

A new method for estimating the demographic history from DNA sequences: an importance sampling approach

Sadoune Ait Kaci Azzou *, Fabrice Larribe and Sorana Froda

Département de Mathématiques, Équipe de Modélisation Stochastique Appliquée (EMOSTA), Université du Québec à Montréal, Montréal, QC, Canada

OPEN ACCESS

Edited by:

Miguel Arenas,
Institute of Molecular Pathology and
Immunology of the University of Porto
(IPATIMUP), Portugal

Reviewed by:

Gregory Bruce Ewing,
École Polytechnique Fédérale de
Lausanne, Switzerland
Stefano Mona,
EPHE (Ecole Pratique des Hautes
Etudes), France

*Correspondence:

Sadoune Ait Kaci Azzou,
Département de Mathématiques,
Équipe de Modélisation Stochastique
Appliquée (EMOSTA), Université du
Québec à Montréal, Case postale
8888, succursale Centre-Ville,
Montréal, QC H3C 3P8, Canada
ait_kaci_azzou.sadoune@uqam.ca

Specialty section:

This article was submitted to
Evolutionary and Population Genetics,
a section of the journal
Frontiers in Genetics

Received: 13 May 2015

Accepted: 20 July 2015

Published: 07 August 2015

Citation:

Ait Kaci Azzou S, Larribe F and Froda
S (2015) A new method for estimating
the demographic history from DNA
sequences: an importance sampling
approach. *Front. Genet.* 6:259.
doi: 10.3389/fgene.2015.00259

The effective population size over time (demographic history) can be retraced from a sample of contemporary DNA sequences. In this paper, we propose a novel methodology based on importance sampling (IS) for exploring such demographic histories. Our starting point is the *generalized skyline plot* with the main difference being that our procedure, *skywis plot*, uses a large number of genealogies. The information provided by these genealogies is combined according to the IS weights. Thus, we compute a weighted average of the effective population sizes on specific time intervals (epochs), where the genealogies that agree more with the data are given more weight. We illustrate by a simulation study that the *skywis plot* correctly reconstructs the recent demographic history under the scenarios most commonly considered in the literature. In particular, our method can capture a change point in the effective population size, and its overall performance is comparable with the one of the *bayesian skyline plot*. We also introduce the case of serially sampled sequences and illustrate that it is possible to improve the performance of the *skywis plot* in the case of an exponential expansion of the effective population size.

Keywords: importance sampling, effective population size, *skywis plot*, coalescent process, serially sampled sequences

1. Introduction

The demographic history of a population leaves its signature in the genome, which means that DNA sequences contain information about the demographic history of the population from which they are sampled. Therefore, it is possible to use genetic data to infer demographic parameters, an issue with important implications in many fields such as public health, epidemiology and conservation biology (Minin et al., 2008).

The first methods for estimating the demographic history from gene sequences were parametric and used coalescent theory. Such methods require a simple demographic model in order to describe the changes in the population size over time in terms of one or more parameters. They are based on importance sampling, e.g., (Slatkin and Hudson, 1991; Stephens and Donnelly, 2000), or Markov Chain Monte Carlo (MCMC) sampling, e.g., (Kuhner et al., 1995, 1998). For example, in the case of exponential growth, the size of the population at any time t measured from the present to the past is given by $N(t) = N(0) \exp(-\beta t)$, and the unknown parameters are $N(0)$ and β .

Usually, in practice, it is not known in advance which demographic model fits the sampled gene sequences. Further, population histories are often more complex than those described by simple parametric models. This has motivated the development of non-parametric and semi-parametric methods for inferring the demographic history from sequence data or from an estimated genealogy (e.g., Fu, 1994; Pybus et al., 2000) without resorting to some previous information about the demographic model.

Our method is nonparametric and is closely related to the family of *skyline plot* methods. The first method in this family was introduced by Pybus et al. (2000), and is referred to as the *classical skyline plot*. The *classical skyline plot* involves two separate steps, see (Ho and Shapiro, 2011): (1) estimating the genealogy from the sequence data and (2) estimating the population history from the estimated genealogy. Step 1 gives an estimated genealogy that includes the relationships among the individuals (tree topology) as well as their times of divergence. Genealogical estimation is done using standard phylogenetic methods under the so-called strict molecular clock. The strict molecular clock condition means that the branch lengths of the tree are proportional to time, with time being measured in mutations, and all lineages evolve at the same rate. It is also possible to estimate a genealogy in a relaxed-clock framework (Drummond et al., 2006). Further, in step 2 in order to estimate the population history from the estimated genealogy, Pybus et al. (2000) apply coalescent theory in a specific way by considering the times of divergence (node times) as coalescent times. When the true population size is constant, this assumption is equivalent to estimating the mean of an exponential distribution using a single realization from this distribution (Minin et al., 2008). This uncertainty is referred to as coalescent error. Further, the single phylogeny of the sequences is assumed to be known without error (i.e., phylogenetic error is assumed to be negligible).

Thus, Pybus et al. (2000) estimate the population size $\hat{N}e_k\mu$, for each coalescent interval $\gamma_k = \mu t_k$, by the product of $\binom{k}{2}$ and γ_k , where μ is the mutation rate per site per generation and γ_k is measured in substitutions. Thus, the *classical skyline plot* produces a piecewise reconstruction of the demographic history that is quite *noisy*, especially in the presence of small intervals when the sampled sequences are similar.

To improve the *classical skyline plot* estimation, several extensions have been proposed. Without being exhaustive, we discuss the extensions that are most relevant to our work.

Strimmer and Pybus (2001) developed a *generalized skyline plot* estimate based on the Akaike Information Criterion correction (AIC) in order to reduce the number of free parameters in the *classical skyline plot*. This method allows multiple coalescent events, i.e., for which little divergence time information is available, to be grouped together. Important developments were obtained in a Bayesian framework. Thus, Drummond et al. (2005) and Opgen-Rhein et al. (2005) use multiple change-point (MCP) models to estimate population size dynamics.

In particular, Drummond et al. (2005) use a Markov chain Monte Carlo (MCMC) sampling procedure that efficiently

samples a variant of the *generalized skyline plot*, given sequence data, and combines these plots in order to generate a posterior distribution of the effective population size through time. Due to the averaging effect of the MCMC sampling, the *Bayesian skyline plot* introduced by Drummond et al. (2005) produces smoother estimates than previous skyline plot methods. Also in the Bayesian framework, Minin et al. (2008) propose an alternative to change-point modeling that exploits Gaussian Markov random fields to achieve temporal smoothing of the effective population size. The advantage of the *skyride* method is that in contrast to estimates given by MCP models, explicit temporal smoothing does not require strong prior decisions like fixing the total number of change points *a priori*.

Finally, Heled and Drummond (2008) introduced the extended Bayesian skyline plot, which permits the analysis of multiple unlinked loci. Increasing the number of independent loci allows the uncertainty in the coalescent to be assessed, leading to an improvement in the reliability of the demographic inference and a substantial reduction in estimation error (Ho and Shapiro, 2011). Further, unlike previous *skyline plot* methods that use a piecewise-constant model, the extended Bayesian skyline plot permits the use of a piecewise-linear model to describe the demographic history, allowing the population size to change continuously along each interval.

In order to estimate the effective population size, we propose a new method in a likelihood-based perspective. Unlike some skyline methods that use a single estimated phylogeny of the sequences, or others that use MCMC approaches, we resort to an efficient importance sampling scheme and our estimate comes to an weighted average over a large number of simulated genealogies, each with a different set of coalescence times. The methodology is described in detail in Section The Skywis Method.

2. Background

2.1. Coalescent Theory

In this section, we present the basic ideas behind the standard coalescent, as well as its extension to the case of fluctuating population size. An introduction to coalescent theory can be found in Nordborg (2003). Coalescent theory allows one to produce genealogies relating the sampled sequences according to a large class of population genetic models. In particular, the classical coalescent process assumes a single, isolated and panmictic population (e.g., a Wright-Fisher model), which evolves with constant (haploid) size N over many generations. For sufficiently large N and a sample size n such that $n \ll N$, the ancestral relationships between the gene sequences can be approximated by Kingman's coalescent (Kingman, 1982).

In short, the ancestry of a sample of sequences is modeled back in time, starting from the current sample and until the most recent common ancestor (MRCA) of the sample is found. At each step in the genealogical tree, one of the following events can occur: (1) two sequences coalesce if they share a common ancestor; (2) one sequence mutates. In the coalescent framework, time is measured in units of N generations, and N is large. The mutation rate μ per sequence per generation is rescaled so that

$\theta = 2N\mu$. Further, one can consider that each pair of lineages coalesces independently as a Poisson process with rate 1, and so, when there are k ancestral lines, coalescent events occur as in a Poisson process with total rate $k(k - 1)/2$ (Stephens, 2000).

In the classical coalescent process, and in the presence of k gene sequences, the waiting time T_k to the next coalescent event is exponentially distributed with rate $\binom{k}{2}$, while the distribution of the time until the first mutation event in any of the k lineages is exponential with parameter $k\theta/2$. Since mutations are assumed to occur independently of coalescence, the waiting time until a mutation or coalescent event is exponentially distributed with parameter

$$\binom{k}{2} + \frac{k\theta}{2} = \frac{k(k - 1 + \theta)}{2}. \tag{1}$$

The classical coalescent framework can be extended to include simple deviations from the idealized Wright-Fisher model, like recombination, fluctuating population size, population structure, and selection. In our paper, we focus on a single extension of the coalescent, namely variable population size.

In the case of non-constant population size, the number of descendants of a sequence in one generation does not follow the Poisson distribution with intensity one (Hein et al., 2005). As a result, when the basic coalescent is used to model a real physical population, the size N of the population in the (haploid) Wright-Fisher model cannot be assumed to be equal to the size of the real population.

Let $N_e(t)$ denote the effective population size at time t with $N_e(0) = N$. The effective population size reflects the number of individuals that contribute offsprings to the descendant generation and is almost always smaller than the census population size. The variable population size coalescent model for contemporary gene sequences was introduced by Griffiths and Tavaré (1994c) and Donnelly and Tavaré (1995). In this case, the coalescence times T_2, T_3, \dots, T_n do not follow independent exponential distributions.

Let $V_k = T_n + \dots + T_k$ be the accumulated waiting time so that the number of sequences pass from n to $k - 1$ sequences, i.e.,

$$V_k = \sum_{\ell=k}^n T_\ell, \tag{2}$$

and let $\Lambda(t)$ the cumulative coalescent rate over time measured relative to the rate at time $t = 0$:

$$\Lambda(t) = \int_0^t \frac{1}{v(u)} du, \tag{3}$$

where $v(t) = N_e(t)/N$, the relative size of $N_e(t)$ to N .

The waiting time until the next event depends only on the time of the previous event by the Markov property. The survival function of the time T_k conditional on $V_{k+1} = v$ is

$$P(T_k > t | V_{k+1} = v) = \exp \left\{ - \binom{k}{2} (\Lambda(t + v) - \Lambda(v)) \right\}, \tag{4}$$

where $v_{n+1} = 0$.

We note that when replacing $\Lambda(t)$ by t (i.e., in the case $N_e(t) = N, t > 0$) in Equation (4), we get the survival function of the exponential distribution. From Equation (4), we obtain the density

$$f_{T_k | V_{k+1}}(t_k | v) = \frac{\binom{k}{2}}{N_e(t_k + v)} \exp \left[- \int_v^{t_k + v} \frac{\binom{k}{2}}{N_e(x)} dx \right]. \tag{5}$$

It is precisely from this equation that Pybus et al. (2000) derived the estimation of the effective population size $\hat{N}e_k$ in the presence of k sequences.

2.2. Importance Sampling

Parameter estimation in population genetic models require optimization of the likelihood of the data given the parameters, $P(\mathcal{D}|\theta)$. The likelihood is then evaluated by:

$$L(\theta) = \int_{\mathcal{G}} P(\mathcal{D}|\mathcal{G}, \theta) P(\mathcal{G}|\theta) d\mathcal{G}, \tag{6}$$

where θ is the collection of parameters (such as population size and migration rates) for the population process. Typically, the objective of the analysis is to estimate these parameters by averaging the likelihood over all possible genealogies. A naive Monte Carlo method for the integral in Equation (6) is given by

$$L(\theta) \approx \frac{1}{J} \sum_{j=1}^J P(\mathcal{D}|\mathcal{G}^{(j)}, \theta), \tag{7}$$

where $\mathcal{G}^{(1)}, \mathcal{G}^{(2)}, \dots, \mathcal{G}^{(J)}$ are an independent sample from $P(\mathcal{G}|\theta)$.

Importance Sampling (IS) allows us to improve the efficiency of the Monte Carlo integration. The main idea of the IS approach is to reduce the inefficiency of the approximation (Equation 7) by concentrating the simulation on the trees that are more likely with the observed data. Instead of choosing histories from the distribution $P_\theta(\mathcal{G})$, we want to sample genealogies from a proposal distribution $Q(\mathcal{G})$ that better supports the observed data, \mathcal{D} . The IS method is based on rewriting (Equation 6) as

$$\int_{\mathcal{G}} P(\mathcal{D}|\mathcal{G}, \theta) \frac{P(\mathcal{G}|\theta)}{Q(\mathcal{G})} Q(\mathcal{G}) d\mathcal{G}. \tag{8}$$

The Monte Carlo approximation of Equation (8) gives

$$L(\theta) \approx \frac{1}{J} \sum_{j=1}^J P(\mathcal{D}|\mathcal{G}^{(j)}, \theta) \frac{P(\mathcal{G}^{(j)}|\theta)}{Q(\mathcal{G}^{(j)})}, \tag{9}$$

where $\mathcal{G}^{(1)}, \mathcal{G}^{(2)}, \dots, \mathcal{G}^{(J)} \sim Q(\mathcal{G})$. Good choices of the distribution $Q(\cdot)$ make this method of approximation much more efficient than (Equation 7). Ideally, we would like to sample from $Q(\mathcal{G}) = P(\mathcal{G}|\mathcal{D})$. However, this is impossible because it supposes perfect knowledge of the likelihood which is not true in practice.

Importance sampling (IS) was first used in this context by Griffiths and Tavaré (1994a,b,c). Stephens and Donnelly (2000) proposed improvements to the method by suggesting an approximation to an optimal proposal distribution for IS, $P(\mathcal{G}|\mathcal{D})$.

3. The Skywis Method

In this section, we describe our estimation method of the effective population size, when n gene sequences are available. The main idea behind our method is to simulate a large number of genealogies and create a weighted average of the effective population sizes, where the most probable genealogies are given larger weight. In short, reconstructing the demographic history from these sequences involves four distinct steps:

1. simulate J genealogies: $\mathcal{G}^{(1)}, \mathcal{G}^{(2)}, \dots, \mathcal{G}^{(J)}$;
2. compute $\widehat{N}e_k^{(1)}, \widehat{N}e_k^{(2)}, \dots, \widehat{N}e_k^{(J)}$ where $\widehat{N}e_k^{(j)}, k = 2, 3, \dots, n$, represents the estimated effective population size for the genealogy $\mathcal{G}^{(j)}$ for each coalescent time $t_k^{(j)}$ (in the presence of k sequences);
3. compute the weights $w^{(1)}, w^{(2)}, \dots, w^{(J)}$, where $w^{(j)}$ represents the weight of the genealogy $\mathcal{G}^{(j)}$ in the likelihood of the data;
4. estimate $\widehat{N}e_k$ based on genealogies $\mathcal{G}^{(1)}, \mathcal{G}^{(2)}, \dots, \mathcal{G}^{(J)}$, by the weighted mean of $\widehat{N}e_k^{(j)}$, for $j = 1, 2, \dots, J$, and $k = 2, 3, \dots, n$, i.e.,

$$\widehat{N}e_k = \sum_{j=1}^J w^{(j)} \widehat{N}e_k^{(j)}. \tag{10}$$

For example, with a variable population size that is expanding from the past to the present, as we progress toward the MRCA one can expect the population size to be smaller, or coalescence times to be shorter, than in the case of a constant population size. This fact, of shorter coalescence times, should be reflected

more faithfully by the most probable genealogies. Since such genealogies receive the largest weights, one can see that through the weighting system the estimator is adapting itself to the information contained in the data.

In what follows we describe our method in full detail, namely:

- how to simulate genealogies;
- how to set the weights;
- how to estimate the effective population size.

3.1. Skywis Plot for Homochronous Sampling

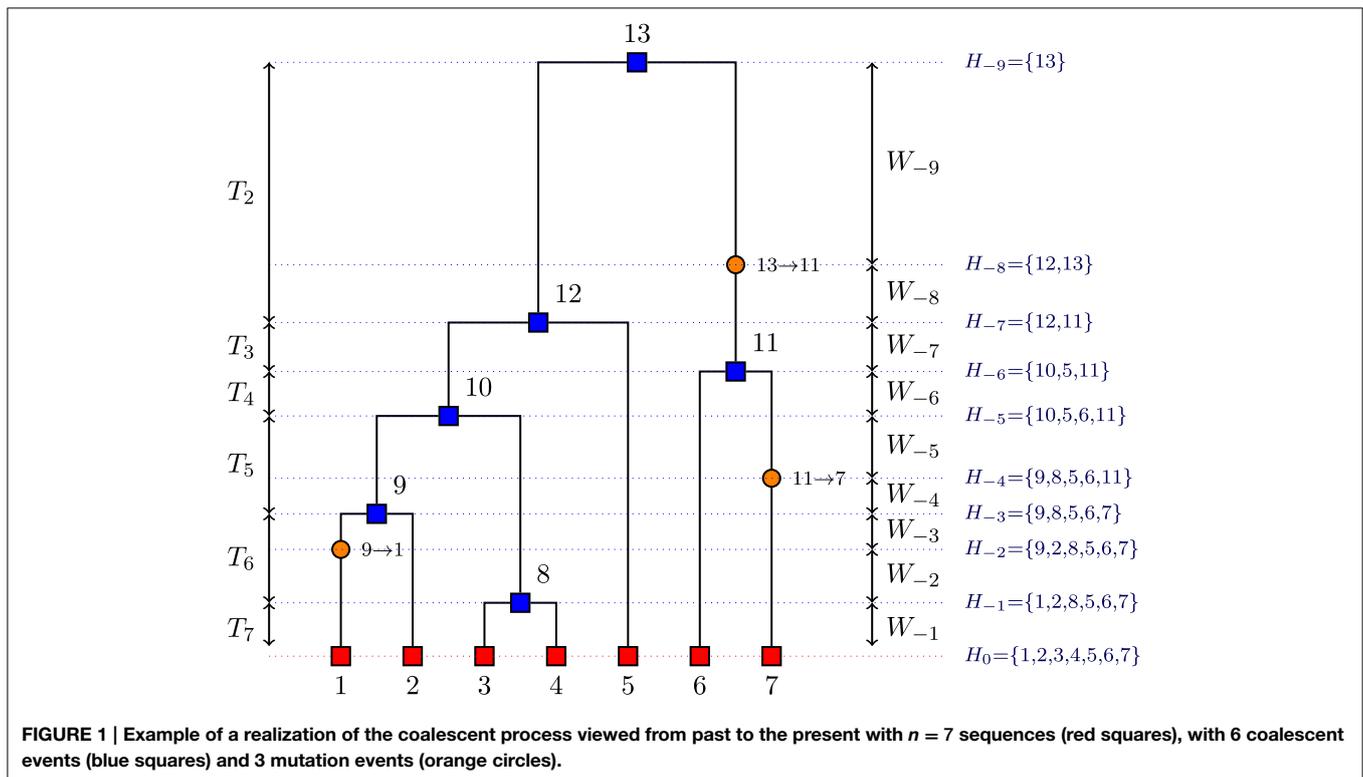
3.1.1. Simulation of Genealogies

In order to generate genealogies we use the proposal distribution $Q(\cdot)$ introduced by Stephens and Donnelly (2000) assuming a constant population size and a finite sites model with known mutation parameters. Given the Stephens and Donnelly (2000) method is crucial to our approach, we describe it briefly.

Let:

- E : the set of possible types of gene sequences;
- H_{-i} : the set of all sequences when event i occurs (coalescence or mutation) where i decreases from the present to the past in steps of 1 for each event (see **Figure 1**);
- $\mathcal{H} = \{H_0, H_{-1}, \dots, H_{-m}\}$: a history of sequences where $H_0 = \mathcal{D}$, m is the total number of events in the history \mathcal{H} , and H_{-m} is a singleton (the MRCA);
- \mathbf{P} : the mutation transition matrix;

In the Stephens and Donnelly (2000) method, the H_{-i} are viewed as states of a Markov process starting at genetic type $H_{-m} \in E$



and ending with $H_0 \in E$. Let \mathbf{P} be the mutation transition matrix. Let $P_{\alpha\beta}$ be the probability of a DNA sequence of type α to mutate to a DNA sequence of type β , and let M_{α}^{β} denote a mutation of a gene sequence from type α to type β according to \mathbf{P} ; let C_{α}^{α} denote a coalescence of two gene sequences of type α . Then, the forward transition probabilities $p_{\theta}(H_i|H_{i-1})$, are defined by Equation (11):

$$p_{\theta}(H_i|H_{i-1}) = \begin{cases} \frac{n_{i-1}^{(\alpha)}}{n_{i-1}} \frac{\theta}{(n_{i-1} - 1 + \theta)} P_{\alpha\beta} & \text{if } M_{\alpha}^{\beta}, \\ \frac{n_{i-1}^{(\alpha)}}{n_{i-1}} \frac{n_{i-1} - 1}{n_{i-1} - 1 + \theta} & \text{if } C_{\alpha}^{\alpha}, \\ 0 & \text{otherwise,} \end{cases} \tag{11}$$

where $n_{i-1}^{(\alpha)}$ is the number of sequences of type α in H_{i-1} , n_{i-1} is the number of sequences in H_{i-1} .

Stephens and Donnelly (2000) consider randomly constructing histories backward in time in a Markov way, from the sample H_0 to an MRCA (single type), according to some backward transition probabilities $q_{\theta}(H_{i-1}|H_i)$ in the class $\mathcal{M} = \{H_{i-1}|P_{\theta}(H_i|H_{i-1}) > 0\}$ with the constraint $q_{\theta}(H_{i-1}|H_i) \propto p_{\theta}(H_i|H_{i-1})$. Their proposed backward transition probabilities $\tilde{q}_{\theta}(H_{i-1}|H_i)$ which define $Q(\cdot)$ are given by Equation (12), namely:

$$\tilde{q}_{\theta}(H_{i-1}|H_i) = \begin{cases} C^{-1} \frac{\theta}{2} \cdot n_i^{(\alpha)} \cdot \frac{\hat{\pi}(\beta|H_i - \alpha)}{\hat{\pi}(\alpha|H_i - \alpha)} \cdot P_{\alpha\beta} & \text{if } M_{\alpha}^{\beta}, \\ C^{-1} \binom{n_i^{(\alpha)}}{2} \cdot \frac{1}{\hat{\pi}(\alpha|H_i - \alpha)} & \text{if } C_{\alpha}^{\alpha}, \\ 0 & \text{otherwise,} \end{cases} \tag{12}$$

where $n_i^{(\alpha)}$ is the number of sequences of type α in H_i , n_i is the number of sequences in H_i , $\{H_i - \alpha\}$ is the set of all sequences in H_i without the chosen sequence α , and $C = n_i(n_i - 1 + \theta)/2$ is a constant of proportionality. The estimated conditional probability $\hat{\pi}(\alpha|H_i)$ is described below.

In the proposed reconstruction, when H_i contains n_i chromosomes, the new type α is obtained by choosing a chromosome from H_i at random and then mutating it a geometric number of times. If $n_i^{(\beta)}$ is the number of chromosomes of type β in H_i , then (Stephens and Donnelly, 2000),

$$\hat{\pi}(\alpha|H_i) = \sum_{\beta \in E} \sum_{m=0}^{\infty} \frac{n_i^{(\beta)}}{n_i} \left(\frac{\theta}{n_i + \theta} \right)^m \frac{n_i}{n_i + \theta} (\mathbf{P}^m)_{\alpha\beta}. \tag{13}$$

In our approach, the genealogies are simulated backwards in time by the following algorithm based on Equation (12):

1. initialize $n_i := n$, where n is the number of DNA sequences at time $t = 0$ (present), and $i = 0$;
2. simulate the time to the next event, W_{-i-1} , as an exponential distribution with parameter $\binom{n_i}{2} + \frac{n_i\theta}{2}$;
3. randomly choose a sequence from H_i ; the chosen sequence type is denoted α ;
4. for each type $\beta \in E$ for which $P_{\alpha\beta} > 0$, compute $\hat{\pi}(\beta|H_i - \alpha)$;
5. compute the quantities x_1 and x_2 , where

$$x_1 = \theta \hat{\pi}(\beta|H_i - \alpha) P_{\beta\alpha} \quad \text{and} \quad x_2 = n_i^{(\alpha)} - 1.$$

Then, choose:

- a coalescence event with probability $\frac{x_2}{(x_1 + x_2)}$;
 - a mutation event (to β) with probability $\frac{x_1}{(x_1 + x_2)}$.
6. depending on the type of event chosen in step 5, we continue as follows:
 - if there is a coalescence event, choose another sequence of type α randomly, and let $n_{i-1} := n_i - 1$;
 - if there is a mutation event, mutate the sequence α into a sequence β , without changing n_i , i.e., let $n_{i-1} := n_i$;
 7. let $i := i - 1$ and continue until $n_i = 1$.

After implementing the above algorithm, the coalescence times that are at the core of our method can be deduced. In the genealogy \mathcal{G} given in **Figure 1**, we can deduce the coalescent times from the event times. For example, $T_7 = W_{-1}$ whereas $T_6 = W_{-2} + W_{-3}$ because we have a mutation event before a coalescence event.

3.1.2. Weights of Genealogies

After generating genealogies using the Stephens and Donnelly (2000) proposal distribution, it is possible to compute the importance weight $w^{(j)}$ for each genealogy $\mathcal{G}^{(j)}$, with $j = 1, 2, \dots, J$. Then $w^{(j)}$ is given by:

$$w^{(j)} = \frac{W^{(j)}}{\sum_{j=1}^J W^{(j)}}, \tag{14}$$

where

$$W^{(j)} = P(\mathcal{D}|\mathcal{G}^{(j)}, \theta) \frac{P(\mathcal{G}^{(j)}|\theta)}{Q(\mathcal{G}^{(j)})}, \tag{15}$$

with

$$Q(\mathcal{G}^{(j)}) = \prod_{i=0}^{-m} \tilde{q}_{\theta}(H_{i-1}|H_i), \tag{16}$$

and

$$P(\mathcal{G}^{(j)}|\theta) = \prod_{i=0}^{-m} p_{\theta}(H_i|H_{i-1}). \tag{17}$$

3.1.3. Estimation of the Effective Population Size

When building genealogies backwards in time, as we move backwards in time, fewer coalescence events occur. As a result, coalescence times close to the present are very short and become larger gradually going back in time. These short coalescence times create an undesirable variability in the estimation of the effective population size. Therefore, we propose to cumulate small coalescence times in order to improve the estimation of the effective population size. These cumulated time intervals are called epochs. To define epochs that get larger as we go backwards in time, we followed (Durbin and Li, 2011), and used a special time scale based on the TMRCA. Forest (2014) adopted the same method.

Finally, we note that the idea of cumulating small coalescence times in order to smooth the graph of the estimator of the effective population size was first proposed by Strimmer and Pybus (2001); it has since become quite standard in the related literature.

Once the genealogies have been simulated using the method described in Section 3.1.1, we cumulate the coalescence times as follows:

- we fix the total number of epochs, n_{cum} , i.e., the total number of time intervals where we estimate the effective population size;
- for each simulated genealogy $\mathcal{G}^{(j)}$, we compute the MRCA time, $T_{\text{MRCA}}^{(j)}$;
- we use formula (Equation 18) proposed by Durbin and Li (2011) in order to define epochs where estimates of the effective population size are computed. In other words, the following time cutting points in a genealogy $\mathcal{G}^{(j)}$, $j = 1, 2, \dots, J$ are used:

$$t_{\text{cut},b}^{(j)} = 0.1 \cdot \exp\left(\frac{b}{n_{\text{cum}}} \cdot \log(1 + 10 \cdot T_{\text{MRCA}}^{(j)})\right) - 0.1, \quad (18)$$

$$b = 1, 2, \dots, n_{\text{cum}},$$

where $t_{\text{cut},n_{\text{cum}}}^{(j)} = T_{\text{MRCA}}^{(j)}$.

For each genealogy, formula (Equation 18) gives the boundaries of the epochs, measured from the present to the past where $b = 0, 1, 2, \dots, n_{\text{cum}}$ (in units of N generations). The boundaries of epochs are different for each genealogy $\mathcal{G}^{(j)}$ and depend on the length of the tree. For example if $T_{\text{MRCA}}^{(1)} = 1$ in units of N generations and $n_{\text{cum}} = 5$, then according to Equation (18), the boundaries of the intervals are 0.0615, 0.1609, 0.3215, 0.581 (backward in time). For example, for the first epoch, this means that we must cumulate coalescence times from T_n until reaching 0.0615 N generations.

The *skyline plot* can be viewed as a method of moments estimator based on the standard coalescence distributions (Strimmer and Pybus, 2001). For a genealogy $\mathcal{G}^{(j)}$, we have:

$$E\left(T_k^{(j)} \cdot \binom{k}{2}\right) = N \text{ (generations)}, \quad (19)$$

because $T_k^{(j)}$ is exponentially distributed as $\exp\left(-\frac{k}{2}\right)$. Therefore, we use the estimate

$$\widehat{N}e_k^{(j)} \approx t_k^{(j)} \binom{k}{2}, \quad j = 1, 2, \dots, J. \quad (20)$$

The expectation of the accumulated waiting time in order to pass from n to ℓ lineages, $T_{n \rightarrow \ell}^{(j)} = \sum_{k=\ell}^n T_k^{(j)}$, is given by (see, for example, Rodrigo et al., 1999)

$$E\left(T_{n \rightarrow \ell}^{(j)}\right) = \frac{2c}{n(n-c)} N \text{ (generations)}, \quad (21)$$

where $c = n - \ell$ represents the number of coalesced sequences. From Equation (21), we can see that we can estimate, using the method of moments, the effective population size for the cumulated time of c coalescence times by:

$$t_{n \rightarrow \ell}^{(j)} \cdot \frac{n(n-c)}{2c}, \quad (22)$$

where $t_{n \rightarrow \ell}^{(j)} = \sum_{k=\ell}^n t_k^{(j)}$, and $c = n - \ell$. In our case, the cumulated waiting times for each genealogy $\mathcal{G}^{(j)}$ are deduced from Equation (18): once the boundaries of the intervals of epochs are computed, the cumulated waiting times, $\Delta t_b^{(j)}$ numbered from present to the past, are derived as:

$$\Delta t_b^{(j)} = t_{\text{cut},b}^{(j)} - t_{\text{cut},b-1}^{(j)}, \quad (23)$$

where $b = 1, 2, \dots, n_{\text{cum}}$; $j = 1, 2, \dots, J$, and $t_{\text{cut},0}^{(j)} = 0$. It follows from Equations (22, 18) that the estimated effective population size for an epoch b , and genealogy $\mathcal{G}^{(j)}$, $j = 1, 2, \dots, J$, is given by:

$$\widehat{N}e_b^{(j)} = \Delta t_b^{(j)} \cdot \frac{d_b^{(j)} \left(d_b^{(j)} - c_b^{(j)} \right)}{2c_b^{(j)}}, \quad (24)$$

where $d_b^{(j)}$ is the number of sequences at the beginning of the $\Delta t_b^{(j)}$ interval, and $c_b^{(j)}$ is the number of cumulated coalescence times in the epoch $\Delta t_b^{(j)}$, $b = 1, 2, \dots, n_{\text{cum}}$, $j = 1, 2, \dots, J$.

The distribution of importance weights of genealogies described by the Equation (15) is an approximation to the posterior distribution $P(\mathcal{G}|\mathcal{D}, \theta)$. As a result, one can approximate quantities of interest related to the tree by forming a weighted average of these quantities over the sampled trees as suggested in Stephens (2001).

In our case, we are interested in the estimation of $E(Ne_b)$, $b = 1, 2, \dots, n_{\text{cum}}$ from the J estimates $\widehat{N}e_b^{(j)}$, $j = 1, 2, \dots, J$ and we let

$$E(Ne_b) \approx \sum_{j=1}^J w^{(j)} \widehat{N}e_b^{(j)}. \quad (25)$$

In our algorithm, the weighted average of $\widehat{N}e_b^{(j)}$ is computed for the same time interval for all $j = 1, 2, \dots, J$ that represent the intersections of epochs for the J simulated genealogies. This way of proceeding gives us weighted estimates of effective population sizes under the assumption that the effective population size is

constant for an epoch. The reason for taking common intervals across genealogies is that $\widehat{N}e_k^{(j)}$ estimates the integral (see Pybus et al., 2000)

$$\left(\frac{1}{t_k} \int_{v_{k+1}^{(j)}}^{t_k^{(j)}+v_{k+1}^{(j)}} \frac{dx}{N_e(x)} \right)^{-1}, j = 1, 2, \dots, J. \quad (26)$$

Therefore, to estimate the integral Equation (26) by a weighted average of estimates from J genealogies, we must use the same time intervals.

For illustration, in **Figure 2**, we assume that two genealogies $\mathcal{G}^{(1)}$ and $\mathcal{G}^{(2)}$ are simulated using the method described in Section 3.1.1 with respective weights $w^{(1)}$ and $w^{(2)}$. Further, we assume that we cumulate coalescence times to obtain $n_{cum} = 3$ epochs. The limits of epochs for a genealogy $\mathcal{G}^{(j)}$ are denoted $t_{cut,b}^{(j)}$, $b = 1, 2$, and the time to the MRCA by $TMRCA^{(j)}$, $j = 1, 2$. The detailed calculation of the weighted effective population size per epochs is summarized in the following table:

Time interval	$\widehat{N}e_\ell, \ell = 1, 2, \dots, 2 \cdot n_{cum}$
$[0; t_{cut,1}^{(2)})$	$w^{(1)}\widehat{N}e_1^{(1)} + w^{(2)}\widehat{N}e_1^{(2)}$
$[t_{cut,1}^{(2)}; t_{cut,1}^{(1)})$	$w^{(1)}\widehat{N}e_1^{(1)} + w^{(2)}\widehat{N}e_2^{(2)}$
$[t_{cut,1}^{(1)}; t_{cut,2}^{(2)})$	$w^{(1)}\widehat{N}e_2^{(1)} + w^{(2)}\widehat{N}e_2^{(2)}$
$[t_{cut,2}^{(2)}; t_{cut,2}^{(1)})$	$w^{(1)}\widehat{N}e_2^{(1)} + w^{(2)}\widehat{N}e_3^{(2)}$
$[t_{cut,2}^{(1)}; TMRCA^{(2)})$	$w^{(1)}\widehat{N}e_3^{(1)} + w^{(2)}\widehat{N}e_3^{(2)}$
$[TMRCA^{(2)}; TMRCA^{(1)})$	$\widehat{N}e_3^{(1)}$

3.2. Skywis for Heterochronous Sampling

The algorithm described in Section 3.1 can be generalized to the case of serially sampled sequences i.e., sequences sampled at different moments in time. Such samples are also called heterochronous. **Figure 3** illustrates a case where we sampled sequences at times $t_0 < t_1 < t_2$, and the time is measured from the present to the past. Let S be the number of instants where we sampled sequences ($S = 3$ in **Figure 3**). Rodrigo and Felsenstein (1999) extend the coalescent likelihood for such heterochronous sequences, a very important issue in the case of rapidly evolving viruses such as HIV. For example, Rodrigo et al. (1999) have estimated, using heterochronous sequences, the viral generation time of HIV type1 (HIV-1). Also, serially sampling rapidly evolving populations is used for dating evolutionary events and divergence times (see e.g., Drummond et al., 2003).

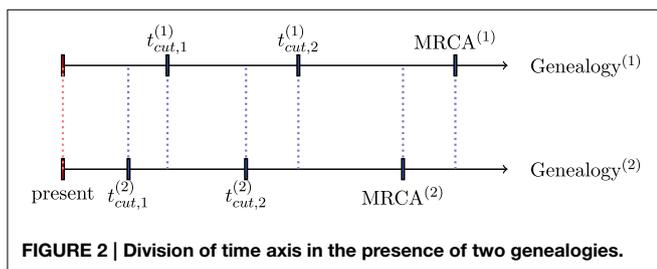


FIGURE 2 | Division of time axis in the presence of two genealogies.

In the presence of serially sampled sequences, we have to adapt the method of Stephens and Donnelly (2000) in order to simulate genealogies in this case. This necessarily involves developing new formulas for the probabilities $p_\theta(H_i|H_{i-1})$ and $\tilde{q}_\theta(H_{i-1}|H_i)$, as given below.

3.2.1. Backward and Forward Probabilities, and Weights of Genealogies

Let $n^{(s)}$ be the number of additional sampled sequences at time t_s , with $s = 1, 2, \dots, S - 1$. The main difference between the algorithm for homochronous sequences presented in Section 3.1, and the new algorithm for heterochronous sequences is that the number of lineages increases at the (known) instants t_s , $s = 1, 2, \dots, S - 1$ where samples of sequences are added. Further, it is necessary to use event times, because the embedded chain differs according to the relative position of these event times with respect to t_s , $s = 0, 1, 2, \dots, S - 1$.

In other words, the probabilities $p_\theta(H_i|H_{i-1})$ and $\tilde{q}_\theta(H_{i-1}|H_i)$ are calculated differently from the case of a single sample of sequences, which has an impact on how the weights of genealogies, $w^{(j)}$, $j = 1, 2, \dots, J$, are computed.

In order to present our results, we introduce these additional notations:

- $D_{i,v} = \{H_i, v\}$: represents the set of all sequences present in the population after the i^{th} event at time v ; this is a generalization of H_i with the specification of the time of event i ;
- \mathcal{E}_s : represents the set of all sequences added at time t_s ;

Our proposal distribution is an adapted version of the Stephens and Donnelly (2000) method for simulating genealogies, to the case of heterochronous sequences. In this case, as mentioned above, we consider that there is a list of pre-specified sampling times t_s , $s = 0, 1, 2, \dots, S - 1$ which are dividing the time axis. In what follows, time is measured from the present to the past and by *event* we mean either a coalescence or a mutation. If an event time v is such that $t_{s-1} < v < t_s$ and the time v' of the next event is such that $v' > t_s$, v' is truncated at t_s , i.e., $v' \leq t_s$. Then, either there is a next event at time $v' \leq t_s$ or the time is truncated at t_s , new sequences are added, and the process starts anew. Thus, from $D_{i,v}$ one can move to either $D_{i-1,v'} = \{H_{i-1}, v'\}$, $v < v' \leq t_s$, where H_{i-1} is obtained from H_i by a coalescence or a mutation, or to D_{i-1,t_s} where $H_{i-1} = H_i + \mathcal{E}_s$. In this last case we add \mathcal{E}_s sequences at time t_s and the process starts anew, with a new set of sequences that includes those at v . The moves of the process (embedded chain) are given by the following formulas, and we consider separately the case $v' < t_s$ and the case $v' = t_s$.

Case 1: $t_{s-1} < v < v' < t_s$.

$$\tilde{q}_\theta(D_{i-1,v'}|D_{i,v}) = \begin{cases} C^{-1} \frac{\theta}{2} n_i^{(\alpha)} \frac{\hat{\pi}(\beta|H_i - \alpha)}{\hat{\pi}(\alpha|H_i - \alpha)} P_{\beta\alpha} & \text{if } M_\alpha^\beta \\ C^{-1} \binom{n_i^{(\alpha)}}{2} \frac{1}{\hat{\pi}(\alpha|H_i - \alpha)} & \text{if } C_\alpha^\alpha \\ 0 & \text{otherwise,} \end{cases} \quad (27)$$

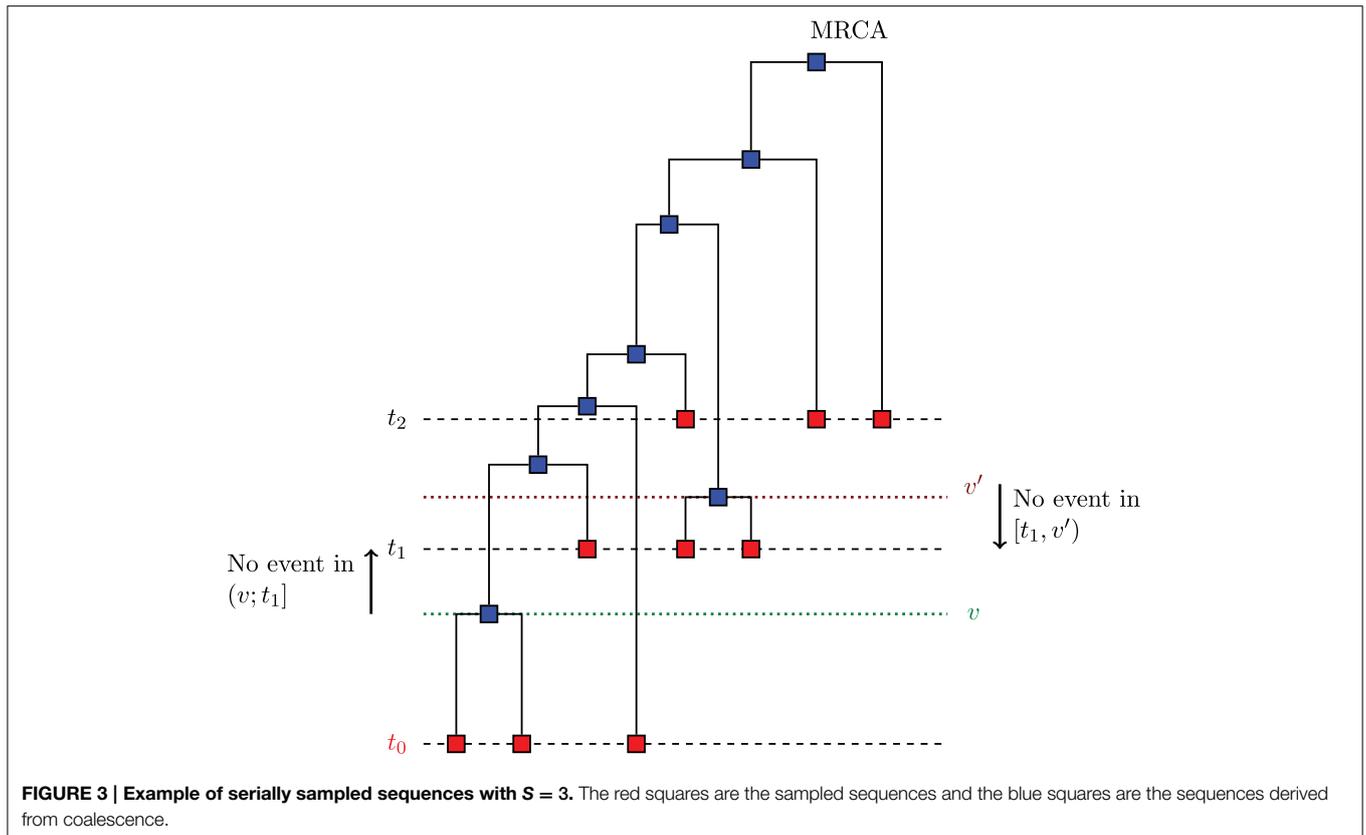


FIGURE 3 | Example of serially sampled sequences with $S = 3$. The red squares are the sampled sequences and the blue squares are the sequences derived from coalescence.

Case 2: $t_{s-1} < v < t_s, v' = t_s$.

$$\tilde{q}_{\theta}(D_{i-1,t_s}|D_{i,v}) = \begin{cases} \Pr(\exists \text{ an event in } (v, t_s]) \cdot C^{-1} \frac{\theta}{2} n_i^{(\alpha)} \frac{\hat{\pi}(\beta|H_i - \alpha)}{\hat{\pi}(\alpha|H_i - \alpha)} P_{\beta\alpha} & \text{if } M_{\alpha}^{\beta}, \\ \Pr(\exists \text{ an event in } (v, t_s]) \cdot C^{-1} \binom{n_i^{(\alpha)}}{2} \frac{1}{\hat{\pi}(\alpha|H_i - \alpha)} & \text{if } C_{\alpha}^{\alpha}, \\ \Pr(\text{no event in } (v, t_s]) & \text{if } H_{i-1} = H_i + \mathcal{E}_s, \\ 0 & \text{otherwise.} \end{cases} \quad (28)$$

Normally (i.e., in homochronous sampling), the waiting time W_{-i-1} from the state $D_{i,v}$ with $t_{s-1} < v < t_s$ to the next event has an exponential distribution with rate $\lambda_i = \binom{n_i}{2} + \frac{n_i\theta}{2}$, where n_i is the number of lineages at time v . Thus, the probability that there are no events in the interval $(v, v'] \equiv (v, t_s]$ is given by the survival function

$$\Pr(W_{-i-1} > t_s - v) = \exp(-\lambda_i(t_s - v)), \quad (29)$$

where W_{-i-1} is the waiting time from state H_i to state H_{i-1} in a process with homochronous sampling.

In the case of heterochronous sequences, the algorithm for simulating the genealogies backward in time is the following:

1. initialize $n_i = n$ and $s = 0$, where n is the number of sampled sequences at time $t_0 = 0$ (present), and s is the index of times where we perform the sampling. Further, initialize the cumulated time t_{cum} to 0;
2. simulate the time to the next event, W_{-i-1} , as an exponential distribution with parameter $\binom{n_i}{2} + \frac{n_i\theta}{2}$; let t_{evt} be the observed value;
3. compute $t_{cum}^* = t_{cum} + t_{evt}$;
4. if $t_{cum}^{(i)} < t_s$ and $t_{cum}^* > t_s$, then
 - let $t_{cum}^{(i-1)} = t_s$;
 - let $n_{i-1} = n_i + n^{(s)}$ (add a sample of sequences at time t_s);
 - let $s := s + 1$ and $i := i - 1$, and go to step 2;
- otherwise, go to step 5;
5. let $t_{cum}^{(i-1)} := t_{cum}^*$ and randomly choose a sequence from n_i ; the chosen sequence type is denoted α ;
6. compute the quantities x_1 and x_2 , where

$$x_1 = \theta \hat{\pi}(\beta|H_i - \alpha) P_{\beta\alpha} \quad \text{and} \quad x_2 = n_i^{(\alpha)} - 1.$$

Then, choose:

- a coalescence event with probability $\frac{x_2}{(x_1 + x_2)}$;
 - a mutation event (to β) with probability $\frac{x_1}{(x_1 + x_2)}$.
7. depending on the result in step 6:

- if there is a coalescence event, choose another sequence of type α randomly, and let $n_{i-1} = n_i - 1$;
- if there is a mutation event, mutate the sequence α into a sequence β , without changing n_i , i.e., let $n_{i-1} = n_i$;

8. let $i = i - 1$ and continue until $n_i = 1$.

After the definition of how to build a genealogy in the case of serially sampled sequences, and the proposal distribution Q , we introduce the probability P of the genealogy by specifying the probability of passing from the state $D_{i-1, v'} = \{H_{i-1}, v'\}$ to the state $D_{i, v} = \{H_i, v\}$ when there are n_{i-1} sequences, and we suppose that an event time v' is such that $t_s < v' < t_{s+1}$ (a coalescence corresponds to a split when viewed from the past to the present). Therefore, as for the backward transition probabilities, we consider separately the case $v > t_s$ and the case $v = t_s$.

Case 1: $t_s < v < v' < t_{s+1}$.

$$p_{\theta}(D_{i, v} | D_{i-1, v'}) = \begin{cases} \frac{n_{i-1}^{(\alpha)}}{n_{i-1}} \frac{\theta}{(n_{i-1} - 1 + \theta)} P_{\alpha\beta} & \text{if } M_{\alpha}^{\beta} \\ \frac{n_{i-1}^{(\alpha)}}{n_{i-1}} \frac{n_{i-1} - 1}{n_{i-1} - 1 + \theta} & \text{if } C_{\alpha}^{\alpha} \\ 0 & \text{otherwise.} \end{cases} \quad (30)$$

Case 2: $t_s < v' < t_{s+1}$ and $v = t_s$.

$$p_{\theta}(D_{i, v} | D_{i-1, v'}) = \begin{cases} \Pr(\exists \text{ an event in } [t_s, v']) \cdot \frac{n_{i-1}^{(\alpha)}}{n_{i-1}} \frac{\theta}{(n_{i-1} - 1 + \theta)} P_{\alpha\beta} & \text{if } M_{\alpha}^{\beta} \\ \Pr(\exists \text{ an event in } [t_s, v']) \cdot \frac{n_{i-1}^{(\alpha)}}{n_{i-1}} \frac{n_{i-1} - 1}{n_{i-1} - 1 + \theta} & \text{if } C_{\alpha}^{\alpha} \\ \Pr(\text{no event in } [t_s, v']) & \text{if } H_i = H_{i-1} - \mathcal{E}_s \\ 0 & \text{otherwise,} \end{cases} \quad (31)$$

where:

- the probability that there are no events in the interval $[t_s, v')$ is given by:

$$\Pr(Wt_{i-1} > v' - t_s) = \exp(-\lambda_i(v' - t_s)). \quad (32)$$

- $n_{i-1}^{(\alpha)}$ represents the number of sequences of type α in $D_{i-1, v'} = \{H_{i-1}, v'\}$;
- $H_i = H_{i-1} - \mathcal{E}_s$: represents the event of adding the set of sequences \mathcal{E}_s at time t_s .

As in the case of homochronous sequences, after computing the probabilities $p_{\theta}(D_{i, v} | D_{i-1, v'})$, and $\tilde{q}_{\theta}(D_{i-1, v'} | D_{i, v})$ for a genealogy $G^{(j)}, j = 1, 2, \dots, J$, the importance weights may be derived from Equations (14–17).

3.2.2. Estimation of the Effective Population Size for Heterochronous Sequences

For heterochronous sequences, the method for producing a *skywis plot* is similar to the one defined in Section 3.1.3. The main difference lies in the definition of epochs in this case ¹. In the presence of S serially sampled sequences, we cumulate the coalescence times as follows:

- for each simulated genealogy $\mathcal{G}^{(j)}$, we compute the MRCA time, $T_{MRCA}^{(j)}, j = 1, 2, \dots, J$;
- we fix the number of epochs at $n_{cum}^{(s)}$ in each time interval $(t_s; t_{s-1})$ where no new sample is added, $s = 1, 2, \dots, S$, $t_S = T_{MRCA}^{(j)}$, and $t_0 = 0$ (present);
- in order to define the epochs, the time cutting points in a genealogy $\mathcal{G}^{(j)}, j = 1, 2, \dots, J$ are computed as follows:

$$t_{cut, b}^{(j, s)} = t_{s-1} + 0.1 \cdot \exp\left(\frac{b}{n_{cum}} \cdot \log(1 + 10 \cdot t_s)\right) - 0.1, \quad (33)$$

where $b = 1, 2, \dots, n_{cum}^{(s)}$ and $s = 1, 2, \dots, S - 1$.

For each genealogy and for each time interval $(t_s; t_{s-1})$, $s = 1, 2, \dots, S$, formula Equation (33) gives the limits of the epochs from the present to the past in units of N generations.

In practice, we performed minor smoothing at times t_s , because the addition of new sequences creates an artificial discontinuity at t_s , $s = 1, 2, \dots, S$. Therefore, the population size in the first epoch after t_s is set to be equal to the effective population size in the epoch preceding the addition of new sequences.

4. Results

To test the ability of our method to capture the demographic signal contained in the DNA sequences, we simulated several demographics scenarios. Further, we compared the results of the *skywis plot* with those of the *generalized skyline plot* that uses single tree, and the *Bayesian skyline plot* that uses MCMC approach. These methods are the closest to our approach.

The DNA sequences were simulated using the *fastsimcoal* program (Excoffier and Foll, 2011) which allows us to consider several demographic scenarios and different mutation models. The genealogies were simulated ² using the method described in Section 3.1.1. In all our simulations, the coalescence times were cumulated into $n_{cum} = \sqrt{n} - 1$ epochs according to the method described in Section 3.1.3, where n represents the number of simulated DNA sequences. After that, we derive the *skywis plot* using Equations (24, 25).

¹The reason we changed the way we define the epochs is that the number of sequences rises at the instants of the serial sampling, so the method used in Section Simulation of Genealogies is not appropriate.

²The simulation of the genealogies was performed using MATLAB programming language (MATLAB, 2013) and the *Parallel Computing Toolbox* which allows parallelization of the simulation of genealogies. This is possible when using IS.

The *generalized skyline plot* was performed as follows:

1. From the DNA sequences generated by *fastsimcoal*, we estimated a phylogeny using the PHYLIP program (the PHYLogeny Inference Package Felsenstein, 1989) using the maximum likelihood method with a molecular clock constraint (we used the *dnamlk* program).
2. Based on the estimated tree produced by PHYLIP, we used the APE package (Paradis et al., 2004) to produce the *generalized skyline plot* according to the optimal strategy for grouping adjacent coalescent intervals introduced by Strimmer and Pybus (2001).

The *Bayesian skyline plot* was performed using the BEAST program, version 1.8.1. In order to reproduce a parametrization which is as close as possible to ours, we used (Hasegawa et al., 1985) substitution model with equal base frequencies, and a strict clock with rate 1.

Below, we present our results according to the demographic models we considered.

4.1. Constant Effective Population Size

In this case, we consider 50 simulated DNA sequences with parameters:

- number of nucleotides: 10,000;
- constant effective population size: 2000 generations;
- no recombination and no population structure;
- mutation rate equals to $2 \cdot 10^{-7}$: therefore ($\theta = 8$);
- JC69 (Jukes and Cantor, 1969) finite sites model.

The estimate of the effective population size (*skywis plot*) is shown in **Figure 4A**. We observe that the *skywis plot* (orange line) gives a relatively smooth curve of the effective population size. Further, the estimation turns around the real value N , with a slight over-estimation close to the present, which can be explained by the fact that when the mutation rate θ is large, the sampled sequences are all different, and we have many mutations before one coalescence; thus, coalescence times are longer, and the corresponding population sizes are larger (see Section Simulation of Genealogies.)

In **Figure 4B** we present the *generalized skyline plot* (in substitution units). In this case, the form of the graph is not recognizable as a constant line.

The *Bayesian skyline plot* is given in **Figure 4C**. In this case, the graph is very smooth and is easily recognizable as a constant line.

4.2. Piecewise Constant Function

In this section, we present results where 25 DNA sequences of length 10,000 nucleotides and mutation rate $\mu = 5 \cdot 10^{-4}$ were simulated under the JC69 mutation model. We assume that the effective population size follows the piecewise constant model function:

$$N_e(t) = \begin{cases} N & \text{if } t < x \\ aN & \text{otherwise,} \end{cases} \quad (34)$$

where $N = N(0) = 10^4$, $x = 5000$ generations, $a = 0.25$ (see **Figure 5**), and the time t is measured from present to the past.

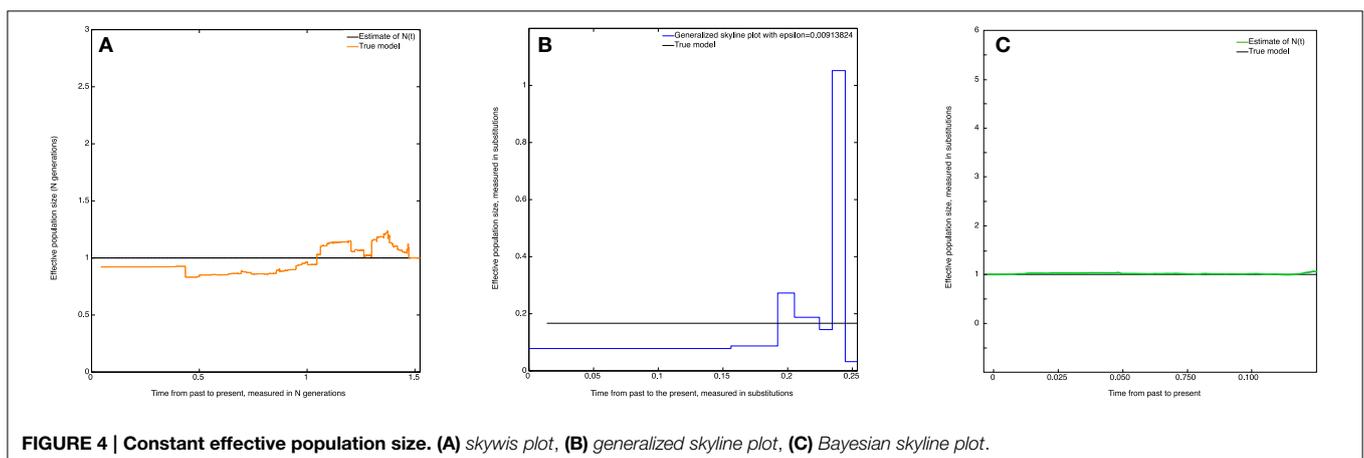
Figure 5A represents the non-parametric estimate (*skywis plot*) of the effective population size for a number of epochs equal to $n_{\text{cum}} = 4$. We note that the *skywis plot* was able to detect well enough the change-point of the size of the actual population which dates back 5000 generations. However, the *skywis plot* seems to overestimate the effective population size for $t > 5000$ generations.

In **Figure 5B** we present the *generalized skyline plot*. The *skywis plot* gives a better result than the *generalized skyline plot* close to the present, while the estimate given by the *generalized skyline plot* is closer to the true value when we approach the MRCA.

The *Bayesian skyline plot* presented in **Figure 5C** is very smooth and generally reproduces the true history except closer to the present, where the Bayesian skyline plot over-smoothes the effective population size.

4.3. Exponential Population Growth

In this section, we suppose that the effective population growth is exponential assuming an instantaneous growth rate that



is proportional to the current population size according to the equation $N_e(t) = N \exp(-\beta t)$ from present to the past.

Using the *fastsimcoal* program, we simulate 50 DNA sequences with the following parameters:

- Number of nucleotides: 1000;
- $N = N_e(0)$ at time $t = 0$: 10,000;
- no recombination, and no population structure;
- mutation rate: $5 \cdot 10^{-7}$ ($\theta = 1$);
- JC69 finite sites model;
- $\beta = 1$ (in generations).

The *skywis plot* for the simulated DNA sequences from the exponential model described above is given in **Figure 6A**.

The result given in **Figure 6A** is quite good in the sense that the size of the effective population decreases steadily from the present to the past and follows the exponential curve quite closely most of the time. However, we note that the estimated effective population size is almost constant from some point in time when approaching the TMRCA. This is explained by the fact that for the last two sequences the theoretical average

time to coalesce represents half the length of the tree, and from this point in time there is no much variability in the estimate of the population size. In particular, this remark led us to consider heterochronous sampling in order to improve the effective population size estimate.

In **Figure 6B** the time is measured in substitution units and we present the *generalized skyline plot*. As before, the *generalized skyline plot* has a fluctuating shape but it exhibits a certain tendency to decrease toward the past. In the end, when we approach the time of the MRCA, the *generalized skyline plot* gives an estimate that is close to the true value.

In **Figure 6C**, we present the *Bayesian skyline plot*. As in the other scenarios, the *Bayesian skyline plot* produces a very smooth curve; in this case it suggests that the population had a mild exponential expansion. However, we note that the curve remains constant closer to the MRCA.

4.4. Exponential Population Growth and Heterochronous Sequences

In order to test the methodology proposed in Section 3.2, we use the same parameters as in Section 4.3, but by assume that the 50 sequences were collected at different moments in time such as:

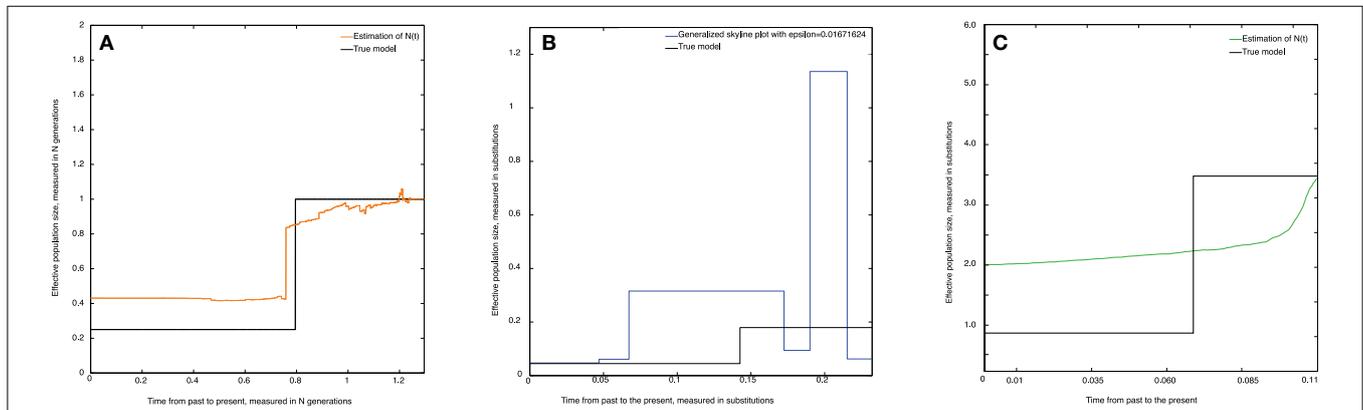


FIGURE 5 | Skywis plot for data simulated from the population model where $N(t) = 10,000$, if $t < 5000$ generations, and $N(t) = 2500$ otherwise (time from the past to the present). (A) skywis plot, (B) generalized skyline plot, (C) Bayesian skyline plot.

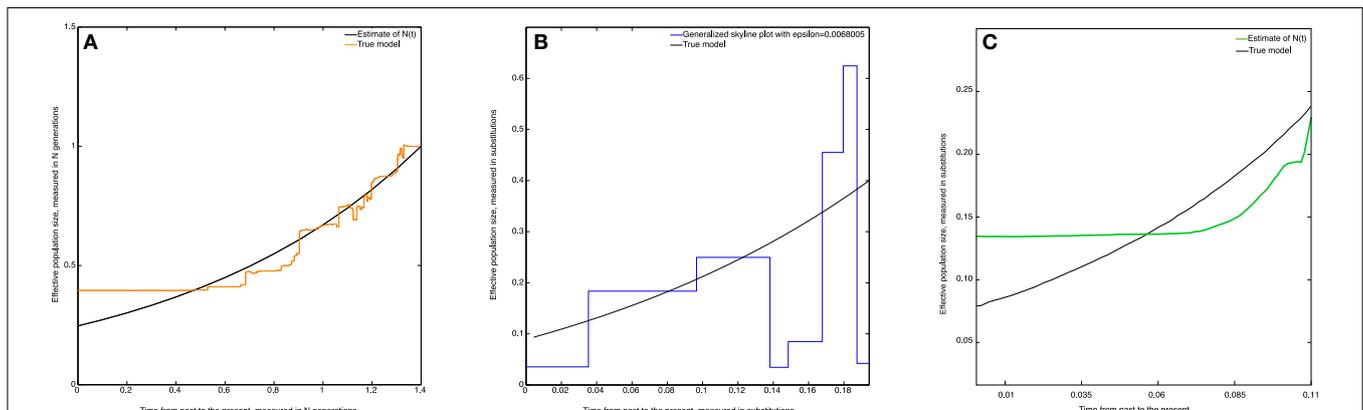
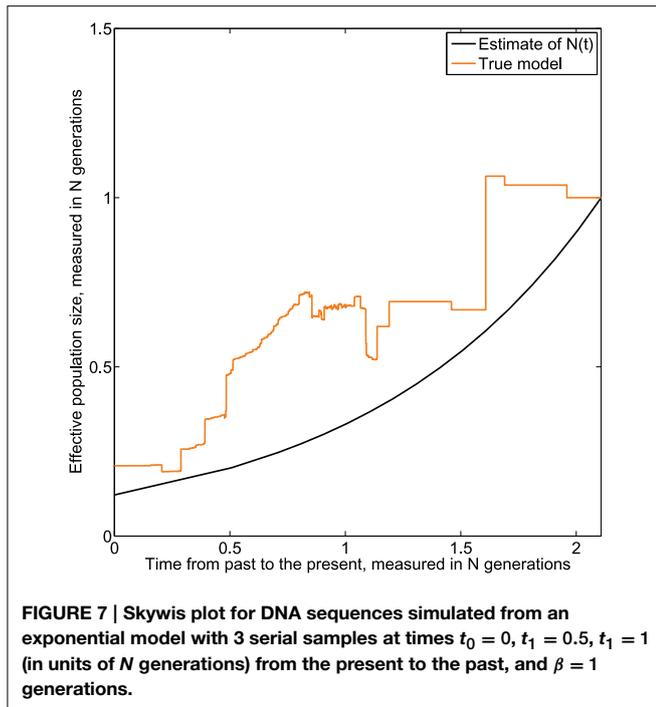


FIGURE 6 | Skywis plot for DNA sequences simulated from an exponential model with $\beta = 1$. (A) skywis plot, (B) generalized skyline plot, (C) Bayesian skyline plot.



- $n_0 = 25$ (present);
- $n_1 = 15$ at time $t_1 = 0.5$ in units of N generations (measured from present to the past);
- $n_2 = 10$ at time $t_2 = 1$ (N generations).

The result given in **Figure 7** suggests that the effective population decreases exponentially from present to the past. Further, we note that the estimated effective population size continues to decrease when approaching the time of the MRCA, which is a net improvement over the homochronous case. This could be explained by the fact that as more sequences are added over time, more information is available as one approaches the MRCA.

5. Discussion

The *skywis plot* is a new flexible method for exploring the demographic history of a sample of DNA sequences based on coalescent theory. Our nonparametric method is likelihood-based and uses IS. More precisely, we generate a large number of genealogies, both their times and their topology; further, we use the importance weights of these genealogies to compute a weighted average of the effective population size per epoch. This allows us to produce estimates that exhibit clear cut population growth tendencies over time, which is the main purpose of this

References

- Donnelly, P., and Tavaré, S. (1995). Coalescents and genealogical structure under neutrality. *Annu. Rev. Genet.* 29, 401–421.
- Drummond, A., Pybus, O. G., and Rambaut, A. (2003). Inference of viral evolutionary rates from molecular sequences.

approach, given that it is nonparametric. In practice, we expect our method to be used as a preliminary procedure that could be supplemented by a parametric analysis.

We present a framework of the new method and test by simulation its ability to capture the demographic signal contained in the DNA sequences under several demographic scenarios. Moreover, we consider both homochronous and heterochronous data using a simple substitution model, JC69 (Jukes and Cantor, 1969). We could also have considered more complicated substitution models, except those that allow variation in evolutionary rates among lineages.

For illustration we present the results given by the generalized skyline plot that uses a single genealogy, and those obtained by the *Bayesian skyline plot* that uses an MCMC approach. Although the *Bayesian skyline plot* is smoother than the *skywis plot*, our estimator is able to capture the shape of the effective population size $N_e(t)$, as well as its main change points, but in some examples it had a (slight) tendency to overestimate the population size as we approached the MRCA. This is not surprising, given that the simulation our estimation method entails first setting a constant population size (where coalescence times are longer) and further operating an adjustment through a weighting system. Further, note that, unlike the methods based on a single tree, it is possible to extend the *skywis plot* and include recombination. Indeed, recombination induces a graph structure rather than a tree, and IS methods in this context already exist (e.g., Fearnhead and Donnelly, 2001).

As a future development, we expect the method to be improved by considering an iterative procedure, in which the present approach would be the first estimation step. As a new approach, the *skywis plot* remains to be tested on more complex demographic models, and models of substitution that could be more realistic, especially for rapidly evolving RNA viruses. Also, the *skywis plot* can be easily extended to include multilocus data because, when there is no recombination, the same IS scheme can be applied.

Funding

The Ph.D. studies of SA were supported in part by scholarships awarded by the ISM (Institut des Sciences Mathématiques) and stipends out of NSERC (Natural Sciences and Engineering Research Council) research grants.

Acknowledgments

We acknowledge the stimulating discussions within the ÉMoStA group.

Adv. Parasitol. 54, 331–358. doi: 10.1016/S0065-308X(03)54008-8

- Drummond, A. J., Rambaut, A., Shapiro, B., and Pybus, O. G. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.* 22, 1185–1192. doi: 10.1093/molbev/msi103

- Drummond, A. J., Simon, Y. W. Ho., Phillips, M. J., and Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biol.* 4:e88. doi: 10.1371/journal.pbio.0040088
- Durbin, R., and Li, H. (2011). Inference of human population history from individual whole-genome sequences. *Nature* 475, 493–496. doi: 10.1038/nature10231
- Excoffier, L., and Foll, M. (2011). Fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics* 27, 1332–1334. doi: 10.1093/bioinformatics/btr124
- Fearnhead, P., and Donnelly, P. (2001). Estimating recombination rates from population genetic data. *Genetics* 159, 1299–1318.
- Felsenstein, J. (1989). Phylip - phylogeny inference package (version 3.2). *Cladistics* 5, 164–166.
- Fu, Y. (1994). A phylogenetic estimator of effective population size or mutation rate. *Genetics* 136, 685–692.
- Griffiths, R. C., and Tavaré, S. (1994a). Simulating probability distributions in the coalescent. *Theor. Popul. Biol.* 46, 131–159.
- Griffiths, R. C., and Tavaré, S. (1994b). Ancestral inference in population genetics. *Stat. Sci.* 9, 307–319.
- Griffiths, R. C., and Tavaré, S. (1994c). Sampling theory for neutral alleles in a varying environment. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 344, 131–159.
- Hasegawa, M., Kishino, H., and Yano, T. (1985). Dating of human-ape splitting by a molecular clock of mitochondrial dna. *J. Mol. Evol.* 22, 160–174.
- Hein, J., Schierup, M., and Wiuf, C. (2005). *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford, UK: Oxford University Press.
- Heled, J., and Drummond, A. (2008). Bayesian inference of population size history from multiple loci. *BMC Evol. Biol.* 8:289. doi: 10.1186/1471-2148-8-289
- Ho, S. Y., and Shapiro, B. (2011). Skyline-plot methods for estimating demographic history from nucleotide sequences. *Mol. Ecol. Resour.* 11, 423–434. doi: 10.1111/j.1755-0998.2011.02988.x
- Jukes, T. H., and Cantor, C. R. (1969). “Evolution of protein molecules,” in *Mammalian Protein Metabolism*, ed H. N. Munro (New York, NY: Academic Press), 21–132.
- Kingman, J. F. C. (1982). On the genealogy of large populations. *J. Appl. Probab.* 19A, 27–43.
- Kuhner, M. K., Yamato, J., and Felsenstein, J. (1995). Estimating effective population size and mutation rate from sequence data using metropolis-hastings sampling. *Genetics* 140, 1421–1430.
- Kuhner, M. K., Yamato, J., and Felsenstein, J. (1998). Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics* 149, 429–434.
- MATLAB. (2013). *Version 8.1.0.604 (R2013a)*. Natick, MA: The MathWorks Inc.
- Minin, V. N., Bloomquist, E. W., and Suchard, M. A. (2008). Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Mol. Biol. Evol.* 25, 1459–1471. doi: 10.1093/molbev/msn090
- Nordborg, M. (2003). “Coalescent theory,” in *Handbook of Statistical Genetics, 2nd Edn.*, eds D. Balding, M. Bishop, and C. Cannings (New York, NY: John Wiley and Sons Ltd.), 602–635.
- Paradis, E., Claude, J., and Strimmer, K. (2004). Ape: analyses of phylogenetics and evolution in r language. *Bioinformatics* 20, 289–290. doi: 10.1093/bioinformatics/btg412
- Pybus, O., Rambaut, A., and Harvey, P. H. (2000). An integrated framework for the inference of viral population history from reconstructed genealogies genetics. *Genetics* 155, 1429–1437.
- Oppen-Rhein, R., Fahrmeir, L., and Strimmer, K. (2005). Inference of demographic history from genealogical trees using reversible jump markov chain monte carlo. *BMC Evol. Biol.* 5:6. doi: 10.1186/1471-2148-5-6
- Rodrigo, A. G., and Felsenstein, J. (1999). “Coalescent approaches to HIV population genetics,” in *Molecular Evolution of HIV*, eds K. Crandall and H. John (Baltimore, MD: University Press), 233–272.
- Rodrigo, A. G., Shpaer, E. G., Delwart, E. L., Iversen, A. K., Gallo, M. V., Brojatsch, J., et al. (1999). Coalescent estimates of hiv-1 generation time *in vivo*. *Proc. Natl. Acad. Sci. U.S.A.* 96, 2187–2191.
- Slatkin, M., and Hudson, R. R. (1991). Pairwise comparisons of mitochondrial dna sequences in stable and exponentially growing populations. *Genetics* 129, 555–562.
- Stephens, M. (2000). Times on trees, and the age of an allele. *Theor. Popul. Biol.* 57, 109–119. doi: 10.1006/tpbi.1999.1442
- Stephens, M. (2001). “Inference under the coalescent,” in *Handbook of Statistical Genetics*, eds D. J. Balding, C. Cannings, and M. Bishop (Chichester: Wiley), 213–238.
- Stephens, M., and Donnelly, P. (2000). Inference in molecular population genetics. *J. R. Stat. Soc.* 62, 605–655. doi: 10.1111/1467-9868.00254
- Strimmer, K., and Pybus, O. G. (2001). Exploring the demographic history of dna sequences using the generalized skyline plot. *Mol. Biol. Evol.* 18, 2298–2305. doi: 10.1093/oxfordjournals.molbev.a003776

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2015 Ait Kaci Azzou, Larribe and Froda. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.