Statistical Methods in Population Genetics and Viral Phylodynamics

by

Caleb Ki

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in the University of Michigan
2022

Doctoral Committee:

Assistant Professor Jonathan Terhorst, Chair
Assistant Professor Yang Chen
Professor Edward Ionides
Professor Sebastian Zollner

Caleb Ki
calebki@umich.edu
ORCID iD: 0000-0001-9089-3735

ACKNOWLEDGMENTS

There are so many people that have supported me and made it possible for me to succeed.

This work would not have been possible without my advisor Jonathan Terhorst. Thank you Jonathan for being always being so patient and so gracious. You have taught me to be a better statistician, coder, and researcher. I have always admired your intelligence, work ethic, and love of research and statistics, and for the past five years you set the standard that I have tried to emulate.

To my parents, Junghee Bang and Ryun Ki, thank you for your unwavering support. You made the decision to move our family the United States over two decades ago. You sacrificed so much to provide a better life for your children, and for that I will be eternally grateful.

To my sister, Seulki Ki, thank you for being my biggest supporter. Thank you for being there for me whenever I need someone to lean on.

To my labmate, Enes Dilber, thank you for the fun conversations over lunch and Slack as well as your collaboration for both homework and research. To the friends I've made along the way in Ann Arbor and to my friends who have supported me from afar, thank you for all for being so funny, so kind. Our planned hangouts and weekend trips always gave me something to look forward to after a long, hard day, week, or semester. In particular, I'd like to thank Byoungwook Jang and Sanjana Gupta. Byoung, thank you for being an excellent roommate and indulging yourself in the unhealthiest of foods with me. Sanjana, thank you for being my best friend these past five years. I don't think I could have made it through grad school without you. Thank for all the memories. I will cherish them forever.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

ABSTRACT

Genetic sequences carry a wealth of information. Scientists and statisticians have utilized genetic variation data to answer a wide range of questions in evolutionary biology and epidemiology. With the advent of high throughput sequencing, the availability of genetic sequence data has exploded this century. While the unprecedented amount of genetic data available presents an opportunity to garner a deeper understanding about viruses and humans, making use of large volumes of genetic data is still a challenging problem.

In what is to follow, we present three methods that tackle various problems analyzing genetic variation data. First, we introduce the framework known as the sequentially Markov coalescent (SMC), which enables likelihood based inference using hidden Markov models (HMMs) where the latent variables represent genealogies. While genealogies are continuous, HMMs are discrete, requiring SMC based methods to discretize genealogies. This discretization often leads to biased and noisy estimates of the population size history. We introduce a method that avoids the need for discretization leading to Bayesian and frequentist inference procedures that are faster and less biased than its predecessors.

Additionally, while coalescent HMMs based on SMC can be decoded in linear time, there does not yet exist a linear time EM algorithm for coalescent HMMs based on SMC', the more accurate approximation. We present a linear time EM algorithm based on SMC'. Advantages of this method include increased accuracy, computation time, uncertainty quantification, and ability to incorporate regularization.

Lastly, we present a new approach for estimating transmission and recovery rates of viruses using genetic sequence data. With the outbreak of the SARS-CoV-2, there are millions of genomic sequences available to analyze, but few methods to exploit the information contained in these sequences. By integrating recent advances in Bayesian inference and differentiable programming with phylodynamics, we provide a method capable of estimating transmission, recovery, and sampling of pathogens using thousands of sequences. We apply our method to SARS-CoV-2 data and find that our estimates of the effective reproductive number closely match other estimates from methods based on public health data.

CHAPTER 1

Introduction

For decades, genetic sequence data has proven to be an invaluable resource for scientists and statisticians in evolutionary biology and epidemiology. Researchers have utilized this data to extract knowledge about evolutionary processes such as natural selection (Fariello et al., 2013; Stern et al., 2021; Dilber and Terhorst, 2022), migration (Cann et al., 1987; Petkova et al., 2016; Al-Asadi et al., 2019), and recombination (Li and Stephens, 2003; Chan et al., 2012; Kamm et al., 2016). Human sequence data have even helped scientists link diseases to different genetic variants (Sun et al., 2022) and illuminate why lactose intolerance is more prevalent in Asia than in Europe (Sahi, 1994; Anguita-Ruiz et al., 2020).

Additionally, viral sequence data have been useful in dating epidemic and pandemic origins (Fraser et al., 2009; Lemey et al., 2006) as well as infection transmissions (Volz et al., 2013a); estimating the reproductive number in hepatitis C (Pybus et al., 2001), HIV (Volz et al., 2009), and influenza (Müller et al., 2020); and lending insight into the efficacy of interventions (Drummond et al., 2001; Stadler et al., 2013). Clearly, scientists have uncovered a large amount of knowledge with the available genetic data, yet there are still many questions in population genetics and phylodynamic inference that remain unanswered.

Today, due to the rapid development in sequencing technology, the world is awash in sequence data. The UK Biobank is a massive data set containing genotype data from approximately 500,000 individuals from the UK (Bycroft et al., 2018). There exists an even larger repository of SARS-CoV-2 sequences, GISAID, which contains 7.5 million sequences and counting (Elbe and Buckland-Merrett, 2017; van Dorp et al., 2021). Moreover, there are over a billion sequences from thousands of different species available on Genbank (Benson et al., 2012). To leverage the rapid influx of data available in order to further understand the evolutionary processes of humans and viruses, there is a need for scalable and accurate methods capable of extracting information from that data.

In this thesis, we present three new contributions to the fields of population genetics and phylodynamic inference that increase our ability to analyze genetic sequence data quickly and accurately.

In the remainder of the chapter, we give a brief overview of the problems in population genetics and phylodynamic inference that we investigate later in the main text. We expound on the differences between human and viral sequence data, illuminating the unique challenges of working with each type of data.

1.1 The Sequentially Markov Coalescent

Regardless of whether we are analyzing viral data or human data (or data from other organisms), at the most basic level, the aim is to understand the underlying biological processes and evolutionary dynamics that govern that data. To that end, probabilistic models of evolution play a central role in analyzing genetic sequence data of all types. Coalescent theory is a mathematical theory of ancestry that serves as the foundation for many probabilistic models in both population genetics and viral phylodynamics.

Kingman (1982) showed that the coalescent model is the limiting process for the Wright-Fisher and Moran models, among others. The coalescent traces the ancestral lineage (the series of ancestors of each gene back through time). The event that two lineages find their most recent common ancestor (MRCA) and merge by finding a common ancestor is called a coalescent event. Each time a coalescent event occurs, the number of lineages decreases. This process repeats until there is just a single lineage left. Given a random sample of n genes, there will be $n-1$ coalescent events. All this information can be encoded in a bifurcating tree where samples are represented as leaves and coalescence events as internal nodes. The branch lengths connecting the trees represent the time to the most recent common ancestor (TMRCA) or the time it took the nodes to coalesce.

If there are i lineages, the waiting time for any pair of lineages to coalesce, T_i, is exponentially distributed with rate parameter $\binom{i}{2}$. In the case where there are only two lineages, T_2 is exponentially distributed with rate 1. In the coalescent model, time is measured in units of $1/(2N_e)$ where N_e is the effective population size. Clearly this means the rate of coalescence and the effective population size are inversely proportional: for large populations, the rate of coalescence is small and for small populations, the rate of coalescence is large. Different from the census population size, the effective population size is the size of an idealized population that has the same value of some parameter (usually genetic drift) in the population of interest. In order for a population to be *ideal*, there must be an equal number of males and females, individuals must be equally likely to produce offspring, and mating must be random. Of course since most real populations deviate from this ideal model of populations, there will be discrepancies between the census population size and the effective population size. Even still, the effective population size is an important parameter to study as it is not only intrinsically interesting, but is also necessary to understand other evolutionary processes such as selection or migration.

If we could link our sample of sequences to the space of trees encoding the ancestry of our samples we could reliably recover the effective population size of the population over time. This can be easily done under the model described above, but in its basic form, the coalescent framework does not account for many of the biological processes found in real data including recombination. Recombination is the exchange of genetic material between two chromosomes from the parents that leads to the offspring having a hybrid chromosome containing genetic material from different sources. As a consequence, various segments of the genome will have distinct ancestral histories. The ancestral history of the sample can longer be encoded in a single gene tree, but instead must be captured in a graph structure known as the ancestral recombination graph (ARG) which encapsulates both coalescence and recombination events (Hudson et al., 1990; Griffiths and Marjoram, 1997).

Because the complexity of ARGs explodes with both the number of sequences and the length of the sequences, the likelihood of function of coalescent models of recombinant sequence data is difficult to calculate and computationally intractable. This makes ARG-based inference of chromosome data extremely challenging. To circumvent this issue, the sequentially Markov coalescent (SMC) is a widely used approximation which allows for efficient calculation of the likelihood (McVean and Cardin, 2005; Marjoram and Wall, 2006). The idea behind SMC is to view the full genealogy as a process along the chromosome rather than a process through time. In place of an ARG, we instead now have a sequence of trees where each tree corresponds to a particular locus along the chromosome, and the distribution of a tree depends only on the previous tree of the sequence. Stretches of loci will share a tree until a recombination event occurs, after which the next stretch of loci will then have a new tree until another recombination event occurs. This framework naturally lends itself to likelihood-based inference using a hidden Markov model (HMM) formulation where the observed sequence is the observed alleles at each locus along the sequence, and the hidden process is the sequence of trees relating the relating the samples at those loci.

While there has been much success in population genetics using coalescent HMMs to infer the effective population size, N_e, over time (Li and Durbin, 2011; Sheehan et al., 2013; Schiffels and Durbin, 2014; Terhorst et al., 2017), they require the hidden state to be discrete. Since genealogies are naturally continuous, this requires discretizing the space of trees. There is no salient way to enforce this discretization, and often the choice of discretization can lead to bias in downstream inference. In Chapter 2, we propose a method enabling SMC-based inference that evades the need for discretization and the biases that come with such an approximation. Compared to existing methods ours is faster and more accurate.

In Chapter 4, we propose a solution to a similar problem. In addition to bias, the required discretization of time in coelescent HMMs can also lead to noisy estimates of the population size history. We present a method that uses regularization that encourages smoothness in population

size history estimates. Our method takes advantage of a linear time decoding algorithm to allow for bootstrapping for uncertainty quantification and an automated cross validation procedure to select the correct level of regularization.

1.2 Bayesian Inference in Viral Phylodynamics

Much like other organisms, we can use methods in population genetics to study the evolutionary processes of viruses. However, viruses are unique in that they have short generation times, so evolutionary and ecological processes occur on the same time scale. This suggests that insight into the evolutionary processes about viruses can lend insight into the epidemiological processes (Pybus and Rambaut, 2009). Thus one application of viral phylogenetics is to infer epidemiological parameters such as the effective reproductive number R. The study of how these processes together impact the patterns of viral genetic variation is called phylodynamic inference.

Like in population genetics, phylodynamic inference usually begins with the reconstruction of a tree. For many problems in population genetics, if we were to know the true gene tree that generated the data, estimation of many parameters in interest would follow easily. One caveat is that confounding processes like natural selection, migration, and population structure can bias these results (Chikhi et al., 2018; Mazet et al., 2016). For viruses this limitation is stronger, as the number and effect of confounding processes is larger than for humans. Even if we were to know the true phylogeny of our sample, there is a many-to-one mapping of evolutionary and epidemiological processes that could have resulted in the phylogeny (Volz et al., 2013b). Another issue that arises in phylogenetics is that multiple reconstructions of the phylogeny can explain the data equally well. To account for this, Bayesian methods have been popular in viral phylodynamics as they can integrate out this so called phylogenetic uncertainty (Drummond et al., 2005; Kühnert et al., 2011).

Bayesian phylogenetic inference procedures typically use Markov chain Monte Carlo (MCMC) algorithms to sample from the posterior. Due to the discrete nature of tree topologies, the state space of tree topologies grows astronomically with the number of tips making sampling difficult. Even with several advancements that have accelerated MCMC, to our knowledge, there are no Bayesian phylogenetic methods that can analyze thousands, let alone the millions of sequences at our disposal. Variational inference (Jordan et al., 1999; Wainwright and Jordan, 2008) is an alternative to MCMC that has been relatively unexplored in phylogenetics. The main idea behind variational inference is frame estimation the posterior as an optimization problem; instead of trying to sample from the posterior, we instead approximate the posterior by finding the distribution that minimizes the Kullback-Leibler divergence from a well known family of distributions that are tractable.

Motivated by the vast amount of SARS-CoV-2 sequences since the inception of the pandemic,

we present a new method in Chapter 3 using variational inference with the ability to rapidly analyze tens of thousands of viral sequences to infer epidemiological parameters to address the limitations of MCMC based Bayesian phylogenetic methods. We conduct a simulation study to demonstrate our method is faster than the current state of the arts tool for Bayesian phylogenetic inference, BEAST, without sacrificing accuracy. We apply our method to SARS-CoV-2 data and find that our method agrees with external estimates of the reproductive number using public health data.

CHAPTER 2

Exact decoding of a sequentially Markov coalescent model in genetics

2.1 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 paper, 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 paper 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 computationally much easier. In particular, this modification allows us to leverage recent innovations in change-point detection, leading to algorithms which not only have less bias than existing approaches, but also outperform them computationally. Of course, some tradeoffs are necessary in order to achieve this feat: 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 paper is organized as follows. In Section 2.2 we formally define our data and model, introduce notation, and survey related work. In Section 2.3 we derive our main results: exact and efficient Bayesian and frequentist algorithms for inferring genealogies from genetic variation data. In Section 2.4 we thoroughly benchmark our method, compare it to existing approaches, and provide an application to real data analysis. We provide concluding remarks in Section 2.5.

2.2 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 et al. (2005) and Durrett (2008).

2.2.1 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

$$p(\mathbf{Y} \mid \varphi) = \int_{A \in \mathcal{A}} p(\mathbf{Y} \mid A) p(A \mid \varphi), \tag{2.1}$$

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 data sets. That is because, for large N and H, there are a huge number of ARGs that could have plausibly generated a given data set, such that the complexity of $\mathcal{A}$ explodes as N and H grow. Indeed, (2.1) 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

$$p(\mathbf{Y} \mid \varphi) \approx \int_{X_1, \ldots, X_N} \pi(X_1 \mid \varphi) p(\mathbf{Y}_1 \mid X_1) \prod_{n=2}^{N} p(\mathbf{Y}_n \mid X_n) p(X_n \mid X_{n-1}, \varphi), \tag{2.2}$$

where $\pi(\cdot \mid \varphi)$ is a stationary distribution for the Markov chain $X_1, \ldots, X_N$, $p(X_n \mid X_{n-1}, \varphi)$ is a transition density, and $[\mathbf{Y}_1 \mid \cdots \mid \mathbf{Y}_N] = \mathbf{Y}$ are the data at each site. If the X_i have discrete support, then this represents a hidden Markov model, whence (2.2) can be efficiently evaluated using the forward algorithm. For estimating φ, EM type algorithms are generally preferred, and these require computing the posterior distribution $p(X_1, \ldots, X_N \mid \mathbf{Y}, \varphi)$.

2.2.2 Demographic inference

To make this problem more concrete, in this paper we focus specifically on computing (2.1) when the chromosomes evolve under selective neutrality, and 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, §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

$$\pi(t) = \eta(t) e^{-\int_0^t \eta(s)\, ds}. \tag{2.3}$$

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 other evolutionary phenomena such as natural selection, or mutation rate variation. Estimation of this function is known in the literature as *demographic inference* (Spence et al., 2018). For the remainder of the paper we will focus on this application. To simplify the notation, we suppress explicit dependence on $N_e(t)$ and capture it implicitly through the function π, and we even suppress dependence on π when it is clear from context.

A number of methods have been proposed for performing demographic inference, using various underlying models and sources of data. One class (Gutenkunst et al., 2009; Bhaskar et al., 2015; Jouganous et al., 2017; Kamm et al., 2017, 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 et al., 2013; Rasmussen et al., 2014; Terhorst et al., 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 et al., 2000; Drummond et al., 2005; Minin et al., 2008; Gill et al., 2013).

2.2.3 Our contribution

As discussed in Section 2.1, 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 2.4.4.) 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.

2.2.4 Notation and model

We now fix necessary notation and define the model that is used to prove our results. For ease of exposition, our results focus on the simplest possible case of analyzing a pair of chromosomes ($H = 2$ in the notation of the previous section). In Section 2.3.4 we describe how to extend our results to larger sample sizes

Assume that 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 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, §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

$$f_{X_{n+1}|X_n}(t \mid s) := p(X_{n+1} \in (t, t+dt) \mid X_n = s) = \delta(t-s)e^{-\rho s} + (1-e^{-\rho s})q(t \mid s), \quad (2.4)$$

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 et al., 2011; Li and Durbin, 2011; Carmi et al., 2014). Importantly, they share the common feature that (2.4) is (approximately, in the case of Li and Durbin, 2011) reversible with respect to the coalescent. That is,

$$\pi(s) f_{X_{n+1}|X_n}(t \mid s) = \pi(t) f_{X_{n+1}|X_n}(s \mid t), \quad (2.5)$$

where π is the stationary measure in equation (2.3). This can be verified in each of the above models by checking the detailed balance condition (Hobolth and Jensen, 2014).

2.2.5 Connection to changepoint detection

Our work is motivated by the observation that (2.4) 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 is hypothesized to have been realized from 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 standard 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 (2.4) makes the problem non-standard from a changepoint perspective. It is tempting to simply ignore it. Indeed, if $q(t \mid s)$ were replaced by some function $\underline{\pi}(t)$ which did not depend on s, then (2.4) 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 efficient methods have been developed to analyze them in both Bayesian (Barry and Hartigan, 1993; Fearnhead, 2006) and frequentist (Jackson et al., 2005; Killick et al., 2012) settings.

2.2.6 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 (2.5). Given this desideratum, the obvious choice for $\underline{\pi}$ becomes

$$\underline{\pi}(t) \propto t\pi(t), \tag{2.6}$$

leading to the modified transition density

$$f^R_{X_{n+1}|X_n}(t \mid s) = \delta(t-s)e^{-\rho s} + (1 - e^{-\rho s})\underline{\pi}(t). \tag{2.7}$$

Checking the detailed balance condition (2.5), we obtain

$$\pi(s)(1 - e^{-\rho s})t\pi(t) \stackrel{?}{=} \pi(t)(1 - e^{-\rho t})s\pi(s), \quad s \neq t. \tag{2.8}$$

Though (2.8) 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.

2.2.7 Prior work

The Markov chain defined by (2.7) 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).

Interpreted broadly, many other lines of research are related to our work, because SMC plays such a fundamental inferential role in genetics. Haplotype copying models (Li and Stephens, 2003) have been used to study natural selection (Voight et al., 2006), ancestry (Price et al., 2009), population structure (Lawson et al., 2012), and population history (Gay et al., 2007); and to perform haplotype phasing and imputation (Scheet and Stephens, 2006; Marchini et al., 2007; Howie et al., 2009). Similarly, PSMC and related methods for inferring population size history (Li and Durbin, 2011; Schiffels and Durbin, 2014; Terhorst et al., 2017; Steinrücken et al., 2019) are now a standard component of population genetic analysis, and have been cited in thousands of papers.

2.3 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 2.6.1–2.6.4 in the supplementary material. For the reader's convenience, the various notations introduced in this section are listed in Table 2.4.

2.3.1 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, §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*

$$p(X_n \mid Y_{1:N}) = \frac{f(X_n)g(X_n)}{\pi(X_n)}. \tag{2.9}$$

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 2.6.5.

2.3.2 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 (2.18) 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) (and essentially to 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

$$p(Y_{u:v}, \overline{R}_{u:v}, R_v) = \int_x x^{\mathbf{1}\{u>1\}} \pi(x) \rho x e^{-\rho \Delta x} \prod_{i=u}^{v} e^{-\theta x} (\theta x)^{Y_i} / Y_i! =: P(u, v). \tag{2.10}$$

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 $\underline{\pi}$; cf. equations 2.2 and 2.7.

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

$$P_{-1}(u, N) = p(Y_{u:N}, \overline{R}_{u:N}) = \int_x x^{\mathbf{1}\{u>1\}} \pi(x) e^{-\rho\Delta x} \prod_{i=u}^{N} e^{-\theta x} (\theta x)^{Y_i} / Y_i!. \tag{2.11}$$

This algorithm can be used whenever (2.10) 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)

$$Q(u) = \sum_{v=u}^{N-1} P(u, v) Q(v + 1) + P_{-1}(u, N) \tag{2.12}$$

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 (2.12) 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 (2.10). Hence, the posterior distribution of

the segment height x conditional on the underlying segmentation is

$$p(x \mid \overline{Y}_{u:v}, R_{u:v}, R_v) \propto x^{\mathbf{1}\{u>1\}} \pi(x) \rho x e^{-\rho \Delta x} e^{-\theta x} x^{\sum_{i=u}^{v} Y_i}. \tag{2.13}$$

If $\pi(x)$ is a gamma (mixture), then (2.13) is also a gamma mixture, and hence easy to sample.

2.3.3 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,

$$X^*_{1:N} := \arg\max_{X_{1:N}} p(X_{1:N}, Y_{1:N}). \tag{2.14}$$

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 the optimization problem (2.14) 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

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

This is the usual Viterbi dynamic program, but defined over a continuous instead of discrete domain. By standard arguments (Bishop, 2006, §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 (2.14) would be solved. In general, it is not obvious how to accomplish this, since $V_n(t)$ is a function, i.e. 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 2.6.3) 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 2.6.6.

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 $t \in [\boldsymbol{\tau}_n[\![k]\!], \boldsymbol{\tau}_n[\![k+1]\!])$,*

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

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 1).

In Section 2.4.2, 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 2.7). 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 2.6.7 we derive a generalization of Proposition 3 which allows us to efficiently compute other paths which are suboptimal with respect to (2.14), but have better pointwise accuracy, thus enabling a range of possible decodings.

2.3.4 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 amongst 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 $\{1, \dots, H\} \times \mathbb{R}_{>0}$, 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 et al., 2013; Spence et al., 2018; Steinrücken et al., 2019).

Modifying our methods to utilize 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.

2.4 Results

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

2.4.1 Insensitivity of the posterior to the prior

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 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.

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 2.6.8.1.

Supplemental Figures 2.3 and 2.4 show the Viterbi path and the posterior heatmap for one run of each scenario of the simulation. From Figure 2.3, 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 2.4). 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 2.5 shows the average absolute error for the two priors over the 25 simulations. In terms of absolute error, the renewal prior does 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. Table 2.6 shows that CSC is slightly better in relative error. However, in general the differences are minor, and both the tables confirm our hypothesis that the posterior is fairly insensitive to the choice of prior.

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 2.6.8.2).

We display the posterior of one pair of chromosomes for all 9 pairs of demographies used as data generation and demographic priors in Figure 2.6. The plots show that regardless of which demographic prior was used, the resulting posteriors all had the same shape. Table 2.10 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 2.11) tell a similar story. Given the large differences between the three demographic models (Figure 2.5), 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.

2.4.2 Comparison of Bayesian and frequentist inferences

In Section 2.3 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 2.4.1. For the Bayesian method we sample 200 paths from the posterior and take the median to compare against the MAP path.

Figure 2.7 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 2.12 and 2.13 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 2.5 and 2.6, the MAP path performs similarly to, and the Bayesian method outperforms, the posterior decoding of the discretized SMC models used in Section 2.4.1.

2.4.3 Empirical time complexity

In Section 2.3.2, 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 2.8 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 2.3.4), the amount of work performed by our algorithms scales linearly in the panel size H. We further verified (Figure 2.9) that the scaling is linear in both panel size (H) and chromosome length (N); in Figure 2.10, 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 (2.12) before the truncation threshold was met (Figure 2.11). On average, the number seems to be bounded by a small constant as the dynamic program (2.12) 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 2.14) 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.

Taken together, these simulations suggest the amount of work performed by our algorithms scales linearly with the number of decoded haplotypes, and, crucially, does not grow with the length of the decoded sequence N. Based on these results, we conjecture that the average case time complexity of our methods is $O(HN)$, which would match the running time of the most efficient existing methods for decoding the SMC (Harris et al., 2014; Palamara et al., 2018). Proving this assertion rigorously appears difficult, and is left to future work.

2.4.4 Applications

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

2.4.4.1 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 2.2.2) 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 et al., 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 et al., 2017) is a generalization of PSMC that does not require phased data which scales to larger sample sizes. Additionally, SMC++ utilizes the more accurate CSC model (cf. Section 2.4.1), whereas PSMC is based on SMC.

As noted in Section 2.1, 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 equation (2.3), 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

$$p(X_{i_1} \mid Y_{1:N}) \times p(X_{i_2} \mid Y_{1:N}) \times \cdots \times p(X_{i_k} \mid Y_{1:N}), \tag{2.15}$$

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

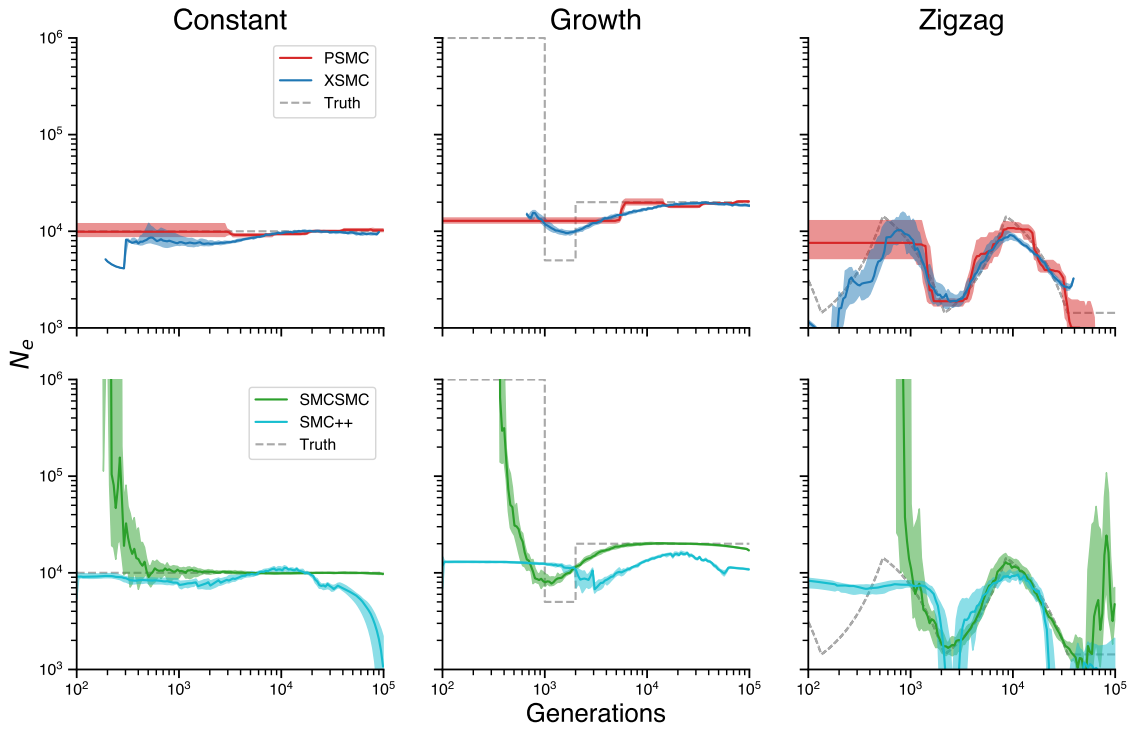

Figure 2.1: Comparison of XSMC, PSMC, SMCSMC, SMC++ on various simulated size histories.

We first compared the performance of our method with PSMC, SMC++, and SMCSMC on simulated data. Figure 2.1 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 (2.15) 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 XSMC exhibit less bias. For PSMC, there is clear bias from the piecewise-constant model class it uses to perform

Table 2.1: 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

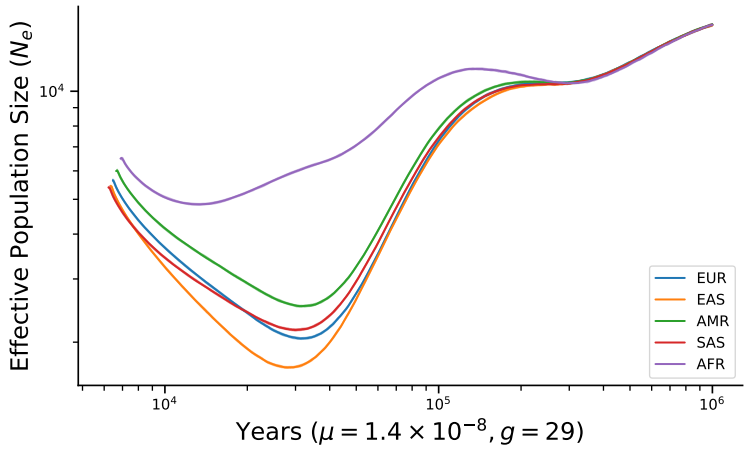

Figure 2.2: Result of fitting XSMC to 1000 Genomes data. For each superpopulation, 20 samples were chosen. Solid line denotes the median across all samples, and shaded bands denote the interquartile range.

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. On the other hand, while SMCSMC does a great job estimating after 10^3 generations, it hallucinates a massive increase towards the present. SMC++ exhibits a slight downward bias towards 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 2,000–1,000 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 et al., 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 towards the present and further back in the past.

Table 2.1 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 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 Fig-

ure 2.2. Broadly speaking, our method agrees with other recently-published estimates (Li and Durbin, 2011; Terhorst et al., 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. 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 minutes on a 12-core workstation. A single human genomes (all 22 autosomes) can be analyzed in about 30 seconds.

2.4.4.2 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 et al., 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 paper, 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 data sets 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 2.2 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 2.3 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.

2.5 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.

Table 2.2: Proportion of segregating sites where XSMC and LS agree on the GNN using the MAP path or posterior mode.

	Panel Size	Chromosome 10	Chromosome 13
MAP	2	0.9129 (0.0073)	0.9166 (0.0070)
	4	0.8473 (0.0014)	0.8533 (0.0015)
	10	0.8365 (0.0021)	0.8440 (0.0027)
	25	0.8423 (0.0032)	0.8413 (0.0033)
	100	0.8619 (0.0044)	0.8468 (0.0067)
Mode	2	0.8435 (0.0128)	0.8457 (0.0128)
	4	0.5989 (0.0025)	0.6006 (0.0024)
	10	0.4102 (0.0015)	0.4083 (0.0014)
	25	0.3564 (0.0013)	0.3557 (0.0020)
	100	0.3362 (0.0014)	0.3381 (0.0019)

Table 2.3: Proportion of segregating sites that XSMC finds the more closely related haplotype than LS conditional on the two methods inferring different haplotypes at that site, using the MAP path or posterior mode.

	Panel Size	Chromosome 10	Chromosome 13
MAP	2	0.5258 (0.0031)	0.5653 (0.0035)
	4	0.5056 (0.0027)	0.5383 (0.0039)
	10	0.4842 (0.0045)	0.5036 (0.0070)
	25	0.5068 (0.0052)	0.5411 (0.0072)
	100	0.5990 (0.0152)	0.5682 (0.0101)
Mode	2	0.5326 (0.0045)	0.5393 (0.0034)
	4	0.5298 (0.0053)	0.5302 (0.0068)
	10	0.5435 (0.0024)	0.5442 (0.0030)
	25	0.5320 (0.0021)	0.5292 (0.0044)
	100	0.5161 (0.0023)	0.5189 (0.0036)

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 2.3.3 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 2.6.6 to perform the pointwise maximization in equation (2.19). 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.

Supplement to "Exact Decoding of the Sequential Markov Coalescent: In the supplement we present supporting lemmas, proofs of the theorems, and additional plots and tables. (pdf)

Code: All of the data analyzed in this paper 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 manuscript is available at `https://terhorst.github.io/xsmc/paper`.

Acknowledgements: This research was supported by the National Science Foundation grant number DMS-2052653, and a Graduate Research Fellowship.

2.6 Appendix

2.6.1 Proof of Proposition 1

The proof requires only a few simple facts from Bayesian analysis.

Fact 1. If $X \sim \Gamma(a, b)$ and $Y \mid X \sim \text{Poisson}(\theta X)$, then $X \mid Y \sim \Gamma(a + Y, b + \theta)$.

Fact 2. If $X \sim \mathcal{M}_\Gamma(K)$ and $Y \mid X \sim \text{Poisson}(X)$, then $X \mid Y \sim \mathcal{M}_\Gamma(K)$.

Fact 3. If $X_n \mid Y_n \sim \mathcal{M}_\Gamma(nK)$ and $\pi \in \mathcal{M}_\Gamma(K)$, then under the renewal approximation (2.7), $X_{n+1} \mid Y_n \sim \mathcal{M}_\Gamma((n+1)K)$.

The first two facts are well-known consequences of conjugacy. To establish the third, note that

$$\begin{aligned} p(X_{n+1} \mid Y_n) &= \int_{X_n} f^R_{X_{n+1}|X_n}(X_{n+1} \mid X_n) p(X_n|Y_n) \\ &= \int_{X_n} [\delta(X_n - X_{n+1})e^{-\rho X_n} + (1 - e^{-\rho X_n})\underline{\pi}(X_{n+1})] p(X_n|Y_n) \\ &= \underbrace{p(X_n = X_{n+1} \mid Y_n)e^{-\rho X_{n+1}}}_{\in \mathcal{M}_\Gamma(nK)} + \underline{\pi}(X_{n+1}) \underbrace{\int_{X_n} (1 - e^{-\rho X_n}) p(X_n|Y_n)}_{=\text{constant}} \qquad (2.16) \\ &\in \mathcal{M}_\Gamma((n+1)K). \end{aligned}$$

Proof of Proposition. By induction on n. The case $n = 1$ follows from Facts 1 and 2. And, if the claim holds for $n = i$, then $X_{i+1} \mid Y_{1:i} \sim \mathcal{M}_\Gamma((i+1)K)$ by Fact 3. Since $Y_{i+1} \perp Y_{1:i} \mid X_{i+1}$, Fact 2 implies

$$(X_{i+1} \mid Y_{1:i}) \mid Y_{i+1} = X_{i+1} \mid Y_{1:i+1} \in \mathcal{M}_\Gamma((i+1)K). \qquad (2.17)$$

□

2.6.2 Proof of Proposition 2

Define $\overrightarrow{\alpha}(X_n) = p(X_n \mid Y_{1:n})$ to be the quantity derived in Proposition 1, and let $\overleftarrow{\alpha}(X_{n+1})$ be obtained by running the forward algorithm from that proposition on the reversed sequence $(Y_N, Y_{N-1}, \ldots, Y_{n+1})$. By reversibility, $\overleftarrow{\alpha}(X_{n+1}) = p(X_{n+1} \mid Y_{n+1:N})$ and hence

$$\begin{aligned} p(X_n \mid Y_{1:N}) &\propto \overrightarrow{\alpha}(X_n) p(Y_{n+1:N} \mid X_n) \\ &\propto \frac{\overrightarrow{\alpha}(X_n) p(X_n \mid Y_{n+1:N})}{\pi(X_n)} \\ &= \frac{\overrightarrow{\alpha}(X_n) \int_{X_{n+1}} p(X_n \mid X_{n+1}) \overleftarrow{\alpha}(X_{n+1})}{\pi(X_n)}. \qquad (2.18) \end{aligned}$$

By Proposition 1, $\overrightarrow{\alpha}(X_n) \in \mathcal{M}_\Gamma(Kn)$ and $\overleftarrow{\alpha}(X_{n+1}) \in \mathcal{M}_\Gamma(K(N-n-1))$. Finally, using the same argument that established equation (2.17),

$$\int_{X_{n+1}} p(X_n \mid X_{n+1}) \overleftarrow{\alpha}(X_{n+1}) \in \mathcal{M}_\Gamma(K(N-n)).$$

□

Remark. Instead of the reversibility argument used to prove Proposition 2, we could have used ideas from the proof of Proposition 1 to derive a sum-of-gammas representation for the rescaled backward function

$$\beta(X_n) = p(Y_{n+1:N} \mid X_n)/p(Y_{n+1:N} \mid Y_{1:n}),$$

whence $p(X_n \mid Y_{1:N}) = \alpha(X_n)\beta(X_n)$. We experimented with this approach, but found that it was numerically unstable for long sequences: whereas the mixture coefficients of $\alpha(X_n)$ live in the simplex, the backwards function $\beta(X_n)$ is not a probability distribution in X_n, and we observed that the mixture coefficients tended to diverge when N was large. It seems that the rational representation (2.9) has superior numerical properties.

2.6.3 Proof of Proposition 3

To prove the result we need a few lemmas. We omit the trivial proofs of the first two.

Lemma 1. *$\mathcal{V}_K$ contains all piecewise constant and piecewise linear functions with K pieces. For all i, $V_i \subset V_{i+1}$. If $c \in \mathbb{R}$ and $f \in \mathcal{V}_i, g \in \mathcal{V}_j$, then $cf \in \mathcal{V}_i$, $f + g \in \mathcal{V}_{i+j}$ and $\max\{f, g\} \in \mathcal{V}_{i+j}$.*

Lemma 2. *Let $e_n(t) := \log p(Y_n \mid X_n = t)$. Then*

$$e_n(t) = -\theta t + Y_n \log(\theta t) - \log Y_n! \in \mathcal{V}_1.$$

Lemma 3. *Suppose that $N_e(t) \in \mathcal{V}_K$ is piecewise constant. Then $\log \pi(t), \log \underline{\pi}(t) \in \mathcal{V}_K$.*

Proof. If $N_e(t)$ is piecewise constant then so too is $\log \eta(t) = -\log N_e(t)$. Also, $R(t) := \int_0^t \eta(s)\, ds \in \mathcal{V}_K$ is piecewise linear on the same set of breakpoints. Hence,

$$\log \pi(t) = \log \eta(t) - R(t).$$

By Lemma 1, $\log \pi(t) \in \mathcal{V}_K$. Similarly,

$$\log \underline{\pi}(t) = \text{const.} + \log t + \log \pi(t),$$

and $\log \underline{\pi}(t) \in \mathcal{V}_K$ using Lemma 1. □

Proof of Proposition. By induction on n. From Lemma 3 we have that $\log \pi(t) \in \mathcal{V}_K$, and from Lemma 2 we have that $e_1(t) \in \mathcal{V}_1$. Thus, for $n = 1$,

$$V_1(t) = \log \pi(t) + e_1(t) \in \mathcal{V}_K$$

as claimed. Again when $N_e(t)$ is piecewise constant then so too is $R(t) := \int_0^t \eta(s)\, ds$. So, For the inductive step, we have

$$V_{n+1}(t) = e_{n+1}(t) + \max\Big\{ \underbrace{-\rho t + V_n(t)}_{(A)}, \underbrace{\log \underline{\pi}(t) + \max_{s \neq t} V_n(s) + \log(\rho s)}_{(B)} \Big\}. \tag{2.19}$$

By the induction hypothesis and Lemmas 1-3, both (A) and (B) are in $\mathcal{V}_{k_1}$ for some k_1. Then, another application of the lemmas shows that in fact the entire right-hand side of (2.19) is in $\mathcal{V}_{k_2}$ for some (possibly larger) k_2. □

2.6.4 Proof of Proposition 4

Proof. In view of equation (2.19), note that for fixed t, we can unwind the recursion $V_{n+1}(t) = e_{n+1}(t) - \rho t + V_n(t)$ until we reach an index i where $(A) < (B)$. By continuity, this index is the same for all $t \in [\boldsymbol{\tau}_n[\![k]\!], \boldsymbol{\tau}_n[\![k+1]\!])$. Denote the vector of such indices associated with each interval by $\mathbf{i}_n$, and let

$$\mathbf{C}_n[\![k]\!] = \max_{s \neq t} V_{\mathbf{i}_n[\![k]\!]}(s) + \log(\rho s) - \prod_{i=\mathbf{i}_n[\![k]\!]}^{n} \log Y_n!.$$

Then

$$V_n(t) = \mathbf{C}_n[\![k]\!] + \prod_{i=\mathbf{i}_n[\![k]\!]}^{n} \log Y_n! + \log \underline{\pi}(t) + \sum_{i=\mathbf{i}_n[\![k]\!]}^{n} e_i(t) - t\rho(n - \mathbf{i}_n[\![k]\!]),$$

so the claim follows by using Lemma 2 to expand the sum. □

2.6.5 Forward recursion constants

In this section we derive the exact mixing weights and scale/shape parameters for the mixture representation proved in Proposition 2. Define $\gamma(x; a, b)$ to be the PDF of the gamma distribution with mean a/b and variance a/b^2,

$$\gamma(x; a, b) = \frac{x^{a-1} e^{-bx}}{\Gamma(a) b^a}.$$

To conserve notation, in this section we use the following array-based conventions for vector expressions:

- Scalar functions operate on vectors in a component-wise manner. For example, if $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^k$ then
$$2\mathbf{x}e^{\mathbf{y}}/\mathbf{z} = \langle 2x_1 e^{y_1}/z_1, \ldots, 2x_n e^{y_n}/z_n \rangle .$$
In particular, for vectors $\boldsymbol{\alpha}, \mathbf{a}, \mathbf{b} \in \mathbb{R}^n$,
$$\boldsymbol{\alpha}\gamma(\cdot; \mathbf{a}, \mathbf{b}) = \langle \alpha_1 \gamma(\cdot; a_1, b_1), \ldots, \alpha_n \gamma(\cdot; a_n, b_n) \rangle .$$
- A binary operation between a scalar and a vector "broadcasts" the scalar to the dimension of the vector. For example, $1 + \mathbf{x} = \langle 1 + x_1, \ldots, 1 + x_n \rangle$.
- To refer to individual entries of subscripted vectors, we will use the notation $x_n[\![i]\!]$. A subvector ("slice") of $\mathbf{x}_n$ of length $i \leq n$ is denoted $\mathbf{x}_n[\![1 : i]\!] = \langle \mathbf{x}_n[\![1]\!], \mathbf{x}_n[\![2]\!], \ldots, \mathbf{x}_n[\![i]\!] \rangle$.
- The sum of all the entries of $\mathbf{x}$ is denoted $\sum \mathbf{x} = \sum_{i=1}^{n} x_i$.

Additionally, we define the following function for later use:

$$f(a, b, c, y) = \frac{\theta^y b^a}{(b + c)^{a+y}} \frac{\Gamma(a + y)}{\Gamma(a)\Gamma(1 + y)}. \tag{2.20}$$

We prove the following theorem in the case where π is a gamma distribution. Extending the proof to gamma mixtures requires no new ideas, only notation; details are left to the reader.

Theorem 1. *Suppose that $\pi(x) = \gamma(x; a_0, b_0)$. For each $n \in [N]$ let $\mathbf{a}_n, \mathbf{b}_n \in \mathbb{R}^n$ be defined by*

$$\begin{aligned} \mathbf{a}_n[\![i]\!] &= 1 + a_0 + \sum_{j=1}^{n} Y_j \\ \mathbf{b}_n[\![i]\!] &= 1 + \theta + (n - i)(\theta + \rho), \end{aligned}$$

and define $\boldsymbol{\alpha}_n^0, \boldsymbol{\alpha}_n \in \mathbb{R}^n$ and $C_n \in \mathbb{R}$ by the recursions

$$\begin{aligned} \boldsymbol{\alpha}_1 &= 1 \\ \boldsymbol{\alpha}_n^0[\![1 : n - 1]\!] &= \boldsymbol{\alpha}_{n-1} f(\mathbf{a}_{n-1}, \mathbf{b}_{n-1}, \theta + \rho, Y_n) \\ \boldsymbol{\alpha}_n^0[\![n]\!] &= f(a_0, b_0, \theta, Y_n) \left(1 - \textstyle\sum \boldsymbol{\alpha}_{n-1} \left[\mathbf{b}_{n-1}/(\mathbf{b}_{n-1} + \rho)\right]^{\mathbf{a}_{n-1}}\right) \\ C_n &= \textstyle\sum \boldsymbol{\alpha}_n^0 \\ \boldsymbol{\alpha}_n &= \boldsymbol{\alpha}_n^0 / C_n. \end{aligned}$$

Then

$$p(X_n \mid Y_{1:n}) = \sum \boldsymbol{\alpha}_n \gamma(x; \mathbf{a}_n, \mathbf{b}_n), \tag{2.21}$$

and additionally $C_n = p(Y_n \mid Y_{1:n-1})$.

Proof. By induction on n. The base case $p(X_1 \mid Y_1)$ follows from conjugacy of the gamma and Poisson distributions (cf. Fact 1 in Appendix 2.6.1.) For the general case, assume that $p(X_n \mid Y_{1:n})$ has the form shown in (2.21). Then

$$\begin{aligned} p(X_{n+1} \mid Y_{1:n+1}) &\propto \int_{X_n} p(Y_{n+1}, X_{n+1}, X_n \mid Y_{1:n}), \\ &= p(Y_{n+1} \mid X_{n+1}) \int_{X_n} p(X_{n+1} \mid X_n) p(X_n \mid Y_{1:n}), \end{aligned} \tag{2.22}$$

where the constant of proportionality $C_{n+1} = p(Y_{n+1} \mid Y_{1:n})$ does not depend on X_{n+1}. Using the transition rule (2.4), this implies

$$\begin{aligned} &\int_{X_n} p(X_{n+1} \mid X_n) p(X_n \mid Y_{1:n}) \\ &= \int_{X_n} [\delta(X_n - X_{n+1}) e^{-\rho X_n} + (1 - e^{-\rho X_n}) \underline{\pi}(X_{n+1})] p(X_n \mid Y_{1:n}) \\ &= e^{-\rho X_{n+1}} p(X_n = X_{n+1} \mid Y_{1:n}) + \underline{\pi}(X_{n+1}) \int_{X_n} (1 - e^{-\rho X_n}) p(X_n \mid Y_{1:n}) \\ &= e^{-\rho X_{n+1}} p(X_n = X_{n+1} \mid Y_{1:n}) + \underline{\pi}(X_{n+1}) \left[1 - \int_{X_n} e^{-\rho X_n} p(X_n \mid Y_{1:n}) \right]. \end{aligned}$$

Now, by the inductive hypothesis and the identity

$$\gamma(x; a, b) x^c e^{-dx} = b^a (b + d)^{-(a+c)} \frac{\Gamma(a + c)}{\Gamma(a)} \gamma(x; a + c, b + d) \tag{2.23}$$

we obtain, for $\boldsymbol{\alpha}'_{\boldsymbol{n}} = \boldsymbol{\alpha}_n \left[\mathbf{b}_n / (\mathbf{b}_n + \rho) \right]^{\mathbf{a}_n}$,

$$\int_{X_n} p(X_{n+1} \mid X_n) p(X_n \mid Y_{1:n}) = \sum \boldsymbol{\alpha}'_n \gamma(X_{n+1}; \mathbf{a}_n, \mathbf{b}_n + \rho) + (1 - \sum \boldsymbol{\alpha}'_n) \, \underline{\pi}(X_{n+1}).$$

Multiplying through by

$$p(Y_{n+1} \mid X_{n+1}) = e^{-\theta X_{n+1}} (\theta X_{n+1})^{Y_{n+1}} / Y_{n+1}!$$

yields

$$p(X_{n+1} \mid Y_{1:n+1}) \propto \sum \boldsymbol{\alpha}_n \underbrace{f(\mathbf{a}_n, \mathbf{b}_n, \theta+\rho, Y_{n+1})}_{(\mathbf{A})} \gamma(X_{n+1}; \mathbf{a}_n + Y_{n+1}, \mathbf{b}_n + \theta + \rho) + \underbrace{f(a_0, b_0, \theta, Y_{n+1})\,(1 - \sum \boldsymbol{\alpha}'_{\boldsymbol{n}})}_{(B)} \gamma(X_{n+1}; 1 + a_0 + Y_{n+1}, b_0 + \theta), \quad (2.24)$$

by (2.20) and (2.23). If we make the additional definitions

$$\mathbf{a}_{n+1}[\![1:n]\!] = a_n + Y_{n+1} \quad (2.25)$$

$$\mathbf{a}_{n+1}[\![n+1]\!] = 1 + a_0 + Y_{n+1} \quad (2.26)$$

$$\mathbf{b}_{n+1}[\![1:n]\!] = \theta + \rho + b_n \quad (2.27)$$

$$\mathbf{b}_{n+1}[\![n+1]\!] = 1 + \theta \quad (2.28)$$

$$C_{n+1}^{-1} = \sum \mathbf{A} + B \quad (2.29)$$

$$\boldsymbol{\alpha}_{n+1}[\![1:n]\!] = C_{n+1}^{-1}\mathbf{A} \quad (2.30)$$

$$\boldsymbol{\alpha}_{n+1}[\![n+1]\!] = C_{n+1}^{-1}B \quad (2.31)$$

then (2.24) can be written as

$$p(X_{n+1} \mid Y_{1:n+1}) = \sum \boldsymbol{\alpha}_{n+1} \gamma(X_{n+1}; \mathbf{a}_{n+1}, \mathbf{b}_{n+1}),$$

completing the proof. The recursive definition for $\boldsymbol{\alpha}_{n+1}$ follows from (2.30) and (2.31), and the representations for $\mathbf{a}_{n+1}$ and $\mathbf{b}_{n+1}$ follow from (2.25)–(2.28). Finally, note that C_{n+1} is precisely the constant of proportionality in (2.22) and therefore equals the conditional evidence $p(Y_{n+1} \mid Y_{1:n})$. □

2.6.6 Optimization in $\mathcal{V}_K$

In this section we derive procedures for finding the pointwise maximum $\max\{f, g\}$ when $f, g \in \mathcal{V}_K$, as well as $\max(f)$ when f is in that class.

2.6.6.1 Computing the maximum

Maximizing (or minimizing) any individual function $f \in \mathcal{V}_K$ is easily accomplished since the derivative of f is of the form $f'(t) = a + b/t$ over each interval. Hence there is at most one critical point at $t = -b/a$ at which the function attains an extreme value; otherwise the maximum occurs

at one of the end points of the interval. We can therefore consider the function value at all possible critical points, as well as the values at the interval endpoints, in order to locate the maximum. This requires $O(K)$ time.

2.6.6.2 Computing the pointwise maximum

Turning to the problem of maximizing/minimizing pairs of functions, by enlarging K if necessary, we can without loss of generality assume that f and g are defined on the same set of breakpoints. Then it suffices to show how to find the zeros (if any) of the function $h = f - g \in \mathcal{V}_K$ over any given interval.

Accordingly, let $h(t) = at + b\log t + c$ for $\tau \in I := [\tau_1, \tau_2)$. By the change of variables $-\log t \to u$ it is equivalent, and slightly simplifies the math, to find the zeros of $h(u) = ae^{-u} - bu + c$ over an arbitrary interval I. Since interchanging the roles of f and g does not change the result, we may also assume that $a \geq 0$, and if $a = 0$ then we may assume that $b \geq 0$.

Let $w = -ae^{c/b}/b$. If $w \geq 0$ then $h(u)$ has a single real root $u_0 = W_0(w) - c/b$, where $W_0(x)$ denotes the principal branch of the Lambert W function (DLMF, §4.13). If $-1/e \leq w_0 < 0$ then $h(u)$ has two real roots, one at u_0 and the other at $u_1 = W_{-1}(w) - c/b$ where W_{-1} is the -1 branch of the Lambert W function.

We will use repeatedly the fact that a trivial solution exists whenever h can be shown to be globally decreasing, since:

- If $h(\tau_2) \geq 0$ then the function is non-negative over I, so the maximum is f.
- If $h(\tau_1) < 0$ then the function is negative over I, so the maximum is g.
- Else the function has a single root $u_0 \in I$, so the maximum is f on $[\tau_1, u_0)$ and g on $[u_0, \tau_2)$.

To find the zeros of $h(u)$, we proceed by cases:

- If $b = 0$:
  - If $a = 0$ then $h = c$, so the maximum over I is either f or g depending on the sign of c.
  - Else ($a \geq 0, b = 0$):
    - If $c \geq 0$ then $h = f - g \geq 0$ so the maximum over I is f.
    - Else, we have $h' = -uae^{-u} + c < 0$ so the function is decreasing.
- If $a = 0$ then we assume that $b \geq 0$. Then $h'(u) = -b \leq 0$, so h is decreasing.
- Else ($a > 0, b \neq 0$):
  - If $b > 0$ then $h'(u) = -ae^{-u} - b < 0$ so h is decreasing.

- Else we have $h''(u) = ae^{-u} > 0$ so the function is convex with a global minimum at $u^* = \log(-a/b)$:
  - If $h(u^*) > 0$ then the function is non-negative so the maximum is f.
  - Otherwise, h is convex with

$$\liminf_{u \to -\infty} h(u) = \liminf_{u \to \infty} h(u) = \infty,$$

    so it has two real roots u_0 and u_1. Without loss of generality assume $u_0 \leq u_1$. There are $\binom{4}{2}$ cases to consider depending on the ordering of u_0, u_1, τ_1, τ_2. For example, if $\tau_1 < u_0 < u_1 < \tau_2$ then h is positive on $[\tau_1, u_0)$, negative on $[u_0, u_1)$ and positive on $[u_1, \tau_2)$, leading to a pointwise maximum function which takes on the values f, g, f on those three intervals. The other five cases are handled similarly, and we omit the details.

The running time of this procedure is $O(1)$ assuming we can evaluate $W_n(w)$ in constant time. Thus, to find the pointwise maximum of f and g when both have are defined on K pieces takes $O(K)$ time.

2.6.7 Non-MAP paths

The MAP path $X_{1:N}^{\text{MAP}}$ solves the optimization problem

$$\begin{aligned} X_{1:N}^{\text{MAP}} &= \arg\max_{Z_{1:N}} p(X_{1:N} = Z_{1:N} \mid Y_{1:N}) \\ &= \arg\min_{Z_{1:N}} \mathbb{E}_{X_{1:N}|Y_{1:N}} \mathbf{1}\{X_{1:N} \neq Z_{1:N}\}, \end{aligned} \tag{2.32}$$

so the Viterbi algorithm can be interpreted as minimizing risk with respect to the loss function $\ell_{\text{MAP}}(x, y) = \mathbf{1}\{x \neq y\}$, where x, y are paths, and $x = y$ if they are equal at every position. This loss function is "global" in that paths incur equal loss irrespective of whether they mismatch the true path at one position or all of them; there is no benefit to improving the match at a particular position.

On the opposite end of the spectrum, the pointwise posterior mode

$$\begin{aligned} X_{1:N}^{\text{PM}} &:= (\arg\max_{Z_1} p(X_1 = Z_1 \mid Y_{1:N}), \ldots, \arg\max_{Z_n} p(X_n = Z_n \mid Y_{1:N}) \\ &= \arg\min_{Z_{1:N}} \sum_{i=1}^{n} \mathbb{E}_{X_i|Y_{1:N}} \mathbf{1}\{X_i \neq Z_i\} \end{aligned}$$

is "local", placing no emphasis on paths that are continuous from one position to the next. Indeed, from Theorem 1 and Appendix 2.6.5, we can see that $\arg\max p(X_i \mid Y_{1:N}) \neq \arg\max p(X_{i+1} \mid Y_{1:N})$ almost surely for all i, so that $X_{1:N}^{\text{PM}}$ has a changepoint at every position and thus vanishingly small prior probability for large N.

For ordinary HMMs, it is possible to algorithmically interpolate between these these two extremes, resulting in paths that achieve better pointwise accuracy than $X_{1:N}^{\text{MAP}}$ and higher prior likelihood than $X_{1:N}^{\text{PM}}$ Yau and Holmes (2013); Lember and Koloydenko (2014). However, these algorithms assume a discrete state space, and it is unclear whether they can be extended to our setting. Instead, we propose a simple modification of our method which has a straightforward interpretation as penalized changepoint detection.

To build the connection, note that we can write the optimization in (2.14) equivalently by representing $X_{1:N}$ by the locations and heights of each segment, $\boldsymbol{\tau}, \mathbf{x} \in \mathbb{R}^K$, such that

$$\begin{aligned} 1 = \tau_1 < \cdots < \tau_K < \tau_{K+1} = N + 1 \\ X_{\tau_k} = X_{\tau_k+1} = \cdots = X_{\tau_{k+1}-1} = x_i, \quad k = 1, \ldots, K. \end{aligned}$$

Then, we can rewrite the complete likelihood as

$$p(X_{1:N}, Y_{1:N}) = p(\boldsymbol{\tau}, \mathbf{x}, Y_{1:N}) = \prod_{k=1}^{K} p(Y_{\tau_k:\tau_{k+1}-1}, \boldsymbol{\tau}, \boldsymbol{x}) \tag{2.33}$$

where

$$p(Y_{\tau_k:\tau_{k+1}-1} \mid \boldsymbol{\tau}, \boldsymbol{x}) = x_k^{\mathbf{1}_{\{k>1\}}} \pi(x_k) (\rho x_k)^{\mathbf{1}_{\{k<K\}}} \frac{e^{-(\rho+\theta)\Delta_k x_k}}{\prod_{i=\tau_k}^{\tau_{k+1}-1} Y_i!} (\theta x_k)^{\bar{Y}_{\tau_k:\tau_{k+1}-1}}.$$

and $\Delta_k = \tau_{k+1} - \tau_k$. Under the renewal approximation, for fixed $\boldsymbol{\tau}$, (2.33) separates into a series of simpler one-dimensional optimization problems:

$$\begin{aligned} \max_{\boldsymbol{\tau}, \mathbf{x}} p(\boldsymbol{\tau}, \mathbf{x}, Y_{1:N}) &= \max_{\boldsymbol{\tau}} \max_{\mathbf{x}} p(\boldsymbol{\tau}, \mathbf{x}, Y_{1:N}) \\ &= \max_{\boldsymbol{\tau}} \prod_{k=1}^{|\boldsymbol{\tau}|} \max_{x_k} p(Y_{\tau_k:\tau_{k+1}-1}, \tau_k, \tau_{k+1}, x_k). \end{aligned} \tag{2.34}$$

where we abused notation to write $|\boldsymbol{\tau}|$ for the dimension of (i.e. the number of changepoints in) $\boldsymbol{\tau}$. Taking the log of equation (2.34), we have that the MAP path equivalently solves

$$\min_{\boldsymbol{\tau}} \sum_{k=1}^{|\boldsymbol{\tau}|} \mathcal{C}_k(Y_{\tau_k:\tau_{k+1}-1}) + \beta|\boldsymbol{\tau}| \tag{2.35}$$

where we defined $\beta = -\log \rho$ and

$$\mathcal{C}_k(Y_{s:t}) = \min_x -\big\{(\mathbf{1}_{\{k>1\}} + \mathbf{1}_{\{k<K\}}) \log x + \log \pi(x) - (\rho + \theta)\Delta_k x + \bar{Y}_{s:t} \log(\theta x)\big\}.$$

Hence, β penalizes segmentations with many changepoints. Above we showed that with $\beta = -\log \rho$, the optimum of (2.35) is exactly $X^{\text{MAP}}_{1:N}$, which is also optimal for (2.32). Other settings of β result in paths which are suboptimal with respect to this objective, but potentially superior by other metrics. In particular, we observed that by setting β lower than $-\log \rho$, thus encouraging the algorithm to find paths with more changepoints than the MAP path, the paths are pointwise superior to $X^{\text{MAP}}_{1:N}$ in the sense of the preceding paragraph.

2.6.8 Simulation Details

In this section, we outline the details of our two simulations regarding the insensitivity of the prior to the posterior.

2.6.8.1 Differences between different SMC models

Under the constant size simulations, the effective population size was set to $N_e(t) = 20,000$ for all t. In the varying case,

$$N_e(t) = \begin{cases} 20,000, & t \geq 3162 \\ 10,000, & 1000 \leq t < 3162 \\ 2,000,000, & t < 1000 \end{cases}.$$

We discretized time into 32 epochs by selecting time points $t_0 = 0 < t_1 < \cdots < t_{32} = \infty$ and setting epoch $I_\epsilon = [t_\epsilon, t_{\epsilon-1})$. After setting the first time point as 0 and the final time point (t_{32}) as ∞, we set $t_1, t_2, \ldots, t_{31}$ as the sequence of 31 evenly log-spaced numbers between 10 and $100,000$ including the endpoints.

In what follows we measure the accuracy of the discretized SMC posterior with respect to the true (simulated) TMRCA at each position. To do this, we assume that coalescence events occur at the expected time of coalescence given that coalescence occurred in that epoch. To perform a fair comparison, even though we know how to solve the renewal model exactly, in this section we compare the time-discretized versions of it and the Markovian model. (See Appendix 2.6.8.3 for a precise description of our metric.)

For each scenario, we used `msprime` Kelleher et al. (2016) to simulate $N = 5 \times 10^6$ base pairs of sequence data for 25 pairs of chromosomes, for a total of $N/w = 5 \times 10^4$ loci. The sequences

were simulated with a per generation mutation rate of $\mu = 1.4 \times 10^{-8}$. Note that in scenarios 3 and 4, $\mu = r$. We calculate the posterior probabilities for the Markovian and renewal approximation using their corresponding transition probabilities. To assess the accuracy of the two priors we measured both absolute and relative error, defined respectively as

$$\text{Err}_A(\hat{\boldsymbol{x}}, \boldsymbol{x}) = \frac{1}{N/w} \sum_{i=1}^{N/w} \mathbb{E}_{\hat{x}_i} |\hat{x}_i - x_i|$$
$$\text{Err}_B(\hat{\boldsymbol{x}}, \boldsymbol{x}) = \frac{1}{N/w} \sum_{i=1}^{N/w} \mathbb{E}_{\hat{x}_i} \left| \log_{10} \left(\frac{\hat{x}_i}{x_i} \right) \right|$$

where x_i is the true TMRCA of the tree at position i and $\hat{x}_i$ is time to coalescence distributed according to the posterior.

To further understand the difference between the priors, we stratified this analysis by quartiles of the true TMRCA. We denote the minimum and maximum TMRCAs as q_0 and q_4, and the first, second, and third quartiles as q_1, q_2, and q_3. We then recalculate the absolute error in quarter j as

$$\text{Err}_A(\hat{\boldsymbol{x}}, \boldsymbol{x}, j) = \frac{\sum_{i=1}^{N/w} \mathbb{E}_{\hat{x}_i} |\hat{x}_i - x_i| \mathbf{1}_{[q_{j-1}, q_j)}(x_i)}{\sum_{i=1}^{N/w} \mathbf{1}_{[q_{j-1}, q_j)}(x_i)}$$

with relative error defined similarly. Due to the length bias of IBD tracts, the number of loci in quarter j will be smaller than the number of loci in quarter $j - 1$. The number of loci in each quarter under the various scenarios is displayed in Table 2.9.

Table 2.7 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 outperforms the Markov approximation by a large margin. 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 2.8 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. The large difference in quarter 4 is expected as under the Markov 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. I.e. $q(t \mid s) \approx 1/s$ when $s \gg t$. In contrast, the distribution under the renewal prior $\pi(t) = e^{-t}$ is more dense for smaller values of t.

In general, outside of the large difference between the methods in quarter 4, the two approximations are comparable, with neither one clearly dominating the other. When the underlying true

TMRCA is smaller, Err_A is the better measure of accuracy, so despite the Markov approximation outperforming the renewal prior in all quarters in terms of Err_B, the renewal prior actually outperforms the Markov approximation in quarters 1 and 2. We conclude from these results that our choice of prior is justified.

2.6.8.2 Effect of the demographic prior

We simulated data under three different demographic models and then analyzed the posterior when each model was used as a prior to infer TMRCAs on data generated from the other models. The standard library for population genetic simulation models, `stdpopsim` Adrion et al. (2019), provides a demographic model of the human population in Africa available as `Africa_1T12` and the zig-zag demography previously mentioned in Section 2.4.4.1 available as `Zigzag_1S14`.

In addition to these two models, we use a model with a constant population size of 2×10^4. We modeled the two non-constant population size history, Africa and zig-zag, using a piecewise constant function of 64 segments instead of a continuous function. The three models are plotted in Supplemental Figure 2.5. The set of time breakpoints used to approximate the size history is also the same set of points we used to discretize time into epochs. Here we discretized time into 64 epochs setting $t_0 = 0$, $t_{64} = \infty$, and the sequence $t_1 < \cdots < t_{63}$ as the sequence 63 evenly log-spaced numbers between 10 and 10^6 including the endpoints.

We then simulated 25 pairs of chromosomes for each model with `msprime` using the human chromosome 20 model with the default flat recombination and mutation maps in conjunction with the demographic models. The per generation per base pair mutation rate and recombination rate for chromosome 20 given by `stdpopsim` are $\mu = 1.29 \times 10^{-8}$ and $r = 1.718 \times 10^{-8}$ respectively. After simulating the data, for each pair of chromosomes generated under each of the models, we used each demographic size history as a demographic prior to calculate the posterior distribution of the TMRCA using the renewal approximation.

2.6.8.3 Expected time to coalescence

In this section we describe how to calculate the expected time to coalescence which we use in the simulations discussed in this Appendix. Suppose we have discretized time into the following set of $m+1$ times points $t_0 = 0 < t_1 < \cdots < t_m = \infty$. Precisely, the distribution of time to coalescence within epoch $I_\epsilon = [t_{\epsilon-1}, t_\epsilon)$ is

$$C_\epsilon \sim t_{\epsilon-1} + Z_\epsilon$$

where Z_ϵ is a truncated exponential in the interval $[0, t_\epsilon - t_{\epsilon-1})$ with parameter $\eta_\epsilon = 1/N_e(t_{\epsilon-1})$. The expectation of Z_ϵ is

$$\mathbb{E}(Z_\epsilon) = \frac{\int_0^\delta z\eta_\epsilon e^{\eta_\epsilon z} dz}{1 - e^{-\eta_\epsilon \delta}} = \frac{1 - e^{-\eta_\epsilon \delta} - \eta_\epsilon \delta e^{-\eta_\epsilon \delta}}{\eta_\epsilon (1 - e^{-\eta_\epsilon \delta})} = \frac{1}{\eta_\epsilon} + \frac{\delta}{1 - e^{-\eta_\epsilon \delta}}$$

where $\delta = t_\epsilon - t_{\epsilon-1}$.

Finally, with some algebra we have that

$$\mathbb{E}(C_\epsilon) = t_\epsilon + \frac{1}{\eta_\epsilon} + \frac{\delta}{e^{-\eta_\epsilon \delta} - 1}.$$

The final epoch $I_m = [t_{m-1}, t_m) = [t_{m-1}, \infty)$ is not bounded above, so the time to coalescence simply follows an exponential random variable with parameter η_{m-1} without truncation. Thus the expected time to coalescence is simply given by

$$\mathbb{E}(C_m) = t_{m-1} + \frac{1}{\eta_m}.$$

Additional figures

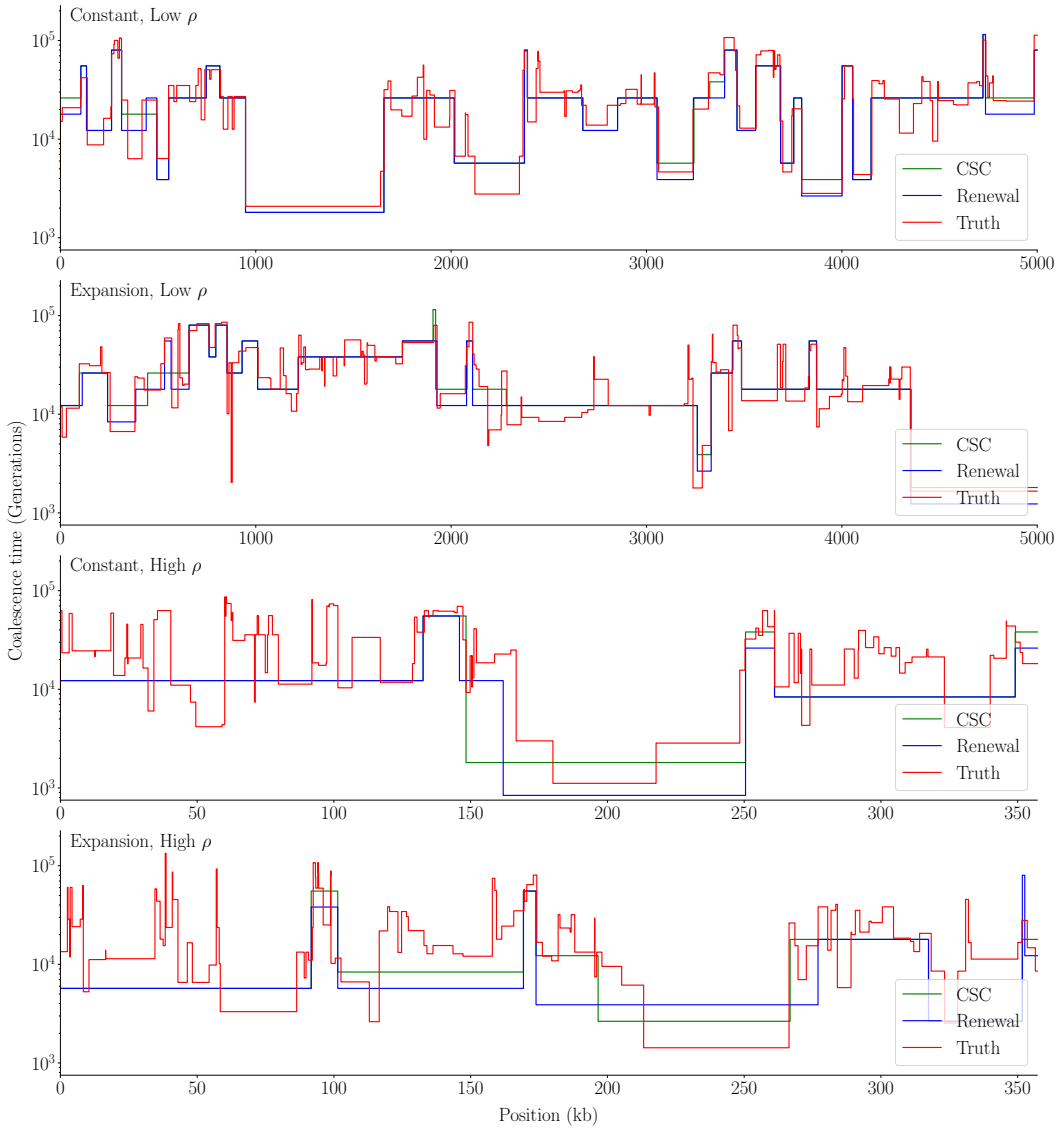

Figure 2.3: Comparison of Viterbi path between conditional Simonsen-Churchill and renewal approximations.

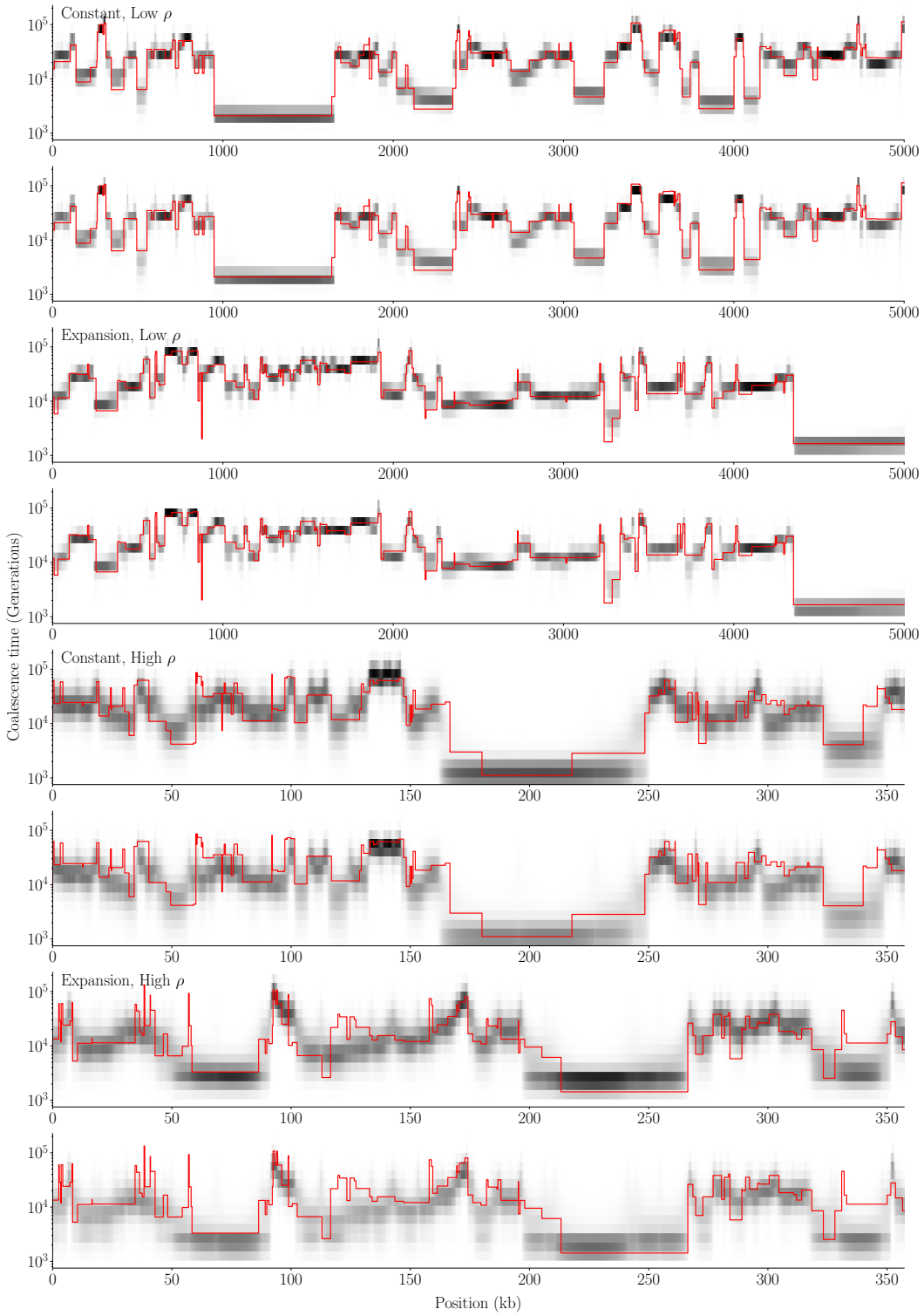

Figure 2.4: Comparison of posterior heatmap between conditional Simonsen-Churchill and renewal approximations. The top panel in each group is the posterior given by the CSC prior and the bottom panel is the posterior given by the renewal prior.

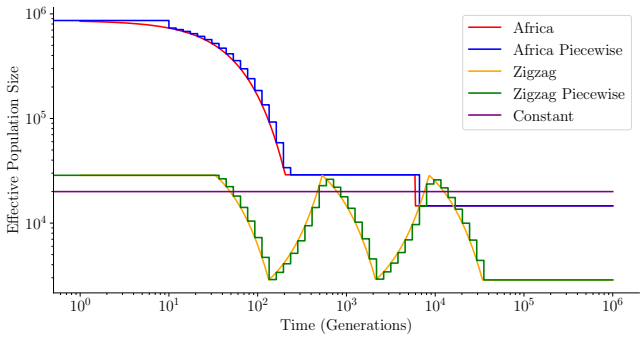

Figure 2.5: The population trajectory under the three models used in the simulation.

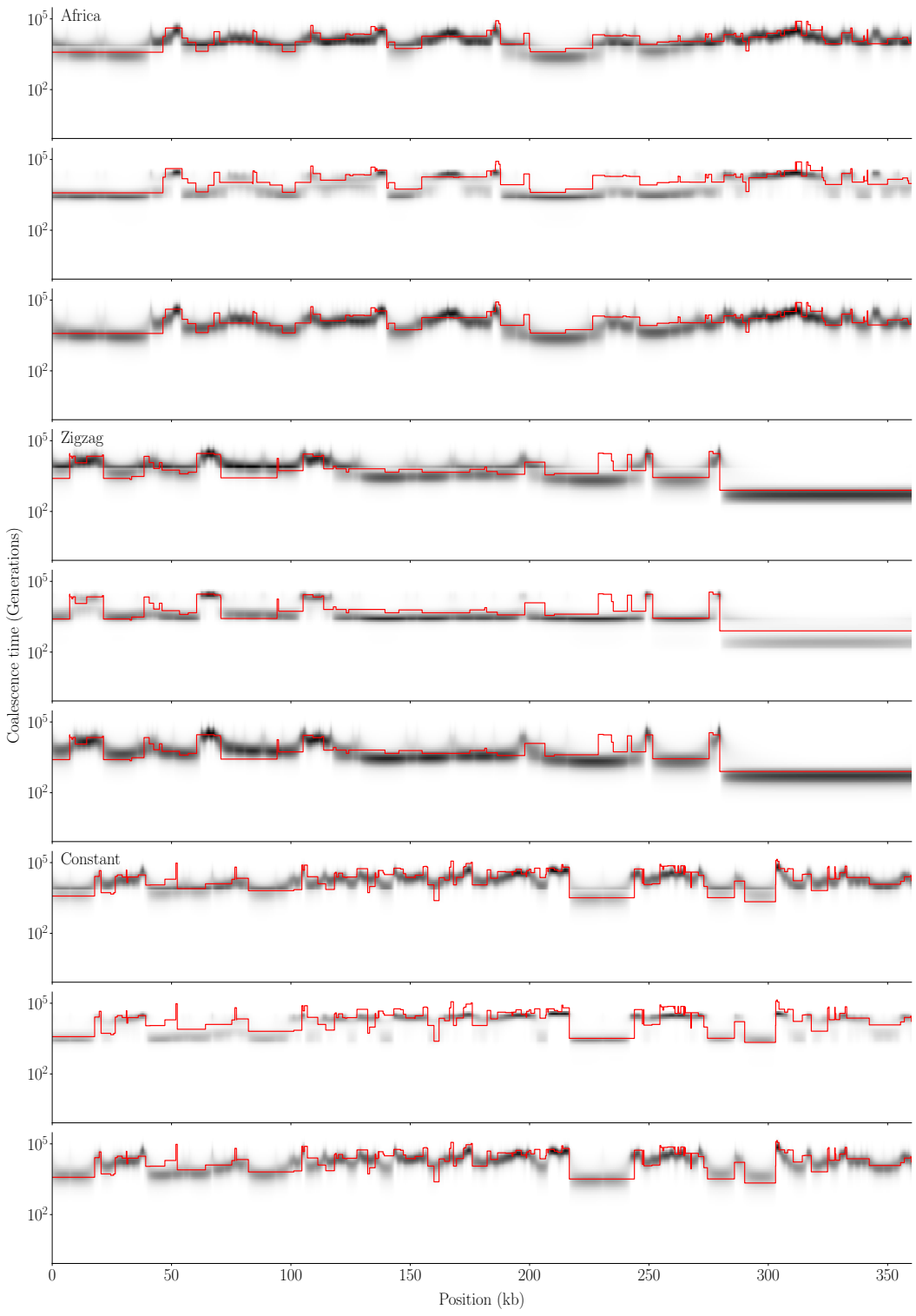

Figure 2.6: Comparison of posterior using different demographic priors. The first three panels were generated by the Africa demography model, the second three by the zigzag model, and the last three by a constant size model. Within each grouping of three, the panels are ordered Africa/zigzag/constant by the demographic prior assumed. The red line is the true TMRCA.

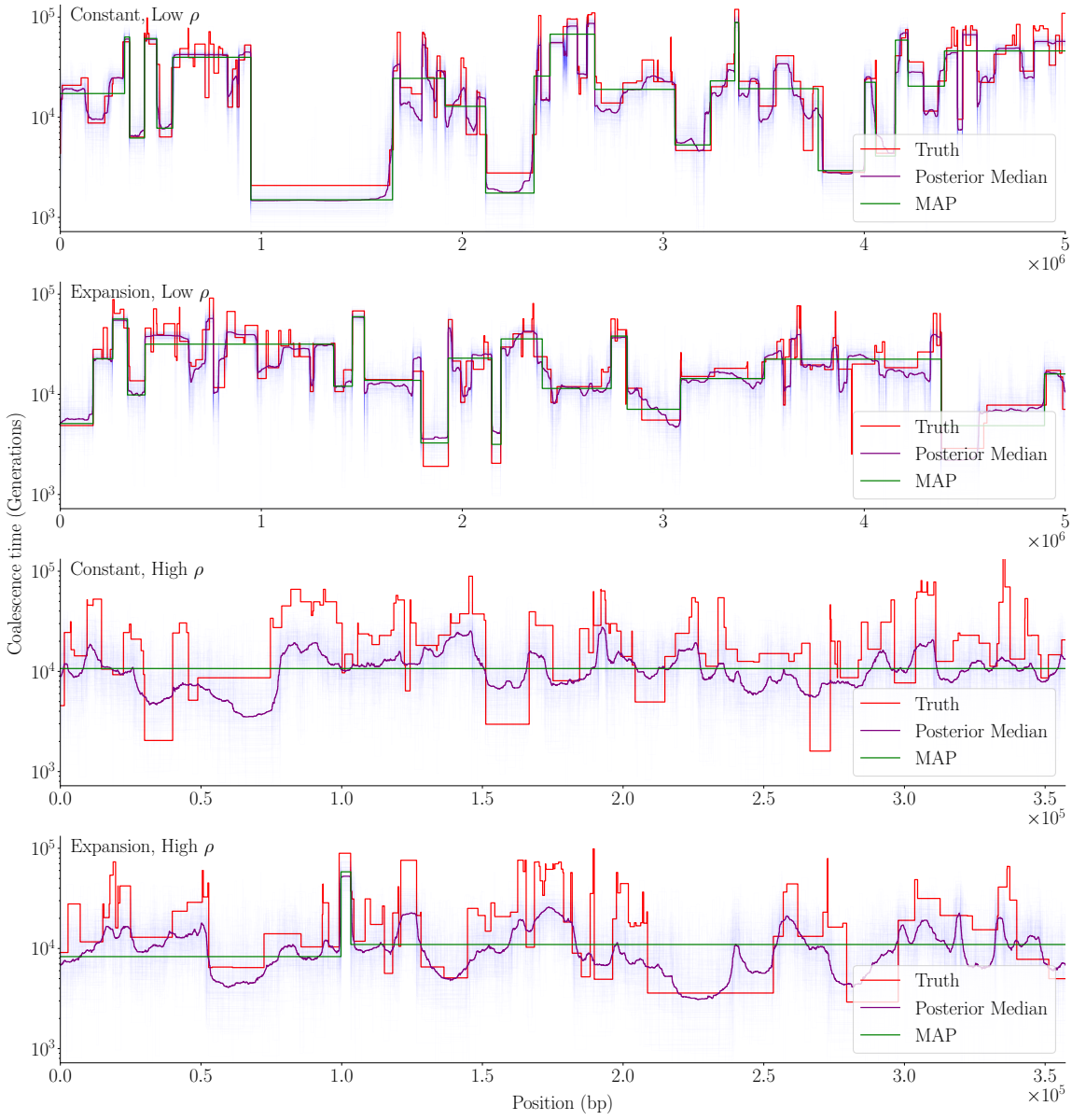

Figure 2.7: Comparison of Bayesian and frequentist method on simulated data. The light purple lines represent sample paths drawn from the posterior.

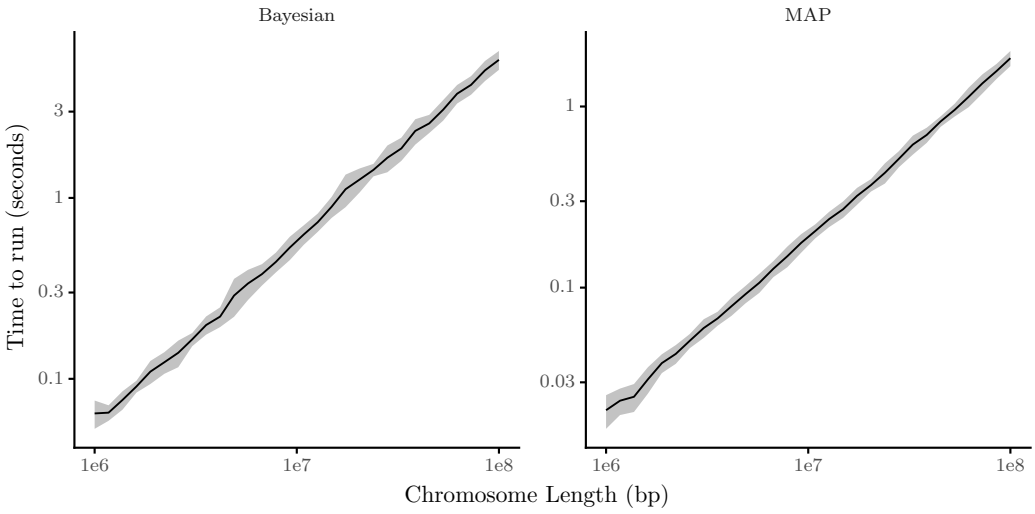

Figure 2.8: Mean running time of Bayesian sampler and MAP decoder over various chromosome lengths on a log-log scale. The bands represent the standard error of the runs.

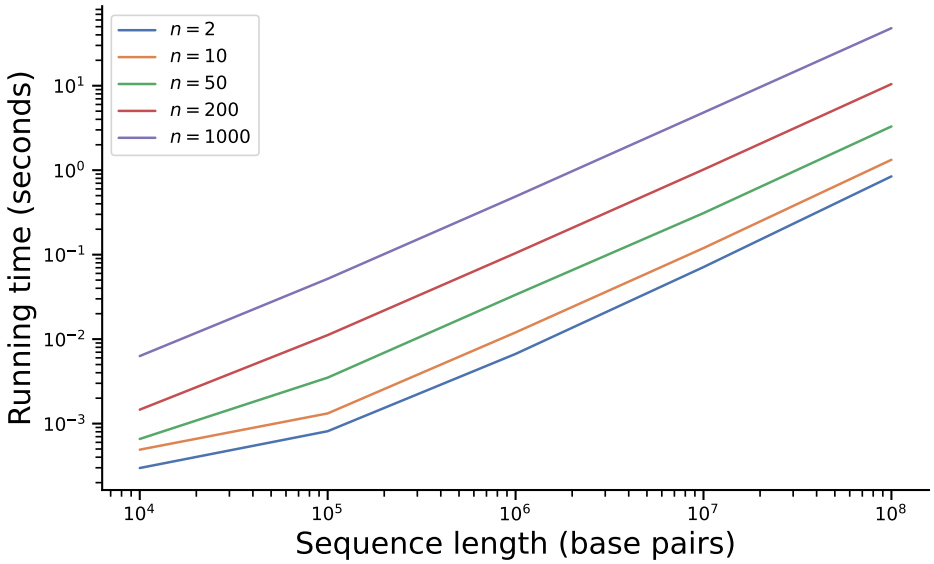

Figure 2.9: Mean running time of MAP decoder over various chromosome lengths and sample sizes on a log-log scale. We simulated chromosomes of indicated lengths and counted the amount of time needed to compute the MAP path for various sample sizes. We repeated this experiment ten times for each setting. In each experiment, the population-scale rates of mutation and recombination were set to $\theta = \rho = 4 \times 10^{-4}$.

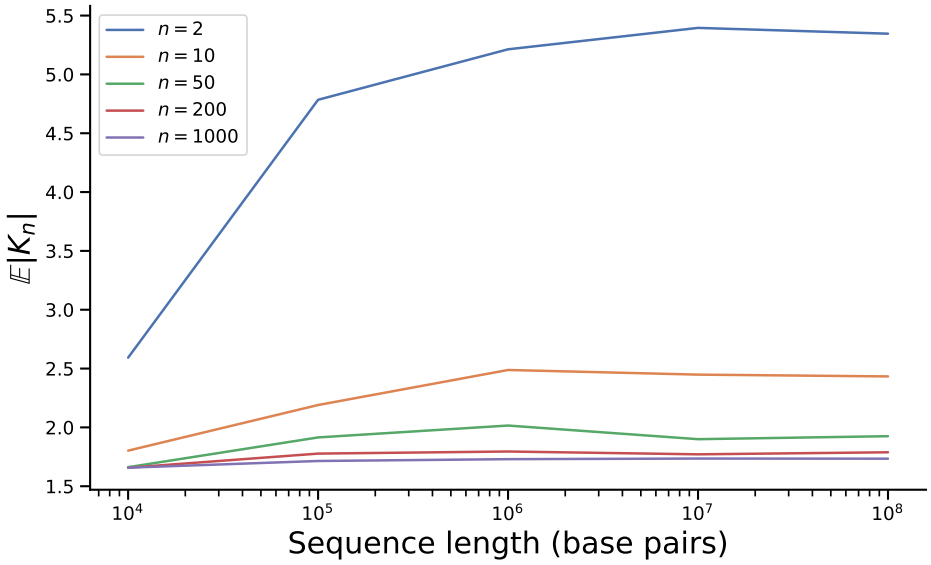

Figure 2.10: Average number of pieces in the piecewise decomposition of Proposition 3. Experimental settings were the same as in Figure 2.9.

.

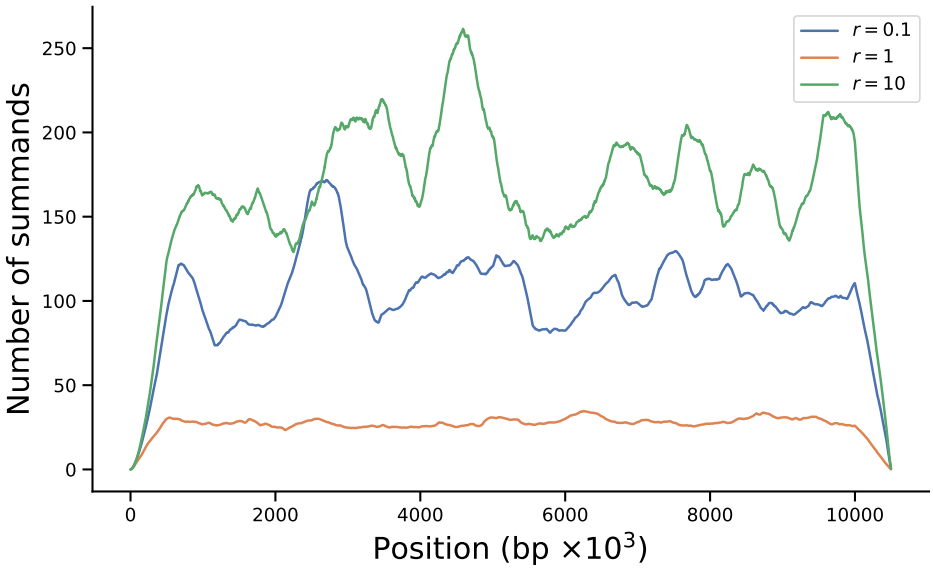

Figure 2.11: Average number of summands considered before truncation in exact Bayesian sampler. We simulated a chromosome of length 10^7 base pairs and counted the number of terms that were evaluated in (2.12) before we reached the truncation condition indicated in the main text. We repeated this experiment ten times over three different settings of the recombination rate: $\rho = \{.1, 1, 10\}\theta$ where $\theta = 4 \times 10^{-4}$ was the scale mutation rate. The plots are shown using a smoothed moving average (window size 500) for clarity.

Table 2.4: List of Symbols

Symbol	Interpretation
$\boldsymbol{Y}$	Matrix of sequence data
H	Number of chromosomes
N	Number of base pairs
A	ARG
$\mathcal{A}$	Support set of ARGs
φ	Evolutionary model
ϕ, N_e	Effective population size
s, t, x	Generations in the past
X_i	TMRCA (Marginal gene tree)
Y_i	Number of differences at locus i between two chromosomes
$\pi(t)$	Marginal distribution of coalescence time
$q(t \mid s)$	Conditional density of t given recombination occurred
$\underline{\pi}(t)$	Density of t given recombination occurred dropping dependence on s
M	Number of hidden states
$M_\Gamma(K)$	Mixture of K gamma distributions
u, v	Position along chromosome
R_v	Event that IBD segment begins at position v
$\overline{R}_{u:v}$	Event that there is not a recombination event between positions u and v
τ	Recombination breakpoints
h	Panel chromosome

Additional tables

Table 2.5: Mean absolute error (Err_A) over 25 runs under each scenario. Standard error in parentheses.

Scenario	1	2	3	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)

Table 2.6: Mean relative error (Err_B) over 25 runs under each scenario. Standard error in parentheses.

Scenario	1	2	3	4
CSC	0.1320 (0.0034)	0.1271 (0.0018)	0.3037 (0.0013)	0.2990 (0.0013)
Renewal	0.1393 (0.0024)	0.1357 (0.0021)	0.3437 (0.0046)	0.3406 (0.0015)

Table 2.7: Mean absolute error (Err_A) over 25 runs under each scenario stratified by quartile. Standard error in parentheses.

	Scenario	1	2	3	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)

Table 2.8: Mean relative error (Err_B) over 25 runs under each scenario stratified by quartile. Standard error in parentheses.

	Scenario	1	2	3	4
CSC	Q1	0.1414 (0.0049)	0.1277 (0.0025)	0.3206 (0.0025)	0.3053 (0.0026)
Renewal	Q1	0.1490 (0.0036)	0.1375 (0.0029)	0.3278 (0.0056)	0.3093 (0.0020)
CSC	Q2	0.1212 (0.0021)	0.1328 (0.0036)	0.2746 (0.0015)	0.2866 (0.0014)
Renewal	Q2	0.1307 (0.0022)	0.1409 (0.0038)	0.3278 (0.0018)	0.3491 (0.0016)
CSC	Q3	0.1272 (0.0022)	0.1297 (0.0045)	0.2795 (0.0016)	0.2857 (0.0019)
Renewal	Q3	0.1333 (0.0024)	0.1362 (0.0048)	0.3743 (0.0021)	0.3853 (0.0020)
CSC	Q4	0.1216 (0.0021)	0.1221 (0.0026)	0.3162 (0.0017)	0.3169 (0.0019)
Renewal	Q4	0.1258 (0.0025)	0.1271 (0.0031)	0.4546 (0.0020)	0.4542 (0.0022)

Table 2.9: Mean counts of loci in each quarter for under each scenario across 25 simulations. Standard error in parentheses.

Scenario	1	2	3	4
Q1	24995.32 (1060.60)	28273.56 (1186.73)	27371.08 (585.02)	27766.12 (370.51)
Q2	12780.96 (830.11)	10312.52 (636.41)	11182.48 (270.56)	10906.12 (213.53)
Q3	8024.28 (480.42)	7304.80 (495.50)	7380.76 (235.32)	7292.56 (167.61)
Q4	4199.44 (323.59)	4109.12 (351.86)	4065.68 (132.49)	4035.20 (111.07)

Table 2.10: Mean absolute error (Err_A) over 25 runs under each scenario. Standard error in parentheses.

	Africa	Zigzag	Constant
Africa	10133.60 (26.40)	10351.75 (29.07)	10651.29 (27.02)
Zigzag	5566.70 (130.60)	5013.63 (118.82)	5762.65 (135.27)
Constant	11624.13 (262.00)	11803.59 (265.48)	11934.81 (269.02)

Table 2.11: Mean relative error (Err_B) over 25 runs under each scenario. Standard error in parentheses.

	Africa	Zigzag	Constant
Africa	0.3287 (0.0004)	0.3931 (0.0004)	0.3545 (0.0004)
Zigzag	0.3669 (0.0041)	0.3538 (0.0044)	0.3776 (0.0039)
Constant	0.3489 (0.0068)	0.3949 (0.0058)	0.3671 (0.0064)

Table 2.12: Mean absolute error (Err_A) over 25 runs. Standard error in parenthesis.

Scenario	1	2	3	4
MAP	6022 (234)	5041 (183)	11916 (293)	12060 (123)
Bayesian	4423 (174)	3846 (140)	8637 (198)	8532 (79)

Table 2.13: Mean relative error (Err_B) over 25 runs. Standard error in parenthesis.

Scenario	1	2	3	4
MAP	0.1332 (0.0048)	0.1224 (0.0041)	0.3388 (0.0035)	0.3447 (0.0029)
Bayesian	0.1098 (0.0052)	0.1112 (0.0034)	0.2584 (0.0067)	0.2278 (0.0012)

Table 2.14: Mean absolute error (Err_A), mean relative error (Err_B), and mean running time over 25 runs varying the truncation cutoff. Standard error in parenthesis.

Truncation Cutoff	Err_A	Err_B	Time
10^{-2}	3219.74 (608.23)	0.6706 (0.1296)	0.0835 (0.0062)
10^{-3}	2979.61 (567.86)	0.5708 (0.1042)	0.1242 (0.0074)
10^{-4}	2905.50 (557.24)	0.4799 (0.0825)	0.1525 (0.0085)
10^{-5}	2877.94 (553.02)	0.4004 (0.0691)	0.1998 (0.0076)
10^{-6}	2864.44 (551.67)	0.3268 (0.0477)	0.3320 (0.0123)

Additional algorithms

Algorithm 1 Exact Viterbi decoding

Require: $\log(\pi(x)) \in \mathcal{V}_K$ ▷ piecewise coalescent prior
Require: $M \in \mathbb{Z}^{N \times H}$ ▷ number of mismatches to panel haplotype h as position n
  $V \leftarrow \text{vector}(H)$ ▷ initialize log likelihoods
  bt $\leftarrow \text{vector}(N)$ ▷ backtracking array
  for all $1 \leq h \leq H$ **do**
    $V[h] \leftarrow \log(\pi)(x)$
    $ibd[h](x) \leftarrow 0$ ▷ length of spanned IBD tract
  end for
  for all $1 \leq n \leq N$ **do** ▷ outer loop over each position
    for all $1 \leq h \leq H$ **do** ▷ add log-emission probability
      $y \leftarrow M[n, h]$
      $V[h] = V[h] + \theta + \rho + yt + \log \Gamma(1 + y) - y \log \theta$
      $ibd[h] \leftarrow ibd[h] + 1$
    end for
    $h^* = \arg\min_{1 \leq h \leq H} \sup_x V[h](x)$ ▷ Computed via Section 2.6.6.1
    $b^* = \sup_x V[h^*](x)$
    bt[n] $\leftarrow (h, V[h^*], ibd[h](b^*))$ ▷ highest probability segment for recombination
    for all $1 \leq h \leq H$ **do**
      $V[h](x) = \max(V[h](x), \log(\pi(x)) + b^*)$ ▷ Computed via Section 2.6.6.2
      for all $x : V[h](x) = \log(\pi(x)) + b^*$ **do**
        $ibd[h](x) \leftarrow 0$ ▷ reset IBD counter for recombinants
      end for
    end for
  end for
  return BACKTRACK(bt) ▷ Algorithm 2

Algorithm 2 Backtracking algorithm

```
function BACKTRACK(bt)
    h, V(x), ibd = bt[N]
    pos = N − ibd
    ret ← [(h, arg max_x V(x))]
    while pos > 0 do
        h, V(x), ibd = bt[pos]
        ret.append((h, arg max_x V(x)))
        pos = pos − ibd
    end while
    return ret
end function
```

CHAPTER 3

Variational phylodynamic inference using pandemic-scale data

3.1 Introduction

The COVID-19 pandemic has demonstrated an important supporting role for phylogenetics in epidemiology and public health, while also creating unforeseen technical and methodological challenges. As the first global public health event to occur in an era of ubiquitous gene sequencing technology, the pandemic has resulted in a data explosion of unprecedented proportions. GISAID, a worldwide repository of SARS-CoV-2 genomic data, currently has over 7.5M samples, with contributions from almost every country (Elbe and Buckland-Merrett, 2017; van Dorp et al., 2021). A phylogenetic representation of this database is believed to be the largest ever constructed (Turakhia et al., 2021a). Existing phylogenetic methods, which were developed and tested on datasets orders of magnitude smaller, are inadequate for pandemic-scale analysis, resulting in missed opportunities to improve our surveillance and response capabilities (Hodcroft et al., 2021; Ye et al., 2021; Morel et al., 2021).

These shortcomings have spurred new research initiatives into phylogenetic inference methods capable of analyzing millions of samples. In particular, there has been significant recent progress in estimating and/or placing novel sequences onto very large phylogenies (Minh et al., 2020; Turakhia et al., 2021a; Aksamentov et al., 2021; Ye et al., 2022a,b). Accurate estimation of the underlying phylogeny has numerous downstream applications, including contact tracing (e.g., Lam-Hine et al., 2021; McBroome et al., 2022), surveillance (e.g., Abe and Arita, 2021; Klink et al., 2021), and improved understanding of pathogen biology (e.g. Majumdar and Sarkar, 2021; Turakhia et al., 2021b).

Another area of active research in phylogenetics, distinct from tree inference, is so-called *phylodynamics*, which seeks to understand how immunological, epidemiological, and evolutionary forces interact to shape viral phylogenies (Volz et al., 2013b). Here, the quantity of interest is typically a low-dimensional parameter vector characterizing the underlying phylodynamic model,

while the phylogeny itself is a nuisance parameter. Of particular interest for the current pandemic are methods that can estimate effective population size and reproduction number of the pathogen from viral genetic data (e.g. Zhou et al., 2020; Lai et al., 2020; Volz et al., 2021; Campbell et al., 2021). Compared to phylogeny estimation, less progress has been made on so-called "phylodynamic inference" at the pandemic scale. This absence motivates the present study.

Bayesian methods are often preferred for phylodynamic inference because there are usually many trees which explain the data equally well. Hence, downstream quantities of interest possess a potentially significant amount of "phylogenetic uncertainty" which is not reflected in frequentist point estimates. Unfortunately, Bayesian phylogenetic procedures inherently scale very poorly: the space of phylogenetic trees grows rapidly, and there are an astronomical number of possible trees to consider, even for relatively small samples. Consequently, on large problems, the workhorse algorithm of field, Markov chain Monte Carlo (MCMC), tends to either conservatively explore very limited regions of tree space, or liberally propose large moves that are often rejected (Whidden and Matsen IV, 2015; Zhang and Matsen IV, 2019).

Even before the pandemic, awareness of the scalability issues surrounding Bayesian phylogenetics was growing (Höhna and Drummond, 2012; Whidden and Matsen IV, 2015; Aberer et al., 2016; Dinh et al., 2017). As a scalable alternative to MCMC, variational inference (VI) has recently garnered some attention in phylogenetics. VI is a general method for sampling approximately from a posterior distribution using techniques from optimization (Jordan et al., 1999). Fourment et al. (2020) used VI to accelerate computation of the marginal likelihood of a fixed tree topology. Fourment and Darling (2019) used the probabilistic programming language STAN to perform variational inference of the Bayesian skyline model (Pybus et al., 2000). Both of the preceding methods only analyze a fixed tree topology, so they cannot account for phylogenetic uncertainty. Simultaneously, Zhang and Matsen IV (2018, 2019); Zhang (2020) have made progress on a full variational approach which includes optimization over the underlying topology. Although these innovations represent significant advances in terms of performance, they still cannot come close to exploiting all of the information contained in a pandemic-scale data set.

3.2 New Approaches

Inspired by these works, and responding to the need for better tooling to study the ongoing pandemic, we devised a method capable of providing accurate and calibrated estimates of the rates of transmission and recovery for COVID-19 using data from tens of thousands of viral genomes. Our approach unites several threads of research in phylogenetics and scalable Bayesian inference. We build on aforementioned advances in variational phylogenetic inference (Fourment and Darling, 2019; Zhang, 2020), as well as recent progress in phylodynamic modeling of infectious diseases

(Stadler et al., 2013), Bayesian stochastic optimization (Hoffman et al., 2013), and differentiable programming (Bradbury et al., 2018). To achieve this level of scalability, our method makes several tradeoffs and approximations which are detailed below. Briefly, we adopt a divide-and-conquer strategy where distant subtrees of a very large phylogeny are assumed to evolve approximately independently, and we further assume that topological estimates of these subtrees are an accurate reflection of their distribution under the prior. We argue that these reasonable approximations in the context of an ultralarge, global phylogeny, and that their combined effect appears to be benign: the resulting estimates closely agree with the existing state of the art on simulated data, and exhibit a remarkable level of concordance with ground-truth estimates on real data, while taking just minutes to produce.

3.3 Results

In this section, we test our method on both simulated and real data, and compare it to the existing implementation of the birth-death skyline model in BEAST.

3.3.1 Simulation

First, we performed a simulation study to evaluate how well VBSKY approximates the posterior distribution compared to BEAST. We studied four different scenarios:

1. Constant: the effective reproductive number stays constant through time;
2. Decrease: there is a sharp drop in the effective reproductive number;
3. Increase: there is a sharp increase in the effective reproductive number; and
4. Zigzag: the effective reproductive number goes through a series of decreases and increases.

We simulated transmission trees using the `R` package TreeSim (Stadler, 2011) and generated sequences data along each tree using the program Seq-Gen (Rambaut and Grass, 1997).

Across all scenarios, the rate of becoming uninfectious, δ is held constant at $\delta(t) = 4$ for all t. The sampling rate is also held constant at $s(t) = 0.25$. Only R is allowed to vary. Under the constant scenario, $R(t) = 1.3$ for all t. In the decrease scenario,

$$R(t) = \begin{cases} 2.25, & t \leq 1 \\ 0.75, & t > 1. \end{cases}$$

In the increase scenario,

$$R(t) = \begin{cases} 1, & t \leq 3 \\ 2.5, & t > 3. \end{cases}$$

In the zigzag scenario,

$$R(t) = \begin{cases} 2.0, & t \in [0,1] \cup (2,3] \\ 0.75, & t \in (1,2] \cup (3,4]. \end{cases}$$

Each simulation was run for four time units, and ten trees were generated under each scenario. Because the sampling process is stochastic in this model, the size of the simulated tree varied from run to run. The minimum (maximum) number of samples in each under the constant, decrease, increase, and zigzag scenarios was 175 (1553), 117 (590), 124 (1075), and 161 (1852), respectively.

We compared the performance of our method to the current state-of-the-art for Bayesian phylogenetic analysis, BEAST (Bouckaert et al., 2019). BEAST allows for the birth-death skyline model to be used as a tree prior, facilitating direct comparison with VBSKY. Because BEAST uses MCMC to estimate the posterior, the number of sequences it can analyze is limited. Therefore, for each simulation, we randomly sampled 100 sequences for BEAST to analyze. We allowed BEAST to run long enough that the effective sample size exceeded 1000 for each evolutionary parameter. Since VBSKY is not limited by sample size, we analyzed all sequences in each simulation, as follows: We set the size of each random subsample to be $b = 100$ tips. The number of trees in the ensemble was set to be the smallest integer such that the number of trees multiplied by 100 was larger than the number of sampled sequences. Under this scheme, each sequence was sampled once on average.

The results of the simulation study are shown in Figures 3.1 and 3.2. Figure 3.1 displays the median of the medians and 95% equal-tailed credible intervals of the simulations under each scenario using BEAST to analyze the data. Figure 3.2 shows the same for VBSKY. Besides a few minor differences, the estimates given using VBSKY are similar to those given by BEAST; both BEAST and VBSKY adequately capture the true value of the effective reproductive number. The credible intervals given by BEAST are wider than those of VBSKY, and do a better job of covering the true model in some cases; we return to this point in Section 3.4. In the decrease scenario, VBSKY is better able to capture the larger value of R earlier in time, while BEAST appears to revert to the prior at times earlier than $t = 0.5$. Because VBSKY allows for more sequences to be analyzed, the method is able to detect transmission events further back in time.

Even though in some cases we analyzed hundreds more sequences using VBSKY than when we used BEAST, the run-time of VBSKY was 71.75 seconds on average for each simulation whereas BEAST took 20 minutes to perform 10,000,000 MCMC steps. The simulation results show that VBSKY is able to get comparable results as BEAST with a much shorter run-time, and in some

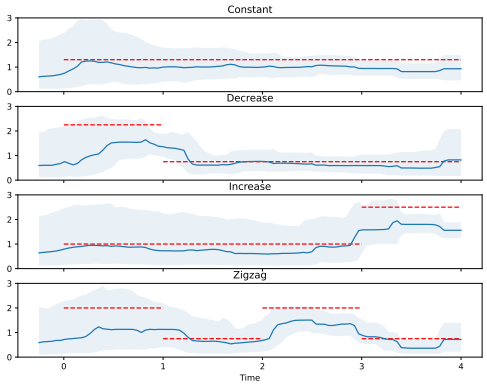

Figure 3.1: Median of the medians and the equal-tailed 95% credible intervals of the posteriors of the effective reproductive number over time of the 10 simulations for each scenario using BEAST. The dotted red line is the true effective reproductive number over time.

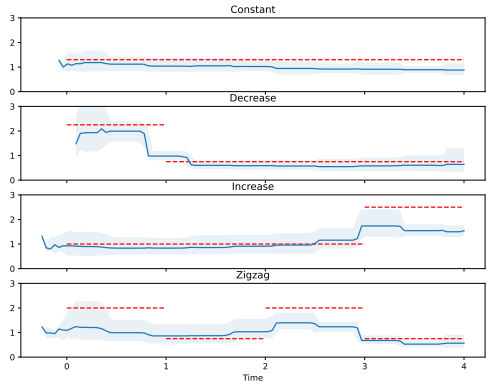

Figure 3.2: Median of the medians and the equal-tailed 95% credible intervals of the posteriors of the effective reproductive number over time of the 10 simulations for each scenario using VBSKY. The dotted red line is the true effective reproductive number over time.

cases like the decrease scenario, VBSKY can produce more accurate estimates than BEAST.

3.3.2 Analysis of the global pandemic

We tested our method on a large, serially-sampled COVID-19 dataset from the GISAID initiative (Elbe and Buckland-Merrett, 2017). At the time this analysis was performed, there were 6.5M SARS-CoV-2 sequences in the database. In addition to the raw nucleotide data, GISAID provides sample time and location information. The collection dates of the sequences range from January 3rd, 2020 to December 8th, 2021.

For our analysis, we chose to study the transmission of COVID-19 of Michigan, Florida, and the entire USA. It is important to study the epidemiology of COVID-19 at the sub-national level as many public health policies such as mask mandates, stay at home orders, vaccine distribution, and other social distancing measures are enforced at the state level. Policies or decisions made in one state may not be detected studying national data. Due to the differences in health policies across states and the reduced frequency of travel during the pandemic, we expect the incidence and prevalence of COVID-19 to vary from state to state. On the other hand, policies are sometimes made at the national level, and more recently travel especially around the holidays has become widespread, so understanding trends at a national level is equally vital.

After filtering the sequences by location, the number of sequences were 81,375, 34,978, and 1,280,563 for Florida, Michigan, and the USA respectively. We noticed that the number of con-

firmed cases increased or decreased based on the day of the week, likely because fewer cases are reported over the weekend. To correct for any inaccuracies in the sample time distribution, we set all sequences sampled in the same calendar week to have the same sample time. We used a fixed molecular clock model with substitution rate $1.12 \times 10^{-3}/\text{bp}/\text{year}$ which is the estimate given by the World Health Organization (WHO) (Koyama et al., 2020).

3.3.2.1 Hyperparameter Tuning

Before proceeding to the analysis, we sought to better understand how the various tuning parameters of our method affected the results. VBSKY has two main tuning parameters that can be adjusted: the number of tips in each subsample (denoted b in the preceding section), and the number of subsamples of the overall dataset $\mathcal{D}$ (denoted S in the preceding section). Increasing either enables us to analyze more sequences, but at the expense of additional computation time.

To understand the effect of the number of trees, we examined the posterior of the effective reproductive number and the sampling rate of Florida and the USA while fixing the number of tips and varying the number of trees. We set the number of tips to be 200 and examined the posterior for each number of trees in the set $\{10, 25, 50, 100, 150\}$. Patients with mild bouts of COVID-19 are generally not infectious after 10 days of symptom onset (Arons et al., 2020; Bullard et al., 2020). The rate of becoming uninfectious is the inverse of the number of infectious days. As one unit of time corresponds to one year, the estimated value for δ is given by $1/10 \times 365 = 36.5$. Using this, we fixed the uninfectious rate to be 36.5 to avoid nonidentifiability issues since we cannot estimate $R, \delta,$ and s simultaneously (Stadler, 2009; Louca and Pennell, 2020). For the GMRF smoothing prior, we chose a relatively uninformative hyperprior distribution with large variance for the parameters of the smoothing prior. In particular, we selected a gamma distribution distribution with parameters $a = b = 0.001,$ giving a mean of 1 and variance of 1000. As a rough estimate of the sampling rate, we also chose the prior for s to be a Beta$(0.02, 0.98)$ distribution with expectation 0.02, as the ratio of sampled sequences to the number of cumulative cases is around 0.02. The remaining priors are shown in the first line of Table 3.1.

Figure 3.3 shows the posterior of R for both Florida and the USA when varying the number of trees. Figure 3.4 shows the posterior for s. The figures indicate a larger difference when the number of trees is 10 compared to any greater number of trees. The median and credible interval for R was much smaller and the median and credible interval for s was much larger closer to the present when the number of trees was 10. The credible intervals when the number of trees was 10 was also much wider. A closer inspection showed that this also seems to be the case when the number of trees is 25, albeit to a smaller degree. When we increased the number of trees to 50, this difference mostly disappeared.

We performed a similar study to understand the effect of varying the number of tips. We fixed

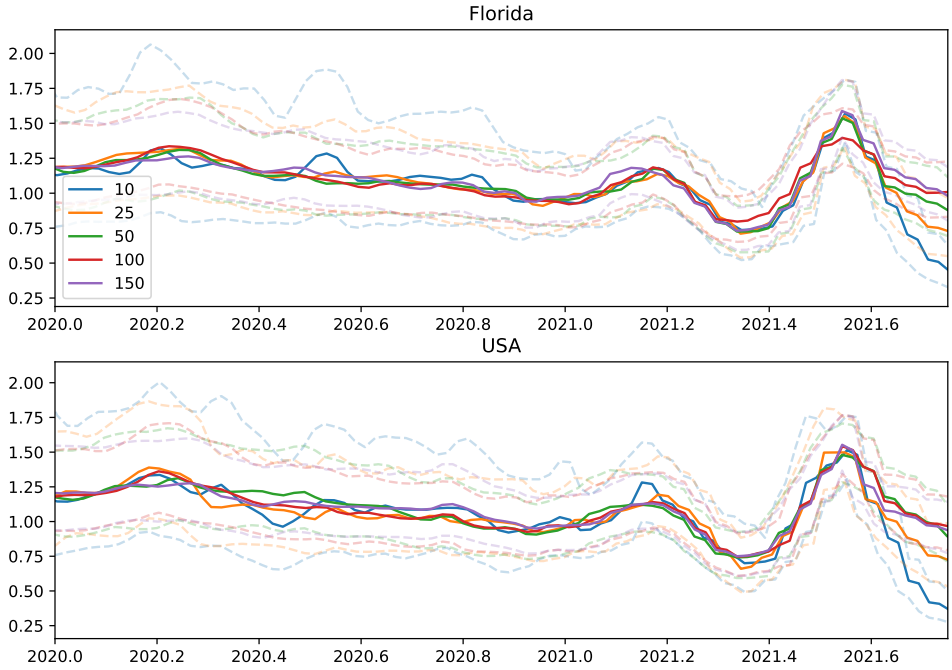

Figure 3.3: Posterior of R while varying the number of trees. Solid lines represent the median and the dotted lines represent the equal-tailed 95% credible intervals.

the number of trees as 50 and adjust the number of tips to values in the set $\{50, 100, 200, 400\}$, and examined the posteriors of R and s while holding δ fixed. Similar to above, varying the number of tips does not appear to have a large effect on the results. Using only 50 tips per tree resulted in a wider credible interval for Florida and the USA for both R and s. Figure 3.5 shows that using 50 tips also leads to flatter estimates for R further back in the past. This is likely the result of trees with fewer tips having fewer transmission events further back in the past which can be used to estimate R.

When comparing the posteriors when the number of tips is 100 or 200, only minor differences appeared. Using 200 tips did seem to lead to better detection of changes in R and s further back in the past. Looking at Figure 3.5, using 400 tips per tree led to a sharper decrease in R towards the present. Figure 3.6 shows that using 400 tips generally led to slightly larger estimates of s at all points in time.

Overall, regardless of the number of tips or trees used, the posterior estimates of both R and s for both Florida and the USA are similar. However, increasing the number of trees decreases the variances in posterior estimates of R and s, and also results in more accurate estimates of both

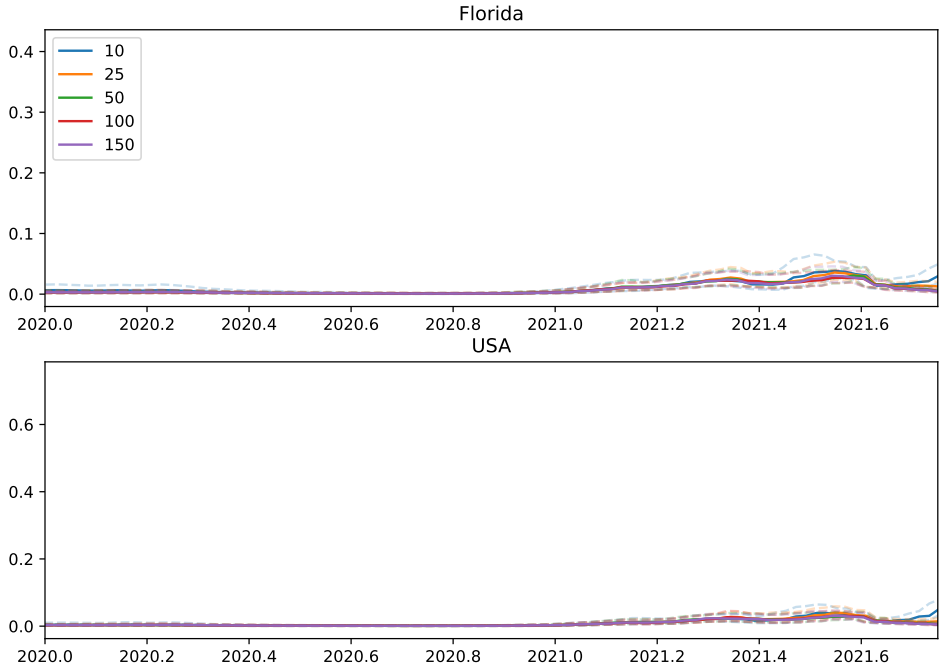

Figure 3.4: Posterior of s while varying the number of trees. Solid lines represent the median and the dotted lines represent the equal-tailed 95% credible intervals.

parameters towards the present. This improvement seems to plateau after increasing the number of trees to 50. Similarly, increasing the number of tips can increase the power to detect changes in R and s further back in the past, but using too many tips can lead to more erratic estimates of the parameters towards the present.

Keeping this in mind while also noting that increasing the number of trees and tips can incur large computational costs, using 50 trees with 200 tips leads to sharper estimates of the posterior without requiring excessive computation.

3.3.2.2 Results

Based on the results from the previous section, we ran VBSKY with 50 subsamples of 200 sequences for a total of 10^4 sequences. We estimated the epidemiological parameters for Florida, Michigan, and the overall USA. State-level results were compared to a "ground truth" estimator of the effective reproductive number which is derived from orthogonal (i.e. non-genetic) public health data sources (Shi et al., 2021). The prior and hyperprior settings for all of the scenarios described below are shown in Table 3.1.

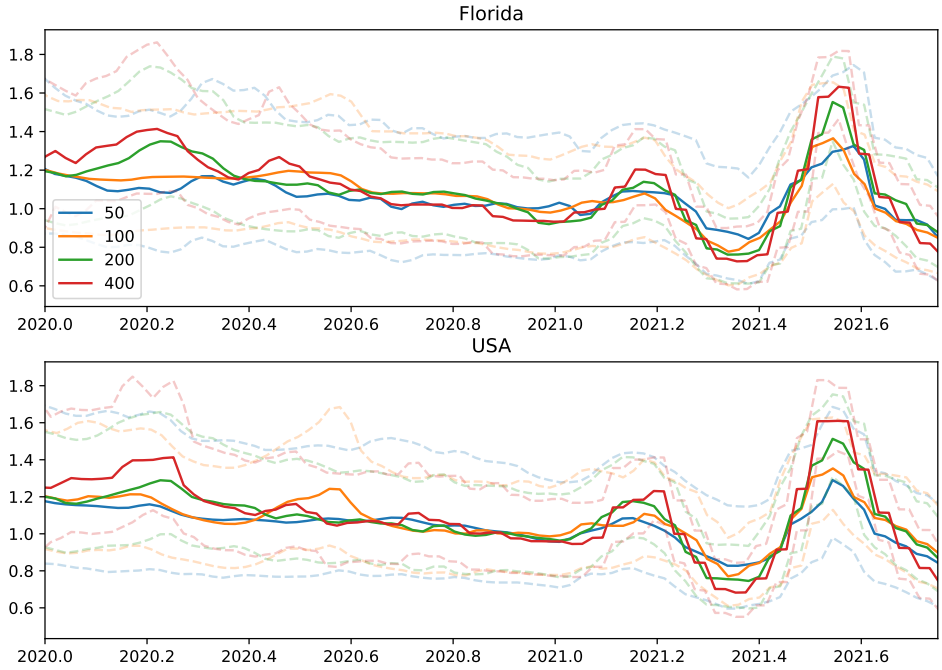

Figure 3.5: Posterior of R while varying the number of tips. Solid lines represent the median and the dotted lines represent the equal-tailed 95% credible intervals.

We first analyzed the data using the same uninformative smoothing hyperpriors as in the hyperparameter study in the previous section ("Uninformative Smoothing" in Table 3.1). Figure 3.7 displays the posterior of R over time for each region for the uninformative smoothing analysis. For Florida (top panel), we see that the estimates for R over time produced by VBSKY matches the results using surveillance data in the recent past. However, earlier in the pandemic, VBSKY does not seem to be able to capture the rise and fall of R but instead provides a flat estimate of the parameter.

In the middle panel (Michigan), we see the VBSKY posterior is very similar to the posterior given by the surveillance data method even looking further back in the past. Looking at the top

Table 3.1: Prior Distributions used in Analyses.

Analysis	R	s	τ_R	τ_s	x_1
Uninformative Smoothing	LogN(1,1)	Beta(.02, .98)	Gamma(.001, 0.001)	Gamma(.001, 0.001)	LogN(-1.2, 0.1)
Less Smoothing	LogN(1,1)	Beta(20, 980)	Gamma(10, 100)	Gamma(10, 100)	LogN(-1.2, 0.1)
Biased Sampling	LogN(1,1)	Beta(20, 980)	Gamma(.001, 0.001)	Gamma(.001, 0.001)	-
Multistrain	LogN(1,1)	Beta(.02, .98)	Gamma(10000, 0.01)	Gamma(.001, 0.001)	LogN(-1.2, 0.1)

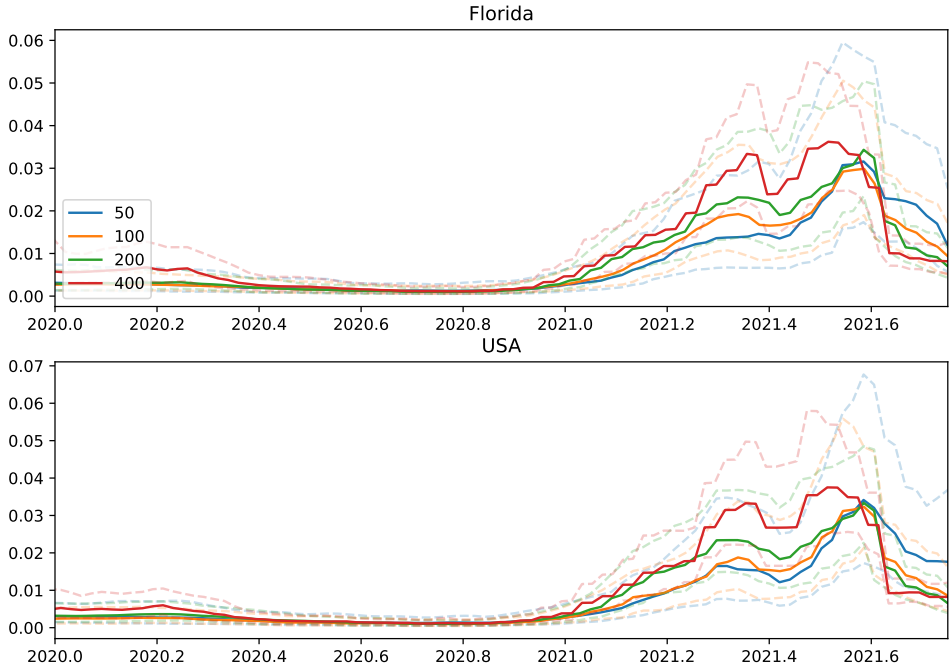

Figure 3.6: Posterior of s while varying the number of tips. Solid lines represent the median and the dotted lines represent the equal-tailed 95% credible intervals.

panel (USA), similar to the results for Florida, the posterior for R is very flat further back in the past. Given that we have seen large rises and falls in the number of cases over time (Figure 3.14), it seems unlikely that the actual value of R is as flat as the method suggests.

One explanation for this performance discrepancy is that the prior may be oversmoothing the estimates of R further back in the past for some of the data sets. Figure 3.13 shows the distribution of sample times for Florida, Michigan, and the USA. Michigan has a larger proportion of sequences sampled early in the pandemic compared to either Florida or the overall USA. Oversmoothing may occur because a lack of samples further back in the past causes the prior to overwhelm the data.

To investigate this, we reran the analysis with stronger hyperpriors designed to reduce the overall amount of smoothness ("Less Smoothing" in Table 3.1). Figure 3.8 shows the posterior when we set the prior of the smoothing parameter to be a gamma distribution with $a = 10$ and $b = 100$, giving a mean of 0.1 and variance 0.001. Looking at the top panel (Florida) of Figure 3.8, we see that the posterior median of R for VBSKY is no longer flat and instead oscillates to better match the results using surveillance data. The bottom panel (USA) also shows the estimates for R for the entire USA are also no longer completely flat further back in the past. The middle panel (Michi-

gan) shows that even with less smoothing, the results for VBSKY in Michigan match well with the surveillance data. When the sample time distribution is unbalanced, as with Florida and the USA, imposing less smoothing can help better capture the signal where the sampling may be more sparse. However, it also widens the credible intervals.

In addition to decreasing the amount of smoothing, we explored the use of a biased sampling scheme to yield sharper estimates further back in the past. The algorithm described in Section 3.5 generates an ensemble of trees by sampling the data randomly without replacement. Hence, if most of the samples were collected in the recent past, most of the trees in the ensemble will have tips from near the present, making it difficult to estimate transmission events further back in time. To verify this, we split the data by the quarter in which the sequence was sampled, where the first quarter of each year was defined to be the first three months (January, February, March) of the year, and so on. Then, instead of randomly sampling to generate the ensemble of trees, for each tree the tips were restricted to only one quarter. We also enforced the number of trees per quarter to be approximately equal. One caveat is that this stratified sampling approach could bias the estimates of the sampling rate.

Figure 3.9 shows the results using the biased sample approach. For this final analysis we reverted some of the smoothing prior changes ("Biased Sampling" in Table 3.1). (Because of convergence issues encountered during model fitting, for this scenario we fixed the origin to 0.3 years prior to the earliest sample date; therefore, no prior on x_1 is listed in the table.) There is a surprisingly close match between our model output and the ground-truth, which we reiterate was estimated using a completely different source of data. The estimates using the biased sampling approach improve the estimates of R further back in the past especially for Florida. Using less smoothing, VBSKY was able to capture the shape of estimates using surveillance data, but the biased sampling approach results in a much closer estimate of R further back in the past. The credible bands produced by VBSKY tend to be narrower, which could reflect either differences in the underlying data or violations of the modeling assumptions described in Section 3.5. Interestingly, both methods appear unable to reject the null hypothesis $R = 1$ except for very early in the pandemic (winter 2020) and very recently (spring-summer 2021). One drawback of the stratified sampling approach is that the estimates of R towards the present seem to be further away from the estimates using surveillance data. While using the biased sampling approach can improve estimates within time periods where sampling is sparse, it can also bias the estimates where sparse sampling is not an issue.

In this section we focused on estimating the effective reproduction number R. A parallel set of estimates for the sampling fraction s are shown in Figures 3.15–3.17.

3.3.2.3 Comparison to BEAST

We ran BEAST on the same data set as in the previous section. BEAST was incapable of analyzing the same number of samples as VBSKY, so to facilitate comparison, we limited the number of sequences we analyzed with BEAST. Both the sample size and the sampling scheme can affect the results of the analysis as well as the mixing time, so we compared how BEAST performed with different combinations of sample sizes and sampling schemes. We ran BEAST with both 100 and 500 sequences. For each sample size, we sampled the most recent sequences by date (contemporary sampling), and we also sampled uniformly at random without any regard to the sample time (random sampling). The XML configuration files we used to run BEAST are included in the supplementary data.

Even after greatly reducing the number of sequences analyzed, accurately sampling from the posterior may still take longer than using VBSKY. We performed both a "short" run for BEAST, where the MCMC sampler is only allowed to run for as long as it took VBSKY to analyze the full data, as well as a "long" run where BEAST was allowed to perform 100 MCMC million iterations, or run for 24 hours, whichever was shorter.

The estimates of the effective reproductive number of the short run for Florida, Michigan, and the USA are displayed in Figures 3.10, 3.18, and 3.19 respectively. The estimates for the long runs are shown in Figures 3.11, 3.20, and 3.21.

For the short runs, depending on the number of samples and the sampling scheme, the results varied widely. Under a short time constraint, the posteriors using 500 tips and both sampling schemes for Florida, 500 tips and recent sampling for Michigan and 500 tips and recent sampling for the USA were mostly flat centered close to 1. The posteriors did not reflect the rise and fall in R that is exhibited in both the surveillance data and VBSKY estimates. In most cases, BEAST is unable to capture any signal further back in the past, and the posterior provided by BEAST does not track the estimates provided by the surveillance data as well as VBSKY.

In the long runs, the issue of completely flat posteriors when using 500 tips mostly disappeared. However, BEAST is only capable of producing comparable results to VBSKY and the surveillance method when analyzing 100 tips sampled uniformly at random, presumably because mixing occurred more rapidly in the time allotted. The long runs also illustrate that uniform random sampling performs better than most-recent sampling when running BEAST. This indicates that having samples throughout time may help infer more transmission events further back in the past rather than having only contemporary sequences. The discrepancy between using 100 tips and 500 tips exists only when the sampling scheme is random. When using contemporary sequences, BEAST is able to complete 100 million iterations. But when random sampling is used, because the MCMC sampler mixes more slowly, BEAST was unable to complete 100 million MCMC moves within 24 hours.

In summary, BEAST performed fairly well when we randomly sample 100 tips, though there was considerable variation between data sets and scenarios. The main difference between VBSKY and BEAST is that the latter was usually unable to capture signal far back in the past. Analyzing more sequences could help, but the computational difficulties that would ensue imply that it is not practical to completely resolve this issue if time is a constraint. Overall, our results indicate that efficiently analyzing thousands of sequences, even using an approximate inference method, generally leads to a sharper posterior which is closer to the ground truth.

3.3.2.4 Strain Analysis

As a supplement to our main analysis, we further investigated the history of different COVID-19 variants. Using GISAID-annotated variant information, we split our data set of Florida, Michigan, and USA sequences into smaller data sets specific to the Alpha and Delta variant and fit our model to each variant.[1] Except for a minor adjustment to the prior on the origin time, we used all the same hyperparameters and priors as in the preceding section. For the GMRF smoothing prior, we chose a hyperior for τ_R to have large expectation to increase smoothing.

The results of our analysis are shown in Figure 3.12 for R and Figure 3.22 for s. The Alpha variant of COVID-19, also known as lineage B.1.1.7, originated in England and was first reported in the USA in early 2021. Using surveillance data, Volz et al. (2021) showed that at the time, the Alpha variant had a transmission advantage over other variants, which is why it came to dominate in the USA in early 2021. There are no samples for the Alpha variant beyond summer 2021, so the estimates for Alpha are truncated at various points during that period depending on the region considered. As shown in Figure 3.14, the number of cases in Michigan, Florida, and the USA all dropped after the first third of the year, corresponding to a decrease in R below one for the Alpha variant. At the same time, the Delta variant was rising in prevalence, such that R is estimated greater than one in all cases until about the third quarter of 2021. Analysis of the sampling fraction over time (Figure 3.22) also shows some interesting trends, for example sampling of the Delta variant in Michigan seems to have been extremely low compared to other areas and strains. Finally, we also explored using other hyperparameter settings to analyze these data, but found that they produced suboptimal results. In particular, without additional smoothing, our model unrealistically estimated that R increased for the Alpha variant throughout the second quarter of 2021, although the credible intervals generally place substantial posterior probability on the event $R < 1$ (Figures 3.23 and 3.24). We noticed that for the Alpha variant, the number of available samples drops severely near the point of truncation. The absence of data would lead to the prior dominating the posterior samples of R. By increasing smoothing, we were able to

[1] At the time this manuscript was written, there were no available sequences from the Omicron variant.

circumvent this issue.

3.4 Discussion

In this paper, we presented the variational Bayesian skyline, a method designed to infer evolutionary models from large phylogenetic datasets. Our method works by fitting a variational Bayesian posterior distribution to a certain approximation of the phylogenetic birth-death model. We showed that, under some simplifying heuristic assumptions, it can be used for posterior inference of epidemiologically relevant quantities such as the effective reproduction number and sampling fraction. We demonstrated that our estimates adhere reasonably closely to alternative approaches such as MCMC, while being significantly faster and therefore able to incorporate large numbers of observations. On real data, we showed how our model corroborates public health surveillance estimates, and could work to fill in the gaps when such data are unavailable.

One shortcoming of our model is that it tends to be overconfident, in the sense that it produces credible intervals which are narrower compared to other methods, and not as well calibrated in simulations. Generally, it is preferable for a method to overcover since this is inferentially more conservative. We believe this behavior is attributable to the heuristics that underlie our approach: since they ignore certain forms of dependence in the data, they create the illusion of a larger sample size than actually exists. We suggest that the credible intervals produce by our method are best interpreted relatively, as showcasing portions of time where the estimates are especially sharp or loose.

Our method could be extended in several ways. Currently, it estimates the tree topology and the continuous variables separately, relying on a distance-based method infer the topology. While faster, distance-based methods are less accurate than likelihood-based methods for tree reconstruction (Kuhner and Felsenstein, 1994). Our method could be potentially extended to unify the estimating procedure for tree topologies and other variables under one variational framework allowing (Zhang and Matsen IV, 2019). We also take random subsamples of data to accelerate our inference. However, the subsampling approach we adopt is very naive, and future work could include developing an improved way strategy for subsampling in phylogenetic problems.

The variational inference scheme we used makes a standard but highly simplified mean-field assumption about the dependence structure of the variational approximating family. We also experimented with other, recent approaches such as normalizing flows (Rezende and Mohamed, 2015), but observed that, consistent with earlier findings (Fourment and Darling, 2019), they did not measurably improve the results and occasionally caused the algorithm to fail to converge. If our approach is adapted to more complex problems, it could be advantageous to revisit this modeling choice.

Currently, our method is restricted to using a strict molecular clock model. Additionally, the substitution models in our method do not currently allow for rate heterogeneity across sites. Allowing for more flexible and complex substitution and clock models could aid in the application of our method to other data sets that evolve differently than COVID-19, when the time scale of the epidemic is much larger.

3.5 Materials and Methods

In this section, we derive our method, which we call variational Bayesian skyline (VBSKY). As the name suggests, VBSKY descends from a lineage of earlier methods designed to infer evolutionary rate parameters from phylogenetic data (Pybus et al., 2000; Drummond et al., 2005; Minin et al., 2008; Gill et al., 2013). Our running example will be inferring the epidemiological history of the COVID-19 pandemic, but the method applies generally to any evolving system that is aptly modeled using a phylogenetic birth-death or coalescent process and approximately meets the assumptions described below.

3.5.1 Notation and model

The data consists of a matrix of aligned sequences $\boldsymbol{\mathcal{D}} = \{A, C, G, T, N\}^{n \times L}$, where n is the number of viral sequences and L is the number of sites, and a vector of times when each sample was collected $\boldsymbol{y} = (y_1, \ldots, y_n)$ where $y_1 \leq \cdots \leq y_n$. Row j of $\boldsymbol{\mathcal{D}}$ corresponds to a sequenced viral genome collected from an infected host at time y_j. Subsamples of *rows* of $\boldsymbol{\mathcal{D}}$ are denoted by $\mathcal{D}_i \in \{A, C, G, T, N\}^{b \times L}$, with corresponding sample times $\boldsymbol{y}^{(i)} = (y_1^{(i)}, \ldots, y_b^{(i)})$, where b is the size of the subsample. We occasionally abuse notation and write $\mathcal{D}_i \subset \boldsymbol{\mathcal{D}}$ to denote a subsample, and $|\mathcal{D}|$ to denote the number of samples contained in a dataset (so e.g. $|\mathcal{D}_i| = b$ above). Phylogenetic trees are denoted by $\mathcal{T} = (\mathcal{T}^{\text{topo}}, \mathcal{T}^{\text{br}})$, which we decompose into a discrete topological component and continuous branch length component. Given n sampled taxa, the topological component $\mathcal{T}^{\text{topo}}$ lives in the space of rooted, labeled bifurcating trees on n leaves, and the branch length component lives in the non-negative orthant $\mathbb{R}_{\geq 0}^{2n-1}$ and gives the length of each edge of the tree (including an edge from crown to origin).

The data are assumed to be generated according to a phylogenetic birth-death skyline model (Nee et al., 1994; Morlon et al., 2011). In this model, samples are related by an unobserved "transmission tree" that records every infection event that occurred during the pandemic. Leaf nodes in the transmission tree represent sampling events, and internal nodes represent events where the virus was transmitted from one host to another. Edges denote periods during which the virus evolved within a particular host, with the length proportional to the amount of evolutionary time

that elapsed between the parent and child nodes. The distribution of the infection tree depends on three fundamental parameters, usually denoted $\mu(t)$, $\lambda(t)$, and ρ, which are respectively the time-varying per-capita rates at which extant lineages in the phylogeny go extinct and speciate, and the fraction of the extant population that was sampled at the present.

Further generalizations (Stadler et al., 2013) incorporate both random and deterministic sampling across time, and it was also shown how phylogenetic BD model can be used for parameter estimation in the susceptible-infected-recovered model (Kermack and McKendrick, 1927) that forms the foundation of quantitative epidemiology. Let $\psi(t)$ denote the rate at which each extant lineage is sampled in the phylogeny. (Henceforth we suppress dependence on time, but all parameters are allowed to be time-varying.) If we assume that sampling is tantamount to recovery (a valid assumption when positive testing leads to quarantine, as is generally the case during the current pandemic), then the overall rate of becoming uninfectious is $\delta = \mu + \psi$; the average time to recovery is $1/\delta$; the sampling proportion is $s = \psi/\delta$; and the effective reproduction number is $R = \lambda/\delta$. Using prior knowledge, it is also common to specify an origin time t_0 when the pandemic began.

Let $\zeta = (R, \delta, s, t_0)$ denote the vector of epidemiological parameters of interest. The hyperprior on ζ is denoted $\pi(\zeta)$. The latent transmission tree describing the shared evolutionary history of all of the sampled pathogens is denoted by $\boldsymbol{\mathcal{T}} = (\boldsymbol{\mathcal{T}}^{\text{topo}}, \boldsymbol{\mathcal{T}}^{\text{br}})$. We assume a simple "strict clock" model, with known rates of substitution, so that no additional parameters are needed to complete the evolutionary model.

We desire to sample from the posterior distribution of ζ given the phylogenetic dataset $\boldsymbol{\mathcal{D}}$. Let $p(\boldsymbol{\mathcal{T}} \mid \zeta)$ denote the likelihood of the transmission tree given the evolutionary model. An expression for $p(\boldsymbol{\mathcal{T}} \mid \zeta)$ can be found in Stadler et al. (2013, Theorem 1), and is reproduced in Appendix 3.6.1 for completeness. The data depend on ζ only through $\boldsymbol{\mathcal{T}}$, so that $p(\boldsymbol{\mathcal{D}} \mid \boldsymbol{\mathcal{T}}, \zeta) = p(\boldsymbol{\mathcal{D}} \mid \boldsymbol{\mathcal{T}})$. Here $p(\boldsymbol{\mathcal{D}} \mid \boldsymbol{\mathcal{T}})$ denotes the "phylogenetic likelihood", which can be efficiently evaluated using the pruning algorithm (Felsenstein, 1981). Putting everything together, the posterior distribution over the unobserved model parameters is

$$p(\zeta, \boldsymbol{\mathcal{T}} \mid \boldsymbol{\mathcal{D}}) \propto p(\boldsymbol{\mathcal{D}} \mid \boldsymbol{\mathcal{T}}) p(\boldsymbol{\mathcal{T}} \mid \zeta) \pi(\zeta). \tag{3.1}$$

3.5.2 Scalable inference

The constant of proportionality in (3.1) is $p(\boldsymbol{\mathcal{D}})$, the marginal likelihood after integrating out all (hyper)parameters and the unobserved tree $\boldsymbol{\mathcal{T}}$. In large phylogenetic data sets, exact evaluation of the marginal likelihood is impossible due to the need to enumerate all possible trees, a set whose cardinality explodes in the number of taxa (Alfaro and Holder, 2006). In practice, methods such as Markov chain Monte Carlo (e.g., Drummond and Rambaut, 2007) which do not require evaluating $p(\boldsymbol{\mathcal{D}})$ are utilized.

Since current phylogenetic MCMC algorithms cannot scale up to pandemic-sized datasets, we propose to modify the inference problem (3.1) using a few heuristics in order to make progress. Let $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_S \subset \boldsymbol{\mathcal{D}}$ be subsamples of $b_1, \ldots, b_S$ rows from the full dataset. If the subsamples are temporally and geographically separated, and $b_i \ll n$, then it is reasonable to suppose that these subsamples are approximately independent conditional on the underlying evolutionary model.

Heuristic 2. *In a very large phylogenetic dataset $\boldsymbol{\mathcal{D}}$, small subsets $\mathcal{D}_1, \mathcal{D}_2 \subset \mathcal{D}$ with $|\mathcal{D}_1|, |\mathcal{D}_2| \ll |\boldsymbol{\mathcal{D}}|$ that are sufficiently separated in space and/or time are approximately independent:* $p(\mathcal{D}_1, \mathcal{D}_2 \mid \zeta) \approx p(\mathcal{D}_1 \mid \zeta) p(\mathcal{D}_2 \mid \zeta)$.

True independence holds, for example, when the clades corresponding to $\mathcal{D}_1, \mathcal{D}_2$ are so distant that a reversible substitution process reaches stationarity on the edge connecting them. While we do not expect this to occur in real data, it seems like a reasonable approximation for studying distant subclades in a large, dense phylogeny which are evolving under a common evolutionary model. An example of the subsampling scheme we have in mind is when $\boldsymbol{\mathcal{D}}$ = "all of the samples collected in Florida" ($n \approx 81,000$), $\mathcal{D}_1$ = "all of the samples collected in Florida during June, 2020" ($b_1 \approx 300$), and $\mathcal{D}_2$ = "all of the samples collected in Florida during June, 2021" ($b_2 \approx 5,100$).

Though incorrect, Heuristic 2 furnishes us with a useful formalism for performing large-scale inference, as we now demonstrate. Using the heuristic, we can approximate the posterior distribution (3.1) as

$$p(\zeta, \mathcal{T}_{1:S} \mid \mathcal{D}_{1:S}) \propto \pi(\zeta) \prod_{i=1}^{S} p(\mathcal{D}_i \mid \mathcal{T}_i) p(\mathcal{T}_i \mid \zeta), \tag{3.2}$$

where we used the array notation $\mathcal{T}_{1:S} \equiv (\mathcal{T}_1, \ldots, \mathcal{T}_S)$ to streamline the presentation. Sampling from (3.2) is easier than sampling from the full posterior (3.1) since it decomposes into independent subproblems, and each subtree $\mathcal{T}_i$ is much smaller than the global phylogeny $\mathcal{T}$. However, the normalizing constant in (3.2) remains intractable even for small trees, so naive sampling would still require expensive MCMC algorithms.

To work around this, we start by rewriting the last term in (3.2) as

$$p(\mathcal{T}_i \mid \zeta) = p(\mathcal{T}_i^{\text{br}} \mid \mathcal{T}_i^{\text{topo}}, \zeta) p(\mathcal{T}_i^{\text{topo}} \mid \zeta).$$

As noted in the introduction, the primary difficulty in Bayesian phylogenetic inference is navigating regions of topological tree space that have high posterior probability. If we could efficiently sample $\hat{\mathcal{T}}_i^{\text{topo}} \sim p(\mathcal{T}_i^{\text{topo}} \mid \zeta)$, then the approximate posterior

$$\hat{p}(\zeta, \mathcal{T}_{1:S}^{\text{br}} \mid \hat{\mathcal{T}}_{1:S}^{\text{topo}}, \mathcal{D}_{1:S}) \propto \pi(\zeta) \prod_{i=1}^{S} p(\mathcal{D}_i \mid \mathcal{T}_i^{\text{br}}, \hat{\mathcal{T}}_i^{\text{topo}}) p(\mathcal{T}_i^{\text{br}} \mid \hat{\mathcal{T}}_i^{\text{topo}}, \zeta) \tag{3.3}$$

would have the property that

$$\mathbb{E}_{\hat{\mathcal{T}}_{1:S}^{\text{topo}}} \hat{p}(\zeta, \mathcal{T}_{1:S}^{\text{br}} \mid \hat{\mathcal{T}}_{1:S}^{\text{topo}}, \mathcal{D}_{1:S}) = p(\zeta, \mathcal{T}_{1:S}^{\text{br}} \mid \mathcal{D}_{1:S}). \tag{3.4}$$

This leads to our second heuristic.

Heuristic 3. *Fitted tree topologies $\hat{\mathcal{T}}_{1:S}^{\text{topo}}$ obtained from subsets $\mathcal{D}_1, \ldots, \mathcal{D}_m$ pairwise satisfying Heuristic 2 are independent and approximately distributed as $p(\mathcal{T}^{\text{topo}} \mid \zeta)$.*

By "fitted trees" we mean trees estimated using any method, including fast heuristic algorithms such as UPGMA, or its extension to serially-sampled time trees (sUPGMA; Drummond and Rodrigo, 2000); maximum likelihood; or simply extracting subtrees from a high-quality, precomputed reference phylogeny (e.g., Lanfear, 2020). The heuristic can fail in various ways: in reality, tree reconstruction algorithms do not necessarily target the correct/any evolutionary prior, and there could be dependence between different trees if they are jointly estimated as part of a larger phylogeny. Also, our current implementation uses the data twice, once to estimate each tree, and again during model fitting to evaluate its phylogenetic likelihood. The tree inference procedure we used to analyze data in this paper is described more fully in the supplement (Section 3.6.2). Note that we only utilize the *topological* information from these procedures; we still perform posterior inference over the branch lengths $\mathcal{T}^{\text{br}}$ as detailed below.

Setting these caveats aside, the point of Heuristic 3 is to endow our posterior estimates with some measure of phylogenetic uncertainty, without resorting to full-blown MCMC in tree space. By (3.4), the approximate likelihood (3.3) is unbiased for $p(\zeta, \mathcal{T}_{1:S}^{\text{br}} \mid \mathcal{D}_{1:S})$, and the latter quantity correctly accounts for phylogenetic variance in the posterior. However, since (3.3) conditions on $\hat{\mathcal{T}}_{1:S}^{\text{topo}}$, all of the remaining parameters to be sampled are continuous, and the problem becomes much easier.

Finally, we point out that our method is not capable generating useful samples from the posterior distribution $p(\boldsymbol{\mathcal{T}} \mid \boldsymbol{\mathcal{D}})$, that is of the overall transmission tree given the original dataset $\boldsymbol{\mathcal{D}}$. But, as noted above, in skyline-type models the main object of interest is the evolutionary posterior $p(\zeta \mid \boldsymbol{\mathcal{D}})$. In Section 3.3, we demonstrate that the heuristic, subsampling-based approach developed here yields a fairly sharp posterior on ζ, while still utilizing a large amount of information from $\boldsymbol{\mathcal{D}}$.

3.5.2.1 Stochastic variational inference

Since (3.3) is a distribution over continuous, real-valued parameters, it is amenable to variational inference (Jordan et al., 1999). As noted in the introduction, variational Bayesian phylogenetic inference has previously been studied by Zhang and Matsen IV (2019); Zhang (2020) and Fourment and Darling (2019). Our approach is most related to the latter since we do not optimize over

the topological parameters of our model in any way. Because we are operating in a different data regime than either of these two pre-pandemic papers, we further incorporated recent advances in large-scale Bayesian inference in order to improve the performance of our method.

Given a Bayesian inference problem consisting of data $\boldsymbol{x}$ and model parameters $\boldsymbol{z}$, traditional VI seeks to minimize the Kullback-Leibler (KL) divergence between the true posterior of interest and family of tractable approximating distributions $\mathcal{Q}$:

$$q^*(\boldsymbol{z}) = \underset{q(\boldsymbol{z})\in\mathcal{Q}}{\arg\min}\, \mathrm{KL}(q(\boldsymbol{z}) \parallel p(\boldsymbol{z} \mid \boldsymbol{x})).$$

We cannot carry out this minimization as the KL divergence still requires evaluating the intractable quantity $p(\boldsymbol{x})$. However,

$$\begin{aligned} \mathrm{KL}(q(\boldsymbol{z}) \parallel p(\boldsymbol{z} \mid \boldsymbol{x})) &= \mathbb{E}(\log q(\boldsymbol{z})) - \mathbb{E}(\log p(\boldsymbol{z} \mid \boldsymbol{x})) \\ &= \mathbb{E}(\log q(\boldsymbol{z})) - \mathbb{E}(\log p(\boldsymbol{x}, \boldsymbol{z})) + \log p(\boldsymbol{x}) \\ &= -\mathrm{ELBO}(q(\boldsymbol{z})) + \text{const.} \end{aligned} \tag{3.5}$$

where the expectations are with respect to the variational distribution q, and

$$\mathrm{ELBO}(q(\boldsymbol{z})) := \mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z})} \left[\log p(\boldsymbol{x}, \boldsymbol{z}) - \log q(\boldsymbol{z})\right] \tag{3.6}$$

is known as the evidence lower bound. Hence, minimizing the divergence between the true and variational posterior distributions is equivalent to maximizing the ELBO.

For VI involving complex (non-exponential family) likelihoods, the ELBO is generally approximated by replacing the first term in (3.6) by a Monte Carlo estimate:

$$\mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z})} \log p(\boldsymbol{x}, \boldsymbol{z}) \approx \frac{1}{B} \sum_{i=1}^{B} \log p(\boldsymbol{x}, \boldsymbol{z}_i); \quad \boldsymbol{z}_1, \ldots, \boldsymbol{z}_B \sim q(\boldsymbol{z}) \text{ i.i.d.} \tag{3.7}$$

where $B = 1$ is a common choice. Each evaluation of the complete likelihood $\log p(\boldsymbol{x}, \boldsymbol{z})$ requires a full pass over the data, which can be prohibitive when the data are large. Stochastic variational inference (SVI; Hoffman et al., 2013) addresses this problem through stochastic optimization. Many Bayesian models naturally factorize into a set of shared, global hidden variables, and sets of local hidden variables which are specific to each observation. Each observation is conditionally independent of all others given its local parameters. Hoffman et al. show how models of this form are well suited to stochastic gradient descent. Specifically, they derive an unbiased gradient estimator of the ELBO (3.6) which operates on a single, randomly sampled data point at each iteration. The algorithm tends to make better progress in early stages when the variational approximation to the

shared global parameters is still quite inaccurate (Hoffman et al., 2013).

By design, the model we derived above is suited to SVI. In equation (3.3), the evolutionary parameters ζ are shared among all datasets, while the branch length parameters $\mathcal{T}_i^{\text{br}}$ are specific to the ith dataset $\mathcal{D}_i$. We therefore refer to ζ as the global parameter, and the vectors of dataset-specific branch lengths $\mathcal{T}_{1:S}^{\text{br}}$ as local parameters. Our algorithm proceeds by iteratively sampling a single dataset $\mathcal{D}_i$ and taking a noisy (but unbiased) gradient step. Note that, because our model is not in the exponential family, we cannot employ the elegant coordinate-ascent scheme originally derived by Hoffman et al.. Instead, we numerically optimize the ELBO using differentiable programming (see below).

3.5.2.2 Model parameterization

It remains to specify our model parameterization and the class of distributions $\mathcal{Q}$ that are used to approximate the posterior. Recall from Section 3.5.1 that the global parameter ζ includes the effective reproduction number $R(t)$, rate of becoming uninfectious $\delta(t)$, and sampling fraction $s(t)$. We follow earlier work (Gill et al., 2013) in assuming that these rate functions are piecewise constant over time, with changepoints whose location and number are fixed *a priori*. The changepoints are denoted $\mathbf{t} = (t_1, \ldots, t_m)$ satisfying $0 = t_0 < t_1 < \cdots < t_m < t_{m+1} = \infty$. Thus,

$$R(t) = \sum_{i=1}^{m+1} R_i \mathbf{1}_{\{t \in [t_{i-1}, t_i)\}}(t),$$

where the transmission rates in each time interval are denoted $\boldsymbol{R} = (R_1, \ldots, R_m) \in \mathbb{R}_{>0}^m$. The rate of becoming uninfectious and sampling fraction are similarly denoted by $\boldsymbol{\delta} \in \mathbb{R}_{>0}^m$ and $\boldsymbol{s} \in [0,1]^m$, respectively. Finally, a Gaussian Markov random field (GMRF) smoothing prior is used to penalize consecutive differences in the log rates (Minin et al., 2008). To account for the fact that each rate parameter may have varying degrees of smoothness and also could be on different scales, each rate parameter has a corresponding precision hyperparameter τ_R, τ_δ, and τ_s.

An extension of the BDSKY model allows for additional sampling efforts at each time t_k. Infected individuals are sampled with probability ρ_k at time t_k. When all sequences are sampled serially without the added sampling effort, $\rho_k = 0$ for $1 \leq k \leq m$. When all sequences are sampled contemporaneously, $\boldsymbol{\psi} = \mathbf{0}$, $\rho_k = 0$ for $1 \leq k \leq m-1$, and $\rho_m > 0$. For our work, we only consider cases where $\rho_k = 0$ for $1 \leq k \leq m-1$. We define b_s as the number of sequences sampled serially, and b_m to be the number of sequences sampled at time t_m. In other words, b_m is the number of contemporaneously sampled sequences at time t_m. Note that $b = b_m + b_s$. The sample times of the b_s serially sampled sequences are denoted by $\tilde{\boldsymbol{y}}^{(i)} = (y_1^{(i)}, \ldots, y_{b_s}^{(i)})$. Because the sequences sampled at t_m have the largest sample time, $\tilde{\boldsymbol{y}}^{(i)}$ is just a truncated version of $\boldsymbol{y}^{(i)}$.

When all sequences are sampled serially, $\boldsymbol{y}^{(i)} = \tilde{\boldsymbol{y}}^{(i)}$. To conserve notation, from this point onward, we will use $\boldsymbol{y}^{(i)}$ to refer to $\tilde{\boldsymbol{y}}^{(i)}$.

The final remaining global parameter is the epidemic origin time t_0. In order for the model to be well defined, this must occur earlier than the earliest sampling time in any of the S subsamples. Therefore, we set $t_0 + x_1 = y_{\text{min}}$, where y_{min} is the earliest sampling time across all subsamples, and place a prior on $x_1 > 0$ as detailed below.

Given the sampling times and estimated tree topology $\hat{\mathcal{T}}_i^{\text{topo}}$, we can identify each local parameter $\mathcal{T}_i^{\text{br}}$ with a vector $\boldsymbol{h}^{(i)} \in \mathbb{R}_{>0}^{b-1}$ giving the height of each internal node when enumerated in preorder. Hence the height of the root node is $h_1^{(i)}$. We follow the parameterizations set forth by Fourment and Darling (2019). In order for a sampled tree to be valid, we must have $h_j^{(i)} < h_{\text{pa}(j)}^{(i)}$ for every j. Here $\text{pa}(j)$ denotes the parent node of node j. This constraint can be met by setting the height of internal node j as $h_j^{(i)} = p_j^{(i)}(h_{\text{pa}(j)}^{(i)} - h_{d(j)}^{(i)})$ where $d(j)$ is the earliest sampled tip from the set of descendants of j and $p_j^{(i)} \in [0, 1]$. Finally, let $x_1^{(i)}$ denote the distance of the root node from the origin measured forward in time. We must have $t_0 < x_1^{(i)} < y_1^{(i)}$ since the root node of $\mathcal{T}_i$ has to be between the origin and the earliest sample time. Therefore we set $x_1^{(i)} - t_0 = r^{(i)} y_1^{(i)}$ for some $r^{(i)} \in [0, 1]$, and calculate the root height $h_1^{(i)}$ from it. Under this parameterization, the set of local variables $\boldsymbol{z}^{(i)} = (p_1^{(i)}, \ldots, p_{b-1}^{(i)}, r^{(i)}) \in [0, 1]^b$ is a set of proportions, with transformations to switch between parameterizations for BDSKY and the observed data likelihood.

3.5.2.3 Variational approximating family

We make a standard mean field assumption, which posits that members of $\mathcal{Q}$ completely factorize into a product of independent marginals. Letting $\boldsymbol{\zeta} = (R_1, \ldots, R_m, \delta_1, \ldots, \delta_m, s_1, \ldots, s_m)$ denote the collection of all global parameters defined above, and recalling the definition of $\boldsymbol{z}^{(i)}$ in the preceding paragraph, we assume that

$$q(\boldsymbol{\zeta}, \boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(m)}) = \prod_i q(\zeta_i \mid \pi_i) \prod_j \prod_k q(z_j^{(k)} \mid \phi_j^{(k)}), \tag{3.8}$$

where we have introduced variational parameters π_i and $\phi_j^{(k)}$ corresponding to each marginal distribution. The distributions $q(\zeta_i \mid \pi_i)$ and $q(z_j^{(k)} \mid \phi_j^{(k)})$ are (suitably transformed) Gaussians, so that $\pi_i, \phi_j^{(k)} \in \mathbb{R} \times \mathbb{R}_{\geq 0}$ each comprises a real location parameter and non-negative scale parameter. In our model, all latent parameters, local or global, are constrained to be positive (e.g., $\boldsymbol{R}, \boldsymbol{\delta}$) or in the unit interval (e.g., $\boldsymbol{s}$, $\boldsymbol{z}^{(i)}$). For each parameter we take q to be an appropriately transformed normal distribution. For positive parameters we use an exponential transformation, and for parameters constrained to be in $(0, 1)$ we use an expit (inverse logistic) transformation.

3.5.2.4 Implementation using differentiable programming

Our Python software implementation uses automatic differentiation in order to efficiently optimize the variational objective function (Kucukelbir et al., 2017; Bradbury et al., 2018). We sample from the variational distribution and estimate the gradient of the (3.7) objective function with respect to the variational parameters $\boldsymbol{\pi}$ and $\boldsymbol{\phi}$ using Monte Carlo integration (cf. eqn. 3.7). Gradients of the phylogenetic likelihood are computed in linear time using the recent algorithm of Ji et al. (2020). The complete fitting algorithm is shown in Algorithm 3.

Algorithm 3 Variational Bayesian Skyline (VBSKY)

Require: Data set $\boldsymbol{\mathcal{D}} \in \{A, C, G, T, N\}^{n \times L}$
  Sampling times $\boldsymbol{y} \in \mathbb{R}^n_{\geq 0}$
  Fixed parameters m, S, b ▷ Number of intervals, number of trees, subsample size
  Step size α
  for all $1 \leq i \leq S$ **do**
    Sample with replacement b times from the data to get subsample $\mathcal{D}_i, \boldsymbol{y}^{(i)}$
    Estimate the tree topology $\hat{\mathcal{T}}_i^{\text{topo}}$
  end for
  Initialize $\boldsymbol{\pi}, \boldsymbol{\phi}$ randomly.
  while not converged **do**
    for all $1 \leq i \leq S$ **do**
      Draw M samples $\boldsymbol{z}^{(i)} \sim q(\cdot \mid \boldsymbol{\phi}^{(i)})$, $\boldsymbol{\zeta} \sim q(\cdot \mid \boldsymbol{\pi})$
      Approximate $\nabla_{\boldsymbol{\phi}^{(i)}} \mathcal{L}$ and $\nabla_{\boldsymbol{\pi}} \mathcal{L}$ using MC integration
      $\boldsymbol{\phi}^{(i)} \leftarrow \boldsymbol{\phi}^{(i)} + \alpha \nabla_{\boldsymbol{\phi}^{(i)}} \mathcal{L}$
      $\boldsymbol{\pi} \leftarrow \boldsymbol{\pi} + \alpha \nabla_{\boldsymbol{\pi}} \mathcal{L}$
    end for
  end while
  return $\boldsymbol{\pi}, \boldsymbol{\phi}$ ▷ Algorithm 3

Acknowledgments

This research was supported by the National Science Foundation (grant number DMS-2052653, and a Graduate Research Fellowship).

Data availability

All of the data analyzed in this manuscript are publicly available. A Python implementation of our method, as well as Jupyter notebooks which reproduce our results, are located at https://github.com/jthlab/vbsky.

3.6 Appendix

3.6.1 Birth Death Skyline

In this section we review the birth-death skyline (BDSKY) model of Stadler et al. (2013). BDSKY is a forward time model which begins with a single individual at time t_0 and ends at t_m. Throughout this section, we will refer to the start of the process as the origin. As with other skyline methods, the parameters of the model are allowed to vary over time. Specifically, given a vector $\mathbf{t} = (t_0, t_1, \ldots, t_m)$ satisfying $0 < t_1 < \cdots < t_m$, parameters are fixed between each t_k and t_{k-1}, and allowed to vary m times. The transmission rates are denoted by the vector $\boldsymbol{\lambda} \in \mathbb{R}^m_{>0}$. Similarly, the death rates are given by the vector $\boldsymbol{\mu}$ and the sampling rates by the vector $\boldsymbol{\psi}$ where each $\mu_k > 0$ and $\psi_k > 0$. In the interval $[t_{k-1}, t_k)$, every infected individual transmits at rate λ_k, recovers at rate μ_i, and is sampled at rate ψ_k. For ease of notation, we denote $\lambda(t), \mu(t)$, and $\psi(t)$ as the transmission rate, uninfectious rate, and sampling rate at time t. We assume that after sampling, the individual can no longer transmit. This assumption holds in reality for many viruses as sampling is often followed by treatment or changes in behavior that would curb or limit spread. For example, those sampled with HIV would undergo antiretroviral therapy or those sampled with COVID-19 would quarantine themselves.

As described in the main text, the BDSKY model also allows for additional sampling efforts at each time t_k. For the reader's convenience we reproduce the notation here. All infected are sampled with rate ρ_k at time t_k. When all sequences are sampled serially without the added sampling effort, $\rho_k = 0$ for $1 \leq k \leq m$. When all sequences are sampled contemporaneously, $\boldsymbol{\psi} = \mathbf{0}$, $\rho_k = 0$ for $1 \leq k \leq m-1$, and $\rho_m > 0$. For our work, we only consider cases where $\rho_k = 0$ for $1 \leq k \leq m-1$. We define b_s as the number of sequences sampled serially, and b_m to be the number of sequences sampled at time t_m. In other words, b_m is the number of contemporaneously sampled sequences at time t_m. Note that $b = b_m + b_s$. The sample times of the b_s serially sampled sequences are denoted by $\tilde{\boldsymbol{y}}^{(i)} = (y_1^{(i)}, \ldots, y_{b_s}^{(i)})$. Because the sequences sampled at t_m have the largest sample time, $\tilde{\boldsymbol{y}}^{(i)}$ is just a truncated version of $\boldsymbol{y}^{(i)}$. When all sequences are sampled serially, $\boldsymbol{y}^{(i)} = \tilde{\boldsymbol{y}}^{(i)}$. To conserve notation, from this point onward, we will use $\boldsymbol{y}^{(i)}$ to refer to $\tilde{\boldsymbol{y}}^{(i)}$. The $b-1$ transmission event times are denoted by $\boldsymbol{x}^{(i)} = (x_1^{(i)}, \ldots, x_{b-1}^{(i)})$ where $0 < x_1^{(i)} < \cdots < x_{b-1}^{(i)}$.

The number of lineages that began before t_k and are extant at t_k is n_k. Any tree $\mathcal{T}_i$ induced by the BDSKY model is described by its tree topology $\mathcal{T}_i^{\text{topo}}$, the transmission times $\boldsymbol{x}^{(i)}$, and the sampling times $\boldsymbol{y}^{(i)}$. Letting S be the event that at we observe at least one sample, the probability density of a tree under the BDSKY model is

$$p(\mathcal{T}_i \mid \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\psi}, \boldsymbol{\rho}, \boldsymbol{t}, S) = \frac{q_1(0)\rho_m^{b_m}}{1 - p_1(0)} \prod_{k=1}^{b-1} \lambda_{I(x_k^{(i)})} q_{I(x_k^{(i)})}(x_k^{(i)}) \prod_{k=1}^{n_s} \frac{\psi_{I(y_k^{(i)})}}{q_{I(y_k^{(i)})}(y_k^{(i)})} \prod_{k=1}^{m} q_{k+1}(t_k)^{n_k}, \quad (3.9)$$

where $I(t) = k$ if $t_{k-1} \leq t < t_k$, and for $k = 1, \ldots, m$ and $t_{k-1} \leq t < t_k$,

$$\begin{aligned}
A_k &= \sqrt{(\lambda_k - \mu_k - \psi_k)^2 + 4\lambda_k\psi_k} \\
B_k &= \frac{(1 - 2(1-\rho_k)p_{k+1}(t_i))\lambda_k + \mu_k + \psi_k}{A_k} \\
p_k(t) &= \frac{\lambda_k + \mu_k + \psi_k - A_k \frac{e^{A_k(t_k - t)}(1+B_k) - (1-B_k)}{e^{A_k(t_k-t)}(1+B_k) + (1-B_k)}}{2\lambda_k} \\
q_k(t) &= \frac{4e^{-A_k(t-t_k)}}{(e^{-A_k(t-t_k)}(1+B_k) + (1-B_k))^2},
\end{aligned}$$

and $p_{m+1}(t_m) = 1$.

3.6.2 Estimating the tree topology

In this section we explain how we estimate the tree topology $\hat{\mathcal{T}}_i^{\text{topo}}$ for each subsample $\mathcal{D}_i$. We employed a simple heuristic method by fitting serial-sample unweighted pair grouping method with arithmetic means (sUPGMA) (Drummond and Rodrigo, 2000). As the name alludes to, sUPGMA is a tree reconstruction algorithm based on the unweighted paired group method with arithmetic means (UPGMA) (Sneath and Sokal, 1973).

We first describe the UPGMA algorithm followed by the sUPGMA algorithm. Both algorithms require a pairwise distance matrix. Taking the sequences of our subsample, $\mathcal{D}_i \in \{A, C, G, T\}^{b \times L}$, we simply take the Hamming distance between all $\binom{b}{2}$ pairs of sequences. That is for a given pair of sequences $s, t \in \mathcal{D}_i$, the distance is $d(s,t) = \sum_{k=1}^{L} \mathbb{1}_{(s_k \neq t_k)}$. For clusters, the mean distance between each element in the clusters. That is

$$d(A, B) = \frac{1}{|A||B|} \sum_{s \in A} \sum_{t \in B} d(s, t),$$

where A and B both represent clusters. At each step the two clusters with the smallest distance between them are combined into a new cluster, and the distances are recalculated between the newly formed cluster and the other clusters. Clusters can be made up of a single sequence. We start with b clusters initially and reduce the number of clusters by one at each step. This is repeated until there is only one single cluster.

While we could naively use UPGMA to get our tree topologies, because UPGMA does not account for sample times, it is possible the algorithm would give us topologies that are impossible given the sample times. We use sUPGMA to ensure that the estimated topology is not only possible, but realistic. Consider our subsample of sample times $\boldsymbol{y}^{(i)} = \{y_1^{(i)}, y_2^{(i)}, \ldots, y_b^{(i)}\}$. Let $d(u_i, v_j)$ be the distance between ith sequence with sample time $y_u^{(i)}$ and the jth sequence with sample time

y_v. We assume that $u < v$. and model $d(u_i, v_j)$ by its expectation, $\mathbb{E}(d(u_i, v_j)) = \Theta_u + \omega(y_v^{(i)} - y_u^{(i)})$, where Θ_u is the expected average distance between any two sequences at time $y_u^{(i)}$, and ω is the expected number of substitutions per unit time. The procedure for sUPGMA is as follows.

1. Estimate the set of parameters $\{\Theta_1, \ldots, \Theta_q, \omega\}$ using regression:

$$d(u_i, v_j) = \sum_{k=1}^{q} \Theta_k X_k + \omega(y_v^{(i)} - y_u^{(i)}) + \epsilon,$$

   where $X_k = 1$ if $k = u$ and $k = 0$ otherwise.

2. Correct the original pairwise distances

$$c(u_i, v_j) = d(u_i, v_j) + \omega(y_u^{(i)} + y_v^{(i)} - 2y_1^{(i)}).$$

3. Cluster using the UPGMA algorithm as described earlier.

We note that while the complete sUPGMA algorithm returns both the tree topology and the branch lengths, we only use this procedure to obtain the tree topology. As branch lengths are continuous variables, we will estimate those using stochastic variational inference. Although there are maximum likelihood based tree reconstruction methods we could use such as IQ-TREE (Minh et al., 2020), since we are only concerned with the topology rather than the entire tree, we prioritize the faster algorithm.

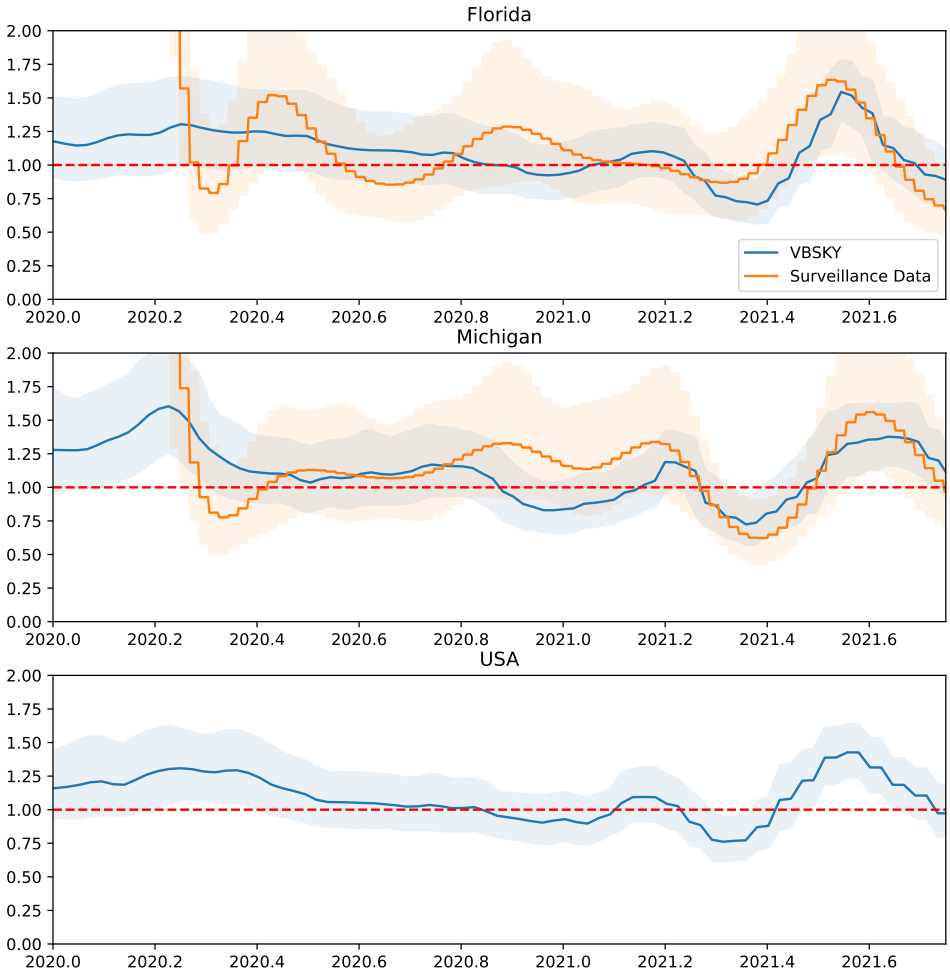

Figure 3.7: Posterior of R for Florida, Michigan, and the USA using an uninformative smoothing prior. VBSKY estimates are in blue. The orange estimates are derived from surveillance data. For each method the posterior median and equal-tailed 95% credible interval are shown. The dotted red line is $R = 1$.

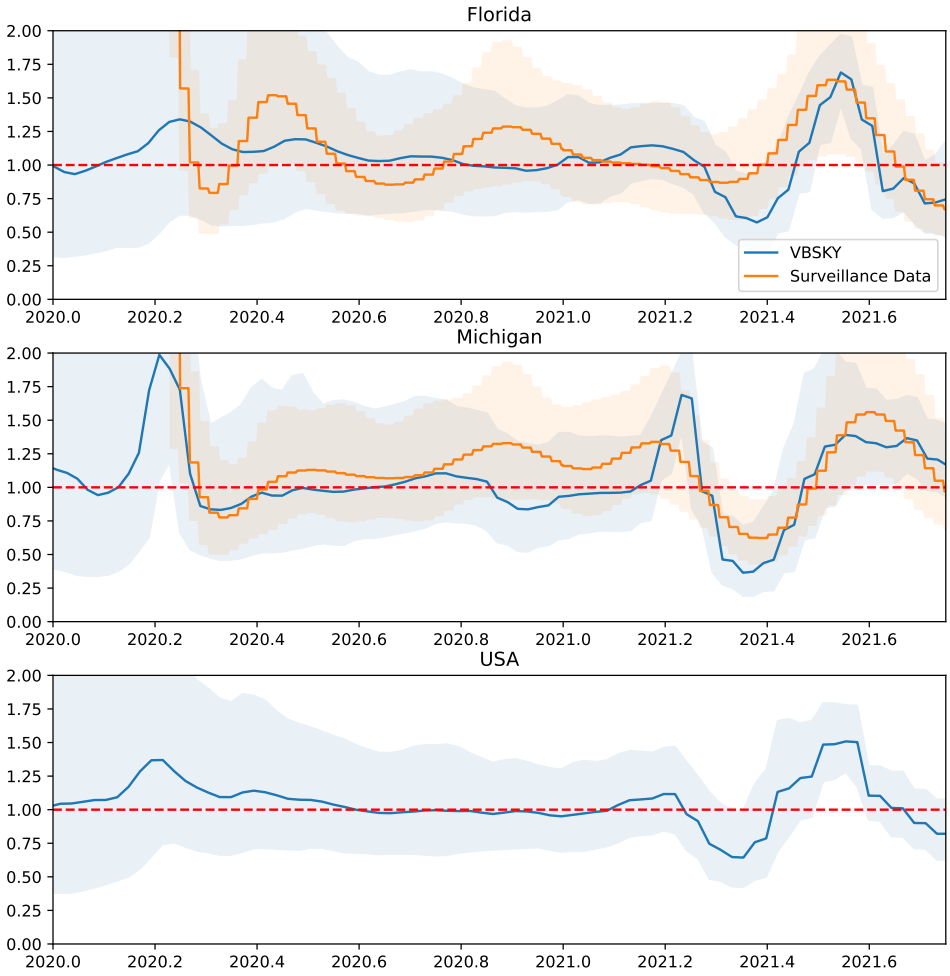

Figure 3.8: Posterior of R for Florida, Michigan, and the USA using less smoothing. VBSKY estimates are in blue. The orange estimates are derived from surveillance data. For each method the posterior median and equal-tailed 95% credible interval are shown. The dotted red line is $R = 1$.

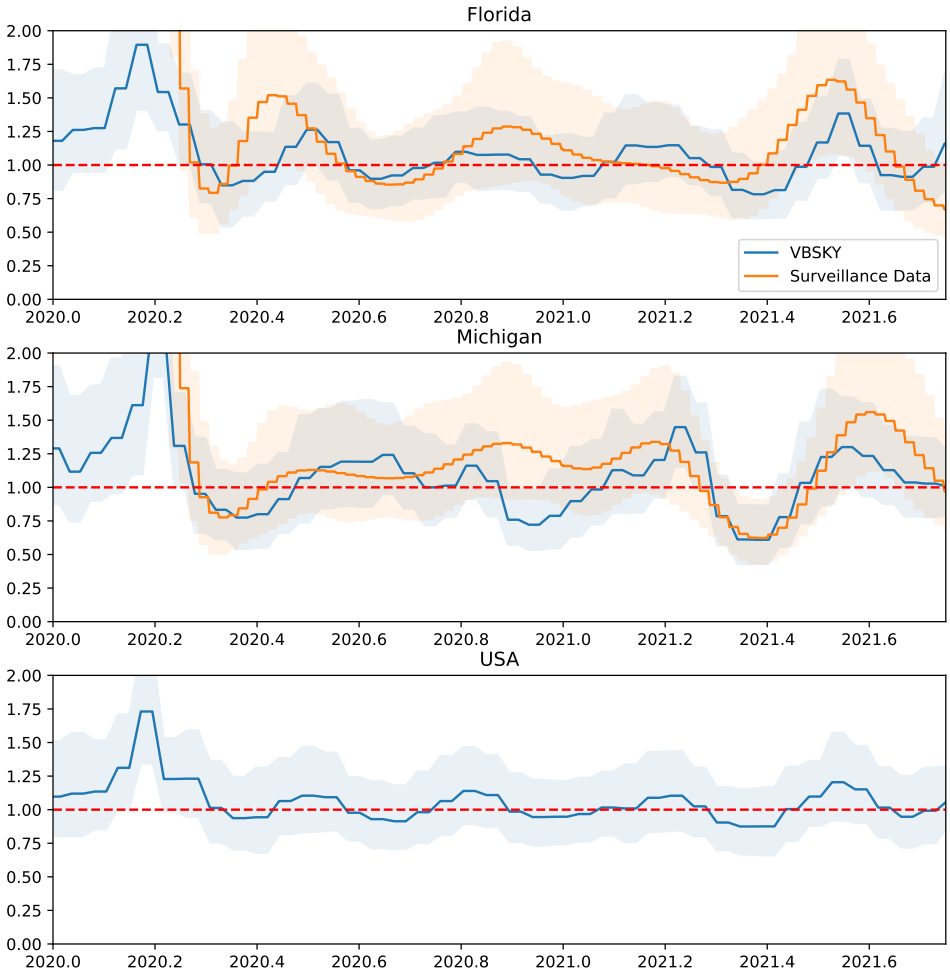

Figure 3.9: Posterior of R for Florida, Michigan, and the USA using biased sampling and a strong prior on s. VBSKY estimates are in blue. The orange estimates are derived from surveillance data. For each method the posterior median and equal-tailed 95% credible interval are shown. The dotted red line is $R = 1$.

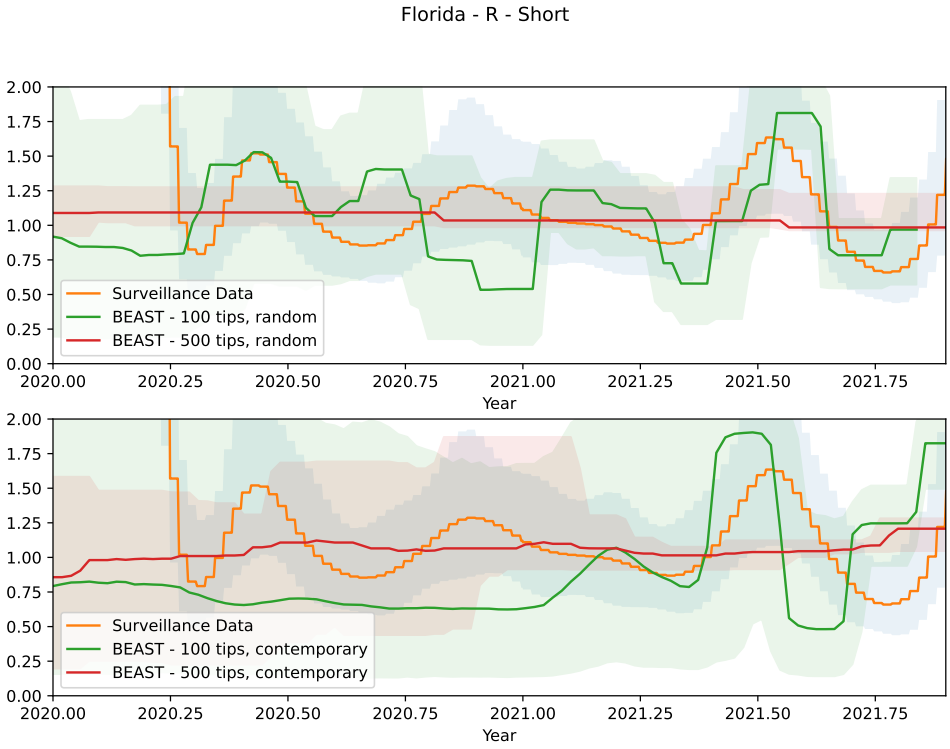

Figure 3.10: The posterior median and equal-tailed 95% credible interval of R for Florida given by BEAST. The top panel contains randomly sampled data, while the bottom contains the most recent available samples. The sampler was allowed to run as long as it VBSKY to analyze the Florida data. This is referred to as the short run in the text.

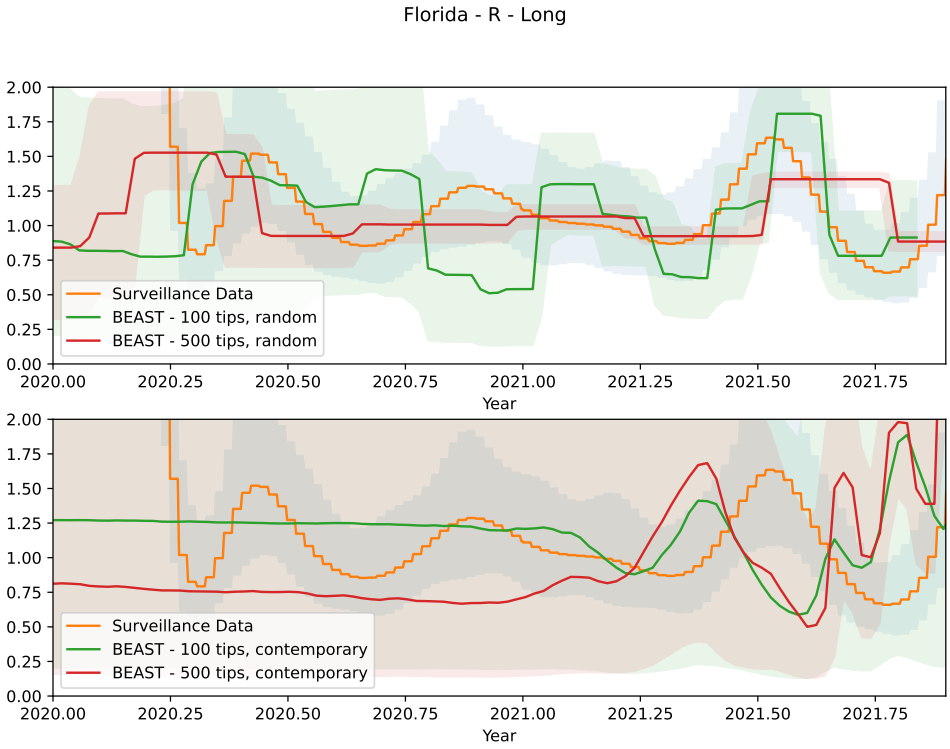

Figure 3.11: The posterior median and equal-tailed 95% credible interval of R for Florida given by BEAST. The sampler was allowed to run for 100 million steps or 24 hours to analyze the data. This is referred to as the long run in the text.

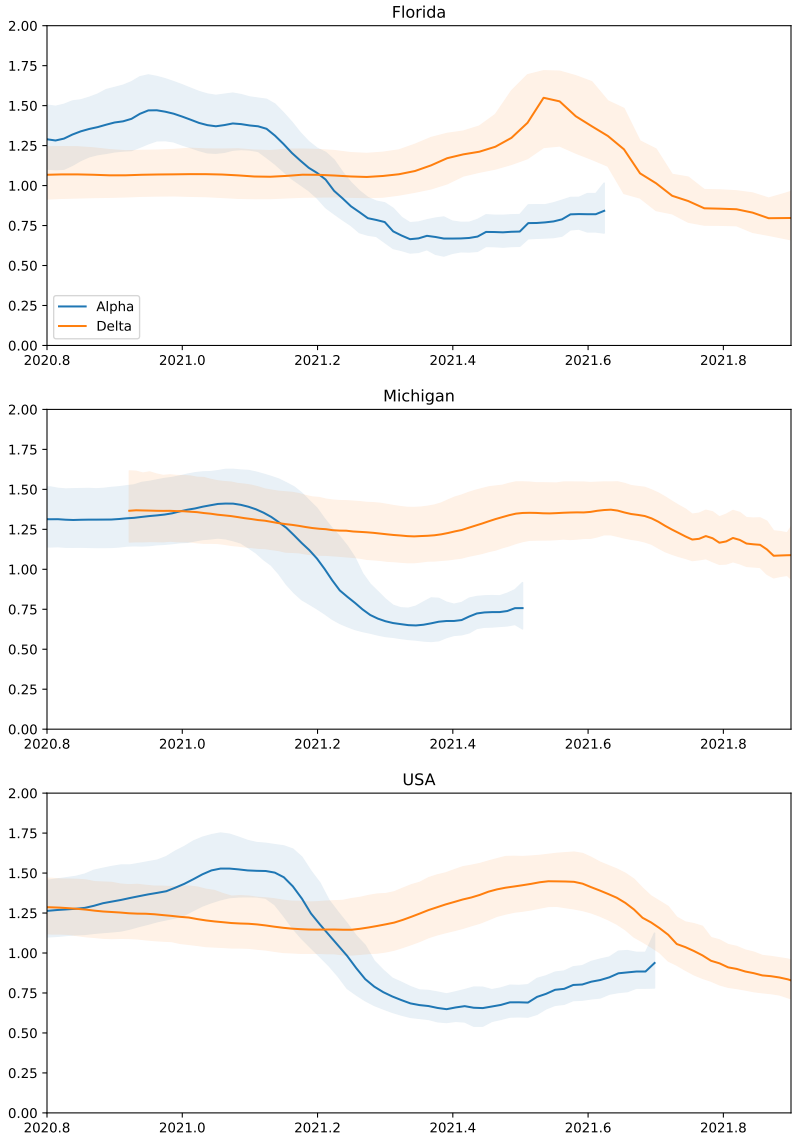

Figure 3.12: The posterior median and equal-tailed 95% credible interval of R for the Alpha and Delta variants.

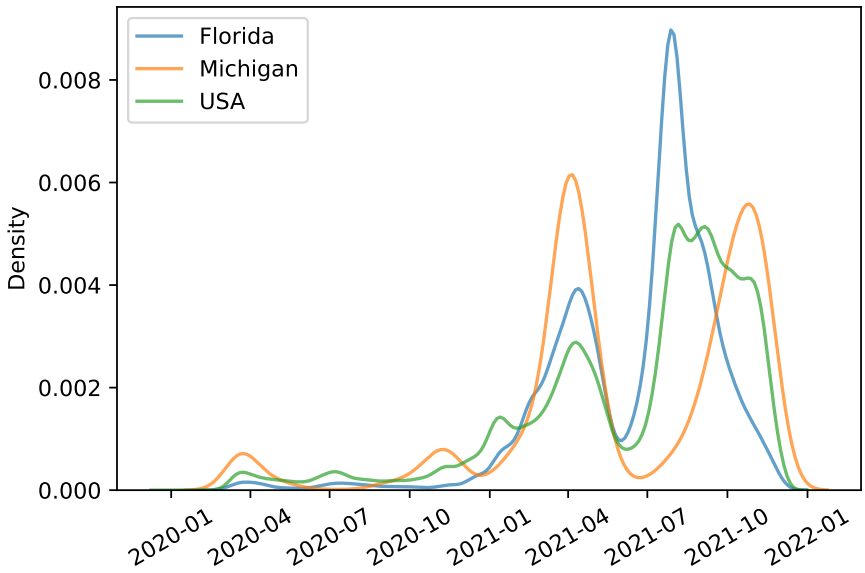

Figure 3.13: Distribution of sample times for Florida, Michigan, and the USA.

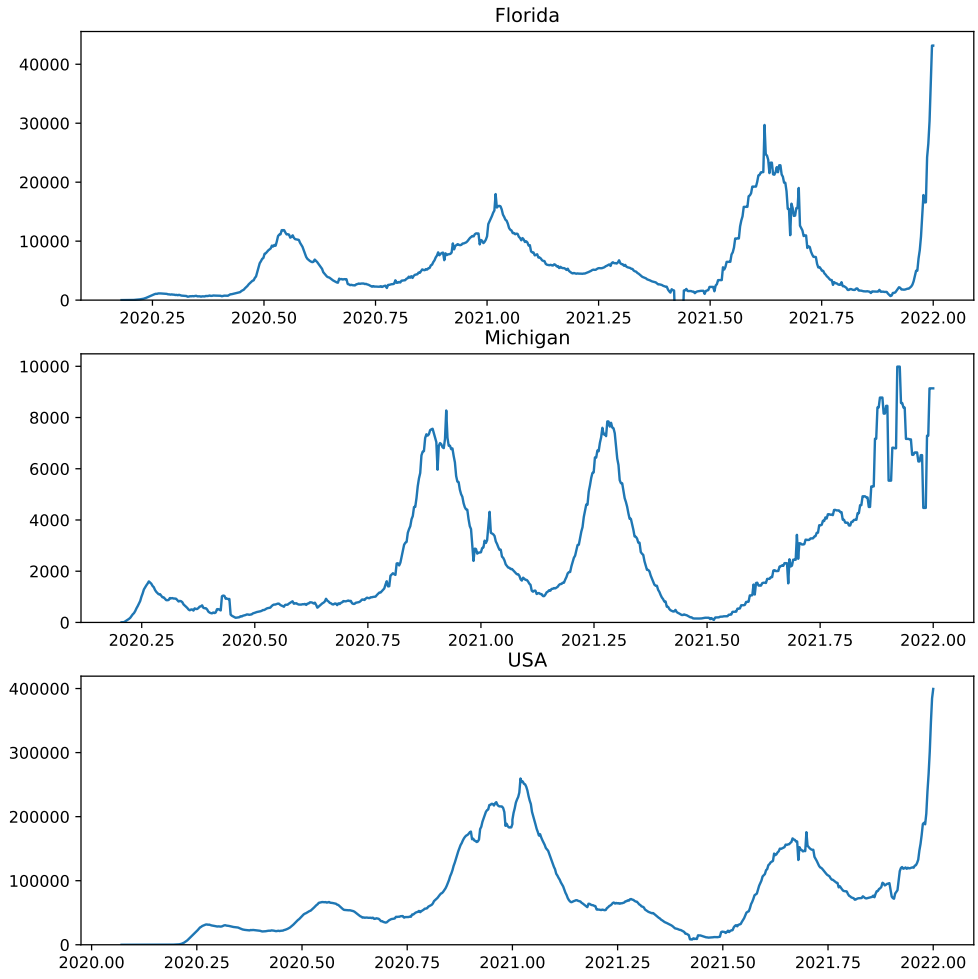

Figure 3.14: Daily new cases of COVID-19 over time for Florida, Michigan, and the USA.

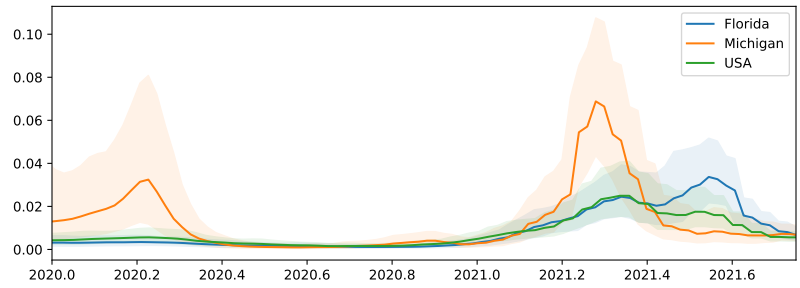

Figure 3.15: The posterior median and equal-tailed 95% credible interval of s for Florida, Michigan, and the USA using an uninformative smoothing prior.

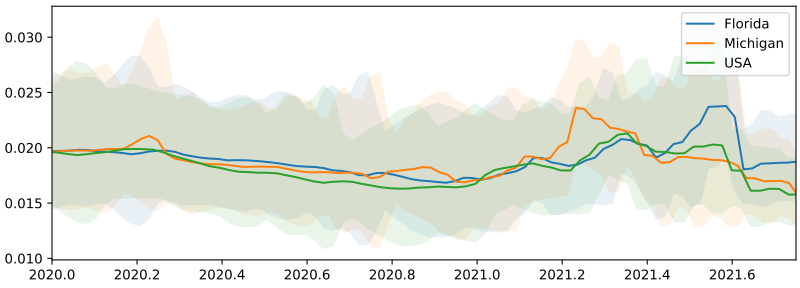

Figure 3.16: The posterior median and equal-tailed 95% credible interval of s for Florida, Michigan, and the USA using less smoothing.

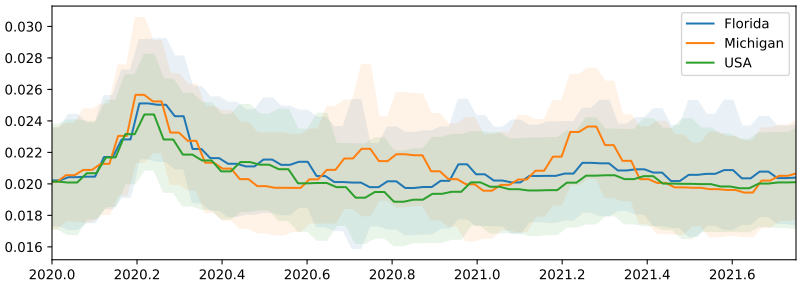

Figure 3.17: The posterior median and equal-tailed 95% credible interval of s for Florida, Michigan, and the USA using biased sampling.

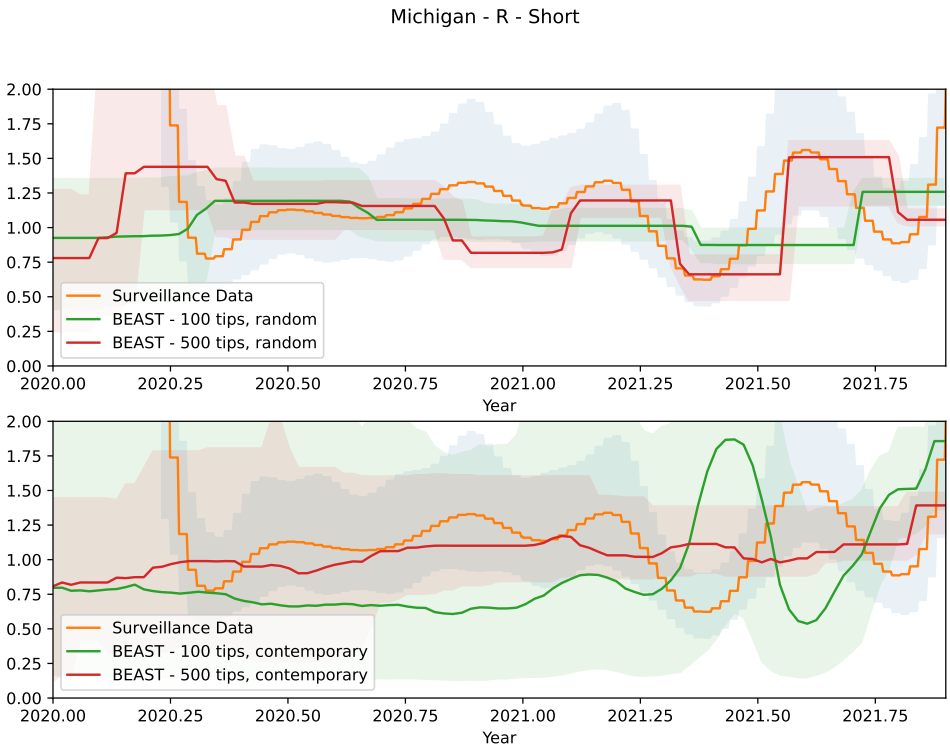

Figure 3.18: The posterior median and equal-tailed 95% credible interval of R for Michigan given by BEAST. The sampler was allowed to run as long as it VBSKY to analyze the Michigan data. This is referred to as the short run in the text.

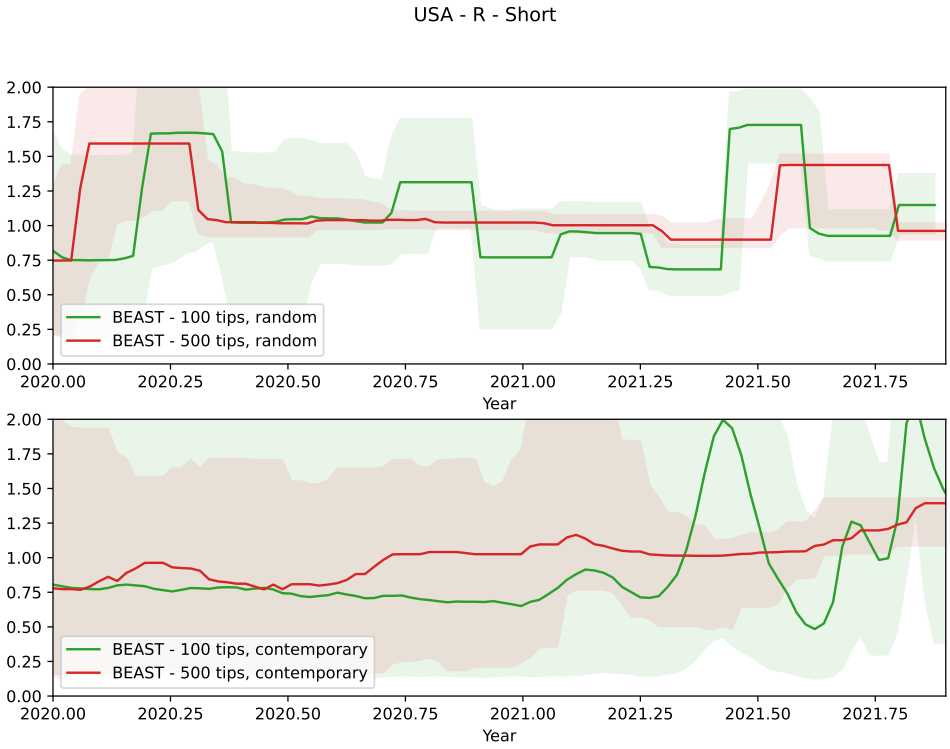

Figure 3.19: The posterior median and equal-tailed 95% credible interval of R for the USA given by BEAST. The sampler was allowed to run as long as VBSKY to analyze the USA data. This is referred to as the short run in the text.

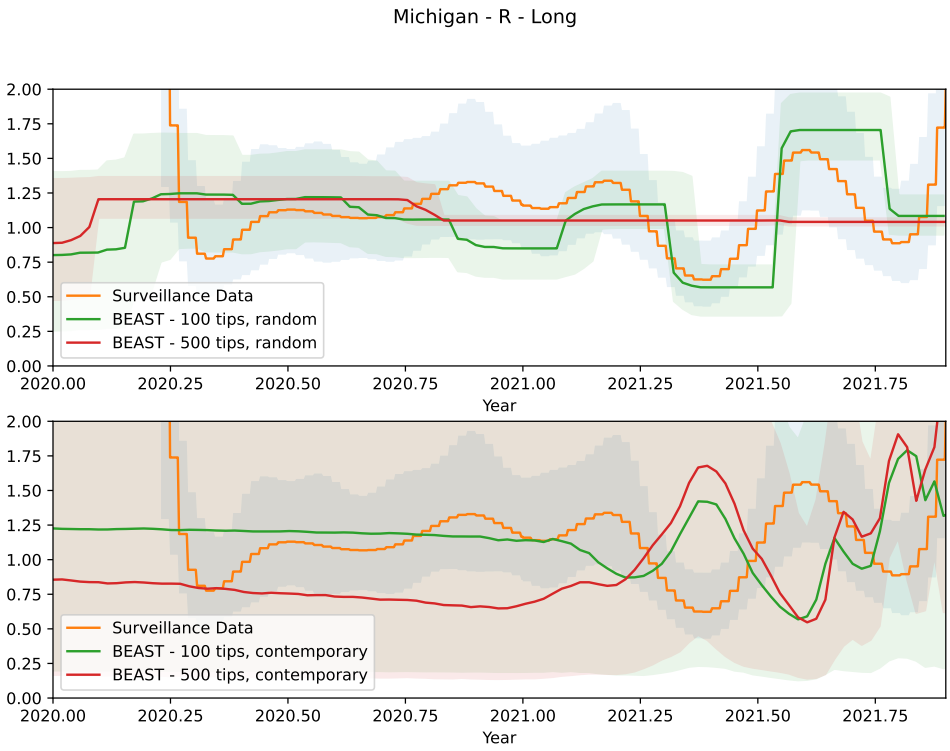

Figure 3.20: The posterior median and equal-tailed 95% credible interval of R for Michigan given by BEAST. The sampler was allowed to run for 100 million steps or 24 hours. This is referred to as the long run in the text.

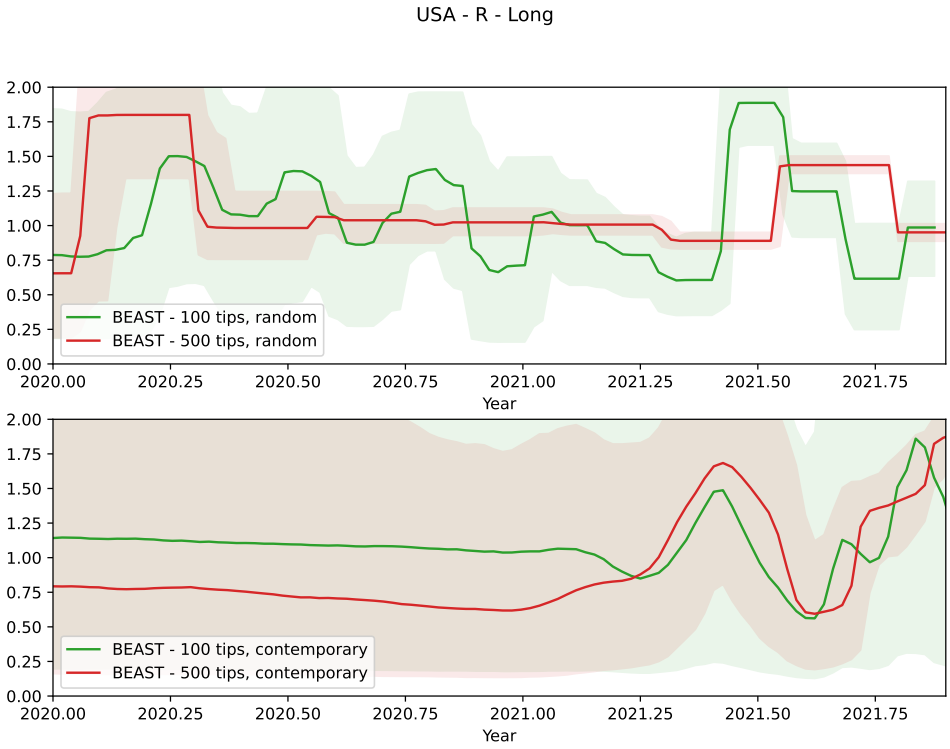

Figure 3.21: The posterior median and equal-tailed 95% credible interval of R for the U.S. given by BEAST. The sampler was allowed to run for 100 million steps or 24 hours. This is referred to as the long run in the text.

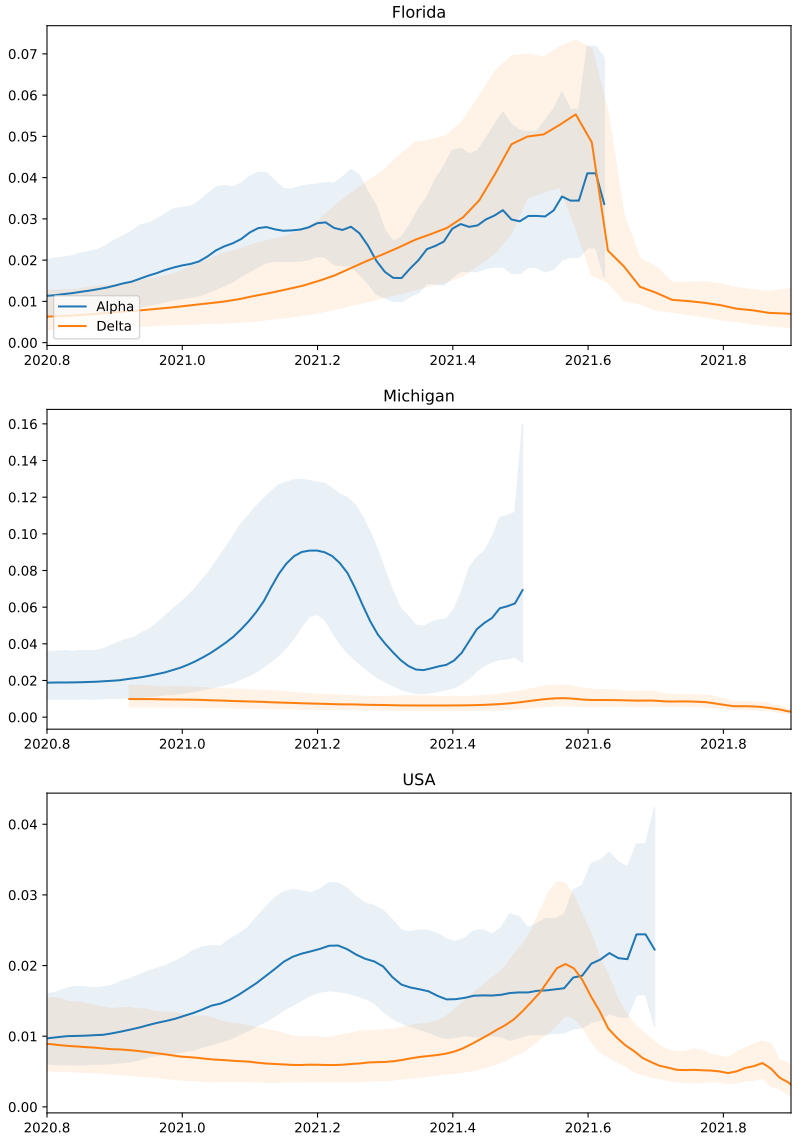

Figure 3.22: The posterior median and equal-tailed 95% credible interval of s for the Alpha and Delta variants.

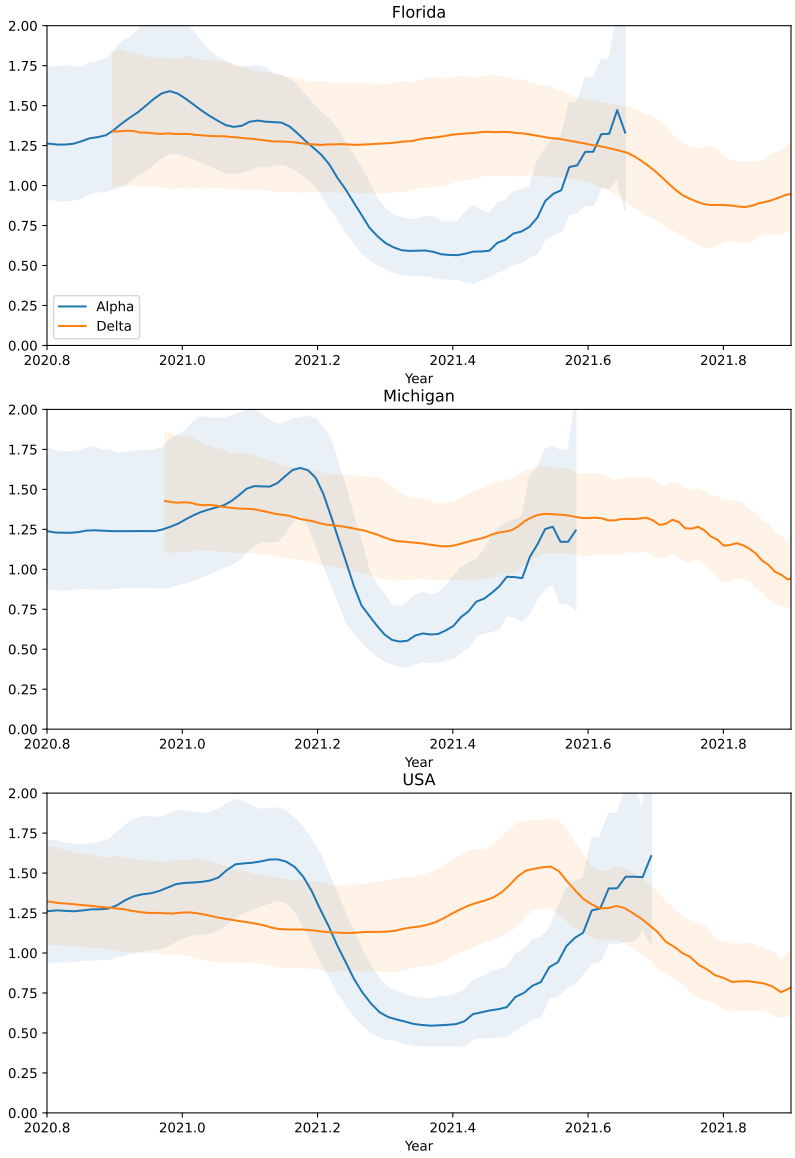

Figure 3.23: The posterior median and equal-tailed 95% credible interval of R for the Alpha and Delta variants.

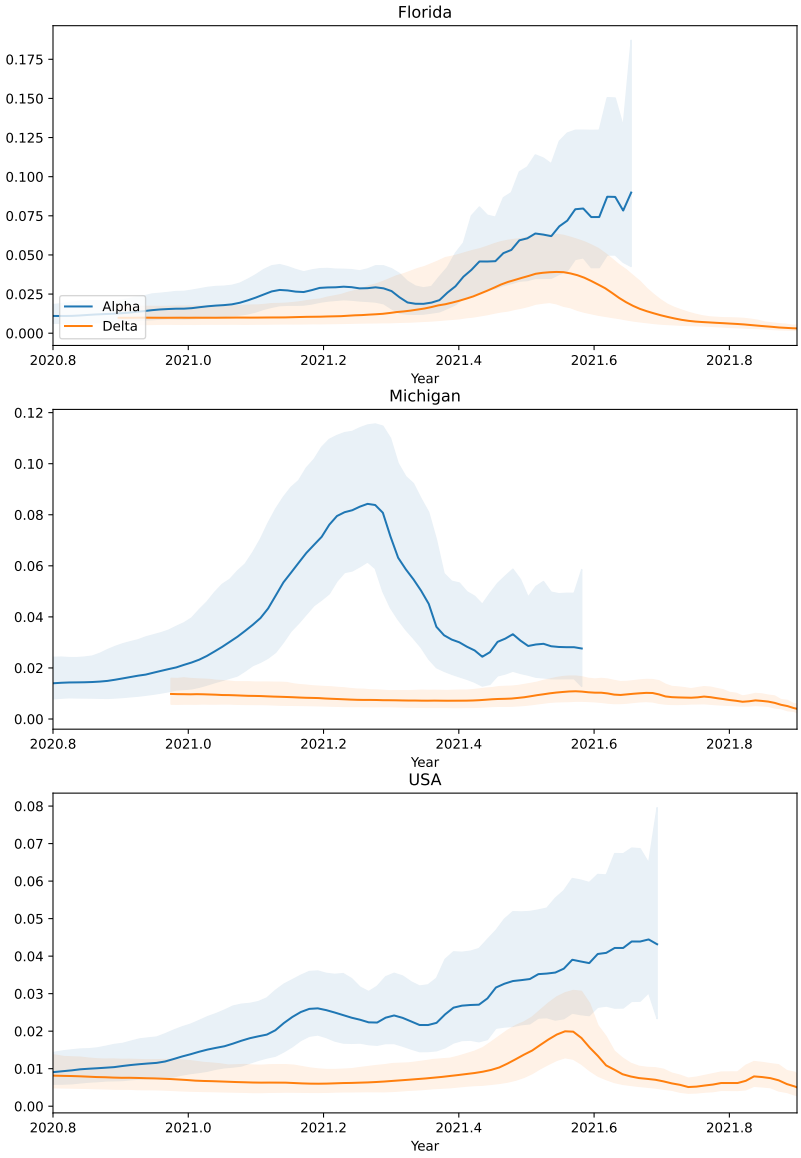

Figure 3.24: The posterior median and equal-tailed 95% credible interval of s for the Alpha and Delta variants.

CHAPTER 4

Inference of population size histories with the fused lasso

4.1 Introduction

The estimation of population size histories using genome variation data has been an important avenue of research in population genetics and has a range of applications. Inference of population size histories can help unravel questions concerning historical events such as the human migration out of Africa and is also paramount to understanding selection. A large class of methods developed with the goal of estimating population size histories are rooted in coalescent theory. Exact models of the coalescent with recombination are computationally intractable as they require integrating over the high dimensional graph structure known as the ancestral recombination graph (ARG; Hudson et al., 1990; Griffiths and Marjoram, 1997). Because of this, approximations to the ARG are necessary to analyze chromosome scale data. Wiuf and Hein (1999) provided the first approximation of the coalescent with recombination as a process along the genome instead of through time; each locus has a local genealogy, and alterations to successive genealogies only occur when a recombination occurs.

Thereafter McVean and Cardin (2005) extended this process with the sequentially Markov coalescent (SMC); as the name implies, after a recombination event the new local genealogy only depends on the genealogy at the previous locus. The Markovian structure of SMC renders the approximation computationally tractable, allowing for likelihood based inference of the population size history without having to integrate over a high-dimensional latent variable by way of hidden Markov models (HMMs). These so-called "coalescent HMMs" have become a fixture in the field of demographic inference (Dutheil et al., 2009; Li and Durbin, 2011; Schiffels and Durbin, 2014; Sheehan et al., 2013; Terhorst et al., 2017). In coalescent HMMs, generally the observed process is the data at each locus and the hidden process is the unobserved genealogy (or an approximation of the genealogy) hypothesized to have generated the observed data at each locus.

Perhaps the most popular coalescent HMM is the Pairwise Sequentially Markov Coalescent (PSMC; Li and Durbin, 2011) which has been used to great success to estimate population size histories in many species (Prado-Martinez et al., 2013; Prüfer et al., 2014; Ekblom et al., 2018). Using only diploid genotype information from a single individual (or, more generally, a pair of phased haplotypes), PSMC returns an estimate of the effective population size through time. Since its development a decade ago, several coalescent HMMs have been developed increasing the number of sequences that can be analyzed. Still PSMC remains a useful method particularly in small sample size situations and has even outperformed the current state of the art SMC based inference method SMC++ (Terhorst et al., 2017) in that regime (Patton et al., 2019). While PSMC is a standard model in demographic inference, it is not without its drawbacks. First, PSMC implements the basic SMC model when more accurate approximations, namely SMC' (Marjoram and Wall, 2006), exist. Second, PSMC cannot consistently recover accurate changes in the recent path. Third, the runtime of PSMC is quadratic in the number of states.

The first two issues were addressed by another method known as Multiple Sequentially Markovian Coalescent (MSMC; Schiffels and Durbin, 2014) and its successor MSMC2 (Malaspinas et al., 2016; Schiffels and Wang, 2020) which use transitions based on SMC'. MSMC can analyze more than a pair of haplotypes, but only tracks the first coalescence time between any pair of haplotypes in the sample. This adaptation gives MSMC the power to detect changes in the population size in the recent past. However, MSMC cannot estimate the the effective population size reliably in very ancient times as it only looks at the first coalescence time across the entire sample. MSMC2 shores up this weakness by modeling the coalescence time of each pair in the sequence with its own HMM. In doing so, MSMC2 is using coalescence time information from all pairs of sequences enabling it to estimate the effective population size both near the present and further back in the past. However, both MSMC and MSMC2 fail to address the final issue using PSMC. Additionally, including PSMC, none of the three aforementioned methods have a formal procedure for uncertainty quantification. Bootstrapping to estimate confidence intervals is possible, but without improving the run-time of these methods, any bootstrap procedure will be slow.

There are of course other limitations with coalescent HMMs not unique to the methods already mentioned. HMMs require the state space of the latent variables to be discrete, so in order to use HMMs for likelihood based inference, time must be discretized. Within each discrete time interval, population sized histories inferred from coalescent HMMs are often assumed to be constant. This discretization can lead to a biased and noisy reconstruction of the population size history. Neighboring epochs may have drastically different values of the effective population size which could suggest demographic events that may not have happened. This is especially the case in time periods where we may not be able to infer many coalescent rates such as when we are only looking at a pair of chromosomes in the recent past. Clearly the time discretization scheme can have large

impact on the estimates of the coalescence rate, yet in practice it is not obvious how to set the discretization and it is often not changed from the default settings in studies that use such methods (Mather et al., 2020).

To rectify these issues we introduce here a new coalescent HMM, QTND, that is both accurate and efficient. Our model QTND takes a pair of haplotypes as input, and tracks the genealogy of a pair of chromosomes as PSMC does. However, like MSMC2, our method can analyze multiple pairs of sequences simultaneously giving it more power to detect changes in population size in the recent past than PSMC. Similar to MSMC and MSMC2, our method uses the SMC' approximation but has a run time that is linear in the number of hidden states. This speedup gives QTND the capacity for fast uncertainty quantification using bootstrap. We also incorporate regularization via the fusion penalty encouraging smoothness between neighboring intervals of time. This should limit the effect of stochastic error from the discretization grid in the estimates of the effective population size. In the rest of the paper, we present the linear time EM algorithm for our coalescent HMM based on SMC' and the explain the regularization in detail.

4.2 Background

The first linear-time inference algorithm for SMC was derived by Harris et al. (2014). They showed that by augmenting the state space of a coalescent HMM to include information on the recombination and back-coalescence process, it was possible to reduce the running times of both the forward-backward and expectation-maximization (EM) algorithms to have linear time complexity in the number of hidden states. The standard algorithms for hidden Markov models has quadratic time complexity in the number of hidden states because of the need to integrate over all possible pairs of transitions between hidden states $1 \leq i, j \leq M$, where M is the overall number of hidden states.

Subsequently it was shown (Wilton et al., 2015; Terhorst et al., 2017) that the SMC' model of Marjoram and Wall is a more accurate approximation to the ancestral recombination graph compared to the original SMC model of McVean and Cardin. Palamara et al. (2018) derived a linear time forward-backward algorithm for SMC' by exploiting some symmetries and numerical properties of the SMC' transition matrix. However, they did not a provide corresponding linear-time EM algorithm, nor is it obvious how to do so using their approach.

Here we follow the probabilistic approach of Harris et al., which can be extended to SMC'. We adopt the notations from their paper throughout this chapter. Table 4.6 provides a summary of the notation used. For example, $\mathbb{P}(R_i, C_{>i} \mid T < i)$ denotes the probability that a recombination occurs in interval i, and the recombinant lineage "floats" back to an earlier interval, given that the existing time to most common recent ancestor (TMRCA) is beneath interval i.

The distinguishing feature of SMC' is that the probability $\mathbb{P}(C_{>j} \mid C_{>j-1})$ is not independent of the current TMCRA T: if $T > j$ then the recoalescence process happens at twice the rate as when $T < j$, but half of the resulting recombinations are silent. In order to account for this fact, we derive suitable expanded systems of recursions which are generalizations of the results of Harris et al..

4.3 The linear-time forward-backward algorithm

In this section we derive a forward-backard algorithm for coalescent HMMs under the SMC' model. As noted above, such an algorithm is already known, but the extension to a linear-time EM algorithm requires additional recursive quantities which were not considered by Palamara et al., whose main interest was in performing posterior decoding. Our algorithm is also differs from the original Harris et al. (2014) algorithm in that it recurses only on rescaled probability distributions, whose entries are always $\mathcal{O}(1)$ (cf. Bishop, 2006, §13.2.4). While mathematically uninteresting, this feature matters in applications, since the joint probability distributions considered by Harris et al.—e.g., $\mathbb{P}(x_{1:\ell}, T_\ell = i)$—will underflow when ℓ is large.

Throughout this section and the next, we use the hat notation to denote quantities which are conditioned on the data, and suppress this conditioning for notational convenience. For example, the forward probability at position ℓ is denoted

$$\hat{f}(T_\ell = i) \equiv \mathbb{P}(T_\ell = i \mid x_{1:\ell}),$$

where $x_{1:\ell}$ denotes the data observed at positions 1 through ℓ (inclusive). Similarly, the backward probability at position $\ell + 1$ is

$$\hat{b}(T_{\ell+1} = i) \equiv \frac{\mathbb{P}(x_{\ell+1:L} \mid T_{\ell+1} = i)}{\mathbb{P}(x_{\ell+1:L} \mid x_{1:\ell})}. \tag{4.1}$$

4.3.1 Forward recursion

Using the above conventions, we can write

$$\begin{aligned}\hat{f}(T_{\ell+1} = i) \propto \xi(x_{\ell+1} \mid T_{\ell+1}) \times \Big\{ & \\ \hat{f}(T_\ell = i) \left[\mathbb{P}(\bar{R} \mid T = i) + \mathbb{P}(R^s_{<i} \mid T = i) + \mathbb{P}(R_{\le i}, C_i \mid T = i)\right] & \\ + \hat{f}(R_{<i}, C_i, T_\ell < i) + \frac{1}{2}\hat{f}(T_\ell > i)\mathbb{P}(R_{\le i}, C_i \mid T > i) & \\ + \hat{f}(T_\ell = i)\mathbb{P}(R_{\le i}, C_i \mid T = i)\Big\} & \end{aligned} \tag{4.2}$$

which extends equation (7) of Harris et al. (2014) by partitioning the event $\{R_{\leq j}, C_j\}$ according to the state of T_ℓ, and accounting for the possibility of silent recombinations. A recursion for the term $\hat{f}(R_{<i}, C_i, T_\ell < i)$ is obtained by partitioning the event $\{T_\ell < i\} = \{T_\ell = i-1\} \cup \{T_\ell < i-1\}$:

$$\hat{f}(R_{<i}, C_i, T_\ell < i) = \hat{f}(R_{<i}, C_i, T_\ell = i-1) + \hat{f}(R_{<i}, C_i, T_\ell < i-1).$$

For the first term of the recursion,

$$\begin{aligned}\hat{f}(R_{<i}, C_i, T_\ell = i-1) &= \hat{f}(T_\ell = i-1)\mathbb{P}(R_{\leq i-1}, C_i \mid T_\ell = i-1). \\ &= \hat{f}(T_\ell = i-1)\mathbb{P}(R_{\leq i-1}, C_{>i-1} \mid T_\ell = i-1)\mathbb{P}(C_i \mid C_{>i-1}, T < i).\end{aligned}$$

For the second term, we have

$$\begin{aligned}\hat{f}(R_{<i}, C_i, T_\ell < i-1) &= \hat{f}(R_{<i-1}, C_i, T_\ell < i-1) \\ &= \hat{f}(R_{<i-1}, C_{>i-1}, T_\ell < i-1)\mathbb{P}(C_i \mid C_{>i-1}, T < i),\end{aligned}$$

which can be solved recursively since

$$\begin{aligned}&\hat{f}(R_{\leq i-1}, C_{>i-1}, T_\ell < i) = \\ &\qquad \hat{f}(R_{\leq i-1}, C_{>i-1}, T_\ell = i-1) + \hat{f}(R_{\leq i-2}, C_{>i-1}, T_\ell < i-1) \\ &\qquad\quad = \hat{f}(T_\ell = i-1)\mathbb{P}(R_{\leq i-1}, C_{>i-1} \mid T_\ell = i-1) + \\ &\qquad\qquad \hat{f}(R_{\leq i-2}, C_{>i-2}, T_\ell < i-1)\mathbb{P}(C_{>i-1} \mid C_{>i-2}, T < i-1),\end{aligned}$$

with base case $f(R_{\leq 0}, C_{>0}, T_\ell < 1 \mid x_{1:\ell}) = 0$.

4.3.2 The backward algorithm

To compute the rescaled backwards probabilities (equation 4.1 above), we follow the suggestion of Harris et al. (2014) and exploit the fact that the sequentially Markov coalescent can be arbitrarily oriented with respect to the direction of the sequence. Hence the probability of the data and the associated conditional probabilities are same regardless of whether we run the model in the standard (i.e., $5' \to 3'$) direction, or in the reverse.

Define the reversed sequence $\mathbf{x}^{\text{rev}}$ by the identity $x_i^{\text{rev}} = x_{L-i+1}$, and let $\hat{f}^{\text{rev}}(T_\ell = i)$ denote the

result of running the forward algorithm derived above on the reversed sequence. Then

$$\begin{aligned}\hat{b}(x_{\ell+1:L}) &= \frac{\mathbb{P}(T_\ell = i \mid x_{\ell+1:L})\mathbb{P}(x_{\ell+1:L})}{\mathbb{P}(T_\ell = i)} \\ &\propto \frac{\hat{f}^{\text{rev}}(T_{L-\ell+1} = i)}{\mathbb{P}(T_\ell = i)\mathbb{P}(x_\ell \mid T_\ell = i)},\end{aligned} \tag{4.3}$$

where the constant of proportionality in the last line is $\mathbb{P}(x_{\ell+1:L})$. To compute this constant, we note that

$$1 = \sum_{i=1}^{d} \hat{f}(T_\ell = i \mid x_{1:\ell})\hat{b}(T_\ell = i)$$

so that $\mathbb{P}(x_{\ell+1:L})$ equals the inner product of the rescaled forwards and backwards probabilities at position ℓ.

4.3.3 Comparison to the algorithm of Palamara et al.

Palamara et al. (2018) also derived a linear time posterior decoding algorithm for the SMC' model. Their algorithm works by modifying the standard recursions for the rescaled backwards pass (e.g., Bishop 2006 §13.2.4), noting that the the below-diagonal entries of the SMC' transition matrix are constant across rows, while the above-diagonal entries have a constant ratio between columns. We observed in practice that the approach of using the "reversed forwards" algorithm outlined in the preceding section tended to be more numerically stable during parameter inference. This is because the rescaled forward algorithm recurses on a probability distribution—namely, the filtering distribution $\mathbb{P}(T_\ell = i \mid x_{1:\ell})$—whereas the standard backward pass recurses on the *conditional* probability distribution shown in equation (4.1). Consequently, the entries of the forward/filtering distribution are always $\mathcal{O}(1)$, while no such guarantee exists for the backwards probabilities. We observed that numerical issues sometimes arose with the linear-time backward algorithm of Palamara et al. during parameter inference, where a numerical optimizer could pass in extreme parameter values, whereas the approach described above appeared to be more robust. Note that this issue was not encountered by Palamara et al., likely because they did not perform parameter inference in their paper.

4.4 Linear time EM algorithm for SMC'

In the preceding section, we derived a posterior decoding algorithm for the SMC' model which is linear in the number of hidden states. However, to use this model for parameter inference, we must also employ the EM algorithm. As already noted by Harris et al. (2014), computing the

expected complete log-likelihood for a hidden Markov model scales quadratically in the number of hidden states, even when the forward and backward algorithms have linear time complexity. This is because of the need to compute the expected number of transitions between all pairs of hidden states conditional on the data.

A linear-time EM algorithm analogous to the one provided by Harris et al. (2014) for the SMC model is also possible for SMC'. It is derived in a similar manner, by conditioning on the interval in which a recombination first occurred, and tracking the back-coalescence process as it proceeds from interval to interval. However, the possibility of silent recombination events in the SMC' model makes these calculations somewhat more involved. To account for this phenomenon, we introduce another auxiliary variable $T^*_{\ell+1}$ denoting the interval into which back-coalescence occurs conditional on a recombination at a given position. Under SMC', we have that $T^*_{\ell+1} = T_{\ell+1}$ with probability one conditional on $T^*_{\ell+1} \geq T_\ell$, but this probability is only one-half when $T^*_{\ell+1} < T_\ell$, since silent recombination, i.e. $T_{\ell+1} = T_\ell$, is equally probable.

We begin by decomposing $\mathbb{P}(R_i, T_{\ell+1} = k, T^*_{\ell+1} = h \mid T_\ell = j)$, the probability distribution of the TMRCA height across a recombination event, similar to equation (3) of Harris et al. This probability depends on the relationship between h, i, j, and k. Under the SMC' model there are a total of six cases to consider, and they are listed in Table 4.1. Compared to SMC, each decomposed term incorporates additional conditioning to account for the dependence of the coalescent probabilities on the position of the current TMRCA. For example, the first row of the table asserts that

$$\mathbb{P}(R_i, T_{\ell+1} = T^*_{\ell+1} = i \mid T_\ell = i) = \mathbb{P}(R_i, C_i \mid T = i);$$

in words, the probability that a recombination in interval i leads to no change in the TMRCA equals the probability of recombining and back-coalescing (either silently or visibly) in i. The most complicated entry of the table is the last row, $i < j < h = k$, where it is the case that a) a recombination occurs in interval i, which is strictly beneath the height of the current TRMCA j; b) the recombinant lineage floats backwards in time past j, and c) continues to float until it reaches interval k. Because the recombination is non-silent, resulting in a change in TMRCA between positions ℓ and $\ell + 1$, it must be the case that $h = k$.

Next, we rewrite the each of the terms in Table 4.1 according to the interval where they must occur. For example, from the last three lines of the table, it can be seen that the likelihood contains a term of the form $\mathbb{P}(R_i, C_{>i} \mid T > i)$ whenever $i < \min(h, j)$. These relations are collected in Table 4.2 (for recombination events), and Table 4.3 (for back-coalescence events).

Finally, to compute the expected log-likelihood, we need to compute the expected number of positions, conditional on the data, where each of the events identified in the left-hand columns of Tables 4.2 and 4.3 occurred. This is equal to the the sum, over all positions, of the posterior probability of each event. These expectations are collected in Tables 4.4 and 4.5, which correspond,

Condition	$\mathbb{P}(R_i, T_{\ell+1} = k, T^*_{\ell+1} = h \mid T_\ell = j)$
$h = i = j = k$	$\mathbb{P}(R_i, C_i \mid T = i)$
$h = i = k < j$ $h = i < j = k$	$\frac{1}{2}\mathbb{P}(R_i, C_i \mid T > i)$
$i = j < k = h$	$\mathbb{P}(R_i, C_{>i} \mid T = i) \times \prod_{m=i+1}^{k-1} \mathbb{P}(C_{>m} \mid C_{>m-1}, T < m)$ $\times \mathbb{P}(C_k \mid C_{>k-1}, T < k)$
$i < h < j = k$ $i < h = k < j$	$\mathbb{P}(R_i, C_{>i} \mid T > i) \times \prod_{m=i+1}^{k-1} \mathbb{P}(C_{>m} \mid C_{>m-1}, T > m)$ $\times \frac{1}{2}\mathbb{P}(C_k \mid C_{>k-1}, T > k)$
$i < h = j = k$	$\mathbb{P}(R_i, C_{>i} \mid T > i) \times \prod_{m=i+1}^{k-1} \mathbb{P}(C_{>m} \mid C_{>m-1}, T > m)$ $\times \mathbb{P}(C_k \mid C_{>k-1}, T = k)$
$i < j < h = k$	$\mathbb{P}(R_i, C_{>i} \mid T > i) \times \prod_{m=i+1}^{j-1} \mathbb{P}(C_{>m} \mid C_{>m-1}, T > m)$ $\times \mathbb{P}(C_{>j} \mid C_{>j-1}, T = j) \times \prod_{m=j+1}^{k-1} \mathbb{P}(C_{>m} \mid C_{>m-1}, T < m)$ $\times \mathbb{P}(C_k \mid C_{>k-1}, T > k)$

Table 4.1: The probability of recombining with to a new TMRCA conditional on the existing TMRCA. The event $T^*_{\ell+1} = h$ denotes that back-coalescence occurred in interval h.

Event: R_i and ...	Contribution to loglik
$i = T_\ell = T_{\ell+1} = T^*_{\ell+1}$	$\mathbb{P}(R_i, C_i \mid T = i)$
$i = T_{\ell+1} = T^*_{\ell+1} < T_\ell$ $i = T^*_{\ell+1} < T_\ell = T_{\ell+1}$	$\frac{1}{2}\mathbb{P}(R_i, C_i \mid T > i)$
$i = T_\ell < T_{\ell+1} = T^*_{\ell+1}$	$\mathbb{P}(R_i, C_{>i} \mid T = i)$
$i < \min(T_\ell, T^*_{\ell+1})$	$\mathbb{P}(R_i, C_{>i} \mid T > i)$

Table 4.2: The probability of recombining with to a new TMRCA conditional on the existing TMRCA. The event $T^*_{\ell+1} = h$ denotes that back-coalescence occurred in interval h.

Event: $R_{<i}$ and ...	Contribution to loglik
$i < T^*_{\ell+i} < T_\ell$	$\mathbb{P}(C_{>i} \mid C_{>i-1}, T > i)$
$i = T_\ell < T^*_{\ell+1}$	$\mathbb{P}(C_{>i} \mid C_{>i-1}, T = i)$
$T_\ell < i < T^*_{\ell+1}$	$\mathbb{P}(C_{>i} \mid C_{>i-1}, T < i)$
$i = T^*_{\ell+1} < T_\ell$	$\mathbb{P}(C_i \mid C_{>i-1}, T > i)$
$i = T^*_{\ell+1} = T_\ell$	$\mathbb{P}(C_i \mid C_{>i-1}, T = i)$
$T_\ell < i = T^*_{\ell+1}$	$\mathbb{P}(C_i \mid C_{>i-1}, T < i)$

Table 4.3: The probability of recombining with to a new TMRCA conditional on the existing TMRCA. The event $T^*_{\ell+1} = h$ denotes that back-coalescence occurred in interval h.

respectively, to the interval where a recombination event occurred; and the intervals beneath which a recombination occurred, and the recombinant lineage either floats past the interval under consideration, or coalesces into it. (Note that there is an implicit summation over ℓ in each entry of the table.)

There is one event which is slightly more difficult to compute: when $i < \min(T_\ell, T^*_{\ell+1})$, it is not true that the data $x_{\ell+1:L}$ are conditionally independent of $x_{1:\ell}$ conditional only on $i < T_\ell$. Hence the argument used to derive the other rows of both tables does not work in this case. To deal with this case, we directly compute the probability of the event $\{R_i, T_\ell > i, T_{\ell+1} > i\}$, conditioned on the data, using another recursion. First, we write

$$\mathbb{P}(R_i, T_\ell > i, T_{\ell+1} > i) = \mathbb{P}(T_\ell > i)\mathbb{P}(R_i, C_{>i}|T > i)\mathbb{P}(T_{\ell+1} > i \mid T_\ell > i, C_{>i}),$$

where to simplify notation we omit the conditioning on the data. Then we consider a recursion for the last term:

$$\begin{aligned}\mathbb{P}(T_{\ell+1} > i, T_\ell > i \mid C_{>i}) = \\ \mathbb{P}(C_{>i+1} \mid C_{>i}, T > i)\mathbb{P}(T_{\ell+1} > i+1, T_\ell > i+1 \mid C_{>i+1}) + \\ \mathbb{P}(T_\ell = i+1, T_{\ell+1} > i+1 \mid C_{>i}) + \mathbb{P}(T_\ell > i+1, T_{\ell+1} = i+1 \mid C_{>i}).\end{aligned}$$

For the first term in the last line, we have

$$\mathbb{P}(T_\ell = i, T_{\ell+1} > i \mid C_{>i-1}) = \hat{f}(T_\ell = i)\hat{b}(T_{\ell+1} > i)\mathbb{P}(C_{>i} \mid C_{>i-1}, T = i)$$

and for the second,

$$\hat{f}(T_\ell > i+1)\mathbb{P}(C_{i+1} \mid C_{>i}, T > i+1)\hat{b}(T_{\ell+1} = i+1).$$

Event: R_i and ...	Expected number of positions: $\sum_\ell \cdots$
$i = T_\ell = T_{\ell+1} = T^*_{\ell+1}$	$\hat{f}(T_\ell = i \mid x_{1:\ell})\mathbb{P}(R_i, C_i \mid T = i)\hat{b}(x_{\ell+1:L} \mid T_{\ell+1} = i)$
$i = T_{\ell+1} = T^*_{\ell+1} < T_\ell$	$\frac{1}{2}\hat{f}(T_\ell > i \mid x_{1:\ell})\mathbb{P}(R_i, C_i \mid T > i)\hat{b}(x_{\ell+1:L} \mid T_{\ell+1} = i)$
$i = T^*_{\ell+1} < T_\ell = T_{\ell+1}$	$\frac{1}{2}\mathbb{P}(R_i, C_i \mid T > i)\sum_{j>i}\hat{f}(T_\ell = j \mid x_{1:\ell})\hat{b}(x_{\ell+1:L} \mid T_{\ell+1} = j)$
$i = T_\ell < T_{\ell+1} = T^*_{\ell+1}$	$\hat{f}(T_\ell = i \mid x_{1:\ell})\mathbb{P}(R_i, C_{>i} \mid T = i)\hat{b}(x_{\ell+1:L} \mid T_{\ell+1} > i)$
$i < \min(T_\ell, T^*_{\ell+1})$	See text

Table 4.4: The probability of recombining with to a new TMRCA conditional on the existing TMRCA. The event $T^*_{\ell+1} = h$ denotes that back-coalescence occurred in interval h.

Event: $R_{<i}$ and ...	Expected number of positions: $\sum_\ell \cdots$
$i < \min(T^*_{\ell+i}, T_\ell)$	See text
$i = T_\ell < T^*_{\ell+1}$	$\hat{f}(T_\ell = i \mid x_{1:\ell})\mathbb{P}(R_{<i}, C_{>i} \mid T = i)\hat{b}(x_{\ell+1:L} \mid T_{\ell+1} > i)$
$T_\ell < i < T^*_{\ell+1}$	$\hat{f}(R_{\leq i-1}, C_{>i-1}, T_\ell < i \mid x_{1:\ell})\mathbb{P}(C_{>i} \mid C_{>i-1}, T < i)\hat{b}(x_{\ell+1:L} \mid T_{\ell+1} > i)$
$i = T^*_{\ell+1} < T_\ell$	$\mathbb{P}(R_{<i}, C_i \mid T > i) \times \Big[\hat{f}(T_\ell > i \mid x_{1:\ell})\hat{b}(x_{\ell+1:L} \mid T_{\ell+1} = i) + \sum_{j>i}\hat{f}(T_\ell = j \mid x_{1:\ell})\hat{b}(x_{\ell+1:L} \mid T_{\ell+1} = j)\Big]$
$i = T^*_{\ell+1} = T_\ell$	$\hat{f}(T_\ell = i \mid x_{1:\ell})\mathbb{P}(R_{<i}, C_i \mid T = i)\hat{b}(x_{\ell+1:L} \mid T_{\ell+1} = i)$
$T_\ell < i = T^*_{\ell+1}$	$\hat{f}(R_{\leq i-1}, C_{>i-1}, T_\ell < i \mid x_{1:\ell})\mathbb{P}(C_i \mid C_{>i-1}, T < i)\hat{b}(x_{\ell+1:L} \mid T_{\ell+1} = i)$

Table 4.5: The probability of recombining with to a new TMRCA conditional on the existing TMRCA. The event $T^*_{\ell+1} = h$ denotes that back-coalescence occurred in interval h.

4.5 Fused Lasso

In this section, we explain in detail how take advantage of our linear time EM algorithm to incorporate regularization. The parameters of our model are the scaled mutation rate θ, recombination rate ρ, and coalescence rates $\boldsymbol{c} = (c_1, \ldots, c_T)$. The coalescence rates are naturally ordered by the discretization interval they belong to. It is reasonable to assume in the absence of a population expansion or bottleneck, coalescence rates in neighboring intervals will be more similar than coalescence rates in intervals that are further apart. The fusion regularization term penalizes differences between successive coalescence rates which can reduce the effect of the discretization scheme on the estimation of the population size history. Formally, if we take $Q(\boldsymbol{c}, \theta, \rho)$ to be the usual function we maximize in the M step of the EM algorithm, we optimize over the following function

$$\boldsymbol{c}^* = \underset{\boldsymbol{c} \in \mathbb{R}^T_{\geq 0}}{\arg\min}\, Q(\boldsymbol{c}) + \alpha \sum_{i=1}^{T-1} |c_{i+1} - c_i|,$$

where α is the regularization parameter. The fusion penalty has the added benefit that changes in the population size are less likely to be due to overfitting whereas in PSMC and other coalescent HMMs it can sometimes be difficult to distinguish between real changes in population size and noise from the model.

We select the best value for α using a cross-validation scheme. We split the data into training and test sets and find the optimal parameters $\boldsymbol{c}_i, \theta_i, \rho_i$ using the training set for each value of the regularization parameter α_i we would like to try. We then score each α_i by the likelihood of model using estimated parameters and the test set, and select the α_i that provides the largest likelihood. Regularization has been used previously for demographic inference in SMC++ (Terhorst et al., 2017) as well as in an allele frequency method (DeWitt et al., 2021). However in these methods regularization requires manual tuning, and there is no built in procedure to find the optimal tuning parameter. Because the EM algorithm for QTND is linear in the number of hidden states, we can try several values of the regularization parameter relatively quickly.

4.6 Appendix

4.6.1 Transition Probabilities

In this section we write out how to recover the probabilities needed to compute transition events. We have M total hidden states where each state is the discretized interval within which coalescence occurs. State i is defined as the interval of coalescence $[t_i, t_{i+1})$. We first begin by defining a Markov chain of the states of a lineage of two loci. The four states are no recombination, floating,

Notation	Meaning
R_i	Recombination occurs in interval i
$R_{<i}$	Recombination in an interval beneath i
$R^s_{<i}$	Silent recombination beneath interval i
C_i	Back-coalescence in interval i
$C_{>i}$	No back coalescence until (at least) interval i
$T = i, T < i, T > i$	The height of the existing TMRCA is in/beneath/above interval i
$\hat{f}(T_\ell = i), \hat{f}(T_\ell > i)$	The value of the (cumulated) rescaled forward recursion at position ℓ.
$\hat{b}(T_{\ell+1} = i), \hat{b}(T_{\ell+1} > i)$	The value of the (cumulated) rescaled backward recursion at position $\ell + 1$.

Table 4.6: Table of notation.

visible recoalescence, silent recoalescence in that order. The rate matrix of this Markov chain is then given by

$$Q_i = \begin{pmatrix} -\rho & \rho & 0 & 0 \\ 0 & -2c_i & c_i & c_i \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \tag{4.4}$$

where ρ is the recombination rate and c_i is the coalescence rate in the interval $[t_i, t_{i+1})$. We write

$$W_i = \int_{t_i}^{t_{i+1}} e^{(t-t_i)Q_i} f(t \mid t \in [t_i, t_{i+1}))dt$$
$$Z_i = \int_{t_i}^{t_{i+1}} e^{(t-t_i)Q_i} f(t \mid t \in [t_i, t_{i+1}))e^{-(t_{i+1}-t_i)c_i}dt,$$

where $f(t \mid t \in [t_i, t_{i+1}))$ is the conditional probability of coalescing at time t given coalescence occurs in $[t_i, t_{i+1})$. These integrals can be computed in closed form because the eigendecomposition is known in closed form. We also define

$$E_i = e^{(t_{j+1}-t_j)Q_i}$$
$$H_i = \prod_{j=1}^{i-1} E_j$$
$$G_i = H_i W_i.$$

Note that each W_i, Z_i, E_i, H_i, G_i are all 4×4 matrices.

The various components of the transition probabilities described in the main text are as follows:

$$
\begin{aligned}
\mathbb{P}(C_{>i} \mid C_{>i-1}, T < i) &= e^{-(t_i - t_{i-1})/c_i} \\
\mathbb{P}(C_i \mid C_{>i-1}, T < i) &= 1 - e^{-(t_i - t_{i-1})/c_i} \\
\mathbb{P}(C_i \mid C_{>i-1}, T = i) &= W_{i,1,2} + W_{i,1,3} + W_{i,1,1} - Z_{i,1,1} \\
\mathbb{P}(C_{>i} \mid C_{>i-1}, T = i) &= Z_{i,1,1} \\
\mathbb{P}(C_i, V \mid C_{>i-1}, T > i) &= E_{i,1,2} \\
\mathbb{P}(C_i, S \mid C_{>i-1}, T > i) &= E_{i,1,3} \\
\mathbb{P}(C_{>i} \mid C_{>i-1}, T > i) &= E_{i,1,1} \\
\mathbb{P}(\bar{R} \mid T = i) &= G_{i,0,0} \\
\mathbb{P}(R_i^s \mid T = i) &= G_{i,0,3} \\
\mathbb{P}(R_{<i}^s \mid T = i) &= \mathbb{P}(R_{<i}^s \mid T > i) = H_{i,0,3} \\
\mathbb{P}(R_i, C_i \mid T = i) &= H_{i,0,0}(W_{i,0,2} + W_{i,0,3} + W_{i,0,1} - Z_{i,0,1}) \\
\mathbb{P}(R_i, C_{>i} \mid T = i) &= H_{i,0,0} Z_{i,0,1} \\
\mathbb{P}(R_{<i}, C_{>i} \mid T = i) &= H_{i,0,1} Z_{i,1,1} \\
\mathbb{P}(R_{<i}, C_i \mid T = i) &= H_{i,0,1} \mathbb{P}(C_i \mid C_{>i-1}, T = i) \\
\mathbb{P}(R_i, C_i, V \mid T > i) &= H_{i,0,0} E_{i,0,2} \\
\mathbb{P}(R_i, C_i, S \mid T > i) &= H_{i,0,0} E_{i,0,3} \\
\mathbb{P}(R_i, C_{>i} \mid T > i) &= H_{i,0,0} E_{i,0,1} \\
\mathbb{P}(R_{<i}, C_i, V \mid T > i) &= H_{i,0,1} E_{i,1,2} \\
\mathbb{P}(R_{<i}, C_i, S \mid T > i) &= H_{i,0,1} E_{i,1,3} \\
\mathbb{P}(R_{<i}, C_{>i} \mid T > i) &= H_{i,0,1} E_{i,1,1}.
\end{aligned}
$$

Now we turn to the stationary probabilities:

$$
\begin{aligned}
\mathbb{P}(C_{>i}) &= e^{-\sum_{j=1}^{i} c_i (t_{i+1} - t_i)} \\
\mathbb{P}C_i &= \mathbb{P}(C_{>i}) - \mathbb{P}(C_{>i-1}) \\
\mathbb{P}(C_{<i}) &= 1 - \mathbb{P}C_i - \mathbb{P}(C_{>i}).
\end{aligned}
$$

CHAPTER 5

Conclusion

In this thesis, we discussed a handful of challenges in analyzing genetic variation data and presented methods that successfully overcome these problems. The challenges discussed are rooted in analyzing genetic data at scale. For analyzing chromosome data, the length of the data renders exact inference under the coalescent with recombination impossible. For viral data, the length of the sequences are much shorter, but there are no current methods that can make use of the large amount of pathogen sequences available. In the rest of this chapter, we summarize and contextualize the contributions of this thesis towards addressing these challenges. We also look towards the future and discuss extensions to the work in this thesis as well as other important, related future directions of research.

Sequentially Markov coalescent

In Chapter 2, we first introduced XSMC, a new method for estimating population size histories. XSMC follows a line of methods based on the sequentially Markov coalescent. SMC is an approximation the coalescent with recombination that balances biological realism with computational tractability. By modeling the coalescent with recombination as a process along the sequence, the approximation naturally lends itself to a hidden Markov model framework. To date there are several methods that have used the SMC based HMM framework in order to infer population size histories (Dutheil et al., 2009; Li and Durbin, 2011; Schiffels and Durbin, 2014; Sheehan et al., 2013; Terhorst et al., 2017).

Unlike previous SMC based methods, XSMC removes the need to discretize genealogies and allows inference of the latent sequence of trees in their natural continuous state. The key insight behind XSMC is that SMC can almost be cast as a change point detection model. To cast our method as a change point model, instead of modeling transitions based on the SMC approximation, we instead use transitions based on the renewal approximation. This permits XSMC to use the already well-established machinery used in change point detection problems.

Rather than fixing a parametric function class, XSMC uses a nonparametric estimator to estimate the effective population size history. Because XSMC uses a nonparametric approach, it does not provide estimates of the effective population size where a coalescence event is not observed. This means that XSMC will only provide estimates after the first inferred coalescence event. Methods like PSMC which are parametric will return an estimate of the effective population even when coalescence events may not have been observed. Another advantage of the nonparametric approach is that it reduces the burden of user to make optimal modeling choices such as selecting the best time discretization for their set of data. Thus, in addition to being accurate and fast, XSMC is also easy to use. However, in general nonparametric approaches require more data than their parametric counterpart. While empirically our survival analysis estimator works well, future work could be done to show what happens when the sequence length is shorter and we can observe less coalescence events.

We showed via simulation how under a range scenarios, XSMC can produce better estimates of the effective population size under a range of different scenarios when compared to other methods. However, this is all under the assumption that the data is from a single, panmictic population. In cases where we have a structured population, it may be prudent to actually allow for correlations in neighboring IBD segments. Extensions to XSMC to allow for this are possible, but computational performance would suffer. Further research should be done to see how robust the method is to this assumption, and how to exactly balance computational performance with accuracy in this setting.

Beyond estimating population size histories, another potential application of our method is downstream phasing and imputation. Many methods for phasing and imputation (Li and Abecasis, 2006; Das et al., 2016; Loh et al., 2016; Browning et al., 2018; Rubinacci et al., 2020) are based on the Li and Stephens (2003) haplotype copying model. We presented some results that demonstrate that XSMC may be more favorable for this type of application than LS. These results are preliminary as we used an approximate metric to substitute phasing and imputation accuracy. So while our initial results are promising, it will be important in the future to actually implement phasing and imputation pipelines using XSMC and compare against the methods that use LS to perform those procedures.

One assumption our method makes is the infinite sites model as it renders likelihood computation easier. Many of the results of our method rely on the fact that the number of positions that differ at a particular locus conditioned on the TMRCA at that locus follows a Poisson distribution. Admittedly while this assumption is standard in population genetics, it is not always grounded in biological realism. It remains an important task to understand how robust our method is to violations of this assumption and under what scenarios we would need to use a more realistic model of mutations. Under a (biallelic) recurrent mutation model, the number of mutations observed at each nonrecombining locus no longer has a Poisson distribution, since an even number of muta-

tions at a given site will result in that site possessing the ancestral state. This poses challenges for XSMC, which gains speed by exploiting the conjugate nature of the gamma and Poisson distributions. New methods to accommodate those settings would have to be developed. Creating models of demographic inference with more complex models of mutation while maintaining scalability is a difficult but nonetheless important task.

Another drawback of XSMC and many other SMC based methods is the assumption of constant recombination and mutation rates across the genome. Even though these methods are robust and can produce accurate estimates of the population size history when the recombination rate is misspecified, it would still be prudent to extend these methods to allow for variable recombination and mutation rates. We also know that the processes that generate real data do not have constant rates of mutation and recombination. Extending XSMC to allow for position specific recombination and mutation rates is straightforward. Doing so could improve our ability to accurately infer population size histories. Additionally we could use this extension to learn spatial or motif specific variation in mutation and recombination rates which has been studied in the past (Harris, 2015; Carlson et al., 2018). This in turn would increase our knowledge about the human genome and possibly provide further insight into the evolutionary process of humans.

In addition to XSMC, we presented the theory behind a new coalescent HMM in Chapter 4. Previously Harris et al. (2014) provided an EM algorithm for their coalescent HMM based on SMC that has linear time complexity on the number of hidden states. We extended this result to provide a linear time EM algorithm for SMC'. This new coalescent HMM which we call QTND tackles many of the issues found in PSMC and other coalescent HMMs. We also equipped QTND with a fusion penalty to encourage smoothness among estimates for the effective population size in neighboring epochs. The fast implementation of the model allows for a quick automated cross-validation procedure helping users select the best regularization parameter for the most accurate inference.

Like XSMC, QTND does not account for population structure in any way. As such analyzing sequences generated in the presence of structure or migration would require these methods to be extended. In particular, augmenting the coalescent HMM of QTND to account for more complex demographic models is important as we will often want to study populations beyond the basic single, panmictic setting. Lastly, we lack a strong theoretical understanding of how coalescent HMMs behave. Many of the insights we insights we have learned about coalescent HMMs is through actually using the methods. For example, we know that PSMC is robust under misspecification of the recombination rate because of experiments. Answering basic theoretical questions about coalescent HMMs could tell how to better use the methods we already have, why some work better, and how to improve them in the future.

Viral Phylogenetics

In Chapter 3, we discussed an important direction of research in Bayesian phylogenetics. Because the state space of tree topologies explodes with the number of taxa, nearly all Bayesian phylogenetic methods rely on MCMC algorithms as they circumvent the need to calculate the marginal likelihood of the data integrating out these trees. These MCMC algorithms are computationally expensive and difficult to accelerate. Thus, the rate at which we are able to gather new genetic variation data of viruses far outpaces our ability to analyze the influx of new data.

VBSKY is a method that seeks to bridge that gap by combining recent work in scalable Bayesian inference, differentiable programming, and phylogenetic analysis that allows for fast phylodynamic inference of thousands of sequences. VBKSY transforms the problem of estimating the posterior from a difficult and laborious sampling procedure to a quicker optimization procedure using variational inference. To do this, VBSKY relies on heuristics to estimate the posterior of the epidemiological parameters without needing to perform posterior inference on the discrete topologies while still accounting for phylogenetic inference.

As stated previously, our method treats the overall phylogeny as a nuisance parameter. While phylogeny estimation is not the primary goal of inference using skyline models, a logical extension of this method would be to coestimate tree topologies along with the branch lengths and epidemiological parameters. There has already been progress in developing variational inference procedures that can estimate entire phylogenies including the topology (Zhang and Matsen IV, 2018, 2019). Additionally Karcher et al. (2021) have taken a divide and conquer approach to reconstruct the original supertree using a variational inference framework. Integrating this line of research with VBSKY would provide a way to perform tree inference along with the skyline parameters. This has potential to not only improve accuracy but would also obviate the need to rely on heuristics to approximately sample tree topologies.

While variational procedures are gaining popularity within the field of viral phylogenetics, using ideas from machine learning like variational inference is a relatively unexplored area in viral phylogenetics. For VBSKY, we can already incorporate ideas like normalizing flows (Rezende and Mohamed, 2015) to allow for correlation between parameters and potentially get more accurate posteriors and better uncertainty quantification. Further, it is clear that if we are to fully make use of the ever increasing available sequence data, new scalable methods are needed. To that end, discovering new and useful ways to integrate scalable ideas from machine learning into viral phylogenetics will be an important line of research.

Instead of using a divide and conquer approach or borrowing ideas from machine learning as we do in VBSKY, another direction to improve inference is to analyze data in real-time. As discussed earlier, there has been recent progress in adding new sequences to already built large phylogenies (Minh et al., 2020; Turakhia et al., 2021a; Aksamentov et al., 2021; Ye et al., 2022a,b).

These methods generally rely on heuristics in order to reduce computation. Less progress has been made on developing similar methods for phylodynamic inference where we can update parameter estimates of a model using new data points. Of course doing so could eliminate the need to rerun computationally expensive methods in the presence of new data. Gill et al. (2020) introduced a method for online inference for phylodynamic models, but the method still relies on MCMC, and thus cannot be used with pandemic scale datasets. From the perspective of planning interventions, quick, accurate, and up-to-date inferences from online phylodynamic inference can ensure that decisions concerning public health can be made with enough information available and in a timely manner.

In our study, we applied VBSKY to SARS-CoV-19 sequences from Michigan, Florida, and the entire United States and found that the estimates of the effective reproductive number given by VBSKY closely align with those given by a method using public health data. While the results from our analysis demonstrate the efficacy of our method, one drawback is that users have to make several modeling decisions that could potentially impact downstream inference. We found that adjusting the smoothing prior and taking different sampling approaches could affect the parameter estimates especially further back in time towards the beginning of the pandemic when the availability of sequences was lower. In addition, we showed results where changing hyperparameters such as the number of trees and tips used could impact the inferred posterior of the parameters.

In Chapter 3, we justified our choice of hyperparameters and studied the difference in various smoothing and sampling schemes through experiments by running multiple analyses. Thus far any claims or justifications about these choices have only been empirical. There is already interest in scrutinizing phylogenetic birth death models through a theoretical lens (Louca and Pennell, 2020; Legried and Terhorst, 2021), but further research is needed to not only understand how robust our method is to model misspecification when the assumption underlying the heuristics prescribed are not met, but also how sensitive the method is to the choice of priors and hyperparameters. Supplementing our results with a theoretical justification of the empirical findings would both improve estimation of parameters in applications and unlock insights on how to further improve the method.

We also ran an additional analyses subsetting SARS-CoV-2 sequences by the strain they belonged to. This application of our method illustrates the one advantage that phylodynamic methods have over methods that only rely on surveillance data: because SARS-CoV-2 variants are characterized by their mutations, this strain split analysis is much easier to perform with phylodynamic methods. Moreover, we can use genetic data to make inference in cases where surveillance data is not available or data is sparse as we can infer transmission and recovery events without having to actually observe them. As our ability to gather sequences and characterize variants of not only SARS-CoV-2 but future diseases increase, this application of our method will continue to be

important.

On a related note, there is a many to one mapping between evolutionary and epidemiological processes to the inferred phylogeny. For this reason, it is important to consider other sources of data that can lend insight into which process generated the phylogeny governing the data. Integrating genetic data with spatiotemporal data and other epidemiological data is an active area of research (Morelli et al., 2012; Lemey et al., 2014; Mate et al., 2015; Dudas et al., 2017; Kraemer et al., 2021). Our divide and conquer stochastic variational inference approach can be used with any model that relates transmission trees inferred from sequences to a set of parameters. In the future, our method can be extended to various other methods that incorporate other types of data to either inform the underlying tree or the set of parameters.

While each chapter in this thesis represents a step forward for various research problems, there still remains many open problems and unexplored research topics in population genetics and viral phylodynamics adjacent to the work presented.

BIBLIOGRAPHY

Takashi Abe and Masanori Arita. Genomic surveillance in Japan of AY. 29—a new sub-lineage of SARS-CoV-2 delta variant with C5239T and T5514C mutations. *medRxiv*, 2021.

Andre J Aberer, Alexandros Stamatakis, and Fredrik Ronquist. An efficient independence sampler for updating branches in bayesian markov chain monte carlo sampling of phylogenetic trees. *Systematic biology*, 65(1):161–176, 2016.

Jeffrey R. Adrion, Christopher B. Cole, Noah Dukler, Jared G. Galloway, Ariella L. Gladstein, Graham Gower, Christopher C. Kyriazis, Aaron P. Ragsdale, Georgia Tsambos, Franz Baumdicker, Jedidiah Carlson, Reed A. Cartwright, Arun Durvasula, Bernard Y. Kim, Patrick McKenzie, Philipp W. Messer, Ekaterina Noskova, Diego Ortega-Del Vecchyo, Fernando Racimo, Travis J. Struck, Simon Gravel, Ryan N. Gutenkunst, Kirk E. Lohmeuller, Peter L. Ralph, Daniel R. Schrider, Adam Siepel, Jerome Kelleher, and Andrew D. Kern. A community-maintained standard library of population genetic models. *bioRxiv*, 2019. doi: 10.1101/2019.12.20.885129.

Ivan Aksamentov, Cornelius Roemer, Emma B Hodcroft, and Richard A Neher. Nextclade: clade assignment, mutation calling and quality control for viral genomes. *Journal of Open Source Software*, 6(67):3773, 2021.

Hussein Al-Asadi, Desislava Petkova, Matthew Stephens, and John Novembre. Estimating recent migration and population-size surfaces. *PLoS genetics*, 15(1):e1007908, 2019.

Michael E Alfaro and Mark T Holder. The posterior and the prior in bayesian phylogenetics. *Annu. Rev. Ecol. Evol. Syst.*, 37:19–42, 2006.

Augusto Anguita-Ruiz, Concepción M Aguilera, and Ángel Gil. Genetics of lactose intolerance: an updated review and online interactive world maps of phenotype and genotype frequencies. *Nutrients*, 12(9):2689, 2020.

Melissa M Arons, Kelly M Hatfield, Sujan C Reddy, Anne Kimball, Allison James, Jesica R Jacobs, Joanne Taylor, Kevin Spicer, Ana C Bardossy, Lisa P Oakley, et al. Presymptomatic SARS-CoV-2 infections and transmission in a skilled nursing facility. *New England journal of medicine*, 382(22):2081–2090, 2020.

Daniel Barry and John A Hartigan. Product partition models for change point problems. *The Annals of Statistics*, pages 260–279, 1992.

Daniel Barry and John A Hartigan. A bayesian analysis for change point problems. *Journal of the American Statistical Association*, 88(421):309–319, 1993.

Dennis A Benson, Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, and Eric W Sayers. Genbank. *Nucleic acids research*, 41(D1):D36–D42, 2012.

A. Bhaskar, Y. X. Rachel Wang, and Y. S. Song. Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data. *Genome Research*, 25(2):268–279, 2015.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387310738.

Remco Bouckaert, Timothy G Vaughan, Joëlle Barido-Sottani, Sebastián Duchêne, Mathieu Fourment, Alexandra Gavryushkina, Joseph Heled, Graham Jones, Denise Kühnert, Nicola De Maio, et al. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS computational biology*, 15(4):e1006650, 2019.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/google/jax.

Brian L Browning, Ying Zhou, and Sharon R Browning. A one-penny imputed genome from next-generation reference panels. *The American Journal of Human Genetics*, 103(3):338–348, 2018.

Jared Bullard, Kerry Dust, Duane Funk, James E Strong, David Alexander, Lauren Garnett, Carl Boodman, Alexander Bello, Adam Hedley, Zachary Schiffman, et al. Predicting infectious severe acute respiratory syndrome coronavirus 2 from diagnostic samples. *Clinical Infectious Diseases*, 71(10):2663–2666, 2020.

Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

Finlay Campbell, Brett Archer, Henry Laurenson-Schafer, Yuka Jinnai, Franck Konings, Neale Batra, Boris Pavlin, Katelijn Vandemaele, Maria D Van Kerkhove, Thibaut Jombart, et al. Increased transmissibility and global spread of SARS-CoV-2 variants of concern as at June 2021. *Eurosurveillance*, 26(24):2100509, 2021.

Rebecca L Cann, Mark Stoneking, and Allan C Wilson. Mitochondrial dna and human evolution. *Nature*, 325(6099):31–36, 1987.

Jedidiah Carlson, Adam E Locke, Matthew Flickinger, Matthew Zawistowski, Shawn Levy, Richard M Myers, Michael Boehnke, Hyun Min Kang, Laura J Scott, Jun Z Li, et al. Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans. *Nature communications*, 9(1):1–13, 2018.

Shai Carmi, Peter R Wilton, John Wakeley, and Itsik Pe'er. A renewal theory approach to ibd sharing. *Theoretical population biology*, 97:35–48, 2014.

Andrew H Chan, Paul A. Jenkins, and Yun S. Song. Genome-wide fine-scale recombination rate variation in Drosophila melanogaster. *PLoS Genetics*, 8(12):e1003090, December 2012.

Lounès Chikhi, Willy Rodríguez, Simona Grusea, Patricia Santos, Simon Boitard, and Olivier Mazet. The iicr (inverse instantaneous coalescence rate) as a summary of genomic diversity: insights into demographic inference and model choice. *Heredity*, 120(1):13–24, 2018.

Sayantan Das, Lukas Forer, Sebastian Schönherr, Carlo Sidore, Adam E Locke, Alan Kwong, Scott I Vrieze, Emily Y Chew, Shawn Levy, Matt McGue, et al. Next-generation genotype imputation service and methods. *Nature genetics*, 48(10):1284–1287, 2016.

William S DeWitt, Kameron Decker Harris, Aaron P Ragsdale, and Kelley Harris. Nonparametric coalescent inference of mutation spectrum history and demography. *Proceedings of the National Academy of Sciences*, 118(21), 2021.

Enes Dilber and Jonathan Terhorst. Robust detection of natural selection using a probabilistic model of tree imbalance. *Genetics*, 01 2022. ISSN 1943-2631. doi: 10.1093/genetics/iyac009. URL https://doi.org/10.1093/genetics/iyac009. iyac009.

Vu Dinh, Arman Bilge, Cheng Zhang, and Frederick A Matsen IV. Probabilistic path hamiltonian monte carlo. In *International Conference on Machine Learning*, pages 1009–1018. PMLR, 2017.

DLMF. *NIST Digital Library of Mathematical Functions*. http://dlmf.nist.gov/, Release 1.0.27 of 2020-06-15. URL http://dlmf.nist.gov/. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds.

Alexei Drummond and Allen G Rodrigo. Reconstructing genealogies of serial samples under the assumption of a molecular clock using serial-sample upgma. *Molecular Biology and Evolution*, 17(12):1807–1815, 2000.

Alexei Drummond, Roald Forsberg, and Allen G Rodrigo. The inference of stepwise changes in substitution rates using serial sequence samples. *Molecular Biology and Evolution*, 18(7): 1365–1371, 2001.

Alexei J Drummond and Andrew Rambaut. Beast: Bayesian evolutionary analysis by sampling trees. *BMC evolutionary biology*, 7(1):1–8, 2007.

Alexei J Drummond, Andrew Rambaut, BETH Shapiro, and Oliver G Pybus. Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular biology and evolution*, 22(5):1185–1192, 2005.

Gytis Dudas, Luiz Max Carvalho, Trevor Bedford, Andrew J Tatem, Guy Baele, Nuno R Faria, Daniel J Park, Jason T Ladner, Armando Arias, Danny Asogun, et al. Virus genomes reveal factors that spread and sustained the ebola epidemic. *Nature*, 544(7650):309–315, 2017.

Richard Durbin. Efficient haplotype matching and storage using the positional burrows–wheeler transform (pbwt). *Bioinformatics*, 30(9):1266–1272, 2014.

R. Durrett. *Probability Models for DNA Sequence Evolution*. Springer, New York, 2nd edition, 2008.

J Dutheil, G Ganapathy, A Hobolth, T Mailund, M Uyenoyama, and M Schierup. Ancestral population genomics: The coalescent hidden Markov model approach. *Genetics*, 183:259–274, 2009.

Robert Ekblom, Birte Brechlin, Jens Persson, Linnéa Smeds, Malin Johansson, Jessica Magnusson, Øystein Flagstad, and Hans Ellegren. Genome sequencing and conservation genomics in the scandinavian wolverine population. *Conservation Biology*, 32(6):1301–1312, 2018.

Stefan Elbe and Gemma Buckland-Merrett. Data, disease and diplomacy: Gisaid's innovative contribution to global health. *Global Challenges*, 1(1):33–46, 2017.

María Inés Fariello, Simon Boitard, Hugo Naya, Magali SanCristobal, and Bertrand Servin. Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. *Genetics*, 193(3):929–941, 2013.

Paul Fearnhead. Exact and efficient bayesian inference for multiple changepoint problems. *Statistics and computing*, 16(2):203–213, 2006.

J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–376, 1981.

Mathieu Fourment and Aaron E Darling. Evaluating probabilistic programming and fast variational bayesian inference in phylogenetics. *PeerJ*, 7:e8272, 2019.

Mathieu Fourment, Andrew F Magee, Chris Whidden, Arman Bilge, Frederick A Matsen IV, and Vladimir N Minin. 19 dubious ways to compute the marginal likelihood of a phylogenetic tree topology. *Systematic Biology*, 69(2):209–220, 2020.

Christophe Fraser, Christl A Donnelly, Simon Cauchemez, William P Hanage, Maria D Van Kerkhove, T Déirdre Hollingsworth, Jamie Griffin, Rebecca F Baggaley, Helen E Jenkins, Emily J Lyons, et al. Pandemic potential of a strain of influenza a (h1n1): early findings. *science*, 324(5934):1557–1561, 2009.

Jo C Gay, Simon Myers, and Gilean McVean. Estimating meiotic gene conversion rates from population genetic data. *Genetics*, 177:881–894, 2007.

Mandev S Gill, Philippe Lemey, Nuno R Faria, Andrew Rambaut, Beth Shapiro, and Marc A Suchard. Improving bayesian population dynamics inference: a coalescent-based model for multiple loci. *Molecular biology and evolution*, 30(3):713–724, 2013.

Mandev S Gill, Philippe Lemey, Marc A Suchard, Andrew Rambaut, and Guy Baele. Online bayesian phylodynamic inference in beast with application to epidemic reconstruction. *Molecular biology and evolution*, 37(6):1832–1842, 2020.

R. C. Griffiths and P. Marjoram. An ancestral recombination graph. In P. Donnelly and S. Tavaré, editors, *Progress in population genetics and human evolution*, volume 87, pages 257–270. Springer-Verlag, Berlin, 1997.

R.C. Griffiths and P. Marjoram. Ancestral inference from samples of dna sequences with recombination. *Journal of Computational Biology*, 3(4):479–502, 1996.

Ryan N. Gutenkunst, Ryan D. Hernandez, Scott H. Williamson, and Carlos D. Bustamante. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, 5(10):e1000695, 2009.

Kelley Harris. Evidence for recent, population-specific evolution of the human mutation rate. *Proceedings of the National Academy of Sciences*, 112(11):3439–3444, 2015.

Kelley Harris, Sara Sheehan, John A Kamm, and Yun S Song. Decoding coalescent hidden Markov models in linear time. In *Proc. 18th Annual Intl. Conf. on Research in Computational Molecular Biology (RECOMB)*, volume 8394 of *LNBI*, pages 100–114. Springer, 2014. (NIHMSID 597680, PMC Pending).

J. Hein, M. H. Schierup, and C. Wiuf. *Gene genealogies, variation and evolution*. Oxford University Press, 2005.

Donna Henderson, Sha Zhu, Christopher B Cole, and Gerton Lunter. Demographic inference from multiple whole genomes using a particle filter for continuous markov jump processes. *PloS one*, 16(3):e0247647, 2021.

Asger Hobolth and Jens Ledet Jensen. Markovian approximation to the finite loci coalescent with recombination along multiple sequences. *Theoretical population biology*, 98:48–58, 2014.

Emma B Hodcroft, Nicola De Maio, Rob Lanfear, Duncan R MacCannell, Bui Quang Minh, Heiko A Schmidt, Alexandros Stamatakis, Nick Goldman, and Christophe Dessimoz. Want to track pandemic variants faster? fix the bioinformatics bottleneck, 2021.

Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(5), 2013.

Sebastian Höhna and Alexei J Drummond. Guided tree topology proposals for bayesian phylogenetic inference. *Systematic biology*, 61(1):1–11, 2012.

Bryan N Howie, Peter Donnelly, and Jonathan Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*, 5(6): e1000529, 2009.

Richard R Hudson et al. Gene genealogies and the coalescent process. *Oxford surveys in evolutionary biology*, 7(1):44, 1990.

Brad Jackson, Jeffrey D Scargle, David Barnes, Sundararajan Arabhi, Alina Alt, Peter Gioumousis, Elyus Gwin, Paungkaew Sangtrakulcharoen, Linda Tan, and Tun Tao Tsai. An algorithm for optimal partitioning of data on an interval. *IEEE Signal Processing Letters*, 12(2):105–108, 2005.

Xiang Ji, Zhenyu Zhang, Andrew Holbrook, Akihiko Nishimura, Guy Baele, Andrew Rambaut, Philippe Lemey, and Marc A Suchard. Gradients do grow on trees: a linear-time o (n)-dimensional gradient for statistical phylogenetics. *Molecular Biology and Evolution*, 37(10): 3047–3060, 2020.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

Julien Jouganous, Will Long, Aaron P Ragsdale, and Simon Gravel. Inferring the joint demographic history of multiple populations: beyond the diffusion approximation. *Genetics*, 206(3): 1549–1567, 2017.

Jack Kamm, Jonathan Terhorst, Richard Durbin, and Yun S Song. Efficiently inferring the demographic history of many populations with allele count data. *Journal of the American Statistical Association*, 115(531):1472–1487, 2020.

John A Kamm, Jeffrey P Spence, Jeffrey Chan, and Yun S Song. Two-locus likelihoods under variable population size and fine-scale recombination rate estimation. *Genetics*, 203(3):1381–1399, 2016.

John A Kamm, Jonathan Terhorst, and Yun S Song. Efficient computation of the joint sample frequency spectra for multiple populations. *Journal of Computational and Graphical Statistics*, 26(1):182–194, 2017.

Michael Karcher, Cheng Zhang, and Frederick A Matsen IV. Variational bayesian supertrees. *arXiv preprint arXiv:2104.11191*, 2021.

Jerome Kelleher, Alison M Etheridge, and Gilean McVean. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS computational biology*, 12(5):e1004842, 2016.

Jerome Kelleher, Yan Wong, Anthony W Wohns, Chaimaa Fadil, Patrick K Albers, and Gil McVean. Inferring whole-genome histories in large population datasets. *Nature Genetics*, 51 (9):1330–1338, 2019.

William Ogilvy Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721, 1927.

Rebecca Killick, Paul Fearnhead, and Idris A Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012.

John FC Kingman. On the genealogy of large populations. *Journal of applied probability*, pages 27–43, 1982.

Galya V Klink, Ksenia Safina, Elena Nabieva, Nikita Shvyrev, Sofya Garushyants, Evgeniia Alekseeva, Andrey B Komissarov, Daria M Danilenko, Andrei A Pochtovyi, Elizaveta V Divisenko, et al. The rise and spread of the SARS-CoV-2 AY. 122 lineage in Russia. *medRxiv*, 2021.

Takahiko Koyama, Daniel Platt, and Laxmi Parida. Variant analysis of SARS-CoV-2 genomes. *Bulletin of the World Health Organization*, 98(7):495, 2020.

Moritz UG Kraemer, Verity Hill, Christopher Ruis, Simon Dellicour, Sumali Bajaj, John T McCrone, Guy Baele, Kris V Parag, Anya Lindström Battle, Bernardo Gutierrez, et al. Spatiotemporal invasion dynamics of sars-cov-2 lineage b. 1.1. 7 emergence. *Science*, 373(6557):889–895, 2021.

Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M Blei. Automatic differentiation variational inference. *The Journal of Machine Learning Research*, 18(1):430–474, 2017.

Mary K Kuhner and Joseph Felsenstein. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular biology and evolution*, 11(3):459–468, 1994.

Denise Kühnert, Chieh-Hsi Wu, and Alexei J Drummond. Phylogenetic and epidemic modeling of rapidly evolving infectious diseases. *Infection, genetics and evolution*, 11(8):1825–1841, 2011.

Alessia Lai, Annalisa Bergna, Carla Acciarri, Massimo Galli, and Gianguglielmo Zehender. Early phylogenetic estimate of the effective reproduction number of SARS-CoV-2. *Journal of medical virology*, 92(6):675–679, 2020.

Tracy Lam-Hine, Stephen A McCurdy, Lisa Santora, Lael Duncan, Russell Corbett-Detig, Beatrix Kapusinszky, and Matthew Willis. Outbreak associated with SARS-CoV-2 B.1.617.2 (delta) variant in an elementary school—Marin County, California, May–June 2021. *Morbidity and Mortality Weekly Report*, 70(35):1214, 2021.

Rob Lanfear. A global phylogeny of SARS-CoV-2 sequences from GISAID, November 2020. URL https://doi.org/10.5281/zenodo.4289383.

D.J. Lawson, G. Hellenthal, S. Myers, and D. Falush. Inference of population structure using dense haplotype data. *PLoS Genetics*, 8(1):e1002453, 2012.

Brandon Legried and Jonathan Terhorst. A class of identifiable birth-death models. *bioRxiv*, 2021.

Jüri Lember and Alexey A Koloydenko. 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):1–58, 2014.

Philippe Lemey, Andrew Rambaut, and Oliver G Pybus. Hiv evolutionary dynamics within and among hosts. *Aids Rev*, 8(3):125–140, 2006.

Philippe Lemey, Andrew Rambaut, Trevor Bedford, Nuno Faria, Filip Bielejec, Guy Baele, Colin A Russell, Derek J Smith, Oliver G Pybus, Dirk Brockmann, et al. Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza h3n2. *PLoS pathogens*, 10(2):e1003932, 2014.

Heng Li and Richard Durbin. Inference of human population history from individual whole-genome sequences. *Nature*, 475:493–496, 2011.

N. Li and M. Stephens. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165:2213–2233, 2003.

Y. Li and G. R. Abecasis. Mach 1.0: Rapid haplotype reconstruction and missing genotype inference. *Am. J. Hum. Genet.*, S79:2290, 2006.

Po-Ru Loh, Petr Danecek, Pier Francesco Palamara, Christian Fuchsberger, Yakir A Reshef, Hilary K Finucane, Sebastian Schoenherr, Lukas Forer, Shane McCarthy, Goncalo R Abecasis, et al. Reference-based phasing using the haplotype reference consortium panel. *Nature genetics*, 48(11):1443–1448, 2016.

Stilianos Louca and Matthew W Pennell. Extant timetrees are consistent with a myriad of diversification histories. *Nature*, 580(7804):502–505, April 2020. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-020-2176-1.

Gerton Lunter. Haplotype matching in large cohorts using the li and stephens model. *Bioinformatics*, 35(5):798–806, 2019.

Swagata Majumdar and Rakesh Sarkar. Mutational and phylogenetic analyses of the two lineages of the omicron variant. *Journal of medical virology*, 2021.

Anna-Sapfo Malaspinas, Michael C Westaway, Craig Muller, Vitor C Sousa, Oscar Lao, Isabel Alves, Anders Bergström, Georgios Athanasiadis, Jade Y Cheng, Jacob E Crawford, et al. A genomic history of aboriginal australia. *Nature*, 538(7624):207–214, 2016.

Jonathan Marchini, Bryan Howie, Simon R Myers, Gil McVean, and Peter Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet*, 39(7):906–13, 2007.

Paul Marjoram and Jeffrey D Wall. Fast "coalescent" simulation. *BMC Genet*, 7:16, 2006.

Suzanne E Mate, Jeffrey R Kugelman, Tolbert G Nyenswah, Jason T Ladner, Michael R Wiley, Thierry Cordier-Lassalle, Athalia Christie, Gary P Schroth, Stephen M Gross, Gloria J Davies-Wayne, et al. Molecular evidence of sexual transmission of ebola virus. *New England Journal of Medicine*, 373(25):2448–2454, 2015.

Niklas Mather, Samuel M Traves, and Simon YW Ho. A practical introduction to sequentially markovian coalescent methods for estimating demographic history from genomic data. *Ecology and evolution*, 10(1):579–589, 2020.

Olivier Mazet, Willy Rodríguez, Simona Grusea, Simon Boitard, and Lounès Chikhi. On the importance of being structured: instantaneous coalescence rates and human evolution—lessons for ancestral population size inference? *Heredity*, 116(4):362–371, 2016.

Jakob McBroome, Jennifer Martin, Adriano de Bernardi Schneider, Yatish Turakhia, and Russell Corbett-Detig. Identifying SARS-CoV-2 regional introductions and transmission clusters in real time. *medRxiv*, 2022.

Gilean AT McVean and Niall J Cardin. Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1459):1387–1393, 2005.

Bui Quang Minh, Heiko A Schmidt, Olga Chernomor, Dominik Schrempf, Michael D Woodhams, Arndt Von Haeseler, and Robert Lanfear. Iq-tree 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular biology and evolution*, 37(5):1530–1534, 2020.

Vladimir N Minin, Erik W Bloomquist, and Marc A Suchard. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Molecular biology and evolution*, 25(7):1459–1471, 2008.

Benoit Morel, Pierre Barbera, Lucas Czech, Ben Bettisworth, Lukas Hübner, Sarah Lutteropp, Dora Serdari, Evangelia-Georgia Kostaki, Ioannis Mamais, Alexey M Kozlov, et al. Phylogenetic analysis of sars-cov-2 data is difficult. *Molecular biology and evolution*, 38(5):1777–1791, 2021.

Marco J Morelli, Gaël Thébaud, Joël Chadœuf, Donald P King, Daniel T Haydon, and Samuel Soubeyrand. A bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data. *PLoS computational biology*, 8(11):e1002768, 2012.

Hélène Morlon, Todd L Parsons, and Joshua B Plotkin. Reconciling molecular phylogenies with the fossil record. *Proc. Natl. Acad. Sci. U. S. A.*, 108(39):16327–16332, September 2011. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1102543108.

Nicola F Müller, Daniel Wüthrich, Nina Goldman, Nadine Sailer, Claudia Saalfrank, Myrta Brunner, Noémi Augustin, Helena MB Seth-Smith, Yvonne Hollenstein, Mohammedyaseen Syedbasha, et al. Characterising the epidemic spread of influenza a/h3n2 within a city through phylogenetics. *PLoS pathogens*, 16(11):e1008984, 2020.

Sean Nee, Robert M. May, and Paul H. Harvey. The reconstructed evolutionary process. *Philosophical Transactions: Biological Sciences*, 344(1309):305–311, 1994.

Pier Francesco Palamara, Jonathan Terhorst, Yun S Song, and Alkes L Price. High-throughput inference of pairwise coalescence times identifies signals of selection and enriched disease heritability. *Nature Genetics*, 50(9):1311–1317, 2018.

Kris V Parag and Oliver G Pybus. Robust design for coalescent model inference. *Systematic biology*, 68(5):730–743, 2019.

Austin H Patton, Mark J Margres, Amanda R Stahlke, Sarah Hendricks, Kevin Lewallen, Rodrigo K Hamede, Manuel Ruiz-Aravena, Oliver Ryder, Hamish I McCallum, Menna E Jones, et al. Contemporary demographic reconstruction methods are robust to genome assembly quality: A case study in tasmanian devils. *Molecular biology and evolution*, 36(12):2906–2921, 2019.

Joshua S. Paul and Yun S. Song. A principled approach to deriving approximate conditional sampling distributions in population genetics models with recombination. *Genetics*, 186:321–338, 2010.

Joshua S. Paul, Matthias Steinrücken, and Yun S. Song. An accurate sequentially Markov conditional sampling distribution for the coalescent with recombination. *Genetics*, 187:1115–1128, 2011. (PMC3070520).

Desislava Petkova, John Novembre, and Matthew Stephens. Visualizing spatial population structure with estimated effective migration surfaces. *Nature genetics*, 48(1):94–100, 2016.

Javier Prado-Martinez, Peter H Sudmant, Jeffrey M Kidd, Heng Li, Joanna L Kelley, Belen Lorente-Galdos, Krishna R Veeramah, August E Woerner, Timothy D O'Connor, Gabriel Santpere, et al. Great ape genetic diversity and population history. *Nature*, 499(7459):471–475, 2013.

Alkes L Price, Arti Tandon, Nick Patterson, Kathleen C Barnes, Nicholas Rafaels, Ingo Ruczinski, Terri H Beaty, Rasika Mathias, David Reich, and Simon R Myers. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet*, 5(6):e1000519, 2009.

Kay Prüfer, Fernando Racimo, Nick Patterson, Flora Jay, Sriram Sankararaman, Susanna Sawyer, Anja Heinze, Gabriel Renaud, Peter H Sudmant, Cesare de Filippo, et al. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, 505(7481):43–49, 2014.

Oliver G Pybus and Andrew Rambaut. Evolutionary analysis of the dynamics of viral infectious disease. *Nature Reviews Genetics*, 10(8):540–550, 2009.

Oliver G Pybus, Andrew Rambaut, and Paul H Harvey. An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics*, 155(3):1429–1437, 2000.

Oliver G Pybus, Michael A Charleston, Sunetra Gupta, Andrew Rambaut, Edward C Holmes, and Paul H Harvey. The epidemic behavior of the hepatitis c virus. *Science*, 292(5525):2323–2325, 2001.

Andrew Rambaut and Nicholas C Grass. Seq-gen: an application for the monte carlo simulation of dna sequence evolution along phylogenetic trees. *Bioinformatics*, 13(3):235–238, 1997.

Matthew D Rasmussen, Melissa J Hubisz, Ilan Gronau, and Adam Siepel. Genome-wide inference of ancestral recombination graphs. *PLoS Genetics*, 10(5):e1004342, 2014.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning*, pages 1530–1538. PMLR, 2015.

Simone Rubinacci, Olivier Delaneau, and Jonathan Marchini. Genotype imputation using the positional burrows wheeler transform. *PLoS genetics*, 16(11):e1009049, 2020.

Timo Sahi. Genetics and epidemiology of adult-type hypolactasia. *Scandinavian Journal of Gastroenterology*, 29(sup202):7–20, 1994.

P. Scheet and M. Stephens. A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.*, 78:629–644, 2006.

Stephan Schiffels and Richard Durbin. Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, 46:919–925, 2014.

Stephan Schiffels and Ke Wang. Msmc and msmc2: the multiple sequentially markovian coalescent. In *Statistical population genomics*, pages 147–166. Humana, New York, NY, 2020.

Sara Sheehan, Kelley Harris, and Yun S Song. Estimating variable effective population sizes from multiple genomes: A sequentially Markov conditional sampling distribution approach. *Genetics*, 194(3):647–662, 2013.

Andy Shi, Sheila M Gaynor, Corbin Quick, and Xihong Lin. Multi-resolution characterization of the covid-19 pandemic: A unified framework and open-source tool. *medRxiv*, 2021.

Peter HA Sneath and Robert R Sokal. *Numerical taxonomy. The principles and practice of numerical classification.* 1973.

Leo Speidel, Marie Forest, Sinan Shi, and Simon R Myers. A method for genome-wide genealogy estimation for thousands of samples. *Nature Genetics*, 51(9):1321–1329, 2019.

Jeffrey P Spence, Matthias Steinrücken, Jonathan Terhorst, and Yun S Song. Inference of population history using coalescent hmms: Review and outlook. *Current opinion in genetics & development*, 53:70–76, 2018.

Tanja Stadler. On incomplete sampling under birth–death models and connections to the sampling-based coalescent. *Journal of theoretical biology*, 261(1):58–66, 2009.

Tanja Stadler. Simulating trees with a fixed number of extant species. *Systematic biology*, 60(5): 676–684, 2011.

Tanja Stadler, Denise Kühnert, Sebastian Bonhoeffer, and Alexei J Drummond. Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proceedings of the National Academy of Sciences*, 110(1):228–233, 2013.

Matthias Steinrücken, Jack Kamm, Jeffrey P Spence, and Yun S Song. Inference of complex population histories using whole-genome sequences from multiple populations. *Proceedings of the National Academy of Sciences*, 116(34):17115–17120, 2019.

Matthew Stephens and Paul Scheet. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am. J. Hum. Genet.*, 76(3):449–62, 2005.

Aaron J Stern, Leo Speidel, Noah A Zaitlen, and Rasmus Nielsen. Disentangling selection on genetically correlated polygenic traits via whole-genome genealogies. *The American Journal of Human Genetics*, 108(2):219–239, 2021.

Benjamin B Sun, Mitja I Kurki, Christopher N Foley, Asma Mechakra, Chia-Yen Chen, Eric Marshall, Jemma B Wilk, Biogen Biobank Team, Mohamed Chahine, Philippe Chevalier, et al. Genetic associations of protein-coding variants in human disease. *Nature*, 2022. URL https://doi.org/10.1038/s41586-022-04394-w.

Jonathan Terhorst, John A Kamm, and Yun S Song. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature genetics*, 49(2):303–309, 2017.

The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

Yatish Turakhia, Bryan Thornlow, Angie S Hinrichs, Nicola De Maio, Landen Gozashti, Robert Lanfear, David Haussler, and Russell Corbett-Detig. Ultrafast sample placement on existing trees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. *Nat. Genet.*, 53 (6):809–816, June 2021a. ISSN 1061-4036, 1546-1718. doi: 10.1038/s41588-021-00862-7.

Yatish Turakhia, Bryan Thornlow, Angie S Hinrichs, Jakob Mcbroome, Nicolas Ayala, Cheng Ye, Nicola De Maio, David Haussler, Rob Lanfear, and Russ Corbett-Detig. Pandemic-scale phylogenomics reveals elevated recombination rates in the SARS-CoV-2 spike region. *bioRxiv*, 2021b.

Lucy van Dorp, Charlotte J Houldcroft, Damien Richard, and François Balloux. COVID-19, the first pandemic in the post-genomic era. *Curr. Opin. Virol.*, 50:40–48, October 2021. ISSN 1879-6257, 1879-6265. doi: 10.1016/j.coviro.2021.07.002.

B. F. Voight, S. Kudaravalli, X. Wen, and J. K. Pritchard. A map of recent positive selection in the human genome. *PLoS Biology*, 4:e72, 2006.

Erik Volz, Swapnil Mishra, Meera Chand, Jeffrey C Barrett, Robert Johnson, Lily Geidelberg, Wes R Hinsley, Daniel J Laydon, Gavin Dabrera, Áine O'Toole, et al. Transmission of SARS-CoV-2 Lineage B.1.1.7 in England: Insights from linking epidemiological and genetic data. *medRxiv*, pages 2020–12, 2021.

Erik M Volz, Sergei L Kosakovsky Pond, Melissa J Ward, Andrew J Leigh Brown, and Simon DW Frost. Phylodynamics of infectious disease epidemics. *Genetics*, 183(4):1421–1430, 2009.

Erik M Volz, Edward Ionides, Ethan O Romero-Severson, Mary-Grace Brandt, Eve Mokotoff, and James S Koopman. Hiv-1 transmission during early infection in men who have sex with men: a phylodynamic analysis. *PLoS medicine*, 10(12):e1001568, 2013a.

Erik M Volz, Katia Koelle, and Trevor Bedford. Viral phylodynamics. *PLoS computational biology*, 9(3):e1002947, 2013b.

Martin J Wainwright and Michael Irwin Jordan. *Graphical models, exponential families, and variational inference*. Now Publishers Inc, 2008.

Jane-Ling Wang. Smoothing hazard rates. *Wiley StatsRef: Statistics Reference Online*, 2014.

Chris Whidden and Frederick A Matsen IV. Quantifying mcmc exploration of phylogenetic tree space. *Systematic biology*, 64(3):472–491, 2015.

Peter R Wilton, Shai Carmi, and Asger Hobolth. The smc' is a highly accurate approximation to the ancestral recombination graph. *Genetics*, 200(1):343–355, 2015.

Carsten Wiuf and Jotun Hein. Recombination as a point process along sequences. *Theoretical Population Biology*, 55(3):248–259, 1999.

Christopher Yau and Christopher C Holmes. A decision-theoretic approach for segmental classification. *The Annals of Applied Statistics*, pages 1814–1835, 2013.

Cheng Ye, Bryan Thornlow, Alexander Michael Kramer, Jakob McBroome, Angie S Hinrichs, Russell Corbett-Detig, and Yatish Turakhia. Pandemic-scale phylogenetics. *bioRxiv*, 2021.

Cheng Ye, Bryan Thornlow, Angie S Hinrichs, Devika Torvi, Robert Lanfear, Russell Corbett-Detig, and Yatish Turakhia. matoptimize: A parallel tree optimization method enables online phylogenetics for SARS-CoV-2. *bioRxiv*, 2022a.

Yongtao Ye, Marcus Shum, Joseph Tsui, Guangchuang Yu, David Smith, Huachen Zhu, Joseph Wu, Yi Guan, and Tommy Tsan-Yuk Lam. Robust expansion of phylogeny for fast-growing genome sequence data. *bioRxiv*, pages 2021–12, 2022b.

Cheng Zhang. Improved variational bayesian phylogenetic inference with normalizing flows. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18760–18771. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/d96409bf894217686ba124d7356686c9-Paper.pdf.

Cheng Zhang and Frederick A Matsen IV. Generalizing tree probability estimation via bayesian networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/b137fdd1f79d56c7edf3365fea7520f2-Paper.pdf.

Cheng Zhang and Frederick A Matsen IV. Variational bayesian phylogenetic inference. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=SJVmjjR9FX.

Tao Zhou, Quanhui Liu, Zimo Yang, Jingyi Liao, Kexin Yang, Wei Bai, Xin Lu, and Wei Zhang. Preliminary prediction of the basic reproduction number of the wuhan novel coronavirus 2019-ncov. *Journal of Evidence-Based Medicine*, 13(1):3–7, 2020.