

RESOURCE ARTICLE

Limits and convergence properties of the sequentially Markovian coalescent

Thibaut Paul Patrick Sellinger  | Diala Abu-Awad | Aurélien Tellier 

Department of Life Science Systems,
Technical University of Munich, Munchen,
Germany

Correspondence

Thibaut Paul Patrick Sellinger, Department
of Life Science Systems, Technical
University of Munich, Liesel-Beckmann
Strasse 2, 85354 Freising, 80333,
Munchen, Germany.
Email: thibaut.sellinger@tum.de

Funding information

Deutsche Forschungsgemeinschaft,
Grant/Award Number: 317616126
(TE809/7-1); Technische Universitat
Munchen

Abstract

Several methods based on the sequentially Markovian coalescent (SMC) make use of full genome sequence data from samples to infer population demographic history including past changes in population size, admixture, migration events and population structure. More recently, the original theoretical framework has been extended to allow the simultaneous estimation of population size changes along with other life history traits such as selfing or seed banking. The latter developments enhance the applicability of SMC methods to nonmodel species. Although convergence proofs have been given using simulated data in a few specific cases, an in-depth investigation of the limitations of SMC methods is lacking. In order to explore such limits, we first develop a tool inferring the best case convergence of SMC methods assuming the true underlying coalescent genealogies are known. This tool can be used to quantify the amount and type of information that can be confidently retrieved from given data sets prior to the analysis of the real data. Second, we assess the inference accuracy when the assumptions of SMC approaches are violated due to departures from the model, namely the presence of transposable elements, variable recombination and mutation rates along the sequence, and SNP calling errors. Third, we deliver a new interpretation of SMC methods by highlighting the importance of the transition matrix, which we argue can be used as a set of summary statistics in other statistical inference methods, uncoupling the SMC from hidden Markov models (HMMs). We finally offer recommendations to better apply SMC methods and build adequate data sets under budget constraints.

KEYWORDS

ancestral recombination graph, kingman coalescent, population genetics

1 | INTRODUCTION

With advances in sequencing technologies, recovering the demographic history of a population has become central to many studies in evolutionary biology, as it allows us to understand the environmental and demographic changes that existing and/or extinct species have

experienced (population expansion, colonization of new habitats, past bottlenecks, migration and admixture events' Arredondo et al., 2020; Bergstrom, 2020; Chikhi et al., 2018; Lord, 2020; Mazet et al., 2016; Palkopoulou, 2018; Rodriguez et al., 2018; Steinrucken et al., 2019). Inferences of demographic history that rely on genomic data (Schraiber & Akey, 2015) can thereafter be linked to archaeological

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2021 The Authors. *Molecular Ecology Resources* published by John Wiley & Sons Ltd.

or climatic data, providing novel insights on the evolutionary history of species (Barroso et al., 2019; Fulgione et al., 2018; Palkopoulou, 2018; Willemsen et al., 2020; Yew, 2018). Current statistical tools can uncover past population size variation Terhorst et al., (2017), Speidel et al., (2019), evidence for migration events (Steinrucken et al., 2019; Wang et al., 2020), as well as the genomic consequences of human activities on wild and domesticated species (Choo, 2016). The inference of demographic changes (e.g. decreasing effective population size) via genome data represents another way to monitor habitat loss or fragmentation (Ekblom et al., 2018; Hendricks, 2018; Kerdoncuff et al., 2020; Oh et al., 2019; Peart, 2020; Poelstra, 2021; Williams et al., 2020). These tools, when used to study the demographic histories of different species in relation to one another Hecht et al., (2018), Oaks et al., (2020), can unveil biological or environmental forces driving changes in species abundance and changes in ecosystem structuring. With the increased accuracy of current methods (Speidel et al., 2019) and the availability of very large and diverse data sets (Cao, 2011; Prado-Martinez, 2013; T. G. P. Consortium, 2012), the inferred demographic history, especially population size variation, is becoming more accessible to better study evolution (Willemsen et al., 2020), though there still exist some challenges (Beichman et al., 2017, 2018; Chikhi et al., 2018). SMC methods Li and Durbin (2011), Schiffels and Durbin (2014), Terhorst et al., (2017), Hecht et al., (2018), Palamara et al., (2018), Barroso et al., (2019), Sellinger et al., (2020), Wang et al., (2020), Ki and Terhorst (2020), which make use of whole genome polymorphism data, are among the most widely used methods for inferring past demography (Beichman et al., 2017; Mather et al., 2020; Spence et al., 2018). Though some works have addressed the limitations of current inference tools based on the SMC Hawks (2017), Chikhi et al., (2018), Rodriguez et al., (2018), Mazet et al., (2016), an in-depth and more comprehensive overview and some evaluations of their sensitivity to violations of the modelling assumptions are still lacking.

The SMC theoretical framework is founded on modelling the Ancestral Recombination Graph (*i.e.* the distribution of genealogies along the genome in presence of recombinations; McVean & Cardin, 2005). The first use of the SMC to infer past changes in population size is the now well-known pairwise sequentially Markovian coalescent (PSMC) method (Li & Durbin, 2011). This method uses the distribution of SNPs along the genome between two haploid genomes to account for and infer recombination and population size variation, assuming neutrality and panmixia. Although PSMC could infer population size variation in time with unprecedented accuracy, while requiring only one unphased sequenced diploid individual, PSMC had limited power in inferring more recent events. In order to address this issue, PSMC has been extended to account for multiple haploid genomes (*i.e.* more than two) in the method known as the multiple sequentially Markovian coalescent (MSMC) (Schiffels & Durbin, 2014). By using more sequences, MSMC better infers recent events and also provides the possibility of inferring population splits using the cross-coalescent rate but requires the data to be phased. Another difference between PSMC and MSMC is that the former is based on SMC theory (McVean & Cardin, 2005) and the latter on a

correction of this theory, known as SMC theory (Marjoram & Wall, 2006; therefore MSMC applied to only two haploid genomes has been defined as PSMC'). Methods developed after MSMC followed suit, with MSMC2 (Malaspina, 2016) 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++ (Terhorst et al., 2017) brings the SMC theory to another level by allowing the use of hundreds of unphased sequences 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 population size variation, its accuracy is increased in recent times (Terhorst et al., 2017). SMC++ is currently the state of the art SMC-based method for big data sets (>20 haploid genomes), but seems to be outperformed by PSMC when using smaller data sets (Patton, 2019).

Despite SMC methods performing very well when using simulated data (especially when using simple single-population models, based on typical data parameters of the human genome Schiffels & Durbin, 2014; Sellinger et al., 2020; Terhorst et al., 2017), we explicit here four reasons for which the method would present biased or poor estimates when applied to real data.

A first fundamental reason is that the accuracy of estimation depends on the ratio of effective recombination over effective mutation rates $\frac{\rho}{\theta}$ Sellinger et al. (2020), Terhorst et al., (2017), Barroso et al., (2019). 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 (e.g. Sellinger et al., 2020). There is no solution to this fundamental limitation of demographic inference methods as the ratio $\frac{\rho}{\theta}$ is fixed for a given species (note that in humans, this ratio is approximately 1).

A second fundamental issue when analysing past demographic events is the confounding role of natural selection (positive, balancing, purifying or background). For example, when unaccounted for, selective sweeps can result in a bottleneck followed by an expansion signature (*i.e.* 'U' shaped demographic history; Schrider et al., 2016). Moreover, background or pervasive positive selection leads to the underestimation of population size coupled with spurious and complex population size variation and a bias towards an expansion signal in the recent past Johri et al., (2021). There is currently no solution for this issue within the SMC theory, though it could be addressed using different theoretical frameworks that are being developed (Johri et al., 2020, 2021; Nakagome et al., 2019; Sheehan & Song, 2016).

Third, a conceptual issue when applying any inference methods in population genomics is the large number of underlying hypotheses of the models Li and Durbin (2011), Schiffels and Durbin (2014), which are potentially violated in genomic data. Several studies address the consequences of hypothesis violation on the accuracy of SMC methods (Chikhi et al., 2018; Hawks, 2017; Mazet et al., 2016; Rodriguez et al., 2018). In particular, unaccounted for population structure, admixture or introgression influence population size variation estimations (Chikhi et al., 2018; Hawks, 2017). We also showed

that ignoring two common traits in both plants and animals which are seed/egg banking and self-fertilization can lead to erroneous estimates of population size changes Sellinger et al., (2020).

Finally, several technical limitations can affect the inference accuracy or bias the results. For example, methods requiring phased data (e.g. MSMC, Schiffels & Durbin 2014) tend to strongly overestimate population sizes in recent time when errors in phasing occur Terhorst et al. (2017). Some methods have been shown to require high coverage for trustworthy results Nadachowska-Brzyska et al. (2016), and even though SMC methods seem robust to genome quality Patton (2019), there may be past demographic scenarios for which this is not the case. Therefore, one should keep in mind that the accuracy of SMC-based methods depends on which of the many underlying hypotheses are prone to being violated in real data sets as well as limitations originating from data quality.

In an attempt to complement previous works, we here study the limits and convergence properties of methods based on the sequentially Markovian coalescent, specifically those that focus on inferring changes in population size. It is important to keep in mind, that although SMC-based models may be theoretically similar, the difference in the model implementation can yield different outcomes when analysing one data set with different methods. In order to address both the theoretical limits and issues linked to the actual computational implementation, we compare four methods: MSMC (Schiffels & Durbin, 2014), MSMC2 (Malaspina, 2016), SMC++ (Terhorst et al., 2017) and eSMC (Sellinger et al., 2020), which we describe in more detail below. We introduce how these methods work, and what the underlying hypotheses are, followed by a definition of the limits of SMC-based methods (i.e. how well they perform theoretically), denoted here as the 'best-case convergence'. This convergence is then compared to results obtained using simulated sequences, so that we can examine the convergence properties in the absence of hypothesis violation. We test several demographic scenarios, as well as study the effect of the optimization function (or composite likelihood) and the time window of the analysis on the estimations of different variables. The effects of commonly violated hypotheses are also tested, 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. Finally, we provide guidelines to interpret abnormal or unexpected results, hinting at specific hypothesis violations, so as to guide users who wish to apply SMC-methodology to their data sets.

2 | METHODS

2.1 | Theoretical foundations of SMC methods

Before detailing how we test the limitations of methods used to infer past variation in population size, it is essential to quickly introduce the theory of the sequentially Markovian coalescent, hidden Markov models and algorithm used for statistical inference. For an additional introduction to the SMC, see Mather et al. (2020).

2.1.1 | The sequentially Markovian coalescent

Inference of past events rely on population genetics theory Wakeley (2020). The population history can be recovered based on the genealogy of sampled individuals Gattepaille et al. (2016). It is assumed that the population follows a neutral model of evolution and that the genealogy of the sample can be described using the Kingman n -coalescent model. This model allows the length of the genealogy in a sample of size n to be connected with the number of polymorphisms observed. In the case of a sample size two, the length of the genealogy until the most recent common ancestor of the sample is directly related to the amount of polymorphic sites.

However, the genealogy varies along the genome due to recombination events, a process which is modelled using the Ancestral Recombination Graph (ARG). The distribution of the ancestral recombination graph of a sample has been described under a Wright-Fisher model in Hudson (1983). Unfortunately, computations under this model can become very intensive with increasing sample size or sequence length Hudson (1983). This computational load can make inferences or simulations intractable. Therefore, a new process has been introduced to model the ARG as an inhomogeneous Poisson process along the sequence Wiuf and Hein (1999). This Poisson process has further been approximated through a Markov chain, implying that all the information necessary to compute the distribution of the genealogy at one position is contained in the genealogy of the previous position (McVean & Cardin, 2005).

2.1.2 | Hidden Markov Models and parameter inference

All SMC methods are in fact Hidden Markov Models (HMM). This means that the observed data are considered to be a signal that is emitted by an underlying, but unobservable, Markov process. Here, the exact genealogy of all individuals from a population/species is unknown; hence, the genealogy can be considered as a latent (hidden/unobserved) variable from which results the observed DNA sequences (i.e. the observed data are conditioned on the unobservable genealogy). Thus, based on SNP data and molecular parameters (e.g. mutation and recombination rates), the ARG can be inferred by considering the genealogy (or coalescence time to the most recent common ancestor) as a hidden state, and the sequence polymorphism data as the observed signal. The series of hidden states (i.e. the coalescent times or times to the most recent common ancestor of a sample) along the genome is therefore assumed to be a Markov process, which we can model using the SMC theory as explained above.

In practice, we are not interested in the hidden states themselves, but the parameters of the Markov process (e.g. population size, recombination rates). These parameters can be inferred by maximizing the likelihood of the modelled Markov process with all parameters calculated from the given sequence data. To do so, there are two main options. One can directly maximize the likelihood through the Forward Algorithm or the Baum-Welch algorithm. As the first option

is computationally very intensive, making optimization intractable for complex models, all SMC methods use the much more tractable Baum-Welch algorithm (described in section 2 of the Supporting Information of Terhorst et al. (2017)). The Baum-Welch algorithm is an Expectation-Maximization algorithm for HMM. Expectation-Maximization (EM) algorithms are iterative and alternate between performing the expectation step, to create an objective function using the current estimates of the parameters, and the maximization step, at which the parameter values maximizing the objective function are computed and updated. However, implementations can differ and the Baum-Welch algorithm is currently based on two different (but very similar) objective functions to infer the model parameters during the maximization step. The possible implementations use either the originally described objective function (denoted here as the complete Baum-Welch algorithm), or with a truncated objective function (here the incomplete Baum-Welch algorithm). The objective function for the complete Baum-Welch algorithm is given by:

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

and the truncated version of the objective function by:

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

with:

- v_{θ^t} : 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 .

Here, $P(X|Z,\theta)$ describes the transition probabilities from state Z to state X , conditional to the set of parameters θ (e.g. the recombination rate and population size). In practice, this probability is represented by a square matrix of size k (k being the number of hidden states). This matrix, which is known as the transition matrix, contains the predicted transition probabilities from one hidden state to another (i.e. the probabilities of coalescence times at a given genomic position, conditioned on the coalescence time at the previous position on the genome) calculated using the SMC theoretical framework. In addition, $E(X,Z|\theta^t)$ is the expected number of transitions of X from Z conditional to the observation and set of parameters θ^t .

$E(X,Z|\theta^t)$ can also be seen as a square matrix of size k , containing all the expected numbers of transitions from one state to another in our data and the set of parameters θ^t . Hence, we call this matrix the expected transition matrix. This matrix (calculated during the Expectation step) can be efficiently computed through the use of the Forward and Backward algorithm, well described in Sand et al., (2013), Terhorst et al., (2017). Intuitively, during the Maximization step, $Q(\theta|\theta^t)$ is maximized when the transition matrix (i.e. $P(X|Z,\theta)$) is similar to the estimated one (i.e. $E(X,Z|\theta^t)$).

2.1.3 | Best-case convergence

In order to measure the theoretical performance of SMC methods, we use a similar approach to Gattepaille et al. (2016), Johndrow and Palacios (2019), in which simulated Ancestral Recombination Graphs (ARG) are used as input. Using sequence simulators such as msprime (Kelleher et al., 2016) or scrm (Staab et al., 2015), one can simulate the Ancestral Recombination Graph (ARG) of a sample, usually given through a sequence of genealogies (e.g. a sequence of trees in Newick format). This exact ARG is then used to build the series of hidden states along the genomes and thus obtain the correct estimated transition matrix of the simulated data (mentioned above). Using this estimated transition matrix built directly from the exact ARG, one can estimate parameters as if the algorithm could correctly infer the hidden states (i.e. build the correct objective function). As the estimation matrix is built from the correct ARG, the results obtained represent the upper bound of performance for these methods. Since, in practice, the correct objective function can never be built (there are biases in estimating the ARG and the estimation matrix will inevitably be inexact), we choose to call this upper bound the best-case convergence. For this study's purpose, a second version of the R package eSMC Sellinger et al. (2020) was developed. This package enables the building of the estimated transition matrix (for eSMC or MSMC) from simulated ARG (or ARG obtained from real data) and can then use this matrix to infer population size variation. The package and its description can be found at: <https://github.com/TPPSellinger/eSMC2>.

2.2 | SMC methods

In this study, we focus on four different SMC-based methods: MSMC, MSMC2, SMC++ and eSMC. As explained above, all these methods are Hidden Markov Models and use whole genome sequence polymorphism data as input. 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 its different convergence properties (Schiffels & Durbin, 2014), (ii) MSMC2 computes coalescence times of all pairwise analyses from a sample of size n and can deal with a large range of sample sizes and sequence lengths (Malaspina, 2016), (iii) SMC++ (Terhorst et al., 2017) is the most advanced and efficient SMC method and lastly,

(iv) eSMC (Sellinger et al., 2020) is a re-implementation of PSMC' (similar to MSMC2). Using eSMC contributes to highlighting the importance of algorithmic translations as it is presently modified to output results and intermediate results necessary for this study. All the command lines to analyse the generated data can be found in the Appendix S2.

2.2.1 | PSMC', MSMC2 and eSMC

PSMC' and methods that stem from it [MSMC2 (Malaspina, 2016) and eSMC (Sellinger et al., 2020)] focus on the coalescence events between only two haploid genomes (or one unphased diploid genome), and, as a result, do 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. The presence or absence of a segregating site along the sequence is used to infer the hidden state (*i.e.* coalescence time). However, the hidden state is only allowed to change in the event of recombination (Wiuf & Hein, 1999). Thus, the population recombination rate ρ constrains the inferred changes of hidden states along the sequence [for a detailed description of the algorithm, see Schiffels and Durbin (2014), Wang et al. (2020), Sellinger et al. (2020)]. MSMC2 uses the complete Baum-Welch algorithm (equation 1), whereas PSMC' uses the truncated version (equation 2).

2.2.2 | MSMC

Unlike other SMC methods, MSMC simultaneously analyses and models the genealogy of 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 analyse more than 10 sequences simultaneously (for a detailed description see Schiffels and Durbin (2014)).

2.2.3 | SMC++

Though conceptually very similar to PSMC', SMC++ is built with different mathematical functions and implementation. SMC++ also uses a more complex signal (*i.e.* observed data) compared to previous methods. Assuming n haploid genomes, SMC++ calculates the sample frequency spectrum of sample size $(n - 2) + 2$, conditioned on the coalescence time of two 'distinguished' haploids and $(n - 2)$ 'undistinguished' haploids. As fully describing SMC++ goes beyond the scope of this study, we direct indefatigable readers to Section 1 of the Supporting Information in Terhorst et al. (2017). In addition SMC++ offers features such as a cubic spline to estimate population size variation to obtain continuous changes in population size, unlike

other models which discretize changes by assuming a piece-wise constant population size.

2.2.4 | Time window

Each tested SMC-based method has its own specific time window, that is the interval of time in the past within which estimations are made. Hidden states are generally defined as discretized intervals of this time window, and as a consequence, boundaries and the length and the number of states implicitly affect the inferred parameters. This complicates one-to-one comparisons of the different methods. Using the updated eSMC package which allows users to set the time window, we test how the defined window affects the accuracy of the inference. We analyse the same data with four different settings: (i) the PSMC' time window Schiffels and Durbin (2014), (ii) a 'long' time window, which goes further in the past and in more recent time, used in MSMC2 (Wang et al., 2020; and similar to the one of the original PSMC Li & Durbin, 2011), (iii) a time window equivalent to the first one (*i.e.* PSMC') shifted by a factor five in the past (*i.e.* multiplied by five) and (iv) a time window equivalent to the first one, but shifted by a factor five in recent time (*i.e.* divided by five).

2.2.5 | Regularization penalty

To avoid inferring unrealistic demographic histories with very large or very rapid variations in populations sizes, SMC++ introduced a regularization penalty. In SMC++, the lower the value of the penalty, the more the estimated population size history becomes flat and tends towards constant population size over time. For comparison, a regularization penalty is also newly introduced in eSMC. Setting the regularization penalty parameter to 0 is equivalent to no penalization, and the higher the parameter value, the more population size variations are penalized (<https://github.com/TPPSellinger/eSMC2> for more details). We tested the effect of regularization on inferences with both methods using simulated sequence data. The sequence data are simulated under sawtooth demographic scenarios with different amplitudes of population size variation.

2.3 | Simulated sequence data

Throughout this study, we simulate different demographic scenarios using either the coalescence simulation program *scrm* Staab et al. (2015) or *msprime* Kelleher et al. (2016). 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* to simulate data for eSMC, MSMC and MSMC2. We use *msprime* to simulate data for SMC++ since *msprime* is more efficient than *scrm* for big sample sizes Kelleher et al., (2016) and can directly output *.vcf* files (which is the input format of SMC++). All the command lines to simulate data can be found in Appendix S1.

2.3.1 | Absence of hypothesis violation

We simulate five different demographic scenarios consisting of changes in population size: 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 for four amplitudes (*i.e.* by how many fold the population size varies: 2, 5, 10 and 50). In the sawtooth demographic scenario, each 'tooth' (*i.e.* episode of expansion/decrease or decrease/expansion) is of the assumed amplitude, thus leading to a variation of fold 4, 25, 100 and 2500, respectively, between the minimum and maximum observed population size. 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} (<https://github.com/stschiff/msmc/blob/master/guide.md>). When analysing simulated sequence data, we simulate sequences of 100 Mb: two sequences for eSMC and MSMC2, four sequences for MSMC and twenty sequences for SMC++ as, according to the guidelines provided for each method, they correspond to the recommended quantity of data required Terhorst et al. (2017), Schiffels and Durbin (2014), Sellinger et al. (2020).

2.3.2 | Calculation of the mean square error (MSE)

Because of differences in time-windows between methods, we evaluate the accuracy of each method by calculating the mean square error (MSE). To do so, we choose ten thousand points uniformly spread across the time window (in \log_{10} scale). We then calculate the MSE by comparing the actual population size and the one estimated by the method at each of the points. We thus have the following formula:

$$MSE = \frac{\sum_{i=1}^{10^4} (y_i - y_i^*)^2}{10^4} \quad (3)$$

where:

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

2.3.3 | Presence of hypothesis violation

We produce data sets simulated under scenarios which are challenging for SMC methods: SNP calling error, variation in mutation and recombination rates along the genome and presence of transposable elements.

2.3.4 | 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 the bioinformatics pipelines (read trimming, quality check, read mapping, and SNP calling), as well as the quality of the reference genome, data coverage and depth of sequencing and species ploidy (Pfeifer, 2017). 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 (Chang & Shapiro, 2016; Slatkin, 2016), SNPs can be simultaneously mistakenly missed or added. We model such events by simulating four sequences of 100 Mb under a 'sawtooth' scenario and then a certain percentage (0, 5, 10 and 25%) of SNPs is randomly added to and/or deleted from the simulated sequences. We then analyse the effect of SNP calling errors on the accuracy of population size variation 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 the SMC methods on a subset of the obtained data.

2.3.5 | Changes in mutation and recombination rates along the sequence

Because the recombination rate and the mutation rate can change along the sequence (Barroso et al., 2019), 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 1.25×10^{-8} per generation per base pair (between 2.5×10^{-9} and 6.25×10^{-8}). We simulate 20 scaffolds of size 2 Mb, as this seems representative of the best available assembly for non-model organisms (*e.g.* Lynch et al., 2017; Stam et al., 2019). We then analyse 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 fivefold) around an average value of 1.25×10^{-8} (the mutation rate being fixed at 1.25×10^{-8} per generation per bp) to study the effect of the assumption of a constant recombination rate along the sequence.

2.3.6 | 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 (Nelson et al., 2017) and simulate them (Kofler, 2018), understanding how their presence/absence influences demographic inferences remains limited. 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 their presence/absence polymorphism in individuals of a population (Ewing, 2015). 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 (Platt et al., 2016). To best capture and mimic the effect of TEs unaccounted for in the data, we altered four simulated haploid sequences of length 20 Mb in four different ways. The first way simulates the effect of unmapped and unaccounted TEs, done by assuming that they exhibit presence/absence polymorphism, hence creating gaps in the sequence. For each individual, we remove small pieces of sequence of different length (1, 10 or 100 kb), so as to remove a percentage (5, 10, 25, 50%) of the original simulated sequence, and thus shorten and fragment the sequence to be analysed. The second way to model the effects of TEs is to consider unmasked TEs. This is done by randomly selecting small pieces of the original simulated sequence (1, 10 or 100 kb) that make up a certain percentage of it (5, 10, 25, 50%) and removing all the SNPs in those regions (i.e. removing mutations from TEs). The removed SNPs are hence structured in many small regions along the genome. Third, we test the consequences of simultaneously having both removed and unmasked TEs in the data set by combining the first two methods. Last, 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 | RESULTS

3.1 | Best-case convergence

In Figure 1, we show the results of the best-case convergence of eSMC under the sawtooth demographic scenario, with similar results obtained for the three other demographic scenarios (bottleneck, expansion and decrease), respectively, displayed in Figures S1–S3. We generally find that increasing the sequence length increases accuracy and reduces variability, leading to better convergence and reducing the mean square error (see Figure 1a–c for eSMC and Table S1). However, when the amplitude of population size variation is too great, the population size variation cannot be retrieved, even when using very large data sets (see Figure 1d). 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 Figure S4, 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 remain badly inferred.

In Figures S5–S9, we show the best-case convergence of MSMC with four genome sequences and generally find that these estimates present a higher variance than eSMC. However, MSMC shows better fits in recent times than eSMC and is better able to retrieve population size variation (see Figure S5d). Scenarios with strong variation of population size (i.e. with large amplitudes) still pose a problem

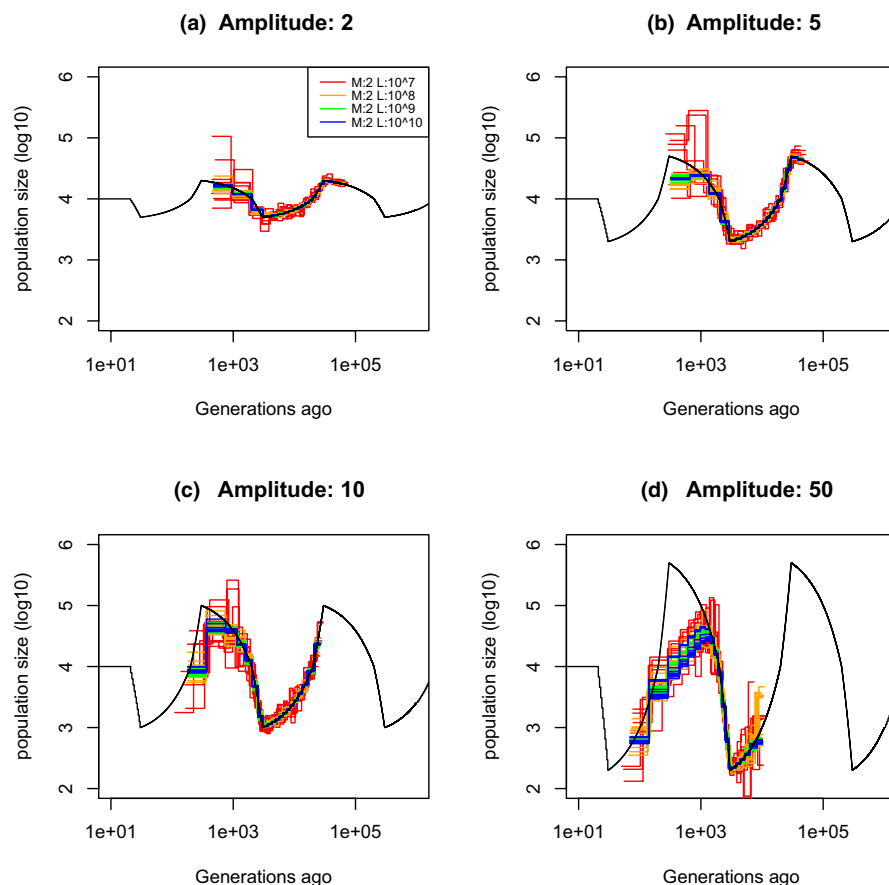


FIGURE 1 Best-case convergence of eSMC. Estimated population size variation using simulated genealogy over sequences of 10, 100, 1000, 10 000 Mb (in red, orange, green and blue, respectively) under a sawtooth scenario (original scenario in black) with 10 replicates for different 'tooth' amplitudes of 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

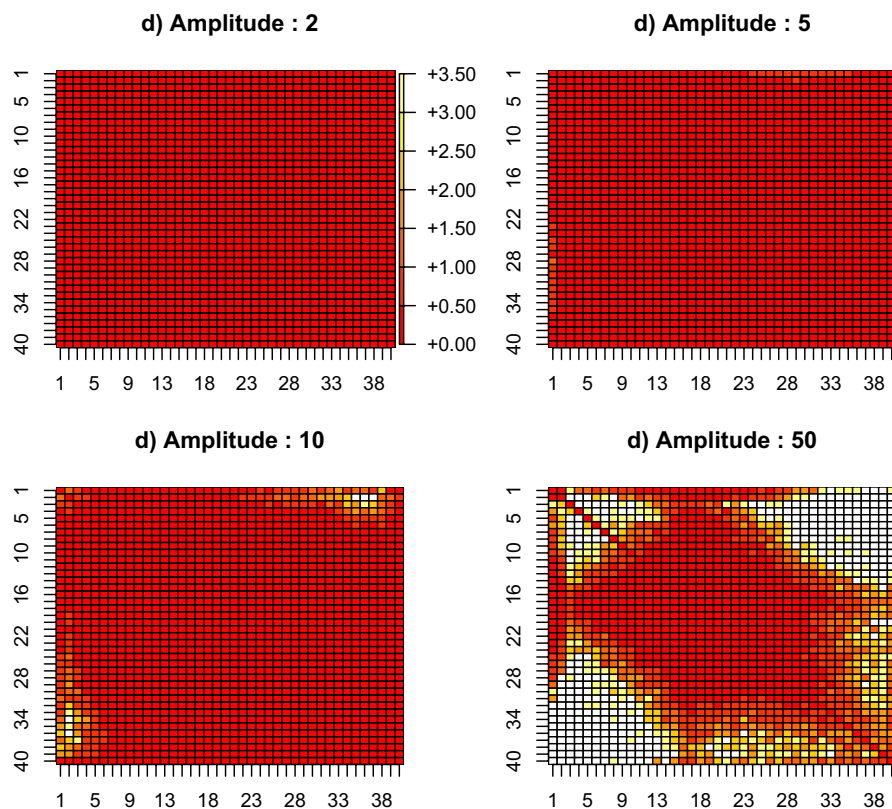


FIGURE 2 Estimated transition matrix in sharp sawtooth scenario. Estimated coefficient of variation of the transition matrix using simulated genealogy over sequences of 10 000 Mb under a sawtooth scenario of 'tooth' amplitude 2, 5, 10 and 50 (in a–d, respectively) each with 10 replicates. Recombination and mutation rates are as in Figure 1. White squares indicate absence of observed transitions (i.e. lack of observed hidden state transitions)

(Figure S9), and no matter the number of estimated parameters, such scenarios are also not retrievable using MSMC.

To better understand these results, we collect the estimated transition matrices (see Methods) from the exact ARGs of Figure 1. We then examine the coefficient of variation (the ratio of the standard deviation to the mean, indicating convergence when equal to 0) at each entry of the matrix calculated from the ten replicates, to study the distribution of the estimated matrices (results are plotted in Figure 2). For small amplitudes of population size variation, convergence is zero at almost all the matrix entries (Figure 2a). However, strong population size variation can lead to partially empty matrices and an increased coefficient of variation (Figure 2d). Unobserved transitions and increased coefficients of variation stem from the reduced probability of coalescence events in those time intervals (i.e. a lack of observation of some hidden states). This therefore results in the increased variability of the inferred parameters, meaning that SMC methods are incapable, even when perfectly inferring the hidden states, to correctly infer the population size variation in such cases. However, it is possible to reduce the coefficient of variation, rendering the inferences less variable, by increasing the sequence length, *that is* the number of observed transitions (see Figure S10).

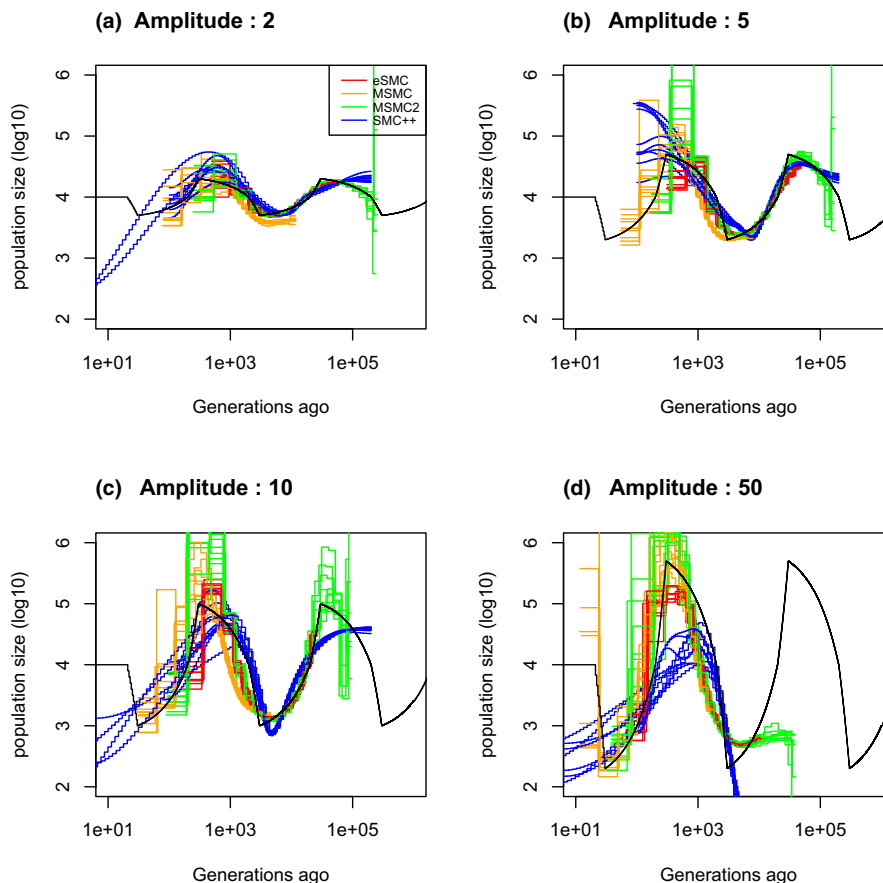
3.2 | Simulated sequence results

3.2.1 | Scenario effect

Having defined the theoretical limitations of eSMC and MSMC using the exact ARG, we now evaluate how these methods

perform when inputting simulated sequence data using the same parameter values as in the previous section. The difference to the previous section is that the SMC methods must additionally estimate the transitions between hidden states, thereby introducing another layer of statistical inference and thus possible noise in the resulting estimations. We first perform two benchmark analyses: the constant population size scenario (Figure S11) and the original sawtooth demographic scenario from Schiffels and Durbin (2014) (Figure S12). eSMC and MSMC2 are both able to retrieve the constant population size scenario, whereas MSMC fails to do so in the far past and SMC++ in recent time (Figure S11). All methods can retrieve the sawtooth demographic scenario, despite SMC++ displaying high variance in recent times (Figure S12). Second, we investigate the effect of amplitude of population size variation as in Figure 1. Results for the sawtooth scenario are shown in Figure 3, where the different models display a good fit, but are not as good as the best-case convergence given the same amount of data (orange line in Figure 1 and Table S1 vs the red line in Figure 3 and Table S2). As predicted by Figures 1 and 2, increasing the amplitude of population size variation diminished inference accuracy (see Table S2 for the MSE). All estimations display low variance and a relatively good fit in the bottleneck and expansion scenarios for small population size variation (see Figures S13a and S14a). However, the strengths of expansions and bottlenecks are not fully retrieved in scenarios with population size variation higher than tenfold the current population size (Figures S13c,d, and S14c,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 (i.e. Estimated transition matrix

FIGURE 3 Estimated demography using simulated sequences as input. Estimated population size variation under a sawtooth scenario (black) with 10 replicates using simulated sequences for different ‘tooth’ amplitudes of population size change: (a) 2, (b) 5, (c) 10 and (d) 50. Two sequences of 100 Mb for eSMC and MSMC2 (in red and green, respectively), 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 1×10^{-8} and 1.25×10^{-8}



vs True estimated matrix). In Figure S15, we show that the hidden states are harder to correctly infer in scenarios with strong population size variation, explaining the higher variance in Figure 3 compared to Figure 1. For the same amount of data, the inevitable inaccuracies in the estimated transition matrix contribute to erroneous population size inferences compared to the best-case convergence.

The variance observed in the inferences is also influenced by the time window, whose effect we tested using eSMC. Increasing the time window results in an increased variance of the inferences, as does shifting the window to more recent times, in the latter case resulting in poor estimations of population size variation (see Figure S16). Shifting the window further in the past does not seem to strongly impact the demographic inferences, though there are consequences on estimations of the recombination rates, as they are greatly over-estimated (Table 1). Concerning the optimization function, we find that the complete Baum-Welch algorithm gives similar results to the incomplete one (Table 1). This result, in addition to results of Figure 1, demonstrates that all the information is contained in the estimated transition matrix.

Adding a regularization penalty to eSMC can drastically impact inferences (Figure S17) and reduces performance quality. When regularization is added, eSMC fails to correctly capture the amplitude of population size variation and with extreme regularization penalty, eSMC infers a constant population size. Yet, adding regularization in SMC++ can increase performance and avoid spurious

variation of population size (Figure S18). However, strong regularization can lead to the inference of constant population size, independently of the underlying demographic scenario, and thus poor estimations.

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 variation in population size (Terhorst et al., 2017). 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 4). However, each method displays the worst fit when $\frac{\rho}{\theta} = 10$ (Table S3). SMC++ seems slightly less sensitive to $\frac{\rho}{\theta}$ than other methods. When calculating the difference between the transition matrix estimated by eSMC and the one built from the actual genealogy (ARG), 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 (*i.e.* staying in the same hidden state), explaining the underestimation of the recombination rate (see Figures S19–S21).

The reason for these results is as follows: if recombination occurs at a higher rate compared to mutation, then it impedes the detection of any recombination events that may have taken place

TABLE 1 Average estimated values for the recombination over mutation ratio $\frac{\rho}{\theta}$ by eSMC over ten repetitions for different sizes of the time window

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	0.79 (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)

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.

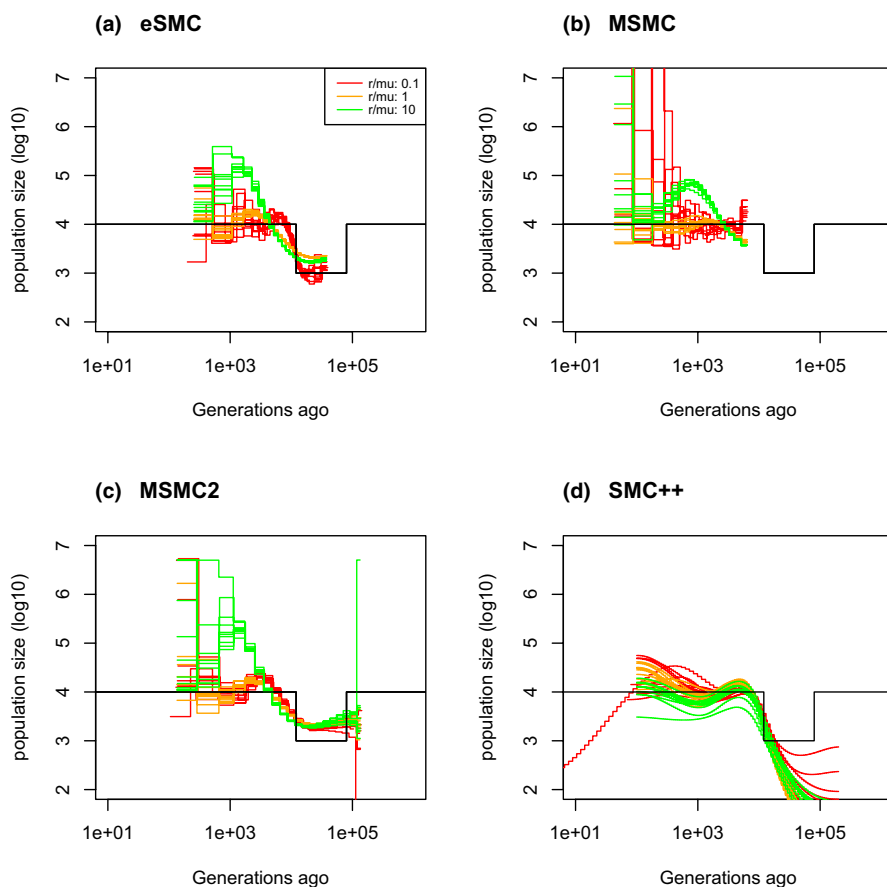


FIGURE 4 Effect of $\frac{\rho}{\theta}$ on inference of population size variation under a bottleneck scenario with 10 replicates using simulated sequences. We simulate two sequences of 100 Mb for eSMC and MSMC2 (in a and b, respectively), four sequences of 100 Mb for MSMC (c) and twenty sequences of 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 population size variations are in red, orange and green, respectively. Sequences are simulated under a bottleneck scenario of amplitude 10 and is represented in black

before the introduction of a new mutation, and thus biases the estimation of the coalescence time (Sellinger et al., 2020; Terhorst et al., 2017).

In some instances, we find it is possible to compensate for a high value (>1) of the ratio $\frac{\rho}{\theta}$ by increasing the number of iterations. Indeed, by doing so, eSMC better infers population size variation (Figure S22), although the correct recombination rate cannot be retrieved (Table 2). MSMC is better able to infer the correct recombination rate than other methods even when $\frac{\rho}{\theta} > 1$, but poorly estimates the variation in population size. MSMC2 and SMC++, however, are insensitive to an increased number of iterations, as the estimated population size variation is not improved (see Figure S22 and Table 2).

3.3 | Simulation results under problematic data and unaccounted for phenomena

3.3.1 | Imperfect SNP calling

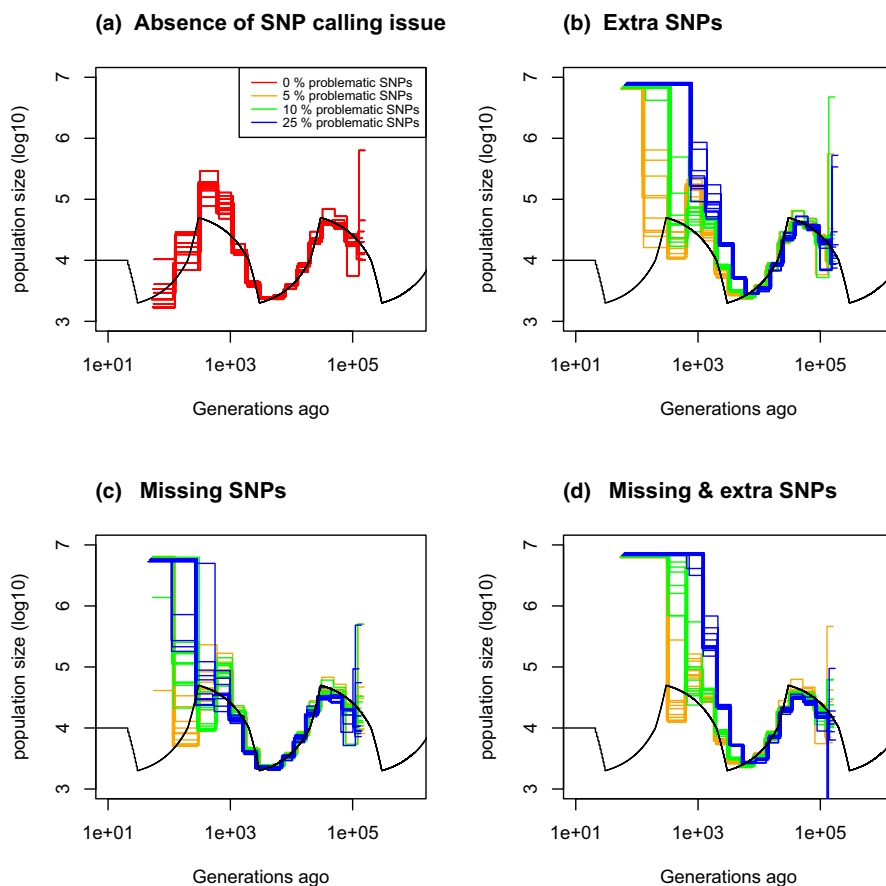
We analyse simulated sequences that have been modified by removing and/or adding SNPs in order to mimic errors in SNP calling. We find that, when using MSMC2, eSMC and MSMC, having more than 10% of spurious SNPs (e.g. low quality filtering) can lead to a strong over-estimation of population size in recent time but that missing SNPs have no effects on inferences in the far past and only mild effects on inferences in recent time (see Figure 5 for MSMC2, Figures S23 and S24 for eSMC and MSMC, respectively).

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

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)

For eSMC, MSMC and MSMC2, we have: set 1: 20 hidden states; set 2: 200 iterations; set 3: 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.

FIGURE 5 Consequences of SNP calling errors. Estimated population size variation 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 population size variation is represented in black. (a) Inferred population size variation in absence of SNP calling issues (red). (b) Inferred population size variation with 5% (orange), 10% (green) and 25% (blue) missing SNPs. (c) Inferred population size variation with 5% (orange), 10% (green) and 25% (blue) additional SNPs. (d) Inferred population size variation with 5% (orange), 10% (green) and 25% (blue) of additional and missing SNPs



The mean square error is displayed in Table S4, demonstrating that the better the filtering quality, the more accurate the population size inferences.

As complementary analyses, we analyse simulated sequences with a Minor Allele Frequency (MAF) threshold. We find that, the more SNPs are removed, the poorer the estimations in recent time (Figure S25), showing the impact of severe ascertainment bias.

3.3.2 | 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 analysed assuming scaffolds share or do not share the same recombination and mutation rates. As shown in Figure 6 (and Table S5), when scaffolds all share the same parameter values, estimated population size variation is accurate in both cases (*i.e.* assuming scaffolds share or not the same mutation and recombination rate). However, when scaffolds are simulated with different parameter values, analysing 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 estimated population size variation is only slightly inaccurate (Figure 6c),

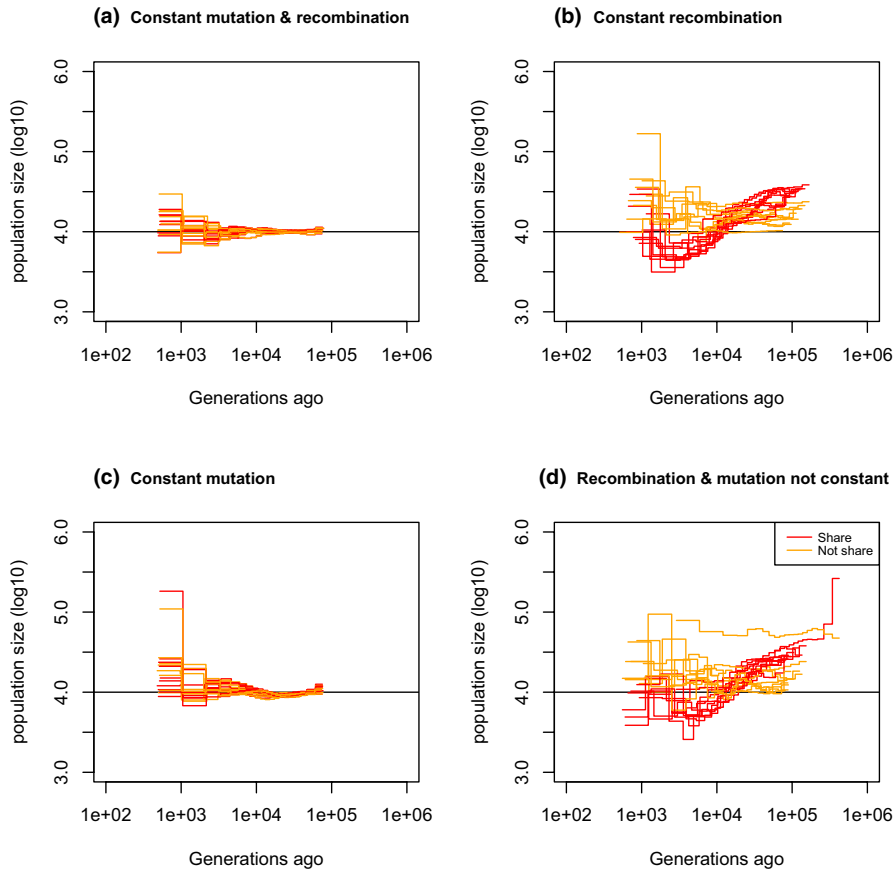


FIGURE 6 Inference of population size from scaffolds sharing or differing in mutation and recombination rates. Estimated population size variation using eSMC under a constant population size scenario with 10 replicates from twenty simulated scaffolds of 2 Mb (sample size of 2) assuming scaffolds share (red) or do not share recombination and mutation rates (orange). The simulated population size variation 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}

whereas, if the mutation rate changes from one scaffold to the other, population size variation is poorly estimated (Figure 6b).

Even if chromosomes are fully assembled, assuming we have one fully assembled scaffold of 40 Mb, 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 Figure S26, population size variation is well inferred, despite an increase in variance and a smooth 'wave' shaped population size variation when sequences are simulated with varying recombination rates throughout the genome compared to those with a fixed recombination rate. 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 N_e even though the true N_e is constant in time.

3.3.3 | How transposable elements affect inferences

Transposable elements (TEs) are present in most species, and are (if detected) taken into account as missing data by SMC methods (Schiffels & Durbin, 2014). Depending on how TEs affect the data set, we find that the different methods are more or less sensitive to TEs, but that they generally all follow similar trends. If TEs are unmapped/removed, there does not appear to be any bias in the estimated population size variation (see Figure 7 and Table S6 for eSMC and Figures S29 and S32 for MSMC and MSMC2, respectively). However, as can be seen from Table 3, there is an overestimation of $\frac{\rho}{\theta}$ and 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). If TEs are present but unmasked in the data set (and thus are not accounted for as missing data by the model; Schiffels and Durbin, 2014), we find that this is equivalent to a faulty calling of SNPs, in which SNPs are missing, hence resulting in population size variation estimations by eSMC similar to those observed in Figure 5a. However, if the size of unmasked TEs increases, different results are obtained. Indeed, in recent times there is a strong underestimation of population size and the model fails to capture the correct population size variation (see Figures S27 and S28 for eSMC, Figures S30 and S31 for MSMC and Figures S33 and S34 for MSMC2). The longer the TEs, the stronger the effect on the estimated population size variation. However, when TEs are detected and correctly masked, there is no effect on inferring population size variation (Figures S35 and S36).

4 | DISCUSSION

Inference methods based on the Sequentially Markovian Coalescent are robust and powerful tools that are being constantly extended to account for more complex scenarios. Here, we test the limits of four methods developed to infer past changes in size under the assumption of a single unstructured population (MSMC, MSMC2, SMC++ and eSMC) and define the parameter ranges in which they can be used more or less confidently. Through the application of the four tested SMC methods, we show that the inferred demographic history depends on

FIGURE 7 Consequences of masking or removing transposable elements (TEs) from data sets. Estimated population size variation by eSMC under a sawtooth 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 scenario is represented in black. Here the TEs are of length 1 kbp. (a) Inferred population size variation in absence TEs. (b) Inferred population size variation where TEs are removed. (c) Inferred population size variation where TEs are masked. (d) Inferred population size variation where half of the TEs are removed and SNPs on the other half are removed. Proportion of the genome made up by TEs is set to 0% (red), 5% (orange), 10% (green), 25% (blue) and 50% (purple)

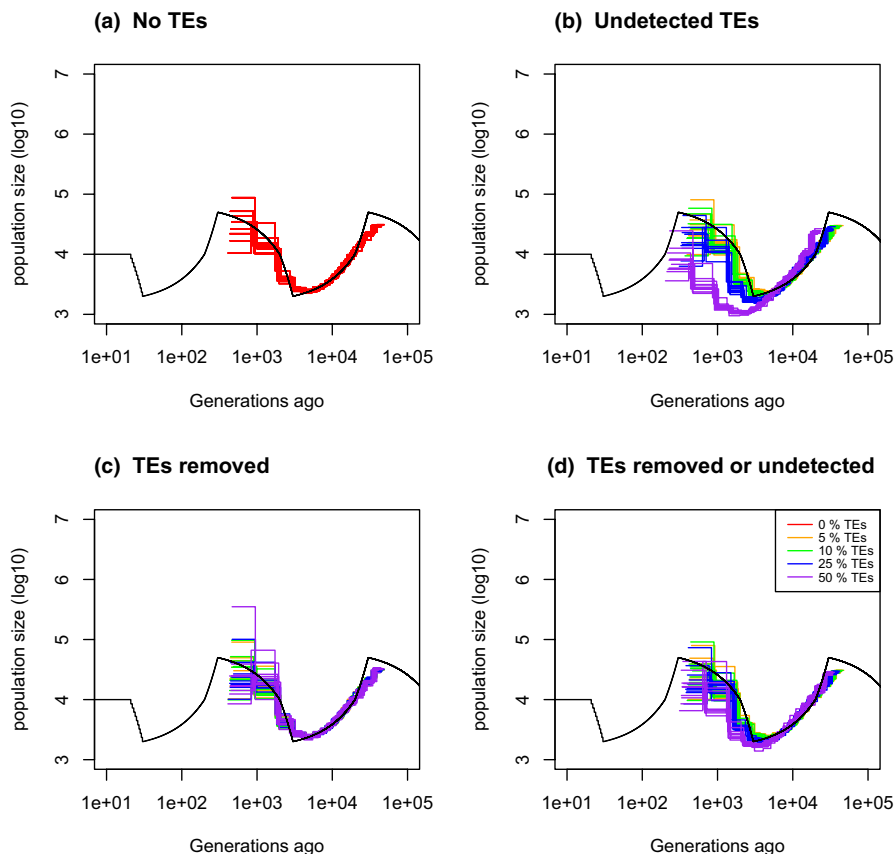


TABLE 3 Average estimated values for the recombination over mutation ratio $\frac{\rho}{\theta}$ over ten repetitions

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
3*1 kb	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)
3*10 kb	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 (0.090)	1.54 (0.11)	1.68 (0.13)
	MSMC2	1	0.87 (0.064)	0.89 (0.067)	0.99 (0.15)	1.13 (0.30)
3*100 kb	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)

The coefficient of variation is indicated in brackets. TEs are of length 1, 10 or 100 kb and are completely removed and the proportion of the genome made up by TEs is 5%, 10%, 25% and 50%.

the scenarios and amplitudes of population size changes, as there are cases where, even in an ideal situation, the current SMC framework is not able to recover the true scenario. Through the objective function of the Baum-Welch algorithm, we demonstrate that all the information is contained in the estimated transition matrix. We also highlight issues that may arise due to technical limitations when using genome data, as well as which assumption violations affect the performance of these methods. By comparing the different methods, we point out the complementarity between these methodologies, with some scenarios being better retrieved when using either MSMC or methods based on PSMC' (e.g. eSMC, MSMC2).

As PSMC', MSMC and other SMC-based methods use the Baum-Welch algorithm, they rely on correctly estimating of the ARG (*i.e.* the genealogies along the sequence), any bias in hidden state inferences decreases the accuracy of inferences. With this knowledge, users are able to associate errors in inference with issues in either the data set (for which there may be solutions, see below) or specific hypothesis violations. The package we have developed can be used to infer population size variation by inputting ARGs (trees in Newick format or sequences of coalescence events), independently of how the ARG has been estimated. As HMMs can be a computational burden under complex models and new methods are developed to accurately infer

genealogies (Kelleher et al., 2019; Ki & Terhorst, 2020; Speidel et al., 2019), we present a renewed interpretation of use of the SMC theory through the use of the estimated transition matrix. This matrix can be used for inference under more complex models as well as for hypothesis testing between different models/scenarios (as in Johndrow & Palacios, 2019). Our approach differs from previous works in that we use the series of hidden states built from the discretization of time summarized in a simple matrix, instead of the more computationally heavy estimations using the actual series of coalescence times (Gattepaille et al., 2016). The estimated transition matrix could thus become a powerful summary statistic in the future, facilitating the development of statistical approaches that encompass more complicated processes. Because new developments offer the possibility of detecting admixture using the SMC framework (Wang et al., 2020), our approach could be extended to account for population structure and migration (to some extent Kim et al., (2020)).

As the above conclusions might result from the fact that all SMC methods rely on the Baum-Welch algorithm for parameter inference (and thus on the accuracy of the HMM to estimate the transition matrix), this could potentially lead the algorithm to fall in local extrema when maximizing the likelihood. Convergence properties of SMC methods based on the direct optimization of the likelihood remain unknown. Therefore, some of the above mentioned issues and limitations could be overcome. The convergence properties of SMC methods based on maximizing the actual likelihood remains to be determined in order to test whether current SMC approaches fail to correctly optimize the likelihood or if limitations originate from the model itself. The first case would require current approaches to better explore the likelihood function to estimate parameters (this can be solved with more computational power), the latter would require to build a new theoretical framework to overcome those limitations to improve accuracy and robustness.

4.1 | General guidelines when applying SMC-based methods

Our aim through this work is to provide guidelines for using SMC-based methods. There are several aspects that must be taken into account when putting together a data set or analysing an existing one. As expected from previous works, the number and size of genome copies used both play an important role in the accurate estimation of population size variation (Gattepaille et al., 2016; Johndrow & Palacios, 2019). However, we find that the amount of data required for an accurate fit depends on the underlying demographic population size scenario, with bottlenecks often posing a problem. Indeed, SMC methods seem to incorrectly infer sudden and strong population size variation, resulting in unreliable inferences. Thus, the amplitude of population size variation also influences the estimation of model parameters, with high amplitudes leading to unobserved hidden state transition (thus partially empty estimated transition matrix), distorting the inferred population size variation. We show that the coefficient of variation of

the estimated transition matrix (using either real or simulated data) can indicate whether the amount of data is enough to retrieve a specific scenario (see also Figure S10).

It is also important to keep in mind that the accuracy of ARG inference by SMC methods 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 (Sellinger et al., 2020; Terhorst et al., 2017). 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 population size variation is of greater importance, PSMC'-based methods should be used. It is also advised to evaluate whether the size of the time window is adequate for the analysis (even though it is often fixed when using the different methods), as increasing its size will also increase the variance of the estimations.

4.2 | Guidelines when applying SMC-based methods on problematic data sets

Simulation results suggest that any variation of the recombination rate along the sequence slightly increases the variance of the results and leads to spurious small waves in population size variation, as expected from previous works Li and Durbin (2011). However, unlike results under the first PSMC method Li and Durbin (2011), if scaffolds do not share similar rates of mutation and recombination, but are analysed together assuming that they do, estimations are very poor. This discrepancy between our results and those in (Li & Durbin, 2011) may be due to the variation of the mutation rate being within a scaffolds in Li and Durbin (2011) compared to between scaffold in the present study. The results in Li and Durbin (2011) could also suggest that analyses based on longer scaffolds would be more robust. However, this problem can be avoided if each scaffold is assumed to have its own parameter values at the cost of increased computation time. Such analyses with varying rates between scaffolds could additionally provide useful insight in unveiling any variation in molecular forces along the genome, albeit in a coarser way than in Barroso et al., (2019). Since the consequences of a varying recombination rate might depend on the topology of the recombination map, we recommend estimating the recombination map (e.g. using iSMC (Barroso et al., 2019) or ReLERN (Adrian, 2020)). If problematic regions are found they can be masked with almost no negative impact on the estimated demography (Figure S35 and S36).

Imperfect data sets, due to technical errors such as SNP calling or to biological characteristics such as presence of transposable elements, can affect the inferences obtained using SMC-based methods. We show that data sets with more than 10% of spurious SNPs

lead to poor estimations of the population size variation, whereas randomly removed SNPs (*i.e.* missing SNPs) have a lesser effect on inferences. We thus recommend to be stringent during SNP calling, as a data set with spurious SNPs is worse than one with missing SNPs. Note, however, that this consideration is valid for inferences under a neutral model of evolution. 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 show that removing random pieces of sequences have no impact on the estimation of population size variation. Taking this into account, removing problematic sections of sequences seems safer than to introduce sequences with SNP call errors or abnormal SNP distributions. However, removing pieces of the 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, have almost no impact, whereas removing many short sections of the sequences lead to a large overestimation of $\frac{\rho}{\theta}$. This result 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 $\frac{\rho}{\mu}$, which remain mainly unexplained.

4.2.1 | Suggestions when building a data set for inference

In order to make the most out of SMC methods, and thus obtain inferences that are as accurate as possible, we recommend simulating a data set corresponding to any suspected demographic scenarios and using, if possible, any available knowledge of the relevant genomic characteristics of the focal species (*e.g.* recombination and mutation rates). From these simulations, the estimated transition matrix for PSMC' or MSMC methods (for a single panmictic population) can be built. Migration could also be accounted for using the approach developed in Wang et al. (2020). The resulting estimated transition matrix and its coefficient of variation can therefore be used to guide users as to the necessary quantity of data needed (sequence size and number of genomes) for a good inference. Yet, it seems overall better to obtain genome assembly and SNP calls of enhanced quality (see Figure 5). Finally, in some cases, SMC-methods are limited by their own theoretical framework resulting in demographic scenarios which cannot be recovered. In such cases, other approaches can be considered (*e.g.* ABC) as they might perform better and are more flexible regarding the demographic scenarios which can be tested. (Beichman et al., 2017; Schraiber & Akey, 2015).

4.2.2 | Evaluating trustworthiness of results

As mentioned above, there are several instances where the past variation in population size may be badly inferred. A simple way to

determine whether the results obtained are indeed trustworthy is by examining the estimated transition matrix. If the matrix is empty in some places (*i.e.* there is 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 in this specific time interval. To measure the accuracy of the inferred demography, one can simulate data under the estimated demographic history and measure how well it is retrieved by the SMC-method Arredondo et al., (2020), Chikhi et al., (2018), Rodriguez et al., (2018).

5 | CONCLUDING REMARKS

Here, we present an extension of the classic SMC framework to help assess the accuracy of inferences when applying these methods to data sets with suspected flaws or limitations. This analysis can be conducted prior to the full analysis of a given data set. We also provide 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 (Gattepaille et al., 2016)). The estimated transition matrix is introduced as a summary statistic, which can be used to capture and recover pieces of the demographic history. This statistic could, in the future, be used in scenarios with migration, without the computational burden of Hidden Markov Models but keeping the high accuracy and resolution of SMC methods.

ACKNOWLEDGEMENTS

This work was funded by Deutsche Forschungsgemeinschaft, project number 317616126 (TE809/7-1) to AT. DAA was funded by the Alexander von Humboldt Stiftung. Version 3 of this preprint has been peer-reviewed and recommended by Peer Community In Evolutionary Biology (<https://doi.org/10.24072/pci.evolbio.100115>). We thank all reviewers for their comments, helping us improve the quality of this manuscript. Open Access funding enabled and organized by Projekt DEAL.


CONFLICT OF INTEREST

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

DATA AVAILABILITY STATEMENT

eSMC2 R package: <https://github.com/TPPSellinger/eSMC2>.

ORCID

Thibaut Paul Patrick Sellinger  <https://orcid.org/0000-0002-8538-7800>

Aurélien Tellier  <https://orcid.org/0000-0002-8895-0785>

REFERENCES

- Adrion, J. R. (2020). A community-maintained standard library of population genetic models. *eLife*, 9, https://doi.org/10.7554/eLife.54967. [Epub ahead of print].
- Arredondo, A., Mourato, B., Nguyen, K., Boitard, S., Rodríguez, W., Noël, C., Mazet, O., & Chikhi, L. (2020). Inferring Number of Populations and Changes in Connectivity under the n-island Model. https://doi.org/10.1101/2020.09.03.282251
- Barroso, G. V., Puzovic, N., & Dutheil, J. Y. (2019). Inference of recombination maps from a single pair of genomes and its application to ancient samples. *PLOS Genetics*, 15(11), e1008449. https://doi.org/10.1371/journal.pgen.1008449
- Beichman, A. C., Huerta-Sanchez, E., & Lohmueller, K. E. (2018). Using genomic data to infer historic population dynamics of nonmodel organisms. *Annual Review of Ecology, Evolution, and Systematics*, 49(1), 433–456. https://doi.org/10.1146/annurev-ecolsys-110617-062431
- Beichman, A. C., Phung, T. N., & Lohmueller, K. E. (2017). Comparison of single genome and allele frequency data reveals discordant demographic histories. *G3-Genes Genomes*. *Genetics*, 7(11), 3605–3620. https://doi.org/10.1534/g3.117.300259
- Bergström, A., McCarthy, S. A., Hui, R., Almarri, M. A., Ayub, Q., Danecek, P., Chen, Y., Felkel, S., Hallast, P., Kamm, J., Blanché, H., Deleuze, J.-F., Cann, H., Mallick, S., Reich, D., Sandhu, M. S., Skoglund, P., Scally, A., Xue, Y., ... Tyler-Smith, C. (2020). Insights into human genetic variation and population history from 929 diverse genomes. *Science*, 367(6484), eaay5012. https://doi.org/10.1126/science.aay5012
- Cao, J., Schneeberger, K., Ossowski, S., Günther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C., Wang, X. I., Ott, F., Müller, J., Alonso-Blanco, C., Borgwardt, K., Schmid, K. J., & Weigel, D. (2011). Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nature Genetics*, 43(10), 956–U60. https://doi.org/10.1038/ng.911
- Chang, D., & Shapiro, B. (2016). Using ancient DNA and coalescent-based methods to infer extinction. *Biology Letters*, 12(2), 20150822. https://doi.org/10.1098/rsbl.2015.0822
- Chikhi, L., Rodríguez, W., Grusea, S., Santos, P., Boitard, S., & Mazet, O. (2018). The IICR (inverse instantaneous coalescence rate) as a summary of genomic diversity: insights into demographic inference and model choice. *Heredity*, 120(1), 13–24. https://doi.org/10.1038/s41437-017-0005-6
- Choo, S. W., Rayko, M., Tan, T. K., Hari, R., Komissarov, A., Wee, W. Y., Yurchenko, A. A., Kliver, S., Tamazian, G., Antunes, A., Wilson, R. K., Warren, W. C., Koepfli, K.-P., Minx, P., Krashennikova, K., Kotze, A., Dalton, D. L., Vermaak, E., Paterson, I. C., ... Wong, G. J. (2016). Pangolin genomes and the evolution of mammalian scales and immunity. *Genome Research*, 26(10), 1312–1322. https://doi.org/10.1101/gr.203521.115
- Eklom, R., Brechlin, B., Persson, J., Smeds, L., Johansson, M., Magnusson, J., Flagstad, O., & Ellegren, H. (2018). Genome sequencing and conservation genomics in the Scandinavian wolverine population. *Conservation Biology*, 32(6), 1301–1312. https://doi.org/10.1111/cobi.13157
- Ewing, A. D. (2015). Transposable element detection from whole genome sequence data. *Mobile DNA*, 24(6). https://doi.org/10.1186/s13100-015-0055-3
- Fulgione, A., Koornneef, M., Roux, F., Hermisson, J., & Hancock, A. M. (2018). Madeiran *Arabidopsis thaliana* reveals ancient long-range colonization and clarifies demography in Eurasia. *Molecular Biology and Evolution*, 35(3), 564–574. https://doi.org/10.1093/molbev/msx300
- Gattepaille, L., Guenther, T., & Jakobsson, M. (2016). Inferring past effective population size from distributions of coalescent times. *Molecular Biology and Evolution*, 204(3), 1191. https://doi.org/10.1534/genetics.115.185058
- Hawks, J. (2017). Introgression makes waves in inferred histories of effective population size. *Human Biology*, 89(1), 67–80. https://doi.org/10.13110/humanbiology.89.1.04
- Hecht, L. B. B., Thompson, P. C., & Rosenthal, B. M. (2018). Comparative demography elucidates the longevity of parasitic and symbiotic relationships. *Proceedings of the Royal Society B: Biological Sciences*, 285(1888), 20181032. https://doi.org/10.1098/rspb.2018.1032
- Hendricks, S., Anderson, E. C., Antao, T., Bernatchez, L., Forester, B. R., Garner, B., Hand, B. K., Hohenlohe, P. A., Kardos, M., Koop, B., Sethuraman, A., Waples, R. S., & Luikart, G. (2018). Recent advances in conservation and population genomics data analysis. *Evolutionary Applications*, 11(8), 1197–1211. https://doi.org/10.1111/eva.12659
- Hudson, R. (1983). Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, 23(2), 183–201. https://doi.org/10.1016/0040-5809(83)90013-8
- Johndrow, J. E., & Palacios, J. A. (2019). Exact limits of inference in coalescent models. *Theoretical Population Biology*, 125, 75–93.
- Johri, P., Charlesworth, B., & Jensen, J. D. (2020). Toward an evolutionarily appropriate null model: Jointly inferring demography and purifying selection. *Genetics*, 215(1), 173–192. https://doi.org/10.1534/genetics.119.303002
- Johri, P., Riall, K., Becher, H., Excoffier, L., Charlesworth, B., & Jensen, J. D. (2021). The impact of purifying and background selection on the inference of population history: Problems and prospects. *Molecular Biology and Evolution*, 1537–1719. https://doi.org/10.1093/molbev/msab050
- Kelleher, J., Etheridge, A. M., & McVean, G. (2016). Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLOS Computational Biology*, 12(5), e1004842. https://doi.org/10.1371/journal.pcbi.1004842
- Kelleher, J., Wong, Y., Wohns, A. W., Fadil, C., Albers, P. K., & McVean, G. (2019). Inferring whole-genome histories in large population datasets (vol 51, pg 1330, 2019). *Nature Genetics*, 51(11), 1660. https://doi.org/10.1038/s41588-019-0523-7
- Kerdoncuff, E., Lambert, A., & Achaz, G. (2020). Testing for population decline using maximal linkage disequilibrium blocks. *Theoretical Population Biology*, 134, 171–181. https://doi.org/10.1016/j.tpb.2020.03.004
- Ki, C., & Terhorst, J. (2020). Exact decoding of the sequentially Markov coalescent. https://doi.org/10.1101/2020.09.21.307355. URL: http://biorxiv.org/lookup/doi/ https://doi.org/10.1101/2020.09.21.307355
- Kim, Y., Koehler, F., Moitra, A., Mossel, E., & Ramnarayan, G. (2020). How many sub populations is too many? exponential lower bounds for inferring population histories. *Journal of Computational Biology*, 27(4), 613–625. https://doi.org/10.1089/cmb.2019.0318
- Kofler, R. (2018). Simulate: Simulating complex landscapes of transposable elements of populations. *Bioinformatics*, 34(8), 1419–1420. https://doi.org/10.1093/bioinformatics/btx772
- Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357), 493–496. https://doi.org/10.1038/nature10231
- Lord, E., Dussex, N., Kierczak, M., Díez-del-Molino, D., Ryder, O. A., Stanton, D. W. G., Gilbert, M. T. P., Sánchez-Barreiro, F., Zhang, G., Sinding, M.-H., Lorenzen, E. D., Willerslev, E., Protopopov, A., Shidlovskiy, F., Fedorov, S., Bocherens, H., Nathan, S. K. S. S., Goossens, B., van der Plicht, J., ... Dalén, L. (2020). Pre-extinction Demographic Stability and Genomic Signatures of Adaptation in the Woolly Rhinoceros. *Current Biology*, 30(19), 3871–3879. https://doi.org/10.1016/j.cub.2020.07.046
- Lynch, M., Gutenkunst, R., Ackerman, M., Spitze, K., Ye, Z., Maruki, T., & Jia, Z. (2017). Population genomics of *Daphnia pulex*. *Molecular Biology and Evolution*, 206(1), 315–332. https://doi.org/10.1534/genetics.116.190611
- Malaspinas, A.-S., Westaway, M. C., Muller, C., Sousa, V. C., Lao, O., Alves, I., Bergström, A., Athanasiadis, G., Cheng, J. Y., Crawford, J.

- E., Heupink, T. H., Macholdt, E., Peischl, S., Rasmussen, S., Schiffels, S., Subramanian, S., Wright, J. L., Albrechtsen, A., Barbieri, C., ... Willerslev, E. (2016). A genomic history of aboriginal Australia. *Nature*, 538(7624), 207. <https://doi.org/10.1038/nature18299>
- Marjoram, P., & Wall, J. (2006). Fast "coalescent" simulation. *BMC Genetics*, 7, <https://doi.org/10.1186/1471-2156-7-16>
- Mather, N., Traves, S. M., & Ho, S. Y. W. (2020). A practical introduction to sequentially markovian coalescent methods for estimating demographic history from genomic data. *Ecology and Evolution*, 10(1), 579–589. <https://doi.org/10.1002/ece3.5888>
- Mazet, O., Rodriguez, W., Grusea, S., Boitard, S., & Chikhi, L. (2016). On the importance of being structured: instantaneous coalescence rates and human evolution-lessons for ancestral population size inference? *Heredity*, 116(4), 362–371. <https://doi.org/10.1038/hdy.2015.104>
- McVean, G., & Cardin, N. (2005). Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B-Biological Sciences*, 360(1459), 1387–1393. <https://doi.org/10.1098/rstb.20053.1673>
- Nadachowska-Brzyska, K., Burri, R., Smeds, L., & Ellegren, H. (2016). PSMC analysis of effective population sizes in molecular ecology and its application to black-and-white ficedula flycatchers. *Molecular Ecology*, 25(5), 1058–1072. <https://doi.org/10.1111/mec.13540>
- Nakagome, S., Hudson, R. R., & Di Rienzo, A. (2019). 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), 2019. <https://doi.org/10.1098/rspb.2018.2541>
- Nelson, M. G., Linheiro, R. S., & Bergman, C. M. (2017). McClintock: An integrated pipeline for detecting transposable element insertions in whole-genome shotgun sequencing data. *G3-Genes Genomes*, 7(8), 2763–2778. <https://doi.org/10.1534/g3.117.043893>
- Oaks, J. R., L'Bahy, N., & Cobb, K. A. (2020). Insights from a general, full-likelihood Bayesian approach to inferring shared evolutionary events from genomic data: Inferring shared demographic events is challenging. *Evolution*, 74(10), 2184–2206. <https://doi.org/10.1111/evo.14052>
- Oh, K. P., Aldridge, C. L., Forbey, J. S., Dadabay, C. Y., & Oyler-McCance, S. J. (2019). Conservation genomics in the sagebrush sea: Population divergence, demographic history, and local adaptation in sagegrouse (*Centrocercus* spp.). *Genome Biology and Evolution*, 11(7), 2023–2034. <https://doi.org/10.1093/gbe/evz112>
- Palamara, P. F., Terhorst, J., Song, Y. S., & Price, A. L. (2018). High-throughput inference of pairwise coalescence times identifies signals of selection and enriched disease heritability. *Nature Genetics*, 50(9), 1311. <https://doi.org/10.1038/s41588-018-0177-x>
- Palkopoulou, E., Lipsen, M., Mallyck, S., Nielsen, S., Rohland, N., Baleka, S., Karpinski, E., Ivancevic, A. M., To, T.-H., Kortschak, R. D., Raison, J. M., Qu, Z., Chin, T.-J., Alt, K. W., Claesson, S., Dalén, L., MacPhee, R. D. E., Meller, H., Roca, A. L., ... Reich, D. (2018). 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. <https://doi.org/10.1073/pnas.1720554115>
- Patton, A. H., Margres, M. J., Stahlke, A. R., Hendricks, S., Lewallen, K., Hamede, R. K., Ruiz-Aravena, M., Ryder, O., McCallum, H. I., Jones, M. E., Hohenlohe, P. A., & Storer, A. (2019). Contemporary demographic reconstruction methods are robust to genome assembly quality: A case study in Tasmanian devils. *Molecular Biology and Evolution*, 36(12), 2906–2921. <https://doi.org/10.1093/molbev/msz191>
- Peart, C. R., Tusso, S., Pophaly, S. D., Botero-Castro, F., Wu, C.-C., Auriolles-Gamboa, D., Baird, A. B., Bickham, J. W., Forcada, J., Galimberti, F., Gemmill, N. J., Hoffman, J. I., Kovacs, K. M., Kunnsaranta, M., Lydersen, C., Nyman, T., de Oliveira, L. R., Orr, A. J., Sanvito, S., ... Wolf, J. B. W. (2020). Determinants of genetic variation across eco-evolutionary scales in pinnipeds. *Nature Ecology & Evolution*, 4(8), 1095. <https://doi.org/10.1038/s41559-020-1215-5>
- Pfeifer, S. P. (2017). From next-generation resequencing reads to a high-quality variant data set. *Heredity*, 118(2), 111–124. <https://doi.org/10.1038/hdy.2016.102>
- Platt, R. N. II, Blanco-Berdugo, L., & Ray, D. A. (2016). Accurate transposable element annotation is vital when analyzing new genome assemblies. *Genome Biology and Evolution*, 8(2), 403–410. <https://doi.org/10.1093/gbe/evw009>
- Poelstra, J. W., Salmons, J., Tiley, G. P., Schübler, D., Blanco, M. B., Andriambeloso, J. B., Bouchez, O., Campbell, C. R., Etter, P. D., Hohenlohe, P. A., Hunnicutt, K. E., Iribar, A., Johnson, E. A., Kappeler, P. M., Larsen, P. A., Manzi, S., Ralison, J. É. M., Randrianambinina, B., Rasoloarison, R. M., ... Yoder, A. D. (2021). Cryptic patterns of speciation in cryptic primates: Microendemic mouse lemurs and the multispecies coalescent. *Systematic Biology*, 70(2), 203–218. <https://doi.org/10.1093/sysbio/syaa053>
- Prado-Martinez, J., Sudmant, P. H., Kidd, J. M., Li, H., Kelley, J. L., Lorente-Galdos, B., Veeramah, K. R., Woerner, A. E., O'Connor, T. D., Santpere, G., Cagan, A., Theunert, C., Casals, F., Laayouni, H., Munch, K., Hobolth, A., Halager, A. E., Malig, M., Hernandez-Rodriguez, J., ... Marques-Bonet, T. (2013). Great ape genetic diversity and population history. *Nature*, 499(7459), 471–475. <https://doi.org/10.1038/nature12228>
- Rodriguez, W., Mazet, O., Grusea, S., Arredondo, A., Corujo, J. M., Boitard, S., & Chikhi, L. (2018). The IICR and the non-stationary structured coalescent: towards demographic inference with arbitrary changes in population structure. *Heredity*, 121(6), 663–678. <https://doi.org/10.1038/s41437-018-0148-0>
- Sand, A., Kristiansen, M., Pedersen, C. N. S., & Mailund, T. (2013). Ziphmmlib: a highly optimised hmm library exploiting repetitions in the input to speed up the forward algorithm. *BMC Bioinformatics*, 14(1), <https://doi.org/10.1186/1471-2105-14-339>. [Epub ahead of print].
- Schiffels, S., & Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, 46(8), 919–925. <https://doi.org/10.1038/ng.3015>
- Schraiber, J. G., & Akey, J. M. (2015). Methods and models for unravelling human evolutionary history. *Nature Reviews Genetics*, 16(12), 727–740. <https://doi.org/10.1038/nrg4005>
- Schrider, D. R., Shanku, A. G., & Kern, A. D. (2016). Effects of linked selective sweeps on demographic inference and model selection. *Genetics*, 204(3), 1207. <https://doi.org/10.1534/genetics.116.190223>
- Sellinger, T. P. P., Abu Awad, D., Moest, M., & Tellier, A. (2020). Inference of past demography, dormancy and self-fertilization rates from whole genome sequence data. *PLOS Genetics*, 16(4). <https://doi.org/10.1371/journal.pgen.1004845>. [Epub ahead of print].
- Sheehan, S., & Song, Y. S. (2016). Deep Learning for Population Genetic Inference. *PLOS Computational Biology*, 12(3), e1004845. <https://doi.org/10.1371/journal.pcbi.1004845>
- Slatkin, M. (2016). Statistical methods for analyzing ancient dna from hominins. *Current Opinion in Genetics & Development*, 41, 72–76. <https://doi.org/10.1016/j.gde.2016.08.004>
- Speidel, L., Forest, M., Shi, S., & Myers, S. R. (2019). A method for genome-wide genealogy estimation for thousands of samples. *Nature Genetics*, 51(9), 1321. <https://doi.org/10.1038/s41588-019-0484-x>
- Spence, J. P., Steinrucken, M., Terhorst, J., & Song, Y. S. (2018). Inference of population history using coalescent hmms: review and outlook. *Current Opinion in Genetics & Development*, 53, 70–76. <https://doi.org/10.1016/j.gde.2018.07.002>
- Staab, P. R., Zhu, S., Metzler, D., & Lunter, G. (2015). scrm: Efficiently simulating long sequences using the approximated coalescent with recombination. *Bioinformatics*, 31(10), 1680–1682. <https://doi.org/10.1093/bioinformatics/btu861>
- Stam, R., Nosenko, T., Hoerger, A. C., Stephan, W., Seidel, M., Kuhn, J. M. M., Haberger, G., & Tellier, A. (2019). 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. <https://doi.org/10.1534/g3.119.400529>
- Steinrucken, M., Kamm, J., Spence, J. P., & Song, Y. S. (2019). 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. <https://doi.org/10.1073/pnas.1905060116>
- T. G. P. Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*.
- Terhorst, J., Kamm, J. A., & Song, Y. S. (2017). Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature Genetics*, 49(2), 303–309. <https://doi.org/10.1038/ng.3748>
- Wakeley, J. (2020). Developments in coalescent theory from single loci to chromosomes. *Theoretical Population Biology*, 133, 56–64. <https://doi.org/10.1016/j.tpb.2020.02.002>
- Wang, K., Mathieson, I., O'Connell, J., & Schiffels, S. (2020). Tracking human population structure through time from whole genome sequences. *PLOS Genetics*, 16(3), e1008552. <https://doi.org/10.1371/journal.pgen.1008552>
- Willemsen, D., Cui, R., Reichard, M., & Valenzano, D. R. (2020). Intra-species differences in population size shape life history and genome evolution. *eLife*, 9, <https://doi.org/10.7554/eLife.55794>. [Epub ahead of print].
- Williams, R. C., Blanco, M. B., Poelstra, J. W., Hunnicutt, K. E., Comeault, A. A., & Yoder, A. D. (2020). Conservation genomic analysis reveals ancient introgression and declining levels of genetic diversity in Madagascar's hibernating dwarf lemurs. *Heredity*, 124(1), 236–251. <https://doi.org/10.1038/s41437-019-0260-9>
- Wu, C., & Hein, J. (1999). Recombination as a point process along sequences. *Theoretical Population Biology*, 55(3), 248–259. <https://doi.org/10.1006/tpbi.1998.1403>
- Yew, C.-W., Lu, D., Deng, L., Wong, L.-P., Ong, R.-H., Lu, Y., Wang, X., Yunus, Y., Aghakhanian, F., Mokhtar, S. S., Hoque, M. Z., Voo, C.-Y., Abdul Rahman, T., Bhak, J., Phipps, M. E., Xu, S., Teo, Y.-Y., Kumar, S. V., & Hoh, B.-P. (2018). 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. <https://doi.org/10.1007/s00439-018-1869-0>

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Sellinger TP, Abu-Awad D, Tellier A. Limits and convergence properties of the sequentially Markovian coalescent. *Mol Ecol Resour*. 2021;21:2231–2248. <https://doi.org/10.1111/1755-0998.13416>