Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/uasa20

Exact Decoding of a Sequentially Markov Coalescent Model in Genetics

Caleb Ki & Jonathan Terhorst

To cite this article: Caleb Ki & Jonathan Terhorst (29 Aug 2023): Exact Decoding of a Sequentially Markov Coalescent Model in Genetics, Journal of the American Statistical Association, DOI: 10.1080/01621459.2023.2252570

To link to this article: https://doi.org/10.1080/01621459.2023.2252570

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2023, VOL. 00, NO. 0, 1–14: Theory and Methods
https://doi.org/10.1080/01621459.2023.2252570

Exact Decoding of a Sequentially Markov Coalescent Model in Genetics

Caleb Ki and Jonathan Terhorst

Department of Statistics, University of Michigan, Ann Arbor, MI

ABSTRACT

In statistical genetics, the sequentially Markov coalescent (SMC) is an important family of models for approximating the distribution of genetic variation data under complex evolutionary models. Methods based on SMC are widely used in genetics and evolutionary biology, with significant applications to genotype phasing and imputation, recombination rate estimation, and inferring population history. SMC allows for likelihood-based inference using hidden Markov models (HMMs), where the latent variable represents a genealogy. Because genealogies are continuous, while HMMs are discrete, SMC requires discretizing the space of trees in a way that is awkward and creates bias. In this work, we propose a method that circumvents this requirement, enabling SMC-based inference to be performed in the natural setting of a continuous state space. We derive fast, exact procedures for frequentist and Bayesian inference using SMC. Compared to existing methods, ours requires minimal user intervention or parameter tuning, no numerical optimization or E-M, and is faster and more accurate. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received October 2020
Accepted August 2023

KEYWORDS
Changepoint; Coalescent; Hidden Markov model; Population genetics

1. Introduction

Probabilistic models of evolution have played a central role in genetics since the inception of the field a century ago. Beginning with foundational work by Ronald Fisher and Sewall Wright, and continuing with important contributions from P.A.P. Moran, Motoo Kimura, J.F.C. Kingman, and many others, a succession of increasingly sophisticated stochastic models were developed to describe patterns of ancestry and genetic variation found in a population. Statisticians harnessed these models to analyze genetic data, initially with the now quaint-seeming goal of understanding the evolution of a single gene. More recently, as next-generation sequencing has enabled the collection of genome-wide data from millions of people, interest has risen in methods for studying evolution using large numbers of whole genomes.

In this article, we study a popular subset of those methods which are likelihood-based; that is, these methods work by inverting a statistical model that maps evolutionary parameters to a probability distribution over genetic variation data. As we will see, exact inference in this setting is impossible owing to the need to integrate out a high-dimensional latent variable which encodes the genome-wide ancestry of every sampled individual. Consequently, a number of approximate methods have been proposed, which try to strike a balance between biological realism and computational tractability.

We focus on one such approximation known as the *sequentially Markov coalescent* (SMC). The sequential or "spatial" formulation of the coalescent was first derived by Wiuf and Hein (1999), and based on their ideas McVean and Cardin (2005) described an efficient Markovian algorithm for performing inference under a coalescent model with recombination. Although the term SMC is often used to refer to McVean and Cardin's original algorithm, there are actually many methods in the literature that are simultaneously (a) sequential, (b) Markov, and (c) approximations of the coalescent with recombination (McVean and Cardin 2005; Marjoram and Wall 2006; Carmi et al. 2014; Hobolth and Jensen 2014). In this article, we therefore use SMC more generally to refer to any method that meets these criteria. In particular, both the influential haplotype copying model of Li and Stephens (2003) and the popular program PSMC (Li and Durbin 2011) for inferring population history are in the family of SMC methods under this definition (Paul and Song 2010).

SMC models lead quite naturally to the use of hidden Markov models (HMMs) to analyze genetic sequence data. However, in order to bring the HMM machinery to bear on this problem, additional and somewhat awkward assumptions are needed. The latent variable in an HMM must have finite support, whereas the latent variable in SMC is a continuous tree. Therefore, the space of trees must be discretized, and, in some cases, restrictions must also be placed on the topology of each tree. In applications, the user must select a discretization scheme, a non-obvious choice which nonetheless has profound consequences for downstream inference (Parag and Pybus 2019).

The main message of our article is that this is not necessary: it is possible to solve a form of the sequentially Markov coalescent exactly, in its natural setting of continuous state space. We accomplish this by slightly modifying the canonical SMC model of McVean and Cardin (2005), in a way that does not greatly impact inference, but renders the problem theoretically and

CONTACT Jonathan Terhorst *jonth@umich.edu* Department of Statistics, University of Michigan, Ann Arbor, MI.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

computationally much easier. In particular, this modification allows us to leverage recent innovations in changepoint detection, leading to algorithms which are computationally efficient and have reduced bias. Of course, some tradeoffs are necessary in order to achieve this: we must place some restrictions on the types of priors that can be used to model the instantaneous rate of coalescence, and, in contrast to existing approaches, the asymptotic running time of our algorithm is not known to us exactly. These restrictions, and their implications for inference, are explored in greater detail below.

The rest of the article is organized as follows. In Section 2 we formally define our data and model, introduce notation, and survey related work. In Section 3 we derive our main results: exact and efficient Bayesian and frequentist algorithms for inferring genealogies from genetic variation data. In Section 4 we thoroughly benchmark our method, compare it to existing approaches, and provide an application to real data analysis. We provide concluding remarks in Section 5.

2. Background

In this section we introduce notation, formalize the problem we want to solve, and survey earlier work. We presume some familiarity with standard terminology and models in genetics; introductory texts include Hein, Schierup, and Wiuf (2005) and Durrett (2008).

2.1. Motivation

Our method aims to infer a sequence of latent genealogies using genetic variation data. To motivate our interest in this, consider first a related problem with a more direct scientific application: given a matrix of DNA sequence data $\mathbf{Y} \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}^{H \times N}$ from $H > 1$ homologous chromosomes each N base pairs long, and an evolutionary model φ hypothesized to have generated these data, find the likelihood $p(\mathbf{Y} \mid \varphi)$. This generic formulation encompasses a wide variety of inference problems in genetics and evolutionary biology; if we could easily solve it, important new scientific insights would result.

Unfortunately, this is not possible using current methods. The difficulty lies in the fact that the relationship between the data $\mathbf{Y}$ and the scientifically interesting quantity φ is mediated through a complex, latent combinatorial structure known as the ancestral recombination graph (ARG; Griffiths and Marjoram 1997), which encodes the genealogical relationships between every sample at every position in the genome. The ARG is sufficient for φ: evolution generates the ARG, and conditional on it, the data contain no further information about φ. Thus, the likelihood problem requires the integration

$$p(\mathbf{Y} \mid \varphi) = \int_{A \in \mathcal{A}} p(\mathbf{Y} \mid A) p(A \mid \varphi), \tag{1}$$

where A denotes an ARG, and $\mathcal{A}$ denotes the support set of ARGs for a sample of H chromosomes. This is a very challenging integral; although a method for evaluating it is known (Griffiths and Marjoram 1996), it only works for small datasets. That is because, for large N and H, there are a huge number of ARGs that could have plausibly generated a given dataset, such that the complexity of $\mathcal{A}$ explodes as N and H grow. Indeed, (1) cannot be computed for chromosome-scale data even for the simplest case $H = 2$.

The sequentially Markov coalescent addresses this problem by decomposing the ARG into a sequence of marginal gene trees $X_1, \ldots, X_N$, one for each position in the chromosome, and supposing that this sequence is Markov. Then, we have

$$p(\mathbf{Y} \mid \varphi) \approx \int_{X_1, \ldots, X_N} \pi(X_1 \mid \varphi) p(\mathbf{Y}_1 \mid X_1) \prod_{n=2}^{N} p(\mathbf{Y}_n \mid X_n) p(X_n \mid X_{n-1}, \varphi), \tag{2}$$

where $\pi(\cdot \mid \varphi)$ is a stationary distribution for the Markov chain $X_1, \ldots, X_N$, the transition density $p(X_n \mid X_{n-1}, \varphi)$ governs the transition from one marginal tree to the next, and $[\mathbf{Y}_1 \mid \cdots \mid \mathbf{Y}_N] = \mathbf{Y}$ are the data at each site.

Even under the Markov assumption, the integral (2) is challenging, since each X_i represents a genealogy. To make the problem tractable, existing methods further assume that these genealogies have special structure. For example, in the widely used program PSMC (Li and Durbin 2011), each "genealogy" has only two leaves, representing the ancestry of a pair of homologous chromosomes, so each $X_i \in \mathbb{R}_{\geq 0}$ can be taken to be a real number representing the height of the corresponding tree. The problem is then further simplified by discretizing time, such that the height of each tree falls into one of a pre-specified collection of discrete intervals. Similarly, the foundational Li-Stephens copying model (Li and Stephens 2003) allows for more than two chromosomes to be analyzed, but assumes that the tree height is fixed to a single, pre-specified value and has a distinctive, "forest-of-trunks" structure (Paul and Song 2010). In both cases, once the state space of the X_i has been made finite, inference methods for hidden Markov models can be employed. Typically these are used to infer φ via the EM or Baum-Welch algorithm, which requires computing the posterior distribution

$$p(X_1, \ldots, X_N \mid \mathbf{Y}, \varphi). \tag{3}$$

2.2. Applications of the Sequentially Markov Coalescent

A number of noteworthy methods in statistical genetics and evolutionary biology depend on this model. Among the most widely used are methods for performing phasing and imputation (Scheet and Stephens 2006; Marchini et al. 2007; Howie, Donnelly, and Marchini 2009). Imputation methods leverage the fact that closely related members of a population tend to share genetic material to fill in missing genotype calls, and are an essential pre-processing step for improving power in genomewide association studies (Huang et al. 2015; Rubinacci et al. 2021). Phasing seeks to resolve diploid genotype calls, which, for technological reasons, are cheapest and easiest to produce, into constituent haplotypes. Phased haploid data is a necessary precursor for most evolutionary studies, and is also used to improve imputation accuracy (Howie et al. 2012). Crucially, through their underlying use of the Li-Stephens haplotype copying model (Li and Stephens 2003), most existing phasing and imputation methods rely on accurate posterior estimates of

local ancestry, $p(X_i \mid \mathbf{Y}, \varphi)$ in the notation of (3). We discuss this connection in further in Section 4.4.2.

Haplotype copying models are also directly used to study evolution, for example to estimate rates of recombination and gene conversion (Li and Stephens 2003; Gay, Myers, and McVean 2007; Chan, Jenkins, and Song 2012), to detect signatures of recent positive selection (Voight et al. 2006; Palamara et al. 2018), or to infer local ancestry (Price et al. 2009; Lawson et al. 2012). These methods aim to fit a particular evolutionary model φ to data using, essentially, (2) and (3), and frequentist or Bayesian model fitting procedures. For example, in their original paper Li and Stephens defined φ to be a sequence of local recombination rates (which enter into the likelihood through the transition density $p(X_n \mid X_{n-1}, \varphi)$ in (2)) and estimated $\widehat{\varphi}$ using the EM algorithm. Similarly, Palamara et al. (2018) compared local posterior distributions $p(X_i \mid \mathbf{Y}, \varphi)$, where i indexes a particular location in the genome, to a genomewide null distribution in order to detect signatures of local adaptation within the last $\sim 10^4$ years.

A problem of particular interest is so-called *demographic inference* (Spence et al. 2018), where φ represents historical fluctuations in population size. In this case, we can identify φ with a function $N_e : [0, \infty) \to (0, \infty)$, such that $N_e(t)$ is the coalescent effective population size t generations before the present (Durrett 2008, sec. 4.4). This function governs the marginal distribution of coalescence time at a particular locus in a sample of two chromosomes. Specifically, setting $\eta(t) = 1/N_e(t)$, the density of this time is

$$\pi(t) = \eta(t) e^{-\int_0^t \eta(s)\, ds}. \tag{4}$$

Note that $\eta(t) = 1$ recovers the well-known case of Kingman's coalescent, $\pi(t) = e^{-t}$, which we treat as the default prior in what follows.

Apart from intrinsic interest in learning population history, it is important to get a sharp estimate of $N_e(t)$ as unmodeled variability in $N_e(t)$ confound attempts to study some of evolutionary phenomena mentioned above, such as natural selection, or mutation rate variation. Many demographic inference methods have been proposed, using various underlying models and sources of data. One class (Gutenkunst et al. 2009; Bhaskar, Wang, and Song 2015; Jouganous et al. 2017; Kamm, Terhorst, and Song 2017; Kamm et al. 2020) infers demographic history using so-called site frequency spectrum data, which is a low-dimensional summary statistic that is computed from mutation data assuming free recombination between markers. A second class of models, which includes ours, are designed to analyze whole-genome sequence data, and extract additional demographic signal from patterns of linkage disequilibrium. These methods are usually based on some form of the sequentially Markov coalescent (Li and Durbin 2011; Sheehan, Harris, and Song 2013; Rasmussen et al. 2014; Terhorst, Kamm, and Song 2017; Schiffels and Durbin 2014; Steinrücken et al. 2019). Another recent development is the emergence of algorithms for inferring complete ancestral recombination graphs using large amounts of sequence data (Speidel et al. 2019; Kelleher et al. 2019), from which the demographic history can be estimated. Finally, there has been significant parallel work in phylogenetics on so-called *skyline models*, which are Bayesian procedures designed to infer population history under the assumption of a nonrecombining genealogy (Pybus, Rambaut, and Harvey 2000; Drummond et al. 2005; Minin, Bloomquist, and Suchard 2008; Gill et al. 2013).

2.3. Our Contribution

As discussed in Section 1, discretizing X_i is unnatural and results in bias. In this work, we derive efficient methods for computing the posterior distribution $p(X_1, \ldots, X_N \mid \mathbf{Y})$, or its "maximum a posteriori" estimate

$$\arg\max_{X_1,\ldots,X_N} p(X_1, \ldots, X_N \mid \mathbf{Y})$$

when each X_i is a tree with continuous branch lengths. (To simplify the formulas, we suppress dependence on the evolutionary model φ until turning to inference in Section 4.4.) That is, unlike existing methods, we do not assume that the set of possible X_i is discrete or finite. For the important case of $H = 2$ chromosomes, our method is "exact" in the sense that it is devoid of further approximations (beyond the standard ones which we outline in the next section). In this case, the gene tree X_i is completely described by the coalescence time of the two chromosomes. For $H > 2$ our method makes additional assumptions about the topology of each X_i, but still retains the desirable property of operating in continuous time.

2.4. Notation and Model

We now fix necessary notation and define the model that is used to prove our results. For simplicity, we first focus on the case of analyzing just one pair of chromosomes ($H = 2$ in the notation of the previous section). In Section 3.4 we describe how to extend our results to larger sample sizes.

Assume that we have sampled a pair of homologous chromosomes each consisting of N non-recombining loci. Meiotic recombination occurs between loci with rate ρ per unit time, and does not occur within each locus. The number of generations backwards in time until the two chromosomes meet at a common ancestor (TMRCA) at locus i is denoted $X_i \in \mathbb{R}_{>0}$. The number of positions where the two chromosomes differ at locus i is denoted by Y_i. Under a standard assumption known as the infinite sites model (Durrett 2008, sec. 1.4), Y_i has the conditional distribution

$$Y_i \mid X_i \sim \text{Poisson}(\theta X_i),$$

where θ is the mutation rate. We assume that both θ and ρ are small. In particular, some of our proofs rely on the fact that $\rho \ll 1$. These are fairly mild assumptions which hold in many settings of interest. For example, in humans, the population-scaled rates of mutation and recombination per nucleotide are $O(10^{-4})$. Conversely, if recombinations are frequent, then there is little advantage in employing the methods we describe here, which depend on the presence of linkage disequilibrium between nearby loci.

The sequentially Markov coalescent is a generative model for the sequence $X_1, \ldots, X_N$, which we abbreviate as $X_{1:N}$ henceforth (and similarly for $Y_{1:N}$). SMC characterizes how shared

ancestry changes when moving from one locus to the next. Assuming there is at most one recombination between adjacent loci, and we can specify an SMC model by the conditional density

$$f_{X_{n+1}|X_n}(t \mid s) := p(X_{n+1} \in (t, t+dt) \mid X_n = s) = \delta(t-s)e^{-\rho s} + (1-e^{-\rho s})q(t \mid s), \quad (5)$$

where $\delta(\cdot)$ is the Dirac delta function, and $q(t \mid s)$ is the conditional density of t given that a recombination occurred and that the existing TMRCA equals s. Various proposals for $q(t \mid s)$ exist in the literature, each with slightly different properties (McVean and Cardin 2005; Marjoram and Wall 2006; Paul, Steinrücken, and Song 2011; Li and Durbin 2011; Carmi et al. 2014). Importantly, they share the common feature that (5) is (approximately, in the case of Li and Durbin 2011) reversible with respect to the coalescent. That is,

$$\pi(s) f_{X_{n+1}|X_n}(t \mid s) = \pi(t) f_{X_{n+1}|X_n}(s \mid t), \quad (6)$$

where π is the stationary measure in (4). This can be verified in each of the above models by checking the detailed balance condition (Hobolth and Jensen 2014).

2.5. Connection to Changepoint Detection

Our work is motivated by the observation that (5) is essentially a changepoint model. Indeed, SMC can be viewed as a prior over the space of piecewise constant functions spanning the interval $[0, N)$; conditional on realizing one such function, say $\xi : [0, N) \to [0, \infty)$, each $X_i = \xi(i-1)$, and the data $Y_{1:N}$ are independent Poisson draws with mean $\mathbb{E}(Y_i \mid X_i) = \theta X_i$. In genetics, each contiguous segment where $X_i = X_{i+1} = \cdots = X_{i+k-1} = \tau$, say, is known as an *identity by descent* (IBD) tract, with *time to most recent common ancestor* (TMRCA) τ; the flanking positions where $X_{i-1} \neq X_i$ and $X_{i+k} \neq X_{i+k-1}$ are called *recombination breakpoints*. In changepoint detection, these are called *segments*, *segment heights* (or just heights), and *changepoints*, respectively. In what follows, we use these terms interchangeably depending on what is most descriptive in a given context.

A common assumption in changepoint detection is that neighboring segment heights are independent, which is to say that $X_i \perp X_{i+1}$ for any i such that $X_i \neq X_{i+1}$. As we will see, this enables fast and accurate algorithms for inferring the sequence $X_{1:N}$. SMC violates this assumption through the conditional density $q(t \mid s)$: the correlation between t and s in (5) makes the problem nonstandard from a changepoint perspective. Although there has been recent work on detecting changepoints in data with dependence between segments (e.g., Fearnhead and Liu 2011; Chan et al. 2021; Shi et al. 2022), particularly in time series, to the best of our knowledge the running time of these methods scales at least quadratically in the length of the underlying sequence.

In our application, sequence length is extremely long (a typical genetic sequence contains millions of observations), so methods with linear running time are essential. Perhaps the simplest way to achieve this is to approximate prior evolutionary model by one which ignores correlations in segment height. Indeed, if $q(t \mid s)$ were replaced by some function $\underline{\pi}(t)$ which did not depend on s, then (5) would become a so-called product partition model (PPM; Barry and Hartigan 1992). In a PPM, a sequence of observations $y_1, \ldots, y_n$ is randomly partitioned into disjoint blocks $(y_1, \ldots, y_{b_1}), (y_{b_1+1}, \ldots, y_{b_2}), \ldots, (y_{b_{k-1}+1}, \ldots, y_{b_k})$, such that the observations in each block are independent of all others. In the identity-by-descent problem described above, each block corresponds to an IBD segment, and the random partition has break points wherever recombinations occurred. PPMs are well-understood, and linear-time approximate methods have been developed to analyze them in both Bayesian (Barry and Hartigan 1993; Fearnhead 2006) and frequentist (Jackson et al. 2005; Killick, Fearnhead, and Eckley 2012) settings.

2.6. A Renewal Approximation

In biological applications, the orientation of the data sequence $Y_{1:N}$ is arbitrary; we could equivalently work with the reversed sequence $Y_N, Y_{N-1}, \ldots, Y_1$ instead. Additionally, both theoretical and empirical evidence overwhelmingly support that Kingman's coalescent is a robust and accurate description of ancestry at a particular gene. For these reasons, it is important that any SMC model maintain the detailed balance condition (6). Given this desideratum, the obvious choice for $\underline{\pi}$ becomes

$$\underline{\pi}(t) \propto t\pi(t), \quad (7)$$

leading to the modified transition density

$$f^R_{X_{n+1}|X_n}(t \mid s) = \delta(t-s)e^{-\rho s} + (1-e^{-\rho s})\underline{\pi}(t). \quad (8)$$

Checking the detailed balance condition (6), we obtain

$$\pi(s)(1-e^{-\rho s})t\pi(t) \stackrel{?}{=} \pi(t)(1-e^{-\rho t})s\pi(s), \quad s \neq t. \quad (9)$$

Though (9) is not true in general, equality holds when both sides are expanded to first-order in ρ, which suffices for the applications we consider here.

The renewal approximation preserves an important piece of prior information concerning the nature of identity-by-descent: an IBD tract with TMRCA x experiences recombination at rate ρx, so more recent tracts are longer, a familiar fact to geneticists. On the other hand, prior information on the correlation between neighboring segment heights is dropped. We hypothesized that, for inference, it is more important that the prior capture the former effect than the latter. This is similar to the observation in changepoint detection that identifying changepoint locations tends to be harder than identifying the corresponding segment heights. Conditional on a given segmentation, finding the most likely segment heights is usually trivial, with a solution that depends mostly on the data and very little on the prior. Thus, it seems most important to encode prior information about the nature of the segmentation itself.

2.7. Prior Work

The Markov chain defined by (8) was previously studied by Carmi et al. (2014), who coined the term renewal approximation. Carmi et al. derived theoretical results and performed simulations to study identity-by-descent patterns produced by

SMC models. They found that the renewal approximation is comparable to other variants of SMC with some inaccuracy mainly in the tails of the IBD distribution. Importantly, these results pertain to the accuracy of these methods as *priors*; they do not necessarily imply that the renewal approximation is inferior for *inference*. Indeed, generally one hopes that "the data overwhelm the prior," so that inferences do not depend strongly on the choice of prior model.

There have been a few papers specifically devoted to improving the efficiency of SMC. Harris et al. (2014) and Palamara et al. (2018) derived $O(MN)$ decoding algorithms for certain SMC models, where M is the number of hidden states (time discretizations) used in the underlying hidden Markov model. Separately, Lunter (2019) recently showed that MAP estimation can be performed for the Li and Stephens model in $O(N)$ time irrespective of the size H of the underlying copying panel, after a preprocessing step that costs $O(HN)$ time (Durbin 2014).

3. Methods

In this section we derive exact representations for the sequence of marginal posterior distributions $p(X_n \mid Y_{1:N})$, $n = 1, 2, \ldots, N$, and efficient algorithms for sampling paths from the posterior density $p(X_{1:N} \mid Y_{1:N})$ and for computing the MAP path

$$X^*_{1:N} = \arg\max_{X_{1:N}} p(X_{1:N} \mid Y_{1:N}).$$

To save space, proofs are deferred to Appendices S1–S2 in the supplementary material. For the reader's convenience, the various notations introduced in this section are listed in Table S1.

3.1. Exact Marginal Posterior

In what follows, we write $f(x) \in \mathcal{M}_\Gamma(K)$ to signify a the probability density f is a mixture of K gamma distributions, with the mixing weights, scale and shape parameters left unspecified. By abuse of notation, we also write $X \sim \mathcal{M}_\Gamma(K)$ to signify that the random variable X is distributed according to such a mixture.

Let $\alpha(X_n) = p(X_n \mid Y_{1:n})$ denote the (rescaled) forward function from the standard forward-backward algorithm for inferring hidden Markov models (Bishop 2006, sec 13.2.4). Our first result shows that, under the renewal approximation, $\alpha(X_n)$ is a mixture of gamma distributions.

Proposition 1. Suppose that $\pi(x) \in \mathcal{M}_\Gamma(K)$. Then $\alpha(X_n) = p(X_n \mid Y_{1:n}) \in \mathcal{M}_\Gamma(nK)$.

Using this result, we can derive a representation for the marginal posterior distribution.

Proposition 2. If $\pi(x) \in M_\Gamma(K)$ then there exists $f(X_n) \in \mathcal{M}_\Gamma(Kn)$ and $g(X_n) \in \mathcal{M}_\Gamma(K(N-n))$ such that

$$p(X_n \mid Y_{1:N}) = \frac{f(X_n)g(X_n)}{\pi(X_n)}. \tag{10}$$

We can also derive exact expressions for the mixing proportions, shape, and scale parameters for $p(X_n \mid Y_{1:n})$, and by extension, the exact algebraic expression for $p(X_n \mid Y_{1:N})$. This requires substantial additional notation and is deferred to Appendix S5.

3.2. Efficient Posterior Sampling

The exact posterior formula derived in Proposition 2 is useful for visualization, or numerically evaluating functionals (e.g., the posterior mean) of the posterior distribution. However, it is less suited to sampling since the denominator does not divide the numerator except when $K = 1$; and even then, sampling requires expanding the numerator in (3) into (as many as) $O(K^2N^2)$ mixture components.

Instead, we provide an algorithm for efficiently sampling entire paths from $p(X_{1:N} \mid Y_{1:N})$. This idea is due to Fearnhead (2006) (see also Barry and Hartigan 1992), with necessary modifications to accommodate our model's dependence between segment length and height.

Let R_v denote the event that a new IBD segment begins at position v, let $\overline{R}_{u:v} := \left(\bigcup_{i=u+1}^{v-1} R_i\right)^C$ denote the event that there is *not* a recombination event between positions u and v (exclusive), and set $\bar{Y}_{u:v} := \sum_{i=u}^{v} Y_i$. The joint likelihood of the data $Y_{u:v}$ and the event that an IBD segment starts at position u and extends $\Delta = v - u + 1$ positions before terminating at position v is

$$\begin{aligned} &p(Y_{u:v}, \overline{R}_{u:v}, R_v) \\ &\quad = \int_x x^{\mathbf{1}\{u>1\}} \pi(x) \rho x e^{-\rho \Delta x} \prod_{i=u}^{v} e^{-\theta x} (\theta x)^{Y_i} / Y_i! =: P(u, v). \end{aligned} \tag{11}$$

A special case for $u = 1$ is necessary because the initial segment height is sampled from the stationary distribution π, while successive segments heights are distributed according to $\underline{\pi}$; see (2) and (8).

For the last segment, we know only that it extended past position N, so we make the special definition

$$\begin{aligned} P_{-1}(u, N) &= p(Y_{u:N}, \overline{R}_{u:N}) \\ &= \int_x x^{\mathbf{1}\{u>1\}} \pi(x) e^{-\rho \Delta x} \prod_{i=u}^{N} e^{-\theta x} (\theta x)^{Y_i} / Y_i!. \end{aligned} \tag{12}$$

Our algorithm can be used whenever (11) can be efficiently evaluated, in particular when $\pi(t)$ is a gamma mixture.

Defining $Q(u) = p(Y_{u:N} \mid R_u)$ and integrating over the location v where the segment originating at position u terminates, we have (Fearnhead 2006, Theorem 1)

$$Q(u) = \sum_{v=u}^{N-1} P(u, v) Q(v+1) + P_{-1}(u, N) \tag{13}$$

which can be solved by dynamic programming starting from $v = N - 1$ in $O(N^2)$ time. When $v - u$ is large, $P(u, v)$ tends to be extremely small, so the summation in (13) can be truncated without loss of accuracy to obtain an algorithm which is effectively linear in N. Except when noted otherwise, we followed Fearnhead's original suggestion, and truncated the summation as soon as $P(u, v)Q(v+1)$ was less than 10^{-4}.

To sample the next recombination breakpoint τ' from the posterior given that the previous breakpoint occurred at location τ, note that

$$\begin{aligned} p(\tau' \mid \tau, Y_{1:N}) &= \frac{p(Y_{1:N}, R_\tau, R_{\tau'}, \overline{R}_{\tau,\tau'})}{p(Y_{1:N}, R_\tau)} \\ &= \frac{p(Y_{1:\tau-1}, R_\tau) p(Y_{\tau:\tau'-1}, R_{\tau'}, \overline{R}_{\tau:\tau'} \mid R_\tau) Q(\tau')}{p(Y_{1:\tau-1}, R_\tau) Q(\tau)} \\ &= P(\tau, \tau' - 1) Q(\tau') / Q(\tau) \end{aligned}$$

for $\tau' = \tau + 1, \ldots, N-1$, with the remaining probability mass placed on the event that there are no more changepoints. If sampling the first changepoint we set $\tau = 1$.

Having sampled a segmentation $0 < \tau_1, \ldots, \tau_K < N$ from the posterior, we then sample heights conditional on this segmentation. Given that observations $u, u+1, \ldots, v-1, v$ are all on the same segment and are flanked by recombinations, the joint probability of the data $Y_{u:v}$, the segment length Δ, and the segment height x, is the integrand in (11). Hence, the posterior distribution of the segment height x conditional on the underlying segmentation is

$$p(x \mid \overline{Y}_{u:v}, R_{u:v}, R_v) \propto x^{\mathbf{1}\{u>1\}} \pi(x) \rho x e^{-\rho \Delta x} e^{-\theta x} x^{\sum_{i=u}^{v} Y_i}. \quad (14)$$

If $\pi(x)$ is a gamma (mixture), then (14) is also a gamma mixture, and hence easy to sample.

3.3. Exact Frequentist Inference

To complement the Bayesian results in the preceding section, we also derive an efficient frequentist method for inferring the *maximum a posteriori* (MAP) hidden state path,

$$X^*_{1:N} := \arg\max_{X_{1:N}} p(X_{1:N}, Y_{1:N}). \quad (15)$$

When $X_1, \ldots, X_N \in \mathcal{X}$ have discrete support, $|\mathcal{X}| = M$, the MAP path can be found in $O(NM^2)$ time using the Viterbi algorithm (Bishop 2006), and in some cases in $O(NM)$ time by exploiting the special structure of the SMC (Harris et al. 2014; Palamara et al. 2018). Our goal is to efficiently solve the optimization problem (15) when $\mathcal{X} = \mathbb{R}_{>0}$.

To accomplish this, we start by defining the recursive sequence of functions

$$\begin{aligned} V_1(t) &= \log \pi(t) + e_1(t) \\ V_n(t) &= \max_s V_{n-1}(s) + \phi(t \mid s) + e_n(t), \quad n \geq 2 \\ V^*_n &= \max_t V_n(t) \end{aligned}$$

where $e_i(t) = \log p(Y_i \mid X_i = t)$, and

$$\begin{aligned} \phi(t \mid s) &= \log p(X_{i+1} = t \mid X_i = s) \\ &= \begin{cases} -\rho t, & t = s \\ \log(1 - e^{-\rho s}) + \log \underline{\pi}(t), & \text{otherwise} \end{cases} \\ &\approx \begin{cases} -\rho t, & t = s \\ \log(\rho s) + \log \underline{\pi}(t), & \text{otherwise}, \end{cases} \quad (\rho \ll 1) \end{aligned}$$

This is the usual Viterbi dynamic program, but defined over a continuous instead of discrete domain. By standard arguments (Bishop 2006, sec. 13.2.5), we have

$$X^*_N = V^*_N = \arg\max_{X_N} \Big[\max_{X_{1:N-1}} p(X_{1:N}, Y_{1:N}) \Big],$$

and the full path $X^*_{1:N}$ can be recovered by backtracing.

Thus, if we could calculate $V_n(t)$ then the optimization problem (15) would be solved. In general, it is not obvious how to accomplish this, since $V_n(t)$ is a function, that is an infinite-dimensional object which cannot be represented by a computer program. However, our next theorem shows that, in fact, each $V_n(t)$ has a finite-dimensional representation.

Definition 1. Let $\mathcal{V}_K$ be the space of all functions $f : [0, \infty) \to \mathbb{R}$ which can be piecewise defined by K functions of the form $t \mapsto at + b \log t + c$. That is, $f \in \mathcal{V}_K$ if and only if there exists there exists an integer K, a vector $\boldsymbol{\tau} \in \mathbb{R}^{K+1}$ satisfying

$$0 = \tau_1 < \tau_2 < \cdots < \tau_{K+1} = \infty,$$

and vectors $\mathbf{a}, \mathbf{b}, \mathbf{c} \in \mathbb{R}^K$ such that

$$f = a_k t + b_k \log t + c_k, \quad t \in [\tau_{k-1}, \tau_k).$$

Proposition 3. Suppose that $N_e(t) \in \mathcal{V}_K$ is piecewise constant. Then for each $n = 1, \ldots, N$, there exists $K_n < \infty$ such that $V_n(t) \in \mathcal{V}_{K_n}$.

The proof of the theorem (Appendix S3) shows that in order to efficiently compute $V_n(t)$ we need to be able to take the pointwise maximum between any two functions in $\mathcal{V}_K$. We provide an $O(K)$ procedure for doing this in Appendix S6.

Our next result establishes the functional form of $V_n(t)$. Each piece of $V_n(t)$ comprises an interval $I \subset \mathbb{R}$ where, conditional on the TMRCA at position n being $t \in I$, the most probable recombination event occurred a certain number of positions ago. In the statement and proof of the theorem, we use double brackets, $[\![\cdot]\!]$, to refer to individual entries of subscripted vectors.

Proposition 4. For each $V_n(t)$, with breakpoints $\boldsymbol{\tau}_n \in \mathbb{R}^{K_n+1}$, there exists vectors $\mathbf{i}_n \in \mathbb{Z}^{K_n}_{\geq 0}$ and $\mathbf{C}_n \in \mathbb{R}^{K_n}$ such that, for $t \in [\boldsymbol{\tau}_n[\![k]\!], \boldsymbol{\tau}_n[\![k+1]\!])$,

$$\begin{aligned} V_n(t) = \mathbf{C}_n[\![k]\!] + \log \underline{\pi}(t) + \bar{Y}_{\mathbf{i}_n[\![k]\!]:n} \log(\theta t) \\ - t(\theta + \rho)(n - \mathbf{i}_n[\![k]\!]) - \theta t. \end{aligned}$$

Hence, up to the constant $\mathbf{C}_n[\![k]\!]$, $V_n(t)$ equals the log-likelihood of $\bar{Y}_{\mathbf{i}_n[\![k]\!]:n}$ given that the most recent recombination event occurred at position $\mathbf{i}_n[\![k]\!]$ and $X_{\mathbf{i}_n[\![k]\!]} = \cdots = X_n = t$.

Complete pseudocode for our algorithm, based on Propositions 3 and 4, is given in the supplement (Algorithm S1).

In Section 4.2, it will be seen that the posterior distribution is sometimes not centered over the MAP path: the latter tends to oversmooth, missing many changepoints, whereas the posterior mode/mean is generally close to the truth (Figure S5). This is a known feature of the Viterbi decoding of a hidden Markov model, and is not specific to our problem setting (Yau and Holmes 2013; Lember and Koloydenko 2014). In Appendix S7 we derive a generalization of Proposition 3 which allows us to efficiently compute other paths which are suboptimal with respect to (15), but have better pointwise accuracy, thus, enabling a range of possible decodings.

3.4. Extension to Larger Sample Sizes

The preceding sections focused on inferring the sequence of TMRCAs in a pair of sampled chromosomes. In modern applications where hundreds or thousands of samples have been collected, methods that can analyze larger sample sizes are desirable.

We can generalize the problem of decoding the pairwise TMRCA among two chromosomes by treating one of the chromosomes as a fixed genealogy, and considering where the other chromosome joins onto this genealogy at each position. Then, more generally, given a "panel" of $H \geq 1$ chromosomes, we can ask where at each position an additional "focal" chromosome joins onto the panel genealogy.

Extending sequentially Markov coalescent methods to larger sample sizes is not trivial for the simple reason that there is more than one possible tree topology to consider when $n > 2$. Instead of inferring a sequence of numbers $X_{1:N}$ (representing the height of a tree with two leaves), as in the preceding sections, one must consider as hidden states the space of edge-weighted binary trees on n leaves. To circumvent this difficulty, we employ a so-called *trunk approximation* (Paul and Song 2010), which supposes that the underlying ancestral recombination graph is a disconnected forest of H trunks extending infinitely far back into the past. The state space of this model is $\{1, \ldots, H\} \times \mathbb{R}_{>0}$, where the first, discrete coordinate describes the panel haplotype onto which a focal haplotype is currently coalesced, and the second, continuous coordinate gives them time at which that coalescence occurrred. Although the trunk assumption is strong, it has proved useful in a variety of settings (Sheehan, Harris, and Song 2013; Spence et al. 2018; Steinrücken et al. 2019).

Modifying our methods to use the trunk approximation is straightforward and amounts to, essentially, replacing the coalescence measure $p(X \in [t, t + dt)) = \pi(t)\, dt$ with the product measure $p((X, h) \in ([t, t + dt), \{i\})) = \pi(Ht)\, dt$ in all of our formulas. (Note that this measure is properly normalized.) In other words, coalescence occurs with each haplotype at rate 1, and conditional on coalescence, it occurs uniformly onto each haplotype.

4. Results

In this section we compare our method to existing ones, benchmark its speed and accuracy, and conclude with some applications.

4.1. Local Ancestry Inference is Comparable to Existing Methods

As described in the introduction, our initial hypothesis was that posterior inferences for the haplotype decoding problem are relatively insensitive to the choice of prior model on the way that the sequentially Markov coalescent transitions from one position to the next. Here we confirm this hypothesis. To study the relationship between the posterior and prior, we compared the renewal model developed above to the conditional Simonsen-Churchill (CSC) model of Hobolth and Jensen (2014). The CSC is the most accurate sequentially Markovian model known in the literature, and other models such as SMC (McVean and Cardin 2005) and SMC' (Marjoram and Wall 2006) are further approximations of it. Hence, CSC and the renewal model can be viewed as the least and most approximative SMC methods, respectively.

To compare models, we used the procedure described in Hobolth and Jensen (2014, sec. 4.4) to compute the transition probability matrix T, where

$$T_{ij} = p(X_{\ell+1} \in [t_j, t_{j+1}) \mid X_\ell \in [t_i, t_{i+1}))$$

is the probability that the TMRCA at site $\ell + 1$ is in the interval $[t_j, t_{j+1})$ given that the TMRCA of an adjacent site is in $[t_i, t_{i+1})$. We then used this transition matrix to perform posterior decoding in a discrete-state coalescent HMM as previously described (Li and Durbin 2011). We compared the CSC and renewal prior under both constant population size and varying population size, as well as when the recombination rate is equal to the mutation rate and when it is lower. Taking all the combinations of the different population size histories and the recombination rate gives us a total of 4 scenarios. Scenarios 1 and 3 have constant population size, and scenarios 2 and 4 have the variable population size. Scenarios 1 and 2 have recombination rate $r = 10^{-9}$, and scenarios 3 and 4 have recombination rate $r = 1.4 \times 10^{-8}$ per base-pair per generation. We bucketed consecutive base pairs into groups of size $w = 100$ and assume that the recombinations occur between these groups. Additional details of our simulation can be found in Appendix S8.1.

Supplemental Figures S1 and S2 show the Viterbi path and the posterior heatmap for one run of each scenario of the simulation. From Figure S1, there is little difference in the Viterbi plot between the CSC and renewal priors. Both priors produce a Viterbi path very similar to the true sequence of TMRCAs. When the recombination rate increases, the Viterbi paths produced by the two priors fail to capture all the recombination events, but are still very similar in their outputs. We performed a similar analysis for the posterior decoding (Figure S2). Again, it is hard to discern any meaningful difference in all scenarios between the two priors. This is especially the case in scenarios 1 and 2 where the recombination rate is lower.

Confirming these qualitative observations, Table 1 shows the average absolute error for the two priors over the 25 simulations. In terms of absolute error, the renewal prior does about as well as the more correct CSC prior. In fact, the renewal prior outperforms CSC under scenarios 3 and 4, the scenarios with higher recombination rate. A potential explanation for this surprising result, suggested by visually inspecting the posterior decoding obtained from the two methods (e.g., Figure S2, bottom panel), is that the signal-to-noise level in the high recombination regime is low enough that ignoring correlations between (noisily) inferred adjacent segments can actually improve estimation. Provisionally, we hypothesize that the renewal approximation acts as a sort of shrinkage prior in the high-noise regime, trading some bias for lower average risk. However, we observed this effect in only a small number of high-recombination settings, and it is not as pronounced when considering relative error (Table S2).

To better understand these results, we also stratified the error measure by quarter of the true TMRCA distribution (Tables 2 and S3). We expected to see a greater difference between the two priors for larger values of the true TMRCA since, under the

Table 1. Mean absolute error (Err_A) over 25 runs under each scenario.

Scenario	Constant N_e Low ρ (1)	Variable N_e Low ρ (2)	Constant N_e High ρ (3)	Variable N_e High ρ (4)
CSC	5686.79 (198.96)	5201.35 (228.23)	12207.96 (316.49)	11949.15 (146.20)
Renewal	5683.52 (192.43)	5212.97 (226.64)	11660.02 (303.80)	11427.61 (147.19)

NOTE: CSC results were obtained from the conditional Simonsen-Churchill model. Renewal results are from our method. Both methods were discretized. Standard error in parentheses.

Table 2. Mean absolute error (Err_A) over 25 runs under each scenario stratified by quartile.

Scenario	Qtr.	Constant N_e Low ρ (1)	Variable N_e Low ρ (2)	Constant N_e High ρ (3)	Variable N_e High ρ (4)
CSC	Q1	2676.74(115.50)	2271.89(126.87)	6932.49(242.41)	6550.15(118.46)
Renewal	Q1	2714.53(117.88)	2330.01(127.42)	5365.78(184.23)	5168.37(77.63)
CSC	Q2	5961.49(111.73)	6263.63(159.11)	13407.48(60.54)	13255.75(45.30)
Renewal	Q2	6061.91(98.98)	6289.53(147.09)	11575.83(44.20)	11549.62(29.92)
CSC	Q3	9679.44(148.74)	9770.56(259.23)	18853.84(41.04)	18811.84(58.56)
Renewal	Q3	9569.68(156.41)	9673.67(283.39)	19620.79(71.97)	19470.72(52.02)
CSC	Q4	15833.47(265.34)	15968.86(426.23)	33105.92(170.73)	33412.66(200.11)
Renewal	Q4	15439.84(322.81)	15760.62(527.12)	40368.10(208.78)	39760.70(241.19)

NOTE: Other details are as in Table 1.

CSC prior, the distribution of tree height of the current segment conditioned on the tree height of the previous segment, $q(t \mid s)$ is approximately uniform in t for large s, that is $q(t \mid s) \approx 1/s$ when $s \gg t$, where under the renewal prior $\pi(t) = e^{-t}$ has an exponential tail. Conversely, since $\lim_{s\to 0} q(t \mid s) = e^{-t}$, the methods should be comparable for recent TMRCAs.

Table 2 contains the mean absolute error over the 25 simulations after stratification. Under scenarios 1 and 2 where the recombination rate is lower, again we see virtually no difference between the two priors across all quarters. Under scenarios 3 and 4 where the recombination rate is higher, we see that in the first and second quarters, the renewal prior actually has lower error compared to CSC. The results are reversed in the third and fourth quarters where the Markov approximation is more accurate than the renewal prior. This trend is mostly mirrored in Table S3 with the mean relative errors. The renewal prior does just slightly worse than the Markov prior under scenarios 1 and 2 across all quarters. Under scenarios 3 and 4 as the underlying true TMRCA increases, so too does the difference in Err_B.

Next, we studied the extent to which the demographic prior $\pi(t)$ affects the resulting estimates. We simulated data under three different demographic models and then measured the resulting accuracy of the posterior when each model was used as a prior to infer TMRCAs on data generated from the other models (details in Appendix S8.2).

We display the posterior of one pair of chromosomes for all nine pairs of demographies used as data generation and demographic priors in Figure . The plots show that regardless of which demographic prior was used, the resulting posteriors all had the same shape. Table S5 shows that in terms of mean absolute error, all three demographic models perform similarly when used as prior, regardless of which one of them in fact generated the data. Relative error measurements (Table S6) tell a similar story. Given the large differences between the three demographic models (Figure S3), if the posterior were sensitive to the demographic model we would expect each column in the table to be quite different from one another. However, this does not seem to be the case; using the correct prior results in an average improvement of a few percent in most cases.

In conclusion, our results suggest that, as long as the chosen prior is not pathological, its effect on inference will be limited.

4.2. Comparison of Bayesian and Frequentist Inferences

In Section 3 we derived various methods for inferring tree heights. Here we compare the Bayesian method where we sample from the posterior and the frequentist method where we take the MAP path. We apply these two methods to the same simulated data from the first simulation in Section 4.1. For the Bayesian method we sample 200 paths from the posterior and take the median to compare against the MAP path.

Figure S5 shows the results of running the two methods on one set of simulated chromosomes under each scenario. The top two panels of the figure show that when the recombination rate is an order of magnitude lower than the mutation rate, both methods give a faithful approximation of the true sequence of TMRCAs. However, the bottom two panels where the recombination rate is larger displays the key difference between the two methods: the MAP path fails to detect many recombination events, whereas the posterior median is an average over many paths so it can detect recombination events that the MAP path cannot.

We use the same measures of absolute and relative we used in the previous sections. For this simulation, we look at the error at each position so $N/w = N$. The results in Tables S7 and S8 show that the posterior median dominates the MAP path. Again, since the MAP path is the most likely single path whereas in the Bayesian method we take the pointwise median of many paths, the MAP path has inferior pointwise accuracy. This result is expected, but it should be noted that when compared to Tables 1 and S2, the MAP path performs similarly to, and the Bayesian method outperforms, the posterior decoding of the discretized SMC models used in Section 4.1.

4.3. Empirical Time Complexity

In Section 3.2, we suggested that by pruning the state space of our methods in certain ways, their running time could be effectively linear in the number of decoded positions. In this section we confirm this by simulations.

We benchmarked our methods on simulated sequences of length $N = 10^4$ to $N = 10^8$. For each length, we simulated 10 pairs of chromosomes. Figure S6 confirms that there is a linear relationship between chromosome length and running time for both the Bayesian sampler method and the MAP decoder. Note that, if decoding against a larger panel of chromosomes (cf. Section 3.4), the amount of work performed by our algorithms scales linearly in the panel size H. We further verified (Figure S7) that the scaling is linear in both panel size (H) and chromosome length (N); in Figure S8, we tracked the quantity K_n defined in Proposition 3, that is the average number of pieces needed to represent the function $V_n(t)$ for each $1 \leq n \leq N$, and found that it too appears to be bounded on average.

We confirmed a similar empirical scaling for the Bayesian algorithm by tracking the number of summands considered in summation (13) before the truncation threshold was met (Figure S9). On average, the number seems to be bounded by a small constant as the dynamic program (13) proceeds from $u = N$ to $u = 1$. It is possible that this truncation strategy could perform poorly for closely related haplotypes which are cosanguineous over long intervals. To investigate this, we simulated 50 chromosomes and selected the two most closely related pairs of haplotypes in terms of overall IBD sharing. We benchmarked the accuracy and runtime of our sampler using various settings for the truncation cutoff. The results (Table S9) suggest that absolute accuracy is fairly unaffected, but relative accuracy does continue to decline as we decrease the threshold from 10^{-2} to 10^{-6}. This is attributable to the fact that we the TMRCA between two closely-related chromosomes is small on average, which inflates relative error.

4.4. Applications

We tested our method on the two most common real-world applications of the sequentially Markov coalescent.

4.4.1. Exact SMC

The pairwise sequentially Markov coalescent (PSMC; Li and Durbin 2011) is a method for inferring the historical population size (i.e., the function $N_e(t)$ defined in Section 2.2) using genetic variation data from a single diploid individual. Although in some settings PSMC has been superseded by more advanced methods which can analyze larger sample sizes (Schiffels and Durbin 2014; Terhorst, Kamm, and Song 2017), it remains very widely used in many areas of genetics, ecology and biology, because it is fairly robust, and does not require phased data, which can be difficult to obtain for species that have not been studied as intensively as humans. SMC++ (Terhorst, Kamm, and Song 2017) is a generalization of PSMC that does not require phased data which scales to larger sample sizes. Additionally, SMC++ uses the more accurate CSC model (see Section 4.1), whereas PSMC is based on SMC.

As noted in Section 1, both PSMC and SMC++ use an HMM to infer a discretized sequence of genealogies. The discretization grid is a tuning parameter which is challenging to set properly—finer grids inflate both computation time and the variance of the resulting estimate, and for a fixed level of discretization, the optimal grid depends on the unknown quantity of interest $N_e(t)$. A poorly chosen discretization can have serious repercussions for inference (Parag and Pybus 2019).

One potential solution to this problem is to employ general algorithms designed to perform inference in continuous state-spaces. Particle filtering is one such example. The sequential Monte Carlo for the sequentially Markov coalescent (SMCSMC; Henderson et al. 2021) is another method that performs demographic inference using particle filtering. However, a potential downside is that it is simulation-based, and potentially very computationally intensive.

Our method proceeds differently from either of these approaches. Recalling (4), we see that inference of $N_e(t)$ is tantamount to estimating (the reciprocal of) $\eta(t)$. In survival analysis, η is known as the *hazard rate function*, and a variety of methods have been developed to infer it (Wang 2014). Thus, if we could somehow sample directly from π, then inference of $N_e(t)$ would reduce to a fairly well-understood problem. While this is impossible in practice, the simulated results shown in the preceding sections inspire us to believe that samples drawn from the posterior $p(X_{1:N} \mid Y_{1:N})$ could serve the same purpose. Concretely, we suppose that a random sample $x_1, \ldots, x_k$ drawn from the product measure

$$p(X_{i_1} \mid Y_{1:N}) \times p(X_{i_2} \mid Y_{1:N}) \times \cdots \times p(X_{i_k} \mid Y_{1:N}), \qquad (16)$$

where the index sequence $i_1, \ldots, i_k \in [N]$ is sufficiently separated to minimize correlations between the posteriors, is distributed as k iid samples from coalescent density. We then use a kernel-smoothed version of Nelson-Aalen estimator (Wang 2014) in order to estimate $\widehat{N}_e(t)$. As a hyperprior on the coalescent intensity function, we simply used Kingman's coalescent, $\pi(t) = e^{-t}$.

We first compared the performance of our method with PSMC, SMC++, and SMCSMC on simulated data. Figure 1 compares the results of running our method, which we call XSMC (e**X**act **SMC**), and the three competing methods on data simulated from three size history functions (plotted as dashed grey lines). We simulated a chromosome of length $N = 5 \times 10^7$ base pairs for 25 diploid individuals (total of 50 chromosomes), and then ran both methods on all 25 pairs. For XSMC, we drew 100 random paths from the posterior distribution, and then sampled marginal TMRCAs from each path according to (16) with 50,000 base pair spacing between sampling locations. The plots show the pointwise median, with the interquartile range (distance between the 25th and 75th percentiles) plotted as an opaque band around the median. For the first two simulations we assumed that the mutation and recombination rates were equal, $\mu = r = 1.4 \times 10^{-8}$ per base pair per generation. For reasons discussed below, we assumed in the third simulation that $r = 10^{-9}$. Both methods were run with their default parameters and provided with the true ratio r/μ used to generate the data.

The left column of the figure ("Constant") depicts the most basic scenario, where the population size is unchanged over time. While all methods do an acceptable job, PSMC and

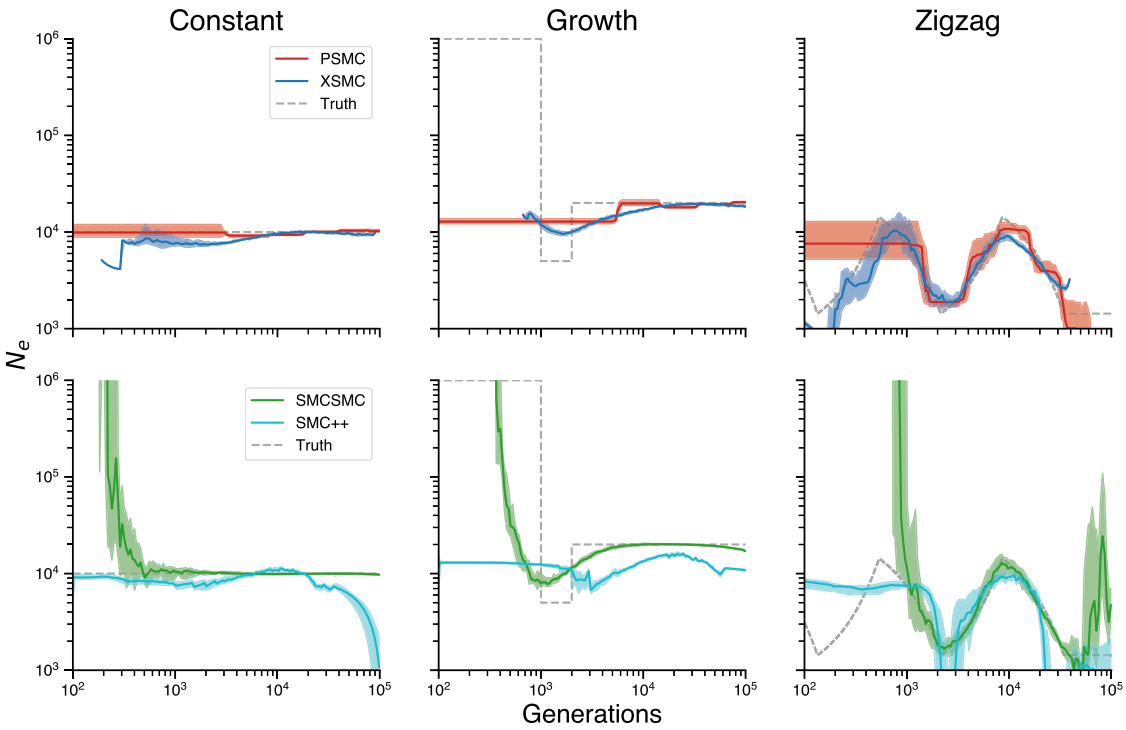

Figure 1. Comparison of XSMC, PSMC, SMCSMC, SMC++ on various simulated size histories.

XSMC exhibit less bias. For PSMC, there is some bias from the piecewise-constant model class it uses to perform estimation. (We note that with its default settings, PSMC actually initializes to the true model in this scenario.) XSMC has a slight downward bias in the recent past, but is otherwise centered over the true values $N_e = 10^4$. Both methods appear slightly biased in the period 10^3-10^4 generations, though in opposite directions. SMCSMC performs well after 10^3 about generations, however, it incorrectly estimates a large increase in N_e toward the present. SMC++ exhibits a slight downward bias toward the recent past and also incorrectly estimates a population crash further back in time.

In the center column ("Growth"), we simulated a cartoon model of recent expansion, in which the population experiences a brief bottleneck from 2000–1000 generations ago, before suddenly increasing in size by two hundredfold. This model is more difficult to correctly infer using only diploid data, because the large recent population size prevents samples from coalescing during this time, depriving methods of the ability to learn size history in the recent past. Nevertheless, XSMC does an acceptable job of showing that the population experienced a dip followed by a sharp increase, though the estimates are oversmoothed. In contrast, PSMC estimates size history that is nearly flat, with no acknowledgement of the bottleneck. SMC++ estimates a similar trajectory as XSMC, but is slightly more downward biased at all points in time. At an initial glance, SMCSMC looks to have most faithfully estimated the population size history. However, the results from the other two scenarios indicate that SMCSMC tends to infer a recent growth in population whether or not it actually occurred. Even so, without considering this feature of the model, SMCSMC returns a similar result to XSMC. This result also illustrates another benefit of the nonparametric approach: XSMC only returns an answer where it actually observes data. Because no coalescence times were observed before $\sim 10^3$ generations when sampling from the posterior, our method does not plot anything outside of that region. This compares favorably with PSMC and related parametric methods (e.g., Schiffels and Durbin 2014; Terhorst, Kamm, and Song 2017; Steinrücken et al. 2019), which have to model $N_e(t)$ over all $0 \leq t < \infty$ in order to perform an analysis, even when the data contain no signal outside of a limited region.

Lastly, in the right-hand column we examined a difficult demography known in the literature as the zigzag model (Schiffels and Durbin 2014). This is a pathological model of repeated exponential expansions and contractions, and is designed to benchmark various demographic inference procedures. We found that with the default setting $\rho = \theta$ used in the preceding two examples, the methods failed to produce good results on the zigzag. We therefore lowered the rate of recombination to $r = 10^{-9}$/bp/generation in order to create more linkage disequilibrium for the methods to exploit. Here, a fairly substantial difference emerges between the two methods. XSMC does the best job of inferring this difficult size history, with accurate results to almost 10^2 generations in the past, and almost no discernible bias. It is also the only method to successfully infer the final population crash in the recent past. In contrast, PSMC and SMC++ return similar results where the methods are able to recover the true value accurately after 10^3 generations. SMCSMC also returns similar results to PSMC and SMC++, but again the method incorrectly infers a population increase both toward the present and further back in the past.

Table 3 displays the total running time in minutes of the four methods of the 75 total simulations across the three different demographies. Each method was parallelized across the simulations and run on a 32-core machine. XSMC and PSMC completed the simulations significantly faster than SMC++ and

Table 3. Total running time of XSMC, PSMC, SMCSMC, and SMC++ in minutes of 75 total simulations on various simulated size histories.

Method	Minutes
SMC++	519.865721
SMCSMC	1840.547969
XSMC	0.891570
PSMC	1.401326

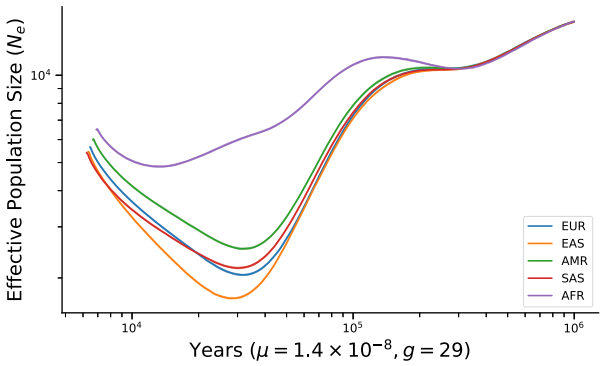

Figure 2. Result of fitting XSMC to 1000 Genomes data. For each superpopulation, 20 samples were chosen. Solid line denotes the median across all samples, and shaded bands denote the interquartile range.

SMCSMC, and between the two methods, XSMC outperformed PSMC computationally by a relatively large margin. The simulation results show that XSMC can deliver high quality estimates of demography more quickly than competing methods.

Encouraged by these results, we next turned to analyzing real data. We performed a simple analysis where we analyzed whole genome data from 20 individuals from each of the five superpopulations (African, European, East Asian, South Asian, and Admixed American) in the 1000 Genomes dataset (The 1000 Genomes Project Consortium 2015). Results are shown in Figure 2. Broadly speaking, our method agrees with other recently published estimates (Li and Durbin 2011; Terhorst, Kamm, and Song 2017), and succeeds in capturing major recent events in human history such as an out-of-Africa event 100–200kya, a bottleneck experienced by non-African populations, and explosive recent growth beginning around 20kya. On the other hand, certain features that have been found in previous analyses (e.g., the peak and drop before 100Kya in Figure 3a of Li and Durbin 2011) are smoothed out by our method, likely due to the novel use of kernel methods here. These estimates could probably be improved with fine-tuning and the use of additional data, but we did not attempt this, the message being that our method has moderate data requirements and produces reasonable results with minimal user intervention. Finally, we note that our method is highly efficient: to analyze all $20 \times 5 \times (3 \times 10^9 \text{Mbp}) \approx 300\text{Gbp}$ of sequence data took approximately 40 min on a 12-core workstation. A single human genomes (all 22 autosomes) can be analyzed in about 30 sec.

4.4.2. Phasing and Imputation

The Li and Stephens (2003) haplotype copying model (hereafter, LS) is an approximation to the conditional distribution of a "focal" haplotype (e.g., a chromosome) given a set of other "panel" haplotypes. It supposes that the focal haplotype copies with error from different members of panel, occasionally switching to a new template due to recombination. Genealogically, this can be interpreted as finding the local genealogical nearest neighbor (GNN) of the focal haplotype within the panel. LS has been used extensively in applications, for example phasing diploid genotype data into haplotypes (Stephens and Scheet 2005) and imputing missing data (Scheet and Stephens 2006; Marchini et al. 2007; Howie, Donnelly, and Marchini 2009). The method's undeniable success is actually somewhat surprising, since it assumes an extremely simple genealogical relationship between the focal and panel haplotypes which ignores time completely (Paul and Song 2010). Hence, while we motivated XSMC as a fast and slightly more approximate SMC prior, it can also be seen as a more biologically faithful version of LS.

We wondered whether our method could be used to improve downstream phasing and imputation. Fully implementing a phasing or imputation pipeline is beyond the scope of this article, so we settled for checking in simulations whether decoding results produced by XSMC were more genealogically accurate than those obtained using LS. We simulated data using realistic models of human chromosomes 10 and 13 (Adrion et al. 2019). We chose these two because chromosome 10 is estimated to have an average ratio of recombination to mutation slightly above 1 ($\rho/\theta = 1.07$), while in chromosome 13 the ratio is slightly below 1 ($\rho/\theta = 0.87$). The ratio of recombination to mutation affects the difficulty of phasing and imputation, with higher ratios leading to less linkage disequilibrium and thus less accurate results. We also explored the effects of varying the size of the haplotype panel. For each chromosome, we simulated 10 datasets with panels of size $H = 2, 4, 10, 25, 100$.

As a proxy for phasing and imputation accuracy, we studied which method identified a genealogical nearer neighbor on average. The GNN at a given position is defined to be any panel haplotype that shares the earliest common ancestor with the focal haplotype. In other words, any panel haplotype that has the smallest TMRCA with the focal haplotype is a GNN. (Note that there may be more than one GNN.) For purposes of accurate phasing and imputation, it is desirable to identify the GNN as closely as possible.

For each simulation we computed the Viterbi path from XSMC and LS, as well as the posterior modal haplotype, and studied the proximity of those paths to the true GNN at each segregating site. Table S10 shows the proportion of segregating sites where XSMC and LS both estimated the same haplotype to be the GNN. For the MAP path, there is a high level of agreement, 80%–90%, between the two methods for both small and large panel sizes. When the panel size is small ($H = 2$), there are few possible choices, and when the panel size is large ($H = 100$) the decoding consists mostly of long, recent stretches of IBD which are fairly easy to estimate. Disagreement is highest for intermediate values $H = 4, 10, 25$ where neither of these effects dominates. At sample size $H = 10$ the methods only agree at about half of segregating sites. The posterior mode appears to be less stable, with the agreement between the two methods decreasing monotonically as the panel size increases, down to agreement at only abouth 1/3rd of sites when $H = 100$.

At the 10%–66% of sites where the methods disagree, the results indicate a statistically significant gain for XSMC

compared to LS. Table S11 shows that conditional on the two methods inferring different haplotypes as the GNN at that site, XSMC finds a genealogical nearer neighbor more often except in one case (chromosome 10, $H = 10$, MAP path.) Using MAP estimation, the advantage of using XSMC increases, as the panel size increases, up to a roughly 6%–10% advantage on chromosome $H = 100$. For the posterior mode, the methods perform more comparably, and the largest difference is on the order of a few percentage points. The performance difference is significantly different from equal odds in almost every case.

5. Conclusion

In this article, we studied the sequentially Markov coalescent, a framework for approximating the likelihood of genetic data under various evolutionary models. We proposed a new inference method which supposes that the heights of neighboring identity-by-descent segments are independent. We showed that this led to decoding algorithms which are faster and have less bias than existing algorithms.

There are several possible extensions to our work. It is straightforward to extend our techniques to allow for position-specific rates of recombination and mutation, which could then be used to infer spatial or motif-specific variation in these processes.

Although we focused here on analyzing data from a single, panmictic population, we can also use posterior samples or MAP estimates to infer more complicated models of population structure. It is also possible to extend some of our techniques to other priors which model correlations between adjacent IBD segments. For the Viterbi decoder, we were able to implement a version of the algorithm in Section 3.3 which works for McVean and Cardin's original SMC model. This could be useful, for example, if analyzing data from a structured population, to the extent that adjacent segments of identity by descent are more likely to derive from members of the same subpopulation. However, the resulting procedure is much more complicated. The Viterbi function $V_n(t)$ no longer has the tractable form derived in Proposition 3. Consequently, we cannot use a simple method like the one in Appendix S6 to perform the pointwise maximization in (4). Instead, numerical optimization must be used instead, resulting in a slower algorithm.

Another interesting possibility is to use our method to estimate ancestral recombination graphs. Recently, there has been a resurgence of interest in inferring ARGs using large samples of cosmopolitan genomic data (Kelleher et al. 2019; Speidel et al. 2019). Although these represent an impressive breakthrough, they rely on heuristic estimation procedures that do not directly model the underlying genealogical process that generates ancestry. Our method provides a new possibility for ARG estimation, by iteratively adding samples onto a sequence of estimated genealogies, but without the need to discretize those genealogies. These and other extensions are the subjects of ongoing work.

Supplementary Materials

In the supplement we present supporting lemmas, proofs of the theorems, and additional plots and tables. (pdf)

Disclosure Statement

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

Funding

This research was supported by the National Science Foundation (grant number DMS-2052653, and a Graduate Research Fellowship), and the National Institute of General Medical Sciences of the National Institutes of Health under award number R35GM151145. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

ORCID

Jonathan Terhorst http://orcid.org/0000-0001-7765-2101

Data Availability Statement

All of the data analyzed in this article are either simulated, or publicly available. A Python package implementing our method is available at *https://terhorst.github.io/xsmc*. Code which reproduces all of the figures and tables in this article is available at *https://terhorst.github.io/xsmc/paper*.

References

Adrion, J. R., Cole, C. B., Dukler, N., Galloway, J. G., Gladstein, A. L., Gower, G., Kyriazis, C. C., Ragsdale, A. P., Tsambos, G., Baumdicker, F., Carlson, J., Cartwright, R. A., Durvasula, A., Kim, B. Y., McKenzie, P., Messer, P. W., Noskova, E., Vecchyo, D. O.-D., Racimo, F., Struck, T. J., Gravel, S., Gutenkunst, R. N., Lohmeuller, K. E., Ralph, P. L., Schrider, D. R., Siepel, A., Kelleher, J., and Kern, A. D. (2019), "A Community-Maintained Standard Library of Population Genetic Models," bioRxiv. [11]

Barry, D., and Hartigan, J. A. (1992), "Product Partition Models for Change Point Problems," *The Annals of Statistics*, 20, 260–279. [4,5]

——— (1993), "A Bayesian Analysis for Change Point Problems," *Journal of the American Statistical Association*, 88, 309–319. [4]

Bhaskar, A., Wang, Y. X. R., and Song, Y. S. (2015), "Efficient Inference of Population Size Histories and Locus-Specific Mutation Rates from Large-Sample Genomic Variation Data," *Genome Research*, 25, 268–279. [3]

Bishop, C. M. (2006), *Pattern Recognition and Machine Learning*, Berlin, Heidelberg: Springer-Verlag. [5,6]

Carmi, S., Wilton, P. R., Wakeley, J., and Pe'er, I. (2014), "A Renewal Theory Approach to IBD Sharing," *Theoretical Population Biology*, 97, 35–48. [1,4]

Chan, A. H., Jenkins, P. A., and Song, Y. S. (2012), "Genome-Wide Fine-Scale Recombination Rate Variation in Drosophila melanogaster," *PLoS Genetics*, 8, e1003090. [3]

Chan, N. H., Ng, W. L., Yau, C. Y., and Yu, H. (2021), "Optimal Change-Point Estimation in Time Series," *The Annals of Statistics*, 49, 2336–2355. [4]

Drummond, A. J., Rambaut, A., Shapiro, B., and Pybus, O. G. (2005), "Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences," *Molecular Biology and Evolution*, 22, 1185–1192. [3]

Durbin, R. (2014), "Efficient Haplotype Matching and Storage Using the Positional Burrows–Wheeler Transform (PBWT)," *Bioinformatics*, 30, 1266–1272. [5]

Durrett, R. (2008), *Probability Models for DNA Sequence Evolution* (2nd ed.), New York: Springer. [2,3]

Fearnhead, P. (2006), "Exact and Efficient Bayesian Inference for Multiple Changepoint Problems," *Statistics and Computing*, 16, 203–213. [4,5]

Fearnhead, P., and Liu, Z. (2011), "Efficient Bayesian Analysis of Multiple Changepoint Models with Dependence across Segments," *Statistics and Computing*, 21, 217–229. [4]

Gay, J. C., Myers, S., and McVean, G. (2007), "Estimating Meiotic Gene Conversion Rates from Population Genetic Data," *Genetics*, 177, 881–894. [3]

Gill, M. S., Lemey, P., Faria, N. R., Rambaut, A., Shapiro, B., and Suchard, M. A. (2013), "Improving Bayesian Population Dynamics Inference: A Coalescent-based Model for Multiple Loci," *Molecular Biology and Evolution*, 30, 713–724. [3]

Griffiths, R. C., and Marjoram, P. (1997), "An Ancestral Recombination Graph," in *Progress in Population Genetics and Human Evolution* (Vol. 87), eds. P. Donnelly, and S. Tavaré, pp. 257–270, Berlin: Springer-Verlag. [2]

——— (1996), "Ancestral Inference from Samples of DNA Sequences with Recombination," *Journal of Computational Biology*, 3, 479–502. [2]

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., and Bustamante, C. D. (2009), "Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data," *PLoS Genetics*, 5, e1000695. [3]

Harris, K., Sheehan, S., Kamm, J. A., and Song, Y. S. (2014), Decoding Coalescent Hidden Markov Models in Linear Time," in *Proc. 18th Annual Intl. Conf. on Research in Computational Molecular Biology (RECOMB)*, Vol. 8394 of LNBI, pp. 100–114, Springer. (NIHMSID 597680, PMC Pending). [5,6]

Hein, J., Schierup, M. H., and Wiuf, C. (2005), *Gene Genealogies, Variation and Evolution*, Oxford: Oxford University Press. [2]

Henderson, D., Zhu, S. J., Cole, C. B., and Lunter, G. (2021), "Demographic Inference from Multiple Whole Genomes Using a Particle Filter for Continuous Markov Jump Processes," *PLoS One*, 16, e0247647. [9]

Hobolth, A., and Jensen, J. L. (2014), "Markovian Approximation to the Finite Loci Coalescent with Recombination Along Multiple Sequences," *Theoretical Population Biology*, 98, 48–58. [1,4,7]

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., and Abecasis, G. R. (2012), "Fast and Accurate Genotype Imputation in Genome-Wide Association Studies through Pre-Phasing," *Nature Genetics*, 44, 955–959. [2]

Howie, B. N., Donnelly, P., and Marchini, J. (2009), "A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies," *PLoS Genetics*, 5, e1000529. [2,11]

Huang, J., Howie, B., McCarthy, S., Memari, Y., Walter, K., Min, J. L., Danecek, P., Malerba, G., Trabetti, E., Zheng, H.-F., UK10K Consortium, Gambaro, G., Richards, J. B., Durbin, R., Timpson, N. J., Marchini, J., and Soranzo, N. (2015), "Improved Imputation of Low-Frequency and Rare Variants Using the UK10K Haplotype Reference Panel," *Nature Communications*, 6, 8111. [2]

Jackson, B., Scargle, J. D., Barnes, D., Arabhi, S., Alt, A., Gioumousis, P., Gwin, E., Sangtrakulcharoen, P., Tan, L., and Tsai, T. T. (2005), "An Algorithm for Optimal Partitioning of Data on an Interval," *IEEE Signal Processing Letters*, 12, 105–108. [4]

Jouganous, J., Long, W., Ragsdale, A. P., and Gravel, S. (2017), "Inferring the Joint Demographic History of Multiple Populations: Beyond the Diffusion Approximation," *Genetics*, 206, 1549–1567. [3]

Kamm, J. A., Terhorst, J., and Song, Y. S. (2017), "Efficient Computation of the Joint Sample Frequency Spectra for Multiple Populations," *Journal of Computational and Graphical Statistics*, 26, 182–194. [3]

Kamm, J., Terhorst, J., Durbin, R., and Song, Y. S. (2020), "Efficiently Inferring the Demographic History of Many Populations with Allele Count Data," *Journal of the American Statistical Association*, 115, 1472–1487. [3]

Kelleher, J., Wong, Y., Wohns, A. W., Fadil, C., Albers, P. K., and McVean, G. (2019), "Inferring Whole-Genome Histories in Large Population Datasets," *Nature Genetics*, 51, 1330–1338. [3,12]

Killick, R., Fearnhead, P., and Eckley, I. A. (2012), "Optimal Detection of Changepoints with a Linear Computational Cost," *Journal of the American Statistical Association*, 107, 1590–1598. [4]

Lawson, D., Hellenthal, G., Myers, S., and Falush, D. (2012), "Inference of Population Structure using Dense Haplotype Data," *PLoS Genetics*, 8, e1002453. [3]

Lember, J., and Koloydenko, A. A. (2014), "Bridging Viterbi and Posterior Decoding: A Generalized Risk Approach to Hidden Path Inference based on Hidden Markov Models," *The Journal of Machine Learning Research*, 15, 1–58. [6]

Li, H., and Durbin, R. (2011), "Inference of Human Population History from Individual Whole-Genome Sequences," *Nature*, 475, 493–496. [1,2,3,4,7,9,11]

Li, N., and Stephens, M. (2003), "Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data," *Genetics*, 165, 2213–2233. [1,2,3,11]

Lunter, G. (2019), "Haplotype Matching in Large Cohorts Using the Li and Stephens Model," *Bioinformatics*, 35, 798–806. [5]

Marchini, J., Howie, B., Myers, S. R., McVean, G., and Donnelly, P. (2007), "A New Multipoint Method for Genome-Wide Association Studies by Imputation of Genotypes," *Nature Genetics*, 39, 906–13. [2,11]

Marjoram, P., and Wall, J. D. (2006), "Fast "Coalescent" Simulation," *BMC Genetics*, 7, 16. [1,4,7]

McVean, G. A., and Cardin, N. J. (2005), "Approximating the Coalescent with Recombination," *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360, 1387–1393. [1,4,7]

Minin, V. N., Bloomquist, E. W., and Suchard, M. A. (2008), "Smooth Skyride through a Rough Skyline: Bayesian Coalescent-based Inference of Population Dynamics," *Molecular Biology and Evolution*, 25, 1459–1471. [3]

Palamara, P. F., Terhorst, J., Song, Y. S., and Price, A. L. (2018), "High-throughput Inference of Pairwise Coalescence Times Identifies Signals of Selection and Enriched Disease Heritability," *Nature Genetics*, 509, 1311–1317. [3,5,6]

Parag, K. V., and Pybus, O. G. (2019), "Robust Design for Coalescent Model Inference," *Systematic Biology*, 68, 730–743. [1,9]

Paul, J. S., and Song, Y. S. (2010), "A Principled Approach to Deriving Approximate Conditional Sampling Distributions in Population Genetics Models with Recombination," *Genetics*, 186, 321–338. [1,2,7,11]

Paul, J. S., Steinrücken, M., and Song, Y. S. (2011), "An Accurate Sequentially Markov Conditional Sampling Distribution for the Coalescent With Recombination," *Genetics*, 187, 1115–1128. [4]

Price, A. L., Tandon, A., Patterson, N., Barnes, K. C., Rafaels, N., Ruczinski, I., Beaty, T. H., Mathias, R., Reich, D., and Myers, S. R. (2009), "Sensitive Detection of Chromosomal Segments of Distinct Ancestry in Admixed Populations," *PLoS Genet*, 5, e1000519. [3]

Pybus, O. G., Rambaut, A., and Harvey, P. H. (2000), "An Integrated Framework for the Inference of Viral Population History from Reconstructed Genealogies," *Genetics*, 155, 1429–1437. [3]

Rasmussen, M. D., Hubisz, M. J., Gronau, I., and Siepel, A. (2014), "Genome-Wide Inference of Ancestral Recombination Graphs," *PLoS Genetics*, 10, e1004342. [3]

Rubinacci, S., Ribeiro, D. M., Hofmeister, R. J., and Delaneau, O. (2021), "Efficient Phasing and Imputation of Low-Coverage Sequencing Data Using Large Reference Panels," *Nature Genetics*, 53, 120–126. [2]

Scheet, P., and Stephens, M. (2006), "A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase," *American Journal of Human Genetics*, 78, 629–644. [2,11]

Schiffels, S., and Durbin, R. (2014), "Inferring Human Population Size and Separation History from Multiple Genome Sequences," *Nature Genetics*, 46, 919–925. [3,9,10]

Sheehan, S., Harris, K., and Song, Y. S. (2013), "Estimating Variable Effective Population Sizes from Multiple Genomes: A Sequentially Markov Conditional Sampling Distribution Approach," *Genetics*, 194, 647–662. [3,7]

Shi, X., Gallagher, C., Lund, R., and Killick, R. (2022), "A Comparison of Single and Multiple Changepoint Techniques for Time Series Data," *Computational Statistics & Data Analysis*, 170, 107433. [4]

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. [3,12]

Spence, J. P., Steinrücken, M., Terhorst, J., and Song, Y. S. (2018), "Inference of Population History Using Coalescent HMMs: Review and Outlook," *Current Opinion in Genetics & Development*, 53, 70–76. [3,7]

Steinrücken, M., Kamm, J., Spence, J. P., and Song, Y. S. (2019), "Inference of Complex Population Histories Using Whole-Genome Sequences from Multiple Populations," *Proceedings of the National Academy of Sciences*, 116, 17115–17120. [3,7,10]

Stephens, M., and Scheet, P. (2005), "Accounting for Decay of Linkage Disequilibrium in Haplotype Inference and Missing-Data Imputation," *American Journal of Human Genetics*, 76, 449–62. [11]

Terhorst, J., Kamm, J. A., and Song, Y. S. (2017), "Robust and Scalable Inference of Population History from Hundreds of Unphased Whole Genomes," *Nature Genetics*, 49, 303–309. [3,9,10,11]

The 1000 Genomes Project Consortium (2015), "A Global Reference for Human Genetic Variation," *Nature*, 526, 68–74. [11]

Voight, B. F., Kudaravalli, S., Wen, X., and Pritchard, J. K. (2006), "A Map of Recent Positive Selection in the Human Genome," *PLoS Biology*, 4, e72. [3]

Wang, J.-L. (2014), "Smoothing Hazard Rates," *Wiley StatsRef: Statistics Reference Online*. [9]

Wiuf, C., and Hein, J. (1999), "Recombination as a Point Process Along Sequences," *Theoretical Population Biology*, 55, 248–259. [1]

Yau, C., and Holmes, C. C. (2013), "A Decision-Theoretic Approach for Segmental Classification," *The Annals of Applied Statistics*, 7, 1814–1835. [6]