementary Figure 8). Our approach can also be used to infer demographic history given an ARG (using trees in Newick format or sequences of coalescence events), independently of how the ARG has been estimated. Our results suggest that whole genome polymorphism data can be summarized in a transition matrix based on the SMC theory to estimate the demographic history of panmictic populations. As new methods can infer genealogies better and faster [156, 92, 116, 131], the estimated transition matrix could become a powerful summary statistic in the future. HMM can be a computational burden depending on the model and model parameters, and estimating genealogy through more efficient methods would still allow the use of SMC theory for parameter estimation or hypothesis testing (as in [56, 71, 87]). In addition, using the work of [176], one could (to some extent [94]) extend our approach to account for

population structure and migration.

We have also demonstrated that the power of PSMC', MSMC, and other SMC-based methods, rely on their ability to correctly infer the genealogies along the sequence (*i.e.* the Ancestral Recombination Graph or ARG). The accuracy of ARG inference by SMC methods, however, depends on the ratio of the recombination over the mutation rate ($\frac{\rho}{\theta}$). As this rate increases, estimations lose accuracy. Specifically, increasing $\frac{\rho}{\theta}$ leads to an over-estimation of transitions on the diagonal, which explains the underestimation of the recombination rate and inaccurate demographic history estimations, as shown in [168, 151]. As a way around this issue, in some cases, it is possible to obtain better results by increasing the number of iterations. MSMC's demographic inference is more sensitive to $\frac{\rho}{\theta}$ but the quality of the estimation of the ratio itself is less affected. This once again shows the complementarity of PSMC' and MSMC. If the variable of interest is $\frac{\rho}{\theta}$, then MSMC should be used, but if the demographic history is of greater importance, PSMC'-based methods should be used. The amplitude of population size variation also influences the estimation of hidden states along the sequences, with high amplitudes leading to a poor estimation of the transition matrix, distorting the inferred demography. We find that increasing the size of the time window increases the variance of the estimations, despite using the same number of parameters, as this results in a small under-estimation of $\frac{\rho}{\theta}$. In addition, the complete and incomplete Baum-Welch algorithms lead to identical results, demonstrating that all the information required for the inference is in the estimated transition matrix.

Finally, we explored how imperfect data sets (due to errors in SNP calling, the presence of transposable elements and existing variation in recombination and mutation rates) could affect the inferences obtained using SMC-based methods. We show that a data set with more than 10% of spurious SNPs will lead to poor estimations of the demographic history, whereas randomly removed SNPs (*i.e.* missing SNPs) have a lesser effect on inferences. It is thus better to be stringent during SNP calling, as false data is worse than missing data. Note, however, that this consideration is valid for demographic inference under a neutral model of evolution, while biases in SNP calling also affect the inference of selection (especially for conserved genes under purifying selection). However, if missing SNPs are structured along the sequence (as would be the case with unmasked TEs), there is a strong effect on inference. If TEs are correctly detected and masked, there is no effect on demographic inferences. It is therefore recommended that checks should be run to detect regions with abnormal distributions of SNPs along the genome. Surprisingly, simulation results suggest that removing random pieces of sequences has no impact on the estimated demographic history. Taking this into account,

when seeking to infer demographic history, it seems better to remove sections of sequences than to introduce sequences with SNP call errors or abnormal SNP distributions. However, removing sequences leads to an over-estimation of $\frac{\rho}{\theta}$, which seems to depend on the number and size of the removed sections. The removal of a few, albeit long sequences, will have almost no impact, whereas removing many short sections of the sequences will lead to a large overestimation of $\frac{\rho}{\theta}$. This consequence could provide an explanation for the frequent overestimation of $\frac{\rho}{\theta}$ when compared to empirical measures of the ratio of recombination and mutation rates $\frac{r}{\mu}$. This implies, that in some cases, despite an inferred $\frac{\rho}{\theta} > 1$, the inferred demographic history can surprisingly be trusted. Note also that as discussed in [151], the discrepancy between $\frac{\rho}{\theta}$ and $\frac{r}{\mu}$ can be due to life history traits such as selfing or dormancy.

Simulation results suggest that any variation of the recombination rate along the sequence does not strongly bias demographic inference but slightly increases the variance of the results and leads to small waves in the demographic history (as a consequence of erroneously estimated hidden state transition events because of the non-constant recombination rate along the sequence), as expected from previous works [103]. However, unlike Li and Durbin's results [103], if scaffolds do not share similar rates of mutation and recombination, but are analyzed together assuming that they do, estimations will be very poor. This could be due to the variation of mutation rate being within a scaffold in their study and the discrepancy between out and their results could suggest analyses based on longer scaffolds to be more robust. However, this problem can be avoided if each scaffold is assumed to have its own parameter values, although this would increase computation time, it could provide useful insight in unveiling any variation in molecular forces along the genome, albeit in a coarser way than in [6]. As we found that non-accounted variation of the recombination rate along the sequence can lead to a spurious two-fold variation of population size, we here provide guidelines to test if small detected variations of population size are to be trusted. Since the consequences of a varying recombination rate might depend on the topology of the recombination map, one first needs to estimate the recombination map (*e.g.* using iSMC [6]). If problematic regions are found they can be removed with almost no negative impact on the estimated demography (Figure 7). Otherwise, the recombination map can be used to simulate sequences *e.g.* using scrm [159]), which can be compared to results obtained for a constant recombination rate. Analyses can be run on both data sets to quantify the effect of the recombination map.

3.4.1 Guidelines when applying SMC-based methods

Our aim through this work is to provide guidelines to optimize the use of SMC-based methods for inference. First, if the data set is not yet built, but there is some intuition concerning the demographic history and knowledge of some genomic properties of a species (*e.g.* recombination and mutation rates), we recommend simulating a data set corresponding to the potential scenarios. From these simulations, the transition matrix for PSMC' or MSMC-based methods can be built using the R package eSMC2. The results obtained can guide users when it comes to the amount and quality of data needed (sequence size and copy number) for a good inference. Beyond being used to guide the building of data sets, it is possible to assess the trustworthiness of results obtained using SMC-based methods on existing data sets. If the estimated transition matrix is empty in some places (*i.e.* no observed transition event between two specific hidden states; white squares in Figure 2), it could suggest a lack of data and/or strong variation of the population size somewhere in time. In order to test the accuracy of the inferred demography, the estimated demographic history can be retrieved and used to simulate a data set with more sequences and/or simulate a demographic history with higher amplitude than the estimated one. The SMC method can then be run on the simulated data in order to check whether using more data results in a matching scenario or if a higher amplitude of population size can indeed be inferred, in which cases the initial results are most probably trustworthy.

As mentioned above, it is better to sequence fewer individuals but to have data of better quality. It is also important to note, that more data is not necessarily always better, especially if there is a risk of spurious SNPs (see Figure 5). In some cases, methods are limited by their own theoretical framework, hence no given data set will ever allow a correct demographic inference. In such cases, other methods based on a different theoretical frameworks (*e.g.* SFS and ABC) might perform better [9, 147].

3.4.2 Concluding remarks

Here we present a simple method to help assess how accurate inferences obtained using PSMC' and MSMC would be when applied to data sets with suspected flaws or limitations. We also provide new interpretations of results obtained when hypotheses are known to be violated, and thus explain why results sometimes deviate from expectations (*e.g.* when the estimated ratio of recombination over mutation is larger than the one measured experimentally). We propose guidelines for building/evaluating data sets when using SMC-based models, as well as a method that

can be used to estimate the demographic history and recombination rate given a genealogy (in the same spirit as Popsicle [71]). The estimated transition matrix is introduced as a summary statistic, which can be used to recover demographic history (more precisely the Inverse Instantaneous Coalescence Rate interpretation of population size variation, when assuming a panmictic population [22, 140]). This statistic could, in the future, be used in scenarios with migration, without the computational load of Hidden Markov Models. When faced with complex demographic histories, or $\frac{\rho}{\theta} > 1$, we show that there are strategies that would allow those wishing to use SMC methodology to make the best use of their data.

Chapter 4

The Sequentially Markovian Beta Coalescent

4.1 Motivation

Many models and methods have been developed to extract information from whole genome sequence data [168, 103, 146, 152, 156, 151, 6, 176]. The common feature of all these methods is their underlying assumption of a Kingman coalescent process [95] to describe the genealogy distribution of a sample. The Kingman coalescent process results itself from the traditional assumption of a Wright-Fisher Model to describe the reproduction mechanism of a population. As genome sequence data starts to be available for many different species [109, 41, 57, 43, 25, 60, 49, 36, 83, 170] with as many different biological trait or life cycle. Hence, for some species the underlying assumption of a Wright-Fisher model is strongly questioned [162, 3, 125, 90, 119]. The Wright-Fisher model (and other models leading to a Kingman coalescence process, *e.g.* the classic Moran Model) might not correctly describe some population's evolutionary process because of their assumptions. More specifically, a key parameter is the distribution of the number of offspring that parents can have. In the Wright-Fisher, due to Binomial sampling, the distribution of offspring number per parent is well approximated by a Poisson distribution with mean 1 and variance 1. To give a biological intuition, it means that most parents will have no, one, or two offsprings, but it is very improbable that one parent has many offspring (of the order of the population size, *e.g.* N/2). However, for some species, it is expected that the variance in reproduction between parents can be much larger than under the Poisson distribution, even under neutrality. Strong selection, dormancy, high fecundity with skewed offspring distribution, extremely strong bottlenecks have been theoretically shown to deviate

from the classic Wright-Fisher model in a way that the genealogies can no longer be described by a Kingman Coalescent process [16, 17, 29, 76, 32, 12, 19]. In such conditions, a new class of process arises to describe the genealogy distribution, a class where multiple individuals can simultaneously coalesce and/or multiple distinguished coalescence events can simultaneously occur [143, 118, 28, 142, 136]. We call this class of genealogical processes Multiple Merger Coalescent (MMC).

Because Kingman coalescent and Λ-coalescents describe different processes and thus have extremely different evolutionary interpretation, it is essential to assess which model best describes the species genealogy [98]. Therefore, methods to distinguish which model best describes the data, and thus the genealogy, are required [98, 97, 33, 68, 113]. Current methods rely on the Site Frequency Spectrum (or derived summary statistic) [98], minor allele frequency [139] or copy number alteration [90] for model selection and parameter estimations . However, this implies that current methods do not directly integrate linkage disequilibrium, although some statistics have been proven to be robust to recombination [98, 139]. Besides, these methods usually need a high sample size (>50) for trustworthy results, which might not be available for non-model species [98, 97, 33].

With the development of the Sequentially Markovian Coalescence theory [115, 111], it becomes tractable (*i.e.* possible) to integrate linkage disequilibrium over chromosomes in inferences [103] based on the Kingman coalescence theory. In addition, it was demonstrated in [149] that if the probability of a parent to have k or more offspring is proportional to k^{α}, where $1 < \alpha < 2$, then the genealogy can be described by the Λ coalescent (a general class of coalescent process describing how and how fast blocks/individuals merge [136, 142]) in which the measure is the Beta(2− α,α) distribution. This coalescent process was thus named the Beta coalescent. If α tends to 2, then the coalescence process tends to a Kingman coalescent up to a scaling constant. If *alpha* tends to one, the model tends to a Bolthausen-Snitzman coalescence process (*i.e.* star shaped). We next define the merging rates of the Beta coalescent. The rate of transition from a state with b lineages (*i.e.* current number of individuals) to $b-n+1$ lineages, *i.e.* a merger of n lineages is :

$$\Lambda_{b,\alpha,b-n+1} = \frac{\binom{b}{n} B(n-\alpha, b-n+\alpha)}{\Gamma(2-\alpha)\Gamma(\alpha)}. \tag{4.1}$$

Thus, the total rate (*i.e.* rate to the next merging event) is :

$$\lambda_{b,\alpha} = \sum_{k=2}^{b} \frac{\binom{b}{n} B(k-\alpha, b-n+\alpha)}{\Gamma(2-\alpha)\Gamma(\alpha)} \tag{4.2}$$

Where k is the number of merging individuals. With decreasing α the probability of having more than 2 lineages merging increases, but when α tends to 2, the probability of having more than 2 lineages merging tends to 0 (*i.e.* Kingman coalescent process).

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 [149], 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 :

$$f(t) = \frac{\lambda}{\chi(t)} e^{-\int_{t_0}^{t} \frac{\lambda}{\chi(s)} ds}, \tag{4.3}$$

which follows as described in [67]. In addition, we chose our scaling to make our coalescent process tend to the Kingman coalescent when α tends to 2.

Hence, to detect and account for multiple merger events along the genome we develop a new Sequentially Markovian Coalescent approach assuming a Beta-coalescent. Our approach can thus approximate the Ancestral Recombination Graph (ARG) using the Sequentially Markovian Coalescent [146]. We build a Sequentially Markovian β Coalescent (SMβC). In addition, the theory describing the exact ARG has been build in [10] allowing the development of a new sequence simulator included in msprime with recombination for multiple merger coalescent models [91].

From the Sequentially Markovian β Coalescent we derived an inference method based on the Multiple Sequentially Markovian Coalescent (MSMC) [146] allowing multiple individuals to simultaneously coalesce. In addition, we modified the underlying hidden markov model to account for spurious multiple mergers originating from the discretization of time. Our model can thus infer recombination rate, population size, and the parameter of the Beta distribution, determining how frequently multiple individuals simultaneously coalesce. We first demonstrate the effect of assuming a Kingman coalescent model when the underlying coalescent model is a Beta coalescent. We then demonstrate the theoretical accuracy of our approach.

4.2 Materials and Methods

In this study, we use three different SMC-based methods: MSMC, MSMC2, eSMC, and our new method SMβC. The basis of novelty in SMβC is that to detect multiple merger events, we need mergers of more than 2 lineages, thus we need to use methods based on multiple sequences (*i.e.* MSMC), and not based on the PSMC. All methods are Hidden Markov Models and use whole genome sequence polymorphism data. The hidden states of these methods are the coalescence times (or first coalescence time for sample size larger than 2) of a sample. In order to have a finite number of hidden states, the hidden states are grouped into x bins (x being the number of hidden states) results of the discretization of time (and an index describing who coalesces for sample size $>$2). The reasons for our model choices are as follows: SMβC to check the convergence properties of our new method and demonstrate its efficiency to uncover multiple merger events. MSMC, from which SMβC is mathematically derived, to compare its convergence properties with SMβC [146]. MSMC2 and eSMC (using the same input file as MSMC) because it computes coalescent times of all pairwise analysis from a sample of size n, and can deal with a large range of data sets [46].

4.2.1 SMC methods

eSMC and MSMC2

eSMC and MSMC2 focus on the coalescence events between only two individuals, and thus do not require phased data. The algorithm goes along the sequence and estimates the coalescence time at each position. Both methods check whether the two sequences are similar or different at each position. If the two sequences are different, this indicates a mutation took place. The absence of mutation (the two sequences are identical) suggests a recent common ancestor. 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 [46, 176, 151].

MSMC

MSMC simultaneously analyses multiple sequences and because of this, MSMC requires the data to be phased. In combination with a second HMM, to estimate the external branch length of the genealogy, it can follow the distribution of the first coalescence event in the sample along sequences. A detailed description of MSMC can be found in [146].

SMβC

SMβC is based on MSMC. Hence it simultaneously analyses multiple sequences and thus also required the data to be phased. It can follow the distribution of the first coalescence event in the sample along sequences assuming a Beta coalescent and therefore allow for more than two individuals to join the first coalescence event. The emission matrix is similar to the one of MSMC. However, currently, SMβC has been derived for up to 4 sequences simultaneously (due to computational load and complexity). The transition probabilities of SMβC for sample size 3 (when time is continuous) are displayed in equation 4.1. A detailed description of SMβC can be found in Appendix A.3.

$$
\begin{cases}
 & p(t,i|s,j,u) = \\
P_s \frac{2\lambda_{2,\alpha}}{\chi_t \binom{M}{2}} e^{-\int_u^t \frac{M\lambda_{M+1,\alpha,M}}{\chi_v} dv} & u < t < s \\
(1-P_s) + P_s(\int_u^s \frac{1}{\chi_k} e^{\int_u^k -\frac{M\lambda_{M+1,\alpha,M}}{\chi_v} dv} dk + \frac{(M-n)\lambda_{(n+1),\alpha,2}}{M\lambda_{(n+1),\alpha,2}+\lambda_{(n+1),\alpha,1}} e^{\int_u^t -\frac{M\lambda_{M+1,\alpha,M}}{\chi_v} dv}) & t = s, m = n \\
P_s \frac{(M-n)\lambda_{(n+1),\alpha,1}}{M(\lambda_{(n+1),\alpha,2}+\lambda_{(n+1),\alpha,1})} e^{-\int_u^s \frac{M\lambda_{M+1,\alpha,M}}{\chi_v} dv} & t = s, m = n+1 \\
P_s \frac{n}{M} \frac{\lambda_{(n+1),\alpha,2}}{s(\lambda_{(n+1),\alpha,2}+\lambda_{(n+1),\alpha,1})} e^{-\int_u^s \frac{M\lambda_{M+1,\alpha,M}}{\chi_v} dv} & t = s, m+1 = n \\
P_s \frac{\lambda_{M,\alpha,(M-m+1)}}{\binom{M}{m}\chi_\alpha} e^{-\int_s^t \frac{\lambda_{M,\alpha}}{\chi_v} dv} e^{-\int_u^s \frac{M\lambda_{M+1,\alpha,M}}{\chi_v} dv} \frac{2\lambda_{(n+1),\alpha,2}}{M(\lambda_{(n+1),\alpha,2}+\lambda_{(n+1),\alpha,1})} & t > s, i = l, j = k
\end{cases}
\tag{4.4}
$$

Where

- 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 ($N_t = \xi_t N_0$)
- $\chi_t = \xi_t^{\alpha-1}$
- M : Number of analyzed sequences (or individuals)
- α : The multiple merger parameter
- $P_s = (1 - e^{-Mrs})$ represents the recombination probability

4.2.2 Simulated data

Simulated ARG

Using the sequence simulators msprime [91], one can simulate the Ancestral Recombination Graph of a sample. From the simulated ARG, one can find what state of the HMM the sample is in at each position of the simulated sequences. Hence, one can build the series of states along the genomes, and build a transition matrix. The transition matrix is a square matrix (of dimension x defined as the number of hidden states) counting the number of transitions from one of the x states to another (it also counts the number of transitions from one state to the same state). Using the transition matrix built directly from the exact ARG, one can estimate parameters of SMC methods as if the HMM could perfectly infer the sequence of hidden states. Hence estimations using the exact transition matrix represent the upper bound of performance for those methods and can thus demonstrate the potential accuracy of a method. Therefore, one can determine the best-case convergence of any SMC method based on HMM. We use msprime to simulate the ARG (also under constant population size) when the underlying coalescent model is a Beta coalescent [149, 10]. We analyze the ARG assuming the recombination rate is known or not (*i.e.* the recombination rate is fixed or set free to be inferred).

4.2.3 Simulated Sequence data

To test the effects of assuming a Kingman coalescent process when the underlying model is a Beta coalescent, we use msprime to simulate sequence data. We simulate sequence data (of diploid individual) under 2 demographic scenarios (constant population size, sawtooth). We first check the effect of multiple mergers on demographic inferences by analyzing simulated sequence data under a Beta coalescent where population size is constant. Under the Beta coalescent model, coalescence time is not scaled linearly to the population size. Thus, to study the effect of varying population size on inferences, we analyze simulated data under a Beta coalescent model where the population undergoes a sawtooth demographic scenario.

4.3 Results

We first study the effect of assuming a Kingman coalescent when the underlying model is a Beta coalescent, results, when population size is constant are displayed in Figure 4.1. All methods fail to correctly infer the population size, due to the scaling discrepancy between the Kingman and Beta coalescence. However, all methods correctly infer a constant population size for α values >1.5. Yet, for smaller α

values, a spurious recent bottleneck can be observed. Results of MSMC2 for beta values smaller than 1.9 are not displayed due to the method crashing for those beta values. Decreasing the parameter of the Beta coalescence (increasing multiple merger probability) increases the variance of inferences. Similar results are observed when sequences are simulated under a sawtooth demographic scenario (Figure 4.2). In addition, assuming a Kingman coalescent model when the underlying model is a Beta coalescent with small a α value will slightly flatten demographic history.

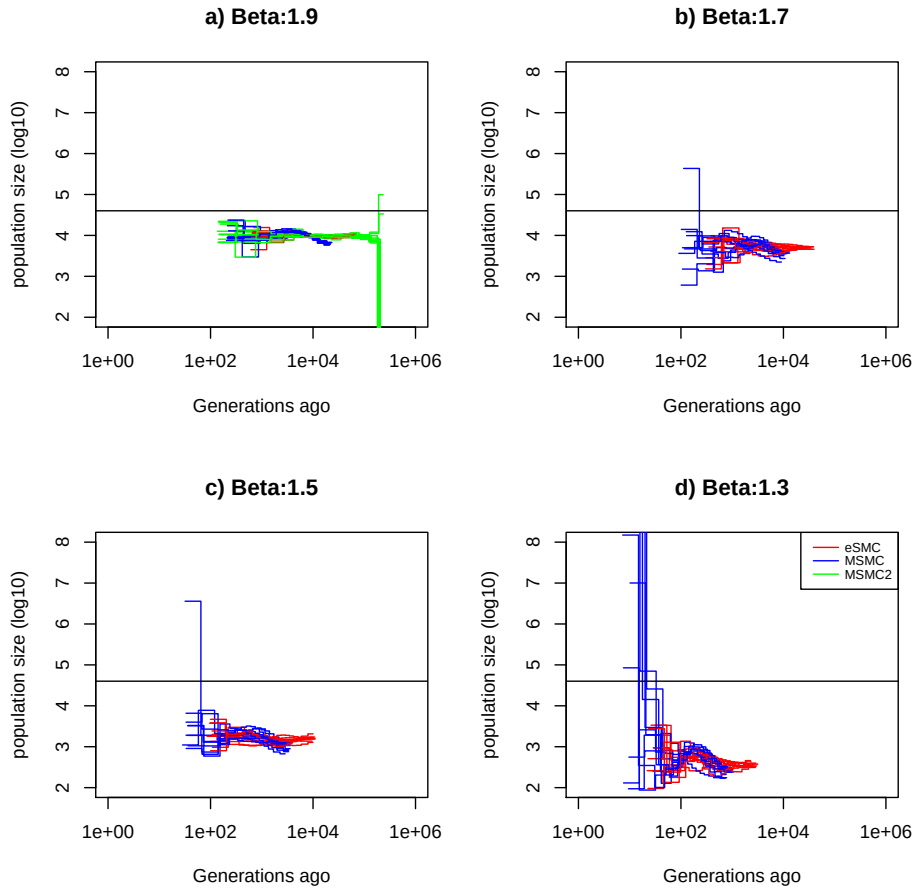

Figure 4.1: **Performance of MSMC, MSMC2 and eSMC under a Beta coalescent.** Estimated demographic history by MSMC, MSMC2 and eSMC using 3 sequences of 100 Mb (respectively in blue, green and red) when population size is constant (black) for different α values, 1.9,1.7,1.5,1.3 respectively in a),b),c) and d). The recombination and mutation rate are set to 1×10^{-8} per generation per bp.

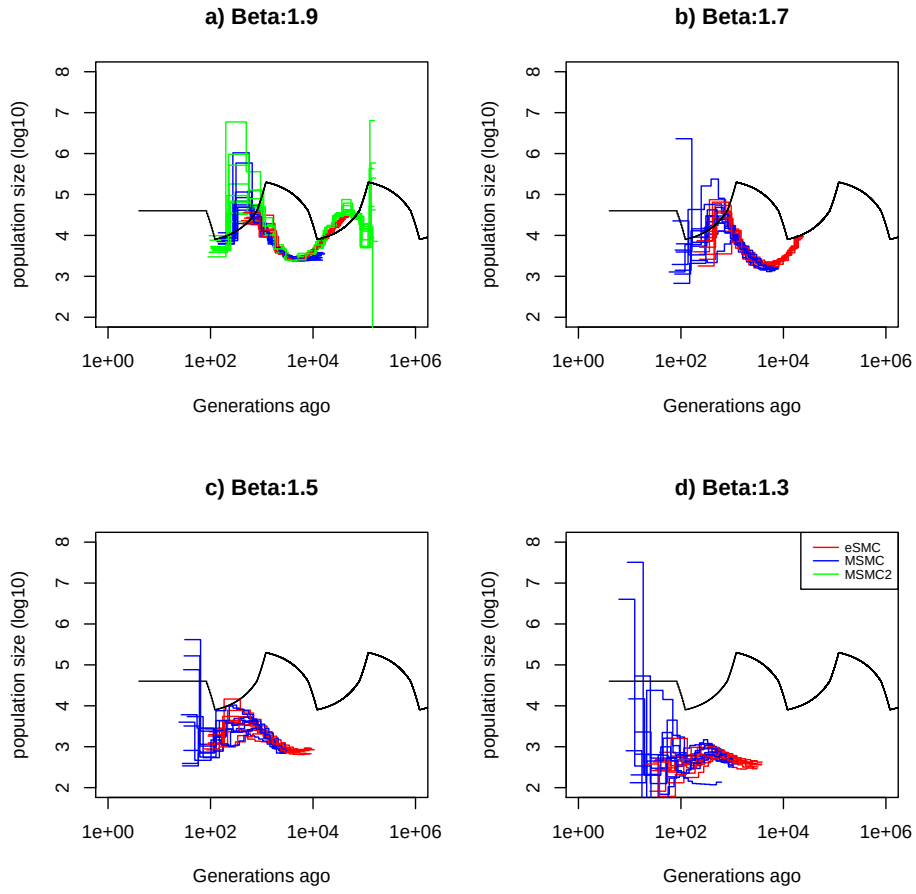

Figure 4.2: **Performance of MSMC, MSMC2 and eSMC under a Beta coalescent.** Estimated demographic history by MSMC, MSMC2 and eSMC using 3 sequences of 100 Mb (respectively in blue, green and red) when population undergoes a sawtooth demographic scenario (black) for different α values, 1.9,1.7,1.5,1.3 respectively in a),b),c) and d). The recombination and mutation rate are set to 1×10^{-8} per generation per bp.

To check if our model can theoretically recover the past demographic history and the parameter α of the Beta coalescent, we give as input the ancestral recombination graph to our new method SMβC. Results for data simulated under a constant population size are displayed in Figure 4.3. For α values high enough ($>$1.5), the constant population size is recovered. However, for smaller values, an extremely high variance is observed, limiting all interpretations. In addition fixing

(*i.e.* constraining) the recombination rate to its value seems to strongly bias inferences and increasing variance. Similar results are obtained when the underlying demographic history is sawtooth shaped (Figure 4.4). Estimated values of α are written in Table 4.1. When the population size is constant and the recombination rate is set free, α is fairly well recovered. However, when population size varies, α can be underestimated. Fixing the recombination rate leads to poor estimations of α.

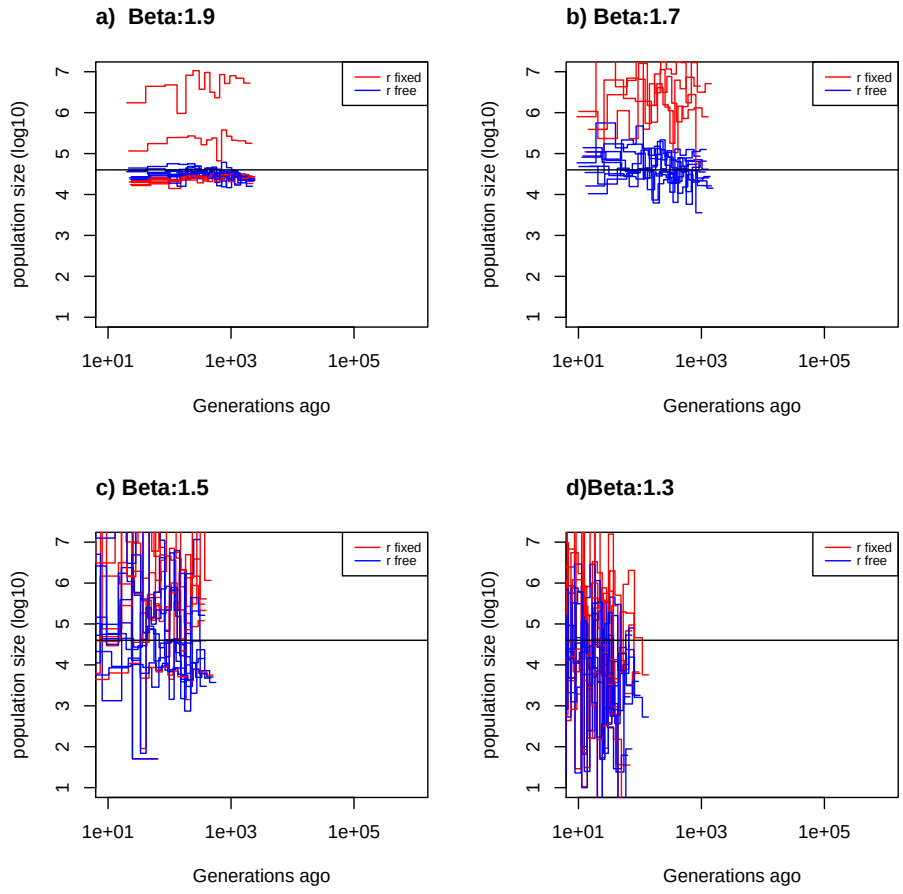

Figure 4.3: **Performance of SMβC under a Beta coalescent.** Best-case convergence estimations of demographic history by SMβC using 3 sequences of 100 Mb with recombination rate fixed or set free to be inferred (respectively in red and blue) when population size is constant (black) under 4 different 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^{-7} per generation per bp.

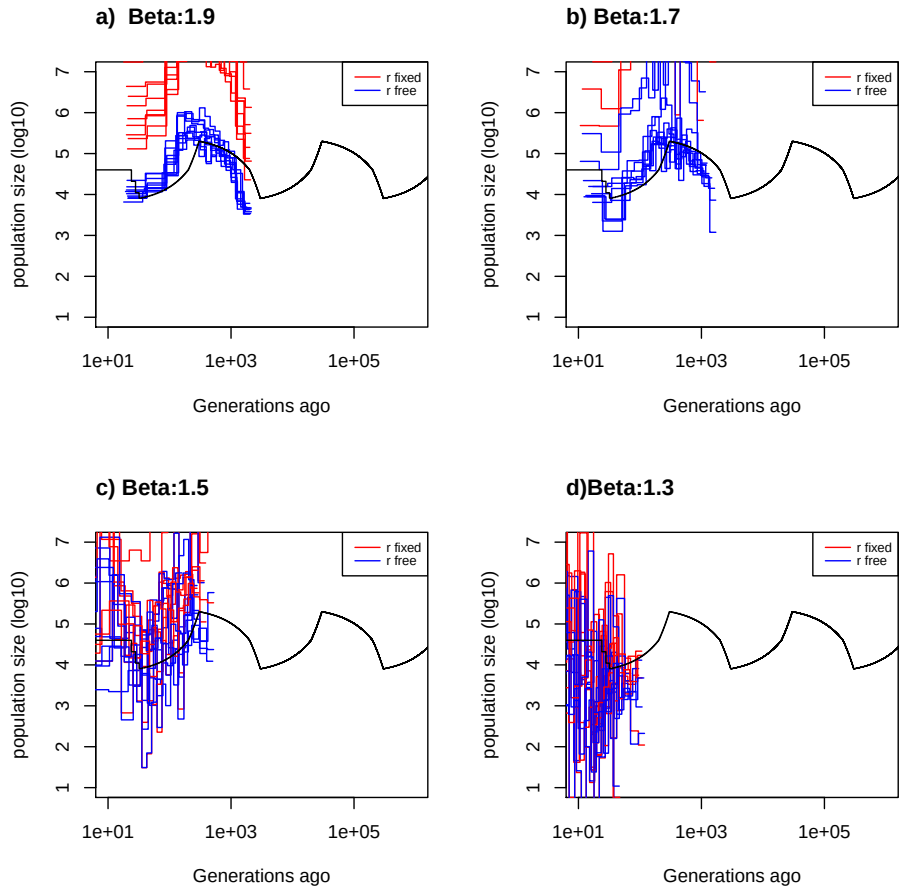

Figure 4.4: **Performance of SMβC under a Beta coalescent.** Best-case convergence estimations of demographic history by SMβC using 3 sequences of 100 Mb with recombination rate fixed of set free to be inferred (respectively in red and blue) when population undergoes a sawtooth demographic scenario (black) 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^{-7} per generation per bp.

scenario	α	estimated α and r free	estimated α and r fixed
Constant	1.9	1.82 (0.03)	1.63 (0.18)
Constant	1.7	1.63 (0.05)	1.24 (0.09)
Constant	1.5	1.53 (0.15)	1.4 (0.18)
Constant	1.3	1.35 (0.06)	1.31 (0.07)
Sawtooth	1.9	1.67 (0.03)	1.3 (0.06)
Sawtooth	1.7	1.59 (0.07)	1.16 (0.06)
Sawtooth	1.5	1.45 (0.08)	1.32 (0.08)
Sawtooth	1.3	1.38 (0.08)	1.32 (0.07)

Table 4.1: Average estimated values of α by SMβC over ten repetitions using 3 sequences of 100 Mb with recombination and mutation rate set to 1×10^{-7} per generation per bp under a Beta coalescent process (with different alpha parameter). The coefficient of variation is indicated in brackets.

In order to test if SMβC could suggest a Kingman coalescent, we analyzed the ancestral recombination graph simulated under Kingman. Results for data simulated under a constant population size are displayed in Figure 4.5. The constant population size is recovered. However, there is a scaling discrepancy between the simulated and estimated demographic history as in Figure 4.1. Similar results are obtained when the underlying demographic history is sawtooth shaped (Figure 4.6). Estimated values of α are written in Table 4.2. Estimated alpha values are higher than 1.85 (when r is set free) suggesting an underlying Kingman coalescent process.

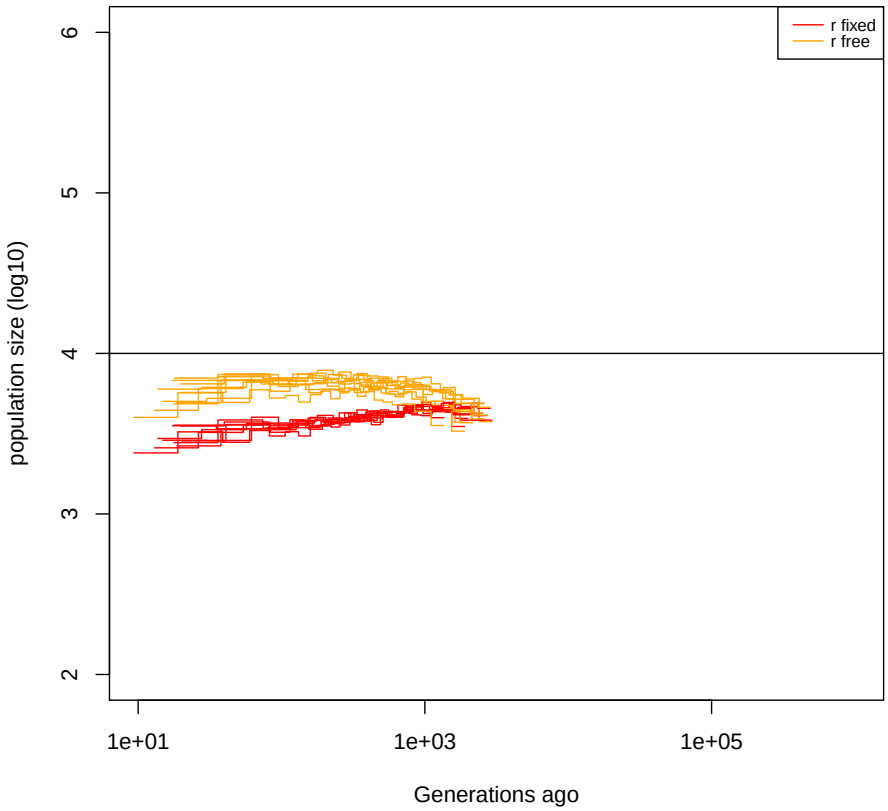

Figure 4.5: **Performance of SMβC under a Kingman coalescent.** Best-case convergence estimations of demographic history by SMβC using 3 sequences of 100 Mb with recombination rate fixed of set free to be inferred (respectively in red and orange) when population size is constant (black) under a Kingman coalescent. The recombination and mutation rate are set to 1×10^{-8} per generation per bp.

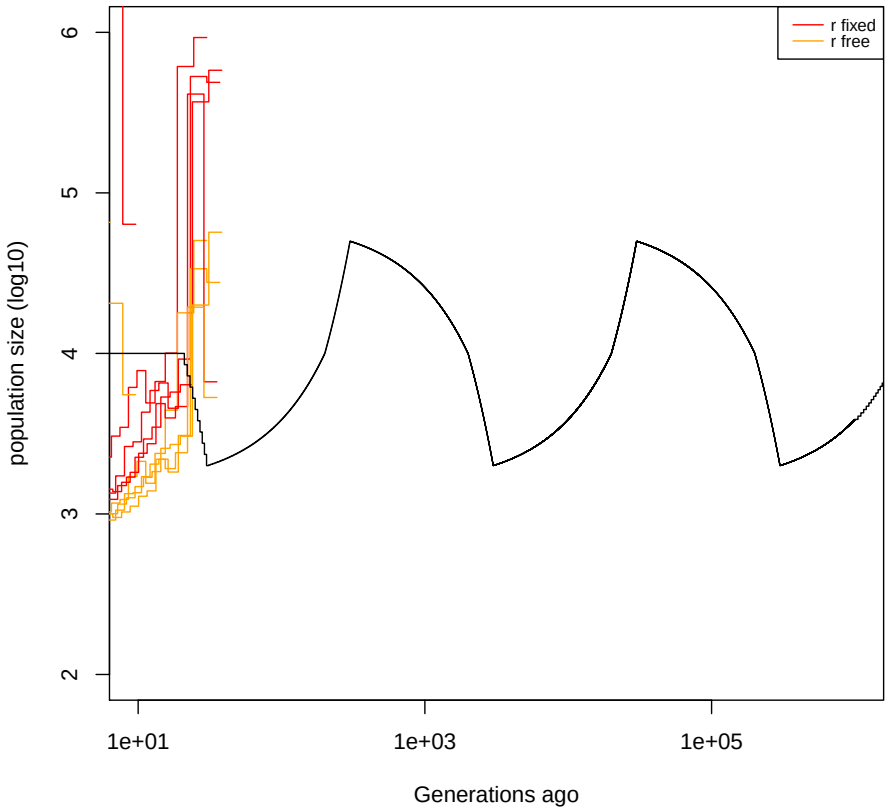

Figure 4.6: **Performance of SMβC under a Kingman coalescent.** Best-case convergence estimations of demographic history by SMβC using 3 sequences of 100 Mb with recombination rate fixed of set free to be inferred (respectively in red and orange) when population undergoes a sawtooth demographic scenario (black) under a Kingman coalescent. The recombination and mutation rate are set to 1×10^{-8} per generation per bp.

scenario	estimated α and r free	estimated α and r fixed
Constant	1.93 (0.005)	1.9 (0.016)
Sawtooth	1.88 (0.009)	1.39 (0.09)

Table 4.2: Average estimated values of α by SMβC over ten repetitions using 3 sequences of 100 Mb with recombination and mutation rate set to 1×10^{-8} per generation per bp under a Kingman coalescent. The coefficient of variation is indicated in brackets.

4.4 Discussion

In this study, we have demonstrated the effect of assuming a Kingman coalescent process when the underlying coalescent process is a Beta coalescent. The scaling discrepancy will lead to an erroneous inference of population size and thus temporal interpretation of inferences. Assuming a Kingman can also lead to flattened demographic history due to the non-linear scaling of the coalescence rate in the Beta coalescence. In addition, a spurious recent bottleneck can be observed under small α values ($<$1.7), which has also been found as a signature similar to selection [148]. Potentially indicating confounding effect with pervasive or strong selection.

When given the ARG and sufficient data, our method can recover the α parameter under a constant population size (as well as recovering the constant population size). However, variations of the population size seem to affect the estimation of α (by underestimating it). Yet, inferences of α still make sense to some extent since the smaller α, the smaller α is inferred. In addition, our model fails to recover the recombination rate from the ARG (overestimating it by a factor between 1 and 2), potentially indicating a scaling issue. This discrepancy could originate from the scaling discrepancy between our implementation and the simulator (our implementation tends to the Kingman coalescent when α tends to 2, not the simulator), but could also originate from the limit of the Markovian hypothesis when dealing with multiple merger [19]. In addition, we observe that smaller α values, increase the variance of estimations, suggesting the needs of increasing the data for smaller α values.

We found our method capable to suggest a Kingman coalescent under a constant population size. However, variations of population size can slightly diminish the inference of α, suggesting that population undergoing strong and rapid variation of population size could mimic the signature of multiple merger events as suggested in [70, 19]. In addition, our method requires phased data, which is known to bias

inferences [168]. One would need to test the effect of phasing error on demographic and α estimations.

A surprising result we found is that multiple merger events seem to not strongly affect the ratio of recombination and mutations ($\frac{\rho}{\theta}$). However, decreasing α (*i.e.* increasing the relatedness between individual) should increase linkage disequilibrium and thus diminish $\frac{\rho}{\theta}$. Intuition would suggest (under neutrality) that multiple merger events (*i.e.* large production of offspring) should increase inbreeding the population, thus results similar to self-fertilization or inbreeding are expected. We failed to find in the literature formulas describing the probability of "effective recombination" (*i.e.* visible) under the beta coalescent. In addition, the potential "invisibility" of recombination events in presence of large offspring production is currently not implemented in msprime. Therefore there is a correction that remains to be derived in our implementation, which could improve the estimation of α and the biological meaning behind our model.

On a brighter tone, new statistics have arisen, capable of suggesting an underlying Beta coalescent process and being largely insensitive to the population size [139]. Thus, prior values of α could be first obtained or hinted and then inputted in SMβC [97, 68]. In addition, inputting/inferring an erroneous α value will lead to extremely strong biased population size inferences (*e.g.* a population of 10 individuals or 10^{10}, depending on the bias on α), resulting in inferences with no biological meaning. Thus, boundaries on the potential values of Beta could easily be calculated by setting realistic boundaries on the population size (*e.g.* $100<$ population size $<10^8$, depending on the species), which could strongly improve inferences accuracy. If the data allows it, the model could be selected using the singleton-tail statistic of [98]. At last, ploidy can strongly affect the Beta coalescence process [11]. Currently, only an implementation for a haploid population is available, but an extension to account for different ploidy is in development.

Chapter 5

Integration of methylation data in the Sequentially Markovian Coalescent

5.1 Motivation

With the simultaneous rise of sequencing technologies and theoretical frameworks [170, 171, 156, 92], it has now become basic routine to estimate the past demographic history of populations and species [147]. Interpreted under the correct model [140, 22, 176, 151], the past demographic history will describe what the species has undergone, help detecting selection signatures [156, 131] and can facilitate the detection of effective population size decrease [178], unveiling endangered populations. Hence, the demographic history is a key information in evolutionary biology, justifying the massive theoretical framework and methodology which has been developed to estimated the past demographic history from whole genome sequencing data [152, 161, 156, 8, 140, 89, 109, 87, 155, 175, 71, 168, 146, 151, 176, 103, 73, 23]. Methodology can be clustered in two main categories [157, 9], those based on the Site Frequency Spectrum (SFS) (or other summary statistics) [73, 109, 8, 155, 175, 13, 138], and those going along the sequence inferring the Ancestral Recombination Graph (ARG) from which model parameters are estimated [103, 146, 168, 156, 6, 176, 151, 152, 161]. Although there is evidence that methods based on the SFS can display higher accuracy in recent times [9], we will focus on the second class of model which has higher resolution and can display high accuracy for small sample [103, 146, 176, 168, 156, 151, 157, 150].

The first method to integrate (*i.e.* account and infer recombination events

while analyzing whole genome sequence data to infer the demographic history is the famous Pairwise Sequentially Markovian Coalescent (PSMC) [103]. This was rendered possible after the work of [180, 115, 111] allowing to model recombinations as a point process along sequences in a tractable way. Ever since PSMC was released, its accuracy in the far past has not been outperformed [156, 53]. However, its limitation to infer events in recent times has been acknowledged [146, 168, 9]. To improve and outperform PSMC, theoreticians and computational biologists have opted for one main strategy: integrating more data. In order to improve inference in recent times, PSMC has been extended to simultaneously account for multiple sequences (*i.e.* more than two) into the method known as the Multiple Sequentially Markovian Coalescent (MSMC) [146]. Methods developed after MSMC followed suit, with MSMC2 [46] extending PSMC by incorporating pairwise analysis, increasing efficiency, and the number of sequences that can be inputted (up to a hundred), resulting in more accurate results. SMC++ [168] brings analyses to another level by allowing the use of hundreds of unphased sequences to increase accuracy in recent times [168]. After SMC++, relate has been developed [156], emancipating itself from the sequentially markovian coalescence theory by using the theory of [105]. With this new theoretical basis, relate can scale up to sample size of multiple thousands, outperforming in recent times (*i.e.* < 1000 generations ago) all other methods.

Currently, most sophisticated methods either require a large number of well sequenced individuals ($>$20) to display good performance in recent times [156, 168, 109]. Yet, such data sets might be unrealistic for biologists working under budget constraints or endangered species. However, it has recently been shown that SMC methods fail to correctly extract all information from genome sequence polymorphism data, partially explaining the poor accuracy in recent times [150]. Theoretical results suggest good performance in recent times could be obtained given an optimum amount of data and by correctly recovering the ancestral recombination graph [150, 71]. Authors show that segregating sites alone are insufficient to correctly recover all information from sequence data [150]. However, current models ignore microsatellites, insertion, deletions, transposable elements, and epigenetics markers [156, 168, 6, 176] which could all be seen as markers and potentially be used and accounted for. Thus, there is a huge potential improvement that would not necessarily increase the monetary cost (as much as sequencing hundreds of individuals) of analysis. As a first step, we here offer a novel method (SMCm) simultaneously integrating sequence and methylation (on cytosine in cytosine-guanine context) polymorphism data leading to increased accuracy in recent times. Our choice is based on the rising availability of bisulfite sequencing data [182] (*i.e.* methylation data).

Studies on plant (and some other species *e.g.* fungus) methylome data suggests that epimutations (or epigenetic modifications) are heritable [65, 100, 185, 50]. Furthermore, studies of the methylated Site Frequency Spectrum (mSFS) of *Arabidopsis thaliana* is well described by theoretical prediction [173, 21], suggesting robust modeling and hypothesis [173]. Based on the Vidalis results, epimutations can be considered neutral at coding regions (while this might be not true for transposable elements) [173, 164, 141]. From those results, it became reasonable to integrate methylation polymorphism information to sequence polymorphism data in order to improve the accuracy of current methods. Because methylation and demethylation rates are much faster than mutation rates [172], methylation polymorphism allows us to estimate the length of short genealogy branches (*i.e.* recent coalescent event). In addition, it also solves a major flaw in most methods, which is the sensibility toward the ratio of recombination over mutation. When methylations (or epimutations) scales much faster than recombination, recombinations which cannot be detected due to the lack of segregating sites (*i.e.* SNPs) can be detected due to methylation polymorphism. Furthermore, methylation polymorphism can help to detect and distinguish demographic scenarios with a strong and rapid variation of population size, as current methods fail to correctly detect them [150].

From the Pairwise Sequentially Markovian Coalescent, we derive an inference method integrating methylation polymorphism [146]. We modify the emission matrix to account for methylation polymorphism allowing for new types of observations as well as changing the original mutation model to now assume a finite site. We also developed a simulator extension, capable of creating sequences which take as input an ARG in the Newick format (*e.g.* outputted by ms and scrm [159, 85]) or from the sequence simulator msprime [91]. In this chapter, we will focus on simulated data, but argue the application to sequence data in the discussion. We first demonstrate the increased accuracy and ability to detect a rapid and strong variation of population size when inferring demographic history with our approach. We then show the increasing capacity to estimate the ARG with the increasing proportion of methylation polymorphism annotated. We then establish the model's capacity to estimate and detect recombination rate higher than the mutation rate.

5.2 Materials and Methods

5.2.1 Methods

In this study, we use two different SMC-based methods: eSMC and our new SMCm. All methods are Hidden Markov Models and use whole genome sequence polymor-

phism data except SMCm which also integrates methylation polymorphism. The hidden states of these methods are the coalescence time of a sample size 2. To have a finite number of hidden states, the coalescence times are grouped into x bins (x being the number of hidden states) resulting from the discretization of time (and an index describing coalescing individuals for sample size >2). The reasons for our model choices are as follows: SMCm to check the convergence properties of our new method and demonstrate its efficiency to better uncover coalescence events. eSMC, from which SMCm is derived, to compare its convergence properties with SMCm to measure the accuracy increase of integrating methylations.

eSMC

eSMC focuses on the coalescence events between only two individuals, and, as a result, does not require phased data. The algorithm goes along the sequence and estimates the coalescence time at each position. Both methods check whether the two sequences are similar or different at each position. If the two sequences are different, this indicates a mutation took place. The absence of mutation (the two sequences are identical) suggests a recent common ancestor. 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 [46, 176, 151].

SMCm

SMCm is based on eSMC, thus it focuses on the coalescence events between only two individuals, hence also does not require phased data. The main, difference is that it accounts for sites that are potentially methylated. Because the model accounts for epimutations, there are five possible observations. As in eSMC, if the two nucleotides are identical at a non methylable site, we indicate this as 0. If the two nucleotides are different, it is indicated as 1 (*i.e.* a mutation occurred). If the methylation state is annotated at a position there are three possible observations. If the two cytosines are unmethylated, it is indicated as a 2. If the two cytosines are methylated, it is indicated as a 3. If at a position a cytosine is methylated and the other one unmethylated, it is indicated as a 3. Depending on the mutation, methylation and, demethylation rates, each observation has a different probability of occurring depending on the coalescence time. The complete description and the probabilities can be found in appendix A.4. SMCm has been implemented in a R package SMCm, only available upon request as the package is currently in development.

5.2.2 Simulated data

Throughout this chapter, we simulate different demographic (constant population size, sawtooth scenario and, bottlenecks) scenarios using the coalescence simulation program scrm [159]. To simulate our sequences with annotated methylation, we first simulate genealogies in the Newick format. We then process the ARG to create sequences to which we add mutations and epimutations. To assess mutations and epimutations we build a mutation and epimutation model. Our model is inspired by the work of [21] which we extended using formulas of [184] (which assumes a Juke-Cantor model) for any sample size. Results for mutations can be found in theorem 1 of appendix A.4. Results for mutations and epimutations can be found in theorem 2 of appendix A.4.

As methylation and demethylation occur at higher rates than recurrent mutations, we expect to increase the accuracy of inferences in recent times. We enlarge the analysis time window to infer demographic history in recent times. To check the effect of integrating epimutations, we simulate data under a sawtooth demographic scenario (sample size two with ten scaffolds of 100 Mb, mutation and, recombination rate are set to 1×10^{-8} per bp per generation.). We simulate data where only 30 % of CG have their methylation sate annotated, which is slightly more than what is expected from the plant model *Arabidopsis thaliana* in CG context, but similar to the proportion of cytosine which can be methylated globally [106, 38, 187, 61]. In addition, to test the effect of the methylation and demethylation rates, we simulate data with methylation rate and demethylation rate of $1 \times 10^{-}4$ and $5 \times 10^{-}4$ respectively, or $1 \times 10^{-}3$ and $5 \times 10^{-}3$ respectively which are rates that have been observed in *Arabidopsis thaliana* [173].

Bottleneck events can be smoothed when inferred by SMC methods [150]. In addition, correctly inferring bottlenecks can be crucial in the field of conservation genetics and, because of small population sizes, the lack of SNPs can lead to poor demographic history inferences. Thus we simulate sequence data (sample size 2 and sequence length of 100 Mb) with mutation and epimutations under a bottleneck scenario with two average population sizes (ten thousand or one thousand) to investigate the effects of adding epimutations on inferences. Mutation and recombination rates are set to 1×10^{-8} per bp per generation. Methylation rate is set to 1×10^{-4} and demethylation rate set to 5×10^{-4}. Again, we simulate data where only 30 % of CG have their methylation sate annotated.

When the recombination rate is higher than the mutation rate, inferences of past demographic history and inferences of the recombination rate can be biased

[168, 151]. Since methylation and demethylation rates are faster than the mutation rate and recombination rate, we study the effect of adding epimutations on inferences when the mutation rate is slower than the recombination rate. To do so we simulate the sequence data under a bottleneck scenario of sample size 2 and with sequences length of 100 Mb. The mutation rate is set to 5×10^{-9} and the recombination rate to 5×10^{-8} per bp per generation. Methylation rate is set to 1×10^{-4} and demethylation rate set to 5×10^{-4}. In addition, we increase the number of iteration to optimize likelihood to 200 as it was shown to increase methods accuracy (Supplementary Figure A.18).

5.3 Results

Results of eSMC and SMCm under the sawtooth demographic history with an enlarged time window are displayed in Figure 5.1. eSMC and SMCm both correctly infer a sawtooth demographic scenario. Yet, SMCm better infers population size variation, especially in recent times. In addition SMCm also better infers the amplitude of population size variation. Similar results are observed when the methylation and demethylation rate are respectively set to 1×10^{-3} and 5×10^{-3} per generation per bp (Supplementary Figure A.23). However, when analyzing smaller data sets (one scaffold of 100 Mb), SMCm displays higher variance in recent times than eSMC (Supplementary Figure A.24).

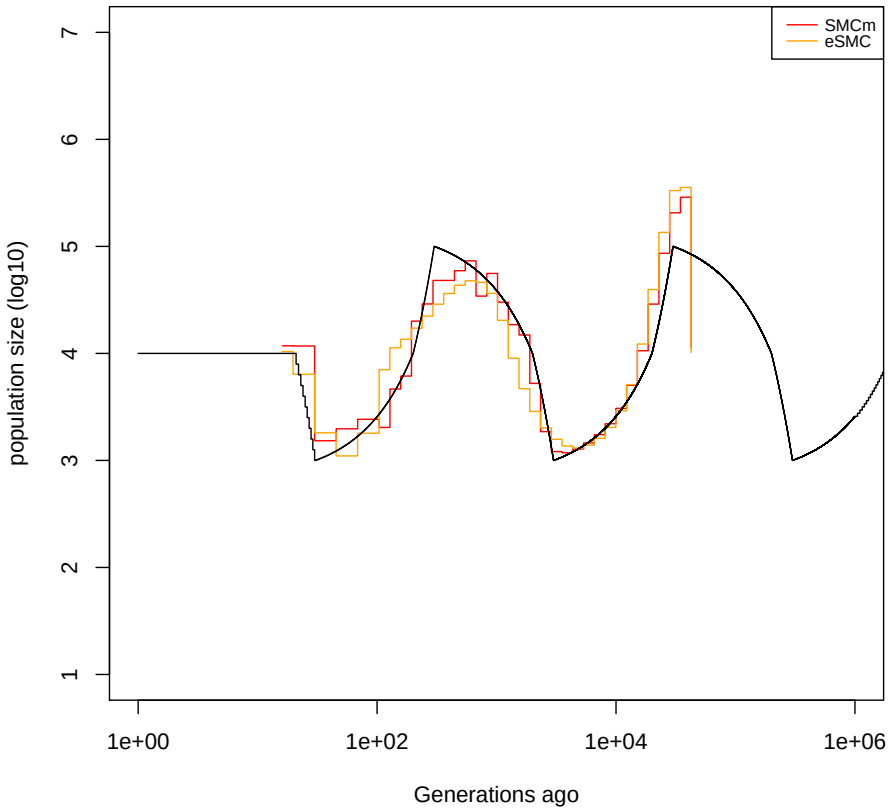

Figure 5.1: **Performance of eSMC and SMCm under a sawtooth scenario.** Estimated demographic history by SMCm and eSMC using 10 scaffolds each of 100 Mb with sample size 2 (respectively in red and orange) under a sawtooth scenario (black). The recombination and mutation rate are set to 1×10^{-8} per generation per bp and the methylation and demethylation rate are respectively set to to 1×10^{-4} and 5×10^{-4} per generation per bp.

Results of eSMC and SMCm under a recent bottleneck are represented in Figure 5.2. Both methods correctly infer the amplitude of population size variation. Yet eSMC fails to infer the sudden variation of population size and infer exponential variation of population size. However, SMCm correctly infers the rapid change of population size undergoing a bottleneck. Similar results are obtained when the population size is ten times smaller (Supplementary Figure A.25).

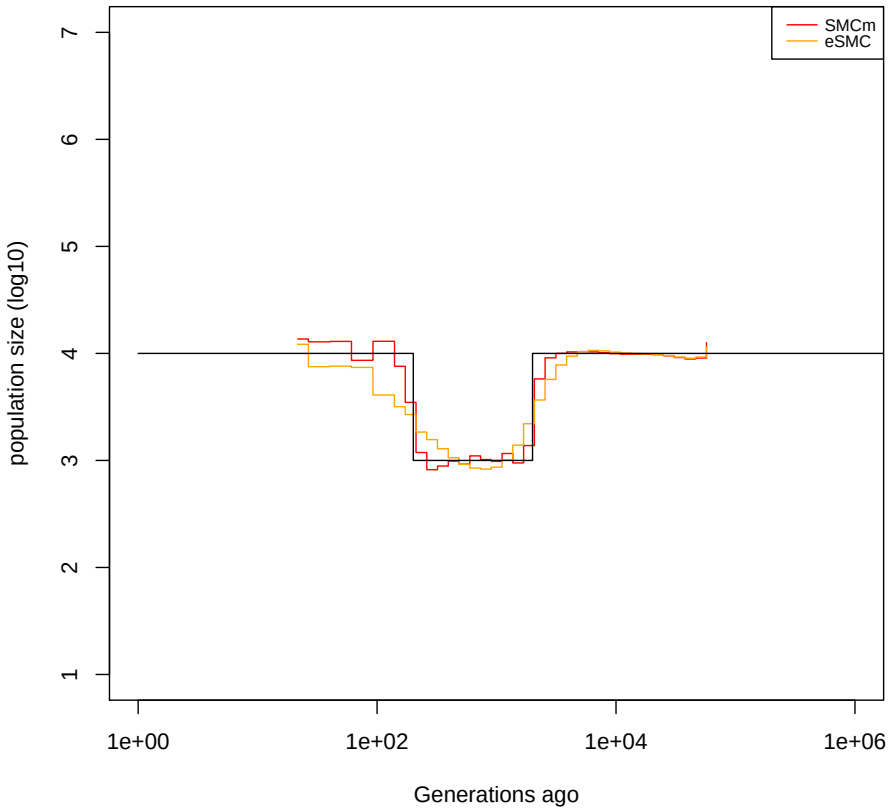

Figure 5.2: **Performance of eSMC and SMCm under a bottleneck scenario.** Estimated demographic history by SMCm and eSMC using 10 scaffolds each of 100 Mb with sample size 2 (respectively in red and orange) under a recent bottleneck (black). The recombination and mutation rate are set to 1×10^{-8} per generation per bp and the methylation and demethylation rate are respectively set to to 1×10^{-4} and 5×10^{-4} per generation per bp.

Results of eSMC and SMCm under a recent bottleneck with $\frac{\rho}{\theta} = 10$ are displayed in Figure 5.3. SMCm correctly infers the population size variation. But eSMC fails to infer the sudden variation of population size and overestimates the population size in recent times. SMCm estimates on average over ten repetitions the ratio $\frac{\rho^*}{\theta} = 9.18$ with a coefficient of variation of 0.041. Yet, eSMC estimates on average $\frac{\rho^*}{\theta} = 7.26$ with coefficient of variation of 0.12.

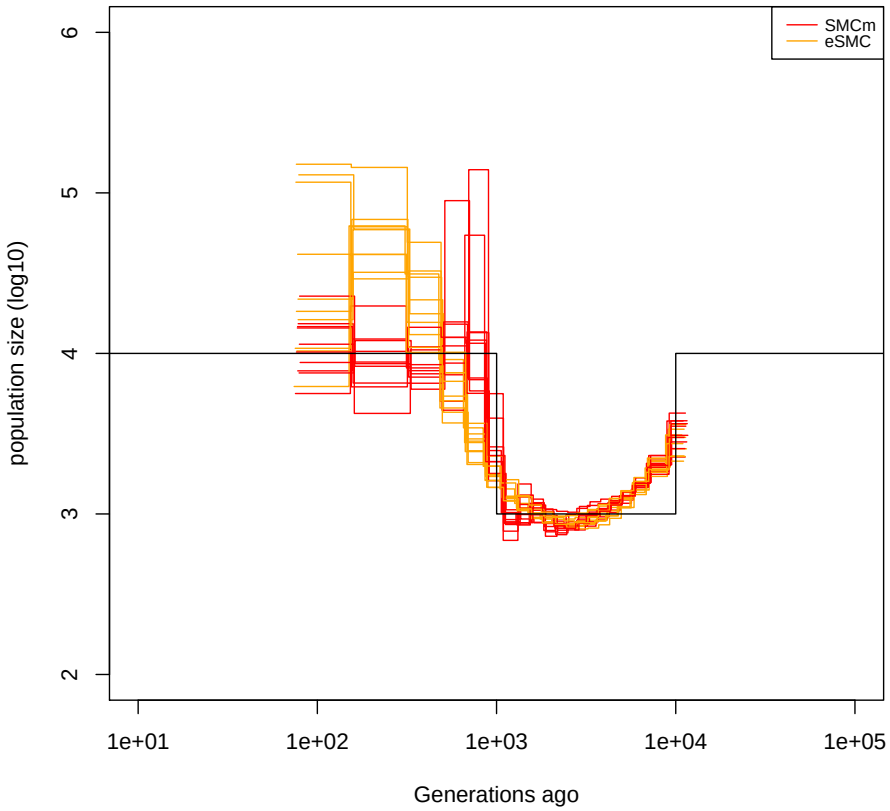

Figure 5.3: **Performance of eSMC and SMCm under a bottleneck scenario.** Estimated demographic history by SMCm and eSMC using simulated sequences of 100 Mb with sample size 2 (respectively in red and orange) under a bottleneck (black). The recombination is set to 1×10^{-7} per generation per bp, the mutation rate to 1×10^{-8} and the methylation and demethylation rate are respectively set to to 1×10^{-4} and 5×10^{-4} per generation per bp.

5.4 Discussion

Throughout this study, we demonstrated the effect of integrating epimutations in the Pairwise Sequentially Markovian Coalescent. Accounting for epimutations increases inferences accuracy in recent times. SMCm outperforms in recent times

any other methods given the amount of data that is used [156, 168]. This indicates that with the use of epimutations, SMCm can better infer hidden states in recent times. However, estimates in the far past seem to be unaffected by the integration of epimutations. Thus, epimutations cannot be used in the hope of increasing analysis accuracy estimating parameters too far in the past (expected due to homoplasy). Obtained results demonstrate that SMCm can also be used to increase the accuracy of inferences in species presenting small diversity (*e.g.* in populations with a small mutation rate or presenting a small effective population size). Yet, additional analysis to test the effect and presence of selection on inferences are required since signatures of selection in *Arabidopsis thaliana* have been found [120] and similar bias found in [148] are expected. We show that adding and accounting for epimutations should only impact inferences in recent times, thus if the discrepancy in the far past is observed when accounting or not for epimutations, it could suggest the presence of unaccounted phenomena (such as selection).

However, our analyses suggest that the increase in accuracy can depend on the rate of methylation and demethylation. As those rates can take many different values from one species to another [173], leading to variation of the ratio of methylation and demethylation. Thus, more analyses are necessary to understand and comprehend the effect of methylation and demethylation rate on inferences. In addition, the effect of methylation and demethylation rate has to be interpreted in the light of the mutation rate, since the effect of integrating epimutations might strongly depend on the ratio of epimutation and mutation speed. Furthermore, SMCm results using small data sets can display high variances, much more than when ignoring epimutations. Hence, SMCm requires larger data sets than eSMC to perform estimations with low variance. Where 100 Mb seems to be a reasonable amount of data to infer demographic history with eSMC (or MSMC2) [150], 1 Gb seems to be more suited for SMCm.

One of the main strengths of integrating epimutations is to overcome issues originating from the recombination rate pointed in [168, 151, 6, 156]. We demonstrate that integrating epimutations helps to correctly infer the recombination rate when the recombination rate is higher than the mutations rate. Based on [151], epimutations could thus help to better measure biological traits where mutations alone would fail. In addition, integrated into iSMC [6], epimutations could potentially help to unveil recombination maps in species presenting a high recombination rate or small diversity. More interestingly, integrating epimutations can lead to correct past demographic inferences although the recombination rate is higher than the mutation rate. This results is of major importance since previous methods presented performance limited by the ratio recombination over mutation [168, 151, 156, 150].

Thus, trustful inferences can now be obtained independently from the ratio of recombination over mutation, opening doors for reliable inferences for many different species.

As an exponential and sudden variation of population size can have different ecological interpretations, it is crucial to be able to distinguish them, which is not possible with current methods [158, 9, 150]. However, we show that integrating epimutations helps to detect a sudden variation of population size and distinguish them from exponential variation. Our results tend more to best-case convergence described in [150, 71] than with sequence polymorphism data alone. In addition, the SMC theory seems to smooth over time the estimations of migration [176]. Integrating epimutations in MSMC-IM could potentially help to better infer migrations events.

At last and most importantly, we deliver a practical demonstration that there is more information in the genome that can be extracted. No current method outperforms the best-case convergence presented in [71, 150]. This implies that current models fail to correctly infer coalescence time along the sequence. Hence, simultaneously integrating epimutations and mutations increases the model's capacity to recover the coalescence time, especially in recent times which can remain problematic [92, 156]. However, results can still be improved as only a fraction of the genome information is currently accounted for. Any heritable mechanism which is currently ignored in analyses could potentially be integrated (*e.g.* insertions and deletions) to improve the model accuracy uncovering new and more reliable estimations. However, this requires to correctly model all the accounted underlying molecular processes in a coherent and unified theory, forcing us to deeper understand the mechanism of genome evolution.

Chapter 6

General discussion and conclusions

6.1 Summary

In this thesis, I focus on simultaneously inferring species history and species specific traits from whole genome sequence data. The current state of the art methods to infer history rely on the Sequentially Markovian Coalescent (SMC) approximation. Yet, these methods are designed for hominid species and make assumptions (*e.g.* Wright-Fisher Model) which are violated in other species or even in humans. I here study if current methods based on the SMC can be extended or complemented to infer and account for species specific traits.

By studying the convergence properties of the Sequentially Markovian Coalescent (SMC) and methods that derive from it, I find that the demographic history of populations that do not undergo a rapid and strong variation of population size can be correctly recovered. Yet, my results also show that even without violating the model's assumptions, some scenarios can never be recovered. This is because the performance of SMC methods relies on their capacity to recover and infer the ancestral recombination graph, which I find to be linked to the ratio of recombinations over mutations ($\frac{\rho}{\theta}$). The higher this ratio, the poorer the inference. Results show that none of the current methods based on the SMC can deliver correct inferences when the ratio value is greater than one (*i.e.* $\frac{\rho}{\theta} > 1$), which seems to be the ratio value in many species. In addition when I test the effect of problems in the data sets (*i.e.* hypothesis violation), I observe drastic biases in the inferences. I find that a non-constant mutation rate along the sequence has the strongest consequence on inferences. Errors in the SNP calling introducing spurious SNPs can also

lead to a strong bias on inferences (contrary to missing SNPs). From the results, I, therefore, recommend stringent filtering and, if possible, high quality reference genomes and sequence data. Lastly, I find that transposable elements can lead to the spurious signature of population size decrease, yet if they are detected and masked, they will not affect the inferences.

Since the accuracy of current SMC methods relies on $\frac{\rho}{\theta}$, I integrated epimutations (*i.e.* methylation and demethylation) in the signal (*i.e.* input data) to increase the amount of information with which to recover the genealogy. My results show that the correct past demographic history and $\frac{\rho}{\theta}$ can be inferred even when $\frac{\rho}{\theta} > 1$. When integrating epimutations, results are accurate enough to distinguish sudden from smooth population size variation, offering a more appropriate interpretation of what changes the population has undergone. Yet, not all species respect the underlying necessary hypothesis (*e.g.*mammals). However, many other genome features are ignored by current methods (*e.g.* insertions and deletions) but if they are correctly integrated and accounted for an increase in accuracy can be obtained. In addition, integrating more genome features in the signal not only strongly improves the inference of the ancestral recombination graph but also leads to a more coherent model describing genome evolution.

When integrating self-fertilization and dormancy, I find they have similar effects as a re-scaling of the coalescence and recombination rate. They can thus be inferred through the comparison of the observed ratio of recombination over mutation ($\frac{\rho}{\theta}$) and the expected one based on the known mutation and recombination rates ($\frac{r}{\mu}$). Therefore my model explains the discrepancy one can observe between $\frac{\rho}{\theta}$ and $\frac{r}{\mu}$, which before remained unexplained and undiscussed. When applied to data, results show dormancy in *Daphnia pulex* and self-fertilization in *Arabidopsis thaliana* at the expected rates. However, my approach alone is not sufficient to distinguish self-fertilization from dormancy and additional experiments, analyses, or prior knowledge are required to distinguish between them. In addition, other biological traits can affect the ratio $\frac{\rho}{\theta}$ (*e.g.* occurrence of clonal events). Hence, the observed $\frac{\rho}{\theta}$ is a result of many biological mechanisms that need to be accounted for to interpret $\frac{\rho}{\theta}$. I currently assume self-fertilization and the dormancy rate to be constant in time, which might not be true [55]. The results of Barroso *et al.* [6] suggest that detecting a variation of $\frac{\rho}{\theta}$ in time is theoretically feasible (under the assumption that the recombination rate is constant along the genome). I derived such a theoretical framework and have implemented it, data analysis remains to be done to demonstrate the accuracy of my approach. Also, as my model assumes the absence of population structure, supplementary studies are required to understand the effect of admixture or migration on the observed $\frac{\rho}{\theta}$. Furthermore,

self-fertilization or dormancy being independent of migration might not be true and will have to be correctly modeled according to species biology (*e.g.* discrepancies between pollen and seed migration rate). As a result, not correctly accounting for self-fertilization or dormancy could lead to erroneous admixture inferences.

In some species, individuals can produce a large number of offspring (in comparison to the population size), increasing the variance in the number of descendants between individuals, as a consequence the life cycle cannot be described by a Wright-Fisher model (*e.g.* in fungi, bivalve mollusks and in some fish species). As a result, the genealogy cannot be described by Kingman coalescent (even a re-scaled one). In this case, the genealogy can be star-shaped (more than two individuals can simultaneously coalesce). I built and implemented a Sequentially Markovian Coalescent model based on the Beta coalescent. In this model, the probability of being "star-shaped" depends on a parameter α which is inferred. I find my model capable of accounting and measuring the probability of the genealogy being star-shaped but does not deliver the "ecological origin" of the star-shaped genealogy (*i.e.* the interpretation of the inferred ancestral recombination graph belongs to the user). Additional analysis of the linkage disequilibrium and knowledge on offspring production in the species are required to distinguish between strong selection, strong bottlenecks, and neutral large variance in offspring production.

6.2 General Discussion

As selection can strongly bias demographic inferences, which are themselves necessary for selection scans, there is a current lack of methodology to account for selection when inferring past demographic events. Since multiple merger events can describe strong positive selection (directional selection) because all individuals will rapidly inherit the advantageous locus and thus simultaneously coalesce to a common ancestor, from the SMβC I am building a model similar to iSMC. Instead of having the recombination rate varying along the sequence, the alpha parameters (*i.e.* probability of a star shaped genealogy) would now become a hidden state. From this model, one could either detect potential regions under selection and remove them to correct demographic inferences (cf chapter 3), or directly account for them while simultaneously inferring the rate at which selection affects linkage disequilibrium. As additional analyses, regions detected or suspected to have a small α parameter (*i.e.* potentially being under selection) can be compared to existing results of selection scans [133] to test the coherence of the methodology. These regions can also be studied more closely and one can check if they correspond to

coding or promoting regions. Furthermore, the strength of selection could be measured or interpreted through the α estimated. Finally, the α parameter could be explained in the light of the simultaneously inferred coalescence time. More precisely, a small α (*i.e.* star shaped genealogy) with a small coalescence time could indicate a region currently under selection. Yet a small α in regions displaying high coalescence time could suggest a region that was under selection at some point in the species history (but no longer is).

Furthermore, selection can originate from many different and non-independent mechanisms (*e.g.* polygenic selection, background selection, fluctuating selection, or balancing selection) which we still fail to fully model and infer. More precisely each type of selection might affect the local $\frac{\rho}{\theta}$, in a different way. For example background selection could decrease the number of observed SNPs (*i.e.* increase $\frac{\rho}{\theta}$) whereas balancing selection could maintain high diversity (*i.e.* decrease $\frac{\rho}{\theta}$). Distinguishing the different mechanisms of selection can be a difficult task but of major importance since these mechanisms have very different ecological or biological interpretations. By building more complex theoretical models describing the topology of the ancestral recombination graph under the different selection models, one could predict and better understand the specific genomic signature of each selection mechanism. Such powerful theoretical results could help us detect and select the correct selection model or mechanism observed (locally or not) in the genome. Yet, the complexity of such analyses could be such that many different scenarios could potentially explain the observed data (*i.e.* identifiability issue). Hence, additional analyses based on different statistics or approaches (*e.g.* based on other population genetics approaches [124]) to select one model from another will be required. New theoretical frameworks might arise (*e.g.* Phase-type theory [82], already applied to balancing selection [186]), offering new tools to solve tedious modeling problems. Integrating more data could also be a solution, or at least a part of the solution as we will discuss below.

One aspect of genome data that is currently not explored is the difference in the information that can be extracted from the genome. Genomes are shaped by many complex molecular and evolutionary forces. Assuming the molecular forces are well understood, then the discrepancy from what is inferred using different genome features could help infer and uncover evolutionary forces. For example, one could use only epimutations to infer past demographic history or only use sequence polymorphism data, a discrepancy between the two inferences could suggest an underlying selection mechanism affecting at least one of the input data sets (*i.e.* genome and/or methylome). Otherwise, inferring similar scenarios could support the neutral hypothesis. By integrating more genomic characteristics in analyses

(*e.g.* microsatellites repetition polymorphism, copy number variation of genes, transposable elements presence/absence polymorphism), and by assuming them being independent or not, new scenarios could be unveiled and help in unveiling specific selection mechanisms.

Integrating more genomic characteristics might require changing time scales, as all molecular mechanisms do not occur at the same rate. For example, complex chromosomic rearrangements are assumed to not occur at the "coalescence" time scale. However, they can occur at the phylogeny scales. Inspired by the work of Marin *et al.* [110], one could build an approach inspired by the SMC (or a more relevant model) to infer the evolutionary history of species along the genome. This approach could also the comparison of species genomes, focusing on genomic discrepancies relevant at the phylogeny time scales. Such an approach, most likely based on the Multi-Species Coalescent [31, 80, 26] (MSC), could complement the work by Marin et al [110], unveiling more complex evolutionary scenarios, giving a new interpretation of species trees and gene genealogies, and deepening our understanding of genome evolution on longer time scales. However, inferences could require strong hypotheses such as nonvariation of molecular rates in time and along the genome, neutrality, or prior knowledge on the mechanisms driving genome evolution at this time scale. Besides, the differences in ecological traits, generation time, and reproduction mechanisms between species might make such analyses unrealistic. Yet overcoming all these issues could render feasible the inference of ecological/biological traits or even genome of ancestral species, a major achievement in the field of biology. Most likely, this approach would only be relevant when analyzing sisters species or clades of recently diverged species (*e.g.* The Tomato clade [134])

Modeling and inferring inter-specific coevolution from genomes remains challenging. However, assuming a system of non-independent species (*e.g.* species belonging to a common trophic network within an ecosystem [18]), one could predict the effect of interacting species on the topology of the ancestral recombination graph (*i.e.* the distribution of genealogies) of each species. Based on those theoretical results, one could study and test species being independent or coevolving by studying the correlation of their ancestral recombination graph or by describing the ancestral recombination graph of a species conditioned to the ancestral recombination graph of the other species. However, similar analyses were made and results are very noisy as shown in [75], thus accurate and coherent modeling with robust statistics are still required to make such approaches relevant. This approach would require integrating the specific biological traits of each species (as mentioned above), which is currently not an easy task, potentially explaining part

of the observed noise.

6.3 Conclusion

As a conclusion, ecological traits can be detected and measured if correctly modeled from genome sequence data. With the development of new theoretical frameworks, it will become possible to simultaneously integrate ecological traits, the variation of population size, admixture, and selection in a unified and coherent model. In order to recover model parameters, I showed that the signal can be enhanced by integrating more genomic characteristics than just sequence polymorphism data. Integrating new genomic characteristics will lead to increasing the time frame in which inferences can be made, potentially going beyond the field of population genetics and reaching the field of phylogenetics.

Appendix A

Appendix

A.1 Appendix of Chapter 2

A.1.1 Supplementary Figure

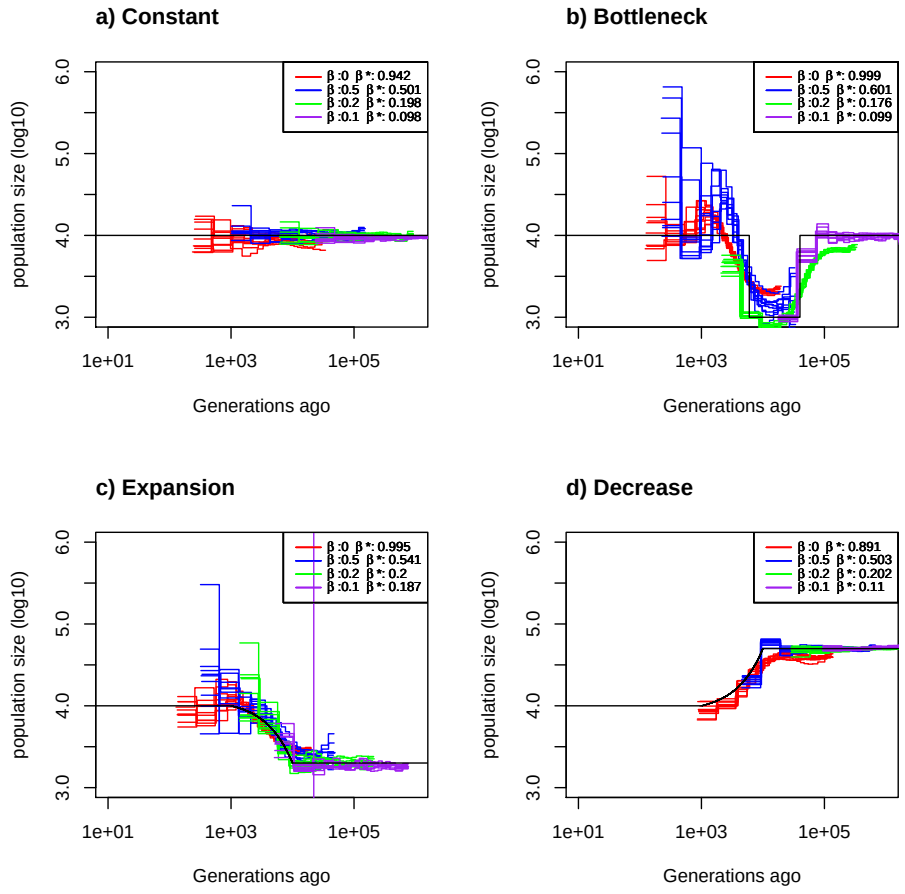

Figure A.1: **Estimated demographic history in four simple demographic scenarios with seed banking.** Estimated demographic history using four simulated sequences of 10 Mb under four different demographic scenarios with 10 replicates. Mutation and recombination rate are set to 1.25×10^{-8} per generation per bp. Simulation were done under four different germination rate β. We have $\beta = 1$ (red), 0.5 (blue), 0.2 (green) and 0.1 (purple). Therefore $\frac{r}{\mu} = 1$ and we respectively have $\frac{\rho}{\theta} = 1$,$\frac{\rho}{\theta} = 0.5$,$\frac{\rho}{\theta} = 0.2$ and $\frac{\rho}{\theta} = 0.1$. The simulated demographic history is represented in black. a) Demographic history simulated under a constant population size. b) Demographic history simulated under a bottleneck. c) Demographic history simulated under an expansion. d) Demographic history simulated under a decrease. In addition we simulated data under four different germination rate β. β* equal the estimated germination rate.

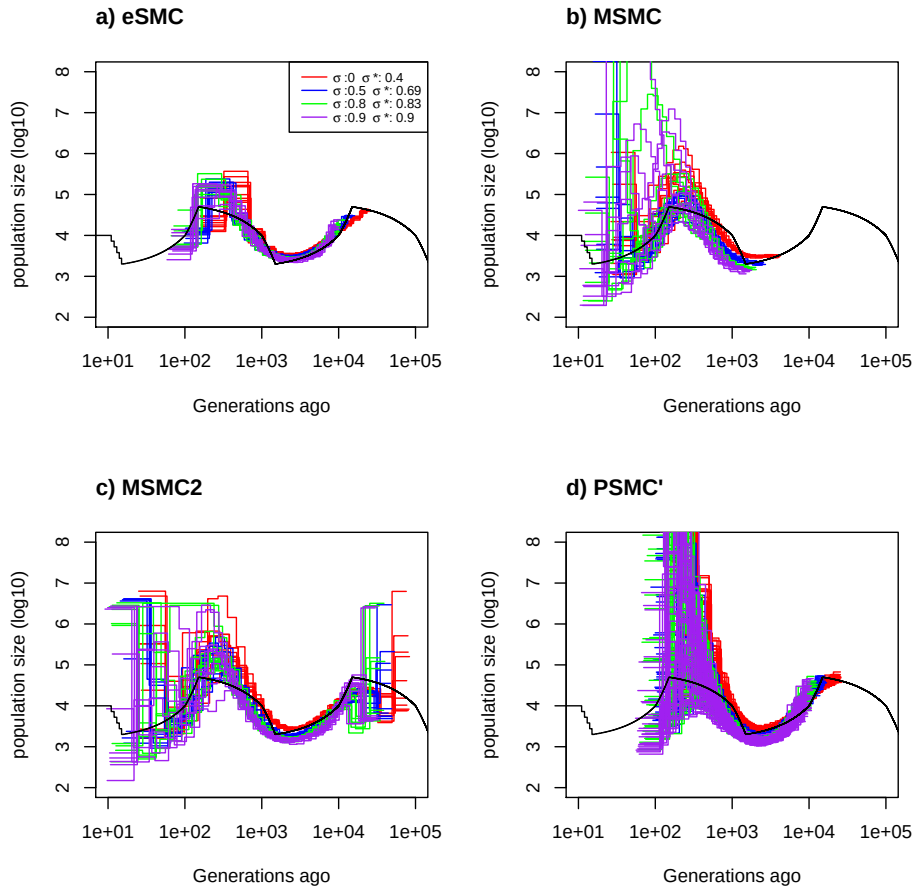

Figure A.2: **Estimated demographic history with selfing under $\frac{r}{\mu} = 5$.**Estimated demographic history using four simulated sequences of 10 Mb and ten replicates under a sawtooth demographic scenario (black). Simulation were done under four different self-fertilization rate σ (0,0.5,0.8 and 0.9). The mutation is set to 1.25×10^{-8} and the recombination rate to 6.25×10^{-8} per generation per bp. Therefore $\frac{r}{\mu} = 5$ and respectively $\frac{\rho}{\theta} = 5$,$\frac{\rho}{\theta} = 2.5$,$\frac{\rho}{\theta} = 1$ and $\frac{\rho}{\theta} = 0.5$. Estimated demographic history are represented for all tested self-fertilization, $\sigma = 1$ (red), 0.5 (blue), 0.2 (green) and 0.1 (purple). The demographic history is estimated using a) eSMC where σ* equals the estimated self-fertilization rate, b) MSMC, c) MSMC2 and d) PSMC'.

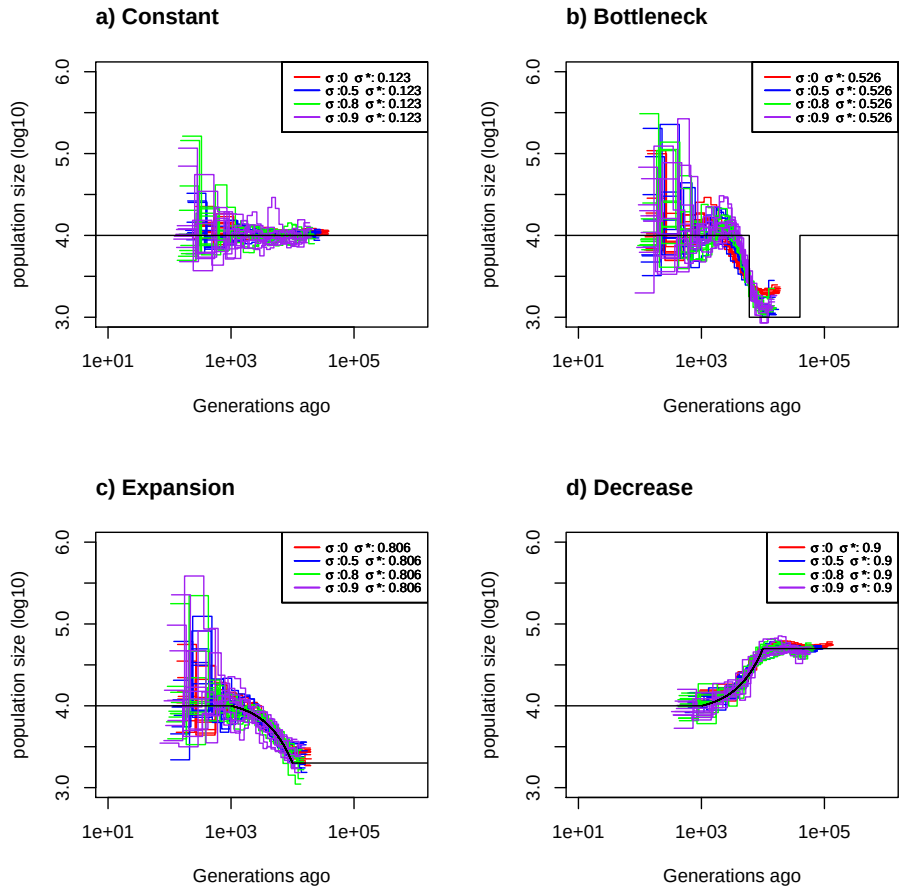

Figure A.3: **Estimated demographic history in four simple demographic scenarios with selfing.** Estimated demographic history using four simulated sequences of 10 Mb under four different demographic scenarios with 10 replicates. Mutation and recombination rate are set to 1.25×10^{-8} per generation per bp. Simulation were done under four different self-fertilization rate σ (0,0.5,0.8 and 0.9). Therefore $\frac{r}{\mu} = 1$ and respectively $\frac{\rho}{\theta} = 1$,$\frac{\rho}{\theta} = 0.667$,$\frac{\rho}{\theta} = 0.333$ and $\frac{\rho}{\theta} = 0.182$. The simulated demographic history is represented in black. a) Demographic history simulated under a constant population size. b) Demographic history simulated under a bottleneck. c) Demographic history simulated under an expansion. d) Demographic history simulated under a decrease. In addition we simulated data under four different self-fertilization rate σ . We have $\sigma = 0$ (red), 0.5 (blue), 0.8 (green) and 0.9 (purple). σ* equal the estimated self-fertilization rate.

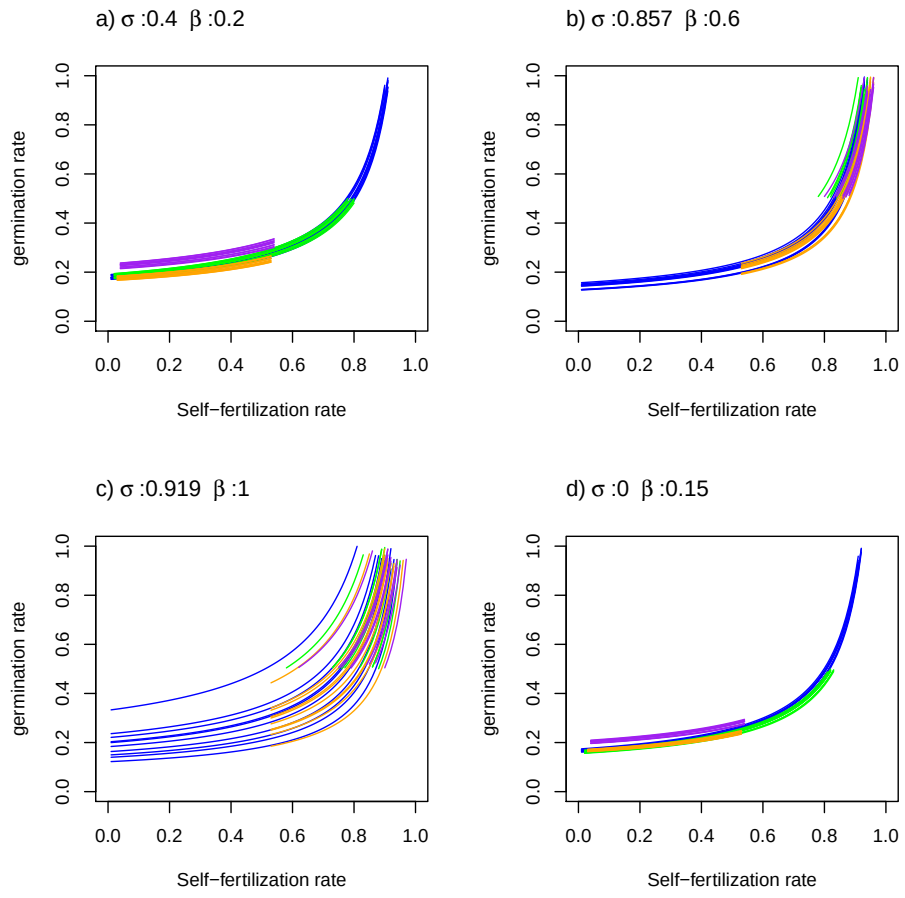

Figure A.4: **Possible selfing and seed banking value where $\frac{r}{\mu} = 1$.** Possible estimated self-fertilization and germination rates because of confounding effect using four simulated sequences of 10 Mb under a sawtooth demographic scenario and four different combinations of germination (b) and self-fertilization (s) rate but resulting in the same $\frac{\rho}{\theta} = 0.15$. Mutation rate is set to 1.25×10^{-8} and recombination rate to 1.25×10^{-8} per generation per bp. Therefore $\frac{r}{\mu} = 1$. The four combination are : a) $\sigma = 0.4$ and $\beta = 0.2$, b) $\sigma = 0.857$ and $\beta = 0.6$, c) $\sigma = 0.919$ and $\beta = 1$ and d) $\sigma = 0$ and $\beta = 0.15$. Hence, for each scenario $\frac{\rho}{\theta} = 0.15$ For each combination of β and σ, eSMC was launched with five different prior settings: ignoring seed banks and self-fertilization (red), accounting for seed banks and self-fertilization but without setting priors (blue), accounting for seed banks and self-fertilization with a prior set only for the self-fertilization rate (green), only for the germination rate (orange) or for both (purple).

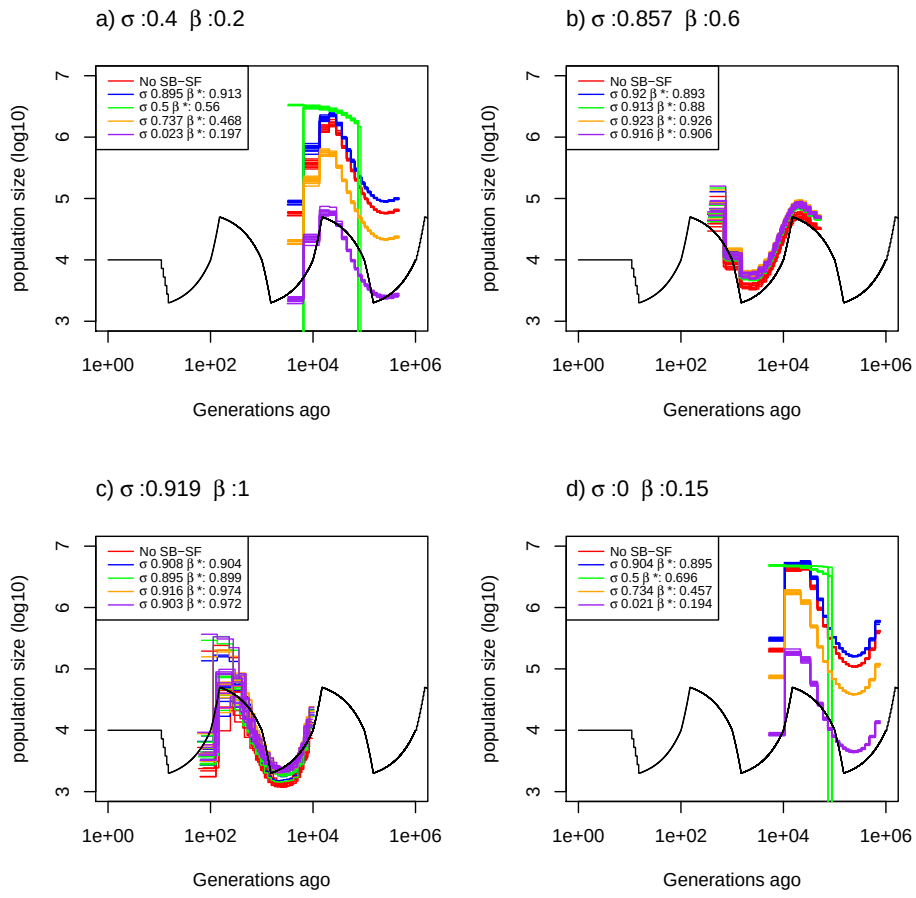

Figure A.5: **Estimated demographic history with selfing and seed banking where $\frac{r}{\mu} = 6.667$.** Demographic history estimated by eSMC for ten replicates using four simulated sequences of 10 Mb under a sawtooth demographic scenario and four different combinations of germination (b) and self-fertilization (s) rate but resulting in the same $\frac{\rho}{\theta} = 1$. Mutation rate is set to 1.25×10^{-8} and recombination rate to 8.335×10^{-8} per generation per bp. Therefore $\frac{r}{\mu} = 6.67$. The four combination are : a) $\sigma = 0.4$ and $\beta = 0.25$, b) $\sigma = 0.75$ and $\beta = 0.6$, c) $\sigma = 0.85$ and $\beta = 1$ and d) $\sigma = 0$ and $\beta = 0.15$. Hence, for each scenario $\frac{\rho}{\theta} = 1$ For each combination of β and σ, eSMC was launched with five different prior settings: ignoring seed banks and self-fertilization (red), accounting for seed banks and self-fertilization but without setting priors (blue), accounting for seed banks and self-fertilization with a prior set only for the self-fertilization rate (green), only for the germination rate (orange) or for both (purple). σ* and β* respectively represent the estimated self-fertilization and germination rate.

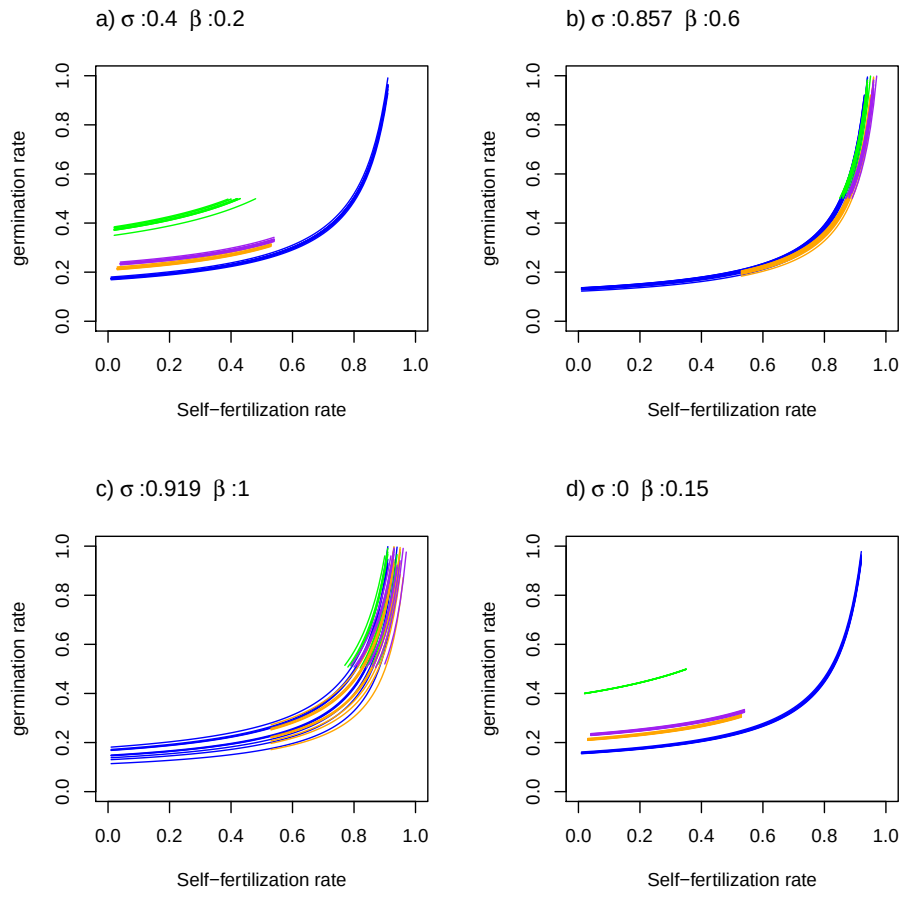

Figure A.6: **Possible selfing and seed banking value where $\frac{r}{\mu} = 6.667$.** Possible estimated self-fertilization and germination rates because of confounding effect using four simulated sequences of 10 Mb under a sawtooth demographic scenario and four different combinations of germination (b) and self-fertilization (s) rate but resulting in the same $\frac{\rho}{\theta} = 1$. Mutation rate is set to 1.25×10^{-8} and recombination rate to 8.335×10^{-8} per generation per bp. Therefore $\frac{r}{\mu} = 1$. The four combination are : a) $\sigma = 0.4$ and $\beta = 0.2$, b) $\sigma = 0.857$ and $\beta = 0.6$, c) $\sigma = 0.919$ and $\beta = 1$ and d) $\sigma = 0$ and $\beta = 0.15$. Hence, for each scenario $\frac{\rho}{\theta} = 1$ For each combination of β and σ, eSMC was launched with five different prior settings: ignoring seed banks and self-fertilization (red), accounting for seed banks and self-fertilization but without setting priors (blue), accounting for seed banks and self-fertilization with a prior set only for the self-fertilization rate (green), only for the germination rate (orange) or for both (purple).

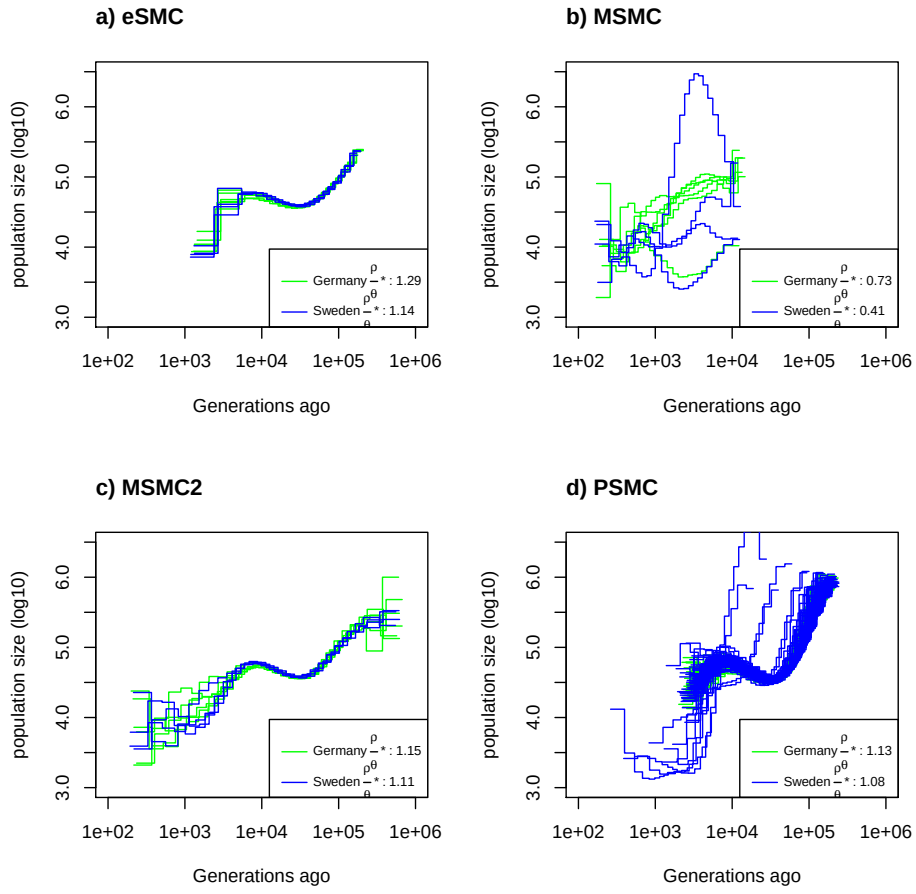

Figure A.7: **Estimated demographic history of *Arabidopsis thaliana* where selfing and seed banking is ignored.** Demographic history of two European (Sweden (blue) and German (green)) populations of *A. thaliana*. Mutation rate is set to 7×10^{-9} per generation per bp and was use as prior for recombination rate. a) Demographic history estimated by eSMC without accounting self-fertilzation or dormancy. b) Demographic history estimated by MSMC. c) Demographic history estimated by MSMC2 . d) Demographic history estimated by PSMC'.

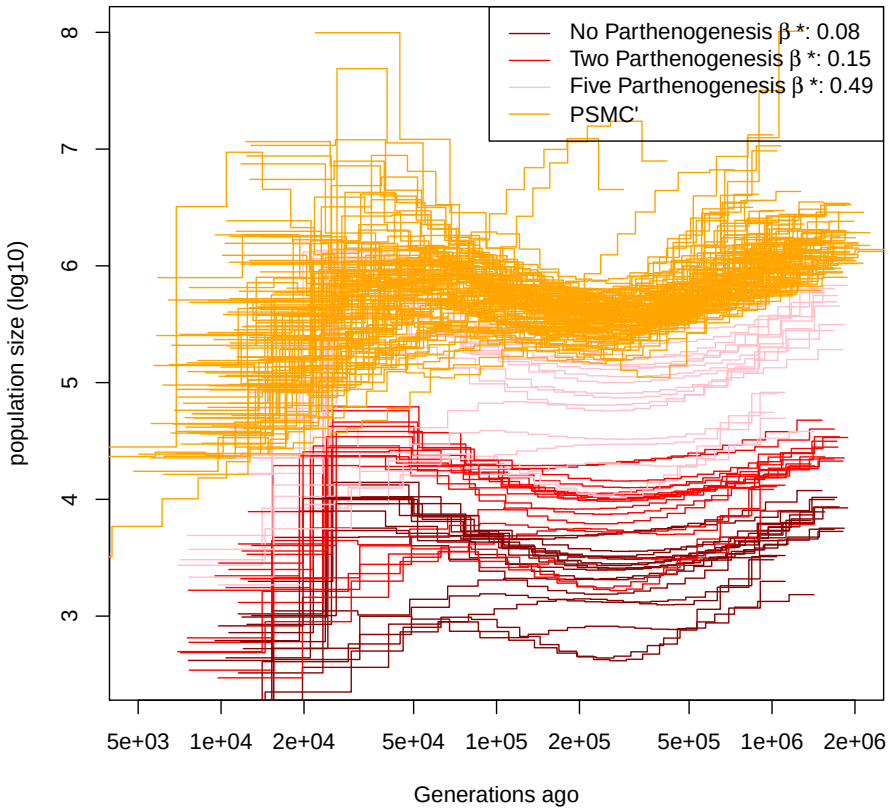

Figure A.8: **Estimated demographic history of *Daphnia pulex*.** Demographic history estimated by eSMC per scaffolds on six individuals of *D. pulex* accounting for egg-banks (β is a variable and $\sigma = 0$). Different assumptions concerning the number of parthenogenetic cycles before the production of the dormant egg are made: Five cycles (pink), two cycles (red) and no parthenogenesis (dark red). Demographic history estimated by PSMC' are ploted in organge. Mutation and recombination rates are respectively set to 4.33×10^{-9} and $\frac{8 \times 10^{-8}}{n_p}$ per generation per bp, where n_p is the number of reproductive cycles per year, parthenogenetic and sexual.

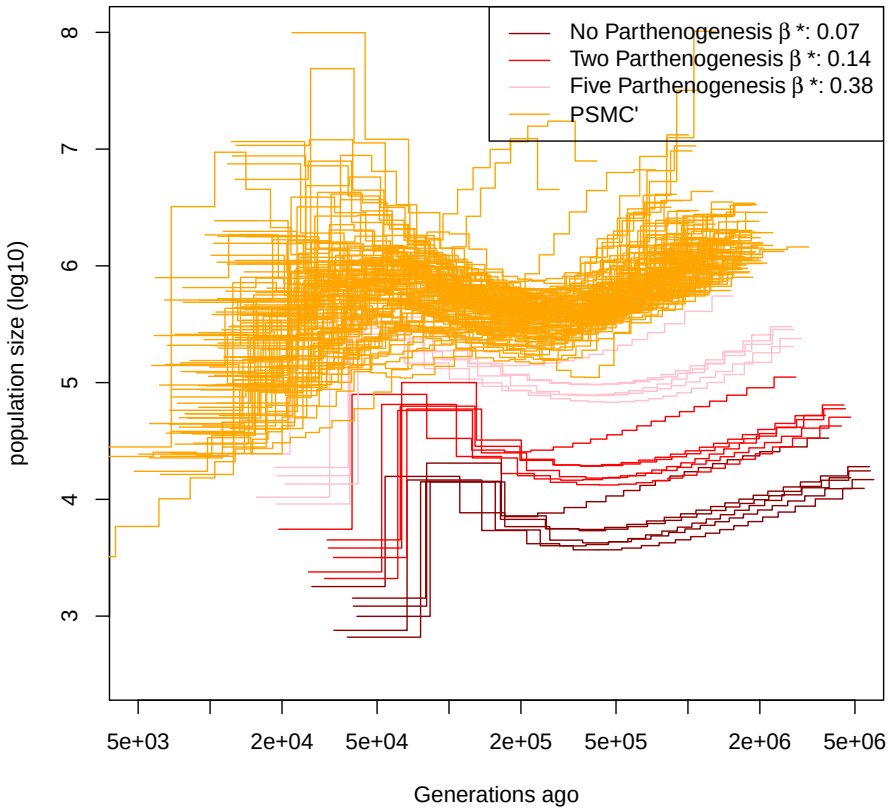

Figure A.9: **Estimated demographic history of *Daphnia pulex*.** Demographic history estimated by eSMC on six individuals of *D. pulex* accounting for egg-banks assuming mutations accumulates 5 times slower during egg stage and active stage (β is a variable and $\sigma = 0$). Different assumptions concerning the number of parthenogenetic cycles before the production of the dormant egg are made: Five cycles (pink), two cycles (red) and no parthenogenesis (dark red). Mutation and recombination rates are respectively set to 4.33×10^{-9} and $\frac{8\times10^{-8}}{n_p}$ per generation per bp, where n_p is the number of reproductive cycles per year, parthenogenetic and sexual.

A.1.2 Model description of eSMC

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

- Hidden States
- The signal (observed data)
- A Transition matrix (Probability of jumping from one state to another)
- An Emission matrix (Probability of observing the data given 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:

- β : the germination rate (expected probability to germinate at every generation, between 0 and 1)
- σ : self fertilization rate (between 0 and 1)
- N_0 : Population at present time
- r : recombination rate per nucleotide per $4N_0$ generations
- μ : Mutation rate per nucleotide per $4N_0$ generations
- u : recombination time (which follows a continuous uniform distribution on the coalescent tree)
- L: Sequence length in bp
- $\rho = r(L-1)$
- $\theta = \mu L$
- N_t : Population size at time t
- χ_t : Scaled population size at time t ($N_t = \chi_t N_0$)

The model's assumptions are :

- Piecewise constant population size
- Infinite site model

- Constant mutation, recombination, germination and self-fertilization rate in time
- Constant mutation and recombination rate along the sequence
- Neutrality
- Wright-Fisher model

Hidden States

We define our hidden states at one position as the coalescent time between the two individuals at that position. We note that coalescent time t (t>0). A transition from a coalescent time s to time t ($t \neq s$) at the next can only occur if a recombination happened in between the two positions.

Observations

Our observation, or the signal, is a sequence of 1s and 0s. This sequence is built from the comparison of two DNA sequences. When going along the sequence, if both nucleotides are similar, then the signal is a 0 (no mutation occurred). If both are different, then a mutation occurred, and the signal is a 1 (Figure 1).

Figure A.10: Schema describing the construction of the signal from phased sequences

Transition Matrix

A transition to state t from state s (t ≠ s) can only occur if there is a recombination event. Assuming recombinations occur along the sequence as a Poisson process,

the recombination probability between two nucleotides is :

$$P(rec|s) = (1 - e^{-\frac{\beta 2(1-\sigma)}{2-\sigma} 2rs}) \tag{A.1}$$

We now Assume that a recombination event occurred at time u (<s) where u follows a uniform distribution between 0 and s. Then three scenarios are possible. Either the new coalescent time t is smaller (t<s), bigger (t>s), or unchanged (t=s). Those scenarios are displayed on Figure 2 A),C) and B).

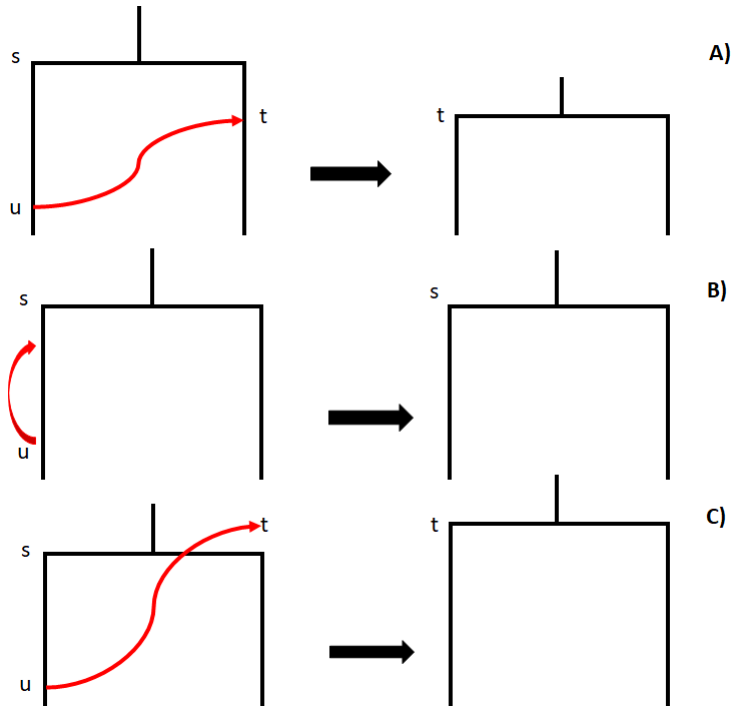

Figure A.11: Schema of the three possible coalescent events after a recombination event

t<s The floating branch resulting from the recombination event coalesces at time t < s . This mean it must not coalesce before time t (including itself). In addition we have u<t. The transition probability is therefore :

$$P(t|s,u) = \frac{2\beta^2}{(2-\sigma)\chi_t}(e^{\int_u^t -\frac{4\beta^2}{(2-\sigma)\chi_v} dv}) \tag{A.2}$$

t=s The floating branch resulting from the recombination event can self coalesce before time t. We therefore have the transition probability :

$$P(s|s,u) = \int_u^s \frac{2\beta^2}{(2-\sigma)\chi_k} e^{\int_u^k -\frac{4\beta^2}{(2-\sigma)\chi_v}dv} dk \tag{A.3}$$

t>s The floating branch resulting from the recombination event must not coalesce (including itself) before time s. Then no coalescent event must happen before time t. We therefore have the transition probability :

$$P(s|s,u) = \frac{2\beta^2}{(2-\sigma)\chi_t} e^{\int_u^s -\frac{4\beta^2}{(2-\sigma)\chi_v}dv} e^{\int_s^t -\frac{2\beta^2}{(2-\sigma)\chi_v}dv} \tag{A.4}$$

Transition probability in continuous time In the end we have :

$$p(t|s,u) = \begin{cases} (1 - e^{-\frac{\beta 2(1-\sigma)}{2-\sigma}2rs}) \frac{2\beta^2}{(2-\sigma)\chi_t} (e^{\int_u^t -\frac{4\beta^2}{(2-\sigma)\chi_v}dv}) & if \quad u < t < s \\ e^{-\frac{\beta 2(1-\sigma)}{2-\sigma}2rs} + (1 - e^{-\frac{\beta 2(1-\sigma)}{2-\sigma}2rs}) \int_u^s \frac{2\beta^2}{(2-\sigma)\chi_k} e^{\int_u^k -\frac{4\beta^2}{(2-\sigma)\chi_v}dv} dk & if \quad t = s \\ (1 - e^{-\frac{\beta 2(1-\sigma)}{2-\sigma}2rs}) \frac{2\beta^2}{(2-\sigma)\chi_t} e^{\int_u^s -\frac{4\beta^2}{(2-\sigma)\chi_v}dv} e^{\int_s^t -\frac{2\beta^2}{(2-\sigma)\chi_v}dv} & if \quad t > s \\ 0 & if \quad otherwise \end{cases} \tag{A.5}$$

Once again, if $\beta = 1$ and $\sigma = 0$, we fall back on the probability from PSMC'.

One can find $p(t|s)$ using the total probability formula which is:

$$p(t|s) = \int_0^s \frac{1}{s} p(t|s,u) du \tag{A.6}$$

As explained before, the hidden state space is finite. We therefore discretize time in n 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 :

$$T_\alpha = -\frac{(2-\sigma)}{2\beta^2} \ln(1 - \frac{\alpha}{n}) \tag{A.7}$$

We therefore have:

$$p(\alpha|s) = \int_{T_\alpha}^{T_{\alpha+1}} p(t|s) dt \tag{A.8}$$

The transition matrix needs is the probability of going from one state to another. Therefore, we need the probability of the coalescent time at the previous position (which is here s), which belongs to the state γ. To do this we simply integrate s over the time interval γ, equivalent to replacing s by the expected coalescent time : t_γ.

Initial Probability We use the equilibrium probability as initial probability. The equilibrium probability is the probability that the coalescent, at the first position, happens in each time interval and is thus given by :

$$\begin{aligned} q_o(\alpha) &= \int_{T_\alpha}^{T_{\alpha+1}} \frac{2\beta^2}{(2-\sigma)\chi_\alpha} e^{\int_0^t \frac{-2\beta^2}{(2-\sigma)\chi_v} dv} dt \\ q_o(\alpha) &= \int_{T_\alpha}^{T_{\alpha+1}} \frac{2\beta^2}{(2-\sigma)\chi_\alpha} e^{\int_0^{T_\alpha} \frac{-2\beta^2}{(2-\sigma)\chi_v} dv} e^{\int_{T_\alpha}^t \frac{-2\beta^2}{(2-\sigma)\chi_v} dv} dt \\ q_o(\alpha) &= e^{\int_0^{T_\alpha} \frac{-2\beta^2}{(2-\sigma)\chi_v} dv} \int_{T_\alpha}^{T_{\alpha+1}} \frac{2\beta^2}{(2-\sigma)\chi_\alpha} e^{\frac{-2\beta^2(t-T_\alpha)}{(2-\sigma)\chi_\alpha}} dt \\ q_o(\alpha) &= e^{\sum_{\eta=0}^{\alpha-1} \frac{-2\beta^2}{(2-\sigma)\chi_\eta} \Delta_\eta} (1 - e^{\frac{-2\beta^2 \Delta_\alpha}{(2-\sigma)\chi_\alpha}}) \end{aligned} \tag{A.9}$$

Calculation of t_γ

$$\begin{aligned} t_\gamma = E[\text{Coalescent time}|\gamma] &= \frac{E[\text{Coalescent time} \cap \gamma]}{P(\gamma)} = \frac{\int_{T_\gamma}^{T_{\gamma+1}} t\Lambda_\gamma e^{-\int_0^t \Lambda_v dv} dt}{q_0(\gamma)} \\ &= \frac{\Lambda_\gamma \int_{T_\gamma}^{T_{\gamma+1}} t e^{-\int_0^{T_\gamma} \Lambda_v dv} e^{-\int_{T_\gamma}^t \Lambda_v dv} dt}{q_0(\gamma)} = \frac{\Lambda_\gamma e^{-\int_0^{T_\gamma} \Lambda_v dv} \int_{T_\gamma}^{T_{\gamma+1}} t e^{-\int_{T_\gamma}^t \Lambda_v dv} dt}{q_0(\gamma)} \\ &= \frac{\Lambda_\gamma \int_{T_\gamma}^{T_{\gamma+1}} t e^{(T_\gamma - t)\Lambda_\gamma} dt}{(1 - e^{-\Delta_\gamma \Lambda_\gamma})} = \frac{T_\gamma - T_{\gamma+1} e^{-\Delta_\gamma \Lambda_\gamma}}{(1 - e^{-\Delta_\gamma \Lambda_\gamma})} + \frac{\int_{T_\gamma}^{T_{\gamma+1}} e^{(T_\gamma - t)\Lambda_\gamma} dt}{(1 - e^{-\Delta_\gamma \Lambda_\gamma})} \\ &= \frac{T_\gamma - T_{\gamma+1} e^{-\Delta_\gamma \Lambda_\gamma}}{(1 - e^{-\Delta_\gamma \Lambda_\gamma})} + \frac{(1 - e^{-\Delta_\gamma \Lambda_\gamma})}{\Lambda_\gamma (1 - e^{-\Delta_\gamma \Lambda_\gamma})} = \frac{T_\gamma - T_{\gamma+1} e^{-\Delta_\gamma \Lambda_\gamma}}{(1 - e^{-\Delta_\gamma \Lambda_\gamma})} + \frac{1}{\Lambda_\gamma} \end{aligned} \tag{A.10}$$

Where :

$$\begin{aligned} \Delta_\gamma &= T_{\gamma+1} - T_\gamma \\ \Lambda_\gamma &= \frac{2\beta^2}{(2-\sigma)\chi_\gamma} \end{aligned} \tag{A.11}$$

Calculation of $p(\alpha|\gamma)$ with $\alpha < \gamma$ We first need $p(t|t_\gamma)$ when $\alpha < \gamma$, which is obtained as described below :

$$\begin{aligned}
p(t|t_\gamma) &= \int_0^t \frac{P_\gamma}{t_\gamma} \frac{2\beta^2}{(2-\sigma)\chi_t} (e^{\int_u^t -\frac{4\beta^2}{(2-\sigma)\chi_v} dv}) du \\
&= \frac{P_\gamma}{t_\gamma} \int_0^t \frac{2\beta^2}{(2-\sigma)\chi_t} (e^{\int_u^t -\frac{4\beta^2}{(2-\sigma)\chi_v} dv}) du \\
&= \frac{P_\gamma}{t_\gamma} (\sum_{\eta=0}^{\alpha-1} \int_{T_\eta}^{T_{\eta+1}} \frac{2\beta^2}{(2-\sigma)\chi_t} (e^{\int_u^t -\frac{4\beta^2}{(2-\sigma)\chi_v} dv}) du \\
&\quad + \int_{T_\alpha}^t \frac{2\beta^2}{(2-\sigma)\chi_t} (e^{\int_u^t -\frac{4\beta^2}{(2-\sigma)\chi_v} dv}) du) \\
&= \frac{P_\gamma}{t_\gamma} (\sum_{\eta=0}^{\alpha-1} \int_{T_\eta}^{T_{\eta+1}} \frac{2\beta^2}{(2-\sigma)\chi_t} (e^{\int_u^{T_{\eta+1}} -\frac{4\beta^2}{(2-\sigma)\chi_v} dv}) (e^{\int_{T_{\eta+1}}^t -\frac{4\beta^2}{(2-\sigma)\chi_v} dv}) du \\
&\quad + \int_{T_\alpha}^t \frac{2\beta^2}{(2-\sigma)\chi_t} (e^{-\frac{(t-u)4\beta^2}{(2-\sigma)\chi_\alpha}}) du) \\
&= \frac{P_\gamma}{t_\gamma} \frac{2\beta^2}{(2-\sigma)\chi_\alpha} (\sum_{\eta=0}^{\alpha-1} e^{\int_{T_{\eta+1}}^t -\frac{4\beta^2}{(2-\sigma)\chi_v} dv} \int_{T_\eta}^{T_{\eta+1}} (e^{-(T_{\eta+1}-u)\frac{4\beta^2}{(2-\sigma)\chi_\eta}}) du \\
&\quad + \int_{T_\alpha}^t e^{-\frac{(t-u)4\beta^2}{(2-\sigma)\chi_\alpha}} du) \\
&= \frac{P_\gamma 2\beta^2}{t_\gamma (2-\sigma)\chi_\alpha} (\sum_{\eta=1}^{\alpha-1} \frac{e^{-\int_{T_{\eta+1}}^t \frac{4\beta^2}{(2-\sigma)\chi_v} dv} (1 - e^{-\Delta_\eta \frac{4\beta^2}{(2-\sigma)\chi_\eta}})}{\frac{4\beta^2}{(2-\sigma)\chi_\eta}} + \frac{(1 - e^{(T_\alpha - t)\frac{4\beta^2}{(2-\sigma)\chi_\alpha}})}{\frac{4\beta^2}{(2-\sigma)\chi_\alpha}})
\end{aligned} \tag{A.12}$$

Where:

$$P_\gamma = (1 - e^{-2rt_\gamma \frac{\beta 2(1-\sigma)}{(2-\sigma)}}) \tag{A.13}$$

We can now calculate $p(\alpha|\gamma)$.

$$
\begin{aligned}
p(\alpha|\gamma) &= P_\gamma \int_{T_\alpha}^{T_{\alpha+1}} p(t|t_\gamma)dt \\
&= P_\gamma \int_{T_\alpha}^{T_{\alpha+1}} \frac{2\beta^2}{t_\gamma(2-\sigma)\chi_\alpha}\Big(\sum_{\eta=1}^{\alpha-1} \frac{e^{-\int_{T_{\eta+1}}^{t}\frac{4\beta^2}{(2-\sigma)\chi_v}dv}(1-e^{-\Delta_\eta\frac{4\beta^2}{(2-\sigma)\chi_\eta}})}{\frac{4\beta^2}{(2-\sigma)\chi_\eta}} \\
&\qquad + \frac{(1-e^{(T_\alpha-t)\frac{4\beta^2}{(2-\sigma)\chi_\alpha}})}{\frac{4\beta^2}{(2-\sigma)\chi_\alpha}}\Big)dt \\
&= P_\gamma \int_{T_\alpha}^{T_{\alpha+1}} \frac{2\beta^2}{t_\gamma(2-\sigma)\chi_\alpha}\Big(\sum_{\eta=1}^{\alpha-1} \frac{e^{-\int_{T_\alpha}^{t}\frac{4\beta^2}{(2-\sigma)\chi_v}dv}e^{-\int_{T_{\eta+1}}^{T_\alpha}\frac{4\beta^2}{(2-\sigma)\chi_v}dv}(1-e^{-\Delta_\eta\frac{4\beta^2}{(2-\sigma)\chi_\eta}})}{\frac{4\beta^2}{(2-\sigma)\chi_\eta}} \\
&\qquad + \frac{(1-e^{(T_\alpha-t)\frac{4\beta^2}{(2-\sigma)\chi_\alpha}})}{\frac{4\beta^2}{(2-\sigma)\chi_\alpha}}\Big)dt \\
&= \frac{P_\gamma 2\beta^2}{t_\gamma(2-\sigma)\chi_\alpha}\Big(\int_{T_\alpha}^{T_{\alpha+1}} \sum_{\eta=1}^{\alpha-1} \frac{e^{(T_\alpha-t)\frac{4\beta^2}{(2-\sigma)\chi_\alpha}dv}e^{-\int_{T_{\eta+1}}^{T_\alpha}\frac{4\beta^2}{(2-\sigma)\chi_v}dv}(1-e^{-\Delta_\eta\frac{4\beta^2}{(2-\sigma)\chi_\eta}})}{\frac{4\beta^2}{(2-\sigma)\chi_\eta}}dt \\
&\qquad + \frac{\Delta_\alpha - \frac{(1-e^{-\Delta_\alpha\frac{4\beta^2}{(2-\sigma)\chi_\alpha}})}{\frac{4\beta^2}{(2-\sigma)\chi_\alpha}}}{\frac{4\beta^2}{(2-\sigma)\chi_\alpha}}\Big) \\
&= \frac{P_\gamma 2\beta^2}{t_\gamma(2-\sigma)\chi_\alpha}\Big(\sum_{\eta=1}^{\alpha-1} \frac{(1-e^{-\Delta_\alpha\frac{4\beta^2}{(2-\sigma)\chi_\alpha}})e^{-\int_{T_{\eta+1}}^{T_\alpha}\frac{4\beta^2}{(2-\sigma)\chi_v}dv}(1-e^{-\Delta_\eta\frac{4\beta^2}{(2-\sigma)\chi_\eta}})}{\frac{4\beta^2}{(2-\sigma)\chi_\alpha}\frac{4\beta^2}{(2-\sigma)\chi_\eta}} \\
&\qquad + \frac{\Delta_\alpha - \frac{(1-e^{-\Delta_\alpha\frac{4\beta^2}{(2-\sigma)\chi_\alpha}})}{\frac{4\beta^2}{(2-\sigma)\chi_\alpha)}}}{\frac{4\beta^2}{(2-\sigma)\chi_\alpha}}\Big) \\
&= \frac{P_\gamma}{t_\gamma 2}\Big(\sum_{\eta=1}^{\alpha-1} \frac{(1-e^{-\Delta_\alpha\frac{4\beta^2}{(2-\sigma)\chi_\alpha}})e^{-\int_{T_{\eta+1}}^{T_\alpha}\frac{4\beta^2}{(2-\sigma)\chi_v}dv}(1-e^{-\Delta_\eta\frac{4\beta^2}{(2-\sigma)\chi_\eta}})}{\frac{4\beta^2}{(2-\sigma)\chi_\eta}} \\
&\qquad + \Delta_\alpha - \frac{(1-e^{-\Delta_\alpha\frac{4\beta^2}{(2-\sigma)\chi_\alpha}})}{\frac{4\beta^2}{(2-\sigma)\chi_\alpha}}\Big) \\
&= \frac{P_\gamma}{t_\gamma 2}\Big(\sum_{\eta=1}^{\alpha-1} \frac{(1-e^{-\Delta_\alpha\frac{4\beta^2}{(2-\sigma)\chi_\alpha}})e^{-\sum_{\zeta=\eta+1}^{\alpha}\frac{4\Delta_\zeta\beta^2}{(2-\sigma)\chi_\zeta}}(1-e^{-\Delta_\eta\frac{4\beta^2}{(2-\sigma)\chi_\eta}})}{\frac{4\beta^2}{(2-\sigma)\chi_\eta}} + \\
&\qquad \Delta_\alpha - \frac{(1-e^{-\Delta_\alpha\frac{4\beta^2}{(2-\sigma)\chi_\alpha}})}{\frac{4\beta^2}{(2-\sigma)\chi_\alpha}}\Big)
\end{aligned}
\tag{A.14}
$$

Where:

$$P_\gamma = (1 - e^{-2rt_\gamma \frac{\beta 2(1-\sigma)}{(2-\sigma)}}) \tag{A.15}$$

Calculation of $p(\alpha|\gamma)$ with $\alpha > \gamma$ We first need $p(t|t_\gamma)$ when $\alpha > \gamma$, which is obtained as described below :

$$\begin{aligned}
p(t|t_\gamma) &= \int_0^{t_\gamma} \frac{P_\gamma 2\beta^2}{t_\gamma(2-\sigma)\chi_t}(e^{\int_u^{t_\gamma} -\frac{4\beta^2}{(2-\sigma)\chi_v}dv} e^{\int_{t_\gamma}^{t} -\frac{2\beta^2}{(2-\sigma)\chi_v}dv})du \\
&= \frac{P_\gamma 2\beta^2}{t_\gamma(2-\sigma)\chi_\alpha} e^{\int_{t_\gamma}^{t} -\frac{2\beta^2}{(2-\sigma)\chi_v}dv} \int_0^{t_\gamma} (e^{\int_u^{t_\gamma} -\frac{4\beta^2}{(2-\sigma)\chi_v}dv})du \\
= \frac{P_\gamma 2\beta^2}{t_\gamma(2-\sigma)\chi_\alpha} e^{\int_{t_\gamma}^{t} -\frac{2\beta^2}{(2-\sigma)\chi_v}dv} &(\sum_{\eta=0}^{\gamma-1} \int_{T_\eta}^{T_{\eta+1}} (e^{\int_u^{t_\gamma} -\frac{4\beta^2}{(2-\sigma)\chi_v}dv})du + \int_{T_\gamma}^{t_\gamma} (e^{\int_u^{t_\gamma} -\frac{4\beta^2}{(2-\sigma)\chi_v}dv})du) \\
= \frac{P_\gamma e^{\int_{t_\gamma}^{t} -\frac{2\beta^2}{(2-\sigma)\chi_v}dv} 2\beta^2}{t_\gamma(2-\sigma)\chi_\alpha} &(\sum_{\eta=0}^{\gamma-1} \int_{T_\eta}^{T_{\eta+1}} (e^{-(T_{\eta+1}-u)\frac{4\beta^2}{(2-\sigma)\chi_v}} e^{\int_{T_{\eta+1}}^{t_\gamma} -\frac{4\beta^2}{(2-\sigma)\chi_v}dv})du \\
&+ \int_{T_\gamma}^{t_\gamma} (e^{-(t_\gamma-u)\frac{4\beta^2}{(2-\sigma)\chi_\gamma}})du) \\
= \frac{P_\gamma 2\beta^2}{t_\gamma(2-\sigma)\chi_\alpha} e^{\int_{t_\gamma}^{t} -\frac{2\beta^2}{(2-\sigma)\chi_v}dv} &(\sum_{\eta=1}^{\gamma-1} e^{-\int_{T_{\eta+1}}^{t_\gamma} \frac{4\beta^2}{(2-\sigma)\chi_v}dv} \frac{(1 - e^{-\Delta_\eta \frac{4\beta^2}{(2-\sigma)\chi_\eta}})}{\frac{4\beta^2}{(2-\sigma)\chi_\eta}} \\
&+ \frac{(1 - e^{(T_\gamma - t_\gamma)\frac{4\beta^2}{(2-\sigma)\chi_\gamma}})}{\frac{4\beta^2}{(2-\sigma)\chi_\gamma}}))
\end{aligned} \tag{A.16}$$

Where:

$$P_\gamma = (1 - e^{-2rt_\gamma \frac{\beta 2(1-\sigma)}{(2-\sigma)}}) \tag{A.17}$$

We can now calculate $p(\alpha|\gamma)$.

$$q(\alpha|\gamma) = P_\gamma \int_{T_\alpha}^{T_{\alpha+1}} q(t|t_\gamma)dt$$

$$= \int_{T_\alpha}^{T_{\alpha+1}} \frac{P_\gamma 2\beta^2}{t_\gamma(2-\sigma)\chi_\alpha} e^{-\int_{t_\gamma}^{t} \frac{2\beta^2}{(2-\sigma)\chi_v}dv} (\sum_{\eta=1}^{\gamma-1} e^{-\int_{T_{\eta+1}}^{t_\gamma} \frac{4\beta^2}{(2-\sigma)\chi_v}dv} \frac{(1-e^{-\Delta_\eta \frac{4\beta^2}{(2-\sigma)\chi_\eta}})}{\frac{4\beta^2}{(2-\sigma)\chi_\eta}}$$
$$+ \frac{(1-e^{\frac{(T_\gamma - t_\gamma)4\beta^2}{(2-\sigma)\chi_\gamma}})}{\frac{4\beta^2}{(2-\sigma)\chi_\gamma}})dt$$

$$= \frac{P_\gamma 2\beta^2}{t_\gamma(2-\sigma)\chi_\alpha} (\sum_{\eta=0}^{\gamma-1} e^{-\int_{T_{\eta+1}}^{t_\gamma} \frac{4\beta^2}{(2-\sigma)\chi_v}dv} \frac{(1-e^{-\Delta_\eta \frac{4\beta^2}{(2-\sigma)\chi_\eta}})}{\frac{4\beta^2}{(2-\sigma)\chi_\eta}} + \frac{(1-e^{\frac{(T_\gamma - t_\gamma)4\beta^2}{(2-\sigma)\chi_\gamma}})}{\frac{4\beta^2}{(2-\sigma)\chi_\gamma}})$$
$$\int_{T_\alpha}^{T_{\alpha+1}} e^{-\int_{t_\gamma}^{T_\alpha} \frac{2\beta^2}{(2-\sigma)\chi_v}dv} e^{-\int_{T_\alpha}^{t} \frac{2\beta^2}{(2-\sigma)\chi_v}dv} dt$$

$$= \frac{P_\gamma 2\beta^2}{t_\gamma(2-\sigma)\chi_\alpha} (\sum_{\eta=1}^{\gamma-1} e^{-\int_{T_{\eta+1}}^{t_\gamma} \frac{4\beta^2}{(2-\sigma)\chi_v}dv} \frac{(1-e^{-\Delta_\eta \frac{4\beta^2}{(2-\sigma)\chi_\eta}})}{\frac{4\beta^2}{(2-\sigma)\chi_\eta}} + \frac{(1-e^{\frac{(T_\gamma - t_\gamma)4\beta^2}{(2-\sigma)\chi_\gamma}})}{\frac{4\beta^2}{(2-\sigma)\chi_\gamma}})$$
$$e^{-\int_{t_\gamma}^{T_\alpha} \frac{2\beta^2}{(2-\sigma)\chi_v}dv} \int_{T_\alpha}^{T_{\alpha+1}} e^{(T_\alpha - t)\frac{2\beta^2}{(2-\sigma)\chi_\alpha}} dt$$

$$= \frac{P_\gamma 2\beta^2}{t_\gamma(2-\sigma)\chi_\alpha} (\sum_{\eta=1}^{\gamma-1} e^{-\int_{T_{\eta+1}}^{t_\gamma} \frac{4\beta^2}{(2-\sigma)\chi_v}dv} \frac{(1-e^{-\Delta_\eta \frac{4\beta^2}{(2-\sigma)\chi_\eta}})}{\frac{4\beta^2}{(2-\sigma)\chi_\eta}} + \frac{(1-e^{\frac{(T_\gamma - t_\gamma)4\beta^2}{(2-\sigma)\chi_\gamma}})}{\frac{4\beta^2}{(2-\sigma)\chi_\gamma}})$$
$$e^{-\int_{t_\gamma}^{T_\alpha} \frac{2\beta^2}{(2-\sigma)\chi_v}dv} \frac{(1-e^{-\Delta_\alpha \frac{2\beta^2}{(2-\sigma)\chi_\alpha}})}{\frac{2\beta^2}{(2-\sigma)\chi_\alpha}}$$

$$= \frac{P_\gamma}{t_\gamma} (\sum_{\eta=1}^{\gamma-1} e^{-\int_{T_{\eta+1}}^{t_\gamma} \frac{4\beta^2}{(2-\sigma)\chi_v}dv} \frac{(1-e^{-\Delta_\eta \frac{4\beta^2}{(2-\sigma)\chi_\eta}})}{\frac{4\beta^2}{(2-\sigma)\chi_\eta}} +$$
$$\frac{(1-e^{\frac{(T_\gamma - t_\gamma)4\beta^2}{(2-\sigma)\chi_\gamma}})}{\frac{4\beta^2}{(2-\sigma)\chi_\gamma}}) e^{-\int_{t_\gamma}^{T_\alpha} \frac{2\beta^2}{(2-\sigma)\chi_v}dv} (1-e^{-\Delta_\alpha \frac{2\beta^2}{(2-\sigma)\chi_\alpha}})$$

(A.18)

Where:

$$P_\gamma = (1-e^{-2rt_\gamma \frac{\beta 2(1-\sigma)}{(2-\sigma)}}) \tag{A.19}$$

Calculation of $p(\alpha|\gamma)$ with $\alpha = \gamma$ Because probabilities sum up to one. We have the following formula:

$$p(\gamma|\gamma) = 1 - (\sum_{\alpha=0}^{\gamma-1} p(\alpha|\gamma) + \sum_{\alpha=\gamma+1}^{n} p(\alpha|\gamma)) \tag{A.20}$$

Emission Matrix

The probability of observing a mutation or not is given by the following formula:

$$\begin{aligned} P(0|\gamma) &= e^{-2\mu t_\gamma} \\ P(1|\gamma) &= 1 - e^{-2\mu t\gamma} \end{aligned} \tag{A.21}$$

Where μ is the mutation rate per nucleotide per N generation and $t\gamma$ the average coalescent time in state γ.

Calculating the objective function of the Baum-Welch Algorithm

To calculate the objective function (or Composite Likelihood (CL)), we first need to define it as :

$$CL = P(Y, X|\beta, \chi, \rho) \tag{A.22}$$

Which is the probability of the signal (Y) and the sequence of Hidden states (X) given the germination rate (β),self-fertilization rate (σ),recombination rate (ρ) and population size per interval (χ). To calculate this probability we use a forward-backward algorithm.

Forward Algorithm The Forward algorithm is an iterative algorithm that calculates at step t the probability :

$$fo_t(i) = P(Y_{1,\dots,t}, X_t = i) \tag{A.23}$$

To calculate this probability we define:

- T : Transition matrix ($T_{i,j} = P(X(t) = j|X(t-1) = i)$)
- O : observation matrix ($O_{i,i} = P(Y(t) = e(t)|X(t) = i)$) where e(t) is the observed data at position t (which can be 0 or 1)

Initialization $fo_1 = q_0 O_1$

Where q_0 is the vector of initial probabilities.

Recursive formula $fo_t = O_t T^T fo_{t-1}$

In case of recurrent patterns in the sequence, a technique has been developed to improve the algorithm efficiency[145]. For example, if there are many repetition of the same observation (repetition of length l), we have :

$$fo_t = O_t T^T fo_{t-1} = O_t T^T O_{t-1} T^T fo_{t-2} = (O_t T^T)^l fo_{t-l} \tag{A.24}$$

To compute $P(O_{1,\dots,L})$ which we call the likelihood (LH), We simply notice that :

$$\sum_i fo_t(i) = \sum_i P(Y_{1,\dots,t}, X_t = i) = P(Y_{1,\dots,t}) \tag{A.25}$$

Which leads to :

$$\begin{aligned} c_t &= \sum_i fo_t(i) \\ fo_t^* &= \frac{fo_t}{c_t} \\ fo_t &= O_t T^T fo_{t-1}^* \\ LH &= \prod_{t=1}^{L} c_t \end{aligned} \tag{A.26}$$

Backward Algorithm The backward algorithm is an iterative algorithm that calculates : $ba_t(i) = P(Y_{t+1,\dots,L}|X_t = i)$

The notations are the same as before. The algorithm is defined as :

Initialization $ba_L = I$

Recursive formula $ba_{t-1} = TO_t ba_t$

In a similar way, if there are repeated observations in the sequence we have:

$$ba_t = TO_{t+1} ba_{t+1} = TO_{t+1} TO_{t+2} ba_{t+2} = (TO_t)^l ba_{t+l} \tag{A.27}$$

Baum-Welch Algorithm

The classic algorithm The Baum-Welch Algorithm is a particular case of the generalized Expectation-Maximization algorithm. At every step the algorithm updates the parameters that maximize the function Q. Where Q is defined at step t as :

$$Q(\theta|\theta^t) = \sum_X P(X|Y,\theta^t) log(P(X,Y|\theta)) \tag{A.28}$$

And so :

$$\theta^{t+1} = argmax_\theta Q(\theta|\theta^t) \tag{A.29}$$

We have :

$$Q(\theta|\theta^t) = \sum_X P(X|Y,\theta^t) log(\prod_{X_1} P(X_1|\theta)^{N(X_1)} \prod_{X,Z} P(X|Z,\theta)^{N(X,Z)} \prod_{X,Y} P(Y|X,\theta^t)^{N(Y,X)}) \tag{A.30}$$

Where :

- $N(X_1)$: number of first positions where the hidden state is X_1
- $N(X,Z)$: number of transitions from state Z to X
- $N(Y,X)$: number of positions with observation Y happening with hidden state X

Which gives us :

$$Q(\theta|\theta^t) = \nu_{\theta^t} log(P(X_1|\theta)) + \sum_{X,Y} E(X,Z|\theta^t) log(P(X|Z,\theta)) + \sum_{X,Y} E(Y,X|\theta^t) log(P(Y|X,\theta)) \tag{A.31}$$

Where:

- ν_θ : The equilibrium probability conditional to the set of parameters θ
- $P(X_1|\theta)$: Probability of the first hidden state conditional to the set of parameters θ
- $E(X,Z|\theta^t)$: Expected number of transitions of X from Z conditional to the observation and set of parameters θ^t
- $P(X|Z,\theta)$: Transition probability from state Z to state X conditional to the set of parameters θ
- $E(Y,X|\theta^t)$ Expected number of observations of type Y that happened during state X conditional to the observation and set of parameters θ^t
- $P(Y|X,\theta)$: Emission probability conditional to the set of parameters θ

The objective function used in [146] is

$$Q^*(\theta|\theta^t) = \sum_{X,Z} E(X,Z|\theta^t) log(P(X|Z,\theta)) \tag{A.32}$$

Calculating $E(X, Z|\theta^t)$

$$E(X, Z|\theta^t) = \frac{\sum_{l=1}^{L-1} fo_t(Z)ba_{t+1}(X)P(Y_{t+1}|X)P(X|Z)}{P(Y_{1,\dots,L}|\theta)} \tag{A.33}$$

Calculating $E(Y, X|\theta^t)$

$$E(Y, X|\theta^t) = \frac{\sum_{t=1}^{L} fo_t(X)ba_t(X)1_{Y_t}}{P(Y_{1,\dots,L}|\theta)} \tag{A.34}$$

Speeding the algorithm We thus have using the approach in [168] (Cf forward and backward algorithm).

$$\begin{aligned} fo_l &= (W^T)^{l-k} fo_k \\ ba_k &= W^{l-k} ba_l \\ W &= TO = PDP^{-1} \end{aligned} \tag{A.35}$$

Calculating $E(Y, X|\theta^t)$ We want $fo_l ba_l$:

$$\begin{aligned} E(Y, X|\theta^t) = \sum_{i=k}^{l-1} fo_i ba_i &= \sum_{i=0}^{l-k-1} diag((W^T)^i fo_k (W^{l-k-i} ba_l)^T) \\ \sum_{i=k}^{l-1} fo_i ba_i &= \sum_{i=0}^{l-k-1} diag((W^T)^i fo_k ba_l^t (W^{l-k-i})^T) \\ \sum_{i=k}^{l-1} fo_i ba_i &= \sum_{i=0}^{l-k-1} diag(((P^{-1})D^i P^T fo_k ba_l^t (P^{-1})^T D^{l-k-i} P^T) \\ \sum_{i=k}^{l-1} fo_i ba_i &= diag((P^{-1})^T A P^T) \\ A &= \sum_{i=0}^{l-k-1} D^i P^T fo_k ba_l^t (P^{-1})^T D^{l-k-i} \\ U &= P^T fo_k ba_l^t (P^{-1})^T \\ A &= \sum_{i=0}^{l-k-1} D^i U D^{l-k-i} \end{aligned} \tag{A.36}$$

We have :

$$\begin{aligned} \sum_{i=0}^{l-k-1} D^i U D^{l-k-i} &= \sum_{i=0}^{m} D^i U D^{m-i} D \\ (\sum_{i=0}^{l-k-1} D^i U D^{l-k-i})_{ab} &= \sum_{i=0}^{m} D^i_{aa} U_{ab} D^{m+1-i}_{bb} = U_{ab} \sum_{i=0}^{m} D^i_{aa} D^{m+1-i}_{bb} \end{aligned} \tag{A.37}$$

We therefore define Q:

$$Q_{ab} = \sum_{i=0}^{m} D^i_{aa} D^{m-i}_{bb} \tag{A.38}$$

In the end:

$$A = (U * Q)D \tag{A.39}$$

Where $*$ stands for the Hadamard product.

Calculating $E(X, Z|\theta^t)$ In a similar way.

$$\begin{aligned} E(X, Z|\theta^t) = \sum_{i=k}^{l-1} \xi_i &= (fo_i(ba_{i+1} Y_{i+1})) * T \\ \xi_{i,ab} &= P(X_i = b, X_{i+1} = a|Y, \theta^t) \end{aligned} \tag{A.40}$$

Which gives us:

$$\begin{aligned} \sum_{i=k}^{l-1} \xi_i &= \sum_{i=k}^{l-1} (fo_i ba_{i+1}^T O) * T \\ \xi_{i,ab} &= P(X_i = b, X_{i+1} = a|Y_{1,\ldots,L}, \theta^t) \end{aligned} \tag{A.41}$$

With repeated observed data we have:

$$\begin{aligned} \sum_{i=k}^{l-1} \xi_i &= \sum_{i=0}^{l-k-1} ((W^T)^i fo_k ba_l^T (W^{k-l-1-i})^T O) * T \\ \sum_{i=k}^{l-1} \xi_i &= \sum_{i=0}^{l-k-1} ((P^{-1})^T D^i P^T fo_k ba_l^T (P^{-1})^T D^{l-k-1-i} P^T) O) * T \\ \sum_{i=k}^{l-1} \xi_i &= (((P^{-1})^T (U * Q) P^T) O) * T \end{aligned} \tag{A.42}$$

Maximizing the objective function

To maximize the Complete Likelihood, as shown before we need to maximize the following value:

$$Q^*(\theta|\theta^t) = \sum_{X,Z} E(X, Z|\theta^t) log(P(X|Z, \theta)) \tag{A.43}$$

To maximize the objective function we use a Barzilai-Borwein spectral method.

A.2 Appendix of Chapter 3

A.2.1 Supplementary Figures

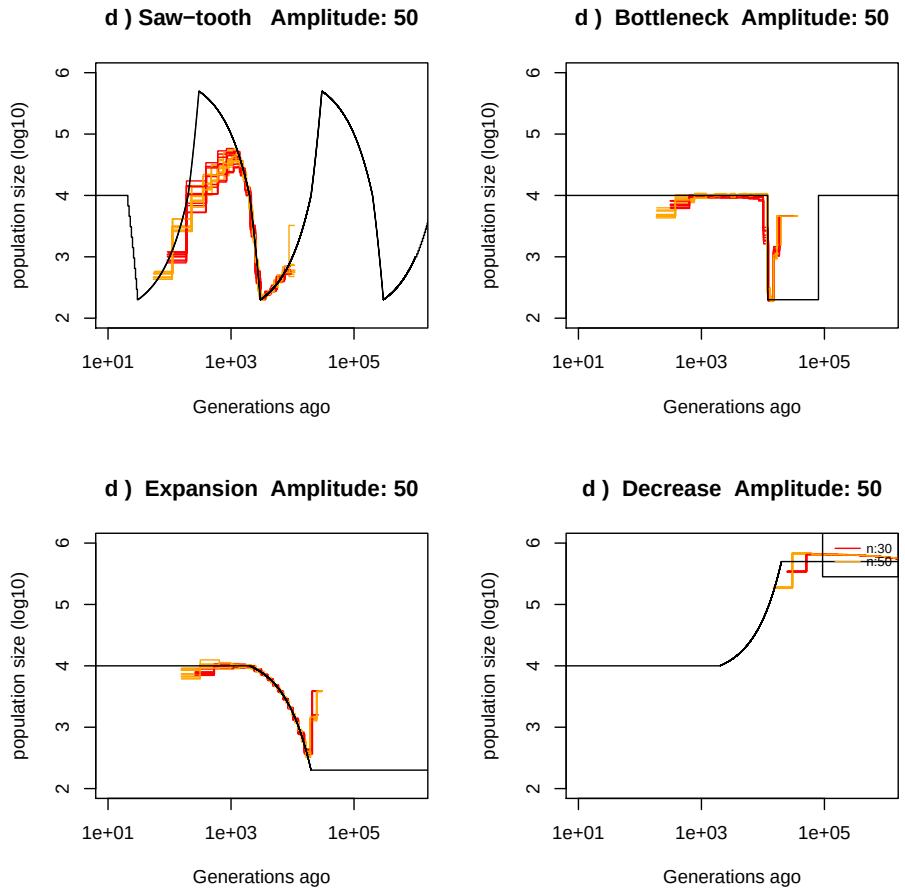

Figure A.12: **Best-case convergence of PSMC'.** Estimated demographic history using simulated genealogy over sequences of 1 Gb using 30 or 50 hidden states (respectively in red,orange) under scenarios with population size of fold 50 (black) with 10 replicates. Recombination rate is set to 1×10^{-8} per generation per bp and mutation rate to 1.25×10^{-8} per generation per bp. a) Demographic history simulated under a sawtooth scenario of strength 50. b) Demographic history simulated under a bottleneck scenario of strength 50. c) Demographic history simulated under a population expansion scenario of strength 50. d) Demographic history simulated under a population decrease scenario of strength 50.

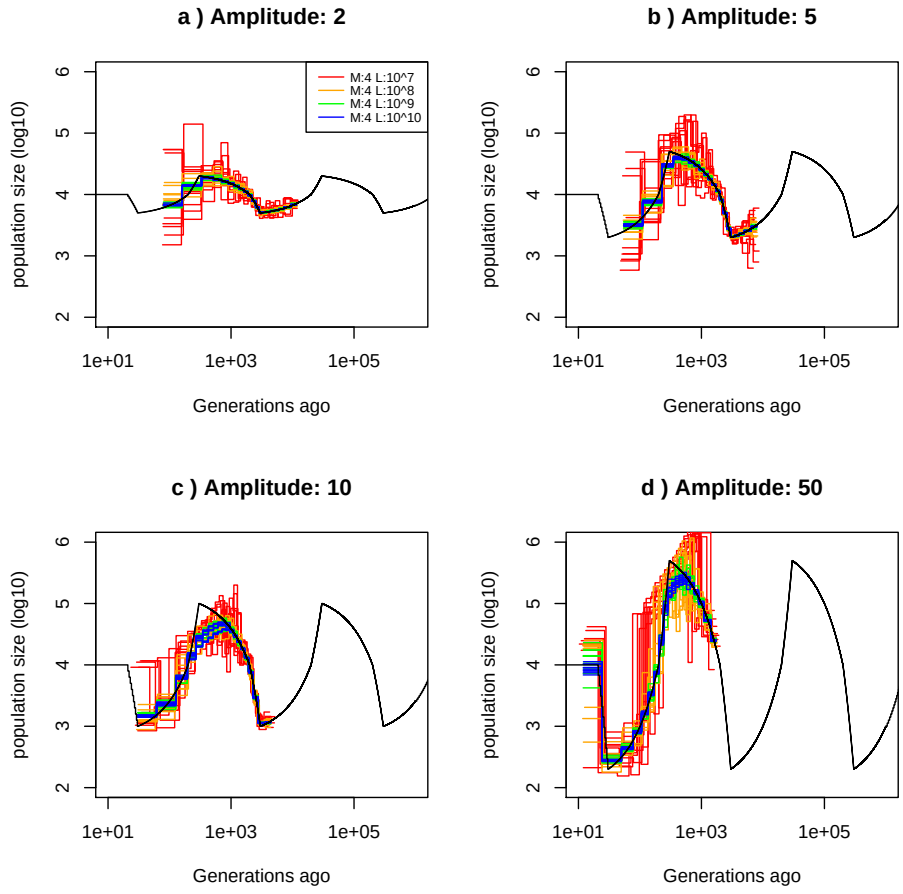

Figure A.13: **Best-case convergence of MSMC.** Estimated demographic history using simulated genealogies of 4 sequences of 10,100,1000,10000 Mb (respectively in red,orange, green and blue) under a sawtooth scenario (black) with 10 replicates for different amplitudes of size change: a) 2-fold, b) 5-fold, c) 10-fold, and d) 50-fold. The recombination rate is set to 1×10^{-8} per generation per bp and the mutation rate to 1.25×10^{-8} per generation per bp.

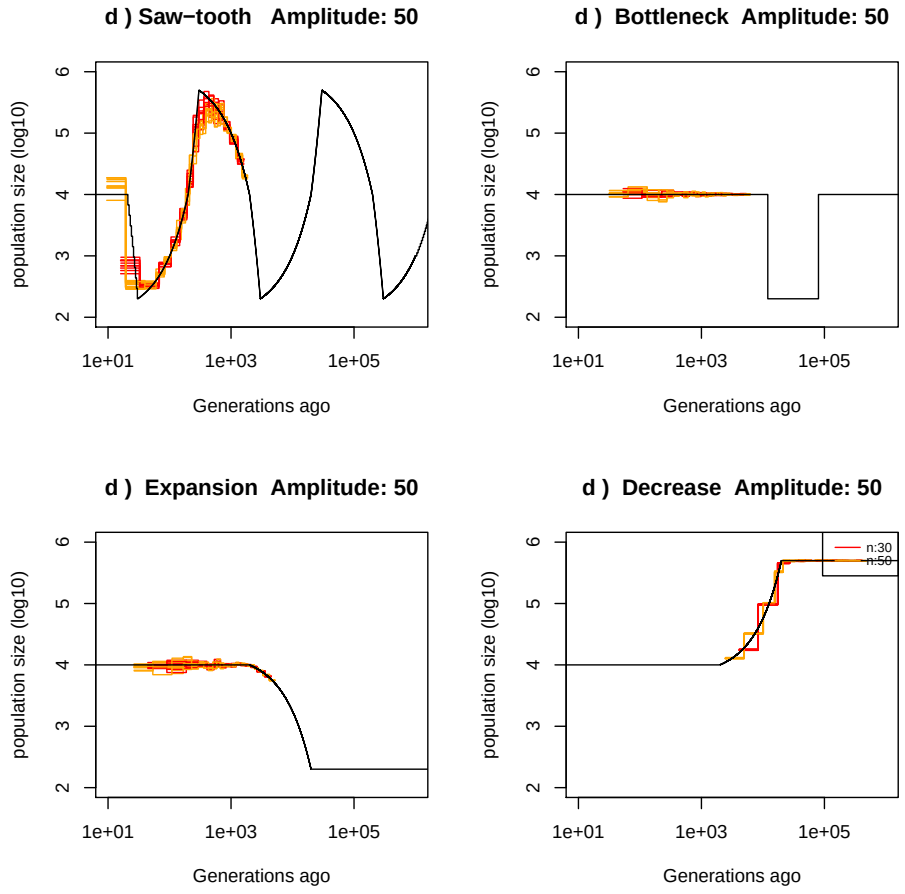

Figure A.14: **Best-case convergence of MSMC.** Estimated demographic history using simulated genealogy over sequences of 1 Gb using 30 or 50 hidden states (respectively in red,orange) under scenarios with variation of population size fold 50 (black) with 10 replicates. Recombination rate is set to 1×10^{-8} per generation per bp and mutation rate to 1.25×10^{-8} per generation per bp. a) Demographic history simulated under a sawtooth scenario of strength 50. b) Demographic history simulated under a bottleneck scenario of strength 50. c) Demographic history simulated under a population expansion scenario of strength 50. d) Demographic history simulated under a population decrease scenario of strength 50.

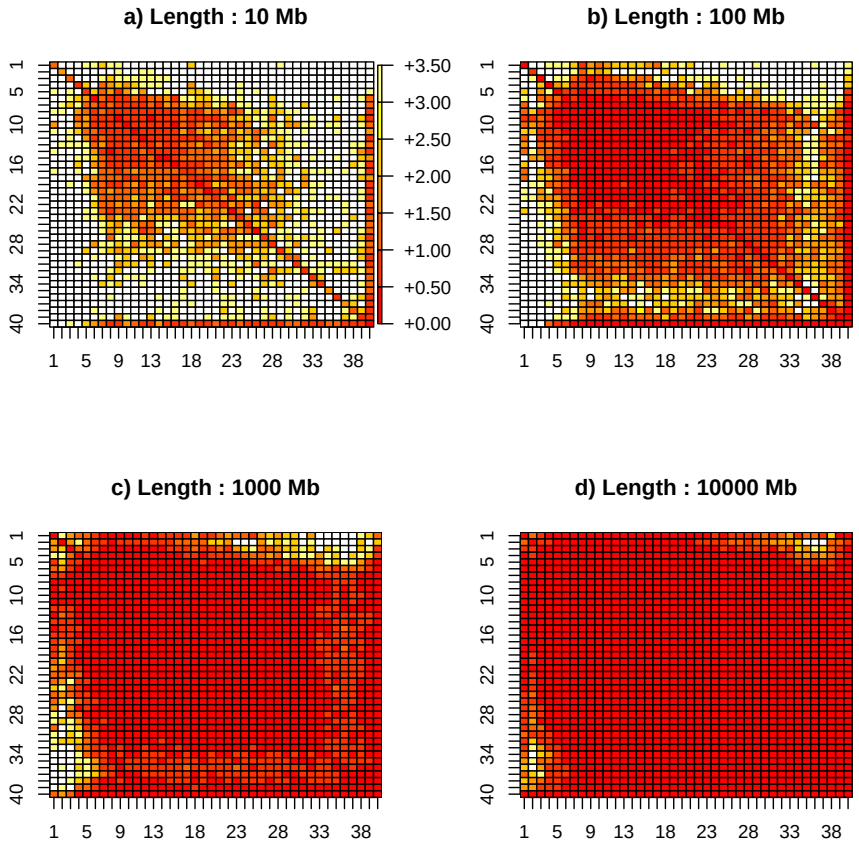

Figure A.15: **Estimated Transition matrix in sawtooth scenario.** Estimated transition matrix coefficient of variation using simulated genealogy over sequences of 10,100,1000,10000 (respectively in a), b) c) and d)) Mb under a sawtooth scenario with 10 replicates. Recombination rate is set to 1×10^{-8} per generation per bp and mutation rate to 1.25×10^{-8} per generation per bp. Demographic history is simulated under a sawtooth scenario of amplitude 10.

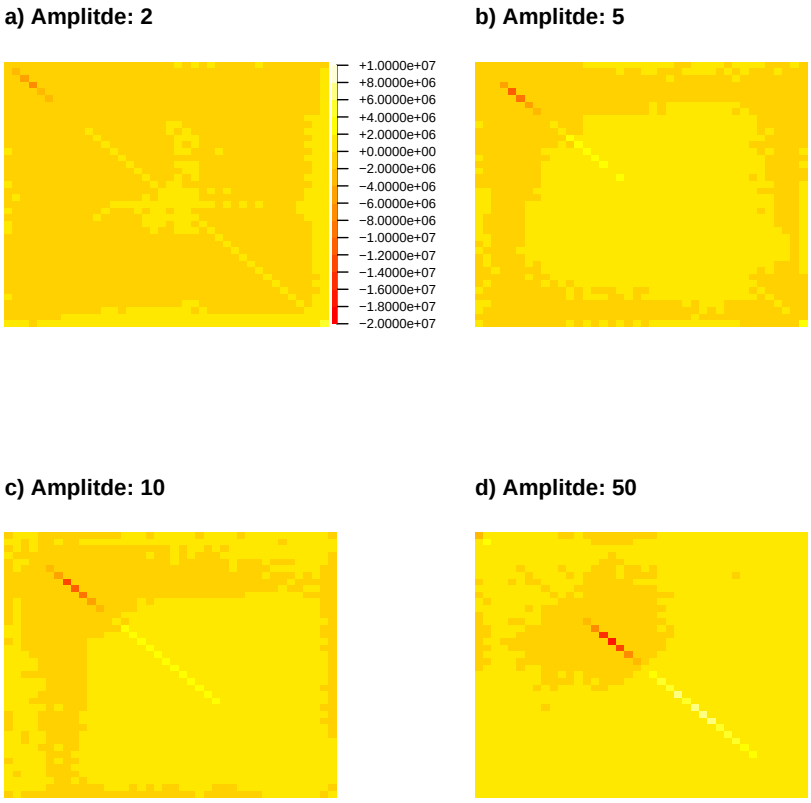

Figure A.16: **Mean difference between estimated and actual transition matrix in sawtooth scenario.** Mean difference between estimated and transition matrix directly build from genealogy using sequences of 100 Mb under a sawtooth scenario of strength 2,5,10 and 50 (respectively in a), b) c) and d)) each with 10 replicates. Recombination rate is set to 1×10^{-8} per generation per bp and mutation rate to 1.25×10^{-8} per generation per bp.

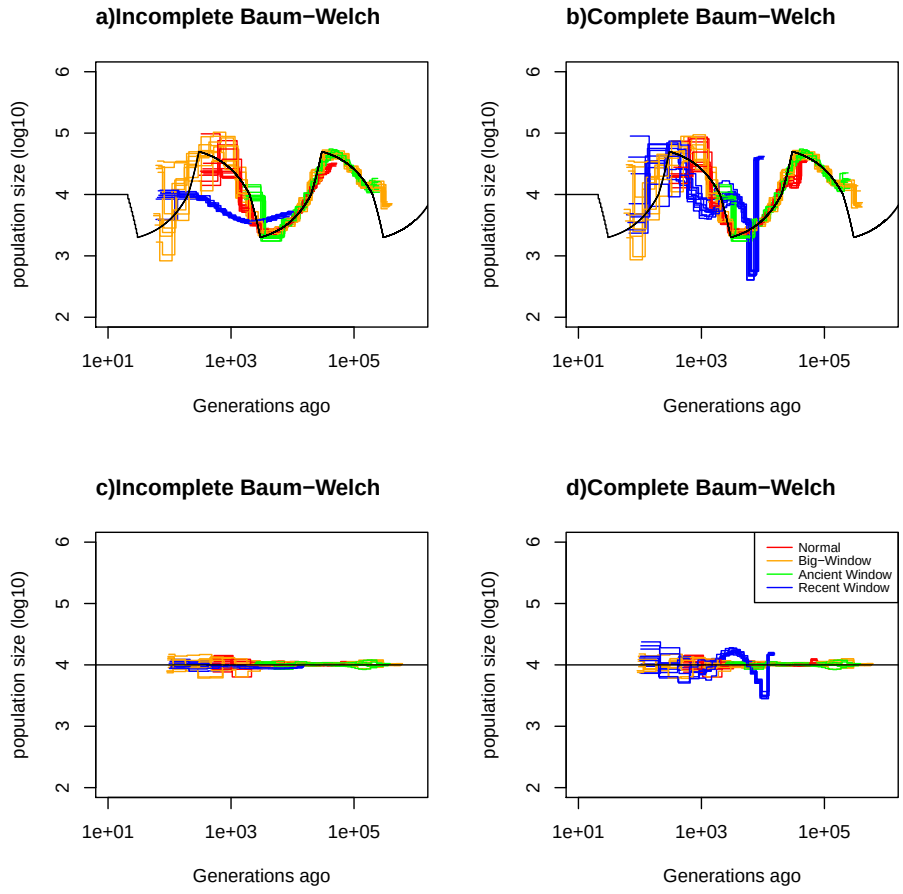

Figure A.17: **Estimated demography using different window and optimization function under a sawtooth and constant population size scenario.** Estimated demographic history under a sawtooth (a), b)) and constant population size (c,d)) scenario with 10 replicates using 4 simulated sequences of length 50 Mb. Estimation with the incomplete optimization function are displayed in a) and c). Estimation with the complete Baum-Welch algorithm are displayed in b) and c). Recombination rate is set to 1×10^{-8} per generation per bp and mutation rate to 1.25×10^{-8} per generation per bp. The simulated demographic history is represented in black. Estimation with the time window of PSMC' are displayed in red, with the time window of MSMC2 in orange, estimation with a PSMC' time window shifted by fold 5 in past are displayed in green and shifted by fold 5 in recent time are displayed in blue.

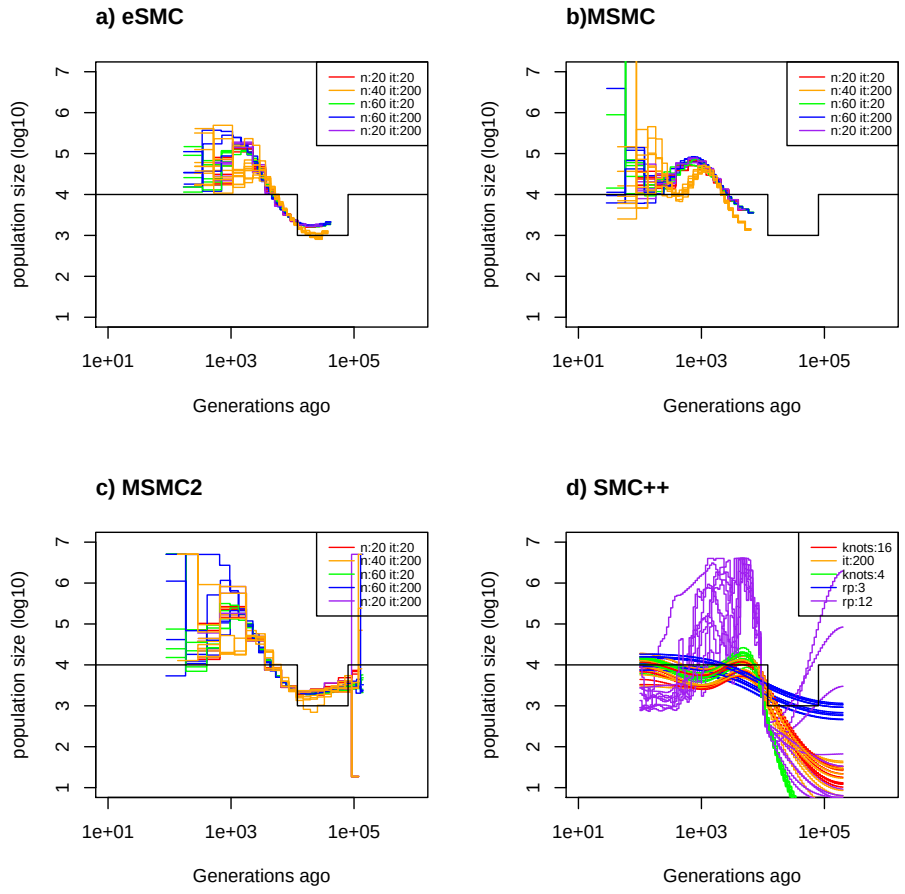

Figure A.18: **Estimated demography of SMC method under a bottleneck scenario.** Estimated demographic history under a bottleneck scenario with 10 replicates using simulated sequences. 2 sequences of 100 Mb for eSMC and MSMC2 (respectively in a) and b)). We use 4 sequences of 100 Mb for MSMC (c)) and 20 sequences of 100 Mb for SMC++ (d)). Recombination rate is set to 1.25×10^{-7} per generation per bp and mutation rate to 1.25×10^{-8} per generation per bp. Demographic history is simulated under a bottleneck scenario of strength 10 and is represented in black. Analysis with eSMC, MSMC and MSMC2 using 20 hidden states are in red, 200 iterations orange, 60 hidden states green, 60 hidden states and 200 iterations in blue, 20 hidden states and 200 iterations in purple. For SMC++, analysis using 16 knots are in red, 200 iterations in orange, 4 knots in green, regularization penalty set to 3 in blue and regularization-penalty set to 12 in purple.

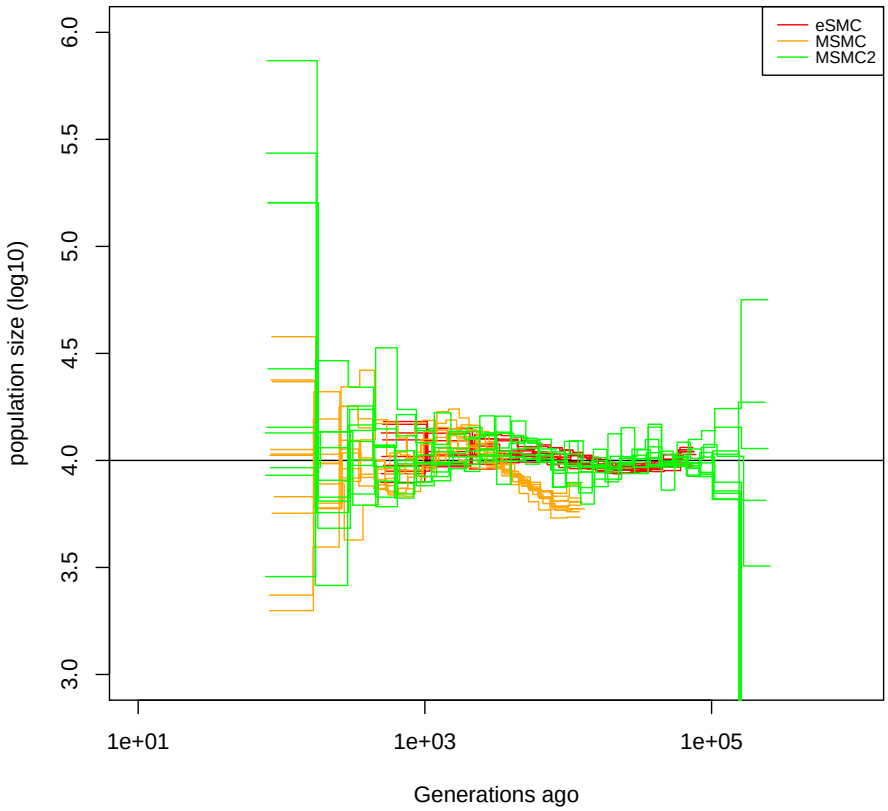

Figure A.19: **Estimated demography of eSMC under a constant population size with recombination rate variation.** Estimated demographic history by eSMC under constant population size (black) with (red) or without (orange) variation of recombination rate along the sequence with 10 replicates using 2 simulated sequences of 40 Mb. Mutation rate is set to 1.25×10^{-8} per generation per bp. Recombination rate changes randomly every 2 Mb between 2.5×10^{-9} and 6.25×10^{-8}. T

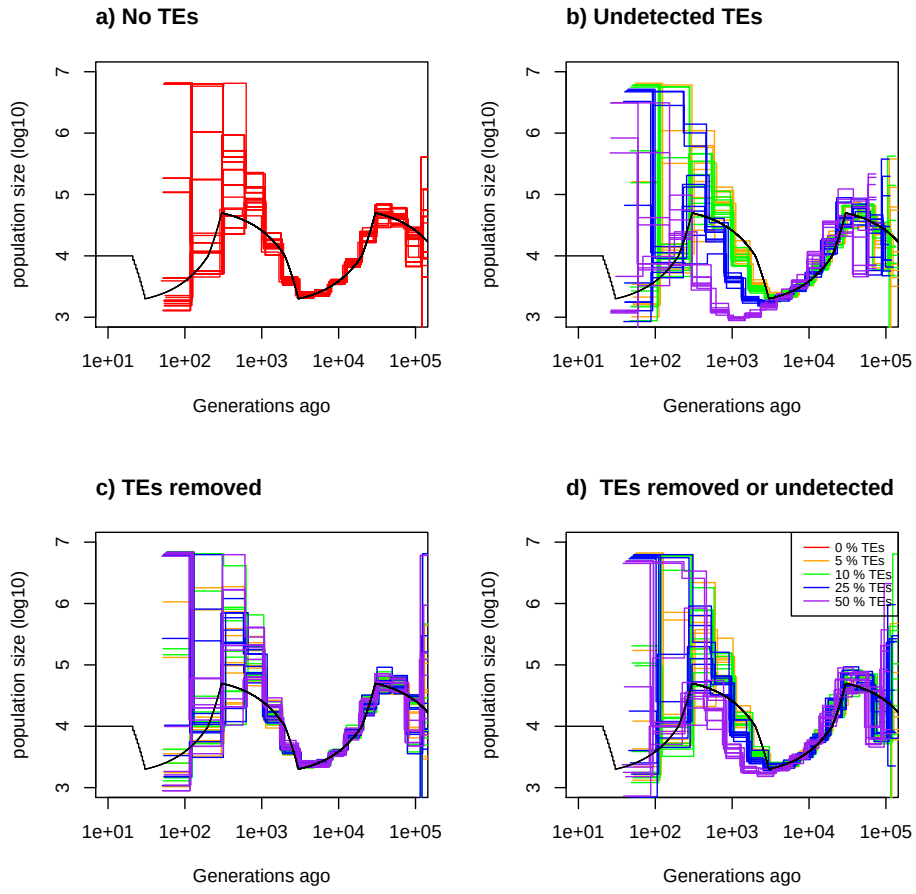

Figure A.20: **Estimated demography of MSMC2 under a sawtooth scenario with transposable elements.** Estimated demographic history by MSMC2 under a sawtooth scenario with 10 replicates using simulated sequences. 4 sequences of 20 Mb. Recombination rate is set to 1.25×10^{-8} per generation per bp and mutation rate to 1.25×10^{-8} per generation per bp . The simulated demographic history is represented in black. Here tansposable element are of length 10kbp. a) Demographic history simulated with no transposable elements. b) Demographic history simulated where transposable element are removed. c) Demographic history simulated where SNPs on transposable are removed. d) Demographic history simulated where half of transposable are removed and SNPs on the other half are removed. Proportion of transpobable element of the genome is set to 0% (red), 5% (orange), 10% (green), 25 % (blue) and 50 % (purple).

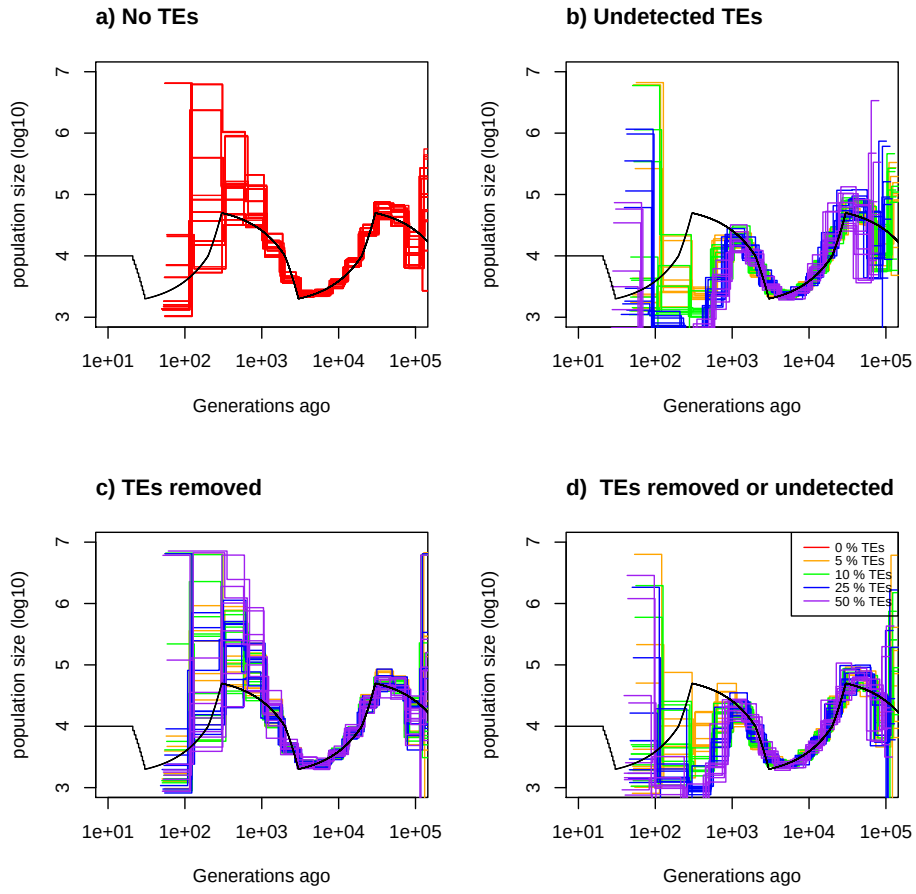

Figure A.21: **Estimated demography of MSMC2 under a sawtooth scenario with transposable elements.** Estimated demographic history by MSMC2 under a sawtooth scenario with 10 replicates using simulated sequences. 4 sequences of 20 Mb. Recombination rate is set to 1.25×10^{-8} per generation per bp and mutation rate to 1.25×10^{-8} per generation per bp . The simulated demographic history is represented in black. Here tansposable element are of length 100 kbp. a) Demographic history simulated with no transposable elements. b) Demographic history simulated where transposable element are removed. c) Demographic history simulated where SNPs on transposable are removed. d) Demographic history simulated where half of transposable are removed and SNPs on the other half are removed. Proportion of transpobable element of the genome is set to 0% (red), 5% (orange), 10% (green), 25 % (blue) and 50 % (purple).

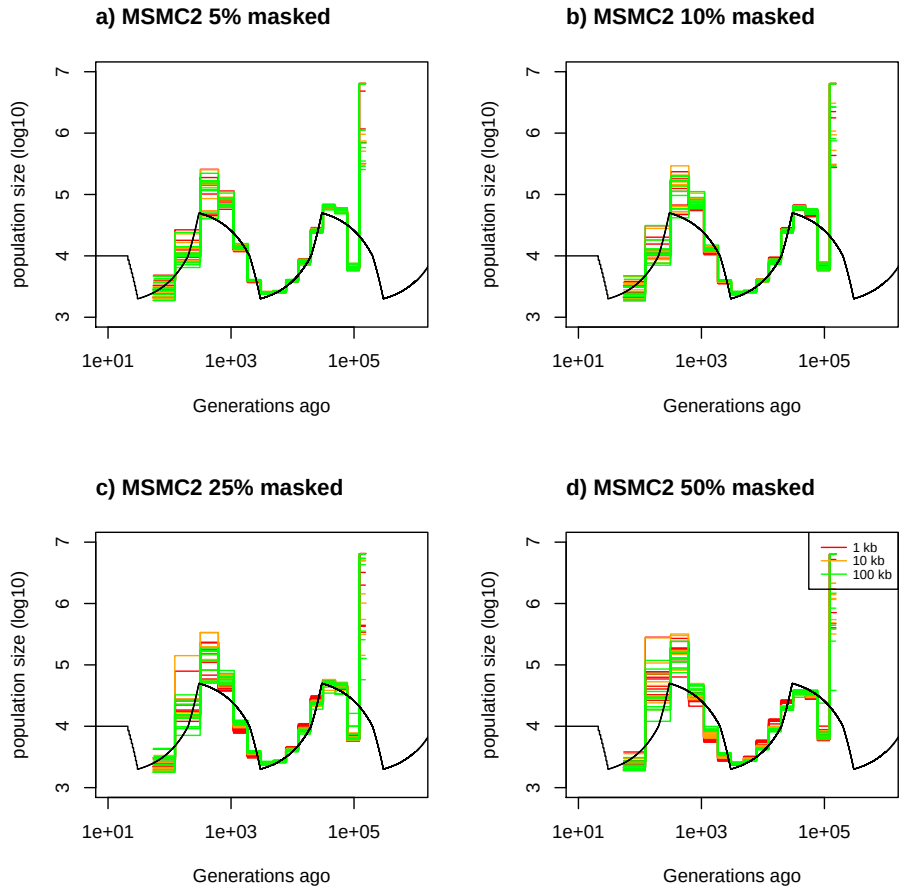

Figure A.22: **Estimated demography of MSMC2 under a sawtooth scenario with masked transposable elements.** Estimated demographic history by MSMC2 under a sawtooth scenario with 10 replicates using simulated sequences. 4 sequences of 20 Mb. Recombination rate is set to 1.25×10^{-8} per generation per bp and mutation rate to 1.25×10^{-8} per generation per bp . The simulated demographic history is represented in black. Here tansposable element are of length 1kbp. a) Demographic history simulated with 5% transposable elements. b) DDemographic history simulated with 10% transposable elements. c) Demographic history simulated with 25% transposable elements. d) Demographic history simulated with 50% transposable elements. Size of transposable elements are set to 1kb (red), 10kb (orange), 100 kb (green).

A.2.2 Supplementary Tables

Figure	eSMC 10 Mb	eSMC 100 Mb	eSMC 1 Gb	eSMC 10 Gb
1 a)	7.23 (0.11)	6.34 (0.040)	6.10 (0.01)	5.99 (0.005)
1 b)	8.59 (0.099)	7.99 (0.012)	7.86 (0.010)	7.86 (0.002)
1 c)	8.96 (0.038)	8.74 (0.011)	8.77 (0.005)	8.76 (0.006)
1 d)	10.36 (0.002)	10.37 (0.001)	10.38 (0.001)	10.38 (0.001)

Table A.1: Average mean square error of Figure 3.1 (in log10). The coefficient of variation is indicated in brackets.

Figure	eSMC	MSMC	MSMC2	SMC++
3 a)	6.82 (0.07)	7.60 (0.05)	8.21 (0.17)	7.57 (0.03)
3 b)	8.16 (0.01)	8.69 (0.08)	10.0 (0.16)	7.84 (0.03)
3 c)	9.27 (0.03)	9.53 (0.08)	11.43 (0.08)	8.86 (0.004)
3 d)	10.0 (0.008)	10.90 (0.12)	11.0 (0.04)	10.46 (0.0001)

Table A.2: Average mean square error of Figure 3.3 (in log10). The coefficient of variation is indicated in brackets.

Figure	$\frac{\rho}{\theta} = 0.1$	$\frac{\rho}{\theta} = 1$	$\frac{\rho}{\theta} = 10$
4 a)	8.21 (0.12)	7.45 (0.06)	9.62 (0.04)
4 b)	15.7 (0.3)	9.46 (0.33)	10.43 (0.20)
4 c)	9.95 (0.24)	8.48 (0.22)	10.65 (0.13)
4 d)	7.78 (0.008)	7.79 (0.003)	7.79 (0.001)

Table A.3: Average mean square error of Figure 3.4 (in log10). The coefficient of variation is indicated in brackets.

% of problematic SNPs	5 a)	5 b)	5 c)	5 d)
0	9.10 (0.10)			
5		12.66 (0.002)	13.0 (0.007)	13.27 (0.001)
10		10.08 (0.11)	12.15 (0.07)	12.55 (0.01)
25		12.83 (0.012)	13.03 (0.001)	13.28 (0.006)

Table A.4: Average mean square error of Figure 3.5 (in log10). The coefficient of variation is indicated in brackets.

Figure	Share	Not share
6 a)	6.58 (0.06)	6.68 (0.08)
6 b)	7.99 (0.04)	8.02 (0.098)
6 c)	7.08 (0.15)	7.43 (0.093)
6 d)	7.99 (0.07)	8.09 (0.09)

Table A.5: Average mean square error of Figure 3.6 (in log10). The coefficient of variation is indicated in brackets.

A.3 Appendix of Chapter 4

A.3.1 Description of the SMβC

The Sequentially Markovian β Coalescent is a Hidden Markov Model based on the Multiple Sequentially Markovian Coalescent (MSMC) where Multiple Merger events are allowed to occur following the Beta 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.
- α : The parameter of the Beta distribution
- ξ_t : Scaled population size at time t ($N_t = \xi_t N_0$)

- $\chi_t = \xi_t^{\alpha-1}$
- M : Number of analyzed sequences (or individuals)

The model's assumptions are :

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

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

$$\lambda_{b,\alpha,b-n+1} = \frac{B(n-\alpha, b-n+\alpha)}{\Gamma(2-\alpha)\Gamma(\alpha)}. \tag{A.44}$$

$$\Lambda_{b,\alpha,b-n+1} = \frac{\binom{b}{n} B(n-\alpha, b-n+\alpha)}{\Gamma(2-\alpha)\Gamma(\alpha)}. \tag{A.45}$$

Thus, the total rate is

$$\lambda_{b,\alpha} = \sum_{k=2}^{b} \frac{\binom{b}{n} B(k-\alpha, b-n+\alpha)}{\Gamma(2-\alpha)\Gamma(\alpha)} \tag{A.46}$$

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 [149], 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

$$f(t) = \frac{\lambda}{\chi(t)} e^{-\int_{t_0}^{t} \frac{\lambda}{\chi(s)} ds}, \tag{A.47}$$

which follows as described in [67].

Hidden States

Our hidden states at one position are define by the first coalescent event's time $t > 0$ at that position and which individuals $\mathbf{i} := (i_1, \ldots, i_n)$ coalesce in the corresponding coalescence. A transition from coalescent time s to time t or a change in the index $\mathbf{i}$ 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 if different from the two other (indicated by i, if the individual i is different from the two other).

Transition Matrix

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

$$P(rec|s) = 1 - e^{-rMs} \tag{A.48}$$

We assume that only one recombination event can occur between two positions. A recombination event in 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 [27, Sect. 5]: Conditional on the mergers of the M other lineages, a binary merger of the "freed" lineage appears with rate $M\lambda_{M+1,\alpha,M}$ and it joins an existing merger of m lineages at some time t with probability $1 - \frac{\lambda_{M+1,\alpha,M+1-m+1}}{\lambda_{M,\alpha,M-m+1}} = \frac{\lambda_{M+1,\alpha,M+1-m}}{\lambda_{M,\alpha,M-m+1}}$, 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.

$\mathbf{t}<\mathbf{s}$ 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 $\mathbf{i} = \{i_1, i_2\}$ (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. (A.47) with rate $M\lambda_{M+1,\alpha,M}$ and the probability that the first merger is indeed merging $\mathbf{i}$, which is $\frac{1}{M}$ (we pick the second lineage

of $\mathbf{i}$ at random from M lineages):

$$f(t, \mathbf{i}|s, \mathbf{j}, u) = \frac{2\lambda_{M+1,\alpha,M}}{M\chi_t} e^{-\int_u^t \frac{M\lambda_{M+1,\alpha,M}}{\chi_v} dv} \tag{A.49}$$

t=s Case 1: a non coalescing individual joins the coalescent event $\mathbf{j} = \{j_1, \ldots, j_n\}$. 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 $\exp(-\int_u^s \frac{M\lambda_{M+1,\alpha,M}}{\chi_v} dv)$ (by integrating Eq. (A.47) with rate $M\lambda_{M+1,\alpha,M}$). Finally, it then needs to join in the coalescent event $\mathbf{j}$, which happens with probability $\frac{\lambda_{M+1,\alpha,M+1-m}}{\lambda_{M,\alpha,M-m+1}}$. This shows that :

$$P(s, \mathbf{i}|s, \mathbf{j}, u) = \frac{(M-n)\lambda_{M+1,\alpha,M+1-m}}{M\lambda_{M,\alpha,M-m+1}} e^{-\int_u^s \frac{M\lambda_{M+1,\alpha,M}}{\chi_v} dv} \tag{A.50}$$

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

Case 2: Recombination occurs in a coalescing individual $i \in \mathbf{j} = \{j_1, \ldots, j_n\}$ of a multiple merger event with $n > 2$ (happens with probability $\frac{n}{M}$). 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

$$P(s, \mathbf{i}|s, \mathbf{j}, u) = \frac{n\lambda_{M+1,\alpha,M+1-m+1}}{M\lambda_{M,\alpha,M-m+1}} e^{-\int_u^s \frac{M\lambda_{M+1,\alpha,M}}{\chi_v} dv} \tag{A.51}$$

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

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. A.49 without the factor $2/M$, integrating over $[u, s]$ yields the probability $(1 - (1/M)\exp(-\int_u^t \frac{M\lambda_{M+1,\alpha,M}}{\chi_v} dv)$ 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. (A.51)).

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 $\mathbf{i}$, 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. (A.51) with $n=2$, and multiply this by the conditional density for merging into any $\mathbf{i}$ of M lineages afterwards. Thus, this is just Eq. (A.47) with rate $\lambda_{M,\alpha}$ multiplied with $\frac{\lambda_{M,\alpha,M-n+1}}{\lambda_{M,\alpha}}$. This leads to

$$f(t,\mathbf{i}|s,\mathbf{j},u) = \frac{2\lambda_{M+1,\alpha,M}}{M\lambda_{M,\alpha,M-1}} e^{-\int_u^s \frac{M\lambda_{M+1,\alpha,M}}{\chi_v}dv} \frac{\lambda_{M,\alpha,M-n+1}}{\chi_t} e^{-\int_s^t \frac{\lambda_{M,\alpha}}{\chi_v}dv} \tag{A.52}$$

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

Full transition probability

$$p(t,\mathbf{i}|s,\mathbf{j},u) = \begin{cases} P_s \frac{2\lambda_{2,\alpha}}{\chi_t M} e^{-\int_u^t \frac{M\lambda_{M+1,\alpha,M}}{\chi_v}dv} & u<t<s \\ (1-P_s) + P_s(\int_u^s \frac{1}{\chi_k} e^{\int_u^k -\frac{M\lambda_{M+1,\alpha,M}}{\chi_v}dv} dk + \frac{(M-n)\lambda_{(n+1),\alpha,2}}{M\lambda_{(n+1),\alpha,2}+\lambda_{(n+1),\alpha,1}} e^{\int_u^t -\frac{M\lambda_{M+1,\alpha,M}}{\chi_v}dv}) & t=s, m=n \\ P_s \frac{(M-n)\lambda_{(n+1),\alpha,1}}{M(\lambda_{(n+1),\alpha,2}+\lambda_{(n+1),\alpha,1})} e^{-\int_u^s \frac{M\lambda_{M+1,\alpha,M}}{\chi_v}dv} & t=s, m=n+1 \\ P_s \frac{n}{M} \frac{\lambda_{(n+1),\alpha,2}}{s(\lambda_{(n+1),\alpha,2}+\lambda_{(n+1),\alpha,1})} e^{-\int_u^s \frac{M\lambda_{M+1,\alpha,M}}{\chi_v}dv} & t=s, m+1=n \\ P_s \frac{\lambda_{M,\alpha,(M-m+1)}}{\binom{M}{m}\chi_\alpha} e^{-\int_s^t \frac{\lambda_{M,\alpha}}{\chi_v}dv} e^{-\int_u^s \frac{M\lambda_{M+1,\alpha,M}}{\chi_v}dv} \frac{2\lambda_{(n+1),\alpha,2}}{M(\lambda_{(n+1),\alpha,2}+\lambda_{(n+1),\alpha,1})} & t>s, i=l, j=k \end{cases} \tag{A.53}$$

Where $P_s = (1-e^{-Mrs})$ represents the recombination probability.

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

$$T_\beta = \frac{-\ln(1-\frac{\beta}{n})}{\lambda_{M,\beta}} \tag{A.54}$$

We therefore have:

$$p(\beta,\mathbf{i}|s,\mathbf{j}) = \int_{T_\beta}^{T_{\beta+1}} p(t,\mathbf{i}|s,\mathbf{j})dt \tag{A.55}$$

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 :

$$\begin{aligned} q_o(\beta, \mathbf{i}) &= \int_{T_\beta}^{T_{\beta+1}} \frac{\lambda_{M,\alpha,(M-m+1)}}{\chi_\beta \binom{M}{m}} e^{-\int_0^t \frac{\lambda_{M,\alpha}}{\chi_v} dv} dt \\ &= \frac{\lambda_{M,\alpha,(M-m+1)} e^{\sum_{\eta=0}^{\beta-1} \frac{\lambda_{M,\alpha}}{\chi_\eta} \Delta_\eta}}{\binom{M}{m} \lambda_{M,\alpha}} (1 - e^{-\Delta_\beta \frac{\lambda_{M,\alpha}}{\chi_t}}) \end{aligned} \tag{A.56}$$

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

$$\begin{aligned} t_{\gamma,j} = E[\text{Coalescent time}|\gamma, j] = \frac{E[\text{Coalescent time} \cap \gamma, j]}{P(\gamma, j)} = \\ \frac{\int_{T_\gamma}^{T_{\gamma+1}} t \lambda_{M,\alpha,(M-n+1)} e^{-\int_0^t \frac{\lambda_{M,\alpha}}{\chi_v} dv}}{\binom{M}{n} q_0(\gamma, j)} dt \\ = \frac{T_\gamma - T_{\gamma+1} e^{-\Delta_\gamma \frac{\lambda_{M,\alpha}}{\chi_\gamma}}}{(1 - e^{-\Delta_\gamma \frac{\lambda_{M,\alpha}}{\chi_\gamma}})} + \frac{\chi_\gamma}{\lambda_{M,\alpha}} \end{aligned} \tag{A.57}$$

Where :

$$\Delta_\gamma = T_{\gamma+1} - T_\gamma \tag{A.58}$$

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

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

We here calculate the transition probabilities from the state γ to a time t in

the time interval β.

$$P(t,i|t_\gamma,j) = \frac{P_\gamma}{t_\gamma}\int_0^t \frac{2\lambda_{2,\alpha}}{\chi_v M} e^{-\int_u^t \frac{M\lambda_{M+1,\alpha,M}}{\chi_v}dv} du$$

$$= \frac{P_\gamma}{t_\gamma}(\sum_{\eta=0}^{\beta-1}\int_{T_\eta}^{T_{\eta+1}} \frac{2\lambda_{2,\alpha}}{\chi_v M} e^{-\int_u^t \frac{M\lambda_{M+1,\alpha,M}}{\chi_v}dv} du + \int_{T_\beta}^{t} \frac{2\lambda_{2,\alpha}}{\chi_v M} e^{-\int_u^t \frac{M\lambda_{M+1,\alpha,M}}{\chi_v}dv} du)$$

$$= \frac{P_\gamma}{t_\gamma}(\sum_{\eta=0}^{\beta-1}\int_{T_\eta}^{T_{\eta+1}} \frac{2\lambda_{2,\alpha}}{\chi_\beta M} e^{-\int_{T_{\eta+1}}^{T_\beta} \frac{M\lambda_{M+1,\alpha,M}}{\chi_v}dv} e^{-\int_{T_\beta}^{t} \frac{M\lambda_{M+1,\alpha,M}}{\chi_v}dv} e^{-\int_u^{T_{\eta+1}} \frac{M\lambda_{M+1,\alpha,M}}{\chi_v}dv} du +$$

$$\int_{T_\beta}^{t} \frac{2\lambda_{2,\alpha}}{\chi_\beta M} e^{-(t-u)\frac{M\lambda_{M+1,\alpha,M}}{\chi_\beta}dv} du)$$

$$= \frac{P_\gamma}{t_\gamma}\frac{2\lambda_{2,\alpha}}{\chi_\beta M}(\sum_{\eta=0}^{\beta-1}\int_{T_\eta}^{T_{\eta+1}} e^{-\sum_{\zeta=\eta+1}^{\beta-1}\Delta_\zeta \frac{M\lambda_{M+1,\alpha,M}}{\chi_\zeta}} e^{-(t-T_\beta)\frac{M\lambda_{M+1,\alpha,M}}{\chi_\beta}} e^{-(T_{\eta+1}-u)\frac{M\lambda_{M+1,\alpha,M}}{\chi_\eta}} du$$

$$+\frac{(1-e^{-(t-T_\beta)\frac{M\lambda_{M+1,\alpha,M}}{\chi_\beta}})}{\frac{M\lambda_{M+1,\alpha,M}}{\chi_\beta}})$$

$$= \frac{P_\gamma}{t_\gamma}\frac{2\lambda_{2,\alpha}}{\chi_\beta M}(\sum_{\eta=0}^{\beta-1} e^{-\sum_{\zeta=\eta+1}^{\beta-1}\Delta_\zeta \frac{M\lambda_{M+1,\alpha,M}}{\chi_\zeta}} e^{-(t-T_\beta)\frac{M\lambda_{M+1,\alpha,M}}{\chi_\beta}} \frac{(1-e^{-\Delta_\eta\frac{M\lambda_{M+1,\alpha,M}}{\chi_\eta}})}{\frac{M\lambda_{M+1,\alpha,M}}{\chi_\eta}}$$

$$+\frac{(1-e^{-(t-T_\beta)\frac{M\lambda_{M+1,\alpha,M}}{\chi_\beta}})}{\frac{M\lambda_{M+1,\alpha,M}}{\chi_\beta}})$$

(A.59)

We then have to integrate t over the time interval β to have the transition

probability from the state γ to the state β.

$$
\begin{aligned}
&P(\beta,i|\gamma,j)\\
&=\int_{T_\beta}^{T_{\beta+1}}\frac{P_\gamma}{t_\gamma}\frac{2\lambda_{2,\alpha}}{\chi_\beta M}\Big(\sum_{\eta=0}^{\beta-1}e^{-\sum_{\zeta=\eta+1}^{\beta-1}\Delta_\zeta\frac{M\lambda_{M+1,\alpha,M}}{\chi_\zeta}}e^{-(t-T_\beta)\frac{M\lambda_{M+1,\alpha,M}}{\chi_\beta}}\frac{(1-e^{-\Delta_\eta\frac{M\lambda_{M+1,\alpha,M}}{\chi_\eta}})}{\frac{M\lambda_{M+1,\alpha,M}}{\chi_\eta}}\\
&\qquad+\frac{(1-e^{-(t-T_\beta)\frac{M\lambda_{M+1,\alpha,M}}{\chi_\beta}})}{\frac{M\lambda_{M+1,\alpha,M}}{\chi_\beta}}\Big)dt\\
&=\frac{P_\gamma}{t_\gamma}\frac{2\lambda_{2,\alpha}}{\chi_\beta M}\Big(\sum_{\eta=0}^{\beta-1}e^{-\sum_{\zeta=\eta+1}^{\beta-1}\Delta_\zeta\frac{M\lambda_{M+1,\alpha,M}}{\chi_\zeta}}\frac{(1-e^{-\Delta_\beta\frac{M\lambda_{M+1,\alpha,M}}{\chi_\beta}})}{\frac{M\lambda_{M+1,\alpha,M}}{\chi_\beta}}\frac{(1-e^{-\Delta_\eta\frac{M\lambda_{M+1,\alpha,M}}{\chi_\eta}})}{\frac{M\lambda_{M+1,\alpha,M}}{\chi_\eta}}\\
&\qquad+\frac{(\Delta_\beta-\frac{(1-e^{-(\Delta_\beta\frac{M\lambda_{M+1,\alpha,M}}{\chi_\beta})})}{\frac{M\lambda_{M+1,\alpha,M}}{\chi_\beta}})}{\frac{M\lambda_{M+1,\alpha,M}}{\chi_\beta}}\Big)\\
&=\frac{P_\gamma}{t_\gamma}\frac{2}{M^2}\Big(\sum_{\eta=0}^{\beta-1}e^{-\sum_{\zeta=\eta+1}^{\beta-1}\Delta_\zeta\frac{M\lambda_{M+1,\alpha,M}}{\chi_\zeta}}(1-e^{-\Delta_\beta\frac{M\lambda_{M+1,\alpha,M}}{\chi_\beta}})\frac{(1-e^{-\Delta_\eta\frac{M\lambda_{M+1,\alpha,M}}{\chi_\eta}})}{\frac{M\lambda_{M+1,\alpha,M}}{\chi_\eta}}\\
&\qquad+(\Delta_\beta-\frac{(1-e^{-(\Delta_\beta\frac{M\lambda_{M+1,\alpha,M}}{\chi_\beta})})}{\frac{M\lambda_{M+1,\alpha,M}}{\chi_\beta}})\Big)
\end{aligned}
\tag{A.60}
$$

Where the recombination probability if defined as:

$$P_\gamma=(1-e^{-Mrt_\gamma}) \tag{A.61}$$

$\gamma<\beta$ We here calculate the transition probabilities from the state γ to a time

t in the time interval β.

$$
\begin{aligned}
&P(t,i|t_\gamma,j)\\
&=\int_0^{t_\gamma}\frac{P_\gamma}{t_\gamma}\frac{\lambda_{M,\alpha,(M-m+1)}}{\binom{M}{m}\chi_\beta}e^{-\int_{t_\gamma}^{t}\frac{\lambda_{M,\alpha}}{\chi_v}dv}e^{-\int_{u}^{t_\gamma}\frac{M\lambda_{M+1,\alpha,M}}{\chi_v}dv}\frac{2\lambda_{(n+1),\alpha,2}}{M(\lambda_{(n+1),\alpha,2}+\lambda_{(n+1),\alpha,1})}du\\
&=\frac{P_\gamma}{t_\gamma}\frac{\lambda_{M,\alpha,(M-m+1)}}{\binom{M}{m}\chi_\beta}(\sum_{\eta=1}^{\gamma-1}\int_{T_\eta}^{T_{\eta+1}}e^{-\int_{t_\gamma}^{t}\frac{\lambda_{M,\alpha}}{\chi_v}dv}e^{-\int_{T_{\eta+1}}^{T_\gamma}\frac{M\lambda_{M+1,\alpha,M}}{\chi_v}dv}e^{-\int_{T_\gamma}^{t_\gamma}\frac{M\lambda_{M+1,\alpha,M}}{\chi_v}dv}\\
&\qquad e^{-\int_{u}^{T_{\eta+1}}\frac{M\lambda_{M+1,\alpha,M}}{\chi_v}dv}\frac{2\lambda_{(n+1),\alpha,2}}{M(\lambda_{(n+1),\alpha,2}+\lambda_{(n+1),\alpha,1})}du\\
&\qquad+\int_{T_\gamma}^{t_\gamma}e^{-\int_{t_\gamma}^{t}\frac{\lambda_{M,\alpha}}{\chi_v}dv}e^{-\int_{u}^{t_\gamma}\frac{M\lambda_{M+1,\alpha,M}}{\chi_v}dv}\frac{2\lambda_{(n+1),\alpha,2}}{M(\lambda_{(n+1),\alpha,2}+\lambda_{(n+1),\alpha,1})}du)\\
&=\frac{P_\gamma}{t_\gamma}\frac{2\lambda_{(n+1),\alpha,2}e^{-\int_{t_\gamma}^{t}\frac{\lambda_{M,\alpha}}{\chi_v}dv}}{M(\lambda_{(n+1),\alpha,2}+\lambda_{(n+1),\alpha,1})}\frac{\lambda_{M,\alpha,(M-m+1)}}{\binom{M}{m}\chi_\beta}(\sum_{\eta=1}^{\gamma-1}e^{-\sum_{\zeta=\eta+1}^{\gamma-1}\Delta_\zeta\frac{M\lambda_{M+1,\alpha,M}}{\chi_\zeta}}\\
&\qquad e^{-(t_\gamma-T_\gamma)\frac{M\lambda_{M+1,\alpha,M}}{\chi_\gamma}}\frac{(1-e^{-\Delta_\eta\frac{M\lambda_{M+1,\alpha,M}}{\chi_\eta}})}{\frac{M\lambda_{M+1,\alpha,M}}{\chi_\eta}}+\frac{(1-e^{-(t_\gamma-T_\gamma)\frac{M\lambda_{M+1,\alpha,M}}{\chi_\gamma}})}{\frac{M\lambda_{M+1,\alpha,M}}{\chi_\gamma}})
\end{aligned}
$$

We then have to integrate t over the time interval β to have the transition probability from the state γ to the state β.

$$
\begin{aligned}
P(\beta,i|\gamma,j)&=\int_{T_\beta}^{T_{\beta+1}}\frac{P_\gamma}{t_\gamma}\frac{2\lambda_{(n+1),\alpha,2}e^{-\int_{t_\gamma}^{t}\frac{\lambda_{M,\alpha}}{\chi_v}dv}}{M(\lambda_{(n+1),\alpha,2}+\lambda_{(n+1),\alpha,1})}\frac{\lambda_{M,\alpha,(M-m+1)}}{\binom{M}{m}\chi_\beta}\\
&(\sum_{\eta=1}^{\gamma-1}e^{-\sum_{\zeta=\eta+1}^{\gamma-1}\Delta_\zeta\frac{M\lambda_{M+1,\alpha,M}}{\chi_\zeta}}e^{-(t_\gamma-T_\gamma)\frac{M\lambda_{M+1,\alpha,M}}{\chi_\gamma}}\frac{(1-e^{-\Delta_\eta\frac{M\lambda_{M+1,\alpha,M}}{\chi_\eta}})}{\frac{M\lambda_{M+1,\alpha,M}}{\chi_\eta}}+\frac{(1-e^{-(t_\gamma-T_\gamma)\frac{M\lambda_{M+1,\alpha,M}}{\chi_\gamma}})}{\frac{M\lambda_{M+1,\alpha,M}}{\chi_\gamma}})dt\\
&=\int_{T_\beta}^{T_{\beta+1}}\frac{P_\gamma}{t_\gamma}\frac{2\lambda_{(n+1),\alpha,2}e^{-\int_{t_\gamma}^{T_\beta}\frac{\lambda_{M,\alpha}}{\chi_v}dv}e^{-\int_{T_\beta}^{t}\frac{\lambda_{M,\alpha}}{\chi_v}dv}}{M(\lambda_{(n+1),\alpha,2}+\lambda_{(n+1),\alpha,1})}\frac{\lambda_{M,\alpha,(M-m+1)}}{\binom{M}{m}\chi_\beta}\\
&(\sum_{\eta=1}^{\gamma-1}e^{-\sum_{\zeta=\eta+1}^{\gamma-1}\Delta_\zeta\frac{M\lambda_{M+1,\alpha,M}}{\chi_\zeta}}e^{-(t_\gamma-T_\gamma)\frac{M\lambda_{M+1,\alpha,M}}{\chi_\gamma}}\frac{(1-e^{-\Delta_\eta\frac{M\lambda_{M+1,\alpha,M}}{\chi_\eta}})}{\frac{M\lambda_{M+1,\alpha,M}}{\chi_\eta}}+\frac{(1-e^{-(t_\gamma-T_\gamma)\frac{M\lambda_{M+1,\alpha,M}}{\chi_\gamma}})}{\frac{M\lambda_{M+1,\alpha,M}}{\chi_\gamma}})dt\\
&=\frac{P_\gamma}{t_\gamma}\frac{2\lambda_{(n+1),\alpha,2}e^{-\int_{t_\gamma}^{T_\beta}\frac{\lambda_{M,\alpha}}{\chi_v}dv}(1-e^{-\Delta_\beta\frac{\lambda_{M,\alpha}}{\chi_\beta}})}{M(\lambda_{(n+1),\alpha,2}+\lambda_{(n+1),\alpha,1})}\frac{\lambda_{M,\alpha,(M-m+1)}}{\binom{M}{m}\lambda_{M,\alpha}}\\
&(\sum_{\eta=1}^{\gamma-1}e^{-\sum_{\zeta=\eta+1}^{\gamma-1}\Delta_\zeta\frac{M\lambda_{M+1,\alpha,M}}{\chi_\zeta}}e^{-(t_\gamma-T_\gamma)\frac{M\lambda_{M+1,\alpha,M}}{\chi_\gamma}}\frac{(1-e^{-\Delta_\eta\frac{M\lambda_{M+1,\alpha,M}}{\chi_\eta}})}{\frac{M\lambda_{M+1,\alpha,M}}{\chi_\eta}}+\frac{(1-e^{-(t_\gamma-T_\gamma)\frac{M\lambda_{M+1,\alpha,M}}{\chi_\gamma}})}{\frac{M\lambda_{M+1,\alpha,M}}{\chi_\gamma}})
\end{aligned}
\tag{A.62}
$$

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

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 in the current first coalescent event (before or after the coalescent event).

$$P(\gamma,i|\gamma,j)=\frac{P_\gamma}{t_\gamma}\int_0^{t_\gamma}\frac{(M-n)\lambda_{(n+1),\alpha,1}e^{-\int_u^{t_\gamma}\frac{M\lambda_{M+1,\alpha,M}}{\chi_v}dv}}{M(\lambda_{(n+1),\alpha,2}+\lambda_{(n+1),\alpha,1})}du+Pc_1+Pc_2$$

$$=Pc_1+Pc_2+\frac{P_\gamma}{t_\gamma}\frac{(M-n)\lambda_{(n+1),\alpha,1}}{M(\lambda_{(n+1),\alpha,2}+\lambda_{(n+1),\alpha,1})}($$

$$\sum_{\eta=}^{\gamma-1}\int_{T_\eta}^{T_{\eta+1}}e^{-\int_u^{T_{\eta+1}}\frac{M\lambda_{M+1,\alpha,M}}{\chi_v}dv}e^{-\int_{T_{\eta+1}}^{t_\gamma}\frac{M\lambda_{M+1,\alpha,M}}{\chi_v}dv}du\int_{T_\gamma}^{t_\gamma}e^{-\int_u^{t_\gamma}\frac{M\lambda_{M+1,\alpha,M}}{\chi_v}dv}du)$$

$$=\frac{P_\gamma}{t_\gamma}\frac{(M-n)\lambda_{(n+1),\alpha,1}}{M(\lambda_{(n+1),\alpha,2}+\lambda_{(n+1),\alpha,1})}(\sum_{\eta=}^{\gamma-1}\frac{(1-e^{-\Delta_\eta\frac{M\lambda_{M+1,\alpha,M}}{\chi_\eta}})}{\frac{M\lambda_{M+1,\alpha,M}}{\chi_\eta}}e^{-\int_{T_{\eta+1}}^{t_\gamma}\frac{M\lambda_{M+1,\alpha,M}}{\chi_v}dv}$$

$$+\frac{(1-e^{-(t_\gamma-T_\gamma)\frac{M\lambda_{M+1,\alpha,M}}{\chi_\gamma}})}{\frac{M\lambda_{M+1,\alpha,M}}{\chi_\gamma}})+Pc_1+Pc_2$$

Pc_1 is the probability that a recombination happens before the first coalescent event in the non coalescing branch, the resulting free branch then coalesces 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, the resulting free branch then coalesces after the current first coalescent event but in the same hidden state (resulting in a multiple merger coa-

lescent because of the discretized time)

$$Pc_1 = \int_{T_\gamma}^{t_\gamma} \frac{P_\gamma}{t_\gamma}\frac{\lambda_{2,\alpha}}{\chi_\gamma M}(\sum_{\eta=0}^{\gamma-1} e^{-\sum_{\zeta=\eta+1}^{\gamma-1}\Delta_\zeta \frac{M\lambda_{M+1,\alpha,M}}{\chi_\zeta}} e^{-(t-T_\gamma)\frac{M\lambda_{M+1,\alpha,M}}{\chi_\gamma}} \frac{(1-e^{-\Delta_\eta \frac{M\lambda_{M+1,\alpha,M}}{\chi_\eta}})}{\frac{M\lambda_{M+1,\alpha,M}}{\chi_\eta}} + \frac{(1-e^{-(t-T_\gamma)\frac{M\lambda_{M+1,\alpha,M}}{\chi_\gamma}})}{\frac{M\lambda_{M+1,\alpha,M}}{\chi_\gamma}})dt$$

$$= \frac{P_\gamma}{t_\gamma}\frac{\lambda_{2,\alpha}}{\chi_\gamma M}(\sum_{\eta=0}^{\gamma-1} e^{-\sum_{\zeta=\eta+1}^{\gamma-1}\Delta_\zeta \frac{M\lambda_{M+1,\alpha,M}}{\chi_\zeta}} \frac{(1-e^{-(t_\gamma-T_\gamma)\frac{M\lambda_{M+1,\alpha,M}}{\chi_\gamma}})}{\frac{M\lambda_{M+1,\alpha,M}}{\chi_\gamma}} \frac{(1-e^{-\Delta_\eta \frac{M\lambda_{M+1,\alpha,M}}{\chi_\eta}})}{\frac{M\lambda_{M+1,\alpha,M}}{\chi_\eta}} + \frac{((t_\gamma-T_\gamma)-\frac{(1-e^{-(t_\gamma-T_\gamma)\frac{M\lambda_{M+1,\alpha,M}}{\chi_\gamma}})}{\frac{M\lambda_{M+1,\alpha,M}}{\chi_\gamma}})}{\frac{M\lambda_{M+1,\alpha,M}}{\chi_\gamma}})$$

$$= \frac{P_\gamma}{t_\gamma}\frac{1}{M^2}(\sum_{\eta=0}^{\gamma-1} e^{-\sum_{\zeta=\eta+1}^{\gamma-1}\Delta_\zeta \frac{M\lambda_{M+1,\alpha,M}}{\chi_\zeta}}(1-e^{-(t_\gamma-T_\gamma)\frac{M\lambda_{M+1,\alpha,M}}{\chi_\gamma}})\frac{(1-e^{-\Delta_\eta \frac{M\lambda_{M+1,\alpha,M}}{\chi_\eta}})}{\frac{M\lambda_{M+1,\alpha,M}}{\chi_\eta}} + ((t_\gamma-T_\gamma)-\frac{(1-e^{-(t_\gamma-T_\gamma)\frac{M\lambda_{M+1,\alpha,M}}{\chi_\gamma}})}{\frac{M\lambda_{M+1,\alpha,M}}{\chi_\gamma}}))$$

(A.63)

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

$$= \frac{P_\gamma}{t_\gamma}\frac{\lambda_{(n+1),\alpha,2}(1-e^{-(T_{\gamma+1}-t_\gamma)\frac{\lambda_{2,\alpha}}{\chi_\gamma}})}{M(\lambda_{(n+1),\alpha,2}+\lambda_{(n+1),\alpha,1})}$$
$$(\sum_{\eta=1}^{\gamma-1} e^{-\sum_{\zeta=\eta+1}^{\gamma-1}\Delta_\zeta \frac{M\lambda_{M+1,\alpha,M}}{\chi_\zeta}} e^{-(t_\gamma-T_\gamma)\frac{M\lambda_{M+1,\alpha,M}}{\chi_\gamma}} \frac{(1-e^{-\Delta_\eta \frac{M\lambda_{M+1,\alpha,M}}{\chi_\eta}})}{\frac{M\lambda_{M+1,\alpha,M}}{\chi_\eta}} + \frac{(1-e^{-\Delta_\gamma \frac{M\lambda_{M+1,\alpha,M}}{\chi_\gamma}})}{\frac{M\lambda_{M+1,\alpha,M}}{\chi_\gamma}})dt$$

(A.64)

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

$$P(\gamma, i|\gamma, j) = \frac{P_\gamma}{t_\gamma} \int_0^{t_\gamma} \frac{n\lambda_{(n+1),\alpha,2} e^{-\int_u^{t_\gamma} \frac{M\lambda_{M+1,\alpha,M}}{\chi_v} dv}}{M(\lambda_{(n+1),\alpha,2} + \lambda_{(n+1),\alpha,1})} du$$

$$= \frac{P_\gamma}{t_\gamma} \frac{n\lambda_{(n+1),\alpha,2}}{M(\lambda_{(n+1),\alpha,2} + \lambda_{(n+1),\alpha,1})} (\sum_{\eta=1}^{\gamma-1} \int_{T_\eta}^{T_{\eta+1}} e^{-\int_u^{T_{\eta+1}} \frac{M\lambda_{M+1,\alpha,M}}{\chi_v} dv} e^{-\int_{T_{\eta+1}}^{t_\gamma} \frac{M\lambda_{M+1,\alpha,M}}{\chi_v} dv} du$$

$$+ \int_{T_\gamma}^{t_\gamma} e^{-\int_u^{t_\gamma} \frac{M\lambda_{M+1,\alpha,M}}{\chi_v} dv} du)$$

$$= \frac{P_\gamma}{t_\gamma} \frac{n\lambda_{(n+1),\alpha,2}}{M(\lambda_{(n+1),\alpha,2} + \lambda_{(n+1),\alpha,1})} (\sum_{\eta=1}^{\gamma-1} \frac{(1 - e^{-\Delta_\eta \frac{M\lambda_{M+1,\alpha,M}}{\chi_\eta}})}{\frac{M\lambda_{M+1,\alpha,M}}{\chi_\eta}} e^{-\int_{T_{\eta+1}}^{t_\gamma} \frac{M\lambda_{M+1,\alpha,M}}{\chi_v} dv}$$

$$+ \frac{(1 - e^{-(t_\gamma - T_\gamma) \frac{M\lambda_{M+1,\alpha,M}}{\chi_\gamma}})}{\frac{M\lambda_{M+1,\alpha,M}}{\chi_\gamma}})$$

(A.65)

$\gamma = \beta, m = n$

$$P(\gamma, j|\gamma, j) = 1 - \sum_{\beta \neq \gamma, i \neq j} P(\beta, i|\gamma, j) \tag{A.66}$$

Emission Matrix

M=3 For $M = 3$, only 4 types of observations. Observation 0, no mutation observed at this position. 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 :

$$P(0|\gamma) = e^{-\mu(Ts)} \tag{A.67}$$

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

$$P(i|\gamma, i) = (1 - e^{-\mu(Ts - 2t_\gamma)}) \tag{A.68}$$

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

$$P(i|\gamma, \bar{i}) = (1 - e^{-\mu(2t_\gamma)}) \tag{A.69}$$

if the first coalescent event involves 3 individuals we have :

$$P(0|\gamma) = e^{-\mu(3t_\gamma)} \tag{A.70}$$

$i \in 1, 2, 3$

$$P(i|\gamma) = 1 - e^{-\mu(3t_\gamma)} \tag{A.71}$$

M=4 If $M = 4$, one can only observe 8 possibilities as we assume an infinite site model. Observation 0, no mutation observed. Observation 1 (2,3,4), individual 1 (2,3,4) is different from all other individual. Observation 5 to 7 , two individuals are different from the other two.

If the first coalescent event involves 4 individuals:

$$P(0|\gamma) = e^{-\mu(4t_\gamma)} \tag{A.72}$$

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

$$P(i|\gamma) = 1 - e^{-\mu(4t_\gamma)} \tag{A.73}$$

If the first coalescent event involves 3 individuals:

$$P(0|\gamma) = e^{-\mu(Ts)} \tag{A.74}$$

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

$$P(i|\gamma, \bar{i}) = (1 - e^{-\mu(t_\gamma)}) \tag{A.75}$$

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

$$P(i|\gamma, i) = (1 - e^{-\mu(Ts-3t_\gamma)}) \tag{A.76}$$

If the first coalescent event involves 2 individuals:

No mutation occurred.

$$P(0|\gamma) = e^{-\mu(Ts)} \tag{A.77}$$

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

$$P(i|\gamma) = (1 - e^{-\mu(t_\gamma)}) \tag{A.78}$$

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

$$P(i|\gamma) = (1 - e^{-\mu(\frac{(Ts-4t_\gamma)}{2}+t_\gamma)}) \tag{A.79}$$

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

$$P(i|\gamma) = e^{-\mu(Ts)} \tag{A.80}$$

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

$$P(i|\gamma) = 0 \tag{A.81}$$

A.4 Appendix of Chapter 5

A.4.1 Supplementary Figures

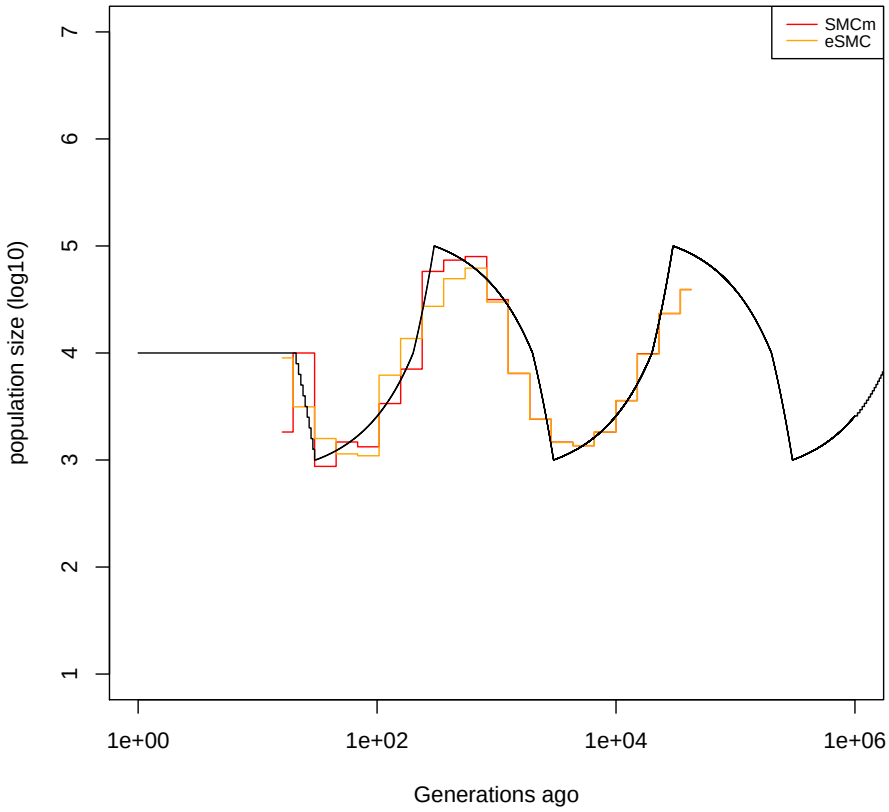

Figure A.23: **Performance of eSMC and SMCm under a sawtooth scenario.** Estimated demographic history by SMCm and eSMC using 10 scaffolds each of 100 Mb with sample size 2 (respectively in red and orange) under a sawtooth scenario (black). The recombination and mutation rate are set to 1×10^{-8} per generation per bp and the methylation and demethylation rate are respectively set to to 1×10^{-3} and 5×10^{-3} per generation per bp.

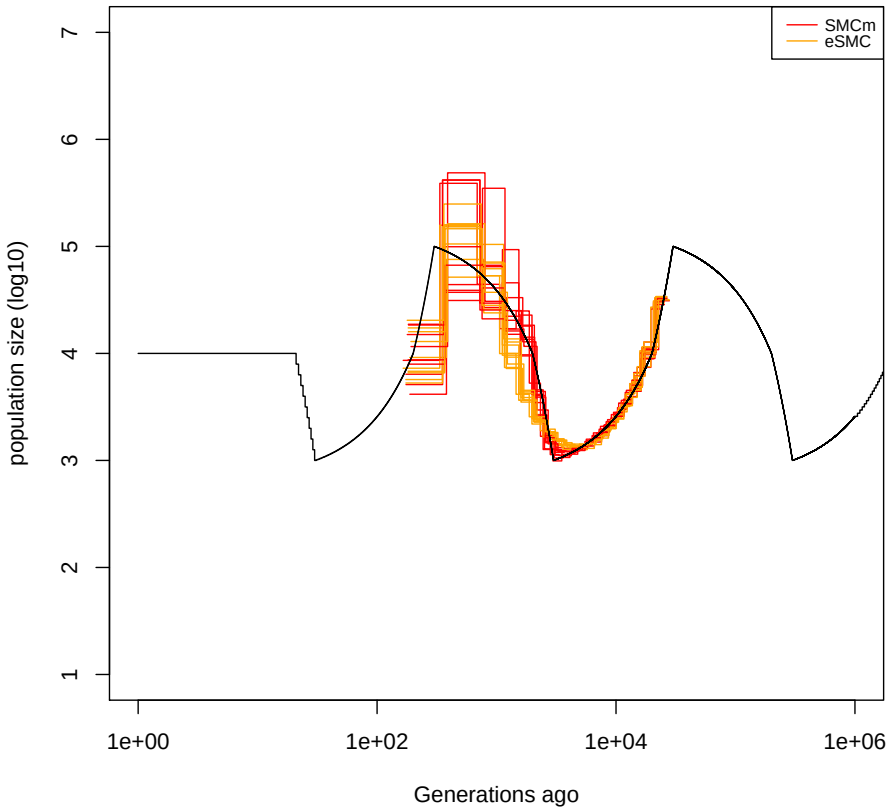

Figure A.24: **Performance of eSMC and SMCm under a sawtooth scenario.** Estimated demographic history by SMCm and eSMC using 2 sequence of 100 Mb (respectively in red and orange) under a sawtooth scenario (black). The recombination and mutation rate are set to 1×10^{-8} per generation per bp and the methylation and demethylation rate are respectively set to to 1×10^{-4} and 5×10^{-4} per generation per bp.

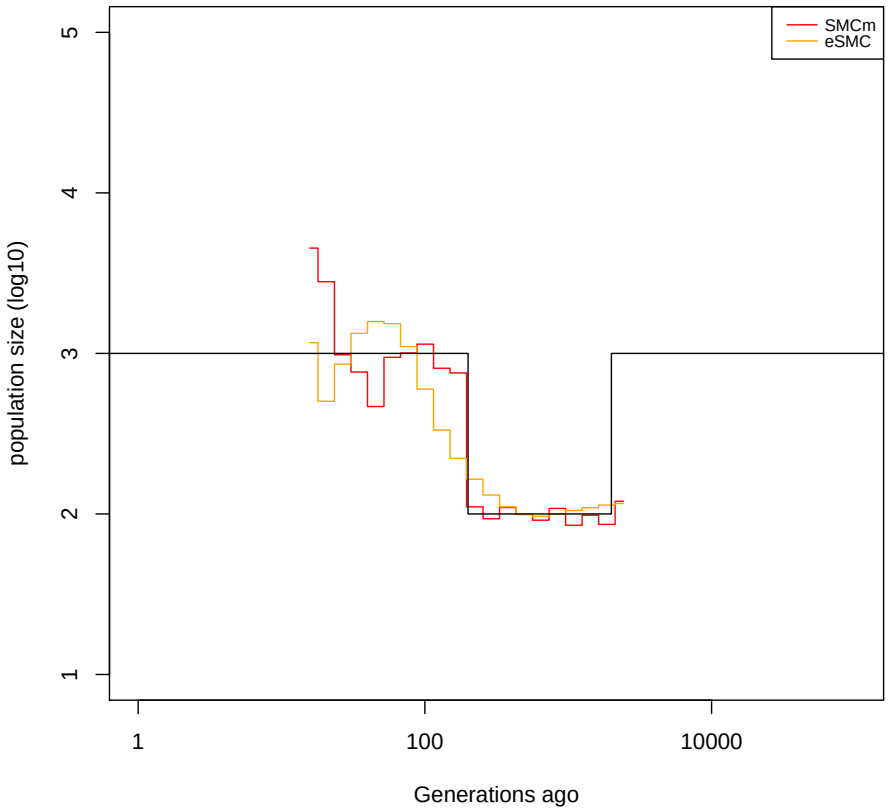

Figure A.25: **Performance of eSMC and SMCm under a bottleneck scenario.** Estimated demographic history by SMCm and eSMC using 10 scaffolds each of 100 Mb with sample size 2 (respectively in red and orange) under a recent bottleneck (black) and where current population size is 1000. The recombination and mutation rate are set to 1×10^{-8} per generation per bp and the methylation and demethylation rate are respectively set to to 1×10^{-4} and 5×10^{-4} per generation per bp.

A.4.2 Theorem 1

When mutation rate, sample size, coalescent times are to important, the infinite site model might no longer be valid. We therefore build a model to describe the distribution of nucleotides under a finite site model. To do so, we build a recursive

formula (along the coalescent tree) to calculate the distribution of nucleotides of the sample at one position. We first model a sequence, at each position and with probability 0.25 the common ancestor (ancestral nucleotidic state) is attributed a nucleotide Xi (*i.e.* A,T,C,G). Then a coalescence event happens and a new individual appears. Its nucleotide is selected with probability equal to its proportion in the sample at that time.

Second step, mutations are added. Let's Note $X_t = (X1_t, X2_t, X3_t, X4_t)$ the number of each nucleotide at step t. Let T_t be the coalescent time at step t. The number of each nucleotide is the sum of t+1 random variable. Four distributions are possible depending on the original nucleotide. The distribution of X_t is thus the sum of four independent random variable conditional to the number of each nucleotide at time t-1. Each of these variable follow a multinomial distribution. We can thus use the convolution formula which gives us :

$$P(X_t|T, X_{t-1}) = \sum_{k=1}^{4} p_{t-1,k} \times (P(X_t|T, X_{k,t-1})) \tag{A.82}$$

Where $X_{k,t-1}$ is X_{t-1} +1 for the number of nucleotide k (new individual). $p_{t-1,k}$ is the probability conditional to X_{t-1} that the new individual is of type k.

This gives :

$$P(X_t|T, X_{t-1}) = \sum_{k=1}^{4} p_{t-1,k} \times (\sum_{n11=0}^{X1_{k,t-1}} .... \sum_{n44=0}^{X4_{k,t-1}} \frac{\prod_{i=1}^{4} Xi_{k,t-1}!}{\prod_{i,j=1}^{4} nij!} \prod_{i,j=1}^{4} p_{ij}^{nij}) \tag{A.83}$$

Where n_{ij} is the number of nucleotides of type i that become of type j. and p_{ij} the probability of turning to nucleotide j from nucleotide i. This formula comes from the fact that each individual (before adding mutations) will give an individual. We therefore have to pick every individual, and sum over all possible combination which will give us X_t

We have $p_{ii} = (0.25 + 0.75 \times e^{-\mu T})$ that we call p and $p_{i \neq j} = (0.25 \times (1 - e^{-\mu T}))$

Thus :

$$P(X_t|T, X_{t-1}) = \sum_{k=1}^{4} p_{t-1,k} \times (\sum_{n11=0}^{X1_{k,t-1}} .... \sum_{n44=0}^{X4_{k,t-1}} \frac{\prod_{i=1}^{4} Xi_{k,t-1}!}{\prod_{i,j=1}^{4} nij!} p^{\sum_{i=1}^{4} nii} (1-p)^{t - \sum_{i=1}^{4} nii}) \tag{A.84}$$

Using Newton formula (and where N= $\sum_{i=1}^{4} nii$):

$$P(X_t|T, X_{t-1}) = \sum_{k=1}^{4} p_{t-1,k} \times (\sum_{n11=0}^{X1_{k,t-1}} .... \sum_{n44=0}^{X4_{k,t-1}} \frac{\prod_{i=1}^{4} Xi_{k,t-1}!}{\prod_{i,j=1}^{4} nij!} \frac{1}{4}^{t} \sum_{d=0}^{N} \sum_{l=0}^{t-N} \binom{N}{d} \binom{t-N}{l} 3^{d} (-1)^{l} e^{-\mu T(d+l)}) \tag{A.85}$$

The integral of a sum is the sum of integral, thus we can integrate over time. Since T follows an exponential distribution of parameter $\binom{t}{2}$, it gives us :

$$P(X_t|X_{t-1}) = \sum_{k=1}^{4} p_{t-1,k} \times (\sum_{n11=0}^{X1_{k,t-1}} .... \sum_{n44=0}^{X4_{k,t-1}} \frac{\prod_{i=1}^{4} Xi_{k,t-1}!}{\prod_{i,j=1}^{4} nij!} \frac{1}{4}^t \sum_{d=0}^{N} \sum_{l=0}^{t-N} \binom{N}{d} \binom{t-N}{l} \frac{3^d(-1)^l \binom{t}{2}}{\binom{t}{2} + \mu(d+l)}) \tag{A.86}$$

Of course we only sum terms where :

- $\sum_{i=1}^{4} nij = Xj_t$
- $\sum_{j=1}^{4} nij = Xi_{t-1}$

Using the total probability formula:

$$P(X_t) = \sum_{X_{t-1}} P(X_t|X_{t-1})P(X_{t-1}) \tag{A.87}$$

Here we have a recursive formula to calculate the distribution.

A.4.3 Theorem 2

We here integrate epimutation (*i.e.* methylation and demethylation of cytosine) to what has been done in theorem 1. We assume that epimutations and mutations are independent and that we are at a position which can be methylated (otherwise cf theorem 1). In addition, we assume that if a nucleotide is replaced by a cytosine, the cytosine is unmethylated. We now have with probability $0.25P_m$ for the initial nucleotide to be a C^* (methylated) and C (unmethylated) with probability 0.25(1-P_m) (assuming we are at equilibrium). Where P_m is defined as:

$$P_m = \frac{\mu_m}{\mu_m + \mu_d} \tag{A.88}$$

Where, μ_m is the methylation rate and μ_d the demethyation rate.

We respectively note X_1,X_2,X_3,X_4,X_5 as A,T,G,C,C*. We first model a sequence, at each position and with probability 0.25 the common ancestor is attributed a nucleotide Xi (*i.e.* A,T,C,G). Then a coalescence event happens and a new individual appears. Its nucleotide is selected with probability equal to its proportion in the sample at that time.

Second step, mutations are added and then epimutations. Let's Note the number of each nucleotide at step t : $X_t = (X1_t, X2_t, X3_t, X4_t, X5_t)$. Let T_t be the

coalescent time at step t. The number of each nucleotide is the sum of t+1 random variable. Four distributions are possible depending on the original nucleotide. The distribution of X_t is thus the sum of four independent random variable conditional to the number of each nucleotide at time t-1. Each of these variable follow a multinomial distribution. We can thus use the convolution formula which gives us :

$$P(X_t|T, X_{t-1}) = \sum_{k=1}^{5} p_{t-1,k} \times (P(X_t|T, X_{k,t-1})) \tag{A.89}$$

Where $X_{k,t-1}$ is X_{t-1} +1 for the number of nucleotide k (new individual). $p_{t-1,k}$ is the probability conditional to X_{t-1} that the new individual is of type k.This gives :

$$P(X_t|T, X_{t-1}) = \sum_{k=1}^{5} p_{t-1,k} \times (\sum_{n11=0}^{X1_{k,t-1}} .... \sum_{n44=0}^{X4_{k,t-1}} \frac{\prod_{i=1}^{4} Xi_{k,t-1}!}{\prod_{i,j=1}^{4} nij!} \prod_{i,j=1}^{4} p_{ij}^{nij}) \tag{A.90}$$

Where n_{ij} is the number of nucleotide of type i that become type j. and p_{ij} the probability of turning to nucleotide j from nucleotide i. We therefore have to pick every individual, and sum over all possible combination which will give us X_t. We have:

- $p_{ii} = (0.25 + 0.75 \times e^{-\mu T})$, if i $\in$ {1,2,3}
- $p_{i\neq j} = (0.25 \times (1 - e^{-\mu T}))$, if j $\in$ {1,2,3}
- $p_{ii} = (0.25 + 0.75 \times e^{-\mu T})(\frac{\mu_d}{\mu_m+\mu_d}(1 + e^{-T(\mu_m+\mu_d)}))$, if i $\in$ {4}
- $p_{ii} = (0.25 + 0.75 \times e^{-\mu T})(\frac{\mu_m}{\mu_m+\mu_d}(1 + e^{-T(\mu_m+\mu_d)}))$, if i $\in$ {5}
- $p_{i\neq j} = (0.25+0.75\times e^{-\mu T})(\frac{\mu_d}{\mu_m+\mu_d}(1-e^{-T(\mu_m+\mu_d)}))$, if j $\in$ {4} and if i $\in$ {5}
- $p_{i\neq j} = (0.25+0.75\times e^{-\mu T})(\frac{\mu_d}{\mu_m+\mu_d}(1-e^{-T(\mu_m+\mu_d)}))$, if j $\in$ {5} and if i $\in$ {4}
- $p_{i\neq j} = \int_0^T \frac{1}{T}(0.25+0.75\times e^{-\mu t})((\frac{\mu_d}{\mu_m+\mu_d} + \frac{\mu_m}{\mu_m+\mu_d}e^{-(T-t)(\mu_m+\mu_d)})dt$, if j $\in$ {4} and if i $\in$ {1,2,3}
- $p_{i\neq j} = \int_0^T \frac{1}{T}(0.25+0.75\times e^{-\mu t})(\frac{\mu_m}{\mu_m+\mu_d}(1-e^{-(T-t)(\mu_m+\mu_d)})dt$, if j $\in$ {5} and if i $\in$ {1,2,3}

From here, the recursive formulas can be obtained using the similar approach as in theorem 1.

A.4.4 Model description of SMCm

SMCm derives from eSMC except for the emission matrix. Hence, for the general HMM description and optimization of the likelihood, cf A.1.

Emission Matrix with sequence and methylation polymorphism

Since the model accounts for sequence and methylation polymorphisms, there are at each position 5 different possible observations when comparing two sequences. The first observation is 0 , corresponding to a non-methylable site where the two nucleotides are identical. 1, if the two nucleotide are different. 2 if it's a methylable site and both are unmethylated. 3, if the site is methylable and both are methylated. Finally, 4 is it's a methylable site and one cytosine is methylated and the other unmethylated. Therefore, from theorem 2 with sample size 2, after approximating the formula assuming methylation state is not affected by mutations we have the following formula:

$$
\begin{aligned}
P(0|\gamma) &= e^{-2\mu t_\gamma} \\
P(1|\gamma) &= 1 - e^{-2\mu t\gamma} \\
P(2|\gamma) &= ((p_d \times ((p_d + ((1-p_d)e^{(-(\mu_d+\mu_m)\times Tc\times Ne))}) \times (p_d + ((1-p_d)e^{(-\theta_m))})) \\
&\quad +((1-p_d) \times (1 - ((1-p_d) + (p_d e^{(-\theta_m))})) \times (1 - ((1-p_d) + (p_d e^{(-\theta_m))})))) \\
P(3|\gamma) &= ((p_d \times ((1 - (p_d + ((1-p_d)e^{(-\theta_m))})) \times (1 - (p_d + ((1-p_d)e^{(-\theta_m))})))) \\
&\quad +((1-p_d) \times (((1-p_d) + (p_d e^{(-\theta_m))})) \times (((1-p_d) + (p_d e^{(-\theta_m))})))) \\
P(4|\gamma) &= ((p_d \times (2 \times (p_d + ((1-p_d)e^{(-\theta_m))}) \times (1 - (p_d + ((1-p_d)e^{(-\theta_m))})))) \\
&\quad +((1-p_d) \times (2 \times ((1-p_d) + (p_d e^{(-\theta_m))}) \times (1 - ((1-p_d) + (p_d e^{(-\theta_m))}))))) \\
p_d &= \frac{\mu_d}{\mu_d + \mu_m} \\
\theta_m &= (\mu_d + \mu_m) \times Tc \times Ne
\end{aligned}
\tag{A.91}
$$

Where μ is the mutation rate per nucleotide per N generation,μ_m the methylation rate per generation,μ_d the demethylation rate per generation and $t\gamma$ the average coalescent time in state γ.

Bibliography

[1] RJ Abbot and MF Gomes. Population genetic-structure and outcrossing rate of Arabidopsis-thaliana (L) HEYNH. *Heredity*, 62(3):411–418, JUN 1989.

[2] V Alekseev and W Lampert. Maternal control of resting-egg production in Daphnia. *Nature*, 414(6866):899–901, DEC 20 2001.

[3] Einar Arnason and Katrin Halldorsdottir. Nucleotide variation and balancing selection at the Ckma gene in Atlantic cod: analysis with multiple merger coalescent models. *PEERJ*, 3, FEB 24 2015.

[4] S.C.H. Barrett. The evolution of plant reproductive systems: how often are transitions irreversible? *Proceedings of the Royal Society B-Biological Sciences*, 280(1765), AUG 22 2013.

[5] Barrett, Spencer C. H. and Arunkumar, Ramesh and Wright, Stephen I. The demography and population genomics of evolutionary transitions to self-fertilization in plants. *Philosophical Transactions of the Royal Society B-Biological Sciences*, 369(1648), AUG 5 2014.

[6] Gustavo V. Barroso, Natasa Puzovic, and Julien Y. Dutheil. Inference of recombination maps from a single pair of genomes and its application to ancient samples. *PLOS Genetics*, 15(11), NOV 2019.

[7] Carol C. Baskin and Jerry M. Baskin. Germination Ecology of Seeds in the Persistent Seed Bank. In *Seeds: Ecology, Biogeography, and Evolution of Dormancy and Germination, 2ND EDITION*, pages 187–276. 2014.

[8] Champak R. Beeravolu, Michael J. Hickerson, Laurent A. F. Frantz, and Konrad Lohse. ABLE: blockwise site frequency spectra for inferring complex population histories and recombination. *Genome Biology*, 19, 2018.

[9] Annabel C. Beichman, Tanya N. Phung, and Kirk E. Lohmueller. Comparison of Single Genome and Allele Frequency Data Reveals Discordant Demographic Histories. *G3-Genes Genomes Genetics*, 7(11):3605–3620, NOV 2017.

[10] Matthias Birkner, Jochen Blath, and Bjarki Eldon. An Ancestral Recombination Graph for Diploid Populations with Skewed Offspring Distribution. *Genetics*, 193(1):255–290, JAN 2013.

[11] Matthias Birkner, Huili Liu, and Anja Sturm. Coalescent results for diploid exchangeable population models l. *Electronic Journal of Probability*, 23, 2018.

[12] Jochen Blath, Adrian Gonzalez Casanova, Noemi Kurt, and Maite Wilke-Berenguer. The seed bank coalescent with simultaneous switching. *Electronic Journal of Probability*, 25, 2020.

[13] Paul D. Blischak, Michael S. Barker, and Ryan N. Gutenkunst. Inferring the Demographic History of Inbred Species from Genome-Wide SNP Frequency Data. *Molecular Biology and Evolution*, 37(7):2124–2136, JUL 2020.

[14] L Brendonck and L De Meester. Egg banks in freshwater zooplankton: evolutionary and ecological archives in the sediment. *Hydrobiologia*, 491(1-3):65–84, JAN 2003.

[15] Sharon R. Browning, Brian L. Browning, Ying Zhou, Serena Tucci, and Joshua M. Akey. Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. *Cell*, 173(1):53+, MAR 22 2018.

[16] 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.

[17] 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.

[18] Lynsey Bunnefeld, Jack Hearn, Graham N. Stone, and Konrad Lohse. Whole-genome data reveal the complex history of a diverse ecological community. *Proceedings of the National Academy of Sciences of the United States of America*, 115(28):E6507–E6515, JUL 10 2018.

[19] Adrián González Casanova, Verónica Miró Pina, and Arno Siri-Jégousse. The Symmetric Coalescent and Wright-Fisher models with bottlenecks. *arXiv:1903.05642 [math]*, September 2020. arXiv: 1903.05642.

[20] Dan Chang and Beth Shapiro. Using ancient DNA and coalescent-based methods to infer extinction. *Biology Letters*, 12(2), FEB 1 2016.

[21] Brian Charlesworth and Kavita Jain. Purifying Selection, Drift, and Reversible Mutation with Arbitrarily High Mutation Rates. *Genetics*, 198(4):1587+, DEC 2014.

[22] Lounes Chikhi, Willy Rodriguez, Simona Grusea, Patricia Santos, Simon Boitard, and Olivier Mazet. The IICR (inverse instantaneous coalescence rate) as a summary of genomic diversity: insights into demographic inference and model choice. *Heredity*, 120(1):13–24, JAN 2018.

[23] Alec J. Coffman, Ping Hsun Hsieh, Simon Gravel, and Ryan N. Gutenkunst. Computationally Efficient Composite Likelihood Statistics for Demographic Inference. *Molecular Biology and Evolution*, 33(2):591–593, FEB 2016.

[24] Bellot S Dann M, Schaeffer H Schepella S, and Tellier A. Mutation rates in seeds and seed-banking influence substitution rates across the angiosperm phylogeny. *bioRxiv*, 2017.

[25] de Manuel etl al. The evolutionary history of extinct and living lions. *Proceedings of the National Academy of Sciences of the United States of America*, 117(20):10927–10934, MAY 19 2020.

[26] James H. Degnan and Noah A. Rosenberg. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *TRENDS IN ECOLOGY & EVOLUTION*, 24(6):332–340, JUN 2009.

[27] Jean-Stephane Dhersin, Fabian Freund, Arno Siri-Jegousse, and Linglong Yuan. On the length of an external branch in the Beta-coalescent. *Stochastic Processes and their Applications*, 123(5):1691–1715, MAY 2013.

[28] P Donnelly and TG Kurtz. Particle representations for measure-valued population models. *Annals of Probability*, 27(1):166–205, JAN 1999.

[29] 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.

[30] Dieter Ebert. *Ecology, epidemiology, and evolution of parasitism in Daphnia*. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information., 2005.

[31] Scott V. Edwards. IS A NEW AND GENERAL THEORY OF MOLECULAR SYSTEMATICS EMERGING? *Evolution*, 63(1):1–19, JAN 2009.

[32] 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.

[33] Bjarki Eldon, Matthias Birkner, Jochen Blath, and Fabian Freund. Can the Site-Frequency Spectrum Distinguish Exponential Population Growth from Multiple-Merger Coalescents? *Genetics*, 199(3):841+, MAR 2015.

[34] Hans Ellegren and Nicolas Galtier. Determinants of genetic diversity. *Nature Reviews Genetics*, 17(7):422–433, JUL 2016.

[35] Bergstrom et al. Insights into human genetic variation and population history from 929 diverse genomes. *Science*, 367(6484, SI):1339+, MAR 20 2020.

[36] Cao et al. Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nature Genetics*, 43(10):956–U60, OCT 2011.

[37] Choo et al. Pangolin genomes and the evolution of mammalian scales and immunity. *Genome Research*, 26(10):1312–1322, OCT 2016.

[38] Cokus et al. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature*, 452(7184):215–219, MAR 13 2008.

[39] Durvasula et al. African genomes illuminate the early history and transition to selfing in Arabidopsis thaliana. *Proceedings of the National Academy of Sciences of the United States of America*, 114(20):5213–5218, MAY 16 2017.

[40] Ekblom et al. Genome sequencing and conservation genomics in the Scandinavian wolverine population. *Conservation Biology*, 32(6):1301–1312, DEC 2018.

[41] Fitak et al. Genomic signatures of domestication in Old World camels. *Communications Biology*, 3(1), JUN 19 2020.

[42] Hendricks et al. Recent advances in conservation and population genomics data analysis. *Evolutionary Applications*, 11(8):1197–1211, SEP 2018.

[43] Humble et al. Chromosomal-level genome assembly of the scimitar-horned oryx: Insights into diversity and demography of a species extinct in the wild. *Molecular Ecology Resources*.

[44] Li et al. Genomic signatures of near-extinction and rebirth of the crested ibis and other endangered bird species. *Genome Biology*, 15(12), 2014.

[45] Mailund et al. A New Isolation with Migration Model along Complete Genomes Infers Very Different Divergence Processes among Closely Related Great Ape Species. *PLOS Genetics*, 8(12), DEC 2012.

[46] Malaspinas et al. A genomic history of Aboriginal Australia. *Nature*, 538(7624):207+, OCT 13 2016.

[47] Mattle-Greminger et al. Genomes reveal marked differences in the adaptive evolution between orangutan species. *Genome Biology*, 19, NOV 15 2018.

[48] McKenna et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303, SEP 2010.

[49] Moura et al. Phylogenomics of the genus Tursiops and closely related Delphininae reveals extensive reticulation among lineages and provides inference about eco-evolutionary drivers. *Molecular Phylogenetics and Evolution*, 146, MAY 2020.

[50] Niederhuth et al. Widespread natural variation of DNA methylation within angiosperms. *Genome Biology*, 17, SEP 27 2016.

[51] Palkopoulou et al. Complete Genomes Reveal Signatures of Demographic and Genetic Declines in the Woolly Mammoth. *Current Biology*, 25(10):1395–1400, MAY 18 2015.

[52] Palkopoulou et al. A comprehensive genomic history of extinct and living elephants. *Proceedings of the National Academy of Sciences of the United States of America*, 115(11):E2566–E2574, MAR 13 2018.

[53] Patton et al. Contemporary Demographic Reconstruction Methods Are Robust to Genome Assembly Quality: A Case Study in Tasmanian Devils. *Molecular Biology and Evolution*, 36(12):2906–2921, DEC 2019.

[54] Prado-Martinez et al. Great ape genetic diversity and population history. *Nature*, 499(7459):471–475, JUL 25 2013.

[55] Tang et al. The evolution of selfing in Arabidopsis thaliana. *Science*, 317(5841):1070–1072, AUG 24 2007.

[56] Wang et al. Glaciation-based isolation contributed to speciation in a Palearctic alpine biodiversity hotspot: Evidence from endemic species. *Molecular Phylogenetics and Evolution*, 129:315–324, DEC 2018.

[57] Yang et al. Chromosome-level reference genome assembly and gene editing of the dead-leaf butterfly Kallima inachus. *Molecular Ecology Resources*.

[58] Ye et al. A New Reference Genome Assembly for the Microcrustacean Daphnia pulex. *G3-Genes Genomes Genetics*, 7(5):1405–1416, MAY 2017.

[59] Yew et al. Genomic structure of the native inhabitants of Peninsular Malaysia and North Borneo suggests complex human population history in Southeast Asia. *Human Genetics*, 137(2):161–173, FEB 2018.

[60] You et al. Variation among 532 genomes unveils the origin and evolutionary history of a global insect herbivore. *Nature Communications*, 11(1), MAY 8 2020.

[61] Zhang et al. Genome-wide high-resolution mapping and functional analysis of DNA methylation in Arabidopsis. *Cell*, 126(6):1189–1201, SEP 22 2006.

[62] Margaret E. K. Evans, Regis Ferriere, Michael J. Kane, and D. Lawrence Venable. Bet hedging via seed banking in desert evening primroses (Oenothera, Onagraceae): Demographic evidence from natural populations. *American Naturalist*, 169(2):184–194, FEB 2007.

[63] MEK Evans and JJ Dennehy. Germ banking: Bet-hedging and varlable release from egg and seed dormancy. *Quarterly Review of Biology*, 80(4):431–451, DEC 2005.

[64] Adam D. Ewing. Transposable element detection from whole genome sequence data. *MOBILE DNA*, 6, DEC 29 2015.

[65] Suhua Feng, Steven E. Jacobsen, and Wolf Reik. Epigenetic Reprogramming in Plant and Animal Development. *Science*, 330(6004):622–627, OCT 29 2010.

[66] Jullien M. Flynn, Frederic J. J. Chain, Daniel J. Schoen, and Melania E. Cristescu. Spontaneous Mutation Accumulation in Daphnia pulex in Selection-Free vs. Competitive Environments. *Molecular Biology and Evolution*, 34(1):160–173, JAN 2017.

[67] Fabian Freund. Cannings models, population size changes and multiple-merger coalescents. *Journal of Mathematical Biology*, 80(5):1497–1521, APR 2020.

[68] Fabian Freund and Arno Siri-Jégousse. The impact of genetic diversity statistics on model selection between coalescents. *Computational Statistics & Data Analysis*, page 107055, 2020.

[69] Andrea Fulgione, Maarten Koornneef, Fabrice Roux, Joachim Hermisson, and Angela M. Hancock. Madeiran Arabidopsis thaliana Reveals Ancient Long-Range Colonization and Clarifies Demography in Eurasia. *Molecular Biology and Evolution*, 35(3):564–574, MAR 2018.

[70] 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.

[71] Lucie Gattepaille, Torsten Guenther, and Mattias Jakobsson. Inferring Past Effective Population Size from Distributions of Coalescent Times. *Molecular Biology and Evolution*, 204(3):1191+, NOV 2016.

[72] Brandon S. Gaut, Danelle K. Seymour, Qingpo Liu, and Yongfeng Zhou. Demography and its effects on genomic variation in crop domestication. *Nature Plants*, 4(8):512–520, AUG 2018.

[73] Ryan N. Gutenkunst, Ryan D. Hernandez, Scott H. Williamson, and Carlos D. Bustamante. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLOS Genetics*, 5(10), OCT 2009.

[74] John Hawks. Introgression Makes Waves in Inferred Histories of Effective Population Size. *Human Biology*, 89(1):67–80, JAN 2017.

[75] Luke B. B. Hecht, Peter C. Thompson, and Benjamin M. Rosenthal. Comparative demography elucidates the longevity of parasitic and symbiotic relationships. *Proceedings Of the Royal Society B-Biological Sciences*, 285(1888), OCT 10 2018.

[76] Dennis Hedgecock and Alexander 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.

[77] Lukas Heinrich, Johannes Mueller, Aurelien Tellier, and Daniel 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.

[78] Chizue Hiruta, Chizuko Nishida, and Shin Tochinai. Abortive meiosis in the oogenesis of parthenogenetic Daphnia pulex. *Chromosome Research*, 18(7):833–840, NOV 2010.

[79] Chizue Hiruta and Shin Tochinai. Spindle Assembly and Spatial Distribution of gamma-tubulin During Abortive Meiosis and Cleavage Division in the Parthenogenetic Water Flea Daphnia pulex. *Zoological Science*, 29(11):733–737, NOV 2012.

[80] Asger Hobolth, Ole F. Christensen, Thomas Mailund, and Mikkel H. Schierup. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. *PLOS Genetics*, 3(2):294–304, FEB 2007.

[81] Asger Hobolth and Jens Ledet Jensen. Markovian approximation to the finite loci coalescent with recombination along multiple sequences. *Theoretical Population Biology*, 98:48–58, DEC 2014.

[82] Asger Hobolth, Arno Siri-Jegousse, and Mogens Bladt. Phase-type distributions in population genetics. *Theoretical Population Biology*, 127:16–32, JUN 2019.

[83] Zhe Hou, Ang Li, and Jianguo Zhang. Genetic architecture, demographic history, and genomic differentiation of Populus davidiana revealed by whole-genome resequencing. *Evolutionary Applications*.

[84] RR HUDSON. Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, 23(2):183–201, 1983.

[85] RR Hudson. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338, FEB 2002.

[86] Philippe Jarne and Josh R. Auld. Animals mix it up too: The distribution of self-fertilization among hermaphroditic animals. *Evolution*, 60(9):1816–1824, SEP 2006.

[87] James E. Johndrow and Julia A. Palacios. Exact limits of inference in coalescent models. *Theoretical Population Biology*, 125:75–93, FEB 2019.

[88] I Kaj, SM Krone, and M Lascoux. Coalescent theory for seed bank models. *Journal of Applied Probability*, 38(2):285–300, JUN 2001.

[89] Marty Kardos, Anna Qvarnstrom, and Hans Ellegren. Inferring Individual Inbreeding and Demographic History from Segments of Identity by Descent in Ficedula Flycatcher Genome Sequences. *Genetics*, 205(3):1319–1334, MAR 2017.

[90] Mamoru Kato, Daniel A. Vasco, Ryuichi Sugino, Daichi Narushima, and Alexander 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.

[91] Jerome Kelleher, Alison M. Etheridge, and Gilean McVean. Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLOS Computational Biology*, 12(5), MAY 2016.

[92] Jerome Kelleher, Yan Wong, Anthony W. Wohns, Chaimaa Fadil, Patrick K. Albers, and Gil McVean. Inferring whole-genome histories in large population datasets (vol 51, pg 1330, 2019). *Nature Genetics*, 51(11):1660, NOV 2019.

[93] Envel Kerdaffrec, Daniele L. Filiault, Arthur Korte, Eriko Sasaki, Viktoria Nizhynska, Uemit Seren, and Magnus Nordborg. Multiple alleles at a single locus control seed dormancy in Swedish Arabidopsis. *ELife*, 5, DEC 14 2016.

[94] Younhun Kim, Frederic Koehler, Ankur Moitra, Elchanan Mossel, and Govind 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.

[95] JFC Kingman. The Coalescent . *Stochastic Processes and their Applications*, 13, 1982.

[96] Robert Kofler. SimulaTE: simulating complex landscapes of transposable elements of populations. *Bioinformatics*, 34(8):1419–1420, APR 15 2018.

[97] Jere Koskela. Multi-locus data distinguishes between population growth and multiple merger coalescents. *STATISTICAL APPLICATIONS IN GENETICS AND MOLECULAR BIOLOGY*, 17(3), JUN 2018.

[98] Jere Koskela and Maite Wilke Berenguer. Robust model selection between population growth and multiple merger coalescents. *Mathematical Biosciences*, 311:1–12, MAY 2019.

[99] Sally C. Y. Lau, Nerida G. Wilson, Catarina N. S. Silva, and Jan M. Strugnell. Detecting glacial refugia in the Southern Ocean. *ECOGRAPHY*, 43(11):1639–1656, NOV 2020.

[100] Julie A. Law and Steven E. Jacobsen. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature Reviews Genetics*, 11(3):204–220, MAR 2010.

[101] Jay T. Lennon and Stuart E. Jones. Microbial seed banks: the ecological and evolutionary implications of dormancy. *Nature Reviews Microbiology*, 9(2):119–130, FEB 2011.

[102] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, JUL 15 2009.

[103] Heng Li and Richard Durbin. Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357):493–U84, JUL 28 2011.

[104] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Proc. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, AUG 15 2009.

[105] N Li and M Stephens. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165(4):2213–2233, DEC 2003.

[106] Ryan Lister, Ronan C. O'Malley, Julian Tonti-Filippini, Brian D. Gregory, Charles C. Berry, A. Harvey Millar, and Joseph R. Ecker. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, 133(3):523–536, MAY 2 2008.

[107] Sverre Lundemo, Mohsen Falahati-Anbaran, and Hans K. Stenoien. Seed banks cause elevated generation times and effective population sizes of Arabidopsis thaliana in northern Europe (vol 18, pg 2798, 2009). *Molecular Ecology*, 19(8):1754, APR 2010.

[108] Michael Lynch, Ryan Gutenkunst, Matthew Ackerman, Ken Spitze, Zhiqiang Ye, Takahiro Maruki, and Zhiyuan Jia. Population Genomics of Daphnia pulex. *Molecular Biology and Evolution*, 206(1):315–332, MAY 2017.

[109] Michael Lynch, Bernhard Haubold, Peter Pfaffelhuber, and Takahiro Maruki. Inference of Historical Population-Size Changes with Allele-Frequency Data. *G3-Genes Genomes Genetics*, 10(1):211–223, JAN 2020.

[110] Julie Marin, Guillaume Achaz, Anton Crombach, and Amaury Lambert. The genomic view of diversification. *Journal of Evolutionary Biology*.

[111] P Marjoram and JD Wall. Fast "coalescent" simulation. *BMC Genetics*, 7, MAR 15 2006.

[112] Niklas Mather, Samuel M. Traves, and Simon Y. W. Ho. A practical introduction to sequentially Markovian coalescent methods for estimating demographic history from genomic data. *Ecology and Evolution*, 10(1):579–589, JAN 2020.

[113] Sebastian Matuszewski, Marcel E. Hildebrandt, Guillaume Achaz, and Jeffrey D. Jensen. Coalescent processes with skewed offspring distributions and non-equilibrium demography. *Genetics*, 2017.

[114] O. Mazet, W. Rodriguez, S. Grusea, S. Boitard, and L. Chikhi. On the importance of being structured: instantaneous coalescence rates and human evolution-lessons for ancestral population size inference? *Heredity*, 116(4):362–371, APR 2016.

[115] GAT McVean and NJ Cardin. Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B-Biological Sciences*, 360(1459):1387–1393, JUL 29 2005.

[116] Sajad Mirzaei and Yufeng Wu. RENT plus : an improved method for inferring local genealogical trees from haplotypes with recombination. *Bioinformatics*, 33(7):1021–1030, APR 1 2017.

[117] M Mohle. A convergence theorem for Markov chains arising in population genetics and the coalescent with selfing. *Advances in Applied Probability*, 30(2):493–512, JUN 1998.

[118] M Mohle and S Sagitov. A classification of coalescent processes for haploid exchangeable population models. *Annals of Probability*, 29(4):1547–1562, OCT 2001.

[119] Ana Y. Morales-Arce, Rebecca B. Harris, Anne C. Stone, and Jeffrey 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.

[120] Aline Muyle, Jeffrey Ross-Ibarra, Danelle K. Seymour, and Brandon S. Gaut. Investigation Gene body methylation is under selection in *Arabidopsis thaliana*. September 2020.

[121] Krystyna Nadachowska-Brzyska, Reto Burri, Linnea Smeds, and Hans Ellegren. PSMC analysis of effective population sizes in molecular ecology and its application to black-and-white Ficedula flycatchers. *Molecular Ecology*, 25(5):1058–1072, MAR 2016.

[122] Shigeki Nakagome, Richard R. Hudson, and Anna Di Rienzo. Inferring the model and onset of natural selection under varying population size from the site frequency spectrum and haplotype structure. *Proceedings Of the Royal Society B-Biological Sciences*, 286(1896), FEB 6 2019.

[123] Michael G. Nelson, Raquel S. Linheiro, and Casey M. Bergman. McClintock: An Integrated Pipeline for Detecting Transposable Element Insertions in Whole-Genome Shotgun Sequencing Data. *G3-Genes Genomes Genetics*, 7(8):2763–2778, AUG 2017.

[124] Svend V. Nielsen, Simon Simonsen, and Asger Hobolth. Inferring Population Genetic Parameters: Particle Filtering, HMM, Ripley's K-Function or Runs of Homozygosity? In Frith, M and Pedersen, CNS, editor, *Algorithms in Bioinformatics*, volume 9838 of *Lecture Notes in Bioinformatics*, pages 234–245, 2016.

[125] Hiro-Sato Niwa, Kazuya Nashida, and Takashi Yanagimoto. Reproductive skew in japanese sardine inferred from dna sequences. *ICES Journal of Marine Science*, 73(9):2181–2189, 2016.

[126] M Nordborg. Linkage disequilibrium, gene trees and selfing: An ancestral recombination graph with partial self-fertilization. *Molecular Biology and Evolution*, 154(2):923–929, FEB 2000.

[127] L Nunney. The effective size of annual plant populations: The interaction of a seed bank with fluctuating population size in maintaining genetic variation. *American Naturalist*, 160(2):195–204, AUG 2002.

[128] Kevin P. Oh, Cameron L. Aldridge, Jennifer S. Forbey, Carolyn Y. Dadabay, and Sara J. Oyler-McCance. Conservation Genomics in the Sagebrush Sea: Population Divergence, Demographic History, and Local Adaptation in Sage-Grouse (Centrocercus spp.). *Genome Biology and Evolution*, 11(7):2023–2034, JUL 2019.

[129] Stephan Ossowski, Korbinian Schneeberger, Jose Ignacio Lucas-Lledo, Norman Warthmann, Richard M. Clark, Ruth G. Shaw, Detlef Weigel, and Michael Lynch. The Rate and Molecular Spectrum of Spontaneous Mutations in Arabidopsis thaliana. *Science*, 327(5961):92–94, JAN 1 2010.

[130] Julia A. Palacios, John Wakeley, and Sohini Ramachandran. Bayesian Nonparametric Inference of Population Size Changes from Sequential Genealogies. *Genetics*, 201(1):281+, SEP 2015.

[131] Pier Francesco Palamara, Jonathan Terhorst, Yun S. Song, and Alkes L. Price. High-throughput inference of pairwise coalescence times identifies signals of selection and enriched disease heritability. *Nature Genetics*, 50(9):1311+, SEP 2018.

[132] Kris V. Parag and Oliver G. Pybus. Robust Design for Coalescent Model Inference. *Systematic Biology*, 68(5):730–743, SEP 2019.

[133] Pavlos Pavlidis, Jeffrey D. Jensen, Wolfgang Stephan, and Alexandros Stamatakis. A Critical Assessment of Storytelling: Gene Ontology Categories and the Importance of Validating Genomic Scans. *Molecular Biology and Evolution*, 29(10):3237–3248, OCT 2012.

[134] James B. Pease, David C. Haak, Matthew W. Hahn, and Leonie C. Moyle. Phylogenomics Reveals Three Sources of Adaptive Variation during a Rapid Radiation. *PLOS Biology*, 14(2), FEB 2016.

[135] S. P. Pfeifer. From next-generation resequencing reads to a high-quality variant data set. *Heredity*, 118(2):111–124, FEB 2017.

[136] J Pitman. Coalescents with multiple collisions. *Annals of Probability*, 27(4):1870–1902, OCT 1999.

[137] Roy N. Platt, II, Laura Blanco-Berdugo, and David A. Ray. Accurate Transposable Element Annotation Is Vital When Analyzing New Genome Assemblies. *Genome Biology and Evolution*, 8(2):403–410, FEB 2016.

[138] Aaron P. Ragsdale and Ryan N. Gutenkunst. Inferring Demographic History Using Two-Locus Statistics. *Genetics*, 206(2):1037–1048, JUN 2017.

[139] Daniel P Rice, John Novembre, and Michael M Desai. Distinguishing multiple-merger from kingman coalescence using two-site frequency spectra. *bioRxiv*, 2018.

[140] Willy Rodriguez, Olivier Mazet, Simona Grusea, Armando Arredondo, Josue M. Corujo, Simon Boitard, and Lounes Chikhi. The IICR and the non-stationary structured coalescent: towards demographic inference with arbitrary changes in population structure. *Heredity*, 121(6):663–678, DEC 2018.

[141] Francois Roudier, Felipe Karam Teixeira, and Vincent Colot. Chromatin indexing in Arabidopsis: an epigenomic tale of tails and more. *TRENDS IN GENETICS*, 25(11):511–517, NOV 2009.

[142] S Sagitov. The general coalescent with asynchronous mergers of ancestral lines. *Journal of Applied Probability*, 36(4):1116–1125, DEC 1999.

[143] S Sagitov. Convergence to the coalescent with simultaneous multiple mergers. *Journal of Applied Probability*, 40(4):839–854, DEC 2003.

[144] P. A. Salome, K. Bomblies, J. Fitz, R. A. E. Laitinen, N. Warthmann, L. Yant, and D. Weigel. The recombination landscape in Arabidopsis thaliana F-2 populations. *Heredity*, 108(4):447–455, APR 2012.

[145] Andreas Sand, Martin Kristiansen, Christian N. S. Pedersen, and Thomas Mailund. zipHMMlib: a highly optimised HMM library exploiting repetitions in the input to speed up the forward algorithm. *BMC Bioinformatics*, 14, NOV 22 2013.

[146] Stephan Schiffels and Richard Durbin. Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, 46(8):919–925, AUG 2014.

[147] Joshua G. Schraiber and Joshua M. Akey. Methods and models for unravelling human evolutionary history. *Nature Reviews Genetics*, 16(12):727–740, DEC 2015.

[148] Daniel R. Schrider, Alexander G. Shanku, and Andrew D. Kern. Effects of Linked Selective Sweeps on Demographic Inference and Model Selection. *Genetics*, 204(3):1207+, NOV 2016.

[149] J Schweinsberg. Coalescent processes obtained from supercritical Galton-Watson processes. *Stochastic Processes and their Applications*, 106(1):107–139, JUL 2003.

[150] Thibaut Sellinger, Diala Abu Awad, and Aurélien Tellier. Limits and Convergence properties of the Sequentially Markovian Coalescent. July 2020.

[151] Thibaut Paul Patrick Sellinger, Diala Abu Awad, Markus Moest, and Aurelien Tellier. Inference of past demography, dormancy and self-fertilization rates from whole genome sequence data. *PLOS Genetics*, 16(4), APR 2020.

[152] Sara Sheehan, Kelley Harris, and Yun S. Song. Estimating Variable Effective Population Sizes from Multiple Genomes: A Sequentially Markov Conditional Sampling Distribution Approach. *Molecular Biology and Evolution*, 194(3):647+, JUL 2013.

[153] Sara Sheehan and Yun S. Song. Deep Learning for Population Genetic Inference. *PLOS Computational Biology*, 12(3), MAR 2016.

[154] Montgomery Slatkin. Statistical methods for analyzing ancient DNA from hominins. *Current Opinion in Genetics & Development*, 41:72–76, DEC 2016.

[155] Chris C. R. Smith and Samuel M. Flaxman. Leveraging whole genome sequencing data for demographic inference with approximate Bayesian computation. *Molecular Ecology Resources*, 20(1):125–139, JAN 2020.

[156] Leo Speidel, Marie Forest, Sinan Shi, and Simon R. Myers. A method for genome-wide genealogy estimation for thousands of samples. *Nature Genetics*, 51(9):1321+, SEP 2019.

[157] Jeffrey P. Spence, Matthias Steinrucken, Jonathan Terhorst, and Yun S. Song. Inference of population history using coalescent HMMs: review and outlook. *Current Opinion in Genetics & Development*, 53:70–76, DEC 2018.

[158] Jeffrey P. Spence, Matthias Steinrucken, Jonathan Terhorst, and Yun S. Song. Inference of population history using coalescent HMMs: review and outlook. *Current Opinion in Genetics & Development*, 53:70–76, DEC 2018.

[159] Paul R. Staab, Sha Zhu, Dirk Metzler, and Gerton Lunter. scrm: efficiently simulating long sequences using the approximated coalescent with recombination. *Bioinformatics*, 31(10):1680–1682, MAY 15 2015.

[160] Remco Stam, Tetyana Nosenko, Anja C. Hoerger, Wolfgang Stephan, Michael Seidel, Jose M. M. Kuhn, Georg Haberer, and Aurelien Tellier. The de Novo Reference Genome and Transcriptome Assemblies of the Wild Tomato Species Solanum chilense Highlights Birth and Death of NLR Genes Between Tomato Species. *G3-Genes Genomes Genetics*, 9(12):3933–3941, DEC 2019.

[161] Matthias Steinrucken, Jack Kamm, Jeffrey P. Spence, and Yun S. Song. Inference of complex population histories using whole-genome sequences from

multiple populations. *Proceedings of the National Academy of Sciences of the United States of America*, 116(34):17115–17120, AUG 20 2019.

[162] Matthias Steinruecken, Matthias Birkner, and Jochen Blath. Analysis of DNA sequence variation within marine species using Beta-coalescents. *Theoretical Population Biology*, 87:15–24, AUG 2013.

[163] Wolfgang 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.

[164] Shohei Takuno, Jin-Hua Ran, and Brandon S. Gaut. Evolutionary patterns of genic DNA methylation vary across land plants. *Nature Plants*, 2(2), FEB 2016.

[165] Aurelien Tellier. Persistent seed banking as eco-evolutionary determinant of plant nucleotide diversity: novel population genetics insights. *New Phytologist*, 221(2):725–730, JAN 2019.

[166] Aurelien Tellier, Stefan J. Y. Laurent, Hilde Lainer, Pavlos Pavlidis, and Wolfgang Stephan. Inference of seed bank parameters in two wild tomato species using ecological and genetic data. *Proceedings of the National Academy of Sciences of the United States of America*, 108(41):17052–17057, OCT 11 2011.

[167] AR Templeton and DA Levin. Evolutionary Consequences of Seed Pools. *American Naturalist*, 114(2):232–249, 1979.

[168] Jonathan Terhorst, John A. Kamm, and Yun S. Song. Robust and scalable inference of population history froth hundreds of unphased whole genomes. *Nature Genetics*, 49(2):303–309, FEB 2017.

[169] Jonathan Terhorst and Yun S. Song. Fundamental limits on the accuracy of demographic inference based on the sample frequency spectrum. *Proceedings of the National Academy of Sciences of the United States of America*, 112(25):7677–7682, JUN 23 2015.

[170] The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, November 2012.

[171] The 1001 Genomes Consortium. 1,135 genomes reveal the global pattern of polymorphism in arabidopsis thaliana. *Cell*, July 2016.

[172] van der Graaf et al. Rate, spectrum, and evolutionary dynamics of spontaneous epimutations. *Proceedings of the National Academy of Sciences of the United States of America*, 112(21):6676–6681, MAY 26 2015.

[173] Amaryllis Vidalis, Daniel Zivkovic, Rene Wardenaar, David Roquis, Aurelien Tellier, and Frank Johannes. Methylome evolution in plants. *Genome Biology*, 17, DEC 20 2016.

[174] R Vitalis, S Glemin, and I Olivieri. When genes go to sleep: The population genetic consequences of seed dormancy and monocarpic perenniality. *American Naturalist*, 163(2):295–311, FEB 2004.

[175] Berit Lindum Waltoft and Asger Hobolth. Non-parametric estimation of population size changes from the site frequency spectrum. *Statistical Applications in Genetics and Molecular Biology*, 17(3), JUN 2018.

[176] Ke Wang, Iain Mathieson, Jared O'Connell, and Stephan Schiffels. Tracking human population structure through time from whole genome sequences. *PLOS Genetics*, 16(3), MAR 2020.

[177] CA 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.

[178] Rachel C. Williams, Marina B. Blanco, Jelmer W. Poelstra, Kelsie E. Hunnicutt, Aaron A. Comeault, and Anne D. Yoder. Conservation genomic analysis reveals ancient introgression and declining levels of genetic diversity in Madagascar's hibernating dwarf lemurs. *Heredity*, 124(1):236–251, JAN 2020.

[179] Peter R. Wilton, Shai Carmi, and Asger Hobolth. The SMC' Is a Highly Accurate Approximation to the Ancestral Recombination Graph. *Molecular Biology and Evolution*, 200(1):343–U637, MAY 2015.

[180] C Wiuf and J Hein. Recombination as a point process along sequences. *Theoretical Population Biology*, 55(3):248–259, JUN 1999.

[181] C Wiuf and J Hein. The ancestry of a sample of sequences subject to recombination. *Molecular Biology and Evolution*, 151(3):1217–1228, MAR 1999.

[182] Katarzyna Wreczycka, Alexander Gosdschan, Dilmurat Yusuf, Bjoern Gruening, Yassen Assenov, and Altuna Akalin. Strategies for analyzing bisulfite sequencing data. *Journal of Biotechnology*, 261(SI):105–115, NOV 10 2017.

[183] Sen Xu, Matthew S. Ackerman, Hongan Long, Lydia Bright, Ken Spitze, Jordan S. Ramsdell, W. Kelley Thomas, and Michael Lynch. A Male-Specific Genetic Map of the Microcrustacean Daphnia pulex Based on Single-Sperm Whole-Genome Sequencing. *Molecular Biology and Evolution*, 201(1):31+, SEP 2015.

[184] ZH Yang. Statistical properties of a DNA sample under the finite-sites model. *Genetics*, 144(4):1941–1950, DEC 1996.

[185] Assaf Zemach, Ivy E. McDaniel, Pedro Silva, and Daniel Zilberman. Genome-Wide Evolutionary Analysis of Eukaryotic DNA Methylation. *Science*, 328(5980):916–919, MAY 14 2010.

[186] Kai Zeng, Brian Charlesworth, and Asger Hobolth. Studying models of balancing selection using phase-type theory. preprint, Evolutionary Biology, July 2020.

[187] Daniel Zilberman, Mary Gehring, Robert K. Tran, Tracy Ballinger, and Steven Henikoff. Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. *Nature Genetics*, 39(1):61–69, JAN 2007.

[188] Daniel Zivkovic and Aurelien Tellier. Germ banks affect the inference of past demographic events. *Molecular Ecology*, 21(22):5434–5446, NOV 2012.