Limits and Convergence properties of the Sequentially Markovian Coalescent

Thibaut Sellinger[1]*, Diala Abu Awad, Aurélien Tellier[1]

[1] Professorship for Population Genetics,

Department of Life Science Systems,

Technical University of Munich

* Corresponding author, thibaut.sellinger@tum.de

Abstract

Many methods based on the Sequentially Markovian Coalescent (SMC) have been and are being developed. These methods make use of genome sequence data to uncover population demographic history. More recently, new methods even allow the simultaneous estimation of the demographic history and other biological variables, extending the original theoretical frameworks. Those methods can be applied to many different species, under different model assumptions, in hopes of unlocking the population/species evolutionary history. Although convergence proofs in particular cases have been given using simulated data, a clear outline of the performance limits of these methods is lacking. We here explore the limits of this methodology, as well as present a tool that can be used to help users quantify what information can be confidently retrieved from given datasets. In addition, we study the consequences for inference accuracy of the violation of hypotheses and assumptions of SMC approaches, such as the presence of transposable elements, variable recombination and mutation rates along the sequence and SNP call errors. We also provide a new interpretation of the SMC through the use of the estimated transition matrix and offer recommendations for the most efficient use of these methods under budget constraints, notably through the building of data sets that would be better adapted for the biological question at hand.

Keywords— Hidden Markov Model, Ancestral Recombination Graph, Population Genetics

1 Introduction

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) [14, 42, 4]. Current statistical tools to estimate the demographic history rely on genomic data [48] and these inferences are often linked to archaeological or climatic data, providing new insights on their consequent genomic signatures [67, 32, 43, 1, 12, 25, 24]. From these analyses, evidence for migration events have been uncovered [25, 5], as have genomic consequences of human activities on other species [9]. Linking demographic history to climate and environmental data greatly supports the field of conservation genetics [10, 17, 39]. Such analyses can help ecologist in detecting effective population size decrease [65], 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 [26]. In addition, studying the demographic histories of different species in relation to one another can unveil latent biological or environmental evolutionary forces [16], 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 an information that is essential in the field of evolution [45, 6]. However, unbiased estimations and interpretations of the demographic history remain challenging [3, 8].

The most sophisticated methods to infer demographic history make use of whole genome polymorphism data. Among the state of the art methods, some are based on the theory of the Sequentially Markovian Coalescent (SMC) developed by [34] after the work of [66], corrected by [30] and first applied to whole genome sequences by [25], who introduced the now well known Pairewise Sequentially Markovian Coalescent (PSMC) method. PSMC allows demographic inference of the whole population 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 and infer recombination and demographic history of a given population, assuming neutrality and a panmictic population. Although PSMC was a breakthrough in demographic inference, it has limited power in inferring more recent events. In order 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) [47]. 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 [34] but on SMC' theory [30], therefore MSMC applied to only 2 sequences has been defined as PSMC'. Methods developed after MSMC followed suit, with MSMC2 [29] 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++ [60] 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 time [60]. 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 [44]. In a similar vein, the Ascertained Sequentially Markovian Coalescent (ASMC) [41] 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 developed by [18].

More recently, a second generation of SMC based methods have been developed. New features have been added to the initial SMC theory, extending their application beyond simply inferring past demography [1, 50, 63]. The development of C-PSMC [16] allows the interpretation of estimated demographic history in the light of coevolution, making the first link between demographic history estimated by PSMC and evolutionary forces (although biological interpretation remains limited). iSMC [1] extends the PSMC theory to account and infer the variation of recombination rate along sequences, unlocking recombination map estimations. An impressive advancement is the development of IS-MSMC, which solves to some extent the population structure problem, allowing accurate and simultaneous inference of the demographic history and population admixture [63]. eSMC [50] 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 which may not be explained under the initial SMC hypotheses can now be explained by the potential presence of measurable phenomena not present in the original PSMC.

New methods have been developed since PSMC, that have been either strongly inspired by the SMC [51, 59] or that are completely dissociated from it [55, 2, 46, 20, 28, 19, 54, 62]. Though there are alternative approaches, methods based on the SMC are still considered state of the art, and remain widely used [31, 3, 56], notably in human evolution studies [56, 44]. However, each described method has its specificity, designed to solve a specific problem using specific data based on different hypothesis.

Although all these methods allow a new and different interpretation of genomic data, none of these methods guarantees unbiased inference, and their limitations have rather underlined how crucial and challenging demographic inference is, highlighting the complementarity and usefulness to use several inference methods on a given dataset.

SMC based methods display very good fits when using simulates data, especially when using simple single population model based on typical human data parameters [60, 47, 50, 63]. However, the SMC makes a large number of hypotheses [25, 47] 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 [8, 61, 3] and several studies address the consequences of hypothesis violation [15, 8, 46, 33, 49]. They bring to light how strongly population structure or introgression influence demographic history estimation if not correctly accounted for [15, 8]. Furthermore, most SMC based methods require phased data (such as MSMC and IS-MSMC), and phasing errors can lead to strong overestimation of population size in recent time [60]. The effect of coverage during sequencing has also been tested in [36], showing the importance of high coverage in order to obtain trustworthy results, and yet, SMC methods seem robust to genome quality [44]. Selection, if not accounted for, can result in a bottleneck signature [49], 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 [52, 37]. More problematic, is the ratio of effective recombination over effective mutation rates $\frac{\rho}{\theta}$. If the ratio is greater than one, biases in estimations are to be expected [60, 1, 50]. 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 rates measured experimentally $(\frac{r}{\mu})$, as the former can be greatly influenced by life-history and this can lead to issues when interpreting results (*e.g.* [50]). It is thus necessary to keep in mind that the accuracy of SMC based methods depends on which of the many underlying hypothesis are prone to being violated by the data sets being used.

In an attempt to complement previous works, we here offer to study the limits and the convergences properties of methods based on the Sequentially Markovian Coalescence. We first define the limits of SMC based methods (*i.e.* how well they perform theoretically), which we will call the theoretical convergence, using a similar approach to [13, 40, 19] by giving the simulated genealogy as input. 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 theoretical limits (and not issues linked with data). We then compare simulation results obtained with the genealogy given as input to results obtained from sequences simulated under the same genealogy, so as to study the convergence properties linked to data sets in the absence of hypothesis violation. 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, our aim is 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.

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. In order to have a finite number of hidden state (and parameters), the hidden states are grouped into x driscretized bins (x being the number of hidden states). The reasons for our model choices are as follows. MSMC, unlike any other method, focuses on the first coalescent event of a sample of size n, and thus exhibits different convergence properties [47]. MSMC2 computes coalescent times of all pairwise analysis from a sample of size n, and can deal with a large range of data sets [55]. SMC++ [60] is the most advanced and efficient SMC method which can make use of hundreds sequences, enabling the use of the SFS along the sequence. Lastly, eSMC [50] 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.

2.1 SMC methods

2.1.1 PSMC', MSMC2 and eSMC

PSMC' and methods that stem from it (such as MSMC2 [29] and eSMC [50]) focus on the coalescence events between only two individuals (or sequences in practice), and, as a result, does not require phased data. The algorithm goes along the sequence and estimates the coalescence time at each position. In order to do this, it checks whether the two sequences are similar or different at each position. If the two sequences are different, this indicates a mutation took place, and, as mutations are considered uncommon, that the common ancestor is far in the past. An 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 [47, 63, 50].

2.1.2 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 sequences. However, MSMC cannot analyze more than 10 sequences simultaneously (due to computational load). A detailed description of MSMC can be found in [47].

2.1.3 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 like a cubic spline to estimate demographic history (*i.e.* not a piece-wise constant population size). The SMC++ algorithm is fully described in [60].

2.1.4 Theoretical convergence

Using sequence simulators such as msprime [21] or scrm [57], 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 defined as the number of hidden states) counting the number of transitions from one of the x state 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 using PSMC' or MSMC as if they could perfectly infer the hidden states. Hence estimations using the exact transition matrix represents the upper bound of performance for those methods. We choose to call this upper bound the theoretical convergence (since it can never be reached in practice). For this study's purpose, a second version of the R package eSMC [50] was developed. This package enables the building of the transition matrix (for PSMC' or MSMC), and can then use it 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.

2.1.5 Baum-Welch algorithm

SMC based method 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 PSMC' 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 PSMC' and MSMC by equation 2:

$$Q(\theta|\theta^t) = \nu_{\theta^t} log(P(X_1|\theta)) + \sum_{X,Y} E(X,Z|\theta^t) log(P(X|Z,\theta)) + \sum_{X,Y} E(Y,X|\theta^t) log(P(Y|X,\theta)) \quad (1)$$

and :

$$Q(\theta|\theta^t) = \sum_{X,Y} E(X,Z|\theta^t) log(P(X|Z,\theta)), \quad (2)$$

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

2.1.6 Time window

Each tested SMC based method has its own specific time window in which estimations are made. As for example, the original PSMC has a time window wider than PSMC'. To measure the effect of the time window we analyze the same data with 4 different time windows. The first time window is the one of PSMC' defined in [47]. The second time window is the one of MSMC2 [63] (similar to the one of the original PSMC [25]), which we call "big" since it goes further in the past and in more recent time than the one of PSMC'. We then define a time window equivalent to the first one (*i.e.* PSMC') shifted by a factor 5 in the past (*first time window multiplied by 5*). The last window is a time window equivalent to the first one shifted by a factor 5 in recent time (*first time window divided by 5*).

2.2 Simulated Sequence data

Throughout this paper we simulate different demographic scenarios using either the coalescence simulation program scrm [57] or msprime [21]. We use scrm for the theoretical convergence as it can output the genealogies in a Newick formart (which we use as input). We use msprime to simulate data for SMC++ since msprime is more efficient than scrm for big sample sizes [21] and can directly output .vcf files (which is the input format of SMC++).

2.2.1 Absence of hypothesis violation

We simulate four demographic scenarios: saw-tooth (successions of population size expansion and decrease), bottleneck, expansion and decrease. Each scenario is tested under four amplitude parameters (*i.e.* by how many fold the population size varies: 2, 5, 10, 50). For each analysis we simulate 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 human on the guide to MSMC, respectively 1.25×10^{-8} and 1×10^{-8} (https://github.com/stschiff/msmc/blob/master/guide.md), all the command lines to simulate data can be found in S1 of the Appendix. For each simulated data set, as previously mentioned, four different algorithms are used to estimate the demographic history and the recombination rate: eSMC, MSMC and MSMC2 and SMC++ (the command lines to launch the analyses can be found in S2 of the Appendix).

2.2.2 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 and many more. Therefore, based on raw sequence data, stringent filters can exclude SNPs (false negatives) or include surious SNPs (false positives). When dealing with complex genomes or ancient DNA [53, 7], SNPs can be simultaneously missed and added. We thus simulate sequences under a "saw-tooth" 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 analyse the variation and bias in SNP call on the accuracy of demographic parameter estimations.

Changes in mutation and recombination rates along the sequence: Because the recombination rate and the mutation rate can change along the sequence[1], 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 changes 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 chose to simulate 20 scaffolds of size 2 Mb, as this can represents the best available assembly for non-model organisms [27, 58]. We then analyze the simulated sequences to study the effect of assuming scaffolds sharing same mutation and recombination rates. In addition, we simulate sequences of 40 Mb (assuming genome 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 which dynamics violate the classic infinite site mutational model for SNPs, and thus potentially affecting the estimation of different parameters. Although methods have been developed to detect [38] and simulate them [23], understanding how their presence/absence affect the demographic estimations remains unclear. TEs are usually masked in the reference genome and thus not taken into account in the mapped individuals due to the redundancy of read mapping for TEs. To best capture and mimic the effect of TEs in data, we altered simulated sequence data in two different ways. Due to the repetitive nature of TEs, it can be difficult using short reads to correctly detect and assemble them, as well as to assess their presence/absence polymorphism across individuals of a population [11]. One way to simulate the effect of TEs is to assume they exhibit presence/absence polymorphism thus creating gaps in the sequence. For each individual, we therefore randomly remove small pieces from the original simulated sequence, thus shortening and fragmenting the whole sequence to be analyzed.

The second way, would be to assume that TEs are masked, a process which we simulate by randomly selecting small pieces of sequence from the original simulated sequence, and removing all the SNPs found in those regions (*i.e.* removing mutations from TEs which could be used for inference but actually are judged to be non-reliable). In the latter, the removed SNPs are structured in many small regions along the genome, and not randomly missing throughout it. We also test the consequences of simultaneously having both removed and masked TEs in the data set.

3 Results

We first study the theoretical accuracy and convergence properties of PSMC' and MSMC methodologies using the sequence exact genealogies. We then analyze the simulated sequences themselves and compare results between different SMC based methods. Lastly, we analyze simulated sequences for which hypotheses made in the SMC framework are violated, so as to study their impact on the accuracy of inference.

3.1 Theoretical convergence

Results of the theoretical convergence of PSMC' under the saw-tooth demographic history are displayed in Figure 1. Increasing the sequence length increases accuracy and reduces variability, leading to a perfect fit (see Figures 1a-c). 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). Similar results are obtained for the three other demographic scenarios (bottleneck, expansion and decrease, respectively displayed in Supplementary Figures 1, 2 and 3). The bottleneck scenario seems especially difficult to infer, requiring large amounts of data, and the stronger the bottleneck, the harder it is to detect it, even with sequence lengths equivalent to 10^{10}bp. In Supplementary Figure 4, we show that even when changing the number of hidden states (*i.e.* number of inferred parameters), some scenarios with very strong variation of population size are badly inferred when using PSMC' based methods.

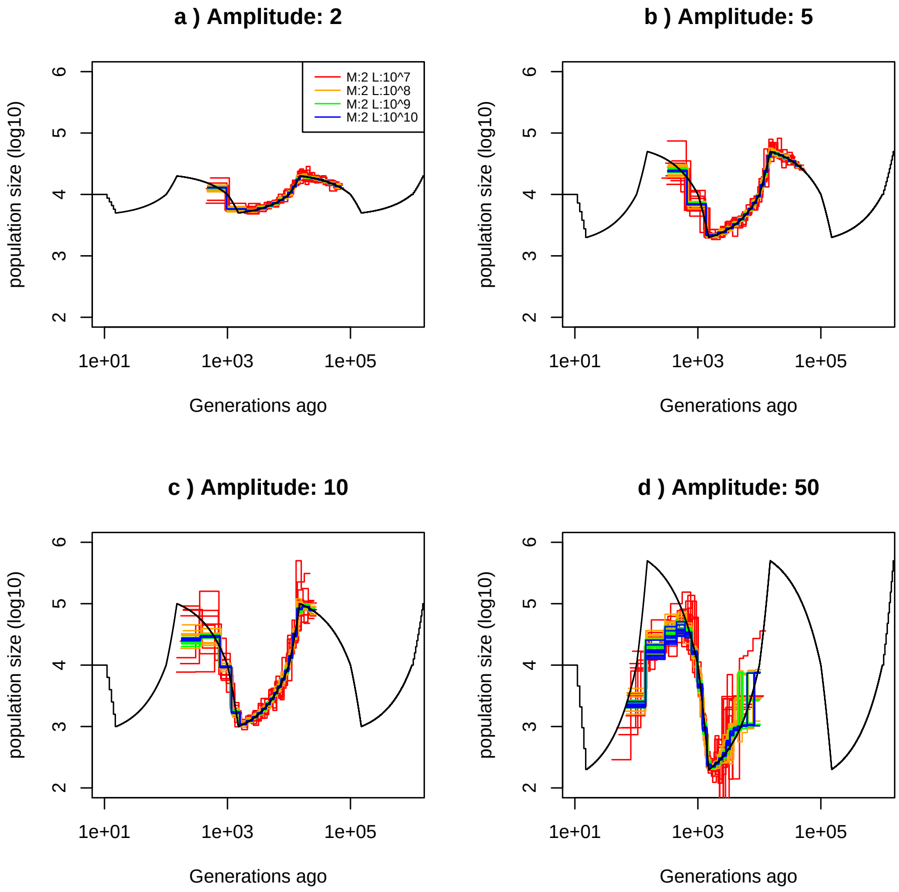

Fig. 1. Theoretical convergence of PSMC' Estimated demographic history using simulated genealogy over sequences of 10,100,1000,10000 Mb (respectively in red,orange, green and blue) under a saw-tooth scenario (black) with 10 replicates for different amplitudes of size change: a) 2-fold, b) 5-fold, c) 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.

In Supplementary Figures 5, 6, 7 and 8, we show the theoretical convergence of MSMC with four genome sequences and generally find that these analyses present a higher variance than PSMC'. However, MSMC shows better fits in recent times than PSMC' and is better able to retrieve population size variation than PSMC' (see Supplementary Figure 5d). Scenarios with strong variation of population size (*i.e.* with large amplitudes) still pose a problem (see Supplementary Figure 9), 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 obtained from the distribution 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 10), but that for scenarios with a large amplitude of population size variation, some hidden state transitions are not at all observed because of a lack of coalescence events occurring in those specific time windows. This results in matrices displaying higher coefficients of variation or no specific transition observed leading to a matrix that is partially empty (Figure 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 window.

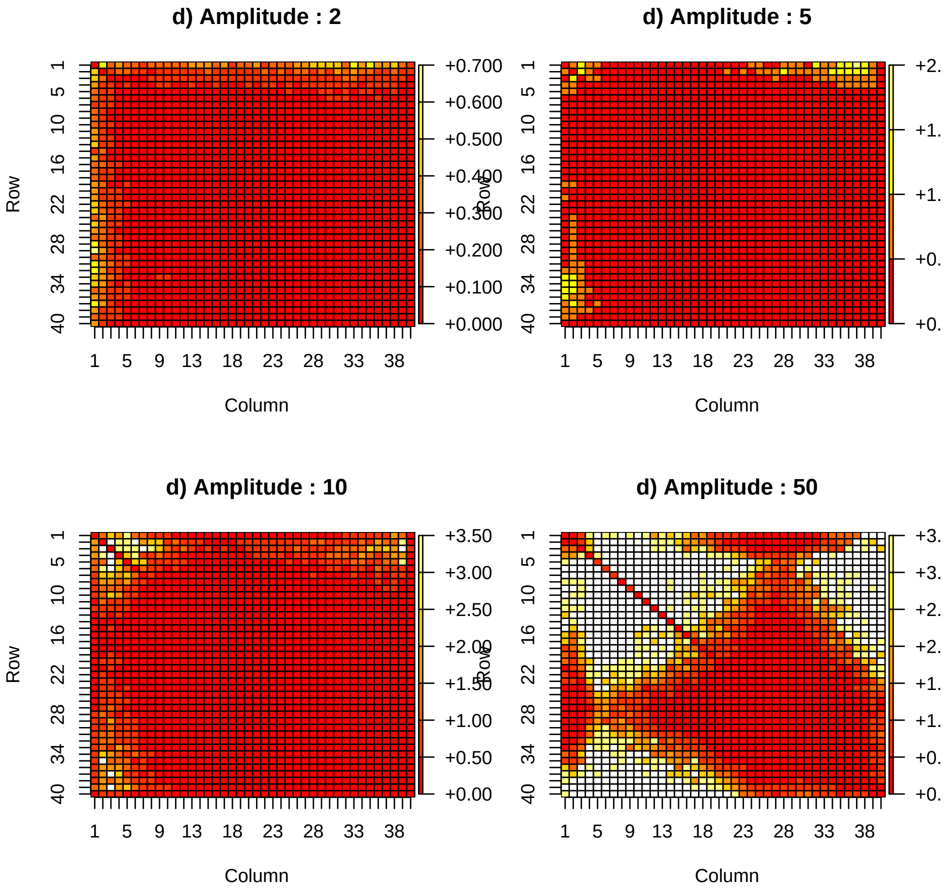

Fig. 2. Estimated transition matrix in sharp saw-tooth scenario Estimated coefficient of variation of the transition matrix using simulated genealogy over sequences of 10000 Mb under a saw-tooth scenario of amplitude 2, 5,10 and 50 (respectively in a, b, c and d) each with 10 replicates with recombination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

3.2 Simulated sequence results

3.2.1 Scenario effect

In the previous section, we explored the theoretical performance limitations of PSMC' and MSMC using trees in Newick format as input. In this section, we evaluate how these methods perform when inputting sequence data simulated under the same scenarios and parameters as above. Results for the saw-tooth scenario are displayed in Figure 3, where the different models display a good fit, but are not as good as expected from the theoretical convergence given the same amount of data (Figure 1 (orange line) vs Figure 3 (red line)). As predicted by Figures 1 and 2, the case with the greatest amplitude of population size variation (Figure 1d) is the least well fitted. All estimations display low variance and a relatively good fit in the bottleneck and expansion scenarios for small population size variation (see Supplementary Figures 11a and 12a). However, the strengths of expansions and bottlenecks are not fully retrieved in scenarios with population size variation equal to or higher than tenfold the current population size (Supplementary Figures 11c-d,and 12c-d). 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 find in scenarios which strong population size variation, explaining the high variance (see Supplementary Figure 13).

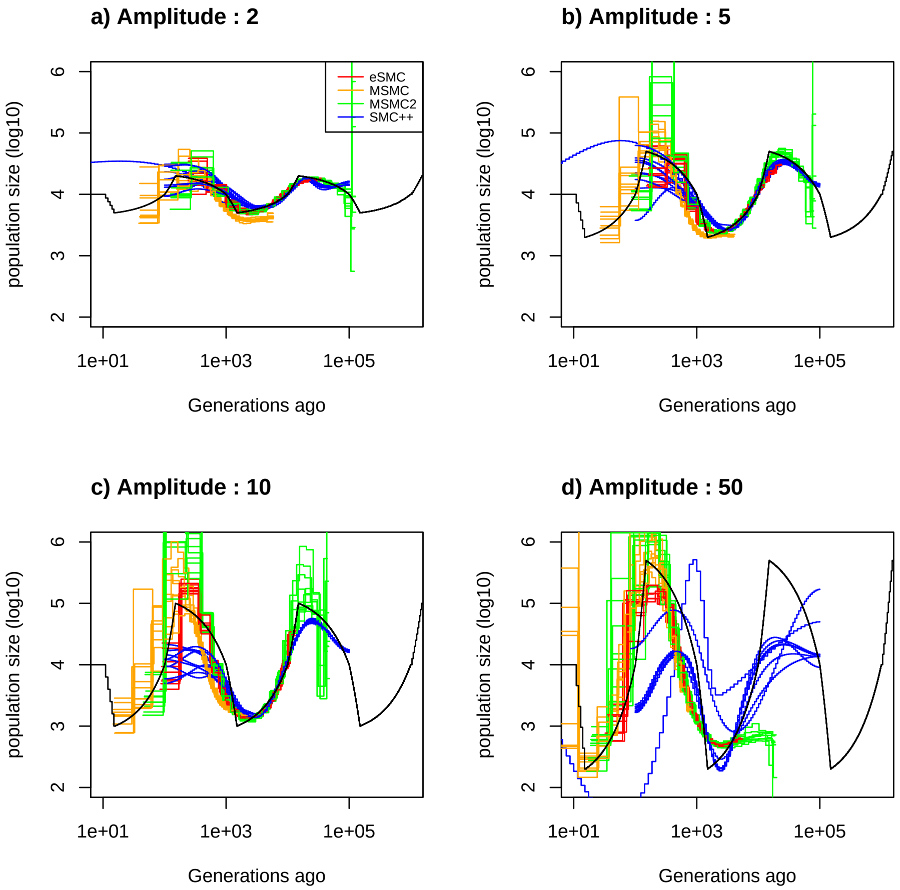

Fig. 3. Estimated demography using simulated sequences as input. Estimated demographic history (black) under a saw-tooth scenario with 10 replicates using simulated sequences for different amplitude of population size change: a) 2, b) 5, c) 10 and d) 50. Two sequences of 100 Mb for eSMC and MSMC2 (respectively in red and green). Four sequences of 100 Mb for MSMC (orange) and 20 sequences of 10 Mb for SMC++ (blue). Recombination and mutation rates are as in previous figures.

366 Increasing the time window results in an increased variance of the inferences

(Supplementary Figure 14). In addition, shifting the window towards more recent time leads to poor demographic estimation, but shifting it further in the past does not seem to bias the demographic estimation (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.

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	Saw-tooth	0.8	0.79 (0.036)	0.72 (0.039)	0.72 (0.042)	0.94 (0.005)
Complete Baum-Welch	Saw-tooth	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 1: Average estimated values for the recombination over mutation ratio $\frac{\rho}{\theta}$ over ten repetitions for different size of the time window. The coefficient of variation is indicated in brackets. four sequences of 50 Mb 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.

3.2.2 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 [60]. 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 [50, 60]. Under the bottleneck scenario, we find that the lower $\frac{\rho}{\theta}$, the better the fit of the inferred demography, but also the higher the variance of the inferences (see Figure 4). SMC++ seems especially sensitive to $\frac{\rho}{\theta}$. When calculating the difference between the transition matrix estimated by eSMC (*i.e.* PSMC') and the one built from the actual genealogy (using Newick trees), we find that, unsurprisingly, changes in hidden states are harder to detect when $\frac{\rho}{\theta}$ increases, leading to an overestimation of hidden states on the diagonal (see Supplementary Figures 15,16 and 17).

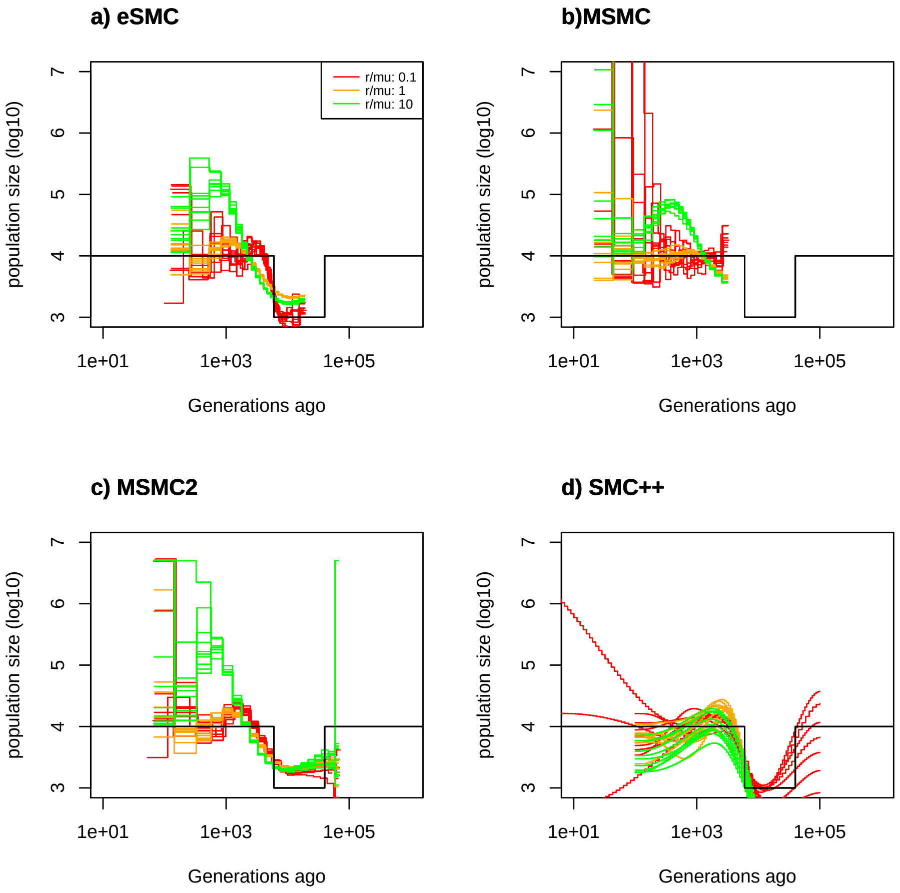

Fig. 4. Effect of $\frac{\rho}{\theta}$ on inference of demographic history. Estimated demographic history under a bottleneck scenario with 10 replicates using simulated sequences. Two sequences of 100 Mb for eSMC and MSMC2 (respectively in a and b). We use four sequences of 100 Mb for MSMC (c) and twenty 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 (see Supplementary Figure 18), although the correct recombination rate cannot be retrieved (Table 2). MSMC is able to retrieve the correct recombination rate despite a high $\frac{\rho}{\theta}$, but poorly estimates the demographic history. The results obtained using MSMC2 and SMC++ are not improved when increasing the number of iterations (see Supplementary Figure 18 and Table 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	0.69 (0.34)	0.60 (0.45)	0.54 (0.15)	0.12 (066)	0.77 (.40)

Table 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 Simulation results under hypothesis violation

3.3.1 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 can lead to a strong over-estimation 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 (see Figure 5 for MSMC2 and Supplementary Figures 19 and 20 for eSMC and MSMC respectively).

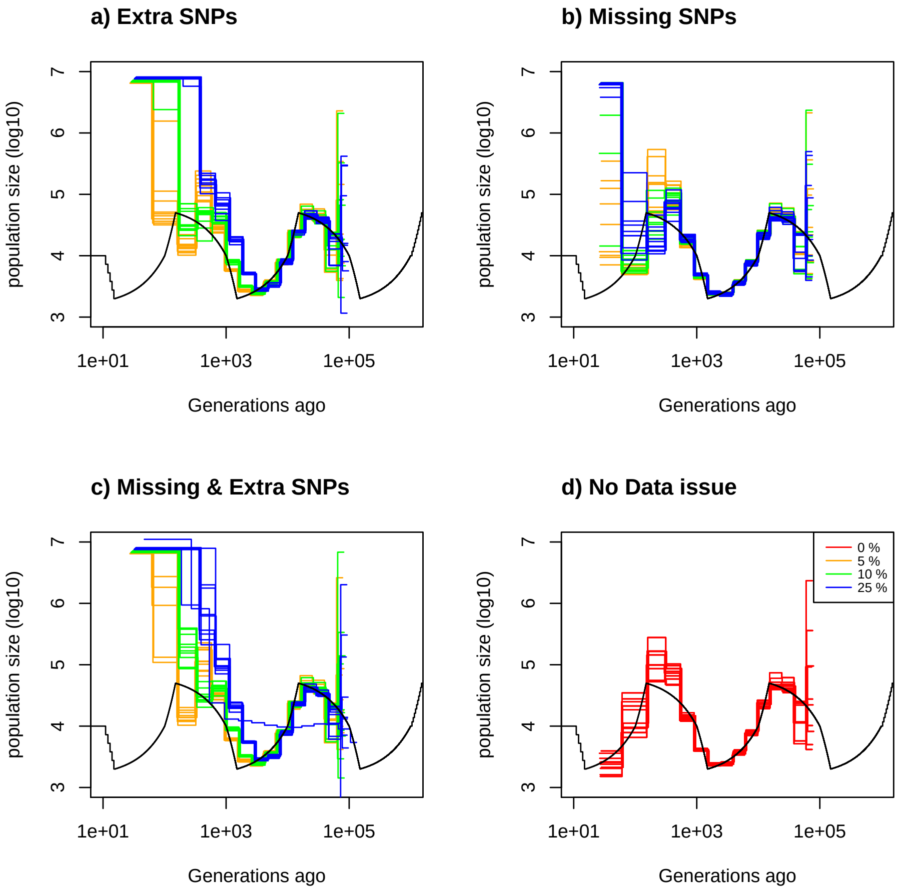

Fig. 5. Consequences of SNP calling errors. Estimated demographic history using MSMC2 under a saw-tooth scenario with 10 replicates using four simulated sequences of 100 Mb. Recombination and mutation rates are as in Figure 1 and the simulated demographic history is represented in black. a) Demographic history simulated with 5% (orange),10% (green) and 25% (blue) missing SNPs. b) Demographic history simulated with 5% (orange),10% (green) and 25% (blue) additional SNPs. c) Demographic history simulated with 5% (orange),10% (green) and 25% (blue) of additional and missing SNPs . d) Demographic history simulated with no SNP call error.

3.3.2 Specific scaffold parameters

We here analyze simulated sequence data where scaffolds either have or do not have the same recombination and mutation rates, and are analyzed assuming scaffolds do share or do not share recombination and mutation rates. We can see on Figure 6 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 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.

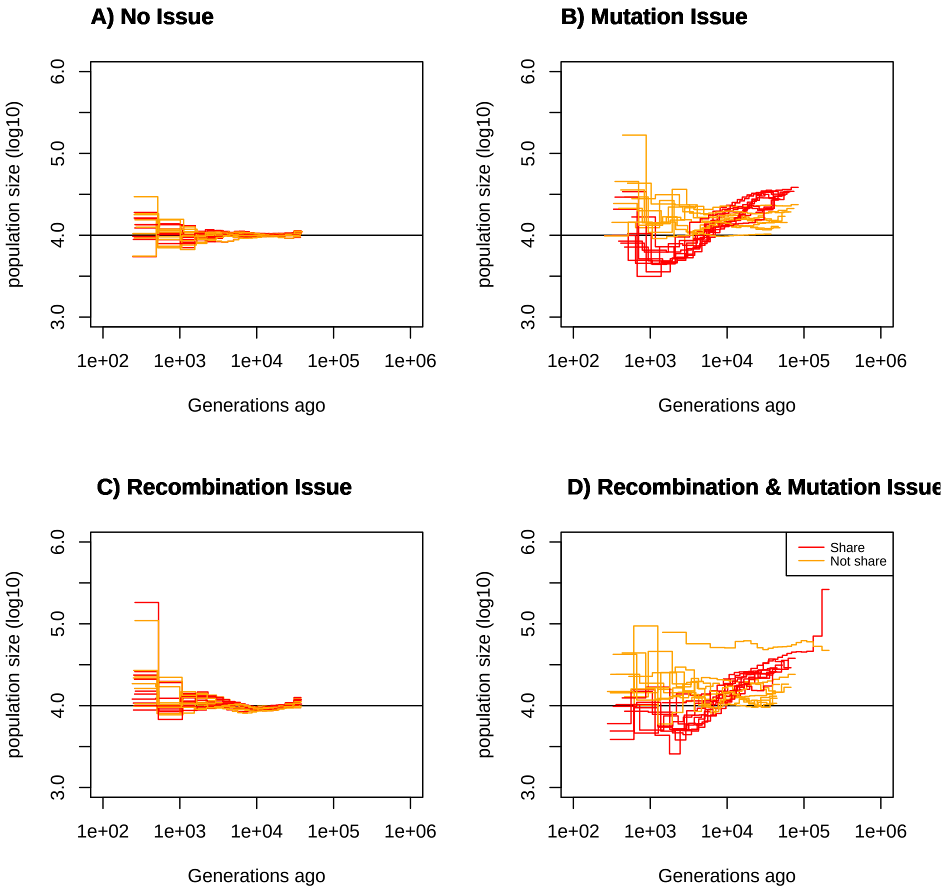

Fig. 6. Estimating demographic history using scaffolds sharing or differing in mutation and recombination rates Estimated demographic history using eSMC under a saw-tooth scenario with 10 replicates using twenty simulated scaffolds of two sequences of 2 Mb assuming scaffolds share (red) or do not share recombination and mutation rate (orange). The simulated demographic history is represented in black, for a) scaffolds share the same parameters, recombination and mutation rates are set at 1.25×10^{-8} , for 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}, for 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 for 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 (*i.e* PSMC'). As can be seen in Supplementary Figure 21, 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.

3.3.3 How transposable elements bias inference

Transposable elements (TEs) are present in most species, and are (if detected) only taken into account as missing data by SMC methods [47]). Depending on how TEs affect the data set, we find that methods are more or less sensitive to them. If TEs are removed from the data set, there does not appear to be any bias in the estimated demographic history when using eSMC (see Figure 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. The smaller the sequences that are removed, the more $\frac{\rho}{\theta}$ is over-estimated (Tables 4 and 5). If TEs are considered to be masked in the data set (and not accounted for missing data by the model), we find that this is equivalent to faulty calling of SNPs, in which SNPs are missing, hence resulting in demographic history estimation by eSMC similar to that observed in Figure 5a. However, if longer parts of the sequences are masked by TEs, different results are obtained (see Supplementary Figures 22 and 23). Indeed, there is a strong underestimation of population size and the model fails to capture the correct demographic history in recent times. The longer the masked parts are, the stronger the effect on the estimated demographic history. Similar results are obtained with MSMC (Supplementary Figures 24, 25 and 26) and

MSMC2 (Supplementary Figures 27, 28 and 29).

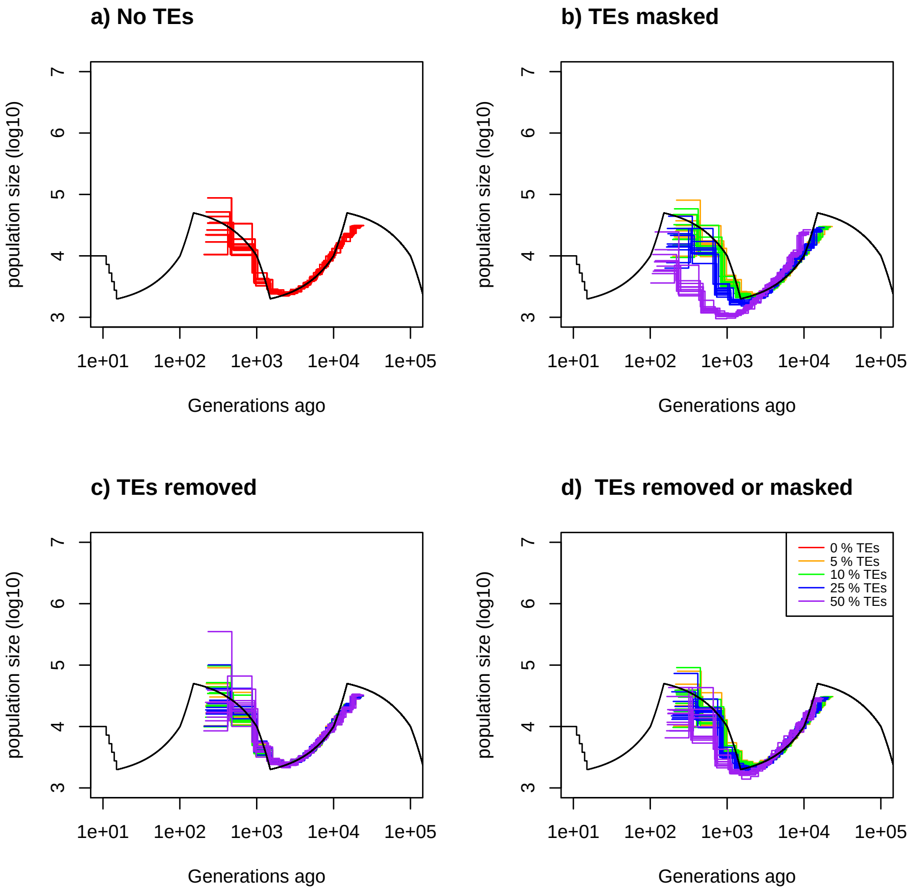

Fig. 7. Consequences of masking or removing transposable elements (TEs) from data sets. Estimated demographic history by eSMC under a saw-tooth scenario with 10 replicates using four simulated sequences of 20 Mb. The recombination and mutation rates are as in Figure 1 and the simulated demographic history is represented in black. Here the tansposable elements are of length 1kbp. a) Demographic history simulated with no transposable elements. b) Demographic history simulated where transposable elements are removed. c) Demographic history simulated where TEs are masked. d) Demographic history simulated where half of transposable are removed and SNPs on the other half are removed. Proportion of transposable element of the genome set to 0% (red), 5% (orange), 10% (green), 25 % (blue) and 50 % (purple).

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
eSMC	1	0.95 (0.021)	0.99 (0.022)	1.16 (0.10)	1.77 (0.36)
MSMC	1	1.31 (0.098)	1.35 (0.11)	1.50 (0.088)	1.91 (0.11)
MSMC2	1	0.87 (0.047)	0.88 (0.049)	1.0 (0.036)	1.35 (0.035)

Table 3: Average estimated values for the recombination over mutation ratio $\frac{\rho}{\theta}$ over ten repetitions. The coefficient of variation is indicated in brackets. TEs are removed and of length 1kb. The proportion of TEs is 5%,10% ,25% and 50%, the results are respectively displayed in column 3 to 6.

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
eSMC	1	0.96 (0.053)	0.98 (0.066)	1.10 (0.18)	1.36 (0.41)
MSMC	1	1.38 (0.074)	1.41 (0.0.090)	1.54 (0.11)	1.68 (0.13)
MSMC2	1	0.87 (0.064)	0.89 (0.067)	.99 (0.15)	1.13 (0.30)

Table 4: Average estimated values for the recombination over mutation ratio $\frac{\rho}{\theta}$ over ten repetitions. The coefficient of variation is indicated in brackets. TEs are removed and of length 10kb. The proportion of TEs is 5%,10% ,25% and 50%.

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
eSMC	1	0.95 (0.047)	0.95 (0.051)	0.98 (0.070)	1.0 (0.12)
MSMC	1	1.36 (0.048)	1.36 (0.062)	1.40 (0.093)	1.49 (0.12)
MSMC2	1	0.87 (0.056)	0.88 (0.050)	0.91 (0.079)	0.91 (0.073)

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

4 Discussion

Throughout this work 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 perfectly estimated, tending to results obtained by [13] or [8]. In [13] and [8] they use the actual series of coalescence time for estimation whereas we use the series of hidden states build from the discretization of time summarized in a simple matrix. However, we find that the amount of data required for a perfect fit depends on the underlying demographic scenario. In addition, some scenarios are better retrieved either with MSMC or PSMC', indicating complementary convergence properties of MSMC and PSMC' 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 a sequence of genealogies (using trees in Newick format or sequences of coalescence events), independently of how the genealogy 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 demographic history. As new methods can infer genealogy better and faster [55, 22, 35, 41], 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 [64, 13, 19]). In addition, using the work of [63], one could potentially extend our approach to account for population structure.

We have also demonstrated that the power of PSMC', MSMC, and other SMC based methods, rely on their ability to correctly infer the genealogy along the sequence (*i.e.* the ancestral recombination graph). The accuracy of the 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 hidden states on the diagonal, which explains the underestimation of the recombination rate and inaccurate demographic history estimations, as shown in [60, 50]. 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 not greatly 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 missing SNPs have a lesser effect. 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 TEs), there is a strong effect on inference. 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 have 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 [50], 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 bias demographic inference but slightly increases the variance of the results and leads to small waves in the demographic history (as consequences of erroneously estimated hidden state transition events because of the non constant recombination rate along the sequence). Those results are similar to the one obtained in [25]. On the other hand, 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 results is surprisingly different than those obtained in [25] (although the variation of mutation rate was within a scaffold in their study). This discrepancy could suggest analyses based on longer scaffold 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. In addition, it could provide useful insight in unveiling any variation in molecular forces along the genome, albeit in a coarser way than in [1].

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 asses 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 1), 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 a 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 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 [3, 48].

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 offer new interpretations of results obtained when hypotheses are known to be violated, and thus offer an explanation as to 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 which can be used to estimate the demographic history and recombination rate given a genealogy (in the same spirit as Popsicle [13]). 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 panmictic population [8, 46]). This statistic could, in future, be used in more complex scenarios, 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.

5 Acknowledgments

This work was funded by $809_7/1 DeutscheForschungsgemeinschaft(https://www.dfg.de/)$ to AT. DAA was funded by the Alexander von Humboldt Stiftung (https://www.humboldt-foundation.de/web/home.html).

References

[1] 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.

[2] 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, Year = 2018, Volume = 19, Month = SEP 25, DOI = 10.1186/s13059-018-1517-y, Article-Number = 145, ISSN = 1474-760X, ORCID-Numbers = Beeravolu Reddy, Champak/0000-0002-*

0800-1994 Frantz, Laurent/0000-0001-8030-3885, Times-Cited = 3, Unique-ID = ISI:000445752300004,.

[3] 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.

[4] Anders Bergstrom, Shane A. McCarthy, Ruoyun Hui, Mohamed A. Almarri, Qasim Ayub, Petr Danecek, Yuan Chen, Sabine Felkel, Pille Hallast, Jack Kamm, Helene Blanche, Jean-Francois Deleuze, Howard Cann, Swapan Mallick, David Reich, Manjinder S. Sandhu, Pontus Skoglund, Aylwyn Scally, Yali Xue, Richard Durbin, and Chris Tyler-Smith. Insights into human genetic variation and population history from 929 diverse genomes. *SCIENCE*, 367(6484, SI):1339+, MAR 20 2020.

[5] 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.

[6] Jun Cao, Korbinian Schneeberger, Stephan Ossowski, Torsten Guenther, Sebastian Bender, Joffrey Fitz, Daniel Koenig, Christa Lanz, Oliver Stegle, Christoph Lippert, Xi Wang, Felix Ott, Jonas Mueller, Carlos Alonso-Blanco, Karsten Borgwardt, Karl J. Schmid, and Detlef Weigel. Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nature Genetics*, 43(10):956–U60, OCT 2011.

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

[8] 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.

[9] Slew Woh Choo, Mike Rayko, Tze King Tan, Ranjeev Hari, Aleksey Komissarov, Wei Yee Wee, Andrey A. Yurchenko, Sergey Kliver, Gaik Tamazian, Agostinho Antunes, Richard K. Wilson, Wesley C. Warren, Klaus-Peter Koepfli, Patrick Minx, Ksenia Krasheninnikova, Antoinette Kotze, Desire L. Dalton, Elaine Vermaak, Ian C. Paterson, Pavel Dobrynin, Frankie Thomas Sitam, Jeffrine J. Rovie-Ryan, Warren E. Johnson, Aini Mohamed Yusoff, Shu-Jin Luo, Kayal Vizi Karuppannan, Gang Fang, Deyou Zheng, Mark B. Gerstein, Leonard Lipovich, Stephen J. O'Brien, and Guat Jah Wong. Pangolin genomes and the evolution of mammalian scales and immunity. *GENOME RESEARCH*, 26(10):1312–1322, OCT 2016.

[10] Robert Ekblom, Birte Brechlin, Jens Persson, Linnea Smeds, Malin Johansson, Jessica Magnusson, Oystein Flagstad, and Hans Ellegren. Genome sequencing and conservation genomics in the Scandinavian wolverine population. *Conservation Biology*, 32(6):1301–1312, DEC 2018.

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

[12] 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.

[13] Lucie Gattepaille, Torsten Guenther, and Mattias Jakobsson. Inferring Past Ef-

fective Population Size from Distributions of Coalescent Times. *Molecular Biology and Evolution*, 204(3):1191+, NOV 2016.

[14] 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.

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

[16] 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.

[17] Sarah Hendricks, Eric C. Anderson, Tiago Antao, Louis Bernatchez, Brenna R. Forester, Brittany Garner, Brian K. Hand, Paul A. Hohenlohe, Martin Kardos, Ben Koop, Arun Sethuraman, Robin S. Waples, and Gordon Luikart. Recent advances in conservation and population genomics data analysis. *Evolutionary Applications*, 11(8):1197–1211, SEP 2018.

[18] 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.

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

[20] 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.

[21] 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.

[22] 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.

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

[24] Sally C. Y. Lau, Nerida G. Wilson, Catarina N. S. Silva, and Jan M. Strugnell. Detecting glacial refugia in the Southern Ocean. *ECOGRAPHY*.

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

[26] Shengbin Li, Bo Li, Cheng Cheng, Zijun Xiong, Qingbo Liu, Jianghua Lai, Hannah V. Carey, Qiong Zhang, Haibo Zheng, Shuguang Wei, Hongbo Zhang, Liao Chang, Shiping Liu, Shanxin Zhang, Bing Yu, Xiaofan Zeng, Yong Hou, Wenhui Nie, Youmin Guo, Teng Chen, Jiuqiang Han, Jian Wang, Jun Wang, Chen Chen, Jiankang Liu, Peter J. Stambrook, Ming Xu, Guojie Zhang, M. Thomas P. Gilbert, Huanming Yang, Erich D. Jarvis, Jun Yu, and Jianqun Yan. Genomic signatures of near-extinction and rebirth of the crested ibis and other endangered bird species. *GENOME BIOLOGY*, 15(12), 2014.

[27] 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.

[28] 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.

[29] Anna-Sapfo Malaspinas, Michael C. Westaway, Craig Muller, Vitor C. Sousa, Oscar Lao, Isabel Alves, Anders Bergstrom, Georgios Athanasiadis, Jade Y. Cheng, Jacob E. Crawford, Tim H. Heupink, Enrico Macholdt, Stephan Peischl, Simon Rasmussen, Stephan Schiffels, Sankar Subramanian, Joanne L. Wright, Anders Albrechtsen, Chiara Barbieri, Isabelle Dupanloup, Anders Eriksson, Ashot Margaryan, Ida Moltke, Irina Pugach, Thorfinn S. Korneliussen, Ivan P. Levkivskyi, J. Vctor Moreno-Mayar, Shengyu Ni, Fernando Racimo, Martin Sikora, Yali Xue, Farhang A. Aghakhanian, Nicolas Brucato, Soren Brunak, Paula F. Campos, Warren Clark, Sturla Ellingvag, Gudjugudju Fourmile, Pascale Gerbault, Darren Injie, George Koki, Matthew Leavesley, Betty Logan, Aubrey Lynch, Elizabeth A. Matisoo-Smith, Peter J. McAllister, Alexander J. Mentzer, Mait Metspalu, Andrea B. Migliano, Les Murgha, Maude E. Phipps, William Pomat, Doc Reynolds, Francois-Xavier Ricaut, Peter Siba, Mark G. Thomas, Thomas Wales, Colleen Ma'run Wall, Stephen J. Oppenheimer, Chris Tyler-Smith, Richard Durbin, Joe Dortch, Andrea Manica, Mikkel H. Schierup, Robert A. Foley, Marta Mirazon Lahr, Claire Bowern, Jeffrey D. Wall, Thomas Mailund, Mark Stoneking, Rasmus Nielsen, Manjinder S. Sandhu, Laurent Excoffier, David M. Lambert, and Eske Willerslev. A genomic history of Aboriginal Australia. *NATURE*, 538(7624):207+, OCT 13 2016.

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

[31] 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.

[32] Maja P. Mattle-Greminger, Tugce Bilgin Sonay, Alexander Nater, Marc Pybus, Tariq Desai, Guillem de Valles, Ferran Casals, Aylwyn Scally, Jaume Bertranpetit, Tomas Marques-Bonet, Carel P. van Schaik, Maria Anisimova, and Michael Kruetzen. Genomes reveal marked differences in the adaptive evolution between orangutan species. *Genome Biology*, 19, NOV 15 2018.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

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

[41] 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.

[42] Eleftheria Palkopoulou, Mark Lipson, Swapan Mallick, Svend Nielsen, Nadin Rohland, Sina Baleka, Emil Karpinski, Atma M. Ivancevici, Thu-Hien To, Daniel Kortschak, Joy M. Raison, Zhipeng Qu, Tat-Jun Chin, Kurt W. Alt, Stefan Claesson, Love Dalen, Ross D. E. MacPhee, Harald Meller, Alfred L. Ro-

car, Oliver A. Ryder, David Heiman, Sarah Young, Matthew Breen, Christina Williams, Bronwen L. Aken, Magali Ruffier, Elinor Karlsson, Jeremy Johnson, Federica Di Palma, Jessica Alfoldi, David L. Adelsoni, Thomas Mailund, Kasper Munch, Kerstin Lindblad-Toh, Michael Hofreiter, Hendrik Poinar, and David Reich. 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.

[43] Eleftheria Palkopoulou, Swapan Mallick, Pontus Skoglund, Jacob Enk, Nadin Rohland, Heng Li, Ayca Omrak, Sergey Vartanyan, Hendrik Poinar, Anders Gotherstrom, David Reich, and Love Dalen. Complete Genomes Reveal Signatures of Demographic and Genetic Declines in the Woolly Mammoth. *Current Biology*, 25(10):1395–1400, MAY 18 2015.

[44] Austin H. Patton, Mark J. Margres, Amanda R. Stahlke, Sarah Hendricks, Kevin Lewallen, Rodrigo K. Hamede, Manuel Ruiz-Aravena, Oliver Ryder, Hamish Mc-Callum, I, Menna E. Jones, Paul A. Hohenlohe, and Andrew Storfer. 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.

[45] Javier Prado-Martinez, Peter H. Sudmant, Jeffrey M. Kidd, Heng Li, Joanna L. Kelley, Belen Lorente-Galdos, Krishna R. Veeramah, August E. Woerner, Timothy D. O'Connor, Gabriel Santpere, Alexander Cagan, Christoph Theunert, Ferran Casals, Hafid Laayouni, Kasper Munch, Asger Hobolth, Anders E. Halager, Maika Malig, Jessica Hernandez-Rodriguez, Irene Hernando-Herraez, Kay Pruefer, Marc Pybus, Laurel Johnstone, Michael Lachmann, Can Alkan, Dorina Twigg,

Natalia Petit, Carl Baker, Fereydoun Hormozdiari, Marcos Fernandez-Callejo, Marc Dabad, Michael L. Wilson, Laurie Stevison, Cristina Camprubi, Tiago Carvalho, Aurora Ruiz-Herrera, Laura Vives, Marta Mele, Teresa Abello, Ivanela Kondova, Ronald E. Bontrop, Anne Pusey, Felix Lankester, John A. Kiyang, Richard A. Bergl, Elizabeth Lonsdorf, Simon Myers, Mario Ventura, Pascal Gagneux, David Comas, Hans Siegismund, Julie Blanc, Lidia Agueda-Calpena, Marta Gut, Lucinda Fulton, Sarah A. Tishkoff, James C. Mullikin, Richard K. Wilson, Ivo G. Gut, Mary Katherine Gonder, Oliver A. Ryder, Beatrice H. Hahn, Arcadi Navarro, Joshua M. Akey, Jaume Bertranpetit, David Reich, Thomas Mailund, Mikkel H. Schierup, Christina Hvilsom, Aida M. Andres, Jeffrey D. Wall, Carlos D. Bustamante, Michael F. Hammer, Evan E. Eichler, and Tomas Marques-Bonet. Great ape genetic diversity and population history. *NATURE*, 499(7459):471–475, JUL 25 2013.

[46] 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.

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

[48] Joshua G. Schraiber and Joshua M. Akey. Methods and models for unravelling human evolutionary history. *NATURE REVIEWS GENETICS*, 16(12):727–740, DEC 2015.

[49] 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.

[50] 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.

[51] 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.

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

[53] Montgomery Slatkin. Statistical methods for analyzing ancient DNA from hominins. *CURRENT OPINION IN GENETICS & DEVELOPMENT*, 41:72–76, DEC 2016.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] 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.

[62] 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.

[63] 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.

[64] Pengcheng Wang, Hongyan Yao, Kadeem J. Gilbert, Qi Lu, Yu Hao, Zhengwang Zhang, and Nan Wang. 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.

[65] 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.

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

[67] Chee-Wei Yew, Dongsheng Lu, Lian Deng, Lai-Ping Wong, Rick Twee-Hee Ong, Yan Lu, Xiaoji Wang, Yushimah Yunus, Farhang Aghakhanian, Siti Shuhada Mokhtar, Mohammad Zahirul Hoque, Christopher Lok-Yung Voo, Thuhairah Abdul Rahman, Jong Bhak, Maude E. Phipps, Shuhua Xu, Yik-Ying Teo, Subbiah Vijay Kumar, and Boon-Peng Hoh. 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.