



HAL
open science

An Efficient Coalescent Model for Heterochronously Sampled Molecular Data

Lorenzo Cappello, Amandine Véber, Julia A Palacios

► **To cite this version:**

Lorenzo Cappello, Amandine Véber, Julia A Palacios. An Efficient Coalescent Model for Heterochronously Sampled Molecular Data. *Journal of the American Statistical Association*, In press, pp.1-13. 10.1080/01621459.2024.2330732 . hal-04568807

HAL Id: hal-04568807

<https://hal.science/hal-04568807v1>

Submitted on 5 May 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License



An Efficient Coalescent Model for Heterochronously Sampled Molecular Data

Lorenzo Cappello, Amandine Véber & Julia A. Palacios

To cite this article: Lorenzo Cappello, Amandine Véber & Julia A. Palacios (17 Apr 2024): An Efficient Coalescent Model for Heterochronously Sampled Molecular Data, Journal of the American Statistical Association, DOI: [10.1080/01621459.2024.2330732](https://doi.org/10.1080/01621459.2024.2330732)

To link to this article: <https://doi.org/10.1080/01621459.2024.2330732>



© 2024 The Author(s). Published with license by Taylor & Francis Group, LLC.



[View supplementary material](#)



Published online: 17 Apr 2024.



[Submit your article to this journal](#)



Article views: 206






[View related articles](#)



[View Crossmark data](#)

An Efficient Coalescent Model for Heterochronously Sampled Molecular Data

Lorenzo Cappello^{a,b,d} , Amandine Véber^c , and Julia A. Palacios^{d,e} 

^aDepartment of Economics and Business, Universitat Pompeu Fabra, Barcelona, Spain; ^bData Science Center, Barcelona School of Economics, Barcelona, Spain; ^cUniversité Paris Cité, CNRS, MAP5, Paris, France; ^dDepartment of Statistics, Stanford University, Stanford, CA; ^eDepartment of Biomedical Data Science, Stanford Medicine, Stanford, CA

ABSTRACT

Molecular sequence variation at a locus informs about the evolutionary history of the sample and past population size dynamics. The Kingman coalescent is used in a generative model of molecular sequence variation to infer evolutionary parameters. However, it is well understood that inference under this model does not scale well with sample size. Here, we build on recent work based on a lower resolution coalescent process, the Tajima coalescent, to model longitudinal samples. While the Kingman coalescent models the ancestry of labeled individuals, we model the ancestry of individuals labeled by their sampling time. We propose a new inference scheme for the reconstruction of effective population size trajectories based on this model and the infinite-sites mutation model. Modeling of longitudinal samples is necessary for applications (e.g., ancient DNA and RNA from rapidly evolving pathogens like viruses) and statistically desirable (variance reduction and parameter identifiability). We propose an efficient algorithm to calculate the likelihood and employ a Bayesian nonparametric procedure to infer the population size trajectory. We provide a new MCMC sampler to explore the space of heterochronous Tajima's genealogies and model parameters. We compare our procedure with state-of-the-art methodologies in simulations and an application to ancient bison DNA sequences. Supplementary materials for this article are available online including a standardized description of the materials available for reproducing the work.

ARTICLE HISTORY

Received July 2022
Accepted January 2024

KEYWORDS

Ancient DNA; Bayesian nonparametric; Effective population size; Gaussian process; Kingman n -coalescent; Markov lumping

1. Introduction

Inference of population size dynamics and other evolutionary parameters from molecular sequence data is an important problem in evolutionary biology and molecular epidemiology of infectious diseases (Ho and Shapiro 2011; Volz, Koelle, and Bedford 2013; Cappello et al. 2022). Phylogenetic models account for the dependence among n DNA sequences \mathbf{Y} and models observed variation through two stochastic processes: an ancestral process of the sample represented by a genealogy \mathbf{g} , and a mutation process with a given set of parameters μ that, conditionally on \mathbf{g} , models the phenomena that have given rise to the sequences.

A standard choice for modeling \mathbf{g} is the Kingman n -coalescent, (Kingman 1982a, 1982b), a model that depends on a parameter called *effective population size* $(N_e(t))_{t \geq 0}$ (henceforth $N_e = (N_e(t))_{t \geq 0}$), a measure of genetic diversity over time. Inference of N_e has important applications in many fields, such as genetics, anthropology, and public health. To give an example of the applications that can be handled with our proposal, we analyze ancient samples of bison in North America (Froese et al. 2017), revisiting the question of why the Beringian bison went extinct (Shapiro et al. 2004). Reproducing Shapiro et al. (2004) analysis with a new dataset is interesting in light of the growing evidence of an arrival of humans in North America earlier than previously estimated (Bourgeon, Burke, and Higham 2017; Becerra-Valdivia and Higham 2020).

In this article, we assume the infinite sites mutation (ISM) model in which multiple mutations never occur at the same nucleotide position (site). A mutation under this model partitions the sample into sequences carrying the mutation and sequences not carrying the mutation. A consequence is that only a subset of the genealogical space has positive likelihood, making the ISM model computationally attractive. Recent work exploits this advantage (Speidel et al. 2019).

Standard Bayesian phylogenetic approaches approximate the augmented posterior $\pi(N_e, \mathbf{g} | \mathbf{Y}, \mu)$ through Markov chain Monte Carlo (MCMC). Here, the genealogy is treated as an auxiliary variable introduced to compute the likelihood in $\pi(N_e, \mathbf{g} | \mathbf{Y}, \mu) \propto P(\mathbf{Y} | \mathbf{g}, \mu) \pi(\mathbf{g} | N_e) \pi(N_e)$. Approximation of the posterior requires the definition of Markov chains (MCs) on genealogies, whose state space is the product space $\mathcal{G}_n \times \mathbb{R}_+^{n-1}$ of tree topologies ($g \in \mathcal{G}_n$) and coalescent times $\mathbf{t} \in \mathbb{R}_+^{n-1}$ (times of coalescence events in the genealogy). Sampling from these distributions is extremely challenging: the target distribution is highly multi-modal, many different topologies have the same likelihood (Sanderson et al. 2015), and mixing times of tree-valued Markov chains are at best polynomial (Simper and Palacios 2022). The issue is exacerbated as the sample size increases because the cardinality of \mathcal{G}_n grows superexponentially with n for the standard coalescent ($|\mathcal{G}_n| = n!(n-1)!/2^{n-1}$). The result is that state-of-the-art methodologies are not scalable to the amount of data available. To resolve this

CONTACT Julia A. Palacios  juliapr@stanford.edu  Department of Statistics, Stanford University, Stanford, CA.

 Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

© 2024 The Author(s). Published with license by Taylor and Francis Group, LLC.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

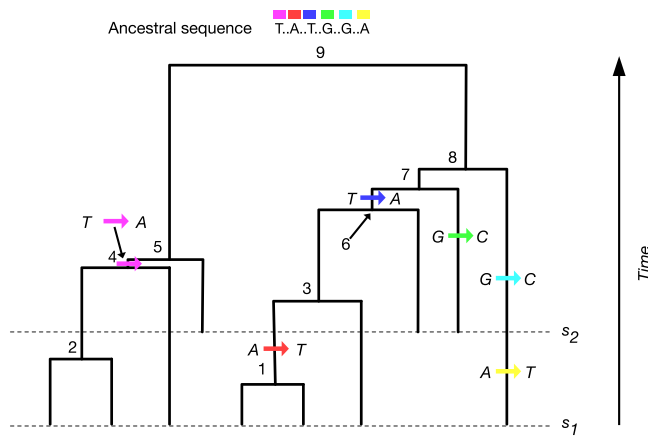


Figure 1. Coalescence and mutation. A genealogy of 10 individuals at a locus of 100 sites is depicted as a bifurcating tree. Six mutations (at different sites) along the branches of the tree give rise to the 10 sequences. Black dots represent the 94 sites that do not mutate in the ancestral sequence. The nucleotides at the polymorphic sites are shown, and the colored arrows depict how ancestral sites are modified by mutation.

computational bottleneck, much research has focused on algorithms better suited for this setting, such as sequential Monte Carlo (Bouchard-Côté, Sankararaman, and Jordan 2012; Wang, Bouchard-Côté, and Doucet 2015; Fourment et al. 2017; Dinh, Darling, and Matsen IV 2017), Hamiltonian Monte Carlo (Dinh et al. 2017), and variational Bayes (Zhang and Matsen IV 2018).

An alternative (or perhaps complementary) solution is to use a lower resolution ancestral (genealogical) process, known as the Tajima n -coalescent (Tajima 1983; Sainudiin, Stadler, and Véber 2015; Palacios et al. 2019). While the tree topologies under Kingman are *labeled* ranked tree shapes with n leaves, the tree topologies under Tajima are *unlabeled* ranked tree shapes with n -leaves (see, e.g., Figure 1), with a space of trees with a drastically smaller cardinality than that of the space of Kingman trees (Disanto and Wiehe 2013). This amounts to taking equivalence classes of Kingman trees, in which the timed ranked tree shape is retained, but leaf labels are removed so that the external tree branches are all considered equivalent. This in turn, requires a new likelihood calculation in which individual DNA sequences are grouped by patterns of shared mutations. Intuitively, the likelihood of Tajima’s genealogies should be more concentrated and lead to more efficient inference. We elaborate on this argument through the following example.

We generated a genealogy with $n = 6$ tips and superimposed three mutations along the branches of the tree at three different sites (Figure 2). The resulting “unlabeled” data consist of one sequence carrying two mutations and two sequences carrying the same one mutation. This is the information used for calculating the likelihood of a Tajima genealogy. The “labeled” data used to compute the likelihood of a Kingman genealogy consist of sequence a carrying two mutations and sequences e and f carrying one mutation. We computed the likelihood of all Kingman and Tajima genealogies with six leaves, all with the same “true” coalescent times. There are 360 Kingman topologies and 16 Tajima topologies with positive likelihood. Figure 2 depicts the distribution of the normalized likelihood values along with their frequencies. Under Kingman’s coalescent, the maximum likelihood value is about 823.2 times larger than

the minimum likelihood value. Under Tajima’s coalescent, this ratio between the maximum and minimum likelihood value is about 3.3. That is, the range of likelihood values is reduced. Moreover, the profiles are remarkably different: under Kingman’s coalescent, there are many trees with a negligible likelihood and a few with higher values; under Tajima’s coalescent, the more frequent likelihood values are closer to the center. Along with the lower cardinality, Tajima’s likelihood profile should make MCMC exploration of tree space more manageable. Indeed, one of the challenges in sampling from multimodal distributions is that MCs cannot fully explore the space because the chains struggle to move between modes due to the proposals getting rejected in low densities regions. The profile seen in Figure 2 will lead to higher acceptance probabilities in Metropolis Hastings, making it easier to move between modes. Importantly, the gains are obtained at no loss of information about N_e . All we lose are the sequence labels in the data and the mapping of mutations to a labeled genealogy. Instead, we retain the unlabeled data and the set of mappings of shared mutations to edges in the unlabeled genealogy. We will be more precise on this aspect in Section 3.

The Tajima coalescent was first used to infer N_e by Palacios et al. (2019), but despite the advances in that article, there are still many challenges to be addressed for the Tajima n -coalescent to be a viable alternative to the Kingman n -coalescent. First, the algorithm for the likelihood calculation of Palacios et al. (2019) can be prohibitively expensive; a loose upper bound of the current algorithm’s complexity is $\mathcal{O}(n!)$. Second, the definition of the likelihood relies on several restrictive modeling assumptions such as no recombination, no population structure, and the fact that all samples are obtained at a single point in time. In this article, we address several of these issues. We introduce a new algorithm for likelihood calculation whose upper bound complexity is $\mathcal{O}(n^2)$, and we extend the Tajima modeling framework to sequences observed at different time points like those at the tips of the genealogy in Figure 1, that is, heterochronous data. Our proposed method infers N_e , mutation rate μ , and can deal with data collected at multiple independent loci. For parsimony, this general setting is studied in the supplementary material.

Out of the many possible directions that may have been pursued from Palacios et al. (2019), the extension to heterochronous data was prioritized for several reasons: (i) data are collected longitudinally in many applications (e.g., ancient DNA and viral phylodynamics), (ii) employing longitudinal data reduces the variance of the estimators of N_e (Rodrigo, Ewing, and Drummond 2007) and (iii) the model becomes identifiable for joint estimation of mutation rates and effective population sizes (Drummond et al. 2002; Parag and Pybus 2019).

Modeling longitudinal data requires the definition of a continuous time Markov chain, which is a lumping of the heterochronous n -coalescent introduced by Rodrigo and Felsenstein (1999). We refer to the lower resolution of this process (the lumped process) as the Tajima *heterochronous* n -coalescent. This process differs from the Tajima n -coalescent (Sainudiin, Stadler, and Véber 2015; Palacios et al. 2019) in that sequences sampled at different time points are not exchangeable. The Tajima heterochronous is a model on partially-labeled genealogies of heterochronous samples.

The rest of the article proceeds as follows. In Section 2, we define the Tajima heterochronous n -coalescent. In Section 3, we

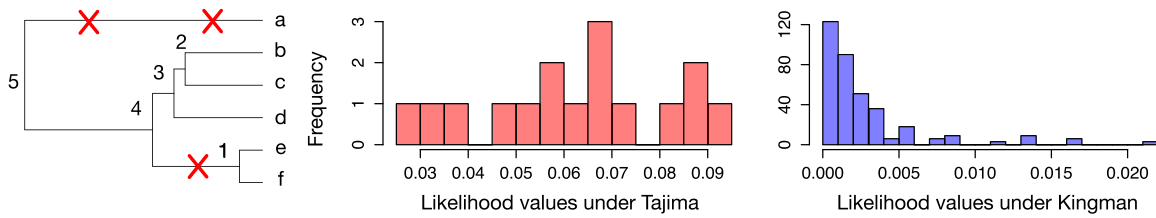


Figure 2. Distributions of the likelihood values under the Kingman and the Tajima n -coalescent for a given dataset. The first plot shows a realization of a genealogy of $n = 6$ samples (tips) with three mutations superimposed (marked as X). The second plot shows the histogram of the likelihood values conditionally on all possible Tajima tree topologies, and the third plot shows the histogram of the likelihood values conditionally on all possible Kingman trees with 6 leaves. We assumed the same coalescent times across all trees.

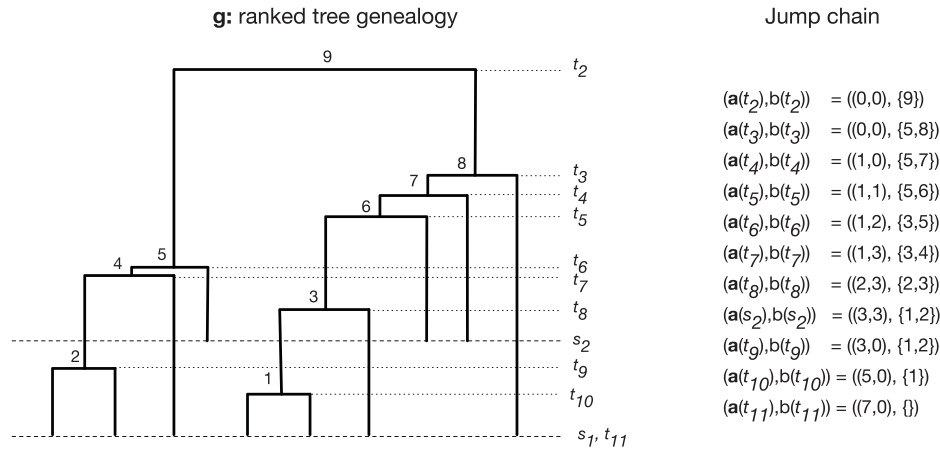


Figure 3. Example of a Tajima heterochronous genealogy and its jump chain. A realization of a Tajima heterochronous n -coalescent with $\mathbf{n} = (7, 3)$ and $\mathbf{s} = (s_1, s_2)$, represented as a ranked tree shape with coalescence and sampling times, denoted \mathbf{g} . The column to the right displays the corresponding jump chain (see the text for notation).

introduce the assumed mutation model, describe the data, define the likelihood and the new algorithm to compute it. Section 4 describes the MCMC algorithm for posterior inference; in Section 5 we discuss the gained efficiency of the proposed approach and in Section 6, we present a comprehensive simulation study outlining how the method works and compares with state-of-the-art alternatives. In Section 7, we analyze modern and ancient bison sequences described in Froese et al. (2017). Section 8 concludes.

2. The Tajima Heterochronous n -coalescent

The Tajima heterochronous n -coalescent is a lumping of the heterochronous n -coalescent (Rodrigo and Felsenstein 1999) that describes the ancestral relationships of a set of n individuals sampled, possibly at different times, from a large population. The set of ancestral relationships of the sample is represented by a ranked and partially labeled genealogy, for example the one depicted in Figure 3 in which internal nodes are ranked bottom-up and tips are partially labeled by their sampling time. We assume that every organism is dated and labeled according to the time in which the organism lived (if ancient, by radiocarbon date) or in which the living organism was sequenced. In this generalization of the Tajima coalescent, each pair of extant ancestral lineages merges into a single lineage at an instantaneous rate depending on the current effective population size $N_e(t)$, and new lineages are added when one of the prescribed sampling times is reached. In this work, we do not model the stochasticity of sampling times but we condition on them being fixed. The

main difference with the isochronous Tajima coalescent is that in the isochronous case, the genealogy is unlabeled, that is, all sequences are sampled at the same time.

Let us introduce some notation and terminology. Let m be the number of sampling time points and n be the total number of samples. Let $\mathbf{n} = (n_1, \dots, n_m)$ denote the number of sequences collected at times $\mathbf{s} = (s_1, \dots, s_m)$, with $s_1 = 0$ denoting the present time, and $s_j > s_{j-1}$ for $j = 2, \dots, m$ (time goes from present toward the past). We refer to sampling group as the set of sequences sampled at a given time; there are m of such groups. The sampling group corresponding to sequences sampled at s_i is labeled by s_i . Let $\mathbf{t} = (t_{n+1}, \dots, t_2)$ be the vector of coalescent times with $t_{n+1} = 0 < t_n < \dots < t_2$; these are the times when two lineages have a common ancestor. By convention, we include $t_{n+1} = 0$ (the present, or the time of the latest sample) despite not being a coalescent time. Such a convention is useful when defining the density of \mathbf{t} . Note that the subscript in t_k does not indicate the current number of lineages, as it is often done in the coalescent literature, but it indicates the number of lineages that have yet to coalesce (some sequences may not have been sampled yet). We use the rank order of the coalescent events (bottom-up) to label the internal nodes of the genealogy. That is, the node corresponding to the coalescent event occurring at time t_n is labeled 1 (see t_{10} in Figure 3), the node corresponding to the coalescence event occurring at time t_{n-1} is labeled 2, etc. We refer to the internal node labels as *vintages* (i.e., rankings), and to lineages that subtend a vintage as *vintaged lineages* to highlight the fact that they are labeled by a given vintage, and thus, distinguishable from all other lineages.

The Tajima heterochronous n -coalescent is an inhomogeneous continuous-time Markov chain $(\mathbf{a}(t), b(t))_{t \geq 0}$ that keeps track of $\mathbf{a}(t)$, a vector of length m whose j th position indicates the number of singletons from sampling group s_j at time t , and $b(t)$ is the set of vintaged lineages at time t . The process starts at $t = 0$ in state $(\mathbf{a}(0) = (n_1, 0, \dots, 0), b(0) = \emptyset)$, jumps deterministically at every sampling time and jumps stochastically at every random coalescent time until it reaches the unique absorbing state $(\mathbf{a}(t_2) = (0, \dots, 0), b(t_2) = \{n - 1\})$ at time t_2 , when all n samples have a single most recent common ancestor at the root (Figure 3). At each sampling time s_i , the state of the Tajima coalescent jumps deterministically as follows:

$$(\mathbf{a}(s_i), b(s_i)) = (\mathbf{a}(s_i-) + n_i \mathbf{e}_i, b(s_i-)),$$

where $f(s_i-)$ denotes the left limit of the function f at s_i and \mathbf{e}_i is the i th unit vector.

Let us now turn to the embedded jump chain at coalescent times. At time t_i , a random pair of extant lineages coalesce to create a new lineage with vintage $n + 1 - i$. All extant lineages coalesce at rate 1 but the transition probability depends on which pair coalesces. Singleton lineages of the same sampling group are indistinguishable and distinguishable from those of other sampling groups and vintaged lineages. All vintaged lineages are uniquely labeled and distinguishable from all the others. Four types of coalescence transitions are possible depending on which lineages are involved: (a) two singletons from the same sampling group coalesce (up to m possible moves for the chain), (b) two singletons from different sampling groups coalesce (up to $m(m-1)/2$ possible moves), (c) one singleton lineage and one vintaged lineage coalesce (up to m possible moves), or (d) two vintaged lineages coalesce (only one possibility because for vintages, the sampling information is irrelevant). Each pair coalesces with the same probability and the transition probabilities at coalescent times are thus given by

$$P[(\mathbf{a}(t_i), b(t_i)) | (\mathbf{a}(t_i-), b(t_i-))] \quad (1)$$

$$= \begin{cases} \frac{\prod_{j=1}^m \binom{a_j(t_i-)}{a_j(t_i-) - a_j(t_i)}}{\binom{\sum_{j=1}^m a_j(t_i-) + |b(t_i-)|}{2}} & \text{if } (\mathbf{a}(t_i), b(t_i)) < (\mathbf{a}(t_i-), b(t_i-)) \\ 0 & \text{otherwise} \end{cases}$$

where $(\mathbf{a}(t_i), b(t_i)) < (\mathbf{a}(t_i-), b(t_i-))$ means that $(\mathbf{a}(t_i), b(t_i))$ can be obtained by merging two lineages of $(\mathbf{a}(t_i-), b(t_i-))$, and $|b|$ denotes the cardinality of the set b .

Observe that the quantity $\sum_{j=1}^m a_j(t_i-) + |b(t_i-)|$ appearing in (1) corresponds to the total number of extant lineages just before the event at t_i . Furthermore, since only two lineages coalesce at time t_i , at most two terms in the product appearing in the numerator of (1) are not equal to one. Finally, if $m = 1$, at any time $t \geq 0$ the vector $\mathbf{a}(t)$ is made of a single integer and (1) degenerates into the transition probabilities of the Tajima isochronous n -coalescent (Sainudiin, Stadler, and Véber 2015; Palacios et al. 2019); on the other hand, if $m = n$, the process degenerates into the Kingman heterochronous n -coalescent since all singletons are uniquely labeled by their sampling times. Figure 3 shows a possible realization from the Tajima heterochronous n -coalescent. Notice that in applications, the

number of observations collected at any given time instance is generally larger than one, and hence, the heterochronous Tajima model would have a smaller state space than the Kingman model on heterochronous data. We investigate this assertion in supplementary material (SM) section 6, where we developed a sequential importance sampling to tackle this combinatorial question.

To define the distribution of the holding times, we introduce the following notation. We denote the intervals that end with a coalescent event at t_k by $I_{0,k}$ and the intervals that end with a sampling time within the interval (t_{k+1}, t_k) as $I_{i,k}$ where $i \geq 1$ is an index tracking the sampling events in (t_{k+1}, t_k) . More specifically, for every $k \in \{2, \dots, n\}$, we define

$$I_{0,k} = [\max\{t_{k+1}, s_j\}, t_k], \quad (2)$$

where the maximum is taken over all $s_j < t_k$, and for every $i \geq 1$ we set

$$I_{i,k} = [\max\{t_{k+1}, s_{j-i}\}, s_{j-i+1}], \quad (3)$$

where the maximum taken over all $s_{j-i+1} > t_{k+1}$ and $s_j < t_k$. We also let $n_{i,k}$ denote the number of extant lineages during the time interval $I_{i,k}$. For example, in Figure 3, in (t_9, t_8) we have $I_{0,8} = [s_2, t_8)$, $I_{1,8} = [t_9, s_2)$ and no $I_{i,8}$ for $i \geq 2$. The vector of coalescent times \mathbf{t} is a random vector whose density with respect to Lebesgue measure on \mathbb{R}_+^{n-1} can be factorized as the product of the conditional densities of t_{k-1} knowing t_k , which reads: for $k = 3, \dots, n + 1$,

$$p(t_{k-1} | t_k, \mathbf{s}, \mathbf{n}, N_e)$$

$$= \frac{C_{0,k-1}}{N_e(t_{k-1})} \exp \left\{ - \int_{I_{0,k-1}} \frac{C_{0,k-1}}{N_e(t)} dt + \sum_{i=1}^m \int_{I_{i,k-1}} \frac{C_{i,k-1}}{N_e(t)} dt \right\}, \quad (4)$$

where $C_{i,k} := \binom{n_{i,k}}{2}$, and the integral over $I_{i,k-1}$ is zero if there are less than i sampling times between t_k and t_{k-1} . The distribution of the holding times defined above corresponds to the same distribution of holding times in the heterochronous Kingman coalescent (Rodrigo and Felsenstein 1999). Although the heterochronous Tajima coalescent takes values on a different state space, it remains true that every pair of extant lineages coalesces at equal rate.

Finally, given \mathbf{n} , \mathbf{s} and \mathbf{t} , a complete realization of the Tajima heterochronous n -coalescent chain can be uniquely identified with an unlabeled binary ranked tree shape g of $\mathbf{n} = (n_1, \dots, n_m)$ samples at (s_1, \dots, s_m) with its $n - 1$ coalescent transitions, so that

$$P(g | \mathbf{t}, \mathbf{s}, \mathbf{n}) = \prod_{i=2}^n P[(\mathbf{a}(t_i), b(t_i)) | (\mathbf{a}(t_i-), b(t_i-))]. \quad (5)$$

Equation (5) gives the prior probability of the tree topology g under the Tajima heterochronous n -coalescent. Putting together (4) and (5), we obtain a prior distribution on the space of genealogies $\mathbf{g} = (g, \mathbf{t})$, which are ranked (unlabeled) topologies equipped with branch lengths

$$\pi(\mathbf{g} | \mathbf{s}, \mathbf{n}, N_e) = P(g | \mathbf{t}, \mathbf{s}, \mathbf{n}) \prod_{k=3}^{n+1} p(t_{k-1} | t_k, \mathbf{s}, \mathbf{n}, N_e). \quad (6)$$

3. Data and Likelihood

3.1. Infinite Sites Model and the Perfect Phylogeny

We assume that the observed data \mathbf{Y} consists of n unlabeled sequences at z polymorphic (mutating) sites at a non-recombining contiguous segment of DNA of organisms with a low mutation rate. Under these assumptions, a widely studied mutation model is the *infinite sites model* (ISM) (Kimura 1969; Watterson 1975) with Poissonian mutation, which corresponds to a Poisson point process with rate μ on the branches of \mathbf{g} such that every mutation occurs at a different site and no mutations are hidden by a second mutation affecting the same site.

An important consequence of the ISM is that \mathbf{Y} can be represented as an incidence matrix \mathbf{Y}_1 and a frequency counts matrix \mathbf{Y}_2 . \mathbf{Y}_1 is a $k \times z$ matrix with 0–1 entries, where 0 indicates the ancestral type and 1 indicates the mutant type; k is the number of unique sequences (or haplotypes) observed in the sample, and the columns correspond to polymorphic sites. \mathbf{Y}_2 is a $k \times m$ count matrix where the (i, j) th entry denotes how many haplotype i sequences belonging to group s_j are sampled. For example, the $n = 10$ sequences defined by the realizations of the ancestral and mutation processes depicted in Figure 1 can be summarized into \mathbf{Y}_1 and \mathbf{Y}_2 displayed in Figure 4(A). Note that we make the implicit assumption that we know which state is ancestral at each segregating site. However, this assumption can be relaxed, see (Griffiths and Tavaré 1995), although the computational cost will substantially increase.

\mathbf{Y}_1 and \mathbf{Y}_2 can alternatively be represented graphically as an *augmented perfect phylogeny* \mathbf{T} . Our likelihood algorithm exploits this graphical representation of the data. The augmented perfect phylogeny representation is an extension of the *gene tree* or *perfect phylogeny* (Gusfield 1991; Griffiths and Tavaré 1994) to the heterochronous case. The standard perfect phylogeny definition leaves out the information carried by \mathbf{Y}_2 . In the augmented perfect phylogeny $\mathbf{T} = (\mathbf{V}, \mathbf{E})$, \mathbf{V} is the set of nodes of \mathbf{T} , and \mathbf{E} is the set of weighted edges. For parsimony, the edge that connects node V_i to its parent node is denoted by E_i . Palacios et al. (2019) also employ a generalization of Gusfield’s perfect phylogeny but their construction differs in that there

is no bookkeeping of sampling information. We define \mathbf{T} as follows:

1. Each haplotype labels at least one leaf in \mathbf{T} . If a haplotype is observed at k different sampling times, then k leaves in \mathbf{T} will be labeled by the same haplotype. The pair (haplotype label, sampling group) uniquely labels each leaf node.
2. Each of the z polymorphic sites labels exactly one edge. When multiple sites label the same edge, the order of the labels along the edge is arbitrary. Some external edges (edges subtending leaves) may not be labeled, indicating that they do not carry additional mutations to their parent node.
3. For any pair (haplotype h_k , sampling group), the labels of the edges along the unique path from the root to the leaf h_k specify all the sites where h_k has the mutant type.

Figure 4(B) plots \mathbf{T} corresponding to \mathbf{Y}_1 and \mathbf{Y}_2 displayed in Figure 4(A). Observe that \mathbf{T} includes sampling information in the leaf labels. In the example, h_c labels two leaves because it is observed at times s_1 and s_2 . The corresponding edges E_3 and E_4 are unlabeled, that is, no mutations are allocated to those edges because the underlying nodes carry identical sequences (same haplotype). We “augment” Gusfield’s perfect phylogeny because the sampling information is crucial in the likelihood calculation.

\mathbf{T} implicitly carries some quantitative information that can be quickly summarized. We denote the number of observed sequences subtended by an internal node V by $|V|$. If V is a leaf node, $|V|$ denotes the frequency of the haplotype h observed at the corresponding sampling time s . Similarly, we denote the number of mutation labels assigned to an edge E by $|E|$. If no mutations are assigned to E , then $|E| = 0$. See Figure 4(C) for an example.

Gusfield (1991) gives an algorithm to construct the perfect phylogeny \mathbf{T}^* in linear time. Constructing \mathbf{T} from \mathbf{T}^* is straightforward since all we need is to incorporate the sampling information and add leaf nodes if a haplotype is observed at multiple sampling times. If we drew \mathbf{T}^* from the data in Figure 4, it would not have node V_4 , but only a single node V_3 labeled by haplotype h_c . A description of the algorithm can be found in the SM section 2.

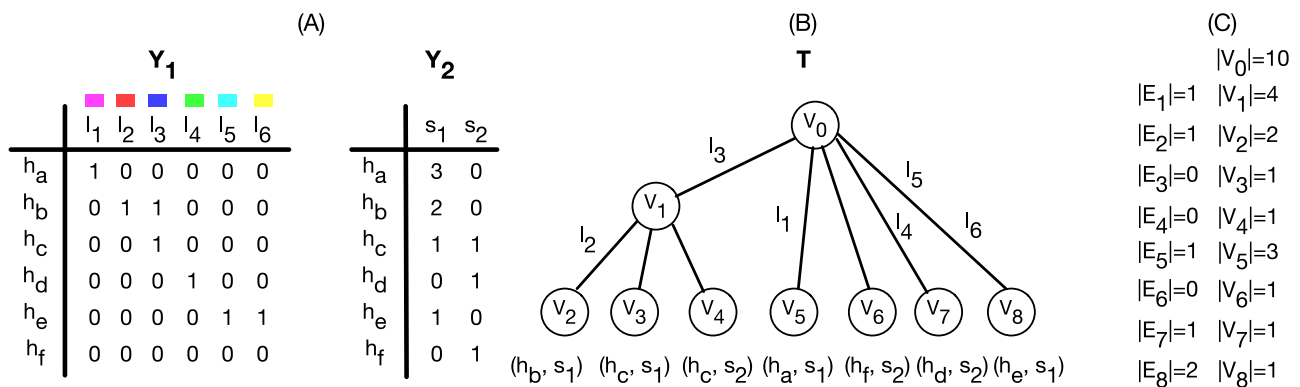


Figure 4. Incidence matrix, frequency matrix and perfect phylogeny representation. Panel (A): data is summarized as an incidence matrix \mathbf{Y}_1 (h denotes the haplotypes, l the segregating sites, the colors correspond to those depicted in Figure 1) and a matrix of frequencies \mathbf{Y}_2 (s denotes the sampling group). Panel (B): \mathbf{T} denotes the perfect phylogeny corresponding to \mathbf{Y}_1 and \mathbf{Y}_2 ; each of the 6 polymorphic sites labels exactly one edge. When an edge has multiple labels, the order of the labels is irrelevant. Each leaf node is labeled by a pair (haplotype, sampling time), with each haplotype possibly labeling more than one leaf nodes. Panel (C): $|E_i|$ corresponds to the number of mutations along the edge subtending node V_i in (B) and $|V_i|$ corresponds to the number of sequences descending from V_i in (B), see the text for details.

3.2. Likelihood

The crucial step needed to compute the likelihood of a heterochronous Tajima genealogy \mathbf{g} is to sum over all possible allocations of mutations to its branches with the corresponding sampling group. This can be efficiently done by exploiting the augmented perfect phylogeny representation of the data \mathbf{T} and by first mapping nodes of \mathbf{T} to subtrees of \mathbf{g} . We note that with Kingman's coalescent, tree leaves are labeled by the sequences so there is a unique possible allocation of mutations to branches. With isochronous Tajima coalescent, leaves are unlabeled and there are potentially multiple allocations of mutations to branches. In Palacios et al. (2019), the authors employ a backtracking algorithm that traverses bottom-up the isochronous perfect phylogeny. Here, we reverse the point of view with a two-steps approach: first, we define a top-down algorithm that uses \mathbf{T} to identify all possible allocations (Section 3.2.1), then the output of the first step is fed into a sum-product algorithm that uses the set of possible allocations to compute the likelihood efficiently (Section 3.2.2).

3.2.1. Allocations

Let \mathbf{a} be a vector of length $n - 1$ that encodes a possible mapping of nodes of \mathbf{T} to subtrees of \mathbf{g} . The i th entry of \mathbf{a} gives the node in \mathbf{T} which is mapped to the subtree with vintage i , \mathbf{g}_i (including the branch that subtends vintage i). Our algorithm first maps all *non-singleton* nodes \mathbf{V} of \mathbf{T} to subtrees of \mathbf{g} , that is, only nodes such that $|V| > 1$ are entries of \mathbf{a} . Singleton nodes in \mathbf{T} ($V \in \mathbf{V}$ such that $|V| = 1$) are treated separately and are initially excluded from the allocation step. For example, Figure 5 shows a possible vector \mathbf{a} whose entries are the non-singleton nodes V_0, V_1, V_2 , and V_5 of \mathbf{T} of Figure 4. We note that nodes can appear more than once in \mathbf{a} , meaning that they can be mapped to more than one subtrees. On the other hand, a single node V_i is not necessarily mapped to all the vintages, leaves and internal

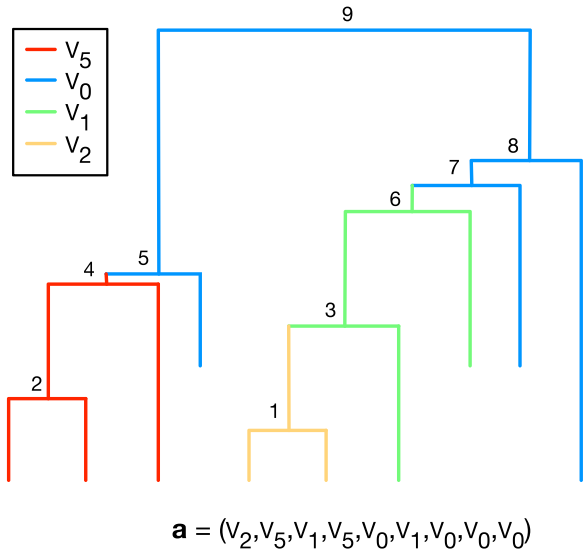


Figure 5. A possible allocation of non-singleton nodes of \mathbf{V} to subtrees of \mathbf{g} . For a given allocation \mathbf{a} (bottom figure), we display how subtrees in \mathbf{g} (identified by the vintage tag at their root—black number in the top figure) are allocated to the nodes of \mathbf{T} . Each color depicts an allocation of a subtree to a node: V_5 (red), V_0 (blue), V_1 (green), and V_2 (yellow).

branches of \mathbf{g}_j ; different nodes may be mapped to some subtrees of \mathbf{g}_j (including external branches), leading to a situation where V_i is mapped to only a subset of the vintages and branches constituting \mathbf{g}_j . For example, in Figure 5, V_1 is mapped to \mathbf{g}_6 and \mathbf{g}_3 , but V_2 is mapped to \mathbf{g}_1 , a subtree of both \mathbf{g}_6 and \mathbf{g}_3 ; hence, V_1 is only mapped to the green part of \mathbf{g}_6 and \mathbf{g}_3 as depicted in Figure 5.

The precise mapping of nodes in \mathbf{T} to subtrees of \mathbf{g} is needed to allocate mutations in \mathbf{T} to branches of \mathbf{g} . The algorithm, described in the SM section 3, outputs all possible allocations. We denote by $\#\mathbf{a}$, the number of possible allocations and a given allocation as \mathbf{a}_i , for $i = \dots, \#\mathbf{a}$.

3.2.2. Likelihood Calculations

To calculate the likelihood, we assume the ISM of mutations and that mutations occur according to a Poisson point process with rate μ on the branches of \mathbf{g} , where μ is the total mutation rate. To compute the likelihood we need to map mutations in \mathbf{T} to branches of \mathbf{g} and this is done for each mapping \mathbf{a}_i of non-singleton nodes of \mathbf{T} to subtrees of \mathbf{g} . For every V in \mathbf{T} such that $|V| > 1$, we define \mathbf{E}_V as the set formed by the edges in \mathbf{T} that subtend singleton children of V and, with the exception of $V = V_0$, \mathbf{E}_V in addition includes the edge that subtends V . For the example in Figure 4(B), $\mathbf{E}_{V_1} = \{E_1, E_3, E_4\}$. Let \mathbf{V}^* be the set of all $V \in \mathbf{V}$ such that $|V| > 1$, then the likelihood function is defined as

$$\begin{aligned} P(\mathbf{Y} \mid \mathbf{g}, N_e, \mu) &= \sum_{i=1}^{\#\mathbf{a}} P(\mathbf{Y}, \mathbf{a}_i \mid \mathbf{g}, N_e, \mu) \\ &= \sum_{i=1}^{\#\mathbf{a}} \prod_{V \in \mathbf{V}^*} P(V, \mathbf{E}_V, \mathbf{a}_i \mid \mathbf{g}, N_e, \mu), \end{aligned} \quad (7)$$

where $P(V, \mathbf{E}_V, \mathbf{a}_i \mid \mathbf{g}, N_e, \mu)$ is the probability of observing the mutations of the \mathbf{E}_V edges along the corresponding branches of \mathbf{g} defined by the mapping \mathbf{a}_i as follows.

If V has no singleton child nodes, then $\mathbf{E}_V = \{E\}$ and

$$P(V, \{E\}, \mathbf{a}_i \mid \mathbf{g}, N_e, \mu) \propto (\mu l)^{|E|} e^{-\mu \mathcal{T}}, \quad (8)$$

where l is the length of the branch in \mathbf{g} that subtends \mathbf{g}_j , j is the largest index such that $\mathbf{a}_{i,j} = V$, and \mathcal{T} denotes the length of the subtree in \mathbf{g} to which V is mapped in \mathbf{a}_i (as described in Section 3.2.1). For example, considering V_2 in Figure 5, we have $\mathcal{T}_2 = 2t_n + (t_{n-2} - t_n)$ and $l = (t_{n-2} - t_n)$ is the length of the branch connecting vintage 1 to vintage 3.

If node V has singleton child nodes,

$$\begin{aligned} P(V, \{E, E_{ch_1}, \dots, E_{ch_k}\}, \mathbf{a}_i \mid \mathbf{g}, N_e, \mu) \\ \propto (\mu l)^{|E|} e^{-\mu \mathcal{T}} \sum_{\mathbf{R} \in \Pi(\mathbf{E}_V)} \prod_{j=1}^k (\mu l_{R_j})^{|E_{ch_j}|}, \end{aligned} \quad (9)$$

where the first term on the r.h.s is defined exactly as the quantity on the r.h.s. of (8), while the second term corresponds to the probability of all possible different matchings between R_1, \dots, R_k , the first k indexes such that $\mathbf{a}_{i,R_j} = V$, and $|E_{ch_1}|, |E_{ch_2}|, \dots, |E_{ch_k}|$, the k numbers of mutations observed on the edges $E_{ch_1}, \dots, E_{ch_k}$ leading to the child nodes of V . In this expression, $\Pi(\mathbf{E}_V)$ is the set of all possible such matchings \mathbf{R} .

Before defining $\Pi(\mathbf{E}_V)$ more precisely, we make two observations. First, not all matchings are possible since not all leaf branches terminate at the same time (heterochronous sampling). Second, it is enough to consider the allocations that contribute to distinct likelihood values, that is, allocations for which the underlying samples are “distinguishable” in the sense that they have a different number of mutations.

We define $\Pi(\mathbf{E}_V)$ as the set of all possible “distinct matchings of number of observed singleton mutations to singleton branches”, that is, allocations which lead to a distinct likelihood values. To construct $\Pi(\mathbf{E}_V)$, we first partition the singleton edges $E_{ch_1}, \dots, E_{ch_k}$ according to the sampling times of the corresponding nodes $V_{ch_1}, \dots, V_{ch_k}$. Let k_j be the number of nodes in $\{V_{ch_1}, \dots, V_{ch_k}\}$ with sampling time s_j , that is, the size of each subset of the partition. We then further partition these subsets by grouping together the edges carrying the same number of mutations (defined as $|E_{ch_1}|, \dots, |E_{ch_k}|$). For each given sampling time s_j , let $k_j^{(1)}, \dots, k_j^{(m_j)}$ denote the cardinalities of the m_j sub-subsets obtained by this procedure, so that $k_j = \sum_{h=1}^{m_j} k_j^{(h)}$. The cardinality of $\Pi(\mathbf{E}_V)$ is then

$$|\Pi(\mathbf{E}_V)| = \prod_{j=1}^m \frac{k_j!}{k_j^{(1)}! \dots k_j^{(m_j)}!}, \quad (10)$$

where the product in (10) is the number of permutations with repetition of the different edges that are compatible with the data in terms of sampling times and numbers of mutations carried. Note that (10) is not the same as eq. (6) in Palacios et al. (2019) because here we account for the different sampling groups. It degenerates into eq. (6) in Palacios et al. (2019) in the isochronous case.

Last, we note that knowing a priori the full matrix \mathbf{A} of all possible allocations, allows to compute efficiently the likelihood (7) via a sum-product algorithm. The set of all possible \mathbf{a} is an $\#\mathbf{a} \times (n - 1)$ matrix \mathbf{A} , where each row is a possible \mathbf{a} ($n - 1$ columns) and the number of rows $\#\mathbf{a}$ is equal to the number of possible allocations. Now, for each $V \in \mathbf{V}^*$ there may be several rows \mathbf{a} of \mathbf{A} such that $P(V, \mathbf{E}_V, \mathbf{a} \mid \mathbf{g}, N_e, \mu)$ is the same, due to the fact that V is mapped to the same subtree in all these allocations. For such a V , one could compute the likelihood corresponding to these r distinct allocations, which we denote by $\mathbf{a}'_1, \dots, \mathbf{a}'_r$, in the following way:

$$\sum_{i=1}^r \prod_{V \in \mathbf{V}^*} P(V, \mathbf{E}_V, \mathbf{a}'_i \mid \mathbf{g}, N_e, \mu) = P(V, \mathbf{E}_V, \mathbf{a}'_1 \mid \mathbf{g}, N_e, \mu) \sum_{i=1}^r \prod_{V \in \mathbf{V}^* \setminus \{V\}} P(V', \mathbf{E}_{V'}, \mathbf{a}'_i \mid \mathbf{g}, N_e, \mu). \quad (11)$$

The exact sum-product formulation of (7) is specific to the observed \mathbf{Y} and \mathbf{A} .

4. Bayesian Model and MCMC Inference

To complete our Bayesian model we now need to specify a prior distribution on $\log N_e$ (the logarithm is used to ensure that $N_e(t) > 0$ for $t \geq 0$). We follow Palacios and Minin (2013), and

place a Brownian motion process as prior on $\log N_e$. We thus have:

$$\begin{aligned} \mathbf{Y} \mid \mathbf{g}, \mu, N_e, \mathbf{n}, \mathbf{s} &\sim \text{Poisson process} \\ \mathbf{g} \mid N_e, \mathbf{s}, \mathbf{n} &\sim \text{Tajima heterochronous } n\text{-coalescent} \quad (12) \\ \log N_e \mid \tau &\sim \text{BM}(0, C(\tau)) \\ \tau &\sim \text{Gamma}(\alpha, \beta) \end{aligned}$$

where $C(\tau)$ is the covariance function of a Brownian process with mean 0 and precision τ . The novelty in (12) is to model the genealogy with the Tajima heterochronous n -coalescent. In terms of modeling N_e , our framework allows for any prior on a piece-wise constant trajectory $\log N_e$. For example, if one expects sudden changes in N_e , the recently proposed prior of Faulkner et al. (2020) could be a good alternative that has shown good empirical results. The posterior distribution

$$\begin{aligned} \pi(\log N_e, \tau, \mathbf{g} \mid \mathbf{Y}, \mu) &\propto P(\mathbf{Y} \mid \mathbf{g}, \log N_e, \mu) \\ &\pi(\mathbf{g} \mid \log N_e) \pi(\log N_e \mid \tau) \pi(\tau), \end{aligned} \quad (13)$$

is approximated via MCMC with Metropolis-within-Gibbs updates. At each MCMC iteration, we jointly update $(\log N_e, \tau)$ via a Split Hamiltonian Monte Carlo (HMC) (Shahbaba et al. 2014) suitably adapted to phylodynamics inference by Lan et al. (2015); then we update the topology g and \mathbf{t} . We propose two Metropolis steps to update g and \mathbf{t} . The latter may also be combined in a single step. The transitions for g and \mathbf{t} are tailored to the Tajima n -coalescent genealogies. To update g , we employ the scheme in Palacios et al. (2019) suitably adjusted for the heterochronous case, with two local proposals that either swap two consecutive coalescent events or swap two offspring, each descending from two different and consecutive coalescent events (Palacios et al. 2019, Figure 4). To update \mathbf{t} , we propose a new sampler (SM, Section S5.2) that accounts for the observed sampling times constraints, an issue specific to heterochronous samples under the ISM assumption (SM, Section S5.1).

Model (12) can be generalized in various ways. We can infer model parameters from data observed at multiple independent loci (without recombination); see SM S11 for details and a simulation study. When the mutation rate is unknown but sequences sampled at different times have accumulated mutations, the model parameters N_e and μ become jointly identifiable (Drummond et al. 2002). In this case, we can add an additional prior distribution to model (12) and an extra MCMC step; see SM S12. Lastly, to incorporate the case of unknown ancestral state, one needs to sum over all possible ancestral states compatible with the data (Griffiths and Tavaré 1995). Other extensions will be mentioned in the discussion.

5. Increased Efficiency with No Loss of Information

We have assumed that the observed data \mathbf{Y} can be summarized as a haplotype incidence matrix \mathbf{Y}_1 and a frequency matrix \mathbf{Y}_2 , that is, the individual samples are labeled by their haplotype and sampling information (see e.g., Figure 4). Under Kingman’s coalescent, every sequence is uniquely labeled, and every tip in the Kingman’s genealogy corresponds to a labeled sequence. Let \mathbf{Y}^{lab} denote the fully labeled data (i.e., the augmented data where each sequence in \mathbf{Y} receives a unique identifying label in

$\{1, \dots, n\}$) and \mathbf{g}^K a labeled ranked tree shape, then the marginal likelihood is $P(\mathbf{Y}^{lab} \mid N_e)$, where we do not condition on μ because, for simplicity, we assume that the mutation rate μ is fixed. If we further assume that all sequences are sampled at the same time point and all observed haplotypes have different frequencies $n_i \neq n_j$ for $i \neq j, i, j = 1, \dots, k$, then we have

$$P(\mathbf{Y} \mid N_e) = \binom{n}{n_1, \dots, n_k} P(\mathbf{Y}^{lab} \mid N_e).$$

The same combinatorial factor above would result even if $n_i = n_j$ as long as the two haplotypes have different numbers of mutations and become distinguishable. If different haplotypes become indistinguishable in terms of number of mutations and frequency, then the combinatorial factor would change by another constant which is independent of N_e . This shows that there is no loss of information when using the partially labeled data \mathbf{Y} . We will use $c(\mathbf{Y})$ to denote the combinatorial factor that counts the number of labeled datasets corresponding to the unlabeled dataset by dropping the sample labels.

The motivation for estimating the posterior of $\log N_e$ by sampling over Tajima's genealogies instead of sampling over Kingman's genealogies is a variance reduction. With labeled data, we are interested in

$$P(N_e \mid \mathbf{Y}^{lab}) = \frac{P(\mathbf{Y}^{lab} \mid N_e)P(N_e)}{P(\mathbf{Y}^{lab})} = \frac{P(N_e)\mathbb{E}_{\mathbf{g}^k}[P(\mathbf{Y}^{lab} \mid \mathbf{g}^k)]}{P(\mathbf{Y}^{lab})}$$

where $\mathbb{E}_{\mathbf{g}^k}[P(\mathbf{Y}^{lab} \mid \mathbf{g}^k)]$ is computed with respect to $P(\mathbf{g}^k \mid N_e)$. Similarly, with unlabeled data, we are interested in

$$P(N_e \mid \mathbf{Y}) = \frac{P(\mathbf{Y} \mid N_e)P(N_e)}{P(\mathbf{Y})} = \frac{P(N_e)\mathbb{E}_{\mathbf{g}}[P(\mathbf{Y} \mid \mathbf{g})]}{P(\mathbf{Y})}$$

where $\mathbb{E}_{\mathbf{g}}[P(\mathbf{Y} \mid \mathbf{g})]$ is computed with respect to $P(\mathbf{g} \mid N_e)$.

Both $\mathbb{E}_{\mathbf{g}^k}[P(\mathbf{Y}^{lab} \mid \mathbf{g}^k)]$ and $\mathbb{E}_{\mathbf{g}}[P(\mathbf{Y} \mid \mathbf{g})]$ are approximated via Monte Carlo. Making the two variances comparable, we wish to show that for any given \mathbf{Y} and \mathbf{Y}^{lab}

$$\frac{\text{var}_{\mathbf{g}^k}[P(\mathbf{Y}^{lab} \mid \mathbf{g}^k)]}{P(\mathbf{Y}^{lab})^2} \geq \frac{\text{var}_{\mathbf{g}}[P(\mathbf{Y} \mid \mathbf{g})]}{P(\mathbf{Y})^2}. \quad (14)$$

Since the distribution of \mathbf{t} is shared in both models, we will assume fixed \mathbf{t} and replace \mathbf{g}, \mathbf{g}^k notation by g, g^k to denote topologies only. Since $P(\mathbf{Y}) = c(\mathbf{Y})P(\mathbf{Y}^{lab})$, we wish to show

$$c(\mathbf{Y})^2 \text{var}_{\mathbf{g}^k}[P(\mathbf{Y}^{lab} \mid \mathbf{g}^k)] \geq \text{var}_g[P(\mathbf{Y} \mid \mathbf{g})]. \quad (15)$$

The simplest cases occur when $n \in \{1, 2, 3\}$. In these cases there is only one g tree topology and so $\text{var}_g[P(\mathbf{Y} \mid \mathbf{g})] = 0$, while $\text{var}_{\mathbf{g}^k}[P(\mathbf{Y}^{lab} \mid \mathbf{g}^k)] \geq 0$ and (15) holds. The next two minimum nontrivial cases are the following.

Example 1 ($n=4$, two haplotypes). Consider two haplotypes with frequencies $n_1 = n_2 = 2$ with m_1 and m_2 mutations in each haplotype, and $m_1 \neq m_2$. We can also fix $\mu = 1$ as it is will account for a proportionality constant. The only tree topology with nonzero likelihood is the tree subtending two cherries. Let

$l_1 = t_2 - t_3$ and $l_2 = t_2 - t_4$, the lengths of the two branches subtending the two cherries. In this case

$$\begin{aligned} \text{var}_g[P(\mathbf{Y} \mid \mathbf{g})] &= c \frac{2}{9} (l_1^{m_1} l_2^{m_2} + l_1^{m_2} l_2^{m_1})^2, \\ \text{var}_{\mathbf{g}^k}[P(\mathbf{Y}^{lab} \mid \mathbf{g}^k)] &= c \left[\frac{1}{18} (l_1^{2m_1} l_2^{2m_2} + l_1^{2m_2} l_2^{2m_1}) \right. \\ &\quad \left. - \frac{1}{324} (l_1^{m_1} l_2^{m_2} + l_1^{m_2} l_2^{m_1})^2 \right]. \end{aligned}$$

where $c = e^{-2\mathcal{T}} / [(m_1!)(m_2!)^2]$ and \mathcal{T} denotes the tree length. Using the two variances, along with $c(\mathbf{Y}) = 4!/(2!2!) = 6$, we can check that that the inequality (15) holds on in this example. In this case, we see that the inequality is strict. Details of the calculations of the variances are given in the SM Section 1.1.

How much bigger $\text{var}_{\mathbf{g}^k}[P(\mathbf{Y}^{lab} \mid \mathbf{g}^k)]$ is than $\text{var}_g[P(\mathbf{Y}^{lab} \mid \mathbf{g})]$ will depend on \mathbf{t}, m_1 and m_2 .

To provide a more general sense of the difference in the variances for this specific example, we sampled 100 time vectors \mathbf{t} assuming $N_e = 1$, and fixed $m_1 = 2$ and $m_2 = 4$. The ratio

$$r = \frac{\text{var}_{\mathbf{g}^k}[P(\mathbf{Y}^{lab} \mid \mathbf{g}^k)]}{\text{var}_g[P(\mathbf{Y} \mid \mathbf{g})]/c(\mathbf{Y})^2}$$

is on average 7.18, with a maximum value of 8.49, and a minimum value of 4.

Example 2 ($n=4$, three haplotypes). Consider three haplotypes with frequencies $n_1 = 2, n_2 = 1$, and $n_3 = 1$ with $m_1 = 1, m_2 = 1$, and $m_3 = 0$ mutations in each haplotype. The two variances are available in closed form in the SM Section 1.2, however, the two analytical expressions are not insightful. The average ratio r based on 100 simulations is 60, taking values from 2 to 320.

While we currently do not have a proof of (14), the previous two examples show the variance reduction in two nontrivial cases. Revisiting the example discussed in introduction (Figure 2), we showed that the likelihood under Tajima is “more concentrated,” in the sense that, for a fixed dataset, the range of possible likelihood values is drastically smaller. We conjecture that this property, along with the cardinality reduction, contributes to a more efficient exploration of the tree space. The Tajima n -coalescent partitions the space of Kingman's trees into equivalence classes, where each Tajima's topology corresponds to a set of Kingman's topologies. That is,

$$P(\mathbf{Y} \mid \mathbf{g}) = c(\mathbf{Y}) \sum_{\mathbf{g}^k: \text{unlabeled}(\mathbf{g}^k)=\mathbf{g}} P(\mathbf{Y}^{lab} \mid \mathbf{g}^k) P(\mathbf{g}^k \mid \mathbf{g}).$$

To compute the likelihood under Tajima, we effectively sum over many topologies with small likelihood. More details on this example are given in the SM section 1.

6. Simulations

In Section 6.1 we compare average runtimes of our likelihood algorithm vis-a-vis the backtracking-based one in Palacios et al. (2019). In Section 6.2, we infer N_e using the Tajima heterochronous n -coalescent, and compare results with our own implementation of the Kingman heterochronous n -coalescent. Both sections rely on simulated data obtained as follows: given

\mathbf{n} , \mathbf{s} , and N_e , we simulate genealogies under the Tajima heterochronous n -coalescent. Given a realized \mathbf{g} and fixed μ , we draw M mutations from a Poisson distribution with parameter μL (L is the length of the tree \mathbf{g} : the sum of all branch lengths of \mathbf{g}) and place them independently uniformly at random along the branches of the timed genealogy. \mathbf{Y}_1 , \mathbf{Y}_2 , and \mathbf{T} are constructed as in Section 3.1.

6.1. Average Runtime

We consider varying sample sizes $n \in (5, 10, 15, 20, 25, 30)$, 10 simulated datasets per each n , and 100 genealogies per dataset. All datasets are simulated under constant population size, and μ is set to have an expected number of mutations of $n/2$. Since the method of Palacios et al. (2019) does not support heterochronous data, we assumed $\mathbf{s} = 0$. We compute $P(\mathbf{Y} | \mathbf{g}, N_e, \mu)$ with the two algorithms for all simulations. Figure 6 plots the average runtimes in seconds, showing that our proposal substantially reduces the average runtime in this specific setup.

6.2. Inference of N_e

We simulate genealogies with three population scenarios: a bottleneck (“bottle”), an instantaneous drop (“drop”), and two periods of constant population size with exponential growth in between (“exp”). For each scenario, we generated genealogies with three numbers of leaves ($n = 14, 35, 70$) and different \mathbf{n} , \mathbf{s} as summarized in Table S1 in the SM Section 7. The mutation parameter is varied to analyze the effect of the number of segregating sites on the quality of the estimation, but in this section it is assumed to be known. A simulation study dealing with unknown μ (i.e., joint estimation of N_e and μ) is given in SM S12. Details of the trajectories and of \mathbf{n} , \mathbf{s} employed are given in SM Section 7. All functions required to simulate genealogies, sequence data and estimate N_e are implemented in the R package `phylodyn`. Likelihood calculation is implemented in Python and called via the R package `reticulate` (Ushey, Allaire, and Tang 2022). MCMC tuning parameters are discussed in SM Section S7 and the validity of the algorithms’ implementation is discussed in SM Section 8.

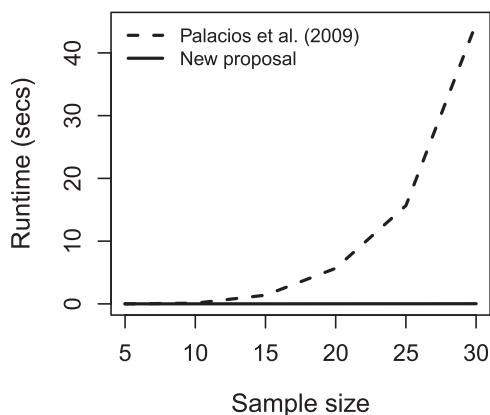


Figure 6. Average runtime (in seconds). The likelihood $P(\mathbf{Y} | \mathbf{g}, N_e, \mu)$ is computed using two algorithms, the one in Palacios et al. (2019) (dashed line) and our proposal. The average is based on 100 genealogies sampled for 10 simulated datasets for each sample size $n \in (5, 10, 15, 20, 25, 30)$.

Comparison to other methods. To our knowledge, there is no publicly available software implementing Bayesian nonparametric inference for N_e under the ISM and variable population size. To test the performance of our model, we implemented a function for computing the likelihood of Kingman’s genealogies for labeled data (we compare the runtimes of this implementation to Tajima in the SM, see Figure S7). For posterior approximation via MCMC, we used the Markovian proposal on the space of ranked labeled topologies of Markovtsova, Marjoram, and Tavaré (2000) (the Tajima proposal has the same rationale but acts on the space of ranked unlabeled topologies). The kernels used to update \mathbf{t} and $\log(N_e)$ are shared between the two implementations. Similarly, we employ the same initialization, removing the labels for Tajima’s chain. We also compare our results to an oracle estimator that infers N_e from the true \mathbf{g} . The oracle estimation is obtained using the method of Palacios and Minin (2012), which is equivalent to model (12) removing the randomness on g and \mathbf{t} . Estimation of N_e in the oracle does not use MCMC but INLA (Rue, Martino, and Chopin 2009). A comparison with two other methodologies implemented in BEAST (Drummond et al. 2012) is in SM section 9. We do not include the results in the main manuscript because these methodologies assume a different mutation model, a different prior on N_e , and a different MCMC scheme. The comparisons with BEAST should be interpreted as validity checks of our implementations and a baseline attainable by state-of-the-art implementations.

Criteria measured. We approximated the posterior distributions under Kingman and Tajima models via MCMC past convergence for a fixed time budget (72 hr). We evaluated trace plots and effective sample sizes (ESS) of \mathbf{t} and of $\log(N_e)$ as empirical assessments of convergence (SM Section S8). In a second study described in SM section 10, we run MCMC past convergence for one million iterations with a larger burn-in, given there are no time constraints. See SM Section S7 for details. There are several limitations that hinder a thorough comparison of empirical mixing among different modeling resolutions. These limitations arise from targeting different posterior distributions, with different initializations and the absence of precise criteria for assessing the achievement of stationarity.

To assess the accuracy in N_e estimate, we employ three criteria. As a measure of bias, we use the sum of relative errors (SRE), $SRE = \sum_{i=1}^k \frac{|\hat{N}_e(v_i) - N_e(v_i)|}{N_e(v_i)}$, where (v_1, \dots, v_k) is a regular grid of k time points, $\hat{N}_e(v_i)$ is the posterior median of N_e at time v_i and $N_e(v_i)$ is the value of the true trajectory at time v_i . To quantify the uncertainty in the estimate, the second criterion is the mean relative width, defined by $MRW = \frac{1}{k} \sum_{i=1}^k \frac{|\hat{N}_{97.5}(v_i) - \hat{N}_{2.5}(v_i)|}{\hat{N}(v_i)}$, where $\hat{N}_{97.5}(v_i)$ and $\hat{N}_{2.5}(v_i)$ are respectively the 97.5% and 2.5% quantiles of the posterior distribution of $N(v_i)$. Lastly, we consider the envelope measure defined by $ENV = \frac{1}{k} \sum_{i=1}^k \mathbf{1}_{\{\hat{N}_{2.5}(v_i) \leq N_e(v_i) \leq \hat{N}_{97.5}(v_i)\}}$, which measures the proportion of the curve that is covered by the 95% credible region; that is, it is a proxy for coverage. We stress that the three metrics are relevant measures of performance as long as they are considered jointly. For example, one can obtain very high coverage (ENV) with wide credible regions (high MRW) but have wildly inaccurate point estimates (SRE). Similarly, narrow credible regions benefit an estimator only if

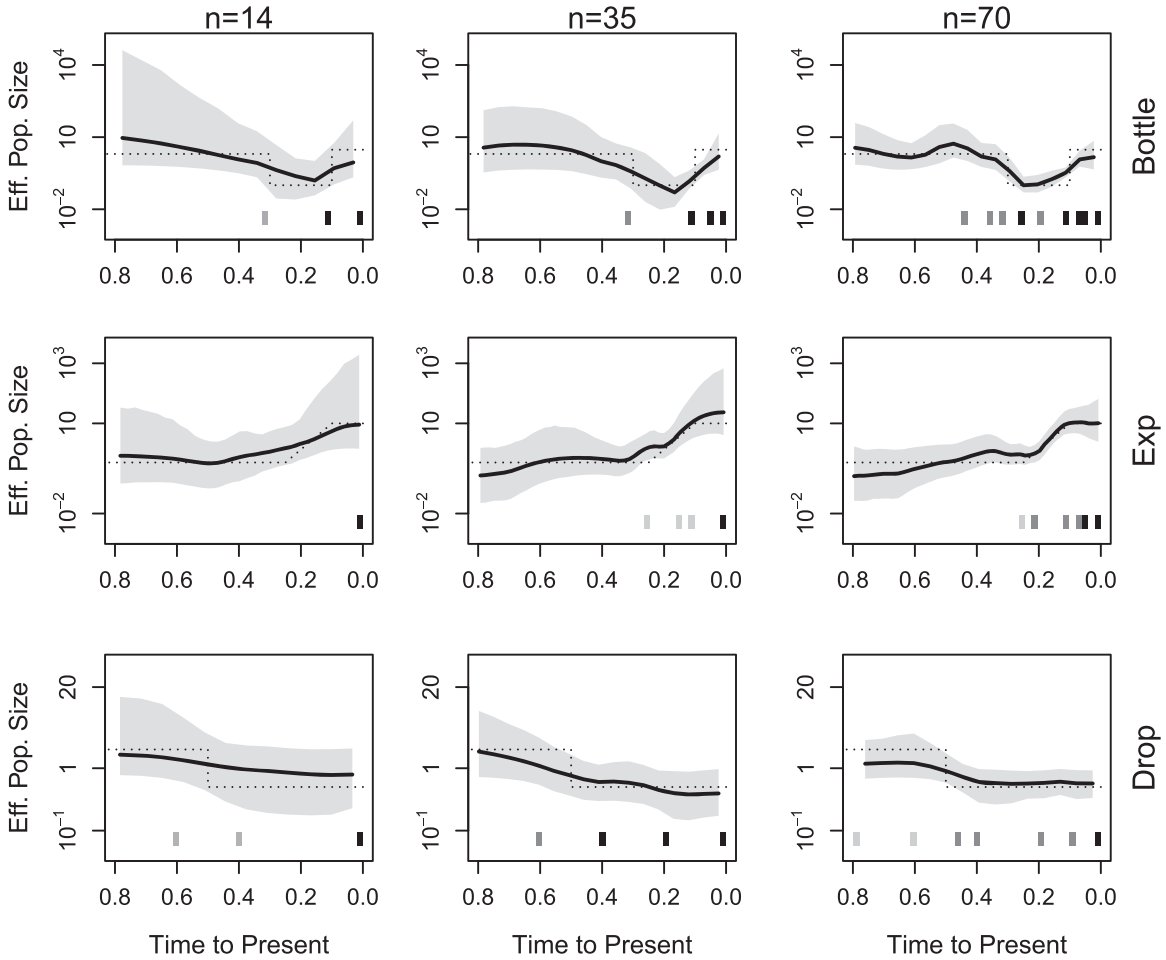


Figure 7. Simulation: effective population size posterior medians from different trajectories and sample sizes for the Tajima-based model. N_e posterior distribution from simulated data with three population size trajectories (rows)—bottleneck (Bottle), exponential growth (Exp) and instantaneous fall (Drop)—different sample sizes (columns)— $n = 14$, $n = 35$, and $n = 70$. Posterior medians are depicted as solid black lines and 95% Bayesian credible intervals are depicted by shaded areas. Dotted lines depict the ground truth. \mathbf{n} and \mathbf{s} are depicted by the heat maps at the bottom of each panel: the squares along the time axis indicate the sampling time, while the intensity of the black color depicts the number of samples. The y-axis is in logarithmic scale.

the estimates remain accurate. In this simulation study we fix $k = 100$, $\nu_1 = 0$ and $\nu_k = 0.6 t_2$.

Another common parameter estimated in coalescent inference is the time to most common ancestor (TMRCA), which corresponds to the largest coalescent time. We compare inferred TMRCA via Tajima and Kingman in the SM Section S13.

Results. Table S2 in SM section S8 summarizes the mean ESS for the 9 simulated datasets achieved with Tajima and Kingman. The high ESSs suggest convergence of the MCs. This is confirmed by the visual inspections of the trace plots (SM Section S8). Tajima has the highest ESS for $\log N_e$ in 5 out of 9 instances, Kingman in 3, and there is one tie. In 6 out of 9 instances, Tajima has the highest ESS for \mathbf{t} , Kingman in 2, and there is one tie. According to this metric, the Tajima chain appears to be more efficient, at least in the sense of achieving stationarity quicker. The results obtained for a fixed number of iterations suggest a more even performance (Tables S4 and S5 in SM section 10). A possible explanation for this result is offered by the larger burn-in used for the second study, suggesting that, in the stationary regime, the two chains have a comparable performance. However, we invite caution given the limitations already discussed and the fact that ESS is only a proxy for convergence.

With regards to N_e estimation, the results of the nine curves estimated with our method are plotted in Figure 7. The SM section 9 includes the plots for Kingman (Figure S9) and the BEAST-based methodologies (Figure S8). True trajectories are depicted as dashed lines, posterior medians as black lines, and 95% credible regions as gray shaded areas. Note that the y axis is logarithmic. Table 1 summarizes SRE, MRW, ENV, and the mean ESS for the 9 simulated datasets achieved with Tajima, Kingman, and “Oracle” for the fixed computational budget runs in all three scenarios.

The patterns are as predicted. As n increases, posterior medians track the true trajectories more closely. It is well known in the literature that abrupt population size changes are the most difficult to recover. The “drop” and “bottleneck” scenarios are less accurate for $n = 14$, as exhibited by the wider credible region. We recover the bottleneck (panel first row and first column), but we do not recover the instantaneous drop (panel first row and third column).

Table 1 quantifies the analysis of Figure 7. First, no method unequivocally outperforms the others. All methods have identical performance for the ENV metric, with the ENV metric decreasing as n increases in some cases and with both models. This is due to the fact that credible regions become

Table 1. Simulation: performance comparison between Tajima, Kingman, and Oracle models.

Label	n	%ENV			SRE			MRW		
		Oracle	Tajima	Kingman	Oracle	Tajima	Kingman	Oracle	Tajima	Kingman
Bottle	14	100	100	100	408.11	175.66	123	20164.85	2298.28	6241.1
	35	99	96	96	155.81	148.33	78.81	203.52	1385.86	148.73
	70	98	88	82	121.34	124.55	98.84	23.33	22.8	17.12
Drop	14	100	100	100	28.78	36.47	38.21	10.54	8.8	6.24
	35	99	96	93	21.27	31.73	67.69	2.96	6.02	24.78
	70	99	92	98	17.1	29.09	34.41	2.13	3.66	4.86
Exp	14	100	100	100	35.94	50.91	53.48	16.56	19.33	1163.38
	35	100	100	100	35.58	112.5	114.42	11.41	116.97	148.147
	70	100	100	100	30.71	43.16	37.31	3.64	3.97	2.75

NOTE: Envelope (ENV), sum of relative errors (SRE), and mean relative width (MRW) for three population trajectories (Bottle, Exp, Drop) and three sample sizes ($n = 14, 35, 70$). Tajima (our model), Kingman (Kingman n -coalescent), Oracle (Palacios and Minin 2012) (known **g**). Bold depicts the method with the best performance (excluding the “oracle”) or within 10% of the best performance. The MCMC was run until convergence.

much narrower, as indicated by MRW. With respect to SRE and MRW, our method has a superior performance in the “drop” and “exp” scenarios, while Kingman-based inference is superior in the “Bottle” scenario. We note that the Tajima methodology is the one that more closely tracks the “Oracle” results (in eight out of nine cases, Tajima has the closest SRE and MRW to the Oracle). We consider this a positive feature given that “Oracle” posterior does not account for the uncertainty in **g** and could be interpreted as a benchmark performance. Surprisingly, both Tajima and Kingman outperform the “Oracle” methodology in certain examples. Finally, in general as pointed out in Palacios et al. (2019) and Cappello and Palacios (2020), and SM Section 6, we would expect Tajima to outperform Kingman in regimes with low mutation rate. In our simulations, Tajima outperforms Kingman in the scenarios with the smallest number of observed mutations (Drop and Exp).

7. North American Bison Data

In this section, we study the long-standing question of whether the population decline of steppe Berigian bison was instigated by human intervention (the overkill hypothesis; Martin 1973) or by environmental changes. Using ancient DNA bison sequences, Shapiro et al. (2004) estimated the start of the population decline to be between 32 and 43 thousand year ago (kya). At the time of the study, the prevailing consensus was that human entered the Americas about 13 kya (“Clovis-first” model, Meltzer 2015). Based on the “Clovis-first” model, Shapiro et al. (2004) suggested that their estimate supported the environmental hypothesis, given that there was not a sufficiently large human population that could have caused the decline. In particular, they hypothesize that the decline may be due to abnormal environmental events preceding the last glacial maxima (LGM), which happened between 25 and 19 kya (Clark et al. 2009). There has been mounting evidence supporting a migration to the Americas preceding that leading to the Clovis population. Latest data show that humans were probably present already before and during the LGM, even if the major expansion happened after the LGM (Becerra-Valdivia and Higham 2020). The authors suggest that studies confuting

the human-driven extinction hypothesis should be thus revisited.

In light of the new discoveries, we deem relevant to reproduce Shapiro et al. (2004) analysis with new bison data recently presented by Froese et al. (2017). Our aim is to assess whether the timing of start of the decline is confirmed by this new data. To our knowledge, there is no phylodynamics analysis of this dataset in the literature. The data differs from that of Shapiro et al. (2004): Shapiro et al. (2004) sequences include 602 base pairs from the mitochondrial control region, while Froese et al. (2017) provide the full mitochondrial genome (16322 base pairs after alignment). We analyzed 38 sequences (10 modern, 28 ancient). Details on the dataset and models used are given in the SM S14

Figure 8 depicts the posterior medians (solid lines) of N_e along with the 95% credible bands (dashed lines) obtained from posterior samples by sampling Tajima’s trees (Tajima, blue colored lines) and Kingman’s trees (GMRF, red colored lines). Both, our method and GMRF, recover the pattern reconstructed by Shapiro et al. (2004): a population expansion followed by a decline. The population decline recovered by GMRF is somewhat sharper than that with Tajima. The two estimates of population expansions are roughly identical, and the credible bands practically overlap. The two methods’ estimates of the population decline are essentially identical: GMRF median time estimate is 29.6 kya, while the median time estimate for our method is 29.7 kya. The estimates obtained analyzing the 2017 data differ substantially from those obtained by analyzing the 2004 data (32–43 kya).

The estimates are also closer to the LGM and recent proofs of human presence in Eastern Beringia (Bluefish Caves, 24 kya, Bourgeon, Burke, and Higham 2017), and the whole Northern and Central America (Ardelean et al. 2020; Bennett et al. 2021). Our results support the Becerra-Valdivia and Higham (2020) call to revisit the question of whether humans have contributed to the extinction of certain animal species, like the steppe Beringian bison. We note that Shapiro et al. (2004) discussed the possibility of human presence before the LGM. At the time of their work, there was an active debate on the date of molecular specimens in the Bluefish Caves. What remains an open question is understanding the number of humans living in the Americas before the LGM.

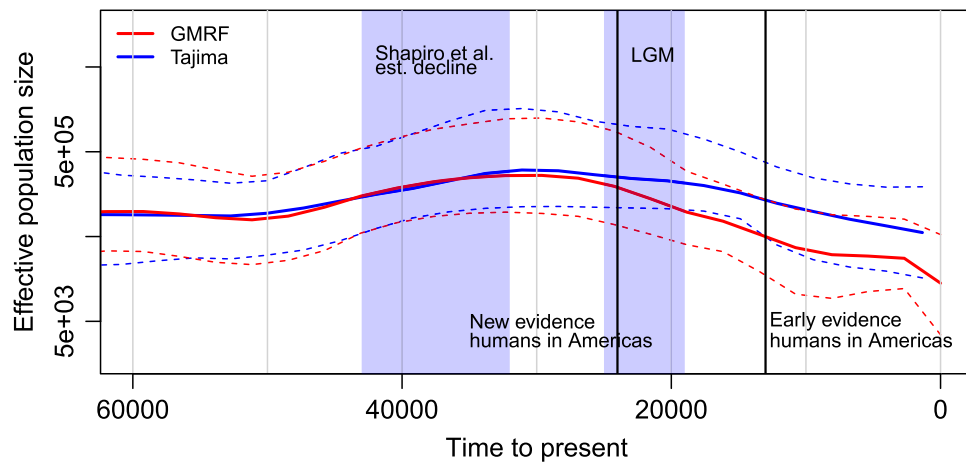


Figure 8. Bison in North America: posterior median estimates from Froese et al. (2017) dataset. Solid lines display estimated posterior medians of N_e obtained from $n = 38$ ancient and modern sequences from North America specimens in Froese et al. (2017) data. Blue depicts our method, and red GMRF. The 95% Bayesian credible region boundaries are depicted by dashed lines. Annotations refer to events of interest commented in the text. Purple shaded area annotations are in the upper part of the panel. The y-axis is in logarithmic scale.

8. Discussion

We have studied an alternative to the Kingman n -coalescent to infer the population size trajectory from serially sampled sequences collected at non-recombining loci. We develop a new efficient algorithm to compute the likelihood of partially labeled ranked trees, and discuss the advantages of using the proposed lower-resolution coalescent process.

Despite the evidence presented, our simulations did not produce irrefutable proof of the advantages of employing such coarsening in terms of faster MCMC convergence. One of the challenges is the fairly comparison of the chains with two different target distributions, $\pi(\mathbf{g}, N_e | \mathbf{Y}, \mu)$ and $\pi(\mathbf{g}^K, N_e | \mathbf{Y}^{lab}, \mu)$. Nevertheless, our goal is not to show the superiority due to faster mixing MC (we tried to design the two proposals as similar as possible), but rather due to some intrinsic properties of the coarsened state space. We showed that in regimes of low mutation rate and large sample size, we tend to perform better. We expect evidence of further gains in simulations with larger sample sizes. Our current implementation in R does not handle sample sizes larger than a couple of hundred sequences and further work needs to be done to optimize current implementation. Further, we impose some limiting assumptions. Accommodating departure from the ISM model and modeling recombination are priority for future work.

Supplementary Materials

The supplementary material file contains the following sections: (S1) Examples of likelihood calculations under Kingman and Tajima coalescent models; and variance calculations of the examples in the main text. (S2) Algorithm for generating the augmented perfect phylogeny. (S3) Mapping of nodes in T to subtrees of g. (S4) Algorithm for Allocation matrix. (S5) Details of the MCMC implementations. (S6) Counting the number of compatible tree topologies under the ISM with heterochronous data. (S7) Simulation details. (S8) Checking the validity of the implementation. (S9) Comparison of Tajima-based inference with state-of-the-art alternatives. (S10) Simulations: fixed number of iterations. (S11) Simulations with multiple loci. (S12) Simulations with unknown mutation rate. (S13) Simulations: time to the most common ancestor. (S14) North American Bison. We detail the dataset and analysis.

We also include a folder with all the codes necessary to reproduce the analysis. For using the method, we suggest looking at the R package *phylodyn* cited in the manuscript.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

The authors gratefully acknowledge partial funding from the France-Stanford Center for Interdisciplinary Studies. JAP acknowledges support from NSF Career award 2143242, NIH R35GM14833801 and the Alfred P. Sloan Foundation. AV acknowledges partial funding from the chaire program Mathematical Modeling and Biodiversity (Ecole polytechnique, Museum National d’Histoire Naturelle, Veolia Environment, Fondation X).

ORCID

Lorenzo Cappello  <https://orcid.org/0000-0001-6682-908X>

Amandine Véber  <https://orcid.org/0000-0003-4464-015X>

Julia A. Palacios  <https://orcid.org/0000-0003-4501-7378>

References

- Ardelean, C. F., Becerra-Valdivia, L., Pedersen, M. W., Schwenninger, J.-L., Oviatt, C. G., Macías-Quintero, J. I., Arroyo-Cabrales, J., Sikora, M., Ocampo-Díaz, Y. Z. E., Rubio-Cisneros, I. I., et al. (2020), “Evidence of Human Occupation in Mexico around the Last Glacial Maximum,” *Nature*, 584, 87–92. [11]
- Becerra-Valdivia, L., and Higham, T. (2020), “The Timing and Effect of the Earliest Human Arrivals in North America,” *Nature*, 584, 93–97. [1,11]
- Bennett, M. R., Bustos, D., Pigati, J. S., Springer, K. B., Urban, T. M., Holliday, V. T., Reynolds, S. C., Budka, M., Honke, J. S., Hudson, A. M., et al. (2021), “Evidence of Humans in North America During the Last Glacial Maximum,” *Science*, 373, 1528–1531. [11]
- Bouchard-Côté, A., Sankararaman, S., and Jordan, M. I. (2012), “Phylogenetic Inference via Sequential Monte Carlo,” *Systematic Biology*, 61, 579–593. [2]
- Bourgeon, L., Burke, A., and Higham, T. (2017), “Earliest Human Presence in North America Dated to the Last Glacial Maximum: New Radiocarbon Dates from Bluefish Caves, Canada,” *PLoS One*, 12, e0169486. [1,11]

- Cappello, L., Kim, J., Liu, S., and Palacios, J. A. (2022), “Statistical Challenges in Tracking the Evolution of SARS-CoV-2,” *Statistical Science*, 37, 162–182. [1]
- Cappello, L., and Palacios, J. A. (2020), “Sequential Importance Sampling for Multi-Resolution Kingman-Tajima Coalescent Counting,” *Annals of Applied Statistics*, 14, 727–751. [11]
- Clark, P. U., Dyke, A. S., Shakun, J. D., Carlson, A. E., Clark, J., Wohlfarth, B., Mitrovica, J. X., Hostetler, S. W., and McCabe, A. M. (2009), “The Last Glacial Maximum,” *Science*, 325, 710–714. [11]
- Dinh, V., Bilge, A., Zhang, C., and Matsen IV, F. A. (2017), “Probabilistic Path Hamiltonian Monte Carlo,” in *International Conference on Machine Learning*, pp. 1009–1018. [2]
- Dinh, V., Darling, A. E., and Matsen IV, F. A. (2017), “Online Bayesian Phylogenetic Inference: Theoretical Foundations via Sequential Monte Carlo,” *Systematic Biology*, 67, 503–517. [2]
- Disanto, F., and Wiehe, T. (2013), “Exact Enumeration of Cherries and Pitchforks in Ranked Trees under the Coalescent Model,” *Mathematical biosciences*, 242, 195–200. [2]
- Drummond, A. J., Nicholls, G. K., Rodrigo, A. G., and Solomon, W. (2002), “Estimating Mutation Parameters, Population History and Genealogy Simultaneously from Temporally Spaced Sequence Data,” *Genetics*, 161, 1307–1320. [2,7]
- Drummond, A., Suchard, M., Xie, D., and Rambaut, A. (2012), “Bayesian Phylogenetics with BEAUti and the BEAST 1.7,” *Molecular Biology and Evolution*, 29, 1969–1973. [9]
- Faulkner, J. R., Magee, A. F., Shapiro, B., and Minin, V. N. (2020), “Horseshoe-based Bayesian Nonparametric Estimation of Effective Population Size Trajectories,” *Biometrics*, 76, 677–690. [7]
- Fourment, M., Claywell, B. C., Dinh, V., McCoy, C., Matsen IV, F. A., and Darling, A. E. (2017), “Effective Online Bayesian Phylogenetics via Sequential Monte Carlo with Guided Proposals,” *Systematic Biology*, 67, 490–502. [2]
- Froese, D., Stiller, M., Heintzman, P. D., Reyes, A. V., Zazula, G. D., Soares, A. E., Meyer, M., Hall, E., Jensen, B. J., Arnold, L. J. et al. (2017), “Fossil and Genomic Evidence Constrains the Timing of Bison Arrival in North America,” *Proceedings of the National Academy of Sciences*, 114, 3457–3462. [1,3,11,12]
- Griffiths, R. C., and Tavaré, S. (1994), “Sampling Theory for Neutral Alleles in a Varying Environment,” *Philosophical Transactions of the Royal Society of London, Series B*, 344, 403–410. [5]
- Griffiths, R., and Tavaré, S. (1995), “Unrooted Genealogical Tree Probabilities in the Infinitely-Many-Sites Model,” *Mathematical Biosciences*, 127, 77–98. [5,7]
- Gusfield, D. (1991), “Efficient Algorithms for Inferring Evolutionary Trees,” *Networks*, 21, 19–28. [5]
- Ho, S. Y. and Shapiro, B. (2011), “Skyline-Plot Methods for Estimating Demographic History From Nucleotide Sequences,” *Molecular Ecology Resources*, 11, 423–434. [1]
- Kimura, M. (1969), “The Number of Heterozygous Nucleotide Sites Maintained in a Finite Population Due to Steady Flux of Mutations,” *Genetics*, 61, 893–903. [5]
- Kingman, J. F. (1982a), “On the Genealogy of Large Populations,” *Journal of Applied Probability*, 19 27–43. [1]
- Kingman, J. F. C. (1982b), “The Coalescent,” *Stochastic Processes and their Applications*, 13, 235–248. [1]
- Lan, S., Palacios, J. A., Karcher, M., Minin, V. N., and Shahbaba, B. (2015), “An Efficient Bayesian Inference Framework for Coalescent-based Nonparametric Phylodynamics,” *Bioinformatics*, 31, 3282–3289. [7]
- Markovtsova, L., Marjoram, P., and Tavaré, S. (2000), “The Age of a Unique Event Polymorphism,” *Genetics*, 156, 401–409. [9]
- Martin, P. S. (1973), “The Discovery of America the First Americans May Have Swept the Western Hemisphere and Decimated Its Fauna within 1000 Years,” *Science*, 179, 969–974. [11]
- Meltzer, D. J. (2015), *The Great Paleolithic War*, Chicago, IL: University of Chicago Press. [11]
- Palacios, J. A., and Minin, V. N. (2012), “Integrated Nested Laplace Approximation for Bayesian Nonparametric Phylodynamics,” in *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence, UAI’12*, pp. 726–735, Arlington, VA: AUAI Press. [9,11]
- Palacios, J. A., and Minin, V. N. (2013), “Gaussian Process-based Bayesian Nonparametric Inference of Population Size Trajectories from Gene Genealogies,” *Biometrics*, 69, 8–18. [7]
- Palacios, J. A., Véber, A., Cappello, L., Wang, Z., Wakeley, J., and Ramachandran, S. (2019), “Bayesian Estimation of Population Size Changes by Sampling Tajima’s Trees,” *Genetics*, 213, 967–986. [2,4,5,6,7,8,9,11]
- Parag, K. V., and Pybus, O. G. (2019), “Robust Design for Coalescent Model Inference,” *Systematic Biology*, 68, 730–743. [2]
- Rodrigo, A. G., Ewing, G., and Drummond, A. (2007), “The Evolutionary Analysis of Measurably Evolving Population Using Serially Smpled Gene Sequences,” in *Reconstructing Evolution: New Mathematical and Computational Advances*, eds. O. Gascuel and M. Steel, pp. 233–272, Oxford: Oxford University Press. [2]
- Rodrigo, A. G., and Felsenstein, J. (1999), “Coalescent Approaches to HIV Population Genetics,” in *The Evolution of HIV*, pp. 233–272, Baltimore, MD: Johns Hopkins University Press. [2,3,4]
- Rue, H., Martino, S., and Chopin, N. (2009), “Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations,” *Journal of the Royal Statistical Society, Series B*, 71, 319–392. [9]
- Sainudiin, R., Stadler, T., and Véber, A. (2015), “Finding the Best Resolution for the Kingman-Tajima Coalescent: Theory and Applications,” *Journal of Mathematical Biology*, 70, 1207–1247. [2,4]
- Sanderson, M. J., McMahon, M. M., Stamatakis, A., Zwickl, D. J., and Steel, M. (2015), “Impacts of Terraces on Phylogenetic Inference,” *Systematic Biology*, 64, 709–726. [1]
- Shahbaba, B., Lan, S., Johnson, W. O., and Neal, R. M. (2014), “Split Hamiltonian Monte Carlo,” *Statistics and Computing*, 24, 339–349. [7]
- Shapiro, B., Drummond, A. J., Rambaut, A., Wilson, M. C., Matheus, P. E., Sher, A. V., Pybus, O. G., Gilbert, M. T. P., Barnes, I., Binladen, J. et al. (2004), “Rise and Fall of the Beringian Steppe Bison,” *Science*, 306, 1561–1565. [1,11]
- Simper, M., and Palacios, J. A. (2022), “An Adjacent-Swap Markov Chain on Coalescent Trees,” *Journal of Applied Probability*, 59, 1243–1260. [1]
- Speidel, L., Forest, M., Shi, S., and Myers, S. R. (2019), “A Method for Genome-Wide Genealogy Estimation for Thousands of Samples,” *Nature Genetics*, 51, 1321–1329. [1]
- Tajima, F. (1983), “Evolutionary Relationship of DNA Sequences in Finite Populations,” *Genetics*, 105, 437–460. [2]
- Ushey, K., Allaire, J., and Tang, Y. (2022), *reticulate: Interface to ‘Python’*, available at <https://rstudio.github.io/reticulate/>, <https://github.com/rstudio/reticulate>. [9]
- Volz, E. M., Koelle, K., and Bedford, T. (2013), “Viral Phylodynamics,” *PLoS Computational Biology*, 9, e1002947. [1]
- Wang, L., Bouchard-Côté, A., and Doucet, A. (2015), “Bayesian Phylogenetic Inference Using a Combinatorial Sequential Monte Carlo Method,” *Journal of the American Statistical Association*, 110, 1362–1374. [2]
- Watterson, G. (1975), “On the Number of Segregating Sites in Genetical Models without Recombination,” *Theoretical Population Biology*, 7, 256–276. [5]
- Zhang, C., and Matsen IV, F. A. (2018), “Variational Bayesian Phylogenetic Inference,” in *International Conference on Learning Representations*. [2]