Title

Bayesian Phylogenetic Inference using Dated DNA Samples with Applications to HIV Latency and Ancient DNA

Author

Nagel, Anna Audrey

Publication Date

2023

Peer reviewed|Thesis/dissertation

Bayesian Phylogenetic Inference using Dated DNA Samples with Applications to HIV Latency and Ancient DNA

By

ANNA A. NAGEL
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Population Biology

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Bruce Rannala, Chair

Michael Turelli

Ziheng Yang

Committee in Charge

2023

Contents

Bayesian Phylogenetic Inference using Dated DNA Samples with Applications to HIV Latency and Ancient DNA

Abstract

A longstanding goal in phylogenetics is to estimate when species, populations, or individuals (in the case of viruses) diverged from a common ancestor. In molecular phylogenetics, which uses sequence data to estimate the topology and branch lengths of phylogenies, the rate of molecular evolution and time are confounded. Only the product of rate and time is identifiable without outside information. Several methods have been used to disentangle rate and time, including fossil calibrations and dated samples (known as tip dating). Tip dating exploits the difference in branch lengths between samples of different known ages to both estimate the rate of molecular evolution and to time calibrate a phylogeny assuming a molecular clock. This dissertation focuses on the development and application of novel tip dating methods to estimate time calibrated phylogenies.

Chapter 1 provides an introduction to relevant areas of phylogenetics. Types of studies that utilize time calibrated phylogenies are described. Then, a brief introduction into Bayesian phylogenetics and MCMC methods is provided. Lastly, phylogenetic time calibration methods are outlined. Chapter 2 applies tip dating to investigate the temporal dynamics of latency in HIV. Effective antiretroviral therapy (ART) for HIV stops HIV from infecting new cells and most infected cells die shortly after infection, leaving HIV clinically undetectable within a patient. However, a small pool of long-lived cells, known as latently infected cells, can persist with the HIV integrated into their genomes for decades. If ART is stopped, these latent cells rapidly repopulate the patient with HIV and lead to disease progression. Due to its clinical relevance, researchers are interested in understanding characteristics of the latent reservoir, such as when individual cells in the reservoir became latent. Because HIV evolves rapidly within hosts, phylogenetic tip dating methods can be used to estimate time calibrated trees for within-host viral datasets. However, viral lineages from latent cells have a much lower mutation rate in comparison to lineages from non-latent sequences. In phylogenies with both latent and non-latent HIV sequences, this difference in mutation rate leads to shorter branch lengths for latent sequences than would be expected given the sampling

times. A novel Bayesian tip dating method is developed that estimates when individual latent lineages became latent using this difference in branch lengths. A method to combine inferences across different regions of the HIV genome is also developed, which accounts for the fact that regions may differ in topology due to recombination. Combined inference greatly improves the accuracy of inferences when using only a few short sequences. The new methods perform better than many alternative heuristic methods and allow for biologically reasonable bounds on inferences, such as enforcing the latency times to be older than the sampling times. Lastly, the empirical utility of the method is demonstrated by analyzing two clinical datasets of patients with HIV.

Chapter 3 develops a method to analyze ancient DNA (aDNA) under the multispecies coalescent (MSC). With the increasing abundance of aDNA sequences, molecular data are now available to investigate the relationships between extinct and extant species, as well as between ancient samples of extant species and their modern relatives. These studies typically treat gene trees as species trees, which can lead to biases in inferred divergence times. The MSC overcomes these issues by explicitly modeling the relationship between gene trees and species trees. However, there are currently no methods that allow for tip dating with multiple sample dates within a species with the MSC; failing to account for sampling dates can also bias divergence time estimation. A method is developed to analyze aDNA under a MSC model, allowing for the inference of divergence times (in time units of both expected number of mutations and calendar time), effective population sizes, and mutation rate using large multilocus datasets with multiple individuals sampled in each population. Simulation studies suggest the new method can estimate the parameters accurately and precisely if the model assumptions are met. It is shown that treating ancient samples as contemporary (mimicking empirical practices) can lead to biases in estimates of divergence times and effective population sizes. Finally, two datasets with extant elephant species and woolly mammoths are analyzed. A strong signal of aDNA degradation was detected in one of the datasets, which likely biased estimates of mutation rate and divergence times. This suggests the need for more careful consideration of the impacts of DNA degradation on downstream analyses.

This dissertation demonstrates the wide utility and diverse potential applications of Bayesian tip dating methods, and provides powerful new methods to analyze empirical datasets on HIV latency and aDNA.

Acknowledgments

Above all, I would like to thank my supervisor, Bruce Rannala. This work would not have been possible without his countless hours of discussion, draft revisions, and programming support. He allowed me an enormous amount of freedom and was extremely generous with his time and input.

This work builds on existing software written by Ziheng Yang and Tomáš Flouri. I am thankful to their allowing me to contribute to their programs and their insights into the software. I would like to thank my dissertation committee members Ziheng Yang and Michael Turelli for their insights in the field and comments on my work.

I would like to thank my labmates Sneha Chakraborty, Mike May, and Jiansi Gao for their feedback on my projects and presentations. Lastly, I want to thank my family, especially my parents, for their support and encouragement throughout my education.

CHAPTER 1

Introduction

1.1. Necessity of Time Calibrated Phylogenies

Inferring time calibrated phylogenies is a long-standing goal in evolutionary biology. Time calibrated phylogenies have branch lengths in units such as years or days, and time is referred to as absolute or calendar time. In contrast, branch lengths of phylogenies that are not time calibrated are typically in units of expected number of DNA or amino acid substitutions. Inferring time calibrated phylogenies is a major area of research in its own right, but it is also the first step in investigating a wide variety of evolutionary questions relating to character evolution, diversification, epidemiology, and population demography. Some of these questions may be investigated with an uncalibrated phylogeny. However, absolute ages are necessary to study evolution relative to other events (or records), such as geologic, climatic, and biotic changes or events impacting disease spread. Here I provide an overview of several important research areas in which time calibrated phylogenies are employed.

1.1.1. Phylogenetic Comparative Methods. Phylogenetic comparative methods (PCMs) seek to associate a phenotype or trait across many taxa in a tree with one or more additional variables, such as other traits, environmental variables, or diversification rates. PCMs often utilize a time calibrated phylogeny. Since taxa share evolutionary history on the internal branches of a phylogeny and evolve from common ancestors, the traits of each taxa are not independent, and standard statistical tests that assume independence cannot be applied. The method of phylogenetic independent contrasts was proposed to account for the correlation of traits among different species due to shared evolutionary history (Felsenstein, 1985). This method tests for associations between continuous traits from species on a phylogeny assuming they evolve according to Brownian motion process. A large body of work extends this concept to other types of data, such as discrete characters, or other models of character evolution, such as Ornstein-Uhlenbeck. Bayesian methods

have been developed to accommodate uncertainty in the phylogeny and ancestral character states (reviewed in Huey *et al.*, 2019; Martins, 1996; Paradis, 2014; Ronquist, 2004). Many of these methods can be applied using either an uncalibrated or calibrated phylogeny. Some view the branch length units as a hypothesis about how evolution occurs and suggest testing the impact of different branch length units on inferences (Pagel, 1994). More generally, researchers may be interested in rates of character evolution over calendar time, which some methods can directly infer if a tree calibrated in units of calendar time is used. It is difficult to interpret rates that are in units of rate of character change per expected number of substitutions as opposed to rate of character change per year, unless the interest is only in the relative rates of evolution. If the question of interest specifically relates to absolute time, then a time calibrated phylogeny is required.

1.1.2. Diversification Rates. Studies of diversification rates are another example where it may be more intuitive to use phylogenies scaled in absolute time. Using absolute time yields results for speciation and extinction rates in per year units rather than expected number of substitutions (which may be hard to interpret). Simple birth-death models estimate rates of speciation and extinction assuming the rates are constant over time and using only extant species (Nee *et al.*, 1994). Other methods ask whether diversification rates vary over time, allowing for the inference of episodic changes in birth or death rates, such as mass extinction events (May *et al.*, 2016). A commonly used method, Bayesian analysis of macroevolutionary mixtures (BAMM), estimates shifts in diversification rates along lineages (Rabosky, 2014). When using a time calibrated tree, this allows researchers to date shifts in diversification rates to specific geologic time periods. For example, Varga *et al.* (2019) found a rapid increase in mushroom diversification rate in the early Jurassic period. However, the statistical correctness of this method has been questioned (Moore *et al.*, 2016).

Researchers are also interested in whether particular traits may impact diversification rates. In the simplest case, under a binary-state speciation and extinction (BiSSE) model, the speciation and extinction rates depend on a two-state character (Maddison *et al.*, 2007). In this model, there are a pair of distinct speciation and extinction rates associated with each character state as well as transition rates between the two character states. As with other methods, using an uncalibrated phylogeny in BiSSE results in estimates with units of expected substitutions that are most useful

for direct comparisons within the model. For example, Goldberg *et al.* (2010) found that speciation rates of self-incompatible species were increased relative to self-compatible species. A family of related speciation and extinction (SSE) models was subsequently developed to allow for more types of character data, such as quantitative traits (QuaSSE) (FitzJohn, 2010), multiple character states (MuSSE) (FitzJohn, 2012), and hidden states that affect diversification (HiSSE) (Beaulieu and O'Meara, 2016) which may impact inferences if the hidden states are not modeled (Rabosky and Goldberg, 2015). Using a time calibrated phylogeny with these models allows of rate estimates to be interpreted in absolute time. This approach was taken by Harvey *et al.* (2020), who used a broad range of SSE models in conjunction with other approaches to investigate drivers of neotropical bird diversity, estimating diversification rates in calendar time.

1.1.3. Epidemiology. Research studying the phylogenetics of infectious diseases and phylodynamics often focuses on inferring the timing of major events or shifts in the pathogen dynamics. While it is possible to study these questions using relative time units, researchers are typically interested in comparing to the timing of outside events measured in calendar time, such as known case numbers by date, changes in policy and behavior, or climactic changes (Pybus and Rambaut, 2009; Volz *et al.*, 2013). For novel pathogens in particular, a key question is often when the disease emerged (Volz *et al.*, 2013). Due to population genetic processes, the time of the most recent common ancestor (tMRCA) of all sequences, which is the root age, is not equivalent to the time of emergence (Volz *et al.*, 2013). The root age depends on which lineages are sampled, and lineages may be lost over time. While emergence time cannot be directly inferred with a time calibrated tree, phylogenies can bound the youngest possible emergence time and possibly suggest realistic divergence times, depending on the samples available (Volz *et al.*, 2013). Emergence time can be investigated in either a local or a global context. For example, a time scaled phylogeny of SARS-CoV-2 from Washington state suggested cryptic transmission occurred early in the outbreak (Bedford *et al.*, 2020). Similarly, a time scaled phylogeny of HIV suggested that HIV started spreading and diversifying within humans around the beginning of 20th century, much earlier than the date that emergence of a novel disease was recognized (Worobey *et al.*, 2008).

Understanding how a disease spreads over time and space is another major area of interest in infectious disease research. These questions are often studied using "mugration" models, which

model migration between discrete geographic areas in the same way that mutation events are treated in traditional phylogenetic models (Lemey *et al.*, 2009), though non-model based approaches, such as directly counting the minimum number of transmission events between areas required on a particular phylogeny, are also used, particularly with smaller trees (e.g. Di Giallonardo *et al.*, 2020; Lu *et al.*, 2020). For example, phylodynamic models revealed that, during the SARS-CoV-2 pandemic, policy changes (such as border closures) led to a decrease in migration rates between countries (Gao *et al.*, 2022). In contrast, the implementation of lockdowns in response to SARS-CoV-2 outbreaks in Italy did not prevent new transmission clusters or transmission between regions (Di Giallonardo *et al.*, 2020). In Guangdong Province, phylogenetic analysis suggested multiple introductions of SARS-CoV-2 lineages into the province in the second half of January 2020, with several lineages likely having later introductions (Lu *et al.*, 2020). As a final example, Dellicour *et al.* (2020) estimated the rate of West Nile virus spread over space and found the rate varied through time with higher rates of spread leading up to an epidemic. In all of these studies, the key findings were obtained by the use of a time calibrated tree.

1.1.4. Demography. Understanding how population size changes over time is of interest when population sizes are expected to be correlated with or driven by environmental changes. For example, researchers have combined data on historical climate and pollen abundance with genetic data to understand changes in woolly mammoth population sizes (MacDonald *et al.*, 2012). Shapiro *et al.* (2004) estimated that the demographic decline of Beringian steppe bison coincided with a warm period which predated the last glacial maximum and evidence of large human populations in this area, arguing environmental changes were the major contributor to population declines.

1.1.5. Summary. The wide range of questions whose answer depends on having a time calibrated phylogeny highlights the need for methods to accurately and precisely infer divergence times. The data available in different systems is highly variable, including sequence data (alone or with tip dates), fossils, morphological character, and molecular datasets. This necessitates a diversity of time calibration methods that can accommodate these difference sources of information.

1.2. Phylogenies, Species Trees, and Gene Trees

With eukaryotes, phylogenetists are usually interested in inferring species trees, which show the relationships between species (or populations) and their divergence times. A species (population) tree shows the idealized relationships between species (populations), assuming that their relationships can be described as a binary tree. A species (population) is a group of interbreeding individuals that have some degree of genetic isolation from individuals of other species (populations).

When inferring a molecular phylogeny, researchers typically use genes sampled from one or more individuals of each species of interest. The phylogeny of a sample of genes reflects the relationship between the sampled genes (the gene tree), rather than the species tree (Hudson, 1990). The relationships between genes may differ from the relationships between species (populations) for many reasons, one of the most important being the coalescent process, described below. Historically researchers have not distinguished between species trees and gene trees; gene trees were treated as equivalent to species trees (Degnan and Rosenberg, 2009). This implicitly assumes that the relationship between genes is equivalent to that between species (populations). Bacterial and viral analyses typically infer gene trees.

1.2.1. The Coalescent. The coalescent is a population genetic model used to describe the relationship between sampled genes in a single panmictic population, including the topology and ages of the most recent common ancestors (MRCAs), also referred to as coalescent times, of genes in the sample. The coalescent process was rigorously described by Kingman (Kingman, 1982a,b). As a simple case, assume there are n diploid individuals sampled at time present from a panmictic population of constant size, N. Assume that $n << N$ and use a continuous approximation to discrete generation times. Going backward in time, the waiting time, t_i, to the coalescent event that decreases the number of lineages from i to $i-1$ is exponentially distributed with rate parameter $\frac{i(i-1)}{2} \times \frac{1}{2N}$. The coalescent is a neutral model. All lineages are equally likely to coalesce, which implies all gene tree topologies are equally likely.

1.2.2. The Multispecies Coalescent. The multispecies coalescent (MSC) is an extension of the coalescent that allows for the coalescent process on a species tree. Starting in each species

(population) on the tips of the species tree, the coalescent process proceeds independently in different species going backward in time until a speciation event. At a speciation event, the lineages in the two daughter species are combined, and the coalescent process continues in the parent species (population) (Rannala and Yang, 2003).

The tMRCA of one locus sampled from each of two populations, A and B, must be older than the divergence time of populations A and B. This is because there is no opportunity for the lineages to coalesce until they are in the same population. This biases divergence time estimates to be too old when treating gene trees as species trees (Angelis and Dos Reis, 2015; Gillespie and Langley, 1979). Additionally, the waiting time to coalesce after the lineages are in the same population may be large; it is possible that the waiting time is larger than the time to the next speciation event going backward in time along the species tree. In this case, there may be lineages sampled from three or more species that persisted in an ancestral population. Since the lineages have equal probability of coalescing in any order, the gene tree topology may not match the species tree topology. The process leading to such gene tree species tree conflict is referred to as incomplete lineage sorting (ILS).

The MSC is a more realistic model for studying species relationships that allows for gene trees and species trees to differ due to the coalescent process. However, there are other possible sources of discordance between gene trees and species trees such as hybridization, horizontal gene transfer, and gene duplication (Boussau and Scornavacca, 2020; Maddison, 1997). The relative contributions of ILS and other sources of gene tree species tree discordance will depend on the species studied (Maddison, 1997). ILS is expected to be common in ancient radiations, such as early bird evolution and cichlids (reviewed in Degnan and Rosenberg, 2009). This is because the differences between gene trees and species trees are likely to be largest due to the MSC when population sizes are large (since the rate of coalescence is then slower) and the period between divergence times is small (since ILS is then more likely). The coalescent process is always occurring within populations while other processes such as gene flow between populations may or may not occur. As such, the MSC model is a first step to describe the differences between gene trees and species and other sources of discordance can be subsequently added to the models.

1.3. Bayesian Phylogenetics

In Bayesian phylogenetics, the goal is typically to infer the posterior probability of the tree and model parameters given the data. In Bayesian statistics, parameters are viewed as random variables, which contrasts with frequentist statistics where parameters are treated as fixed/unknowns (see Casella and Berger (2002) for more extensive background). Since parameters are viewed as random variables in Bayesian inference, there are probability distributions for each parameter. Using Bayes theorem, the posterior probability of the phylogenetic tree, T, and the model parameters, θ, given the data, $P(T, \theta|D)$, is equal to the probability of the data, D, given the tree and model parameters, $P(D|T, \theta)$, times the prior probability of the tree and model parameters, $P(T, \theta)$, divided by the marginal probability of the data, $P(D)$ (Yang and Rannala, 2012). In mathematical terms,

$$P(T, \theta|D) = \frac{P(D|T, \theta)P(T, \theta)}{P(D)}. \tag{1.1}$$

The data are typically a multiple sequence alignment. The model parameters will depend on the specific model used for inference, but usually include parameters of the model of sequence evolution and parameters specifying the branch lengths. The denominator of this equation, $P(D)$, is not tractable to calculate analytically because it involves a multidimensional integral over all of parameter space. However, this quantity is a constant, so Eqn. 1.1 can be rewritten as

$$P(T, \theta|D) \propto P(D|T, \theta)P(T, \theta). \tag{1.2}$$

This allows for the use of Markov chain Monte Carlo (MCMC) to approximate the posterior distribution by simulating samples from the distribution without the need to calculate the marginal probability of the data (see section 1.4).

1.4. Markov Chain Monte Carlo

Markov chain Monte Carlo (MCMC) is numerical technique used to approximate distributions. In Bayesian phylogenetics, MCMC is used to avoid the need to calculate the intractable denominator in the posterior probability (see section 1.3). The Metropolis algorithm (Metropolis *et al.*, 1953) for MCMC, which is standard in Bayesian phylogenetics, proceeds as follows. Using the same notation as section 1.3, first choose initial values for all parameter in the model and the tree, θ and T,

respectively. Propose a change to one or more parameters in the model, so the new parameters are θ'. Then calculate the acceptance ratio, α,

$$\alpha = \min\Big(1, \frac{P(D|T,\theta')}{P(D|T,\theta)} \times \frac{P(T,\theta')}{P(T,\theta)}\Big). \tag{1.3}$$

Accept or reject the move with probability α and record the parameter values. These steps are repeated proposing changes to the topology of the tree, T, in addition to the model parameters, θ, until the desired number of samples (see section 1.4.2) is obtained. The number of samples in each state is proportional to the posterior probability of the state. If the probability of proposing a change from state θ to state θ', written $q(\theta'|\theta)$ does not equal the probability of proposing the reverse change, i.e. a change from state θ' to θ, written $q(\theta|\theta')$, the Metropolis-Hasting algorithm (Hastings, 1970) must instead be used. In this case, a correction factor called the Hastings ratio, $\frac{q(\theta|\theta')}{q(\theta'|\theta)}$, is required and the acceptance ratio becomes

$$\alpha = \min\Big(1, \frac{P(D|T,\theta')}{P(D|T,\theta)} \times \frac{P(T,\theta')}{P(T,\theta)} \times \frac{q(\theta|\theta')}{q(\theta'|\theta)}\Big). \tag{1.4}$$

This accounts for the fact that if some moves are more likely to be proposed than other, some states would be visited more often simply because they are proposed more often.

The beginning of the MCMC is referred to as the burn-in. During this phase of the MCMC, the parameters are not drawn from the true posterior distribution because the chain has not had time to move away from the initial parameter values, which are chosen somewhat arbitrarily. When MCMC results are summarized, the burn-in is removed from the analysis. Since MCMC samples are autocorrelated, often the samples are thinned by only recording the MCMC samples every given number of iterations. This can greatly reduce output file sizes without significant impacts on inferences. The basic MCMC algorithm has been modified in many ways (see Robert and Casella (2004)).

1.4.1. Why MCMC Methods are Hard: Numerical Challenges. MCMC methods face a number of challenges for developers and users, especially when using complex phylogenetic models. Many aspects of MCMC algorithms, such as the proposal density or the initial parameter values,

are arbitrary but consequential, parameters are often correlated, and parameters may imposed constraints on each other.

1.4.1.1. *Initial Values.* The initial parameter values in an MCMC are arbitrary, but they may impact properties of the MCMC. Any set of parameters that fit the model and its constraints are permissible initial parameter values for an MCMC. However, initial parameters values that have very low posterior probability may lead to slow convergence to the posterior or a failure to converge.

1.4.1.2. *MCMC Proposals.* The choice of what types of MCMC proposals to implement is also somewhat arbitrary. Any move is allowed that both proposes parameters that are valid in the model (the constraints are met) and has a Hastings ratio that can be calculated to achieve detailed balance. For example, in a very simple case MCMC moves change only one parameter at a time. For a single continuous parameter, many possible simple proposal distributions exist, such as uniform, normal, or Laplace distributions. Each of these distributions also has parameter(s) that impact the mean and variance of the proposal density. The choice of proposal distribution and its parameters impacts the frequency with which moves are accepted in the chain (Yang and Rodríguez, 2013). Proposal distributions which propose larger changes to the parameter value, such as proposal distributions with larger variance, tend to lead to decreased frequencies of acceptance and higher correlation in the sampled values. If the proposal distributions often propose very small changes to parameter values, the moves are very likely to be accepted because the new posterior probability is very close to the posterior probability of the previous parameters. This can lead to very slow mixing of the chain because it takes many moves to achieve even moderate changes of the parameter values. Ideally, moves are accepted at an intermediate frequency. More complex moves change multiple parameters at once, leading to even more choices on how to construct the proposal and additional covariance parameters of the proposal density.

1.4.1.3. *Correlated Parameters.* Parameters are often correlated in the posterior distribution. For example, substitution rate estimates are correlated with divergence time estimates. In MSC models, the effective population size of a population is often correlated with the divergence time of the parent population (Rannala and Yang, 2003). Correlations between parameters are one reason moves are used that change multiple parameters; this allows for a correlation between parameters to be preserved, leading to more moves being accepted.

1.4.1.4. *Constraints Among Parameters.* Parameters also impose constraints on other parameters in the model. For example, a daughter node is always younger than its parent node and gene tree coalescent times of genes sampled from different species are always older than the tMRCA of the species. Constraints are another reason complex proposals are used; if only one parameter were changed at a time, this could allow only very small moves that do not violate the model assumptions. For example, with many loci, there are many gene trees in the species tree. To propose changes to the species tree divergence time, all of coalescent times between species must remain older than the divergence time. However, some coalescent times will be close to the divergence time, so only small moves would be allowed without jointly proposing changes to the coalescent times. Developing proposal moves that jointly update a parameter and the parameters that constrain it may greatly improve mixing in some cases. For example, Rannala and Yang (2003) developed complex MCMC proposals specifically designed to address mixing problems cause by constraints. However, these moves are often very complex and may propose large changes to the parameters, leading to large changes in the likelihood, which may decrease the probability of acceptance. Constraints may also induce correlations between parameters, such as between node ages of parent and daughter nodes.

1.4.1.5. *MCMC Mixing: Multiple Modes in the Posterior.* Challenges such as correlations and constraints among parameters can lead to poor mixing of the MCMC; the chain can get stuck in particular regions of parameter space or move slowly between regions of parameter space (Nascimento *et al.*, 2017). In this case, the MCMC is not sampling from the true posterior distribution, but only a portion of it. Sometimes these issues can be overcome by running the MCMC longer. However, this increases the computational cost and does not always work (Nascimento *et al.*, 2017).

1.4.1.6. *Computational Expense.* MCMC can be extremely computationally expensive, taking days or weeks for some analyses to run. The computational cost scales with the number of loci and the number of site patterns in the sequence alignment, meaning that larger dataset are more expensive to analyze. While Bayesian inference will always be slower and more computationally expensive than many ad hoc methods, efficient proposals can greatly decrease the required run time by decreasing the number of iterations required for the MCMC to converge.

1.4.2. MCMC Convergence. It can be difficult to determine when an MCMC has converged to the posterior distribution. Two main issues are whether the chain is sampling from the true posterior distribution and whether the chain has enough samples.

1.4.2.1. *Sampling from the True Posterior Distribution.* If the MCMC is densely sampling one region of parameter spaces, this does not mean that there is not another area of parameter space with much higher posterior probability. Rather, the chain may not have found other areas of parameter space with higher posterior probability. If the chain has converged it is sampling all regions of parameter space in proportion to their posterior probability. One common approach to check convergence is to run two or more MCMCs using the same data and priors, but starting with different initial parameters. Often initial parameters are random, but sometimes users may specify some parameters, such as the tree topology. In this case, the user may specify different starting tree topologies for each run (given the MCMC program is estimating the topology and not conditioning on a fixed topology). After running the MCMCs, the posterior distributions are compared. MCMCs that have converged should have very similar posterior distributions from independent runs. Trace plots, which show the parameter value on the y-axis and the MCMC iteration on the x-axis, can be used to look for unfavorable behavior in the MCMC. Large episodic changes in parameter values may indicate poor mixing, where the chain is stuck in part of parameter space but infrequently jumps to a new region. Colloquially, the trace plot should look like a "fuzzy-caterpillar", meaning the trace plot does not trend up or down, but has substantial variation about a constant median over the iterations.

1.4.2.2. *Sample Size.* Samples from MCMCs are not independent; samples from consecutive iterations are autocorrelated because moves are proposed from previous states of the chain. Due to the non-independence of MCMC samples, statistics have been proposed to define what would be the equivalent number of independent samples, known as the effective sample size (ESS), of each parameter. ESS is an estimate based on the autocorrelation between samples and describes the number of samples from a random sample that would have the same uncertainty as the samples from the MCMC (Geyer, 1992). To calculate the ESS, a few other statistics must first be calculated. The mean of a MCMC sample is used to estimate the mean of the posterior distribution. The mean

has variance

$$v_{\text{MCMC}} = v_{\text{IND}}[1 + 2(\rho_1 + \rho_2 + \dots)],$$

where v_{IND} is the variance of an independent sample from the posterior distribution with the same size as the MCMC from the MCMC and ρ_i is the correlation between the samples of the MCMC taken at i iterations apart (Nascimento *et al.*, 2017). The efficiency of an MCMC is defined by the ratio of the variance of independent samples to samples from the MCMC,

$$\text{Eff} = \frac{v_{\text{IND}}}{v_{\text{MCMC}}}.$$

An Eff of a half means that $2n$ samples are required to obtain a sample with the same variance as n sample drawn independently from the posterior distribution. The effective sample size is then

$$\text{ESS} = \text{Eff} \times n,$$

where n is the number of samples (Nascimento *et al.*, 2017). Larger ESS values indicate better sampling from the posterior distribution. However, there is not an agreed ESS value for determining convergence (Nascimento *et al.*, 2017). All of these tests of MCMC convergence are subjective; it may be very obvious when an MCMC has not converged, but less clear if it has converged.

1.4.3. Validation of MCMC methods. Due to the complexity of the models and proposals, the implementation of Bayesian phylogenetic programs requires careful validation. A basic check is to run the program without data (using a constant likelihood), which is equivalent to sampling from the prior distribution. The MCMC output can then be compared against the specified prior distribution on each parameter. Another check is to perform inference using increasing amounts of simulated data, such as longer sequences or more loci, depending on the program. As the amount of data increases, the inferred parameter values should center on the true parameter values used to simulate the data with decreasing credible interval sizes. Bayesian simulation is a more sophisticated check of a program's implementation (Flouri *et al.*, 2022). In Bayesian simulation, parameter values are drawn from the prior. Data are simulated using those parameter values, and the parameter values of the simulated data are inferred using the inference program. This is repeated many times and the posterior distributions from all replicates are combined. This mixture

distribution across posterior distributions should match the prior distributions for each parameter if the program is implemented correctly. Bayesian simulation checks both the implementation of the MCMC proposals and the calculation of the priors and likelihood.

1.5. Considerations when using Bayesian Methods

1.5.1. Strengths of Bayesian Methods. Despite the difficulties of Bayesian phylogenetics, there are many advantages. Likelihood based methods make full use of the information in the data and are consistent if the model is correct (Yang and Rannala, 2012). That is, as the amount of data increases to infinity, the estimates will converge to the true value. Likelihood based methods are also more efficient than other phylogenetic methods. A more efficient estimator produces an unbiased estimate with lower variance than other estimators (Yang and Rannala, 2012). Bayesian methods also yield easily interpretable credible intervals (sets). The true value should fall in the 95% credible interval (set) with frequency 95% given the model is correct and the data are informative (Huelsenbeck and Rannala, 2004). There is not a natural or obvious way to measure uncertainty in topology using maximum likelihood, whereas in Bayesian inference probabilities can be assigned to different tree topologies. In contrast, bootstrap values are used as a measure of confidence on nodes in the tree in maximum likelihood, but are difficult to interpret (Felsenstein and Kishino, 1993; Hillis and Bull, 1993). Moreover, many of the standard models used to estimate time calibrated phylogenies are not tractable with other approaches, such as maximum likelihood, without using heuristics (e.g. Sanderson, 2002).

1.5.2. Criticisms of Bayesian Methods. Bayesian methods have been criticized on several grounds. The important criticisms are: the prior distributions can be arbitrary; the model may be misspecified (having adverse effects on inference).

1.5.2.1. *Prior Models.* Bayesian methods require priors on all model parameters. Some see this as a weakness, since the inferences will be influenced by the choice of priors. Others see this as a strength, since models can be informed by outside information. Historically, prior models available were often based on tractability (conjugate priors), rather than biology. The use of MCMC offers an alternative to the use of conjugate priors, allowing for different prior distributions to be used that better match the biology. However, MCMC can have mixing issues with certain priors. For

users, specifying priors can be challenging, especially for complex models. Often researchers have little prior biological knowledge regarding some of the model parameters. Ideally, diffuse priors would be chosen for these parameters and the posterior distribution would be relatively insensitive to the choice of priors. However, in practice program defaults for priors are often used, which can have large impacts on inferences when the defaults are too informative and do not reflect realistic uncertainty in the parameter values. For example, using an exponential distribution as the prior on branch lengths, which was the default prior in the program MrBayes, led to an overly informative prior and produced inferred branch lengths that were unreasonably long (Rannala *et al.*, 2011). Similarly, the default prior in BEAST for the migration rates and the number of migration events for phylodynamic analyses is strongly informative and is biologically unrealistic in many cases, leading to poor inferences on pathogen dispersal rates (Gao *et al.*, 2023). However, the impacts of priors can be assessed by comparing the results using different priors. Often the shape of the prior distribution does not have a large impact on the results if the prior is sufficiently diffuse and the data are informative.

1.5.2.2. *Model Specification.* Bayesian inference assumes the model is correct and very unfavorable behavior can result when this assumptions is violated. A classic example is the star tree paradox. If data are generated under a star tree and Bayesian inference is used to infer the tree topology under a binary tree model, with increasing amount of data, the posterior distribution becomes increasingly concentrated on one of the binary trees, even though all the trees are equally wrong (Lewis *et al.*, 2005; Suzuki *et al.*, 2002; Yang and Rannala, 2005). This is a very clear case of model violation and solutions have been proposed to overcome these issues, such as allowing for the possibility of polytomies in the model. However, phylogenetic models can be very complex and it not always clear how model violations may impact inferences. With infinite data, when comparing two wrong models where one is less wrong, the less wrong model will be preferred. The wrongness of a model is usually measured with Kullback–Leibler divergence. However, with large but finite datasets, the more wrong model may be preferred with high confidence. In this case, Bayesian model selection is overconfident (Yang and Zhu, 2018). These issues are not unique to Bayesian inference and will arise for other likelihood based methods as well (Yang and Rannala, 2005).

1.6. Time Calibration of Phylogenies

Inferring a time calibrated phylogeny, even when the topology is assumed to be known, is a non-trivial task. When sequence data alone are used to infer a phylogeny using a likelihood based method, the branch lengths are usually estimated in units of expected number of substitutions. Since the expected amount of molecular evolution between two lineages depends on the product of the substitution rate and the divergence time, only the product of rate and time is identifiable without outside information. There are several approaches to separate rate and time. All are based on the idea of a molecular clock, which says that the amount of molecular evolution is approximately proportional to time (reviewed in Bromham and Penny, 2003). A simple option is to assume the substitution rate is known, allowing branch lengths in units of expected number of substitutions to be directly converted into calendar time. This is unrealistic as the substitution rate is never known and must always be estimated from outside sources. Additionally, rate variation among lineages is relatively common, especially for species with more ancient divergence times, which undermines this approach (reviewed in Ho, 2020). Another option is to use fossils or geologic events with "known" ages (estimated from outside sources) to calibrate the tree, which is typically done in a Bayesian framework. Geologic events may place bounds on the possible divergence times, but are not equivalent to node ages (Forest, 2009). Fossils need to placed on the tree using sparse morphological datasets. Even large morphological datasets contain much less information than typical molecular datasets and realistic models of morphological evolution are difficult to develop. It is also not usually known on exactly which part of the tree fossils should be placed. They could be tips on their own branch which went extinct, fall along branches in the tree (intermediate ancestors), or be on internal nodes (recent common ancestors). Similar to geologic events, fossils may constrain the ages of divergence events, but may not directly date a node in the tree.

A third option to calibrate phylogenies is tip dating, which uses the difference in branch lengths of sequences sampled through time to estimate a substitution rate and convert relative time to absolute time. With tip dating, branches sampled earlier in time have fewer opportunities for mutations to occur; no mutations can occur after sampling. Lineages sampled later may accumulate mutations during the period between sampling events. This leads to shorter branch lengths for sequences sampled earlier in time. Even simple methods, such as assuming a known substitution rate, will not

work with such datasets unless sample ages are explicitly accounted for. To have sufficient power to estimate the substitution rate, the difference in branch lengths must be reasonably large; this could result from either a high mutation rate or a long period of time between sampling events. The possibility of tip dating has long been recognized for HIV and other viruses (Li *et al.*, 1988; Rambaut, 2000). Modern tip dating methods are typically Bayesian methods and are widely used with viral data, especially RNA viruses due to their rapid evolution. The use of tip dating with eukaryotic DNA has been more limited, since only ancient samples potentially have a large enough difference in branch lengths to infer the substitution rate. With ancient DNA (aDNA), BEAST allows for tip dating to be combined with fossil calibrations (Suchard *et al.*, 2018), resulting in a phylogeny informed by both (e.g. van der Valk *et al.*, 2021).

1.6.1. Bayesian Inference of Gene Tree Node Ages. Bayesian phylogenetic models require a prior model on the tree topology and divergence times. Two major groups of explicit prior models that have been used are coalescent models (Drummond *et al.*, 2002) and birth-death models (Rannala and Yang, 1996). Some of these models accommodate tip-dating or fossil calibrations.

1.6.1.1. *Coalescent Based Models.* Though the Kingman coalescent originally described contemporary samples, the coalescent was extended to accommodate sequences sampled over time (Rodrigo and Felsenstein, 1999). Coalescent priors are extremely common in the analysis of population genetic tip dated sequences, and a wide variety of coalescent priors are available in BEAST, one of the most common tip dating programs (Suchard *et al.*, 2018). Coalescent based models that infer gene tree topologies but not species tree topologies are best suited for either single species or pathogen data.

Many complex models based on the coalescent have been developed to relax the potentially unrealistic assumption of a constant population size through time. A common parametric coalescent model is exponential growth (Slatkin and Hudson, 1991), which was generalized by Griffiths and Tavare (1994). Model testing can be used to evaluate whether one of these models fit the data better than a model of constant population size. Although more flexible than a constant population size model, these models still may not be flexible enough to accommodate realistic fluctuations in population sizes. Skyline plot models are a group of non-parametric coalescent models which allow for changes in population size through time. These models may be very useful if the question of

interest is specifically the demographic history of a population. The classic skyline plot model was not Bayesian (Pybus *et al.*, 2000), but it has been extended in a Bayesian framework (Drummond and Rambaut, 2007; Drummond *et al.*, 2005; Gill *et al.*, 2012; Minin *et al.*, 2008; Parag *et al.*, 2020).

1.6.1.2. *Birth-Death Based Models.* A separate group of prior models is based on the birth-death process. A Yule model is a pure birth process where every lineage branches at a rate λ (Yule, 1925). This is a special case of the birth-death model (Feller, 1939), where each lineage bifurcates at rate λ and dies at rate μ. Both of these models are used as priors for node ages, typically for species trees or slowly evolving bacteria or viruses. However, they require that all tree tips are contemporary (Nee *et al.*, 1994). These models have been extended to accommodate non-ultrametric trees, such as trees that including fossils or pathogen data.

The birth-death sequential sample (BDSS) model is a tip dating method that was developed for rapidly evolving populations, such as viruses or other pathogens (Stadler and Yang, 2013). Births, deaths, and sampling occur at constant rates for each lineage. Following Nee *et al.* (1994), at time present a fraction of all remaining lineages are sampled (Stadler and Yang, 2013). This method allows for the estimation of a time calibrated tree using samples continuously collected over time. The fossilized birth-death (FBD) process is a prior for trees with both extinct and extant taxa and is very similar to the BDSS model (Heath *et al.*, 2014). Lineages either speciate, become extinct, or become observed fossils at constant rates. Additionally, a proportion of lineages are sampled at time present. This can be used as a prior for total evidence dating (Gavryushkina *et al.*, 2016; Zhang *et al.*, 2015), which jointly infers the tree topology and branch lengths for extinct and extant species using morphological characters, sequences, and fossil calibrations. This draws on methods for tip dating, morphological evolution, and molecular clock dating (Pyron, 2011).

1.6.2. Bayesian Inference of Species Tree Node Ages. To directly infer species tree node ages, a MSC model is used. The MSC serves as a prior on the gene tree topologies and branch lengths conditional on the species tree topology and branch lengths. The gene trees are integrated over in the MCMC. A separate prior is used for the speciation times and species tree topology. This model is only implemented in a few inference programs including BPP (Flouri *et al.*, 2018) and *starBeast* (Douglas *et al.*, 2022), a BEAST2 package (Bouckaert *et al.*, 2019), and can only be used on a relatively small species tree. Only a constant population size demographic model

is available in BPP and starBEAST3, and there are a few demographic models in starBEAST2 (Ogilvie *et al.*, 2017).

1.7. Focal Questions in Bayesian Phylogenetic Tip Dating

1.7.1. HIV Latency. Bayesian phylogenetic tip dating models have been extensively applied to pathogen datasets to estimate divergence times between sequences (Kühnert *et al.*, 2011; Rieux and Balloux, 2016). However, they have not been used to estimate the times of other types of epidemiological events, such as latent integration times in HIV (though see recently published Jones and Joy (2023)). Latent integration is a process whereby HIV virions are integrated into a host cell, but enter a state of reversible inactivity where new virions are not produced. Latency effectively pauses sequence evolution of HIV. When latent lineages are sampled, this leads to branch lengths that are shorter than otherwise expected given the sampling times. This may allow for the estimation of the integration times of individual proviral sequences using a tip dating model. To date, estimation of integration times has used heuristic methods, rather than tip dating models (Abrahams *et al.*, 2019; Bruner *et al.*, 2016; Jones and Joy, 2020; Jones and Poon, 2017; Jones *et al.*, 2018; Pankau *et al.*, 2020). Understanding characteristics of latent lineages is of interest because the existence of a reservoir of latently infected cells is the reason current HIV therapies can treat, but not cure HIV.

1.7.2. Ancient DNA. Tip dating and MSC methods have largely been developed and implemented independently. This is reasonable for pathogen data, as the MSC is not typically applied to this kind of data. However, as aDNA sequence data is becoming increasingly available (e.g. Nielsen *et al.*, 2017; Orlando *et al.*, 2021; Ramos-Madrigal *et al.*, 2016; Rasmussen *et al.*, 2010; Soubrier *et al.*, 2016), dated sequences are also being used to investigate relationships between species (e.g. van der Valk *et al.*, 2021). There is currently only one program that allows for inference with the MSC using tip dated sequences (Douglas *et al.*, 2022). However, it requires that all samples from a particular species be sampled at the same time, which is unrealistic for aDNA samples. Both the failure to model sampling date with tip dated sequences and the failure to use the MSC with data from multiple species can lead to biases in divergence times estimates (Angelis and Dos Reis, 2015; Gillespie and Langley, 1979; Li *et al.*, 1988; Rieux and Balloux, 2016). Moreover, tip dating

provides an alternative to fossils when calibrating a phylogeny. As such, a fully general model and inference program with tip dating and the MSC should be developed to analyze multispecies datasets with aDNA.

CHAPTER 2

Bayesian Phylogenetic Inference of HIV Latent Lineage Ages Using Serial Sequences

HIV evolves rapidly within individuals, allowing phylogenetic studies to infer histories of viral lineages on short time scales. Latent HIV sequences are an exception to this rapid evolution, as their transcriptional inactivity leads to negligible mutation rates in comparison to non-latent HIV lineages. This difference in mutation rates generates potential information about the times at which sequences entered the latent reservoir, providing insight into the dynamics of the latent reservoir. A Bayesian phylogenetic method is developed to infer integration times of latent HIV sequences. The method uses informative priors to incorporate biologically sensible bounds on inferences (such as requiring sequences to become latent before being sampled) that many existing methods lack. A new simulation method is also developed, based on widely-used epidemiological models of within-host viral dynamics, and applied to evaluate the new method– showing that point estimates and credible intervals are often more accurate than existing methods. Accurate estimates of latent integration dates are crucial in relating integration times to key events during HIV infection, such as treatment initiation. The method is applied to publicly available sequence data from 4 HIV patients, providing new insights regarding the temporal pattern of latent integration.

2.1. Introduction

A major obstacle to the development of a cure for HIV has been the presence of latently infected cells. HIV is a retrovirus that integrates its genome into the host genome. During latent infection, the integrated provirus is in a reversible state of transcriptional inactivity. Latently infected cells are not targeted by current treatment methods, namely antiretroviral therapy (ART). Consequently, treatment must be continued for life or reactivation of latent cells will lead to a rapid rebound in viral load and disease progression (Davey *et al.*, 1999). A detailed understanding of the dynamic

processes of seeding, reseeding, and decay of the latent reservoir through the inference of latent integration dates for individual proviruses will allow researchers to better understand the nature of the reservoir as they work toward a cure for HIV.

HIV infects immune cells, specifically CD4+ cells, such as helper T cells and macrophages. Most infected cells die quickly (Ho *et al.*, 1995; Wei *et al.*, 1995). In contrast, memory T cells have a long half-life of 4.4 years and can thus establish a latent reservoir for HIV (Siliciano *et al.*, 2003). Memory T cells may be infected directly or an activated T cell may revert back to a quiescent state (Dufour *et al.*, 2020). Latently infected memory T cells can be activated by antigens, leading to the activation of the HIV provirus (Siliciano and Greene, 2011). Effective ART prevents infections of new host cells but does not prevent previously infected cells from producing virions. HIV can therefore persist hidden in memory cells for decades, even with effective ART (Siliciano *et al.*, 2003).

The latent reservoir is initially formed within days of infection and continues to be reseeded over time (Chun *et al.*, 1998; Verhofstede *et al.*, 2004; Whitney *et al.*, 2014). However, the extent to which the composition of the reservoir changes over time is unclear. There is strong evidence that there is not ongoing cycles of viral replication during ART (Dinoso *et al.*, 2009; Hatano *et al.*, 2011; McMahon *et al.*, 2010), so it is very unlikely the HIV reservoir is replenished from ongoing replication during ART. Some studies concluded that the latent reservoir that exists during ART is mostly seeded shortly before treatment initiation (Abrahams *et al.*, 2019; Brodin *et al.*, 2016; Pankau *et al.*, 2020), while others have concluded that the reservoir is continuously seeded until treatment initiation (Jones *et al.*, 2018). However, some of these results are difficult to interpret as a variety of mechanisms could account for these patterns. The timing of the formation of the latent reservoir is ultimately an empirical question that can be studied in multiple ways. Experimental techniques, such as the use of quantitative viral outgrowth assays (QVOAs), allow researches to obtain sequence data from individual cells known to be latent. In addition to further experimental work, reconstructing the ages of latent lineages can in principle be done by analyzing the patterns of variation observed among sampled sequences and applying phylogenetic methods designed to estimate sequence divergence times with serial sequence samples (Abrahams *et al.*, 2019; Bruner *et al.*, 2016; Jones and Joy, 2020; Jones and Poon, 2017; Jones *et al.*, 2018; Pankau *et al.*, 2020).

The focus of this paper will be the development of new statistical and computational methods to accurately date the integration times of sampled latent sequences.

A variety of heuristic methods have been developed to estimate integration times using a combination of RNA sequences from serially sampled actively replicating sequences and RNA or DNA from putative latent sequences. All methods rely on a fixed estimate of the gene tree topology for the HIV sequences and some require branch lengths. Jones et al. developed a distance method that used linear regression (LR) to estimate the mutation rate from root-to-tip distances and sampling dates for non-latent sequences. This mutation rate is then used to estimate the latent integration dates (Jones *et al.*, 2018). This method relies on a molecular clock, and is not used if the clock is rejected. Jones and Joy developed a related method, estimating mutation rate in the same way but estimated internal node ages using a maximum likelihood (ML) approach using a specified mutation rate (Jones and Joy, 2020). To et al. developed a distance method using a least squares (LS) approach to estimate mutation rates and date internal nodes and tips with unknown ages To *et al.* (2016). Their method requires the sequence length for estimating confidence intervals, but not the alignment. It was designed for extremely large phylogenies, but is applicable to HIV latency datasets as well. Abrahams et al. used multiple heuristic methods to date latent sequences. In one method, the distance from the closest sequence to the latent sequence, d, is determined, and the age of the latent sequence is assigned based on the sample time of the majority of sequences within $2d$ of the latent sequence (Abrahams *et al.*, 2019). A similar method traverses the tree from the latent sequence toward the root of the tree until a node with 90% bootstrap support is found with at least one pre-treatment sequence. Then a latency time is assigned based on the most common sampling time of the pre-treatment sequences descendant from the well supported node (Abrahams *et al.*, 2019). The two methods used by Abrahams et al. may be very sensitive to the number of sequences sampled and the sampling times. Simulation studies suggest that LS may out-perform all of these methods (Jones and Joy, 2020; To *et al.*, 2016). An alternative to these existing methods could be developed based on established parametric phylogenetic models that use tip dating for estimating and calibrating phylogenies of viral data, and are potentially more accurate (Rambaut, 2000; Stadler and Yang, 2013).

It has been difficult to evaluate the statistical performance of current methods for inferring integration times of latent HIV since existing simulation methods are biologically unrealistic. During the acute phase of infection, viral load grows exponentially shortly after infection, peaking within several weeks (Deeks *et al.*, 2015). Then the viral load falls one to two orders of magnitude before reaching a quasi-steady state. During this chronic phase of infection, the viral load remains relatively unchanged or rises only slowly until the onset of AIDS. In contrast, simulation methods that have been used to evaluate methods for dating integration events largely ignore the underlying population dynamics of HIV. Some assume a constant rate birth-death process while others use a compartmental model with logistic growth (Jones and Joy, 2020; Jones *et al.*, 2018). Epidemiologists use more complex models, typically ordinary differential equations (ODEs), to describe HIV viral dynamics (Nowak and Bangham, 1996; Perelson and Ribeiro, 2013; Phillips, 1996). These models produce population trajectories that more closely match empirical observations, especially during acute infection, but the models have yet to be used in simulations to generate within-host HIV sequence data. The time period of acute infection is known to be important in establishing the latent reservoir (Chun *et al.*, 1998), and this peak dynamic should be incorporated into simulation methods used to test inference methods aimed at estimating latency times.

We propose a full likelihood Bayesian inference method to infer the latent integration date of HIV sequences, conditional on the phylogenetic tree topology. The method assumes it is known *a priori* which sequences are derived from latent proviruses and which are from non-latent viruses. This is possible when sequencing RNA from untreated patients and using QVOAs which stimulate the production of virus from latently infected cells. Additionally, we develop a simulation method based on existing viral dynamic models of HIV to test the performance of the inference method. The simulation model is parameterized using estimates from empirical datasets that produce realistic viral population dynamics (Stafford *et al.*, 2000).

2.2. Inference Methods

A new program, HIVtree, was developed by modifying an existing program, MCMCtree, to infer latent integration dates (Stadler and Yang, 2013). MCMCtree is a Bayesian phylogenetic inference program which estimates a time calibrated tree using viral sequences with serial samples

given a fixed tree topology. It uses Markov chain Monte Carlo (MCMC) to estimate the model parameters. HIVTREE incorporates additional parameters, the latent integration times, into the model. The program also estimates the originally defined parameters in MCMCTREE, including substitution model parameters, substitution rate, and the internal node ages.

HIVTREE assumes *a priori* that certain sequences are known to be latent while others are known not to be. Every sequence must also have a known sample date. In addition, every latent sequence has an unknown latent integration date. The youngest possible latent integration date is the sample time, and internal nodes cannot be latent. There is an optional bound on the oldest possible latent integration time, which could correspond to the oldest possible infection time. The model assumes that latent lineages have a mutation rate of zero, and all other lineages follow strict molecular clock. For calculating the likelihood, the latency time is treated as if it were the sample date for a non-latent lineage. This acts to reduce the tip age to be the time the sequence became latent (Fig. S4).

2.2.1. Markov Chain Monte Carlo (MCMC). HIVTREE adds an additional step to the MCMC to estimate the latent times. In MCMCTREE, proposals to non-root internal node ages are bounded above by the age of the parent node and below by the age of the oldest daughter node. A new time for each internal node is proposed within these bounds, the acceptance ratio is calculated, and the move is either accepted or rejected (Stadler and Yang, 2013). In HIVTREE, in addition to bounds on nodes, latent times are bounded above by the age of the parent node and below by the sample time. This ensures that the sequence becomes latent before it is sampled and that internal nodes cannot be latent. If the optional bound on latent integration times is used, the younger of the parent node age and the bound is used as the bound. Similar to MCMCTREE, for each latent time, a move is proposed within these bounds, the acceptance ratio is calculated, and the move is either accepted or rejected (Fig. 2.1). Other than the difference in bounds, the proposal moves for the internal nodes and the latency times are identical. For the mixing step, the latency time is treated as equivalent to the sample date. The mixing step was not modified from MCMCTREE (Stadler and Yang, 2013).

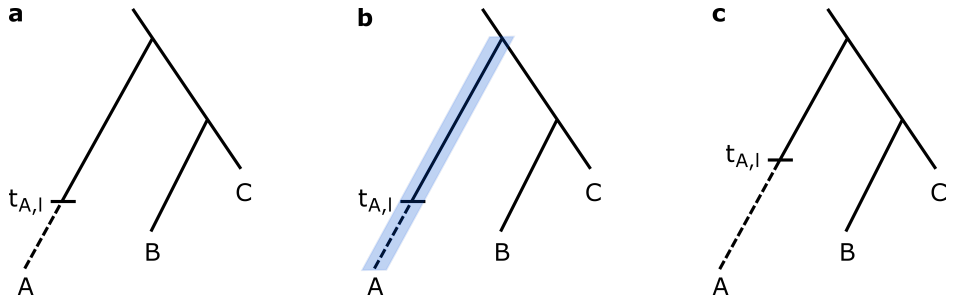


FIGURE 2.1. Proposal steps in the MCMC for latency times. Tips B and C correspond to non-latent sequences. At some time in the past, $t_{A,l}$, lineage A became latent. The dashed line shows when the lineages was latent. (a) Starting from the current latent time, (b) a new time can be proposed anywhere between the sample time and the age of the parent node, shown in blue. (c) Once a time is proposed, the move can be accepted or rejected. In this case, the move is accepted and the time is updated. For the calculation of the likelihood, the branch lengths correspond to the length of the solid lines only.

2.2.2. Prior on Distribution of Times. Two new root age priors were implemented in HIVTREE. HIVTREE and MCMCTREE both require the user to specify the priors in backward time. The time of the last sample is considered to be time zero, and earlier times are positive. The programs also require a specification of a time unit transformation. For example, consider HIV data with the sample times specified in days. A time unit of 1000 days means that 0.365 is equivalent to a year in the prior specification. A shifted gamma prior, $\Gamma(\alpha, \beta)$, is implemented as the root age prior. The distribution is shifted by adding the earliest sample time to the variable. This ensures there is no density for a root age younger than the sample ages. The gamma distribution parameters must also be chosen with the time unit transformation going backward in time. An option for a more informative prior is a uniform prior with narrow hard bounds (zero tail probability), $\mathrm{U}(a, b)$. There is no explicit prior on the internal nodes ages which is equivalent to a uniform prior on the possible node ages given the constraints from the sampling dates and the root age. This is in contrast to MCMCTREE, which uses a birth-death-sequential-sampling prior (Stadler and Yang, 2013). Since the sampling prior is not explicit and the rank order of the nodes and the constraints jointly determine the prior, the MCMC must be run without data in order to recover the prior for the internal nodes, latency times, and root age. The distribution of the root age when the MCMC is run without data will not be equivalent to the user specified prior (Fig. 2.2). This results from the lack of an explicit prior on the internal nodes and latency times and from not explicitly

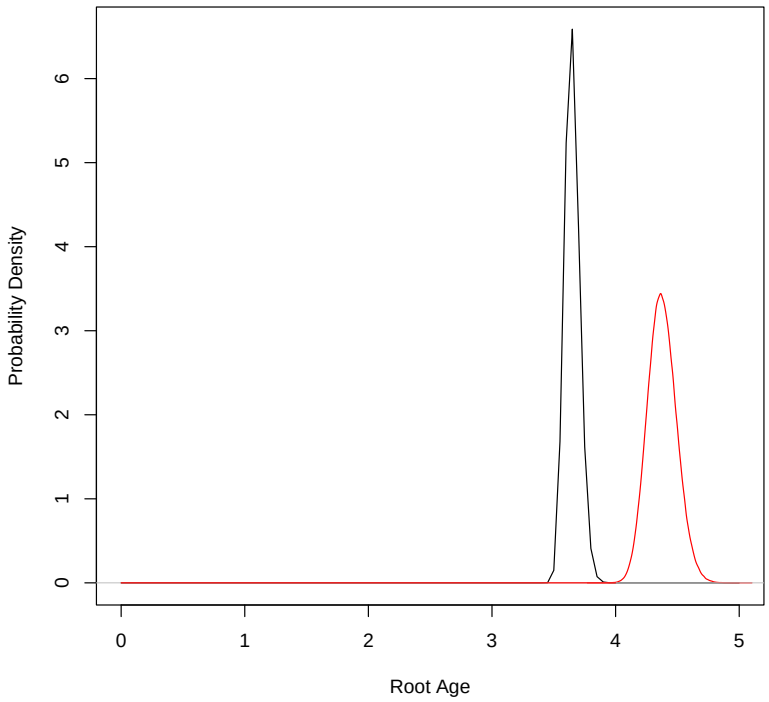


FIGURE 2.2. The user input prior for the root age is not the same as the prior determined by running HIVTREE without data. The black line shows the user input root age prior of Gamma(36.5, 100) on a tree with a last sample time of 3285 days before present with a time unit of 1000. This gives as mean root age of 3.65 in the time units used in HIVTREE. This is the same as all of the simulated trees in our analyses. Using a simulated dataset for C1V2, a tree topology was inferred with RAXML-NG and outgroup rooted. This tree was used to run HIVTREE under the prior. The red line shows the results, in which the root age is older than the user input prior.

conditioning on the tip ages (described below). This effect is similar to constraints imposed by fossil calibrations (Rannala, 2016). The mean root age will be older than the expectation of the prior distribution. The parameters of the gamma distribution can be modified to achieve a desired mean and variance for the root age. Using a uniform prior with a wide interval is discouraged due to this effect (an induced prior age of the root that is very old).

2.2.3. Explanation of Bias in the Root Age Prior. There are two distinct causes underlying the difference between the user specified prior for the root age and the induced prior on

the root age obtained by running HIVTREE under the prior (without data). One cause is that the distributions of the latency times, internal nodes and the root age are not independent. The "constant factor" prior applied to the latency times and internal nodes in our initial implementation does not explicitly condition on the root age. The other cause is that the prior does not explicitly condition on the tip dates.

2.2.4. Non-independence of Node Ages. To understand the first cause of the bias in the root age prior, we first consider a simple case without tip dating.

2.2.4.1. *Theory.* Consider a labeled history with n tips where all tips are contemporary and l latent samples. Define the root age to be T and rank order the nodes and latent sampling times, we subsequently refer to either node times or latent sampling times as events. Let x_i be the difference of height between events i and $i-1$, $i \in (2, \ldots, n+l)$, where $i = n+l$ denotes the tips, $i = n+l-1$ is the youngest event, etc (see Fig. 2.3). We assume that the unordered events have a uniform distribution on the interval $(0, T)$. Suppose that $T = 1$, then the joint density of the x_i for a set of rank ordered events is

$$f(x_2, \ldots, x_{n+l} | T = 1) = \Gamma(n + l - 1),$$

which integrates to 1 over the simplex as required. Note that if $T = 1$ is fixed then the prior ratio is 1. Now suppose that $T \neq 1$, then conditioning on T we can transform the node ages as

$$y_i = x_i T.$$

The inverse is

$$x_i = y_i / T,$$

and the Jacobian matrix has positive diagonals with

$$\frac{\partial x_i}{\partial y_i} = \frac{1}{T}.$$

We omit x_2 because it is not an independent random variable and is determined by the remaining variables as

$$x_2 = 1 - \sum_{i=3}^{n+l} x_i.$$

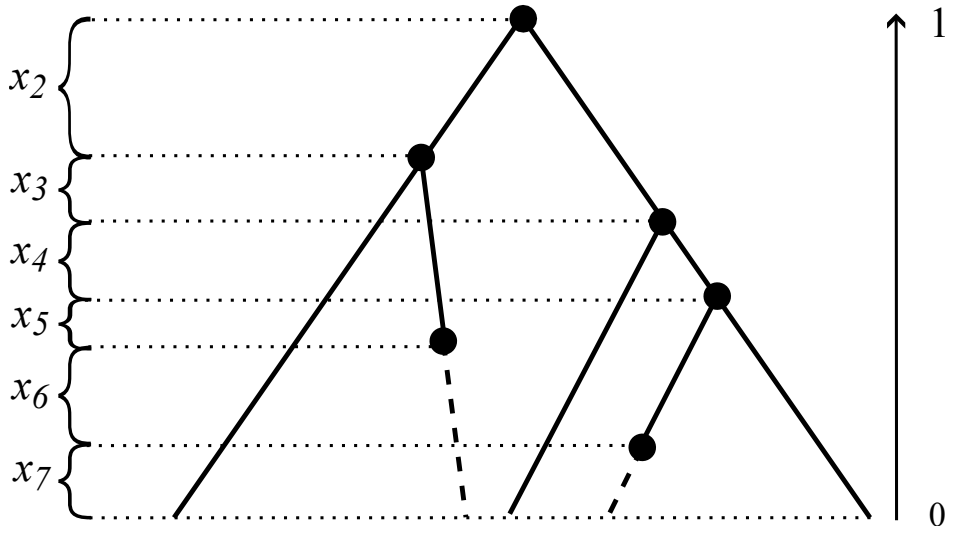


FIGURE 2.3. The tree has five samples, two of which are latent. The latent samples are indicated by the dashed lines in the tree. The latent integration time are shown by the transition to a solid line within a branch. The x_i are shown for this labeled history. Since there are 5 samples and 2 latent times, i ranges from 2 to $2+5=7$. Note that $x_2 = 1 - x_7 - x_6 - x_5 - x_4 - x_3$.

The determinant of the Jacobian is then

$$\left| \frac{1}{T^{(n+l-2)}} \right|,$$

and the conditional probability density of $\mathbf{x} = \{x_i\}$ given T is

$$f(\mathbf{x}|T) = \frac{\Gamma(n+l-1)}{T^{(n+l-2)}}.$$

Suppose that T follows a gamma density

$$g(T|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} T^{\alpha-1} \mathrm{e}^{-\beta T}.$$

The joint probability of T and $\mathbf{x}$ is

$$h(\mathbf{x},T|\alpha,\beta) = g(T|\alpha,\beta) f(\mathbf{x}|T) = \beta^\alpha T^{(\alpha-n-l+1)} \mathrm{e}^{-\beta T} \frac{\Gamma(n+l-1)}{\Gamma(\alpha)}.$$

2.2.4.2. *Impact of the Prior Ratio on the Root Age Bias.* If we were to integrate over $\mathbf{x}$ on the simplex defined by $x_i \in (0,T)$ without using the probability density function, the integral would

be

$$\int_{\mathbf{x}} d\mathbf{x} = \frac{T^{(n+l-2)}}{(n+l-2)!}$$

Now suppose that we treat $f(\mathbf{x}|T) = 1$ as in our "constant factor" prior, and normalize the resulting density in the MCMC. This is equivalent to using the distribution

$$h^*(\mathbf{x}, T|\alpha, \beta) = g(T|\alpha, \beta)\frac{1}{C},$$

where

$$C = \int_T \int_{\mathbf{x}} g(T|\alpha, \beta) d\mathbf{x} dT = \beta^{(2-l-n)} \frac{\Gamma(\alpha + n + l - 2)}{\Gamma(\alpha)\Gamma(n + l - 1)}.$$

The marginal density of T is then

$$g^*(T|\alpha, \beta) = \int_{\mathbf{x}} g(T|\alpha, \beta)\frac{1}{C} d\mathbf{x} = \frac{\beta^{(n+\alpha+l-2)} \mathrm{e}^{-\beta T} T^{(n+l+\alpha-3)}}{\Gamma(n + l + \alpha - 2)}.$$

The mean of this density is

$$\mathbb{E}(T) = \frac{\alpha + n + l - 2}{\beta},$$

and the variance is

$$Var(T) = \frac{\alpha + n + l - 2}{\beta^2}.$$

Thus both the mean and variance of T increase with increasing n. For $n = 2$ and $l = 0$ the results match the gamma density as expected. In the prior ratio when proposing changes to T all terms cancel that do not involve T and we are left with

$$T^{(n+l+\alpha-3)} \mathrm{e}^{-\beta T}.$$

If this is multiplied by the Jacobian determinant we are left with the correct density in terms of T,

$$T^{(n+l+\alpha-3)} \mathrm{e}^{-\beta T} \times \frac{1}{T^{(n+l-2)}} = T^{(\alpha-1)} \mathrm{e}^{-\beta T}.$$

If $T*$ is the proposed root age, the prior ratio would be multiplied by $(T^{(n+l-2)}/T*^{(n+l-2)})$.

In the model used in HIVTREE by default, the Jacobian terms for the prior ratio were not used in the analyses for this paper. Instead, the prior was adjusted to produce a reasonable induced prior. The prior with Jacobian terms is available as an option in the program.

2.2.5. Tip Dates. Analytical results are not available for the distribution of the root age conditional on the tip dates when all the dates are not contemporary. However, the distribution of the root age can be numerically estimated with rejection simulation. Consider a three tip tree with one latent time, such as shown in Fig. 2.1a. Let T be the root age and t_{BC} be the age of the node that is parent to nodes B and C. Let t_A, t_B, and t_C be the sample times of nodes A, B, and C, respectively. To simulate the prior distribution of the node ages and latency times, draw a root age from the user specified prior distribution. Then draw the latency time, $t_{A,l}$ from $U(0, T)$. Draw t_{BC} from $U(0, T)$. If $t_{A,l} < t_A$ or $t_{BC} < \max(t_B, t_C)$, reject the sample. Otherwise, store the times and repeat the procedure until the desired number of samples is obtained.

2.2.6. Comparison between HIVtree results using Jacobian Prior Ratio and Rejection Simulations. To show that the combination of (1) tip dates and (2) dependencies among node ages, latency sampling times and the root age, are the two factors causing the induced prior on root age to differ from the specified prior the calculation of the prior ratio was used as an option in HIVTREE. HIVTREE was run without data using the tree topology in Fig. 2.1a with $t_A = 300$, $t_B = 100$, and $t_C = 400$. The root age prior was a shifted gamma with $\alpha = 5$ and $\beta = 10$. The tip date unit was 1000. The rejection simulation procedure was used to sample from the prior distribution of the node ages and latency times.

2.2.7. Bayesian Simulation. Bayesian simulation was conducted to validate our implementation of HIVTREE. Bayesian simulation draw parameter values of a model from their prior distributions, simulate data with those parameter values, and then infer the posterior distribution of the parameters from the combined data. The theoretical expectation is that with a large number of replicates, the combined samples from the posteriors of all the replicates will match the prior distributions of the parameters (Flouri *et al.*, 2022). A four-tip, asymmetric tree was used with one latent time (Fig. 2.9). 4,000 simulations were conducted. For each simulation, three times were drawn from U(0, 3650) and rank ordered. The mutation rate was drawn from $\Gamma(2, 200)$ and divided by 1,000 to get a per day mutation rate. The times and mutation rate were used to simulate sequences of 100 base pairs under a Jukes-Cantor model the DNA simulation program described in this manuscript. The sequence at the root was drawn from the stationary distribution. HIVTREE

was run using a root age prior of U(3.6499, 3.6501), tip date units of 1000, and Jukes-Cantor substitution model. The mutation rate prior was $\Gamma(2, 200)$. The burnin was 100,000 iterations and 15,000 samples were taken with 1 sample every 25 iterations. Two MCMCs were run for each of the 4,000 simulated datasets. Convergence was checked by comparing the means and the 95% equal tail probability credible sets between the two replicate MCMCs. If the means differed by less than 0.01, which is 10 days, or the 95% credible set bounds differed by less than 0.10, which is 100 days, for all of the time parameters, the run was considered to have converged. Runs that did not converge were excluded from the figures. The results of all of the MCMCs were combined and kernel density estimation was used to summarize the distribution using the kdensity function in the R package kdensity (Moss and Tveten, 2019).

2.2.8. Combining Inferences Across Genes. HIVTREE only allows single locus inferences and assumes no recombination within a locus. However, recombination is common in HIV, meaning the whole genome cannot be analyzed assuming a single gene tree topology. However, the entire HIV genome is incorporated in the host cell genome at the same time, meaning different regions of genome share the same latent integration times. Let $X = \{x_i\}$ be sequence data for n loci, where x_i are sequence data at locus i. Let T be a latency time that is shared across loci. The remaining parameters of the gene tree may be different due to recombination between loci. The posterior density of T is

$$f(T|X) = \frac{P(X|T)f(T)}{\int P(X|T)f(T)dT}.$$

If we ignore the correlation between gene trees due to limited recombination and treat the loci as independent, as is generally done in phylogenetics, the posterior density can be written as

$$f(T|X) = \frac{\prod_{i=1}^{n} P(x_i|T)f(T)}{C_A},$$

where C_A is the marginal probability of the data (which is a constant),

$$C_A = \int \prod_{i=1}^{n} P(x_i|T)f(T)dT.$$

We want to calculate the posterior probability of T for each locus separately using MCMC and subsequently combine them to obtain a posterior density for all the loci. To do this we formulate the above equation as a product of the marginal posterior of T for each locus,

$$f(T|X) = \prod_{i=1}^{n} \left[\frac{f(T|x_i)}{f_i(T)} \right] \times f(T) \times \frac{\prod_{i=1}^{n} C_i}{C_A}, \tag{2.1}$$

where $f_i(T)$ is the prior on T for the ith locus and $f(T)$ is the desired prior for the combined posterior. The last term is a proportionality constant that insures the posterior density integrates to 1. C_i is the marginal probability of the data for an individual gene,

$$C_i = \int P(x_i|T) f(T) dT.$$

A simple example illustrating this general approach to combine posteriors using a normal distribution is provided in Appendix A.

In our analyses, n independent MCMC analyses are run (with and without using the likelihood) and kernel density estimation is used to estimate $f(T|X_i)$ and $f_i(T)$, respectively, for $i = 1, \ldots, n$. The estimated kernel functions are then used to evaluate equation 2.1 up to an unspecified proportionality constant (see methods). Simulations were used to evaluate the performance of this approach to combine posteriors.

This method may be used on regions of the genome that are not complete genes. For simplicity, the term gene tree will be used to describe a phylogeny inferred using data from any region of the genome, but genomic region will be used rather than gene to describe a part of the genome that may not produce a complete functional product.

2.3. Simulation Design

A stochastic simulation based on existing ordinary differential equations was developed to simulate tree topologies of sampled latent and active HIV sequences.

2.3.1. Deterministic Model. Here we describe the deterministic model of HIV population dynamics that will serve as the large-population analog of our stochastic model (see below). Let $T(t)$ be the number of uninfected target cells at time t. Let $T^*(t)$ be the number of productively infected cells at time t. Let $L(t)$ be the number of latently infected, replication-incompetent cells at

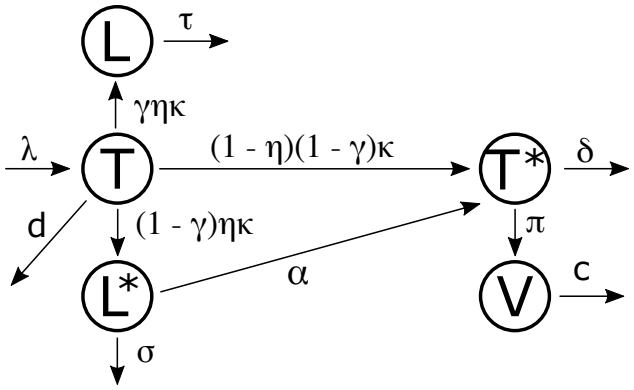


FIGURE 2.4. Within-host viral dynamics model

time t. Let $L^*(t)$ be the number of latently infected, replication-competent cells at time t. Let $V(t)$ be the number of virions at time t (Fig. 2.4). Actively infected target cells that are replication-incompetent are not modeled. Define λ to be the rate at which uninfected target cells are produced and d to be the per cell rate at which they die. Let δ be the per cell rate at which actively infected cells die. Latent replication-competent cells and replication-incompetent cells die at constant per cell rates of σ and τ, respectively. Let γ be the proportion of newly infected cells that are replication-incompetent. Let η be the proportion of newly infected cells that are latently infected and $(1-\eta)$ be the proportion of newly infected cells that are actively infected. Let κ be the rate constant for target cells/virion pairs resulting in infected cells. Productively infected cells must be replication-competent and are produced at a rate equal to product of the rate constant κ, the number of virions, the number of uninfected cells, the proportion of cells that are replication-competent, and the proportion of cells that are actively infected. The rate of production of latent replication-competent cells is calculated similarly, except that the proportion of cells that are latently infected is used rather than the actively infected population. For replication-incompetent latent cells, the rate of production is equal to the product of the rate constant κ, the number of virions, the number of uninfected cells, the proportion of cells that are replication-incompetent, and the proportion of cells that are latently infected. When an infected cell is produced, an uninfected cell is lost, since the uninfected cell becomes the infected cell. This is true for actively infected cells and both types of latently infected cells. T cells that are infected with replication-incompetent virus are not explicitly tracked and remain uninfected T cells. This assumes that replication-incompetent

infection does not change the death rate of the T cells and allows for super infection (infection with more than one provirus).

Latent replication-competent cells can reactivate and become actively infected cells. This occurs at a constant per cell rate of α. HIV virions, V, are produced at a rate proportional to the concentration of actively infected cells, with rate constant π. The virions are cleared at a constant per virion rate of c. This model is shown graphically in Fig. 2.4 and gives the following set of equations:

$$\begin{aligned}
\frac{dT(t)}{dt} &= \lambda - dT(t) - (1-\gamma(1-\eta))\kappa T(t)V(t) \\
\frac{dT^*(t)}{dt} &= (1-\eta)(1-\gamma)\kappa T(t)V(t) - \delta T^*(t) + \alpha L^*(t) \\
\frac{dV(t)}{dt} &= \pi T^*(t) - cV(t) \\
\frac{dL^*(t)}{dt} &= (1-\gamma)\eta\kappa T(t)V(t) - \alpha L^*(t) - \sigma L^*(t) \\
\frac{dL(t)}{dt} &= \gamma\eta\kappa T(t)V(t) - \tau L(t)
\end{aligned}$$

The solutions to these equation are obtained by numerical analysis using the function ode in the R package deSolve (Soetaert *et al.*, 2010) and the parameters in Table 2.6.

If an individual begins antiretroviral therapy (ART), the infection of new cells is prevented. In the model, this corresponds to κ becoming zero, assuming ART is perfectly effective.

2.3.2. Stochastic Model. Viral dynamics were modeled using a continuous-time Markov chain with instantaneous rates as previously described in the deterministic model. For example, let A be the event that a birth of an uninfected cell occurs in the time interval Δt. Then,

$$P(A) = \lambda \Delta t$$

The process is modeled as a jump chain. Only one event can occur in a small interval Δt, and the number of viruses, or of any cell type, can only change by one in that interval. The waiting time between birth events of uninfected cells is exponentially distributed with mean waiting time $\frac{1}{\lambda}$. The instantaneous rates and waiting time between other events are determined similarly. The

total rate of events, $R(t)$, is given by the sum of the rates of all possible events.

$$\begin{aligned} R(t) =& \lambda + (d + (1 - \gamma(1 - \eta))\kappa V(t))T(t) + (\delta + \pi)T^*(t) \\ &+ (\alpha + \sigma)L^*(t) + \tau L(t) + cV(t) \end{aligned}$$

The waiting time between any event is exponentially distributed with mean $\frac{1}{R(t)}$. Given that an event occurs, the probability the event was a birth of an uninfected cell, for example, is given by the ratio of the rate of birth events of uninfected cells and the total rate of events, $\frac{\lambda}{R(t)}$. The probabilities of other events are determined similarly. The simulation allows antiretroviral therapy (ART) to be initiated, leading to a decrease in the parameter κ. ART is assumed to be perfectly effective, and thus $\kappa = 0$. All other parameters remain the same. The initiation time of ART is prespecified. Methods for simulating continuous time Markov Chains are well established, e.g., Ross (1997), and our simulation of the birth-death process follows the standard simulation approach pioneered by Kendall Kendall (1950).

2.3.3. Simulation of Tree Topologies. The stochastic model was implemented as a C program. In the program, the parent daughter relationship of all of the viruses in a tree structure is tracked. The cell or virus type (e.g. T^*, V, L, or L^*) and amount of time latent in each branch is also tracked. The simulation is initialized with a single actively infected cell. Each time a virus is born, an actively infected cell is randomly selected to branch into two daughter lineages. One lineage is an actively infected cell and the other an active virus. Each time a virus or cell dies, an existing virus or cell of that type is randomly removed from the tree. When a virus latently infects a cell, a virus is randomly chosen to branch into an infected cell and a virus. This is designed to follow the conventional ODE models, even though a single virus cannot infect multiple cells in real systems. This is likely inconsequential, since the waiting time for a virus to die is short, and thus the probability a virus infects multiple cells is very small. Replication-competent latent viruses may be reactivated, meaning they become actively infected cells. Extinction is considered to be analogous to a failure to establish infection. In this case, the simulation is restarted. At pre-specified times, a pre-specified number of active viruses and latently infected cells are sampled.

FIGURE 2.5. Simulation parameters

Parameter	Description	Value	Citation
λ	Birth rate of uninfected cells	170 $\frac{\text{cell}}{\text{mL}\times\text{day}}$	(Stafford *et al.*, 2000)
d	Death rate of uninfected cells	0.017 $\frac{1}{\text{day}}$	Stafford *et al.* (2000)
κ	Transition rate from uninfected to actively infected cells	8.0×10^{-7} $\frac{\text{mL}}{\text{virion}\times\text{day}}$	Stafford *et al.* (2000)
δ	Death rate of actively infected cells	0.31 $\frac{1}{\text{day}}$	Stafford *et al.* (2000)
π	Viral birth rate	730 $\frac{\text{virions}}{\text{cell day}}$	Stafford *et al.* (2000)
c	Viral clearance rate	3 $\frac{1}{\text{day}}$	Stafford *et al.* (2000)
η	Proportion of newly infected cells that are latent	1.16×10^{-3}	Chun *et al.* (1997)
α	Rate of activation of replication-competent, latent cells	5.7×10^{-5} $\frac{1}{\text{day}}$	Hill *et al.* (2014); Luo *et al.* (2012)
γ	Proportion of viruses that are defective	0.95	Bruner *et al.* (2016)
σ	Death rate of latent, replication-competent cells	5.2×10^{-4} $\frac{1}{\text{day}}$	Peluso *et al.* (2020)
τ	Death rate of latent, replication-incompetent cells	1.1×10^{-4} $\frac{1}{\text{day}}$	Peluso *et al.* (2020)

FIGURE 2.6. The parameters from Stafford *et al.* (2000) are for patient 7. κ is typically estimated as the rate constant of new infections of replication-competent cells, which is $\kappa(1-\gamma)(1-\eta)$ in this model. Thus, the empirical estimates of κ, as presented in the table, is divided by $(1-\gamma)(1-\eta)$ to obtain the parameter value used in the model.

Replication-competent and incompetent cells are not distinguished during sampling. Sampling is equivalent to a death event for all sampled lineages.

2.3.4. Parameter Values. Parameter values were determined using empirical estimates. Since many of the parameters are not independent and choosing parameters independently can lead to unrealistic patterns of viral load change over time, parameters obtained from a single patient and study were used for as many of the parameters as possible (Stafford *et al.*, 2000). The remaining parameters are taken from the literature (Table 2.6). η is fixed such that there are 1.4×10^6 replication competent latent cells in 5L of blood at equilibrium (Chun *et al.*, 1997). The initial concentration of uninfected target cells is assumed to be 10 cell/μL (Stafford *et al.*, 2000). Initially there is a single actively infected cell. All other cell and virus populations have size zero.

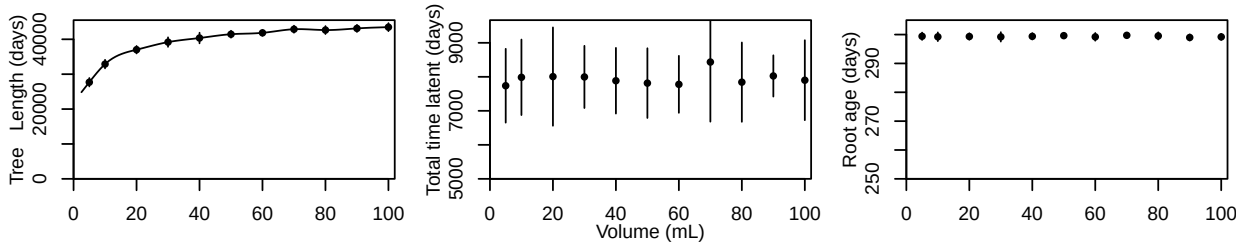


FIGURE 2.7. Impact of simulation volume on properties of genealogies. 50 active and 20 latent viruses were sampled at 75, 100, 200, and 300 days. 10 simulations were run for each simulation volume. Other simulation parameters match those in Table 2.6. Standard error is shown.

In principle, the simulation method described above would allow the entire viral population within a host to be simulated. However, this is not computationally tractable due to the simulation time and memory usage. ODEs of viral dynamics in HIV typically describe the changes in concentrations of cells and viruses per mL of blood. If properties of the viral genealogies become independent of the simulation size as the simulation size increases, it may be reasonable to use a simulation volume much smaller than the total blood volume in an adult. To determine whether this was the case, the impact of simulation size was examined by simulating genealogies generated with different blood volumes while keeping the number sampled sequences constant. Tree length increases and then plateaus as the simulation size increases. Other tree metrics, including root age and total time spent in latency, also showed no trend with volume (Fig. 2.7). Thus, 100 mL was used as the simulation volume.

2.3.5. Simulation of Sequence Data. A separate C program was written to simulate DNA sequences given a sampled tree with branch lengths and a latent history. Sequences are simulated in the typical manner, assuming independent substitutions among sites, starting at the root of the tree and simulating forward in time toward the tips of the tree. The simulator accommodates models as general as the GTR+Γ substitution model (Tavaré, 1986; Yang, 1993). No substitutions can occur while a lineage is latent. Recombination and indels were not simulated. Typically, regions with many indels are removed from alignments. Not including indel in the simulation model has the likely effect of making the sequences slightly longer and slightly more informative than if indels were simulated, but parts of the alignment were removed. The program allows an outgroup with a

node age of zero to be simulated. The sequence at the root is specified by a FASTA format input file (from an existing HIV sequence, for example).

2.3.6. Estimation of DNA Substitution Model Parameters. To select DNA substitution model parameters to use in the simulations, parameters were inferred from empirical datasets for four genomic regions using MCMCtree (Yang, 2007). Alignments for *nef*, *tat*, *C1V2*, and *p17* were taken from a studies on longitudinal Cytotoxic T-lymphocyte (CTL) responses from the LANL HIV special interest alignments (Liu *et al.*, 2006, 2007, 2011). This patient (code PIC1362) was infected in 1998, was a homosexual male, and participated in a study at University of Washington Primary Infection Clinic. The patient had sequences samples taken at 18 time points and was untreated at the time of the study.

To root the tree, sequences from four patients were selected using the LANL database to use as outgroups (GenBank accession numbers: AY331284, AY331289, AB078005, JN024426). The best outgroup is not always clear in phylogenetic studies. Multiple outgroups were used to compare of the effect of rooting on substitution rate estimates. All four of these patients were infected within 2 years of PIC1362, were likely infected on the west coast of the United States, have sexual transmission as a risk factor, were untreated at the time of sampling, and had all four genomic regions were available. The outgroup sequences were combined with the existing alignments using the SynchAlign tool on the LANL HIV database. This resulted in 16 alignments, one for each genomic region outgroup pair. Then, sites with more than 75% gaps were removed from the sequences using a custom R script. This was done to remove problematic regions of the alignments, particularly in *C1V2*.

To obtain parameter estimates, maximum likelihood trees were inferred with RAxML-NG (Kozlov *et al.*, 2019) under an HKY+Γ model (Hasegawa *et al.*, 1985; Yang, 1993) and outgroup rooted. The outgroups were removed from each of the alignments and the maximum likelihood trees. MCMCtree was used to infer the substitution model parameters and substitution rate for each genomic region with each outgroup rooting Yang (2007). An HKY+Γ model with 15 rate categories was used. The prior for κ in the HKY model was $\mathrm{G}(8, 1)$. The prior for among site rate variation was $\alpha \sim \mathrm{G}(1, 1)$. A time unit of 1000 was used with a rate prior of $\mathrm{G}(2, 200)$, or 10^{-6} substitutions per base per day. A birth-death-sequential-sampling model was used with parameters

Region	HXB2 start	HXB2 end	μ	α	κ	π_A	π_C	π_G	π_T
C1V2	6213	7037	3.56×10^{-5}	0.4294	6.9801	0.35322	0.17636	0.21123	0.259191
nef	8797	9414	1.34×10^{-5}	0.4878	8.9138	0.30641	0.21240	0.28265	0.19853
p17	817	1207	8.9×10^{-6}	0.5306	10.6361	0.39393	0.18392	0.25040	0.17175
tat	5831	5962	9.9×10^{-6}	0.7283	7.1751	0.29841	0.21021	0.23449	0.25689

TABLE 2.1. DNA simulation parameters. μ is in units of expected number of substitutions per day per base. The genomic regions simulated do not cover the entire genes.

$\lambda = 2$, $\mu = 1$, $\rho = 0$, and $\psi = 1.8$ (Stadler and Yang, 2013). A root age prior was U(1, 10), meaning the root age was 1000 to 10000 days prior to the last sample time, with 0.01 tail probabilities (Yang and Rannala, 2006).

5 replicates of MCMCTREE were run for each genomic region outgroup pair. Each MCMC was run with a burnin of 1000, sample frequency of 2, and 10000 samples. The estimates from each of the 5 replicate MCMCTREE runs were similar in all cases, indicating the MCMC converged. The point estimate of the substitution rate and the 95% HPD interval bounds for the substitution rate were averaged over the 5 replicates. In most cases, each outgroup produced similar mutation rate estimates for a given genomic region. The outgroup rooting with the smallest 95% HPD interval of the substitution rate divided by substitution rate was used to provide parameters for DNA simulation. However, for *nef*, outgroup 1006 had a much different rooting than the other outgroups. CS2 and PIC55751 had the same root location. Of those two, the one with the smaller 95% HPD interval of the substitution rate divided by substitution rate was used. This resulted in JN024426 being selected as the outgroup for all genomic regions. The first replicate MCMC run of MCMCTREE with JN024426 as the outgroup rooting was used for parameters estimates for each genomic region. This included the estimates of α, κ, μ, and the stationary frequencies (Table 2.1).

The HXB2 sequence was used at the root sequence for the simulation of each region (Table 2.1). However, no bases were removed inside the sequence, as done in the original alignment in regions with over 75% gaps. An HKY model was used for the simulation since the parameters inferences were made with an HKY model MCMCTree.

2.3.7. Sampling and Simulation Parameters. 100 trees were simulated using the stochastic simulator. 50 viruses are sampled every year for 9 years. At year 9, ART begins. 100 latent

cells are sampled at year 10. For each of these 100 phylogenies, 30 alignments for each of four genomic regions were generated with the DNA simulator using an outgroup. Empirically estimated DNA substitution parameters were used, as described above. The estimated substitution rate and length varied among the simulated regions, with *C1V2* having the highest substitution rate ($\mu = 3.56 \times 10^{-5}$ per base per day) and the most sites ($n = 825$) and *nef* having the next highest substitution rate ($\mu = 1.34 \times 10^{-5}$ per base per day) and number of sites ($n = 618$). *p17* has a slightly lower substitution rate than *tat* ($\mu = 8.9 \times 10^{-6}$ per base per day versus $\mu = 9.9 \times 10^{-6}$ per base per day), but more sites ($n = 391$ versus $n = 132$) (Table 2.1). *C1V2* had the lower α for the Γ rate variation model, meaning it has the highest variance in the substitution rate among sites. For each phylogeny and alignment, the sequences and phylogenies were then subsampled three times to generate three trees and three corresponding alignments. Specifically, 10 viruses were subsampled every year for 9 years. 20 were subsampled at 10 years of infection. In total, 300 tree topologies were simulated, each with 20 latent and 90 non-latent randomly sampled sequences. This led to a total of 300 topologies $\times$ 30 alignments $\times$ 4 regions $= 36,000$ simulated datasets.

2.3.8. Effect of the Number of Non-latent Samples on Method Performance. The effect of tree size on the inference of latent samples was examined by changing the number of non-latent samples at each sample time. Using the simulated trees and alignments used in the main simulation analysis, the subsampling was changed from having 10 to 10, 15 or 20 non-latent sequence sampled every year for ten years. This results in a larger phylogenetic tree with the same number of latent sequences for each tree. Each tree was subsampled only one time for each number of non-latent sequences, rather than three times in the main analysis.

2.4. Analysis of Simulated Datasets

HIVTREE was compared with three existing methods, least squares dating (LS) (To *et al.*, 2016), linear regression (LR) (Jones *et al.*, 2018), and pseudo maximum likelihood (ML) (Jones and Joy, 2020) using simulated datasets.

2.4.1. Maximum Likelihood Tree Inference and Rooting. To analyze the simulated datasets a rooted tree topology was first inferred for use by HIVTREE and other heuristic programs. Maximum likelihood trees were inferred with RAxML-NG using an HKY+Γ model and outgroup

rooted (Hasegawa *et al.*, 1985; Kozlov *et al.*, 2019). 25 parsimony and 25 random starting trees were used for the tree search. The outgroup was removed from the inferred tree. Both the LS and Bayesian methods use the outgroup rooted tree. For the ML method, the tree was re-rooted using root to tip regression available in the R package ape prior to analysis (Paradis *et al.*, 2004; Rambaut, 2000). The LR method re-roots the tree using root to tip regression as part of the analysis. For LS, the sampling time was used as an upper bound for the latent lineages and the lower bound was 45 days prior to infection, while the active lineages were constrained to their sampling time. The ML and LR methods do not include additional constraints.

2.4.2. Bayesian Inference. For HIVTREE analyses of simulated data, an HKY+Γ model was used with 5 rate categories and the prior $\kappa \sim \Gamma(8, 1)$ (Hasegawa *et al.*, 1985). The prior for among site rate variation was $\alpha \sim \Gamma(4, 8)$. A time unit of 1000 was used with a substitution rate prior of $\Gamma(2, 200)$, meaning the mean was 10^{-5} per base per day. The root age prior was $\Gamma(36.5, 100)$. The latent times were bounded at 3.695, which is equivalent to 45 days prior to infection.

2.4.3. MCMC Settings for Simulation Analysis. Two MCMCs were run with different seeds for each analysis to check for convergence. The MCMC was sampled every other iteration for 30,000 samples with a burn in of 2,500. Thus a total of $30000 \times 2 + 2500 = 62,500$ iterations were run. The internal node ages of the two replicate MCMCs were compared for each analysis. If the mean age difference between the two replicate MCMCs was more than 10 days for more than 8 internal nodes, 20 days for more than 4 internal nodes, or 100 for any internal nodes, the MCMCs are considered to not have converged. A total of 234 pairs of MCMCs did not converge out of 36,000 pairs run. For each pair of MCMCs that did not converge, another 2 MCMCs were run with different seeds with 60,000 samples. Of those, 7 pairs of MCMCs did not converge. Those MCMCs were rerun again with different seeds, a burnin of 10000 iterations, and were run for 240,000 iterations, sampling every other iteration. All of these runs met the above convergence criteria.

2.4.4. Combining Posterior Estimates from HIVtree. For combining results in Bayesian analyses of the simulated and empirical datasets, the function kdensity in the kdensity R package was used for kernel density estimation of the posterior distribution and the prior distribution of

each latent time (Moss and Tveten, 2019). The posteriors and priors for each genomic region were multiplied according to equation 2.1, using a uniform distribution between the sample time and the upper bound for the oldest possible integration date as the desired prior. The resulting function was normalized by finding the proportionality constant using the integrate function. For the simulated datasets, the integral bounds were set to the bounds on the latent time in HIVTREE, which was the sample time and 45 days prior to infection. The 0.025 and 0.975 quantiles were found using the invFunc function in the R package GoFKernel (Pavia, 2015). The mean for the joint posterior was found using the integrate function. For the simulated datasets, this analysis was conducted on only a third of the trees from the main simulation analysis due to the highly demanding computations involved.

2.4.5. Effect of the Number of Non-latent Samples on Method Performance. As preliminary analysis did not show any trend with the other methods, this analysis was only run for the *p17* datasets with HIVTREE. For the analyses with HIVTREE, the priors were the same as in the main simulation analyses with HIVTREE. The MCMCs were run with a burnin of 5,000 iterations, sampling every other iteration and sampling a total of 50,000 times. Two replicate MCMCs were run for each analysis. The difference between the mean times of the internal nodes was compared. The MCMCs were considered to have converged if this difference was no more than 10 days for at most 10% of the internal nodes, 20 days for at most 5% of the internal nodes, and no more than 100 days for any of the internal nodes. All MCMCs met the convergence criteria.

2.4.6. Existing Methods. The LR method used scripts available at: https://github.com/cfe-lab/phylodating. This method uses a linear model to estimate the latent integration dates. The ML method used scripts available at: https://github.com/brj1/node.dating/releases/tag/v1.2. This method uses a pseudo-maximum likelihood approach to estimate the latent integration times by fixing the mutation rate and then using maximum likelihood to estimate the integration dates. The driver script provided by Jones et al. is available at: https://github.com/nage0178/HIVtreeAnalysis. The LS method was obtained from: https://github.com/tothuhien/lsd-0.3beta/releases/tag/v0.3.3. This method uses a

least squares approach to minimize the difference between the branch lengths and sample dates and infer unknown ages.

2.5. Empirical Dataset and Analysis

2.5.1. Jones et al. Dataset. Sequences originally published by Jones et al. (2018) were taken from GenBank (accession nos. MG822917-MG823179), and separated into patient 1 and patient 2 (Jones *et al.*, 2018). The sequences from patient 1 were aligned using MAFFT (version 7.453) using the default settings (Rozewicki *et al.*, 2019). The sequences from patient 2 did not need to be aligned. The relative sample dates were determined using the collection date.

2.5.2. Abrahams et al. Dataset. Alignments for patients 217 and 257 originally published by Abrahams et al. (2019) were available from https://github.com/veg/ogv-dating/tree/master/results/alignments (Abrahams *et al.*, 2019). There were multiple alignments for each data set and the "fasta_combined.msa" alignments were used. The week of sampling is included in the sequence name. Using the supplemental data table, the relative dates of sampling in units of days were determined. For some patients, there were multiple visit dates in the same week. In this case, the first visit date was used as the sample date for all sequences collected during that week. For each alignment, sequences were subsampled to include 10, 15, or 20 sequences from each pre-ART each collection time point and all outgrowth virus sequences. If less than the desired number of sequences were available at a given time point, all of the available sequences were used. While the sequences were aligned, some of the alignments had many gaps. Sites in the alignments were removed if they had more than 75%, 85%, or 95% gaps. Thus, for each of 8 starting alignments, 9 alignments were created. However, some of the alignments with gap removal were identical. Thus, a total of 46 unique alignments were created. HIVTREE requires the sampling date to be at the end of the sequence name. Thus, the sequence names from the original publications were modified for our analyses.

2.5.3. Empirical Data Analysis. For all empirical data sets, RAxML-NG was run using an HKY+Γ model (Kozlov *et al.*, 2019). 25 parsimony and 25 random starting trees were used for the tree search. Trees were rooted using root to tip regression using the rtt function in the ape

package available in the R package ape prior to analysis (Paradis *et al.*, 2004; Rambaut, 2000). Each of the four methods were run on all datasets.

2.5.3.1. *Jones et al. Dataset.* For the first dataset (Jones *et al.*, 2018), HIVTREE was run with a root age prior of $\Gamma(8, 60)$ for patient 1 and $\Gamma(15, 50)$ for patient 2. These priors were chosen to have an induced prior when running without data with a variance of several years and a mean several years prior to diagnosis. Latent integration times were bounded 10 years prior to diagnosis, as a very conservative oldest possible bound. In the HIVTREE analysis, an HKY+Γ model was used with 5 rate categories with the prior $\kappa \sim \Gamma(8, 1)$. The prior for among site rate variation was $\alpha \sim \Gamma(4, 8)$. A time unit of 1000 was used with a substitution rate prior of $\Gamma(5, 1000)$, meaning the mean was 5×10^{-6} per base per day. HIVTREE was run with a burnin of 5,000 iterations, with 70,000 samples, sampling every other iteration. Two replicate MCMCs were run for each dataset. Convergence was checked by confirming no more than 5% of the mean internal nodes ages differed by more than 10 days between replicate MCMCs, 2.5% differed by more than 20 days, or any of the internal nodes differed by more than 100 days. Both pairs of MCMCs met this convergence criteria. For the LS analysis, latent integration times had the same bounds of 10 years prior to diagnosis and the sample times. This dataset only sampled one gene, so estimates from multiple genes could not be combined.

2.5.3.2. *Abrahams et al. Dataset.* For the second dataset (Abrahams *et al.*, 2019), the LS and HIVTREE analyses bounded the latent times at the infection times and the sample times. In the HIVTREE analysis, an HKY+Γ model was used with 5 rate categories with the prior $\kappa \sim \Gamma(8, 1)$. The prior for among site rate variation was $\alpha \sim \Gamma(4, 8)$. A time unit of 1000 was used with a substitution rate prior of $\Gamma(2, 200)$, meaning the mean was 10^{-5} per base per day. The root age prior was $\Gamma(0.25, 110)$ for all datasets. This prior was chosen to have a relatively wide variance on the root age with a mean slightly before the infection time as well as a large variance on the latent integration times. As described in the Prior Model section, the root ages are older than the given prior when run without data, and they are also different for each dataset. When running the MCMC under the prior, small changes to the prior appeared to cause little change to the posterior distribution of the latent integration times.

Two replicate runs of HIVtree were run for each analysis. A burnin of 8,000 was used with samples taken every other iterations for a total of 80,000 samples. Thus, the MCMC was run for 168,000 iterations. Convergence of the MCMCs was checked by comparing the mean ages of the internal node ages. If more than 5% of the mean internal nodes ages differed by more than 10 days between replicate MCMCs, 2.5% differed by more than 20 days, or any of the internal nodes differed by more than 100 days, the MCMC was considered to not have converged. Two pairs of MCMCs did not converge. These were rerun with a a total of 150,000 samples, sampling every other iteration with a burnin of 8,000 iterations. Convergence was checked again with the same criteria as previously. Both pairs MCMCs had converged.

The estimates from multiple genomic regions for the second dataset were only combined for the tree with 10 non-latent sequences per sampling time and sites with gaps in over 75% of the sequences were removed from the alignment.

2.5.4. Program Availability. The gene tree and the DNA simulation software packages are available at: https://github.com/nage0178/HIVtreeSimulations. The HIVtree software package is available at: https://github.com/nage0178/HIVtree. Scripts to produce the results in this paper are available at: https://github.com/nage0178/HIVtreeAnalysis.

2.6. Simulation Results

2.6.1. Validation of Inference Program. The prior distributions from the rejection simulation and HIVtree are very similar (Fig. 2.8), indicating these are these two factors explain the induced root age prior. The average posterior distributions from the Bayesian simulation are in good agreement with the prior distributions (Figs. 2.10 - 2.13). This suggests the program is implemented correctly.

2.6.2. Agreement between the Deterministic and Stochastic Models. For large population sizes, the stochastic model and the deterministic (ODE) model are expected to produce similar results for the population size as a function of time given the parameters and initial values are such that the population does not go extinct in the stochastic simulation. This is because we have designed the stochastic simulator to have an expected population size equal to the predicted population size for the deterministic model at any point in time and the relative variance of the

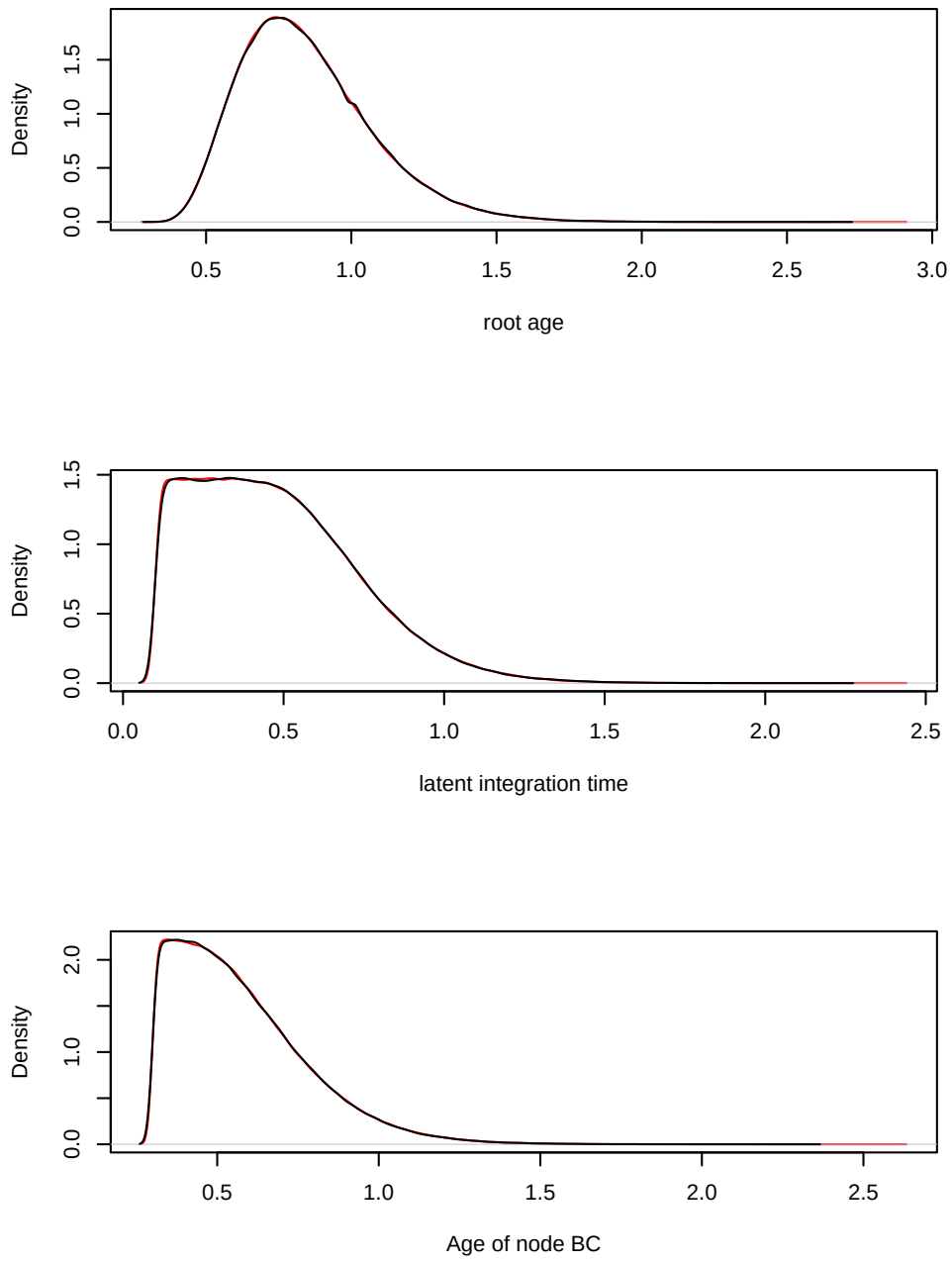


FIGURE 2.8. The prior distributions of the node ages and latent integration time from HIVTREE, in black, and with rejection simulation, in red, are shown. The distributions from HIVTREE and the rejection simulation are in good agreement.

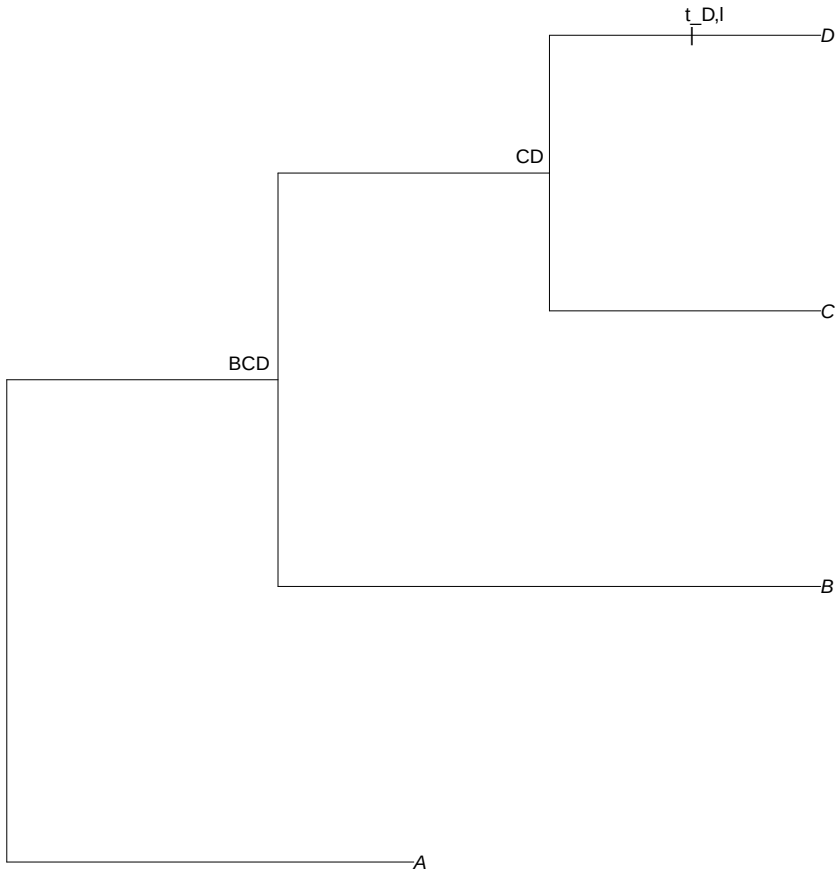


FIGURE 2.9. Tree used for the Bayesian simulation is asymmetric with one latent sequence, D. Sequences B, C, and D were sampled at time 3650 days and sequence A was sampled at 1825 days. The MCMC estimates the times of BCD, CD, and the latent time, t_D,l. The MCMC also estimates the root age, but the prior was very informative so the results are not shown.

stochastic model decreases with increasing population size. Populations sizes are in good agreement when there is no extinction (Fig. 2.14). Cases of extinction are common, but are not considered further.

2.6.3. Simulation Analysis. Here we compare the statistical performance of HIVTREE and several other existing methods when analyzing simulated datasets with known latency times.

2.6.3.1. *Comparisons on a Fixed Tree Topology.* HIVTREE was compared with three existing methods, least squares dating (LS) (To *et al.*, 2016), linear regression (LR) (Jones *et al.*, 2018), and pseudo maximum likelihood (ML) (Jones and Joy, 2020) using simulated datasets. The effect

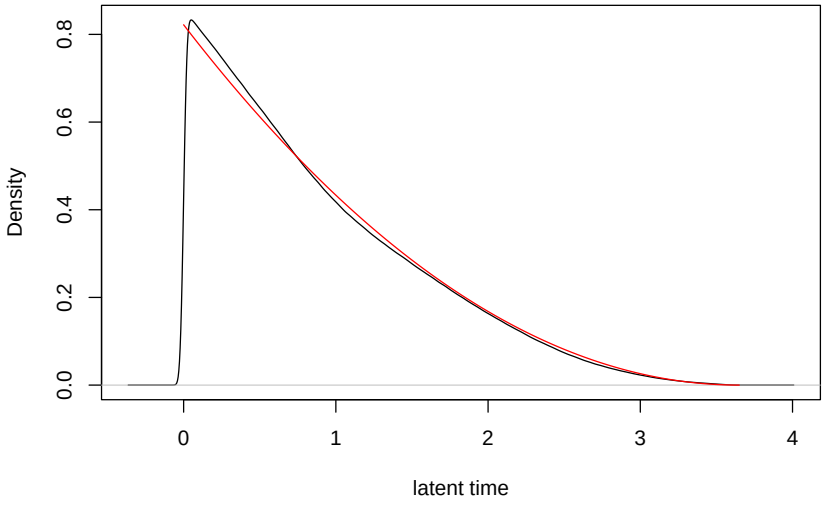


FIGURE 2.10. The KDE of the latent time t_D,1 is shown in black with the prior density shown in red. The prior is beta distributed with $\alpha = 1$ and $\beta = 3$ transformed to be on the interval [0, 3.650]. The figure is in the units reported by MCMC, where 1 is a 1000 days and the last samples were taken at time zero.

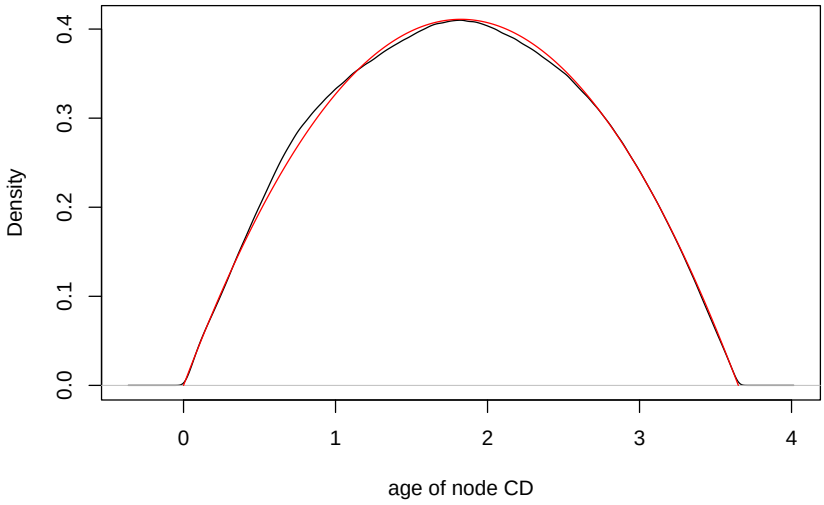


FIGURE 2.11. The KDE of the age of node CD,1 is shown in black with the prior density shown in red. The prior is beta distributed with $\alpha = 2$ and $\beta = 2$ transformed to be on the interval [0, 3.650]. The figure is in the units reported by MCMC, where 1 is a 1000 days and the last samples were taken at time zero.

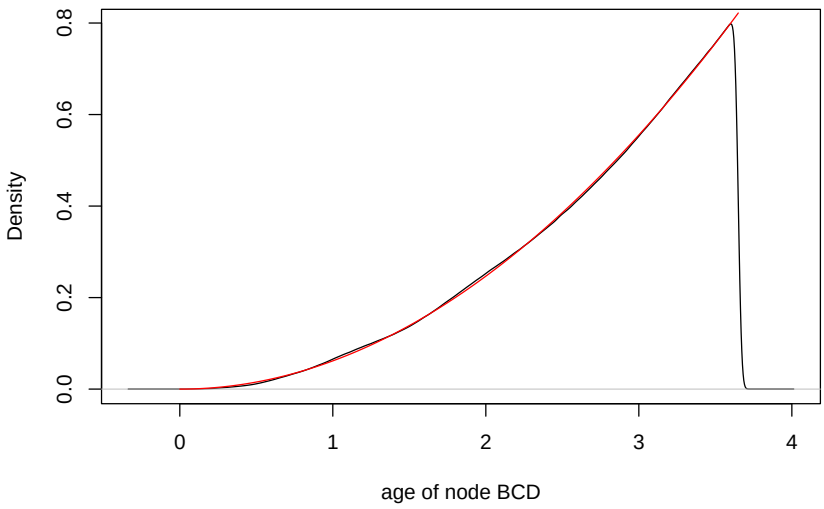


FIGURE 2.12. The KDE of the age of node BCD is shown in black with the prior density shown in red. The prior is beta distributed with $\alpha = 3$ and $\beta = 1$ transformed to be on the interval $[0, 3.650]$. The figure is in the units reported by MCMC, where 1 is a 1000 days and the last samples were taken at time zero.

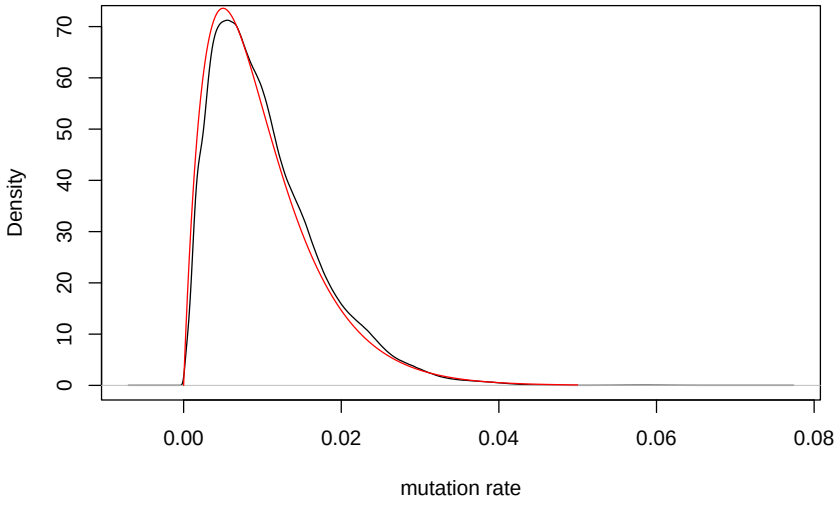


FIGURE 2.13. The KDE of the mutation rate is shown in black with the prior density shown in red. The prior is $\Gamma(2, 200)$, in units of expected substitutions per 1000 days.

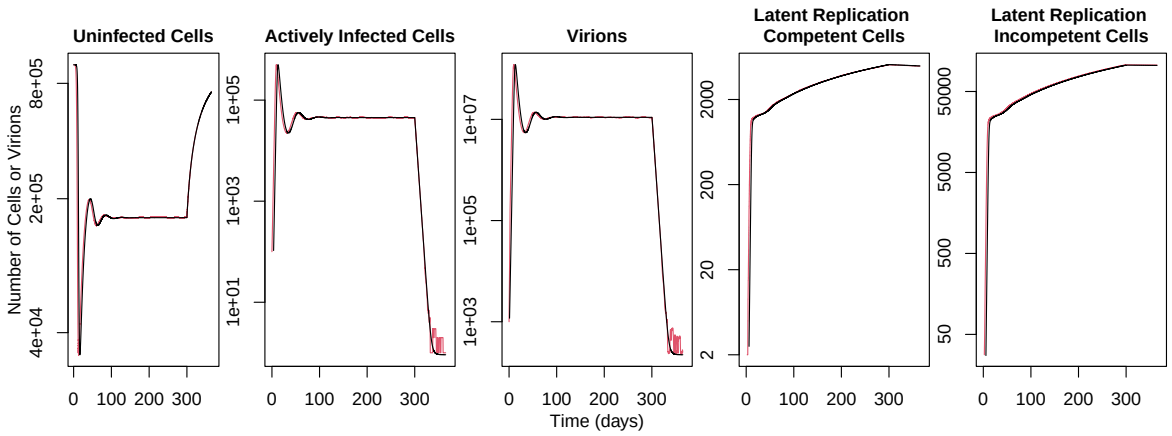


FIGURE 2.14. Predicted population sizes in the deterministic model and observed population sizes in the stochastic model are very similar. For both models, a blood volume of 10 mL was modeled using the parameters listed in Table 2.6. The initial population sizes are 10^4 target cells/mL, 1 actively infected cell/mL, and 10 virions/mL. ART was started at 300 days. The deterministic model is shown in black, and one realization of the stochastic simulation is shown in red. In comparison to the initial conditions described in the text, a larger number of actively infected cells was used to limit the stochastic effects of small population sizes, allowing for a comparison when the virus is unlikely to become extinct.

of variation among the independently simulated sequences on point estimates of latent tip ages can be seen by comparing the estimates for a given latent tip in a fixed tree. Even with *C1V2*, the most informative genomic region simulated, there is considerable variation in the estimated latency time for a given latent tip (Fig. 2.15). The variation is even larger for the other genomic regions (Fig. 2.16). The estimated times for a single latent tip sometimes differs from the true value by a decade or more for both the LR and ML methods. The LS method has fewer extreme estimates, which are prevented by bounds on the integration times. LS allows for upper and lower bounds for each individual latent sequence while ML has the same upper bound on all latent sequences, which is the last sample time. The LR has no bounds on the inferred integration time, potentially allowing the latent sequences to be formed either after the sequence was sampled or before an individual was infected. Both outcomes are highly unlikely.

2.6.3.2. *Combined Inferences Across Genes.* The posterior distribution for each latent time is inferred separately for each genomic region when using HIVTREE. When the marginal densities are

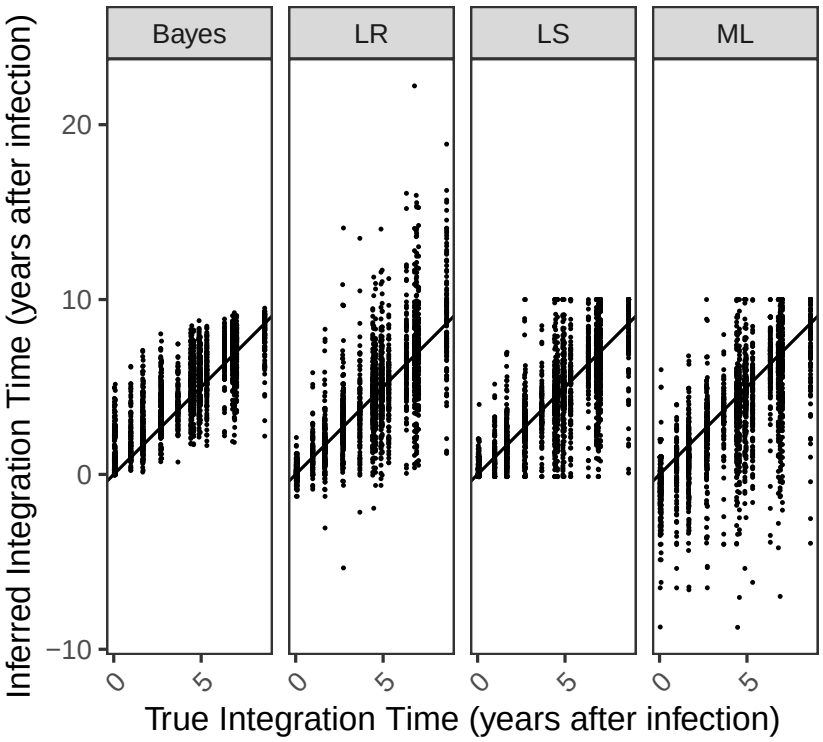


FIGURE 2.15. For all 30 alignments simulated for *C1V2* on a fixed tree, the inferred integration dates are shown for each method. If the methods performed perfectly, all points would fall on the line, which is has an intercept of 0 and slope of 1. The units are years after infection.

combined across the regions, the posterior densities become narrower and closer to the true value (Fig. 2.17). The other methods presented here do not allow such information sharing.

2.6.3.3. *Numerical Issues Combining Posteriors.* For a small subset of simulated data, numerical issues prevented estimation of a combined latent integration time . In one case, no error messages resulted but the proportionality constant was on the order of 10^{-12}. Likely due to numerical issues, this caused the mean latent integration time to be estimated to a value on the order of 10^6, which has zero prior probability. This latent integration inference was removed from the analysis. Out of inferences for 90,000 latent integration times, 63 other analyses combining latent integration times from all four genomic regions produced error messages related to non-integrable functions, and did not produced an estimate. This occurred for three latent times in the analysis of two genes only. One additional run produced an estimate, but suffered from obvious numerical

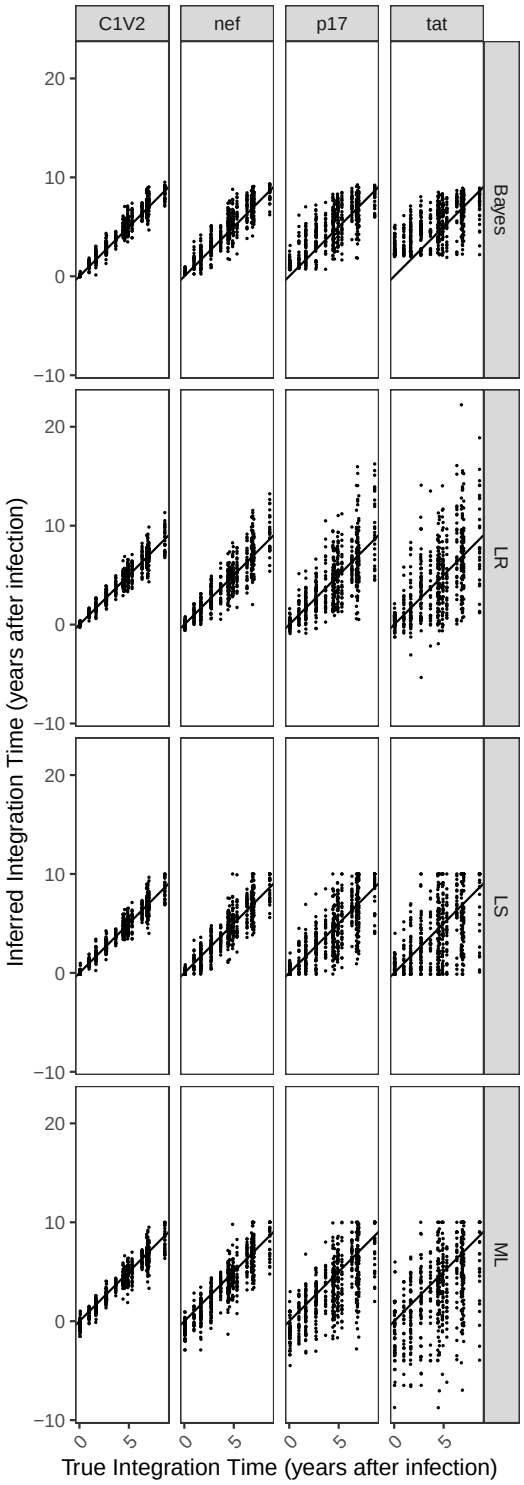


FIGURE 2.16. For a fixed tree topology, there are 30 latent integration times for each of the 30 alignments for a given genomic region. The line has slope 1 and intercept 0.

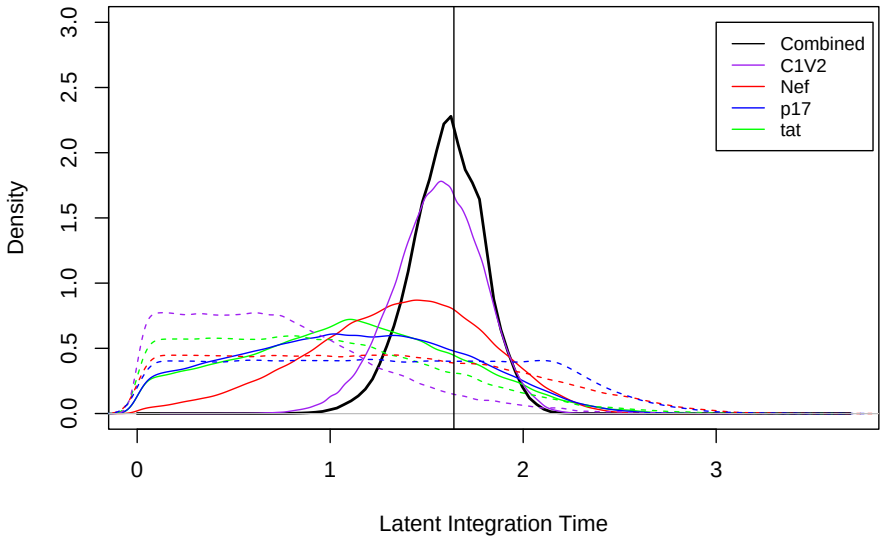


FIGURE 2.17. Joint posterior density for a single latency time across all simulated regions. Each solid colored line shows the marginal posterior density for a single latency time for different genomic regions. The dashed colored lines show the marginal prior densities, which result from running the MCMC without data. The solid black line shows the estimate with the genomic regions combined. The vertical line is the true latent integration time. The MCMC was run for 500,000 iterations, sampling every other iteration. This results in smoother curves than the shorter MCMC runs used in the larger analysis of simulated data, but results are very similar.

issues. These cases were removed from further analysis. These likely result when the posterior distributions from different genomic regions are non-overlapping.

2.6.3.4. *Summary of Method Performance.* Root mean square error (RMSE) is a useful measure of method performance that includes both bias and variance and is directly comparable across methods. RMSE is lowest for *C1V2* and highest for *tat* for all analyses (Fig. 2.18a). For *C1V2*, the average RMSE among methods, from lowest to highest, is Bayesian (0.67), LS (0.74), LR (0.77) and ML (0.86). All the methods are the least biased for *C1V2* and most biased for *tat* (Fig. 2.18b). The average bias for the ML and LS methods are more negative for the shorter, slower evolving genomic regions (-1.64 and -0.47 years respectively for *tat*), while the Bayesian and LR method have a positive bias on average (0.78 and 0.12 years respectively for *tat*). The trend for the mean square error (MSE) is similar to the trend for RMSE (Fig. 2.19).

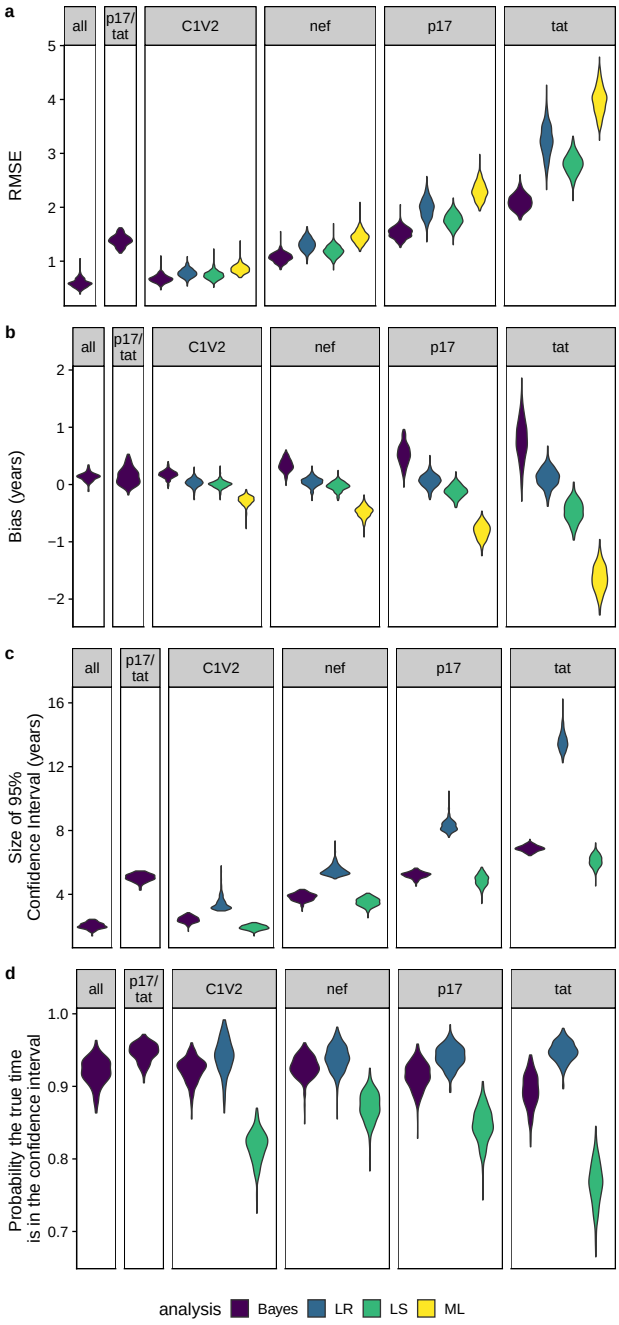


FIGURE 2.18. For each of fixed tree topologies, the root mean square error (RMSE), bias, and size of the 95% confidence/credibility interval was averaged across all 900 latent times for each genomic region analysis combination. Each violin plot is made using 300 data points, corresponding to the average from each of the 300 fixed tree topologies. For the Bayesian combined analysis of either all of the genomic regions or only *p17/tat*, only a third of the fixed tree topologies were analyzed.

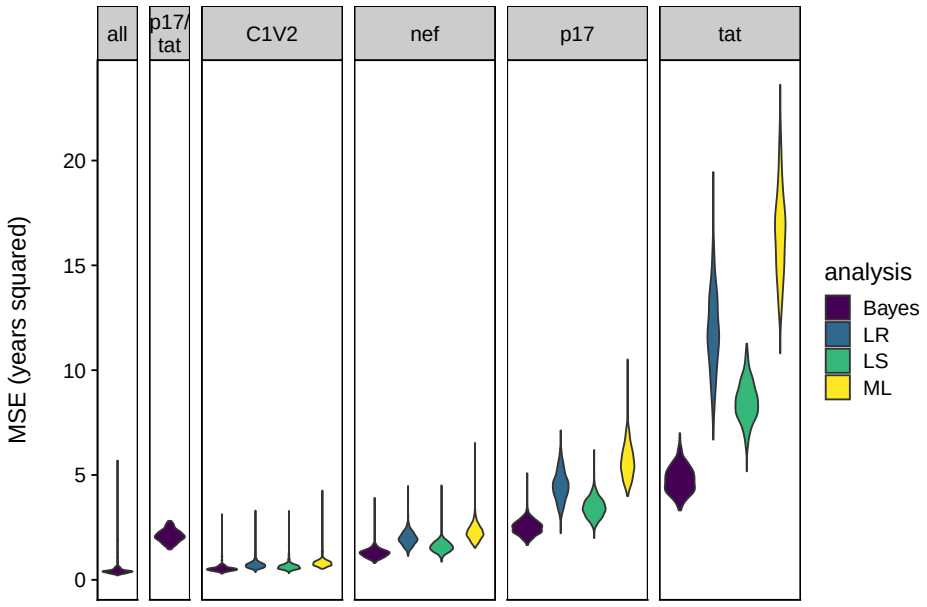


FIGURE 2.19. For each of fixed tree topologies in the main analysis, the mean square error (MSE) was averaged across all 900 latent times for each genomic region analysis combination. Each violin plot is made using 300 data points, corresponding to the average from each of the 300 fixed tree topologies. For the Bayesian combined analysis of either all of the genomic regions or only p17/tat, only a third of the fixed tree topologies were analyzed. This plot uses the same datasets and inferences as figure 3 in the main text.

The probability that the true value falls in the 95% confidence interval (or 95% highest posterior density interval for Bayesian analysis) was also considered (Fig. 2.18d). The Bayesian method has comparable average coverage probabilities for *C1V2* and *nef* of 92% and 93%, respectively, with the lowest coverage probability for *tat* (89%). The average size of the 95% credible set for the longest and shortest sequences, *C1V2* and *tat*, are 2.4 years and 6.9 years, respectively. LR has the highest coverage, with a coverage probability of 94% for *C1V2* and 95% for *tat*. However, LR has very large confidence intervals (Fig. 2.18c). The mean sizes of the 95% confidence interval is 3.4 years and 13.6 years for *C1V2* and *tat*, respectively. In contrast, LS shows lower coverage probabilities but smaller confidence intervals. LS has its highest average coverage probability for *nef* (87%), but drops to 77% for *tat* (Fig. 2.18d). For the longest genomic region, *C1V2*, the average coverage probability is only 82%. This is likely due to the much smaller confidence interval size. The size of the 95% confidence interval is much larger for the LR method than either the LS or Bayesian

methods (Fig. 2.18c). The LS and Bayesian methods have similar size confidence intervals, but the Bayesian method is more likely to contain the true value in the 95% confidence interval (has higher average coverage probability). The ML method has the largest RMSE and bias on average for all regions and does not provide confidence intervals.

For the Bayesian method, when the inferences are combined across all four genomic regions, the average size 95% credible set is 144 days smaller on average than with *C1V2* alone. The average probability the true integration time is in the 95% credible set is similar to the results for the longest genomic region. When the two shortest genes, *p17* and *tat*, are combined, the average size of the 95% credible set is slightly smaller than with *p17* alone (60 days), but the probability the true value is in the 95% credible set increases from 91% with *p17* alone to 95% in the combined analysis (Fig. 2.18c,d). The average RMSE is slightly smaller for the combined analysis of all genes (0.59) than with *C1V2* alone (0.67). The average RMSE is smaller when *p17* and *tat* are combined (1.39) than with *p17* alone (1.53).

2.6.3.5. *Effect of Number of Non-latent Samples.* The number of non-latent sequences at each sampling time does not have a large impact on bias (Fig. 2.20), MSE (Fig. 2.21), size of the 95% confidence intervals (2.22), or the probability the inferred integration times fall within the 95% confidence intervals or credible sets (Fig. 2.23) for any of the methods.

2.7. Empirical Results

We applied each of the four methods to HIV data sets from two studies of serial sampled HIV sequences. The first data set is comprised of *nef* sequences for two patients (Jones *et al.*, 2018). For each patient, plasma HIV RNA was sequenced multiple times over a period of almost a decade either pre-treatment or during incompletely suppressive dual ART. After the initiation of combination ART (cART), samples from the putative reservoir were taken from at least two time points. Samples consisted of HIV RNA sequences sampled during viral blips and proviral DNA collected from whole blood and peripheral blood mononuclear cells (PBMC). The second data set has three regions of *env* for both the patients analyzed (217 and 257) and *gag* and *nef* sequences for one patient (257) (Abrahams *et al.*, 2019). For both patients, virus was sequenced from the

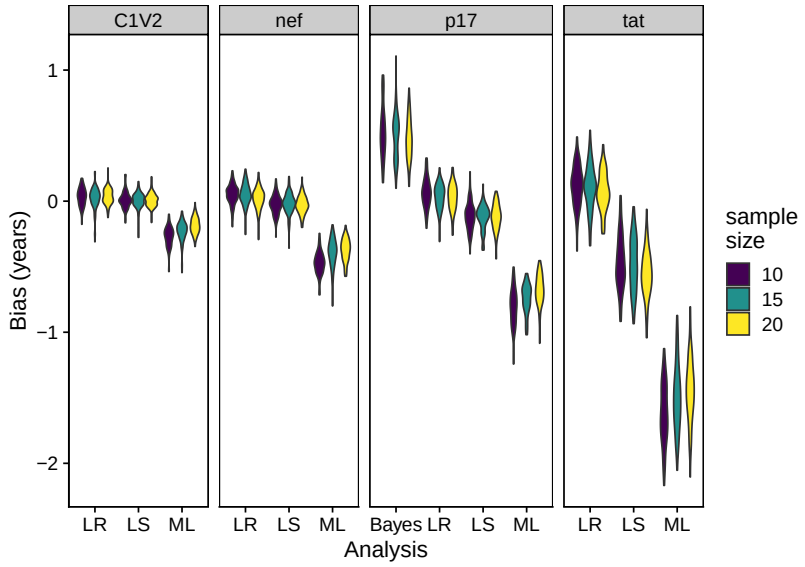


FIGURE 2.20. The bias for each simulated region using each of four analysis is shown. Each data point in the violin plot is the average bias of 20 latent times in each of 30 alignments with a fixed topology. There are a total of 100 fixed topologies for each violin plot. The number of non-latent sequences sampled at each of 9 sampling time points is indicated by the color. While the longest and most quickly evolving genomic region, *C1V2*, has the lowest bias for all methods and the shorter, more slowly evolving genomic regions have greater bias, there is not a consistent trend in bias by the sample size.

plasma multiple times over several years prior to ART initiation. After ART initiation, viral RNA was isolated from the supernatant of quantitative viral outgrowth assays.

The inferred latent integration times for the patients in the first dataset obtained using HIVTREE span over a decade (Fig. 2.24), similar to estimates obtained using other methods (Fig. 2.25). However, ML and LR infer integration times that occur after the sampling time in some cases (Fig. 2.25). For the second dataset, the point estimates, especially for the early sample times (11.1 for patient 1 and 17.9 for patient 2), tend to be concentrated near the time of ART initiation. The combined point estimates for the latency times inferred using HIVTREE appear loosely clustered around the time ART began for patient 257, with narrower credible sets than the analyses on individual genomic regions (Fig. 2.26). These patterns for patient 217 are less clear, possible due to fewer genomic regions and fewer latent sequences (Fig. 2.27).

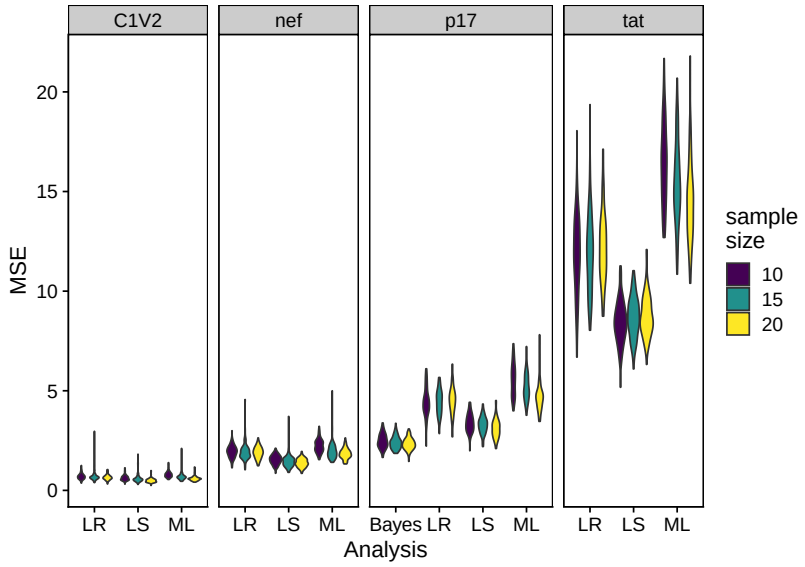


FIGURE 2.21. Each data point in the violin plot is the average MSE of 20 latent times in each of 30 alignments with a fixed topology. There are a total of 100 fixed topologies for each violin plot. The number of non-latent sequences sampled at each of 9 sampling time points is indicated by the color. There is not a consistent trend in MSE by the sample size.

Each figure (Fig. 2.28 - 2.43) show the inferred integration date for each method, LR, LS, ML, and HIVTREE with different levels of gap removal and sample sizes. Each figure is for a single patient and genomic regions. Some figures have two levels of gap removal instead of three because gap removal at different levels resulted in identical alignments. Thus, only the non-redundant results are shown. The names for the regions of the genome (e.g. ENV_4, NEF_1) match those in the original alignment names. Sometimes LS gives very large confidence intervals, covering the entire area between the bounds for a sequence (Fig. 2.28, 2.31), while in other cases the confidence intervals are smaller than LR.

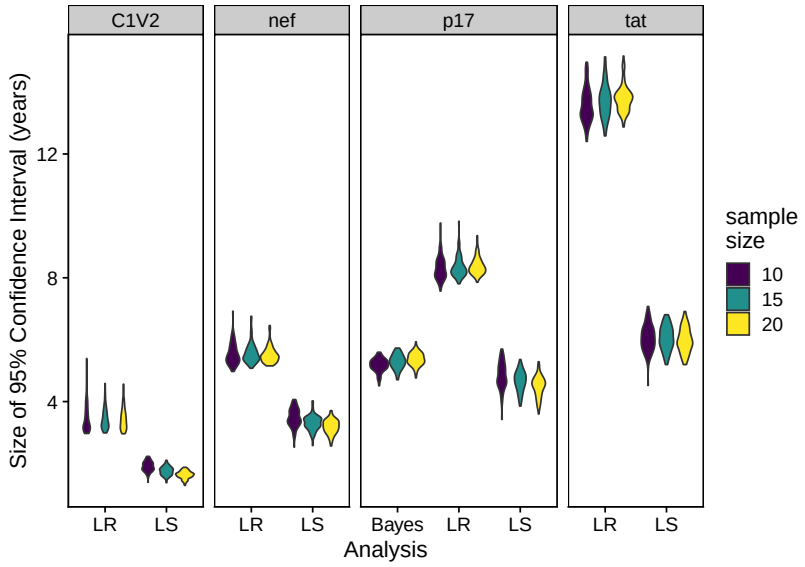


FIGURE 2.22. Each data point in the violin plot is the average size of the 95% confidence intervals (or credible sets for the Bayesian method) of 20 latent times in each of 30 alignments with a fixed topology. There are a total of 100 fixed topologies for each violin plot. The number of non-latent sequences sampled at each of 9 sampling time points is indicated by the color. The longest and most quickly evolving genomic region, *C1V2*, has smaller confidence intervals for all methods. The sample size does not have a large effect on the size of the confidence intervals.

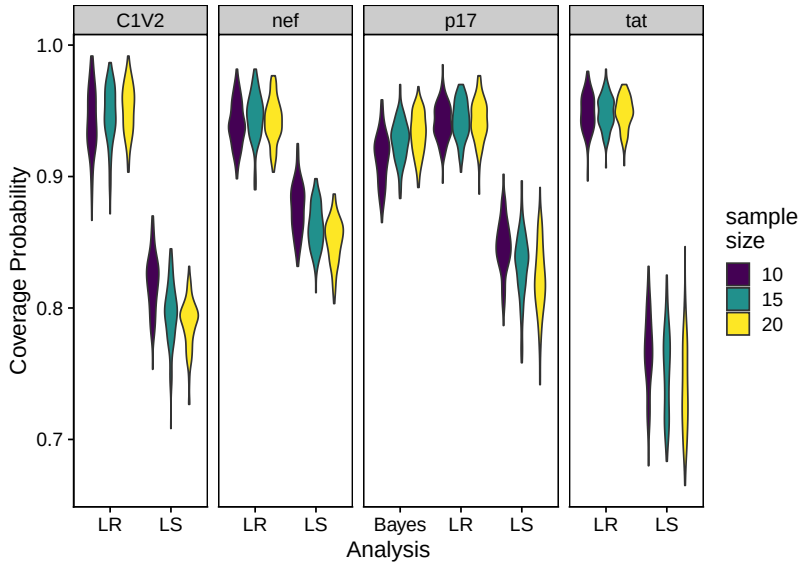


FIGURE 2.23. Each data point in the violin plot is the probability the true latent time falls within the 95% confidence intervals (or 95% highest posterior density set) for 20 latent times in each of 30 alignments with a fixed topology. There are a total of 100 fixed topologies for each violin plot. The number of non-latent sequences sampled at each of 9 sampling time points is indicated by the color. For the LS method, the probability decreases when the region is shorter with a lower mutation rate, but does not vary predictably with sample size. The ML method is not shown since it does provide confidence intervals or credible sets.

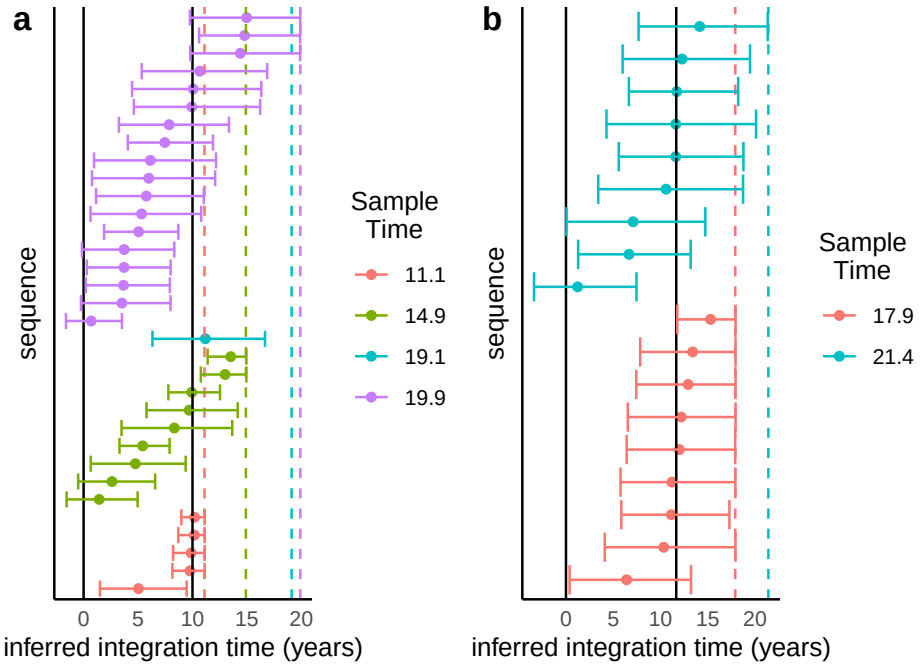


FIGURE 2.24. Panels (a) and (b) show the inferred latent integration times, in units of years after diagnosis, for patients 1 and 2, respectively, inferred using HIVTREE to analyse sequence data for the *nef* gene locus. A dot indicates the posterior mean and bars represent the 95% credible interval. The solid vertical lines indicate the positive test date (left) and time of cART initiation (right) for each patient. The colored dashed vertical lines indicate the sample times.

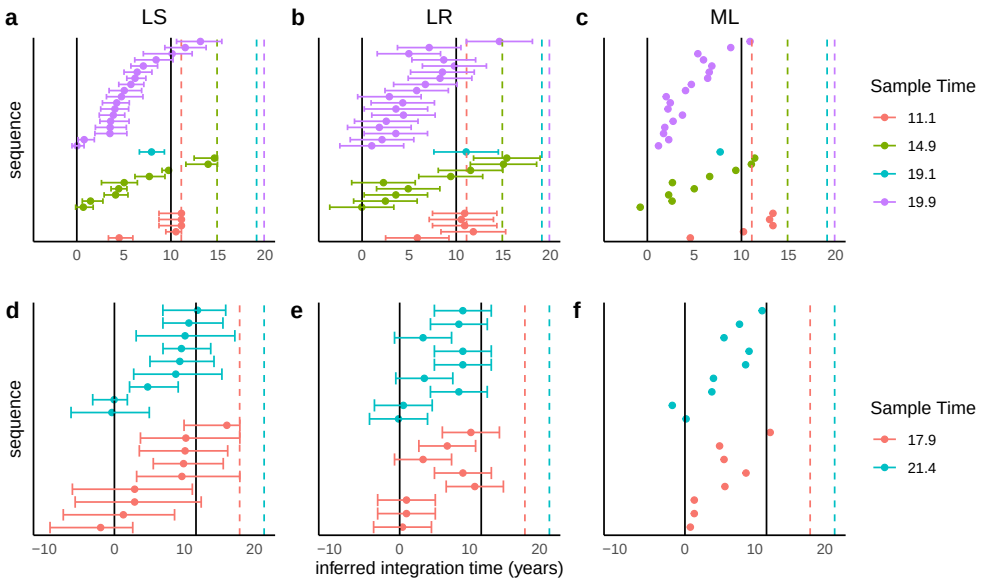


FIGURE 2.25. (a-c) and (d-f) show the inferred integration dates for each sequence from patient 1 and 2, respectively. (a,d), (b,e), and (c,f) show inferences from LS, LR, and ML, respectively. The vertical lines show the first positive date (left) and start of cART (right). The bar show 95% confidence intervals for LS and LR. Confidence intervals are not inferred in the ML method. With sample time 11.1 for patient 1, three of the latent integration times inferred with ML and one with LR are after the sampling date. The LS method is bounded at the sample time, but those sequences are inferred to have been integrated at the sample time.

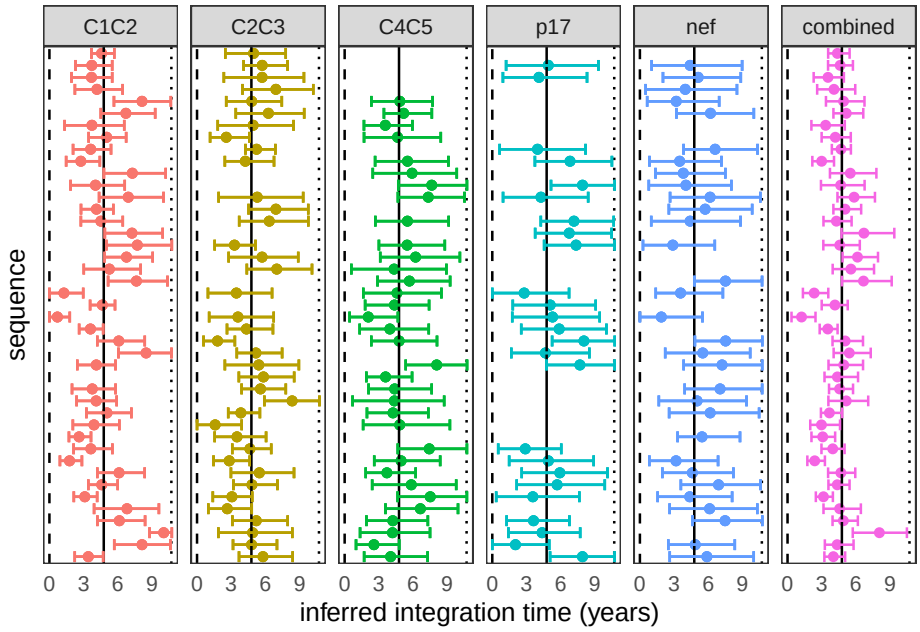


FIGURE 2.26. The five panels to the left each show the integration times inferred using HIVTREE for a single genomic region locus. The panel to the right shows the inferred integration times when posterior distributions for the five loci are combined. A dot indicates the posterior mean and bars represent the 95% credible interval, in units of years after diagnosis. The results are from patient 257 (Abrahams *et al.*, 2019). 10 non-latent sequence were used as each available timepoint and sites with more than 75% gaps were removed from the alignment prior to analysis. The dashed line shows the infection time, the solid line shows the start of ART, and the dotted line shows the sample time.

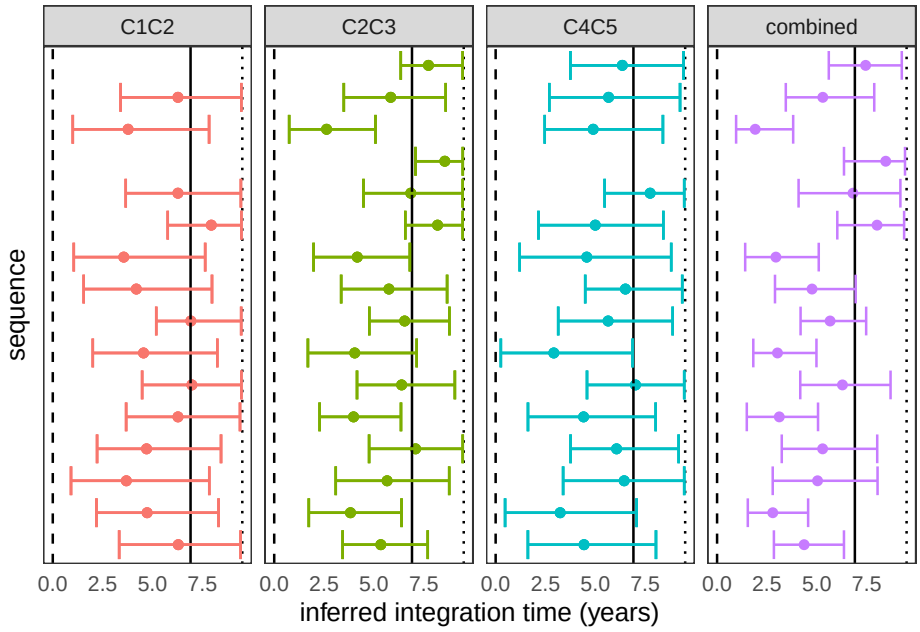


FIGURE 2.27. Each of the left three panels shows the integration times inferred using HIVTREE for a single sequence. The panel on the right shows the inferred integration times when the posterior estimate for the three sequences are combined. The results are from patient 217 (Abrahams *et al.*, 2019). 10 non-latent sequence were used as each available timepoint and sites with more than 75% gaps were removed from the alignment prior to analysis. The dashed line shows the infection time, the solid line shows the start of ART, and the dotted line shows the sample time.

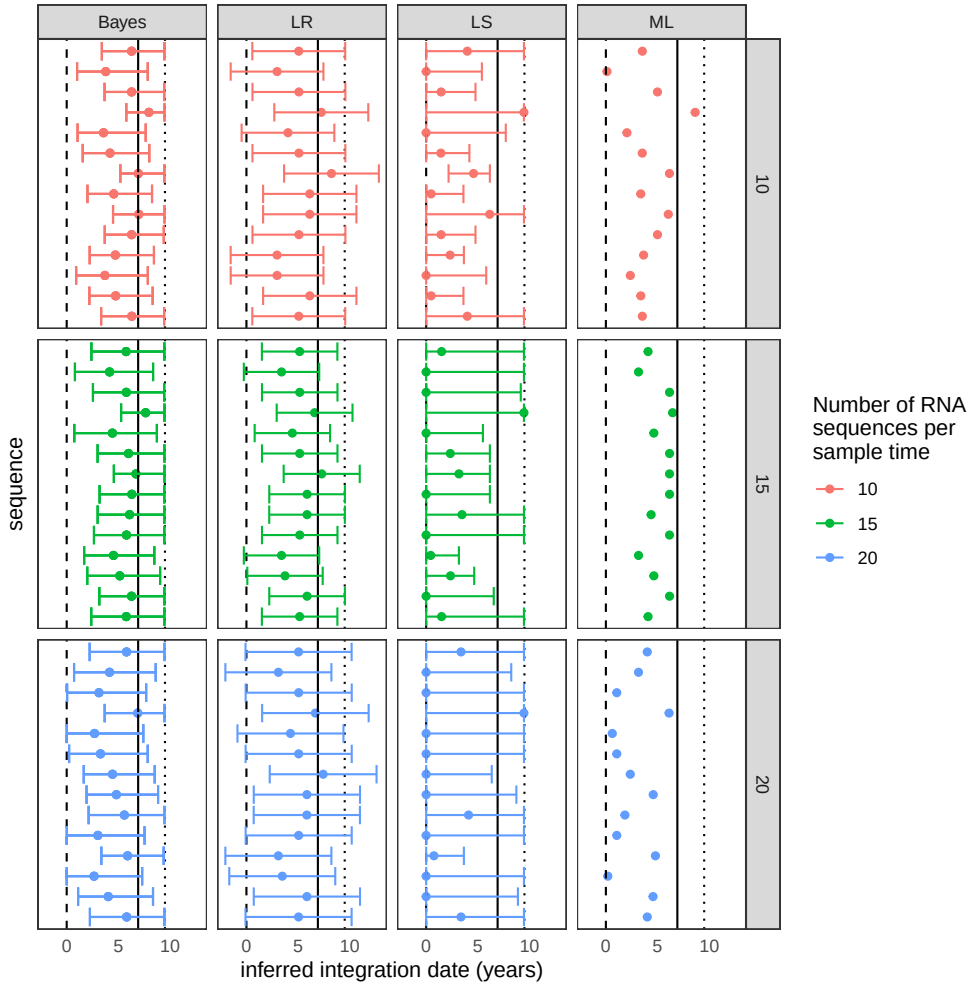


FIGURE 2.28. The inferred latent integration dates for Env_2 from patient 217 are shown for each method. 95% confidence intervals are shown for the LR and LS methods, and the 95% credible interval is shown for HIVTree. Sequences are shown in the same order in each panel. The vertical lines show the time of infection (dashed), time of treatment start (solid) and the time of sampling (dotted). The color shows the number of RNA sequences subsampled from the original alignment at each sample time. If fewer sequences were available then the number indicated by the color at a given time, all available sequences were used. Sites with greater than 75% missing gaps have been removed from the alignment.

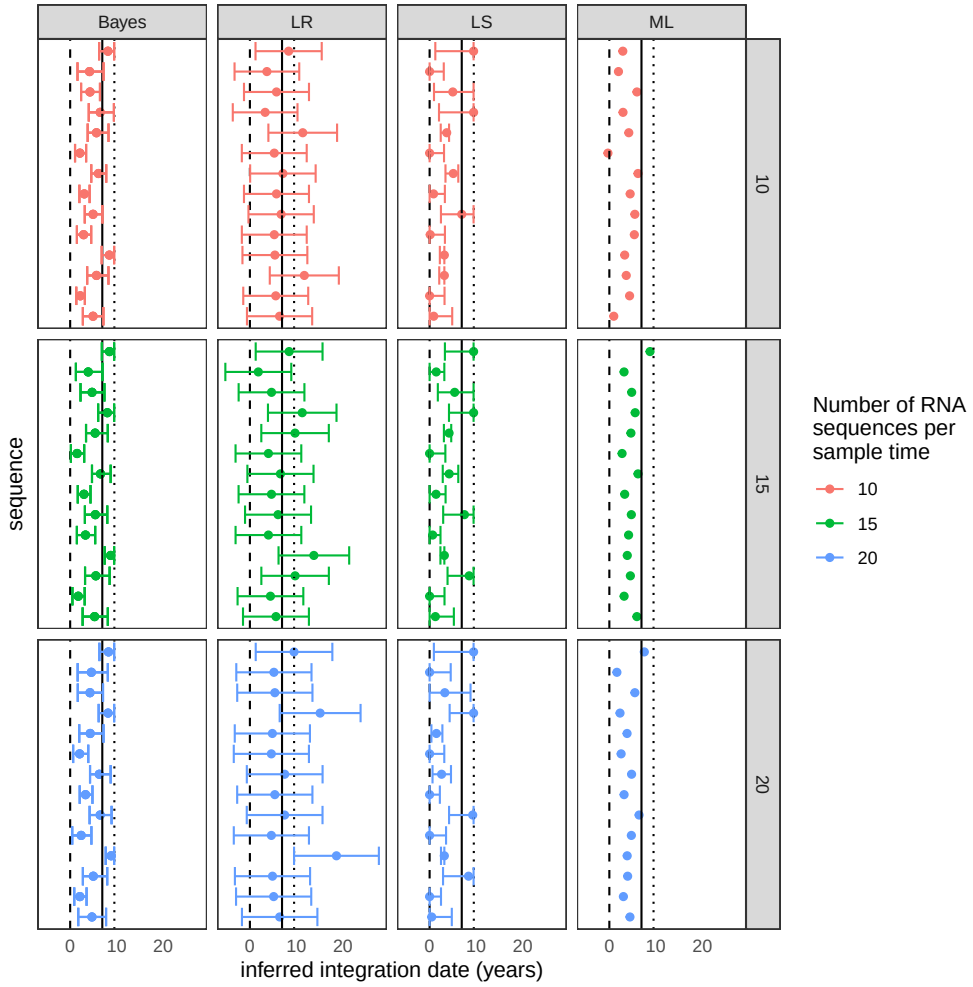


FIGURE 2.29. The inferred latent integration dates for Env_2 from patient 217 are shown for each method. 95% confidence intervals are shown for the LR and LS methods, and the 95% credible interval is shown for HIVTree. Sequences are shown in the same order in each panel. The vertical lines show the time of infection (dashed), time of treatment start (solid) and the time of sampling (dotted). The color shows the number of RNA sequences subsampled from the original alignment at each sample time. If fewer sequences were available then the number indicated by the color at a given time, all available sequences were used. Sites with greater than 95% missing gaps have been removed from the alignment.

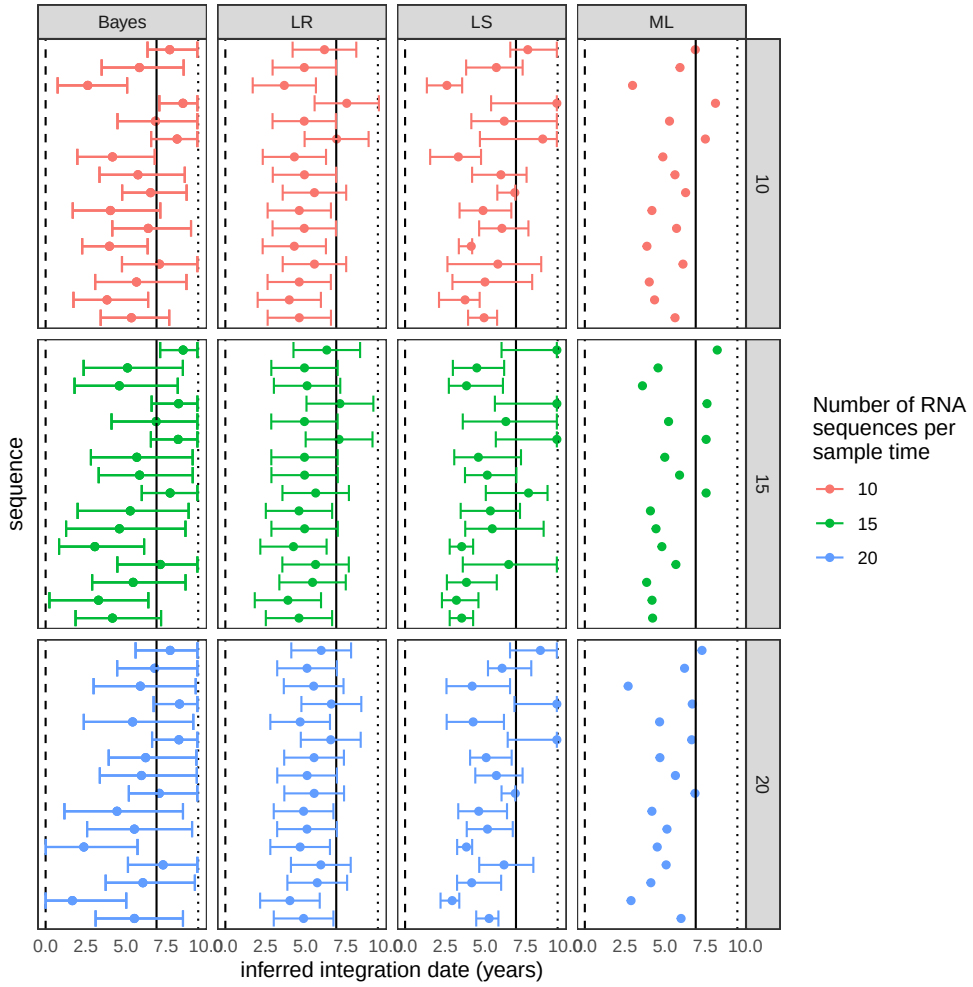


FIGURE 2.30. The inferred latent integration dates for Env_3 from patient 217 are shown for each method. 95% confidence intervals are shown for the LR and LS methods, and the 95% credible interval is shown for HIVTree. Sequences are shown in the same order in each panel. The vertical lines show the time of infection (dashed), time of treatment start (solid) and the time of sampling (dotted). The color shows the number of RNA sequences subsampled from the original alignment at each sample time. If fewer sequences were available then the number indicated by the color at a given time, all available sequences were used. Sites with greater than 75% missing gaps have been removed from the alignment.

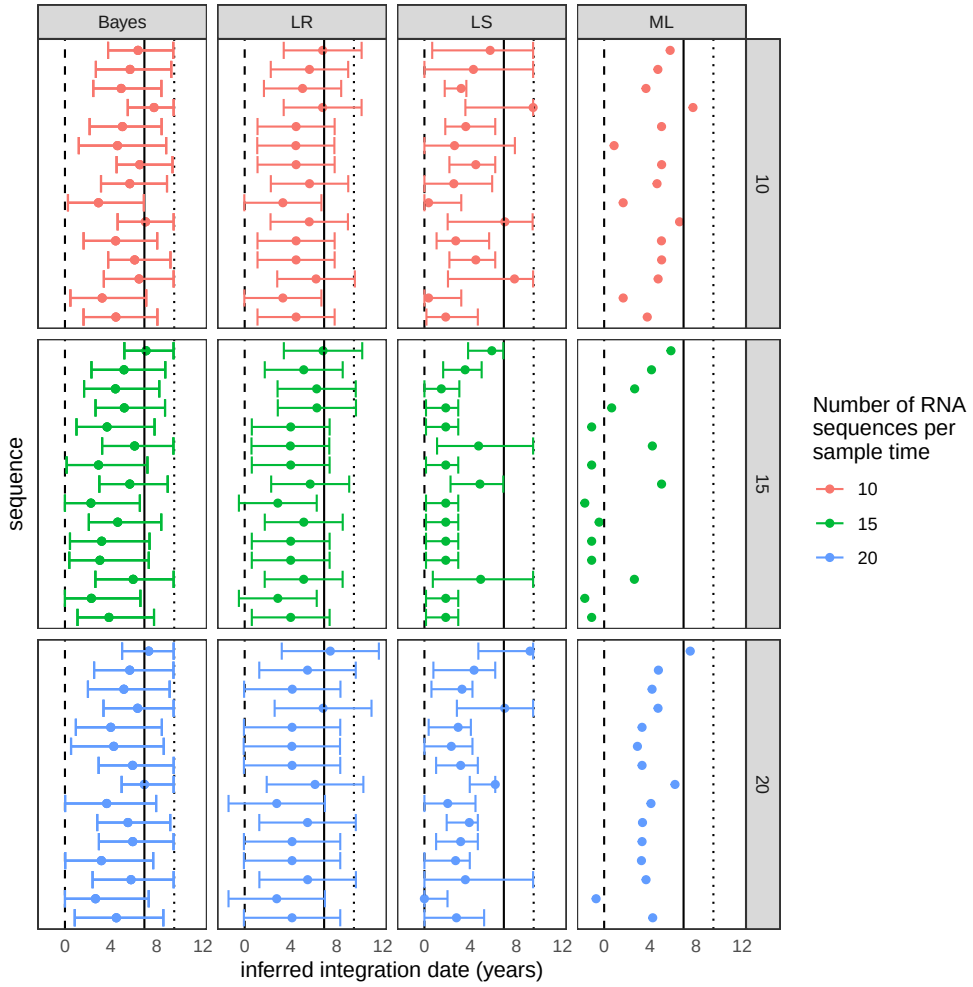


FIGURE 2.31. The inferred latent integration dates for Env_4 from patient 217 are shown for each method. 95% confidence intervals are shown for the LR and LS methods, and the 95% credible interval is shown for HIVTree. Sequences are shown in the same order in each panel. The vertical lines show the time of infection (dashed), time of treatment start (solid) and the time of sampling (dotted). The color shows the number of RNA sequences subsampled from the original alignment at each sample time. If fewer sequences were available then the number indicated by the color at a given time, all available sequences were used. Sites with greater than 75% missing gaps have been removed from the alignment.

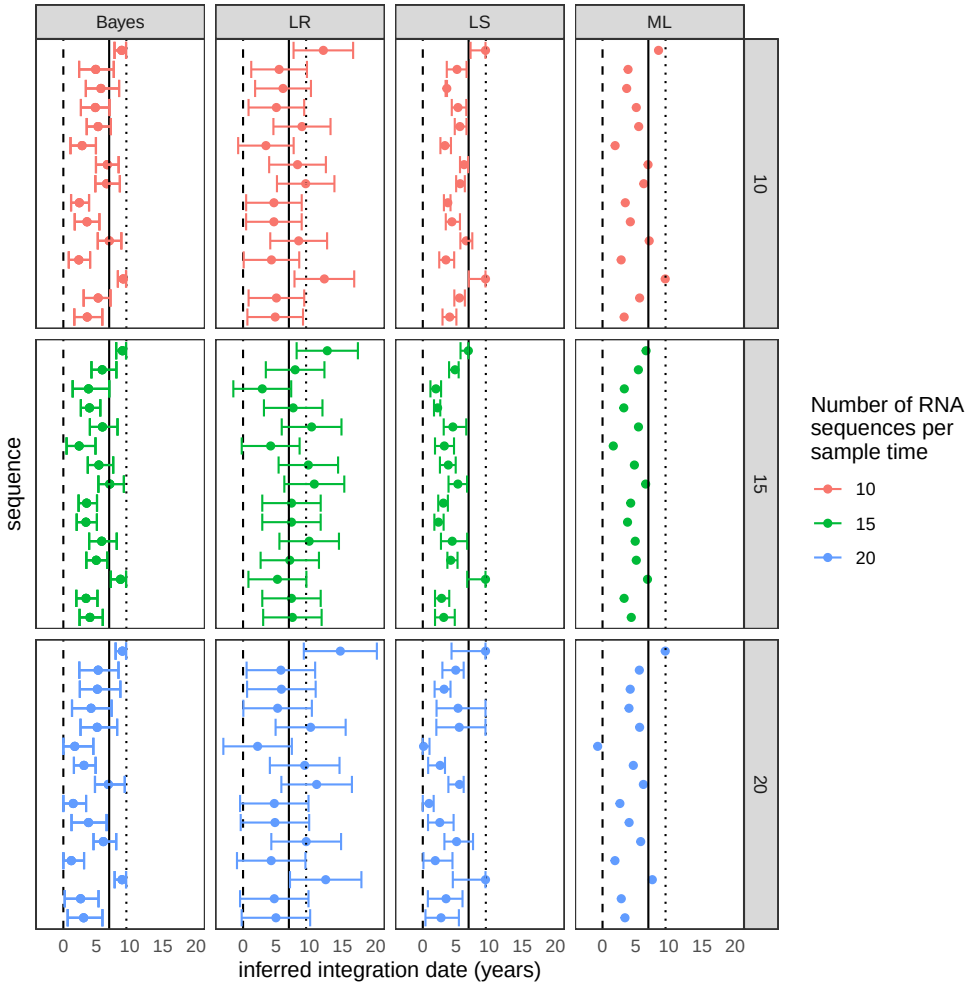


FIGURE 2.32. The inferred latent integration dates for Env_4 from patient 217 are shown for each method. 95% confidence intervals are shown for the LR and LS methods, and the 95% credible interval is shown for HIVTree. Sequences are shown in the same order in each panel. The vertical lines show the time of infection (dashed), time of treatment start (solid) and the time of sampling (dotted). The color shows the number of RNA sequences subsampled from the original alignment at each sample time. If fewer sequences were available then the number indicated by the color at a given time, all available sequences were used. Sites with greater than 95% missing gaps have been removed from the alignment.

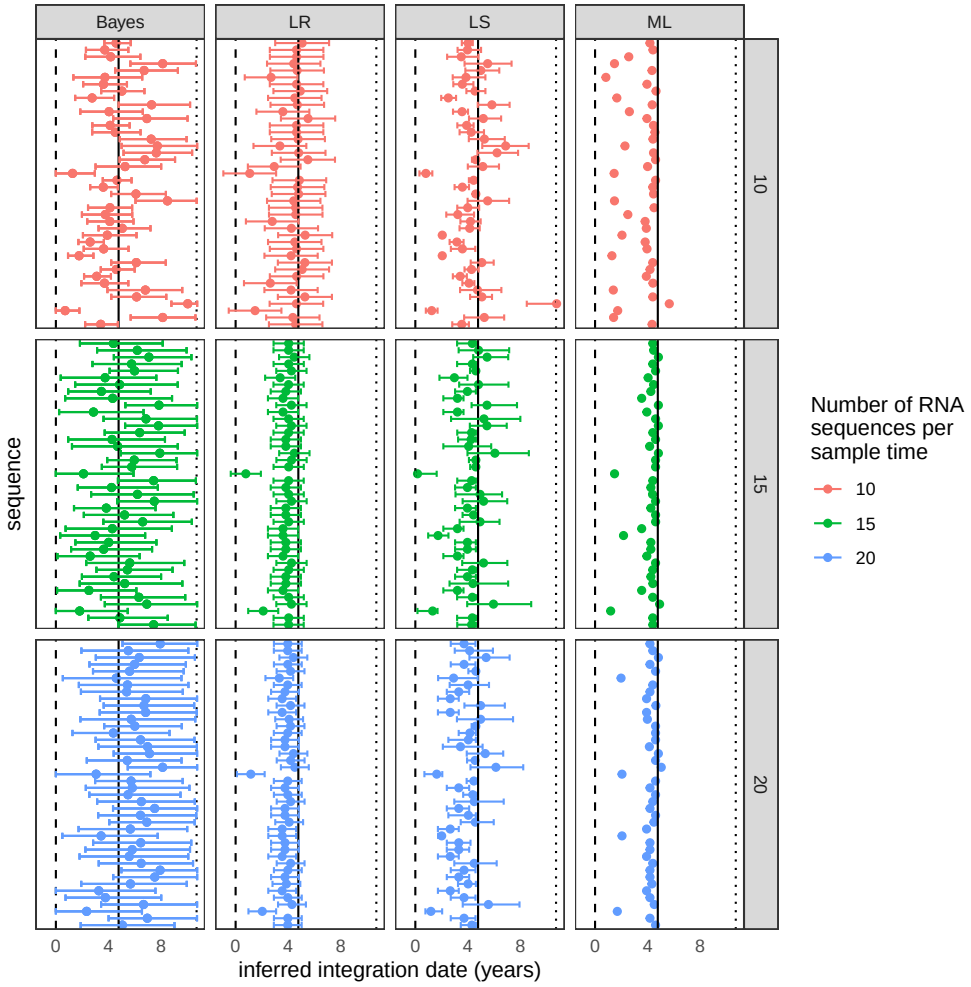


FIGURE 2.33. The inferred latent integration dates for Env_2 from patient 257 are shown for each method. 95% confidence intervals are shown for the LR and LS methods, and the 95% credible interval is shown for HIVTree. Sequences are shown in the same order in each panel. The vertical lines show the time of infection (dashed), time of treatment start (solid) and the time of sampling (dotted). The color shows the number of RNA sequences subsampled from the original alignment at each sample time. If fewer sequences were available then the number indicated by the color at a given time, all available sequences were used. Sites with greater than 75% missing gaps have been removed from the alignment.

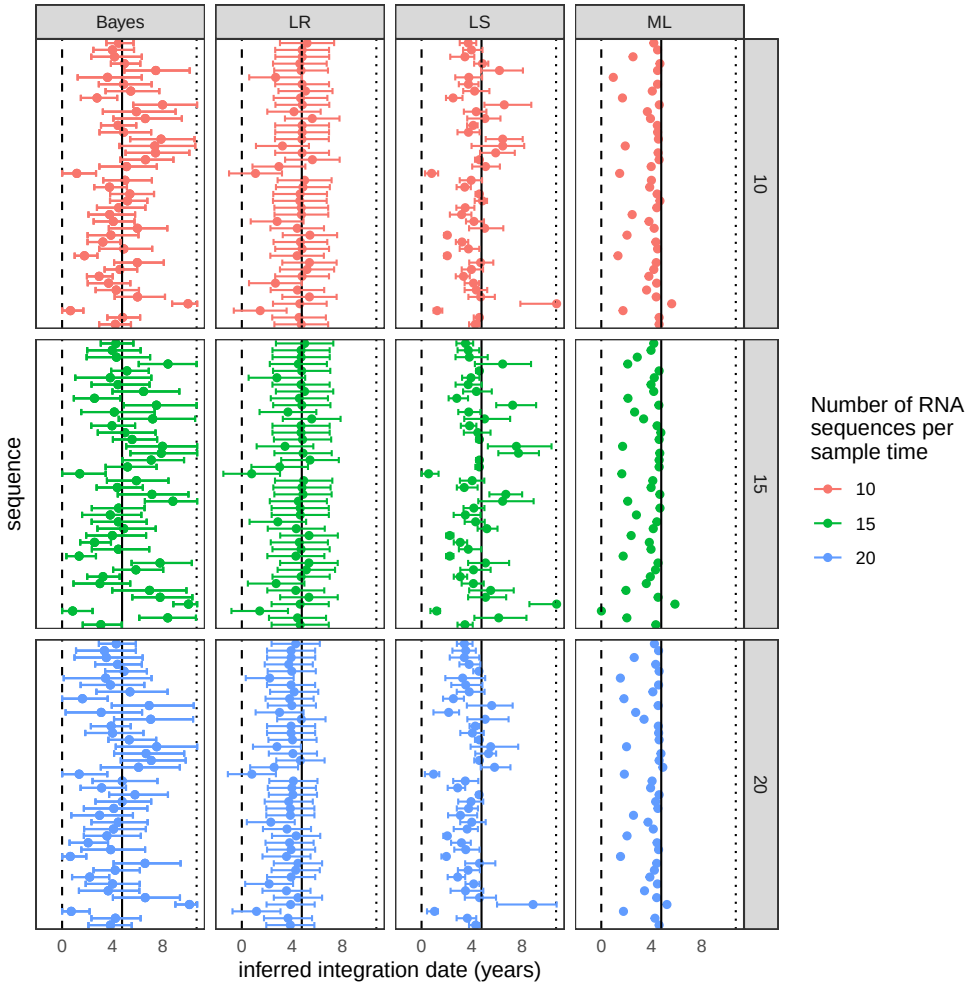


FIGURE 2.34. The inferred latent integration dates for Env_2 from patient 257 are shown for each method. 95% confidence intervals are shown for the LR and LS methods, and the 95% credible interval is shown for HIVTree. Sequences are shown in the same order in each panel. The vertical lines show the time of infection (dashed), time of treatment start (solid) and the time of sampling (dotted). The color shows the number of RNA sequences subsampled from the original alignment at each sample time. If fewer sequences were available then the number indicated by the color at a given time, all available sequences were used. Sites with greater than 85% missing gaps have been removed from the alignment.

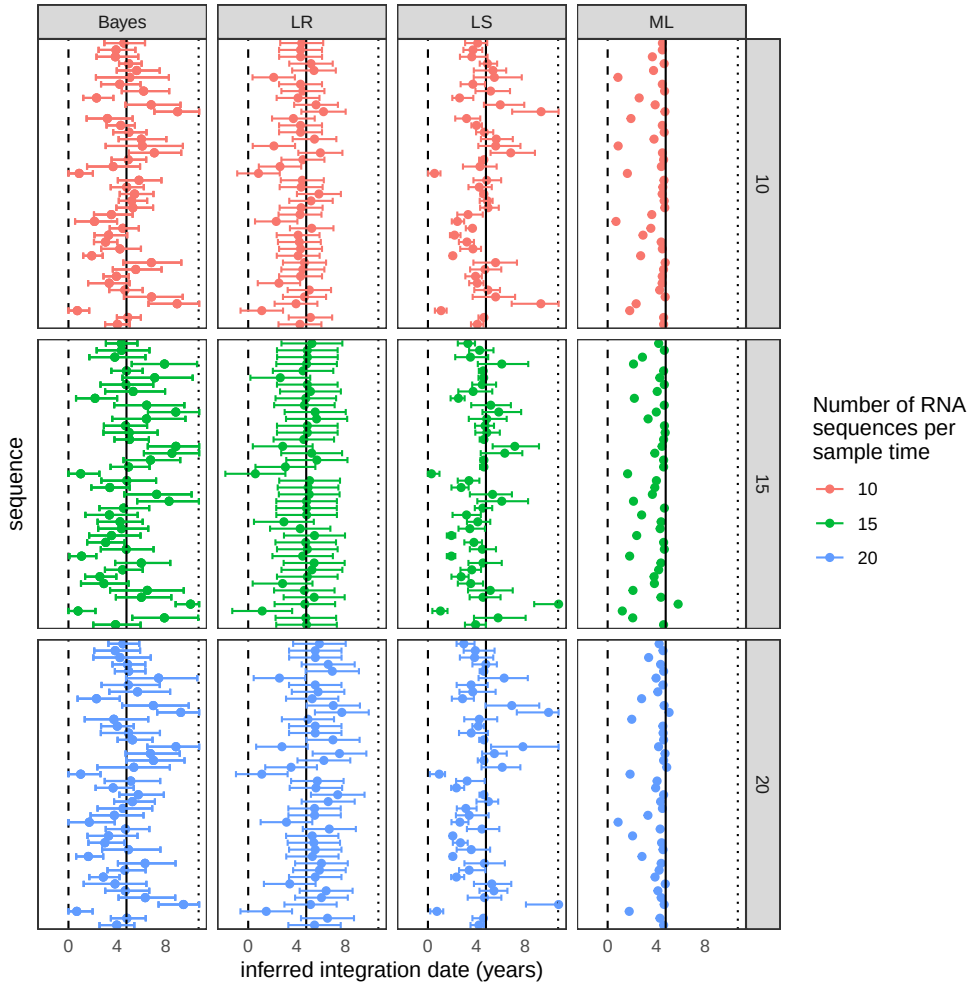


FIGURE 2.35. The inferred latent integration dates for Env_2 from patient 257 are shown for each method. 95% confidence intervals are shown for the LR and LS methods, and the 95% credible interval is shown for HIVTree. Sequences are shown in the same order in each panel. The vertical lines show the time of infection (dashed), time of treatment start (solid) and the time of sampling (dotted). The color shows the number of RNA sequences subsampled from the original alignment at each sample time. If fewer sequences were available then the number indicated by the color at a given time, all available sequences were used. Sites with greater than 95% missing gaps have been removed from the alignment.

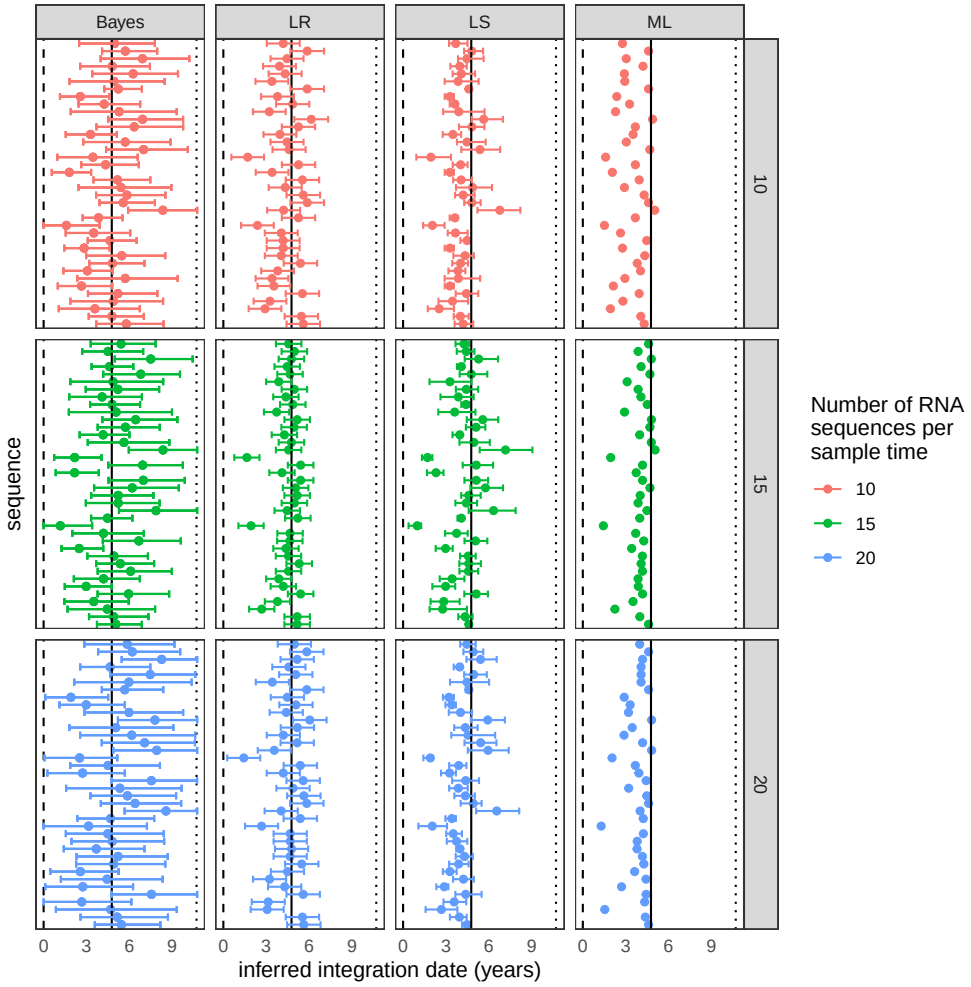


FIGURE 2.36. The inferred latent integration dates for Env_3 from patient 257 are shown for each method. 95% confidence intervals are shown for the LR and LS methods, and the 95% credible interval is shown for HIVTree. Sequences are shown in the same order in each panel. The vertical lines show the time of infection (dashed), time of treatment start (solid) and the time of sampling (dotted). The color shows the number of RNA sequences subsampled from the original alignment at each sample time. If fewer sequences were available then the number indicated by the color at a given time, all available sequences were used. Sites with greater than 75% missing gaps have been removed from the alignment.

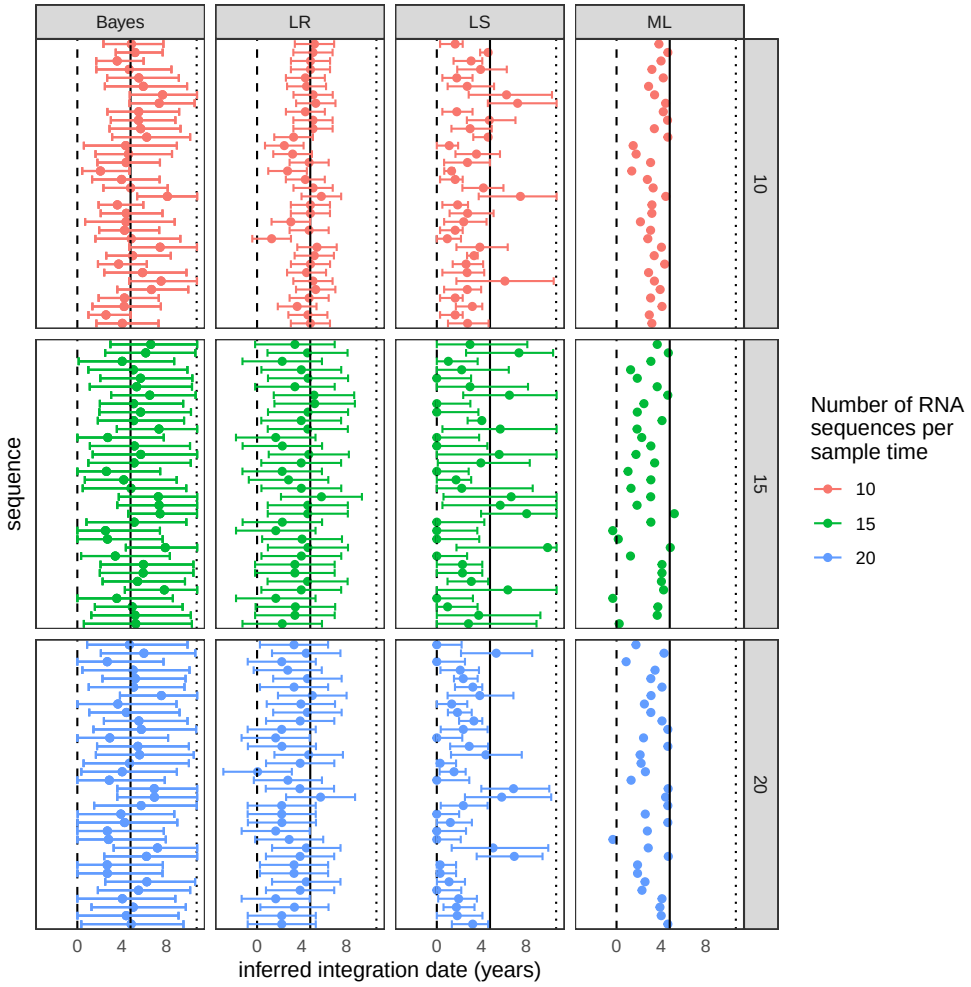


FIGURE 2.37. The inferred latent integration dates for Env_4 from patient 257 are shown for each method. 95% confidence intervals are shown for the LR and LS methods, and the 95% credible interval is shown for HIVTree. Sequences are shown in the same order in each panel. The vertical lines show the time of infection (dashed), time of treatment start (solid) and the time of sampling (dotted). The color shows the number of RNA sequences subsampled from the original alignment at each sample time. If fewer sequences were available then the number indicated by the color at a given time, all available sequences were used. Sites with greater than 75% missing gaps have been removed from the alignment.

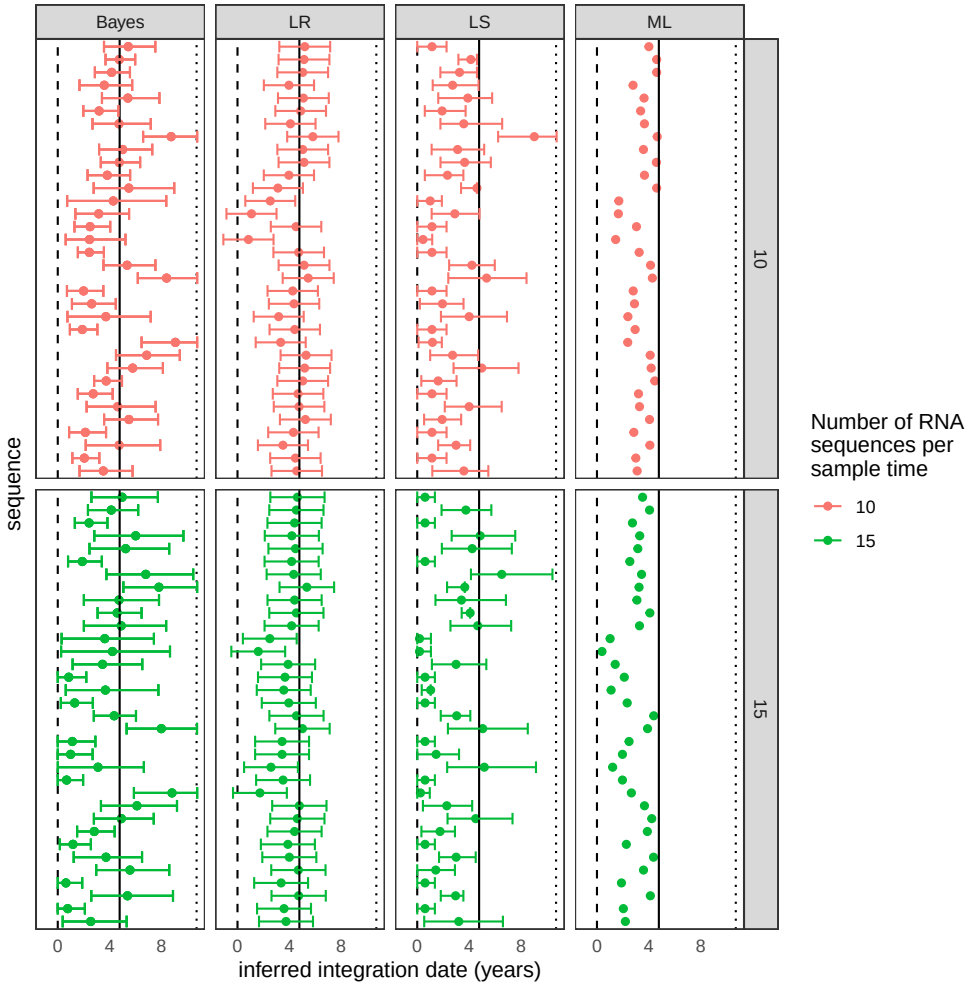


FIGURE 2.38. The inferred latent integration dates for Env_4 from patient 257 are shown for each method. 95% confidence intervals are shown for the LR and LS methods, and the 95% credible interval is shown for HIVTree. Sequences are shown in the same order in each panel. The vertical lines show the time of infection (dashed), time of treatment start (solid) and the time of sampling (dotted). The color shows the number of RNA sequences subsampled from the original alignment at each sample time. If fewer sequences were available then the number indicated by the color at a given time, all available sequences were used. Sites with greater than 85% missing gaps have been removed from the alignment.

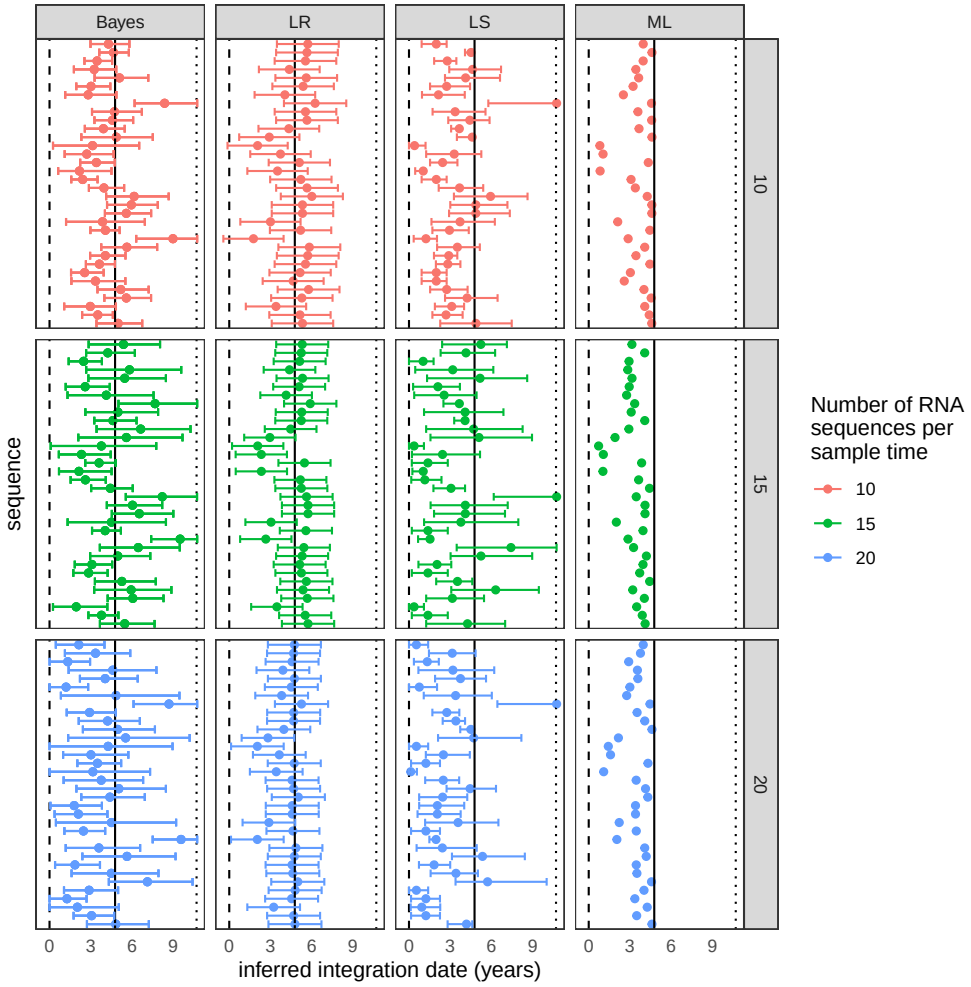


FIGURE 2.39. The inferred latent integration dates for Env_4 from patient 257 are shown for each method. 95% confidence intervals are shown for the LR and LS methods, and the 95% credible interval is shown for HIVTree. Sequences are shown in the same order in each panel. The vertical lines show the time of infection (dashed), time of treatment start (solid) and the time of sampling (dotted). The color shows the number of RNA sequences subsampled from the original alignment at each sample time. If fewer sequences were available then the number indicated by the color at a given time, all available sequences were used. Sites with greater than 95% missing gaps have been removed from the alignment.

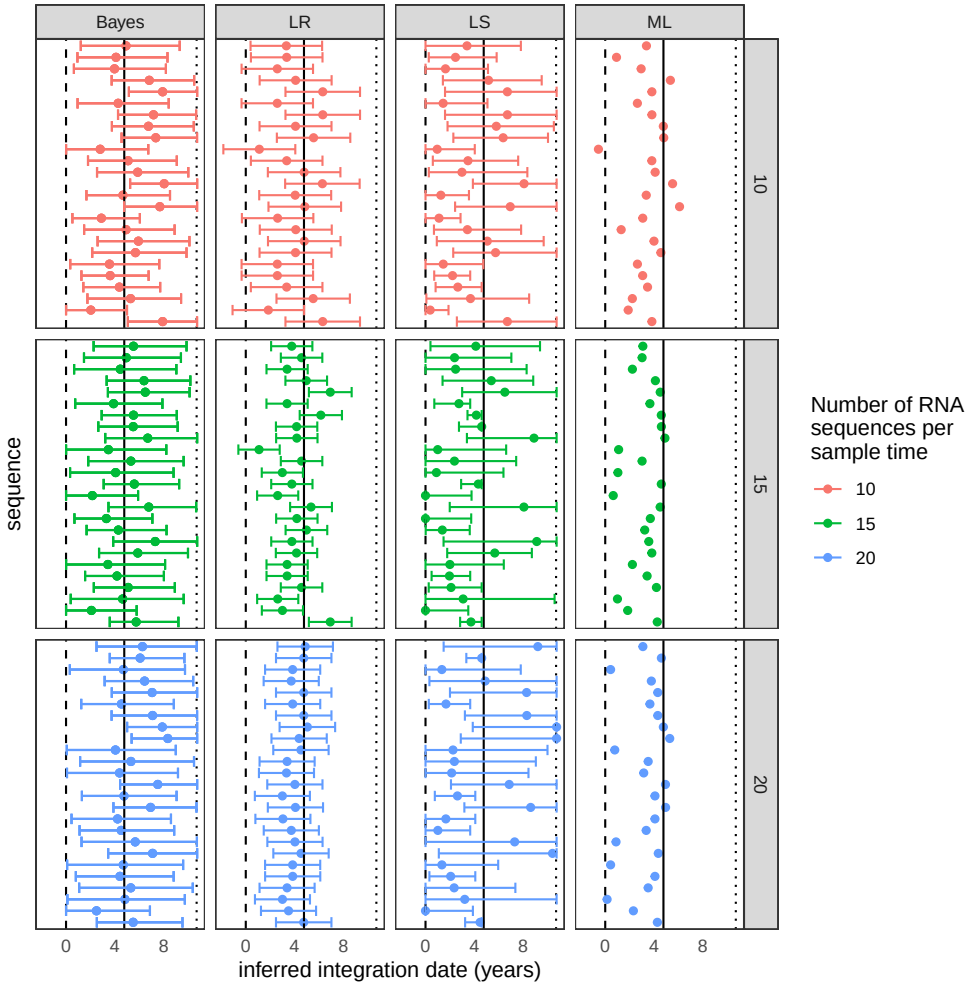


FIGURE 2.40. The inferred latent integration dates for GAG_1 from patient 257 are shown for each method. 95% confidence intervals are shown for the LR and LS methods, and the 95% credible interval is shown for HIVTree. Sequences are shown in the same order in each panel. The vertical lines show the time of infection (dashed), time of treatment start (solid) and the time of sampling (dotted). The color shows the number of RNA sequences subsampled from the original alignment at each sample time. If fewer sequences were available then the number indicated by the color at a given time, all available sequences were used. Sites with greater than 75% missing gaps have been removed from the alignment.

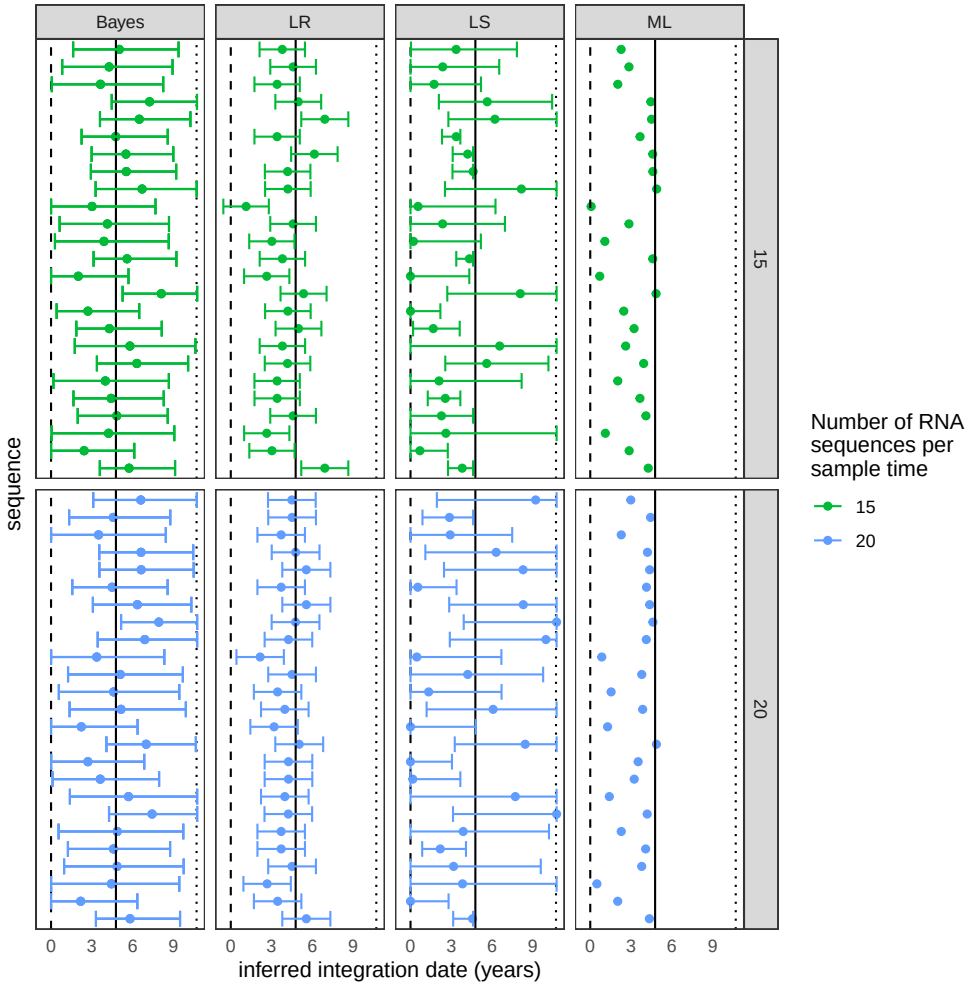


FIGURE 2.41. The inferred latent integration dates for GAG_1 from patient 257 are shown for each method. 95% confidence intervals are shown for the LR and LS methods, and the 95% credible interval is shown for HIVTree. Sequences are shown in the same order in each panel. The vertical lines show the time of infection (dashed), time of treatment start (solid) and the time of sampling (dotted). The color shows the number of RNA sequences subsampled from the original alignment at each sample time. If fewer sequences were available then the number indicated by the color at a given time, all available sequences were used. Sites with greater than 95% missing gaps have been removed from the alignment.

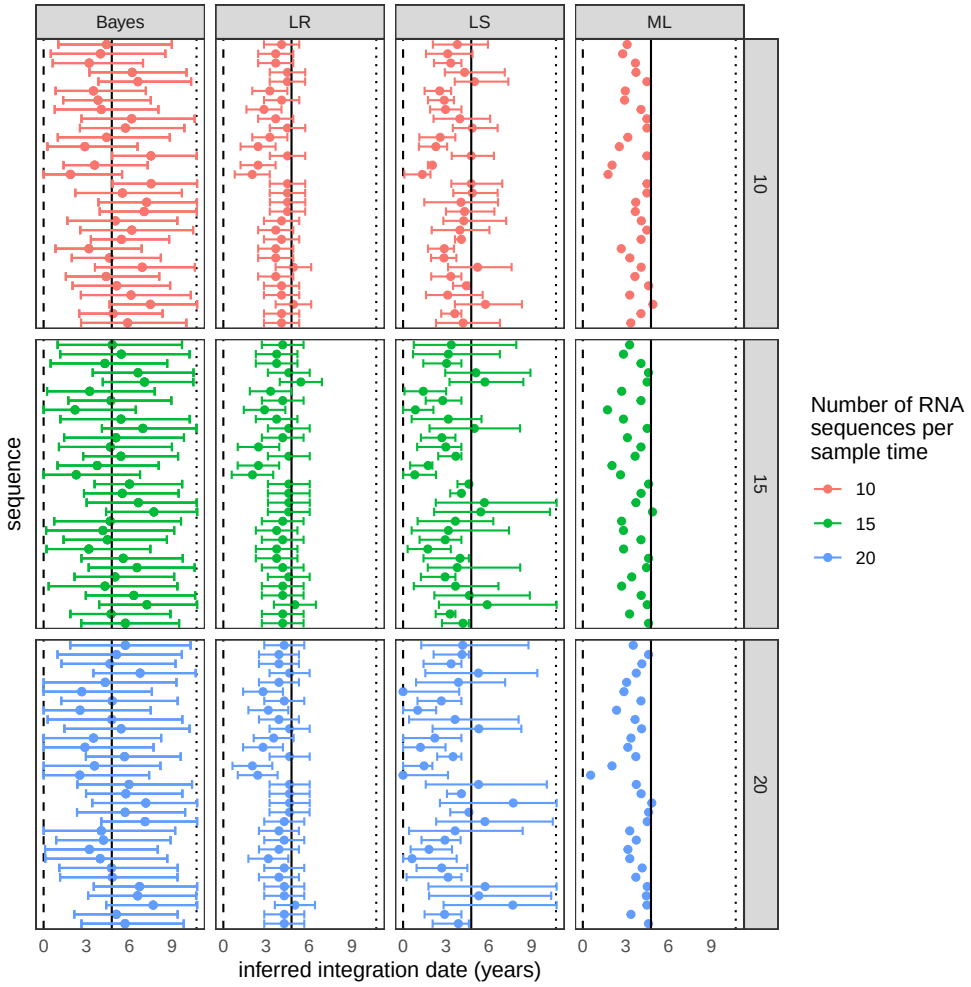


FIGURE 2.42. The inferred latent integration dates for NEF_1 from patient 257 are shown for each method. 95% confidence intervals are shown for the LR and LS methods, and the 95% credible interval is shown for HIVTree. Sequences are shown in the same order in each panel. The vertical lines show the time of infection (dashed), time of treatment start (solid) and the time of sampling (dotted). The color shows the number of RNA sequences subsampled from the original alignment at each sample time. If fewer sequences were available then the number indicated by the color at a given time, all available sequences were used. Sites with greater than 75% missing gaps have been removed from the alignment.

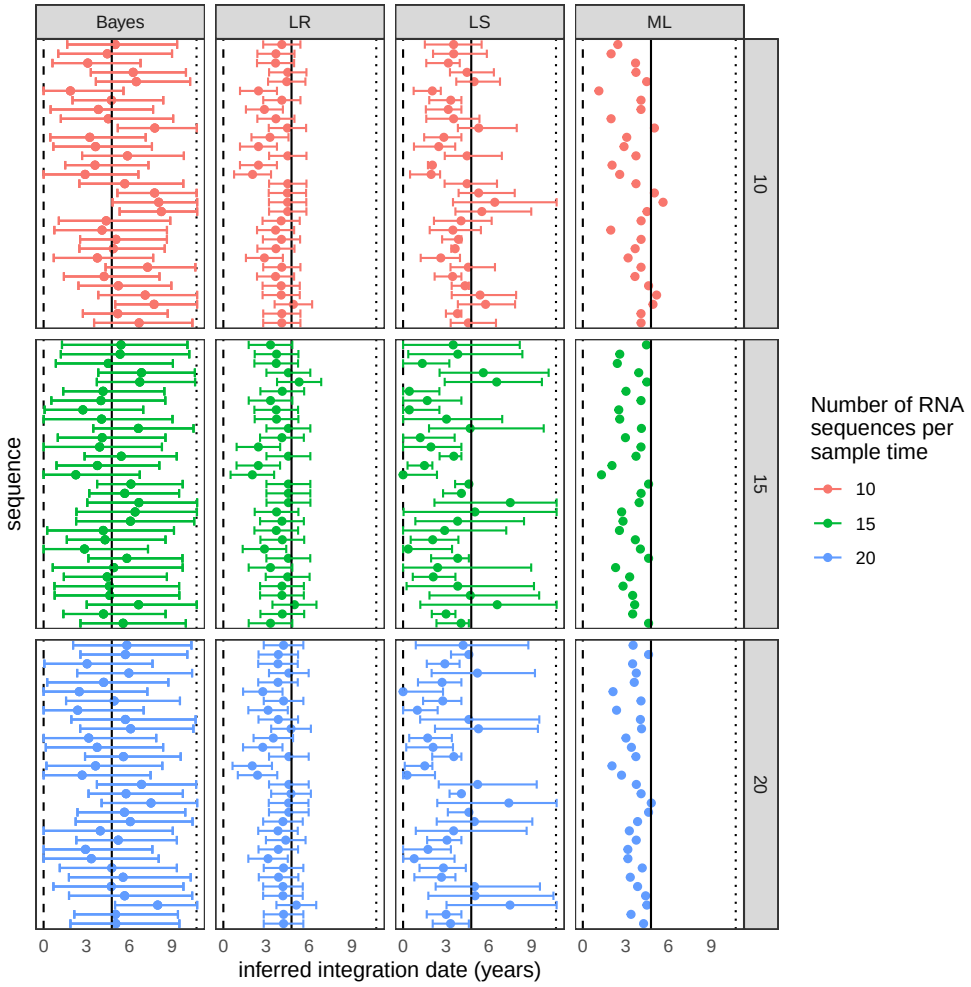


FIGURE 2.43. The inferred latent integration dates for NEF_1 from patient 257 are shown for each method. 95% confidence intervals are shown for the LR and LS methods, and the 95% credible interval is shown for HIVTree. Sequences are shown in the same order in each panel. The vertical lines show the time of infection (dashed), time of treatment start (solid) and the time of sampling (dotted). The color shows the number of RNA sequences subsampled from the original alignment at each sample time. If fewer sequences were available then the number indicated by the color at a given time, all available sequences were used. Sites with greater than 95% missing gaps have been removed from the alignment.

2.8. Discussion

Here, we have described both a phylogenetic method to infer latent integration times and a new method to simulate sequence data based on within-host viral dynamics. While there is currently only a limited amount of data suitable for this method, future research will be able to utilize this tool to make statistically justified conclusions about latent integration times. Our method does not directly address how the composition of the reservoir changes over time and the rates at which latent sequences enter the latent reservoir. However, accurate dating of the entry of individual sequences into the latent reservoir will likely be necessary to answer these questions.

HIVTREE performs better than existing methods by a variety of metrics. The method has smaller credible/confidence intervals on average than alternative methods, while still containing the true value with high probability, resulting in more precise interval estimates of the integration dates. The RMSE of HIVTREE was slightly lower on average than the other methods for the largest gene, the average RMSE is comparable across all methods, with a difference of about 2 months between the method with the lowest (Bayes) and highest (ML) RMSE. The larger RMSE for HIVTREE for small regions is likely due to the influence of the uninformative prior in low information cases which increases bias; more informative priors based on other sources of information could help reduce RMSE.

HIVTREE has several improvements over existing methods. It allows for biologically relevant bounds on latent integration times, such as requiring the latent times be older than the sample times with an option to bound the integration times at the time of infection. Among the alternative methods, only the LS method allows for such bounds. Bayesian inference also provides a sensible way to combine estimates across genes or genomic regions, while allowing for potentially different gene tree topologies due to recombination. This results in more precise estimates, especially when the sequences available are short. There is currently no alternative to the HIVTREE method for jointly inferring latency times using multiple loci, nor is there a clear way to do so. Because each locus is relatively short, combining information across loci can greatly improve precision of latency time estimates. Lastly, Bayesian methods have the advantage, shared by other likelihood based methods, of well known statistical properties, such as statistical efficiency and consistency. By treating an alignment as data, HIVTREE allows for full use of the available sequence data in the

inference, whereas the other methods only use an inferred phylogenetic tree which may not be a sufficient statistic.

The simulation model in the study allows for the possibility of a lineage becoming latent, reactivating, and potentially becoming latent again. However, the inference model assumes that latency can only occur at the tips of the tree. It is possible that some of the lineages in the trees have ancestral periods of latency that are not accounted for in the inference model. This does not appear to have a large impact on the results, as the inference method performs well on average. The simulation model did not include cases where an individual was on effective treatment and then ceased treatment. In this case, most of the virus in blood may be derived from sequences that were latent at some point in the past. For this reason, we caution against using this method for patients with a history of multiple periods of treatment with periods of uncontrolled viremia in between.

There are several avenues for improvement of HIVtree. In the current paper, to use data from multiple loci in HIVtree the marginal distributions for the latent integration times were combined. A more formal method to combine data across loci would be to jointly analyze the loci in a single model, allowing the MCMC to integrate over the node ages in each of gene trees separately while constraining the latent integration times to be the same for sequences derived from an individual infected cell. It would also be better to integrate over the different tree topologies, rather than fix the tree topology as is done in this method and existing methods. This would be most easily implemented in a program that accommodates multilocus data and estimates gene trees, such as bpp (Flouri *et al.*, 2018), rather than the parent program of HIVtree, MCMCtree.

Furthermore, despite the desirability of a diffuse prior on the node ages and latent times, the prior model in HIVtree seems to be too informative in some cases. The rank order of the nodes and the serial sampling cause the average root age of the phylogeny in the prior to be older than the user input prior. If the root age is constrained, such as by using a uniform prior, the latent times are pushed closer to the present time by the prior, which introduces a bias to the latent inferences (unpublished preliminary analysis). This means that constraining the root age to be close to the true age can increase the influence of the prior, leading to worse estimates of the latent times. Similar effects driven by constraints among node ages have been previously noted for fossil

calibrations and serially sampled data (Stadler and Yang, 2013; Yang and Rannala, 2006). However, the effects appear to be more pronounced when the root ages are close to the the serially sampled sequences, as can result from within-host viral data. While there may be quite informative outside knowledge on the age of the root for HIV, such as the time of infection, we currently caution against forcing the root age to match the infection time when using HIVTREE because this may induce bias in estimates of latent virus integration times.

The difference between the user input prior distribution on the root age and the prior observed when running the MCMC without data appears to be larger with the empirical data than with the simulated datasets. While the exact cause of this discrepancy is unknown, it may be related to the ladder-like tree topologies of the empirical data or the sampling times of the sequences. A different prior may improve some of these limitations. One option would be a serial sample coalescent prior with changing populations sizes (Minin *et al.*, 2008; Rodrigo and Felsenstein, 1999). This would also be more sensible to implement in a program which includes coalescent models, such as bpp. Such a prior could also allow for the incorporation of information on viral population sizes (such as from well described viral dynamic models) and knowledge of the time of infection.

The viral dynamic simulation method developed in this paper is based on well-studied models of HIV population dynamics within hosts. This is likely to be more realistic than traditional methods used to simulate phylogenies, such as constant rate birth-death processes, and it follows standard epidemiology approaches for studying viral dynamics. However, this model does not incorporate selection or recombination, which are known to be important in HIV evolution and effect tree topology. The method produces trees that are more star-like, with short internal branches, than those typically inferred in empirical studies of HIV sequences. Future work should focus on simulating selection and recombination, as well as other aspects of HIV biology, such as clonal proliferation of latently infected immune cells, which may impact tree topology and latent histories. Additionally, researchers should investigate different priors for inference that may more accurately model HIV biology and produce trees that more closely match the empirical observations, such as the ladder-like nature of many within-host HIV phylogenies.

2.9. Funding

A.A.N. was supported by the National Science Foundation Graduate Research Fellowship Program under Grant No.2036201. This research was supported by National Institutes of Health Grant GM123306 to Bruce Rannala.

CHAPTER 3

Bayesian Inference Under the Multispecies Coalescent with ancient DNA sequences

Ancient DNA (aDNA) is increasingly being used to investigate questions such as the phylogenetic relationships and divergence times of extant and extinct species. If aDNA samples are sufficiently old, expected branch lengths (in units of DNA substitutions) are reduced relative to contemporary samples. This can be accounted for by incorporating sample ages into phylogenetic analyses. Existing methods that use tip (sample) dates infer gene trees rather than species trees, which can lead to incorrect or biased inferences of the species tree. Methods using a multispecies coalescent (MSC) model overcome these issues. We developed an MSC model with tip dates and implemented it in the program BPP. The method performed well for a range of biologically realistic scenarios, estimating calibrated divergence times and mutation rates precisely. Simulations suggest that estimates can be best improved by prioritizing sampling of many loci and more ancient samples. Incorrectly treating ancient samples as contemporary in analyzing simulated data, mimicking a common practice of empirical analyses, led to large systematic biases in model parameters, including divergence times. Two genomic datasets of mammoths and elephants were analyzed, demonstrating the method's empirical utility.

3.1. Introduction

Ancient DNA (aDNA) sequences are increasingly available for many species due to advances in sequencing technology. Whole genome sequences from aDNA exist for several groups of extinct species, including neanderthals (Green *et al.*, 2010), woolly and Columbian mammoths (Palkopoulou *et al.*, 2015, 2018), woolly rhinoceros (Lord *et al.*, 2020) and cave bears (Fortes *et al.*, 2016). Genome sequences from aDNA also exist for many extant species, for example humans (Nielsen *et al.*, 2017; Rasmussen *et al.*, 2010) and maize (Ramos-Madrigal *et al.*, 2016). More limited aDNA data are

available for an even wider variety of species such as bison (Soubrier *et al.*, 2016), polar bears (Miller *et al.*, 2012), pigs (Horsburgh *et al.*, 2022) and many plants and pathogens (Orlando *et al.*, 2021). These data have opened the door to new ways to investigate long-standing questions in phylogenetics and population genetics, such as phylogenetic relationships between extinct and extant species, their divergence times, and their demographic and migration histories.

A key feature that distinguishes aDNA from modern DNA is the (potentially large) differences in ages among sampled aDNA sequences; in conventional studies of modern DNA all samples are contemporary. The importance of accounting for the sampling date of non-contemporary sequences has long been recognized for viral sequences, in particular RNA viruses (Drummond *et al.*, 2003). Due to the high substitution rates of RNA viruses, lineages that are sampled later have more opportunities for substitutions to occur, creating differences in expected branch lengths between lineages descended from a common ancestor, even under a strict molecular clock. With molecular sequence data, the amount of evolution observed is determined by the product of substitution rate and time. Sequences sampled at different times may have detectable differences in expected substitutions if either the mutation rate is high (as with viral data) or the time interval between sampling events is large (as with older aDNA samples). Similar to fossil calibrations, sampling dates provide information about substitution rates, allowing absolute divergence times (e.g., days or years) and absolute substitution rates to be jointly estimated (Li *et al.*, 1988; Rambaut, 2000).

Whole genomes of aDNA contain much information for detecting even small differences of expected numbers of substitutions; one might speculate that increasing the number of loci will improve estimates of parameters such as absolute divergence times and mutation rate even with younger samples because each locus is an independent source of information. As more loci are added, the expected difference in branch lengths between lineages sampled at different times is more precisely estimated, thus improving estimates of both mutation rate and absolute divergence times. An advantage of dating with aDNA samples over fossil calibrations is that the position of the sample in the phylogeny can potentially be inferred from the sequence data whereas fossils must be assigned to ancestral nodes based solely on sparse morphological characters and are probably frequently misassigned.

Another reason to develop statistical models for analyzing aDNA is the potential for biased estimates if sample dates are ignored. Several studies have analyzed aDNA by treating all the samples (including aDNA) as contemporary (e.g. Palkopoulou *et al.*, 2018; Rohland *et al.*, 2010). This should lead to underestimation of divergence times. It is poorly understood how great the absolute time interval between samples must be before it affects inference when sampling dates are not explicitly modeled.

3.1.1. Analyses of aDNA without Tip Dates. Population samples of aDNA have been analyzed using several methods which do not explicitly use sampling dates. Two of these, pairwise sequential Markovian coalescent (PSMC) (Li and Durbin, 2011) and coalHMM (Mailund *et al.*, 2012), are commonly used methods for inferring ancestral demography (past effective population size through time) based on an approximation to the coalescent process with recombination. However, both allow inference for small samples only (2 or 3 individuals). In order to estimate population sizes in calendar time, mutation rate and generation time are treated as known in PSMC, though both are uncertain. When two or a few individuals have been sampled that share an ancestral population, researchers have used PSMC independently on the samples and then aligned the demographic histories inferred with PSMC to determine when the populations diverged. This is problematic because data from different individuals are analyzed independently and divergence times are not estimated directly.

With population genetic data from multiple species, multispecies coalescent (MSC) models in MCMCcoal (an early version of BPP) have been used to infer divergence times and effective population sizes with aDNA, treating the ancient sequences as if they were contemporary (Rohland *et al.*, 2010). The effect of ignoring sample ages is unknown for programs such as coalHMM and MCMCcoal, but may bias inferences.

3.1.2. Analyses of aDNA with Tip Dates. The programs BEAST and BEAST2 explicitly use sample dates in the analysis of aDNA (Bouckaert *et al.*, 2019; Suchard *et al.*, 2018). With data from a single species, BEAST can infer demographic histories on an absolute time scale (Bouckaert *et al.*, 2019; Suchard *et al.*, 2018) and accommodate multilocus datasets, allowing different tree topologies for different genes and requiring some parameters to be shared across genes. However,

BEAST does not employ the MSC to infer divergence times, and ignores the difference between the gene trees and the species tree. One option is to use the divergence times of different clades in gene trees as an estimate of the species divergence time (e.g., Chang *et al.*, 2017). This approach overestimates divergence times since the common ancestor of a gene must be older than the common ancestor of the species (Angelis and Dos Reis, 2015; Gillespie and Langley, 1979).

The package starBEAST3 in BEAST2 uses an MSC model and can accommodate tip dates (Douglas *et al.*, 2022), estimating divergence times, effective population sizes and mutation rate. However, it assumes all samples from a particular species are the same age, which is uncommon for aDNA samples. Approximate Bayesian computation (ABC) has also been used with aDNA to investigate complex histories that include migration and can in principle use tip dates. ABC has a number of practical issues, however, such as the need to identify good summary statistics and severe limitations on the size of the possible state space (number of populations, etc.) (Lintusaari *et al.*, 2016).

3.1.3. Prospects for MSC Analysis of aDNA. Phylogenetic methods based on the multi-species coalescent (MSC), such as BPP and starBEAST3, provide a more realistic model to analyze sequence data from multiple species or populations. These methods can estimate divergence times and effective population sizes and a variety of migration and hybridization histories. The BPP program allows analyses of datasets of thousands of loci, many individuals per population and multiple populations (or species) (Flouri *et al.*, 2018). Moreover, the methods are statistically consistent assuming the model is correct and make complete use of all information available in the data.

Here, we describe an MSC model with tip dates that allows any number of distinct sampling times within each population (or species). We implement this model in the Bayesian phylogenetic inference program BPP. We assess the performance of the method using simulations under a variety of population histories and investigate the impact of incorrectly treating ancient sequences as contemporary. We apply the new method to analyze two elephant and mammoth nuclear DNA and mtDNA datasets.

3.2. Methods

3.2.1. Theory: Overview of the MSC Model with Tip Dating. The standard MSC assumes that all sequences are sampled at time present, which is incorrect when using ancient samples. Here, the MSC is modified to allow a joint analysis of ancient and modern samples. We assume a fixed species tree topology, Ψ. Let Θ be the vector of parameters of the species tree, $\Theta = \{\boldsymbol{\tau}, \boldsymbol{\theta}\}$, where $\boldsymbol{\tau}$ is the vector of speciation times and $\boldsymbol{\theta}$ is the vector effective population sizes. Both $\boldsymbol{\tau}$ and $\boldsymbol{\theta}$ are measured in units of expected number of mutations. Let $\boldsymbol{X} = \boldsymbol{x}_i$ be the sequence data at locus i. Let $\boldsymbol{G} = \{G_i\}$ be the gene trees, where G_i is the gene tree at locus i and includes both the topology and coalescent times.

Let the mutation rate for all loci be μ. Since the MSC models within population genetic processes, the rate of DNA change measured is the mutation rate rather than the substitution rate. The assumption of a constant μ may be best suited for non-coding datasets. We assume a strict molecular clock. All parameters in the model are scaled by the expected number of mutations. By estimating μ, we can convert between calendar time and time in expected number of mutations. Let $y^{\triangle}$ denote a parameter y in units of years before present (ybp). For example, let $\tau^{\triangle}$ be speciation time in ybp and μ have units of mutations per year. Then, $\tau^{\triangle} = \tau/\mu$.

The joint posterior probability of the divergence times, effective population sizes, and gene trees is given by

$$f(\Theta, \boldsymbol{G}, \mu | \boldsymbol{X}, \Psi) = \frac{f(\boldsymbol{G}|\Theta, \Psi) \cdot P(X|\boldsymbol{G}, \mu) \cdot f(\mu, \Theta)}{P(\boldsymbol{X}|\Psi)}$$

This is analogous to Eqn. 4 in Rannala and Yang (2003). We need to calculate the gene-tree density $f(\boldsymbol{G}|\Theta, \Psi)$ when the sampled tips of the gene tree are not contemporaneous (have different absolute ages). In a single population, let there be E distinct sampling epochs. The sampling times in expected number of mutations are denoted t_{si} and are ordered such that $t_{s1} < t_{s2} < ... < t_{s(E-1)} < t_{sE}$. $t_{si}^{\triangle}$ denotes the age of the samples in years before present. Let t_{s0} be the time that the population ends (either time present or the age of the next population divergence event) and $t_{s(E+1)}$ be the speciation time of the ancestral population. At sample time t_{si}, there are m_i lineages sampled. Let the number of lineages surviving to time t_{si} excluding the number of new lineages sampled at time t_{si} be denoted n_i. Let $m_0 = 0$ and the waiting time for the coalescent event which

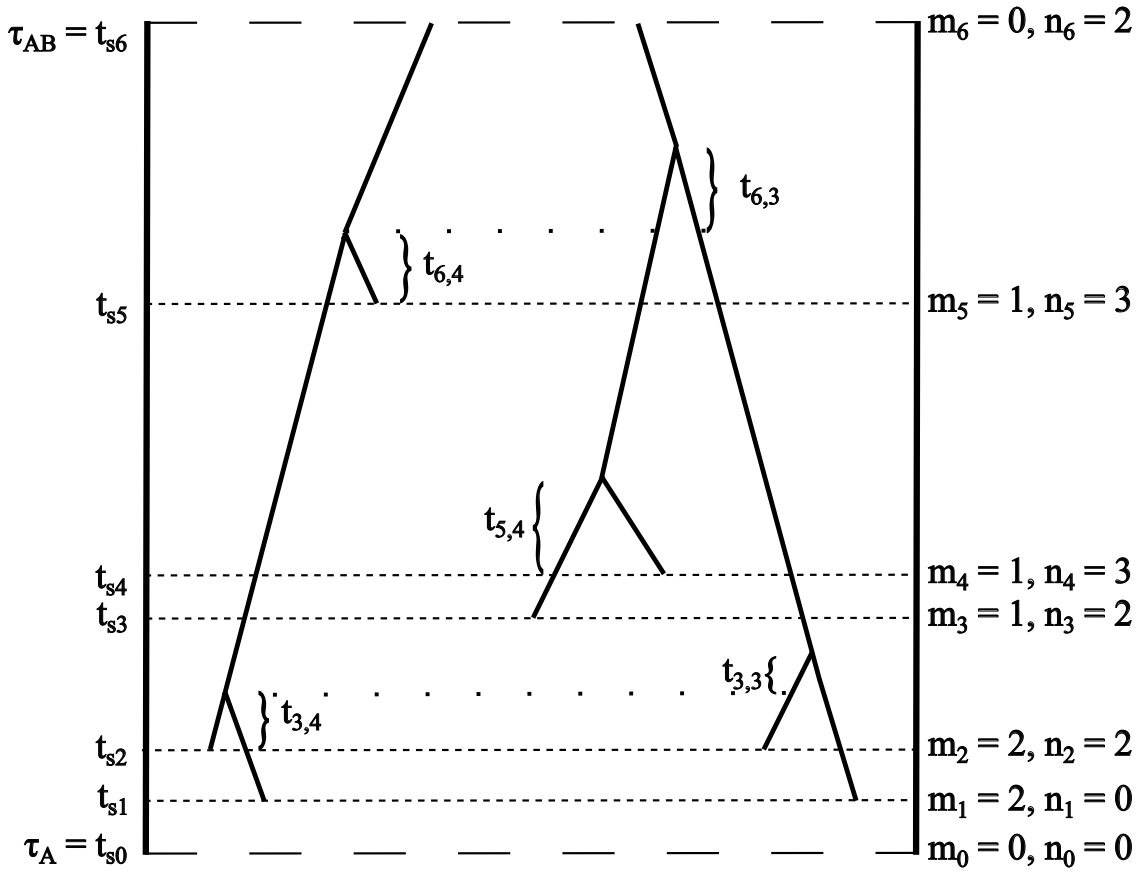


FIGURE 3.1. Part of a gene tree in population A. Going backward in time, population A begins at time $\tau_A = t_{s0}$ and ends at time $\tau_{AB} = t_{s6}$, shown with the large dashed lines. Samples are taken at five distinct times, t_{si}, $i = 1, 2, \ldots, 5$, indicated by the small dashed lines. Time is split into 6 epochs in which no samples are added. Between t_{si} and $t_{s(i-1)}$ (i.e. between sampling events) the number of lineages can only decrease. The number of lineages existing at t_{si} equals the number of lineages sampled (m_i) plus the number surviving to t_{si} (n_i). For example, 3 lineages survive to time t_{s5}, so $n_5 = 3$. One lineage is sampled at time t_{s5}, so $m_5 = 1$. Waiting times until coalescent events are written with two subscripts. First index denotes the epoch and the second the number of lineages before the coalescent event. For example, during epoch (t_{s5}, t_{s6}) the waiting time until the first coalescent event is $t_{6,4}$, where the 6 refers to the sixth epoch and the 4 refers to coalescent event than reduces the number of lineages from 4 to 3. The curly braces show the waiting times to coalescent events. If a coalescent event has already occurred within an epoch, the waiting time to the next coalescent event starts at the time of the last coalescent event, shown by the dotted line.

reduces the number of lineages from k to $k-1$ during the interval $t_{s(i-1)}$ to t_{si} be denoted $t_{i,k}$

(Fig. 3.1). The sample times and coalescent times can be converted to years before time present using the mutation rate. For example, $t_{si}^{\triangle} \times \mu = t_{si}$.

The gene tree density given in Rannala and Yang (2003) is modified to account for the sampling times. We split the time duration for each population into epochs (time intervals) within which no new samples are added. The gene tree probability density, given the species tree, the τs, the θs, and the sampling times, is then given as a product over populations and for each population over the epochs. The probability density of the gene tree for one population is

$$f(\boldsymbol{G}|\theta,\Psi) = \prod_{i=1}^{E+1}\left(\prod_{j=n_i+1}^{n_{i-1}+m_{i-1}}\left[\frac{2}{\theta}\exp\left\{-\frac{j(j-1)}{\theta}t_{i,j}\right\}\right]\right.$$
$$\left.\times\exp\left\{-\frac{n(n-1)}{\theta}\left(t_{si}-\left[t_{s(i-1)}+\sum_{j=n_i+1}^{n_{i-1}+m_{i-1}}t_{i,j}\right]\right)\right\}\right).$$

The root population does not have a time $t_{s(E+1)}$. The density for the root population is given by

$$f(\boldsymbol{G}|\theta,\Psi) = \prod_{i=1}^{E}\left(\prod_{j=n_i+1}^{n_{i-1}+m_{i-1}}\left[\frac{2}{\theta}\exp\left\{-\frac{j(j-1)}{\theta}t_{i,j}\right\}\right]\right.$$
$$\left.\times\exp\left\{-\frac{n(n-1)}{\theta}\left(t_{si}-\left[t_{s(i-1)}+\sum_{j=n_i+1}^{n_{i-1}+m_{i-1}}t_{i,j}\right]\right)\right\}\right)$$
$$\times\prod_{j=2}^{n_E+m_E}\left[\frac{2}{\theta}\exp\left\{-\frac{j(j-1)}{\theta}t_{(E+1),j}\right\}\right].$$

The density of the complete gene tree at every locus is found by multiplying the population densities.

3.2.2. The MCMC Algorithms. We implemented the MSC model with dated tips in the Bayesian inference program BPP. Markov chain Monte Carlo (MCMC) is used to sample from the joint posterior distribution. A fixed species tree topology without migration is assumed. It is also assumed that each sample can be assigned *a priori* to a population. No samples are from ancestral populations. New proposals and modifications to existing proposals in the MCMC are described.

3.2.2.1. *Mutation Rate.* The sample times are specified by the user in units of calendar time before present. These values are assumed to be known and do not change during the MCMC. The times are multiplied by μ to convert them to units of expected number of mutations, as all of the

calculations in BPP are in these units. When proposals are made to μ, all of the sample times (in units of expected mutations) must be updated to preserve the absolute sample times.

$$t^*_{si} = t_{si} \times \mu^*/\mu, \tag{3.1}$$

where the superscript $*$ indicates a proposed value. This ensures the absolute sample times are constant. Since each sample is assigned to a population, the divergence times impose constraints on the possible values of μ. Change in μ must not move the sample between populations. More specifically,

$$t^{\triangle}_{si} \times \mu^* = t^*_{si} < \tau$$

This gives a local upper bound for μ^* as $\min(\tau/t^{\triangle}_{si})$ for all samples in a population. The minimum of this bound over all loci for all populations gives the global upper bound used in the proposal. The lower bound is an arbitrarily small positive number. We propose a new mutation rate, μ^*, on a log scale with sliding window, reflecting at the bounds (Yang, 2014, p. 221-226)

$$\mu^* = \mu \times c = \mu \times e^{\epsilon x}, \tag{3.2}$$

where ϵ is the fine-tune parameter (or step size) and x is a random variable drawn from a Bactrian Laplace distribution (Yang and Rodríguez, 2013). A Bactrian distribution is a mixture of two unimodal distributions, giving a bimodal distribution. This move has a proposal ratio of c (Yang, 2014, p. 225). The tip dates in units of expected mutations undergo a transformation given by the following equation

$$t^*_{si} = t_{si} \times \frac{\mu^*}{\mu} = t^{\triangle}_{si} \times \mu^*$$

Updating the tip ages in units of expected number of mutations without updating the coalescent times can lead to the coalescent times being younger than their daughter nodes, which is not possible. This type of move could be rejected, but that leads to poor mixing. To improve mixing of the MCMC, we jointly update the coalescent times in the populations when updating tip dates. Let b_i be the age (in expected number of mutations) of the oldest sample that is descendant from

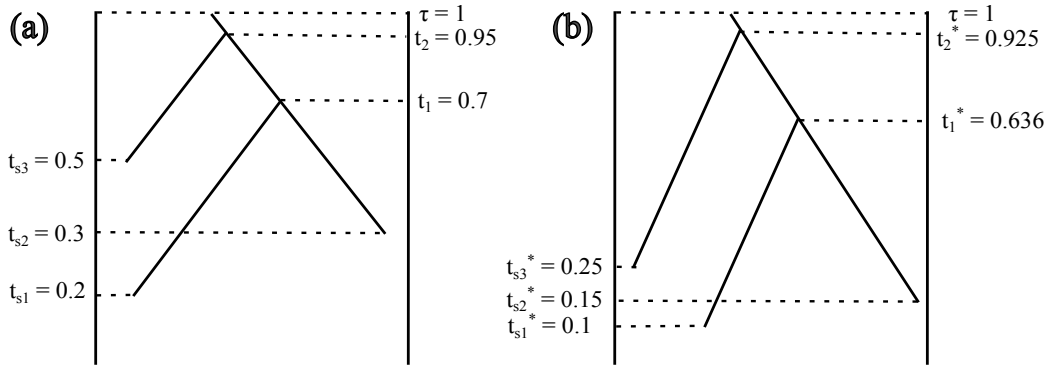


FIGURE 3.2. Part of a gene tree within a population (**a**) before and (**b**) after a new mutation rate is proposed. Here $\mu/\mu^* = 2$ and the tip dates are updated to be: $t_{b1} = \max(t_{s1}, t_{s2}) = 0.3$, $t^*_{b1} = \max(t^*_{s1}, t^*_{s2}) = 0.15$, $t_{b2} = \max(t_{s1}, t_{s2}, t_{s3}) = 0.5$, and $t^*_{b2} = \max(t^*_{s1}, t^*_{s2}, t^*_{s3}) = 0.25$. Not drawn to scale.

a node i in the gene tree. We keep the age of t_i relative to b_i and τ constant (Fig. 3.2).

$$\frac{\tau - t_i^*}{\tau - b_i^*} = \frac{\tau - t_i}{\tau - b_i}$$

Let $h_i = (\tau - t_i)/(\tau - b_i)$ and rearranging the equation,

$$t_i^* = \tau - h_i \times (\tau - b_i^*) \tag{3.3}$$

To derive the proposal ratio,

$$J(h) = \det \frac{\partial(t_1, t_2, ...t_n)}{\partial(h_1, h_2, ..., h_n)} = \prod_{i=1}^{n} (\tau - b_i^*)$$

Proposing the change to μ on a log scale has a proposal ratio of c. The proposal ratio for the move is thus

$$c \times \frac{J(h^*)}{J(h)} = c \times \prod_{i=1}^{n} \frac{\tau - b_i^*}{\tau - b_i}$$

It is possible for this move to propose times such that a daughter node is older than a parent node in the gene tree. In this case, the move is rejected.

For example, consider the gene tree embedded in the species tree of Fig. 3.2. The sample time or coalescent time, in expected number of mutations, is labeled for each node. A new value of μ is proposed using Equation 3.2. The sample times (t_{s1}, t_{s2}, t_{s3}) are updated using Equation 3.1. Then the coalescent times (t_1, t_2) are updated using equation 3.3, resulting in the gene tree in Fig. 3.2b.

3.2.2.2. *Divergence times.* The speciation times, τ, are proposed so that the sample times bound the possible node ages. The age of a node is constrained above by the age of the parent node, τ_u, and below by the oldest daughter node τ_l. Samples cannot change populations, imposing an additional constraint on speciation times. For a given population, t_{sE} is the oldest sample across all loci. Since the samples only occur in tip populations of the species tree, there $\tau_l = 0 \leq t_{sE}$. The speciation time for the parent population is thus bounded below by t_{sE}. As in the previous implementation, a proposed move that is outside of the bounds is reflected to be within bounds.

3.2.2.3. *Gene tree SPR.* The subtree-pruning-and-regrafting (SPR) proposal applied to gene trees (Rannala and Yang, 2017) is modified to allow for dated samples. In the implementation without sample dates, a node in the gene tree is selected to be pruned. The branch between the node and the parent node is removed. This could remove either a single node or several nodes in a subtree. To choose a time to reattach the subtree, a bound on the youngest possible reattachment time is found. If the population in which the node exists has nodes that are not part of the subtree, the bound is equal to the node age of the pruned node. If the population does not have nodes which are not part of the subtree, the bound is the speciation time for the youngest ancestral population which has gene tree nodes that are not part of the subtree (Fig. 3.3). The upper bound is an arbitrarily large number. A reattachment time is proposed and reflected at the bounds.

With dated tips, it is possible that a population will have gene-tree nodes that are not part of the subtree, but are older than the proposed time. This may occur when the pruned node is younger than all samples that are not part of the subtree (Fig. 3.3). In this case, the move is rejected.

As an example, consider the gene tree and species tree in Fig. 3.3a. If the node sampled at time t_{s1} pruned, the lower bound on reattachment is t_{s1}. It is possible to propose a time between t_{s1} and t_{s2}. In this case the move is rejected as there are no branches on which to attach in this time interval. If the node sampled at time t_{s2} is pruned, the lower bound is t_{t2}, and there will always be at least one branch (leading to the node at t_{s1}) on which to attach. The node in population B could also be pruned. The lower bound for attachment is τ, as there are no other nodes in population B. Similarly, the node at time t_1 could be pruned and have a lower bound for attachment of τ. In Fig. 3.3b, the node at time t_{s1} is pruned, and a time t_1^* is proposed for reattachment. In this case

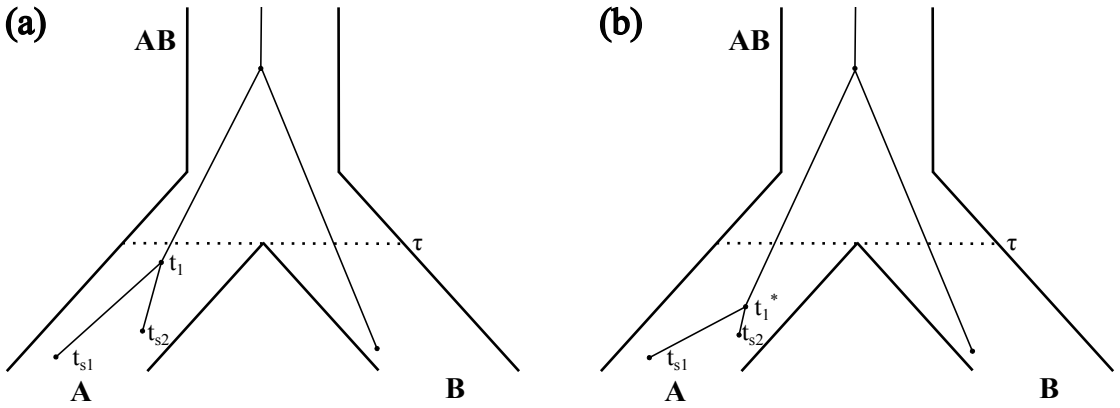


FIGURE 3.3. A gene tree to illustrate the gene-tree SPR move. If the sample at time t_{s1} is pruned, the lower bound on the reattachment time is t_{s1} since there are nodes in population A that are not in the subtree. If the reattachment time it less than t_{s2}, there is no possible reattachment point so the move is rejected. If the sample at time t_{s2} is pruned instead, the lower bound on reattachment is t_{s2}. Since t_{s1} is younger than t_{s2}, it will always be possible to attach the sample at time t_{s2} at the proposed time. If the sample from population B is pruned instead, the lower bound on the reattachment time is τ since there are no other nodes within population B.

the topology of the gene tree did not change. If t_1^* were older than τ, the node could also have been grafted to the branch from the node in population B.

3.2.2.4. *Other Proposals.* The proposals to the gene tree coalescent times and the proposal on θ did not require modifications. The mixing proposal, which jointly changes multiple parameters but does not change the likelihood (Thorne *et al.*, 1998), is turned off in the current implementation.

3.2.3. Simulation Method. The simulation method in BPP was modified to accommodate serial sampling. The dates are specified in units of expected number of mutations and given in an input file. Simulation works similarly the standard MSC simulation with a few extra steps. When simulating the MSC without tip dates, times for the coalescent events are drawn from an exponential distribution with the rate determined by the number of lineages within a population. When a coalescent event occurs, two lineages are randomly chosen to coalesce and the number of lineages decreases by one. This continues until either there is only one sequence in the population or the time drawn is older than the population divergence time. In either case, the time is reset to be the population divergence time, the number of lineages from the two populations are combined and the simulation continues backward in time until the root population only has one lineage. With

tip dates, the simulation starts with the youngest sample time, rather than at time zero. Every time a coalescent time is drawn, it must be checked if the time is older than either the population divergence time or next oldest sampling event. In the former case, the simulation proceeds in the same way as without tip dating. In the later case, the time is set to the next oldest sampling event to determine all of the lineages that the sampling event are added to the lineage count, and the simulation proceeds.

3.2.4. Validation of the Implementation. We have extensively tested our simulation and MCMC implementations. Each MCMC proposal was tested by running under the prior, which is equivalent to setting the likelihood of the data to one. The MCMC results were compared against the analytical results for the prior distributions when these were known. However, the tip dates impose constraints on τs and μ, changing their prior distribution so that the 'effective' priors used by the algorithm differ from the user-specified gamma prior. This is similar to the situation in Bayesian relaxed-clock dating where the effective priors on divergence times differ from user-specified fossil-calibration densities (Rannala, 2016). In our tests we used rejection simulation to determine the effective prior.

An independent simulation program was written to sample from the effective prior for a four-tip symmetric tree and a four-tip asymmetric tree. For both we assume the tree topology is fixed and the tip ages in years before present are known.

For the asymmetric tree, the simulation works as follows. A mutation rate is drawn from the prior distribution. The sample dates in expected number of mutations are calculated. A root age is drawn from the prior. Two node ages are drawn on a uniform distribution between zero and the root age. The times are rank ordered to determine the node ages. If the node ages are younger than the sample dates in a daughter population, the move is rejected. Otherwise, the times are stored. This is repeated until the desired number of samples has been obtained. With an symmetric tree, the simulation works similarly except that the ages for the two (non-root) internal nodes are drawn independently from a uniform distribution between zero and the root age.

3.2.5. Bayesian Simulation to Validate the Implementation of the MCMC Algorithms. Bayesian simulation is a technique to assess the correctness of a Bayesian inference program, in which a set of parameters of the model are drawn from their prior distributions and then used to simulate a replicate dataset. Then, the inference program is used to analyze each dataset using the priors from which the parameters were drawn, to generate the posterior of the parameters. When the posteriors from replicate datasets are combined, the mixture distribution (or average posterior) should match the prior distribution (Flouri *et al.*, 2022).

Bayesian simulation was conducted on a four-tip symmetric tree with five individuals per species. Sample times were drawn from a uniform distribution between 0 and 50,000 years before present. The sample times were the same for all replicate datasets. Each replicate dataset had 100 loci that were 1000 base pairs in length. Sequence data was simulated with the Jukes-Cantor model (Jukes and Cantor, 1969). As noted above, the prior distribution for some of the parameters in the model is not known analytically. Given the fixed set of sample times and species tree, the rejection simulation method was used to draw parameters from the prior distribution of the τs and μ. The θs were drawn using their analytical prior distributions. We simulated 3000 replicate datasets. The root age was assigned the prior $\Gamma(10, 100)$, the mutation rate had $\mu \sim \Gamma(10, 10^8)$, and $\theta \sim \Gamma(8, 2000)$. The mean of the distribution $\Gamma(\alpha, \beta)$ is α/β with variance α/β^2.

3.2.5.1. *MCMC.* Bayesian simulations were conducted with 3000 replicate datasets. The parameters are described in the main text. Each MCMC was sampled 400,000 times, sampling every 4 iterations with 80,000 iterations of burn-in. Two MCMCs were run for each dataset to check convergence. Convergence was checked by comparing posterior samples from the two MCMCs for each set of parameters. A two-sample t-test was used to compare the posterior means in the two chains.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{\frac{2}{n}}}$$

where

$$s_p = \sqrt{\frac{s_{X_1}^2 + s_{X_2}^2}{2}}$$

In the standard two-sample t-test, X_i are the sample means, $s^2_{X_i}$ are the unbiased estimators of the variance and n is the sample size. There are $2n-2$ degrees of freedom. Since the samples were not independent, rather than using the total number of samples in the MCMC, $n = 10,000$ was used as the sample size. The test was performed on estimates of all of the θs and τs. If there was a significant difference between the samples for any variable, the run was considered to not have converged. Additionally, any pairs of MCMCs that had effective sample sizes lower than 200 for any the θs and τs were considered to not have converged. This resulted in 408 datasets with MCMCs that did not converge. All runs that did not converge were re-run with different seeds and a burnin of 200,000 iterations. Convergence was checked again using the same criteria. There were 213 datasets that did not converge on this second analysis and these were excluded from the results (e.g. the plot summaries in Fig. 3.6 & 3.7).

Assessing convergence of MCMC is non-trivial, and these methods of checking convergence were spot checked for MCMCs that did or did not converge. The trace plot, the effective sample sizes, and plots of kernel density estimation were further visually examined for these spot checked cases.

3.2.6. Inference with Extinct Species.

3.2.6.1. *Simulations: Nuclear DNA.* To investigate the performance of the method with extinct species, sequence data were simulated for a four-species symmetric tree, with either one or two extinct species (Fig. 3.4a). We used $\theta = 0.001$ or 0.0001 for all populations, which may be representative of great apes (Kaessmann *et al.*, 2001). For each extant population 3 diploid individuals were sampled, with two phased sequences per locus. For each extinct population either three or six diploid individuals were sampled, with two phased sequences per locus. Datasets had 10, 100, 500, or 2000 loci of 1000 sites each. Sequence data were simulated with a Jukes-Cantor model; for closely related species that experience few multiple substitutions a more complex model is unnecessary. The mutation rate μ was assumed constant across loci with rate 10^{-9} mutations per year. For each of the extinct populations, the sample date for each individual was drawn from $U(0,1)$. The extinct populations were assumed to have become extinct 5,000 years before present. The date for each individual was rescaled to be between 5,000 and 10,000 or 5,000 and 50,000 years before present. The number of samples for each extinct species, the number of extinct species, number of loci, value of θ, and age of the samples were examined factorially. For each set of conditions, 20

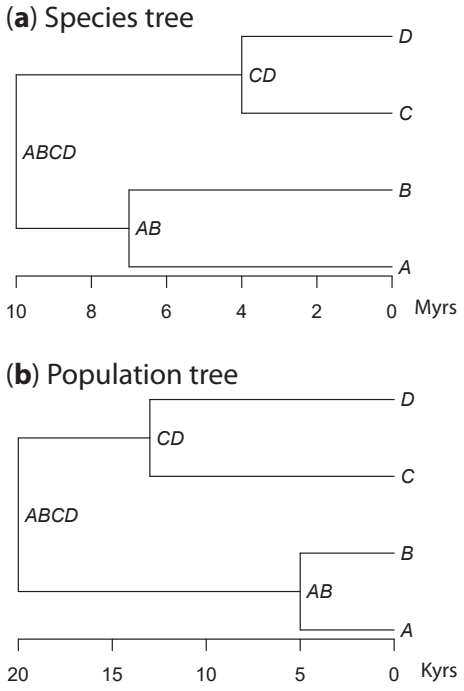


FIGURE 3.4. (**a**) The tree used to simulate data with either species A or both A and C extinct. In both cases, the root was 10 million years old, node AB was 7 million years and node CD was 4 millions years old. The extinction occurred at 5000 years before present. (**b**) The tree used to simulate recent population divergences. The root age is 20 kyr. The age of node AB is 5 kyr and the age of node CD is 13 kyr.

replicate datasets were simulated. For one replicate, the uniform draws to determine the sampling dates were the same for all of the loci and date ranges. This may mimic the scenario of sampling the same individuals and collecting more loci from them. Relative to the 3-individual datasets, three individuals with sampling dates were added in the 6-individual datasets. Note that sequence data and coalescent times were simulated independently for each dataset and differ among datasets.

A new version of BPP implementing models for dated samples, branched from the BPP version 4.6, was used for inference on the simulated datasets. The root age prior was $\Gamma(10, 1000)$. The mutation rate prior was $\mu \sim \Gamma(10, 10^{-10})$. The θ prior was $\Gamma(2, 2 \times 10^4)$ and $\Gamma(2, 2 \times 10^5)$ for θ equal to 0.001 and 0.0001, respectively. The priors were chosen to have the means centered at the true parameter values.

3.2.6.2. *Simulations: Mitochondrial DNA.* Using the same tree as the nuclear DNA simulations (Fig. 3.4a), data were simulated with parameters similar to mitochondrial DNA. Specifically, each

individual has a single locus that was 16,000 base pairs in length (Boore, 1999) with $\mu = 10^{-8}$ mutations per year. 10 individuals were sampled for each extant population. 10, 20, or 100 individuals were sampled for each extinct population. θ was either 0.0025 or 0.00025 for all populations θ and the number of individuals sampled in the extinct populations were varied factorially. As in the nuclear datasets, the dates from the 10 individual datasets matched 10 of the individuals in the 20 individual datasets, and the dates from the 20 individual datasets matched 20 of the dates in the 100 individual datasets.

The mutation rate was assigned the prior $\mu \sim \Gamma(10, 10^{-9})$. The prior for population sizes was $\theta \sim \Gamma(2.5, 10^3)$ and $\Gamma(2.5, 10^4)$ for the larger and smaller values of θ, respectively. The age of the species-tree root had the prior $\tau \sim \Gamma(4, 400)$. Other priors remained the same as in the previous analyses.

3.2.7. Inference of Recent Population Divergences. To investigate the ability of the method to estimate recent divergence times, data were simulated using a four tip tree with a root age of 20 kyr (Fig. 3.4b). Three individuals were sampled per population, each with two phased sequences per locus. Sample ages were drawn between 0 and the divergence time for each population. Datasets were simulated with either 10, 100, 500, or 2000 loci. θ was either 0.001 or 0.0001. The number of replicate datasets simulated for each number of loci was 20. The sample dates were redrawn for each of the 20 replicate datasets.

The root age was assigned the prior $\tau \sim \Gamma(20, 10^6)$. The mutation rate was assigned the prior $\mu \sim \Gamma(10, 10^{-10})$. The prior for θ was $\Gamma(10, 10^4)$ and $\Gamma(10, 10^5)$, for the high and low values of θ, respectively. As before, the prior means match the true parameter values. Note that the root age and μ have to be compatible with the fixed sample dates and their effective priors after the truncation differ from the specified gamma distributions.

3.2.8. Treating Ancient Samples as Contemporary. To examine the effects of ignoring sample dates, the simulated datasets were reanalyzed with all of the sample dates set to zero. The BPP program with tip dating options implemented was also used for these analyses and all priors, including the mutation rate prior, remained the same.

3.2.9. MCMCs.

3.2.9.1. *MCMC settings.* The all MCMCs run were sampled 400,000 of times, sampling every 4 of iterations. The burnin was 160,000 iterations. Two independent MCMCs were run for each dataset.

3.2.9.2. *Convergence.* Convergence was checked by comparing the results between the independent MCMCs. Convergence was checked using criteria similar to the Bayesian simulations, except that an n of 2000 was used in the two-sample t-test and differences in the means of all parameters (θs, τs, $\tau^{\triangle}$s, and μ) were required to not be significantly different between the replicate MCMCs. All parameters except μ were required to have an effective sample size of at least 200 in both MCMC replicates to be considered as converged. Runs that did not converge were re-run with different seeds and 600,000 samples, sampled every 4th iteration. The burnin length was not changed. The same test was conducted after re-running the MCMCs, except that the ancestral population sizes and the root age in expected number of mutations were not checked (these parameters converged more slowly than other parameters, and were not central to the results) and a two-sample t-test sample size of $n = 200$ was used. The root age in time before present was included in the convergence criteria and appeared to converge more quickly than root age in expected number of mutations in some cases. The mitochondrial simulations and recent population divergence simulations that did not meet the convergence criteria were removed from the results. These comprised no more than half of any set of 20 replicate simulations. The other MCMCs that did not meet the convergence criteria were re-run with different seeds and 1,200,000 samples, sampled every 4th iteration. These tended to be the larger datasets with 500 or 2000 loci. The convergence was assessed again using the same test as was used for the first MCMC re-runs (ancestral population sizes and the root age in expected number of mutations were not checked and a two-sample t-test sample size of $n = 200$ was used). The simulations that did not meet the convergence criteria were removed from the results. These simulations comprised no more than half of simulation replicates for any set of simulation parameters.

3.2.10. Empirical Analysis of Mammoths and Elephants.

3.2.10.1. *Mitochondrial Dataset.* The mitochondrial alignment from van der Valk *et al.* (2021) was downloaded (see supplement). The van der Valk dataset includes forest (*Loxodonta cyclotis*),

Accession No.	Species	Age
KY616982.1	*Loxodonta africana*	modern
KY616977.1	*Loxodonta africana*	modern
KY616974.1	*Loxodonta africana*	modern
AB443879.1	*Loxodonta africana*	modern
MT636097.1	*Loxodonta cyclotis*	1533 (417 ybp)
MT636095.1	*Loxodonta cyclotis*	1533 (417 ybp)
MT636093.1	*Loxodonta cyclotis*	1533 (417 ybp)
KY616981.1	*Loxodonta cyclotis*	modern
KY616980.1	*Loxodonta cyclotis*	modern
KY616975.1	*Loxodonta cyclotis*	modern
KJ557423.1	*Loxodonta cyclotis*	modern
NC_020759.1	*Loxodonta cyclotis*	modern
DQ316068.1	*Elephas maximus*	modern
OP575307.1	*Elephas maximus*	modern
OL628830.1	*Elephas maximus*	modern

FIGURE 3.5. Additional samples used in the mitochondrial analysis downloaded from GenBank.

savanna (*Loxodonta africana*), and Asian (*Elephas maximus*) elephants, woolly mammoths (*Mammuthus primigenius*), Columbian mammoths (*Mammuthus columbi*), and mammoths not identified to the species level. Sequences of unknown age or from unknown species were removed from the dataset. Sequences of Columbian mammoths were also removed, as researchers have suggested a potential hybrid origin (van der Valk *et al.*, 2021). This resulted in 10 elephant sequences and 69 woolly mammoth sequences. The calibrated sample dates published in the original papers were used.

Additional sequences were downloaded from GenBank, including four savanna elephants, eight forest elephants, and three Asian elephants (Fig. 3.5). The sequences were realigned with MUSCLE (v3.8.425) using the default settings (Edgar, 2004). Sites in the alignment with more than 25% missing data were removed. These were almost entirely sites at the beginning or end of the alignment. Three sequences from forest elephants were recovered from a ship that sank. The shipwreck year was used as the sample ages for these specimens (Fig. 3.5). All other extant species sequences were assigned sample ages of zero.

A HKY+$\Gamma(4)$ substitution model was used (Hasegawa *et al.*, 1985; Yang, 1994) to account for the extreme transition/transversion rate bias due to DNA degradation. The prior for θ was

$\Gamma(2, 200)$. The prior for τ was $\Gamma(45, 1000)$. The prior for μ was $\Gamma(10, 10^9)$. The reasoning for the prior choices is described below

3.2.10.2. *Priors for Mitochondrial Dataset.* The θ prior was determined by calculating the average pairwise divergence between all contemporary samples within a species across all possible pairs. Gaps were removed prior to calculating pairwise divergence. A relatively broad prior was chosen to reflect the large difference in average pairwise divergence in the different species.

Species	pairwise divergence
Asian	0.0034
Forest	0.013
Savannah	0.026

To find a prior for the root τ, the average pairwise divergence was found between all pairs of Asian and Forest elephants sequences and Asian and Savannah elephant sequences. The prior was chosen to have a mean close to the average pairwise divergence, with a relatively large variance to reflect the prior uncertainty in the parameter value.

Species 1	Species 2	pairwise divergence
Asian	Forest	0.047
Asian	Savannah	0.048

3.2.10.3. *Nuclear Dataset.* The dataset from Rohland *et al.* (2010) was reanalyzed using BPP. The dataset has three extant species: Asian, forest, and savanna elephants; and two extinct species: woolly mammoths and American mastodons (*Mammut americanum*). There are 347 loci, averaging 106 base pairs in length. One individual was sampled per species. The mastodon data is phased, but has one sequence for each individual at each locus, and all other sequences are unphased. The woolly mammoth samples are dated to approximately 43,000 years before present and the mastodon sample to between 50,000 and 130,000 years before present (Rohland *et al.*, 2007; Römpler *et al.*, 2006).

Analyses were conducted using either 50,000, 90,000 or 130,000 years before present as the sample date for the mastodon. The analysis was also repeated without the mastodon sample, both due to the uncertain age and concerns about DNA degradation, as described in original analysis of this dataset (Rohland *et al.*, 2010). The JC model substitution model was used. The prior for τ

was $\Gamma(35, 1000)$ and $\Gamma(7, 1000)$ with and without the mastodon sample, respectively. The prior for θ was $\Gamma(2, 2000)$ and the prior for μ was $\Gamma(5, 10^{10})$. The reasoning for the prior choices is described below.

3.2.10.4. *Priors for Nuclear Dataset.* To choose appropriate parameters for the root age prior, all of the loci were concatenated for each species. The average pairwise divergence between sequences from the mammoth and elephant species and the mastodon was calculated to specify a prior for the dataset with the mastodon. The average pairwise divergence between all pairs of species that are not sisters was calculated to choose a prior for the dataset without the mastodon. Gaps were removed from the two sequences being compared prior to calculating pairwise divergence and "n" was treated as a gap. When ambiguity codes existed in the sequences, equal probability was given to all possible bases indicated in the ambiguity code. This method will give an overestimate of root age, as the coalescent times must be older than the speciation time. However, this should give a reasonable order of magnitude for the prior mean. The variance was chosen such that there was a broad distribution around the mean, since there is not strong prior information about the speciation times in expected number of mutations.

Species 1	Species 2	pairwise divergence
Asian	Forest	0.0072
Asian	Savannah	0.0070
Mammoth	Forest	0.0069
Mammoth	Savannah	0.0068
Asian	Mastodon	0.037
Forest	Mastodon	0.036
Mammoth	Mastodon	0.036
Savannah	Mastodon	0.036

To obtain a prior for θ, the pairwise divergence between within a population was calculated for all populations with unphased data. Sites with ambiguity codes were considered to be heterozygous in the individual and not due to sequencing error. As before, concatenated sequences were used and all gaps were removed prior to calculating the pairwise divergence. The μ prior was chosen

to have a mean of 5×10^{-9} based on the priors and justifications used in previous analyses of this dataset (Rohland *et al.*, 2010).

Species	pairwise divergence
Asian	0.0012
Forest	0.0024
Mammoth	0.0008
Savannah	0.0006

3.2.10.5. *Convergence.* Convergence was assessed in tracer by comparing the distributions of all parameters in the pairs of replicate MCMCs and examining the trace plot.

3.2.11. Code Availability. The BPP version used in this manuscript is available at `https://github.com/bpp/bpp/tree/devAnna`. Simulation scripts and BPP control files are available at `https://github.com/nage0178/tipDating_analysis`.

3.3. Results

The correctness of the implementation was tested with Bayesian simulations. The statistical performance of the method was tested using two population histories, a history of ancient species divergence and a recent population divergence, each with four populations. On a four population tree, the method estimates the three divergence times in units of years ($\tau^{\triangle}$) and expected number of mutations (τ), the seven effective population sizes (θ), and the mutation rate (μ). Simulated nuclear datasets were used for both histories and simulated mitochondrial datasets were used for the species divergence. The effect of treating the aDNA sequences as contemporary was investigated for all datasets. Two elephant and mammoth datasets were analyzed with the new method.

3.3.1. Bayesian Simulations. The data generated for the Bayesian simulations were very informative about the speciation times and the mutation rate (Fig 3.6). There was also information about the population sizes in the tip populations. However, there was very little information about the ancestral population sizes, as the posterior distributions very closely resembled the prior distributions. The combined posterior distributions of the MCMCs closely matched the prior distributions for all parameters (Fig 3.7). This suggests the program is correctly implemented. For parameters for which the data are more informative, such as the τs (as seen by a low variance in

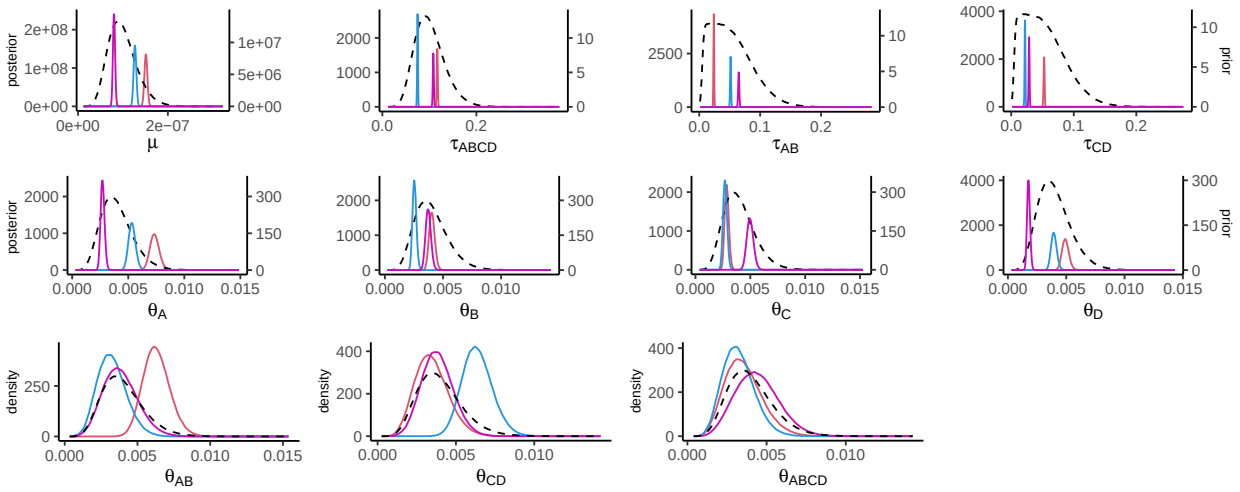


FIGURE 3.6. Posterior distributions for individual replicates in the Bayesian simulation. Each solid colored line shows the posterior distributions of representative replicates. The dotted lines show the prior distributions. The y-axis scale is different for the prior and posterior distributions for the first two rows.

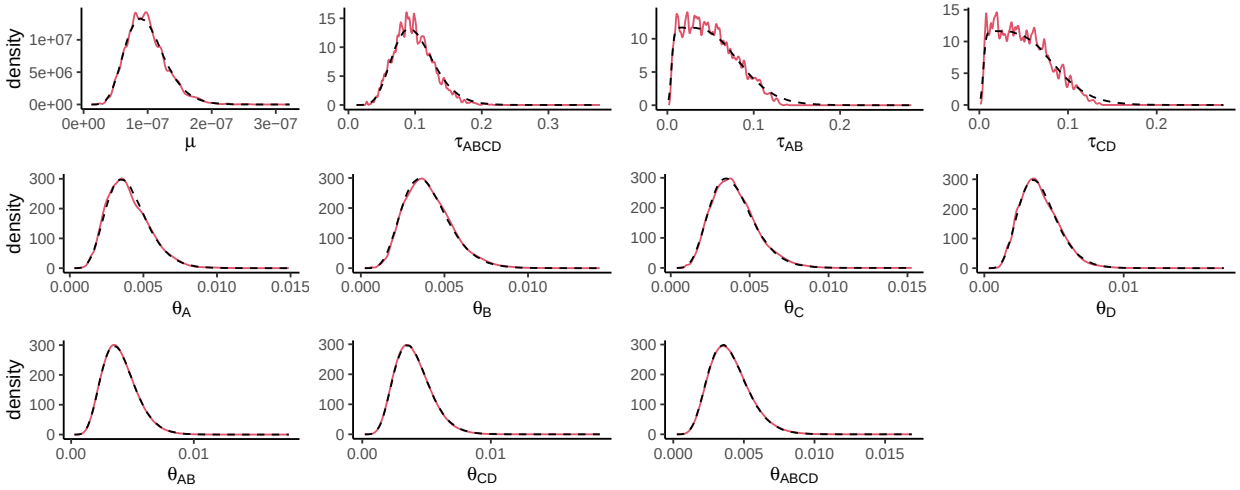


FIGURE 3.7. Bayesian simulation priors (black dotted line) and average posterior (solid red line) distributions for each parameter in the model.

the posterior distributions for individuals replicates), the combined distributions are less smooth as expected.

3.3.2. Simulations: Species Divergence.

3.3.2.1. *Inference under the Correct Model.* Here we examine the effects of the number of loci and the number of sequences (sampled individuals) on estimation of mutation rate (μ) and divergence times (τs), obtained from simulated nuclear and mitochondrial sequences. As the number of loci increased with nuclear sequences and the number of samples increased with mitochondrial sequences, estimates of $\tau^{\triangle}$ improved (Fig. 3.8a, 3.9a, &3.10b). This improvement is a result of better estimates of both μ and τ with more loci (Fig. 3.8b,c & 3.9b,c). Going from 500 to 2000 loci, the average size of the 95% HPD interval decreases much more for μ than τ. The 95% credible intervals were much smaller for the nuclear analysis with many loci than for the mitochondrial analysis with many individuals. The coverages (frequency at which the true parameter value was contained in the 95% credible set) for all datasets with 2000 loci were 97.9% for all divergence times ($\tau^{\triangle}_{ABCD}$, $\tau^{\triangle}_{AB}$, $\tau^{\triangle}_{CD}$) and 97.6% μ, respectively. The coverages for all mitochondrial analyses were 97.8%, 97.8%, 97.6%, and 97.6% for $\tau^{\triangle}_{ABCD}$, $\tau^{\triangle}_{AB}$, $\tau^{\triangle}_{CD}$, and μ, respectively.

The precision and accuracy of estimates of μ in the most informative case (2000 loci) were most impacted by the age range of the samples, with older dates giving more precise estimates (Figs. 3.11, 3.12). Increasing the number of samples for each extinct species and the number of extinct species also improved estimates of μ but to a lesser degree, with the former (number of samples) having the greatest impact. The trends for the estimates of μ are similar with the mitochondrial datasets (Fig. 3.12). Using a smaller true value of θ in the simulations for all populations improved estimates of μ and τ (Figs. 3.8& 3.9).

3.3.2.2. *Biases when Ancient Samples were Treated as Contemporary.* Here we examine the potential negative impacts on estimates of μ, τ and θ if ancient samples are treated as contemporary (e.g., with sample dates set to zero) when analyzing the simulated nuclear sequences. Both μ and $\tau^{\triangle}$ were poorly estimated when ancient samples were treated as contemporary (Fig. 3.8d, 3.10b) with increased widths of credibility intervals and estimates of θ for extinct species that were biased to be too large (Fig. 3.13, 3.14). Without tip ages, the posterior distribution of μ is the same as the prior distribution because μ and τ are not identifiable in this case – only their product can be estimated.

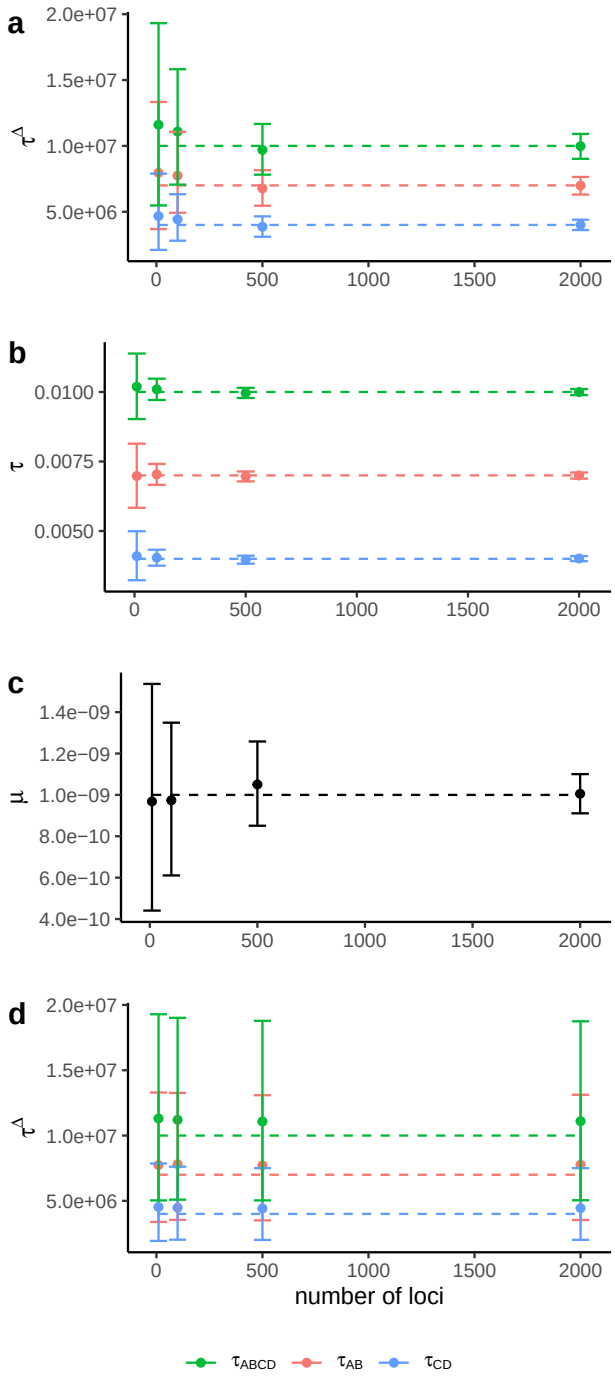


FIGURE 3.8. Average posterior means and 95% HPD CIs (bars), over 20 replicate datasets, of (**a**) divergence times in mutations, (**b**) divergence times in years, (**c**) mutation rate, and (**d**) divergence times in years when the samples are treated as contemporary. The data were simulated under the model of figure 3.4a with two extinct species (A and C), sample dates are between 5,000 and 50,000 years, and $\theta = 0.0001$. The dashed lines show the true parameter values.

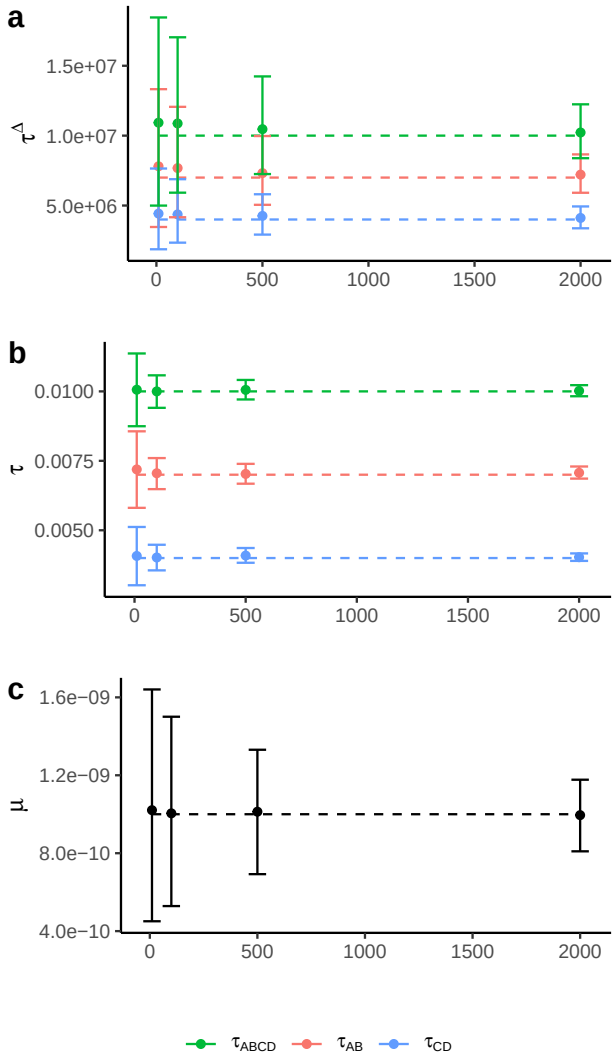


FIGURE 3.9. Average posterior means and 95% HPD CIs (bars), over 20 replicate datasets, of (**a**) divergence times in mutations, (**b**) divergence times in years, and (**c**) mutation rate. The data were simulated under the model of figure 1a with two extinct species (A and C), sample dates are between 5,000 and 50,000 years, and $\theta = 0.001$.

3.3.3. Simulations: Population Divergence.

3.3.3.1. *Inference under the Correct Model.* Here we examine the effects on inference of μ, τ and θ of increasing the number of loci when considering populations that have recently diverged. There is much less information in this case and priors have more influence on the posterior, even with 2000 loci (Fig. 3.10 & 3.16). As the number of loci increased, estimates of population divergence time (τ)

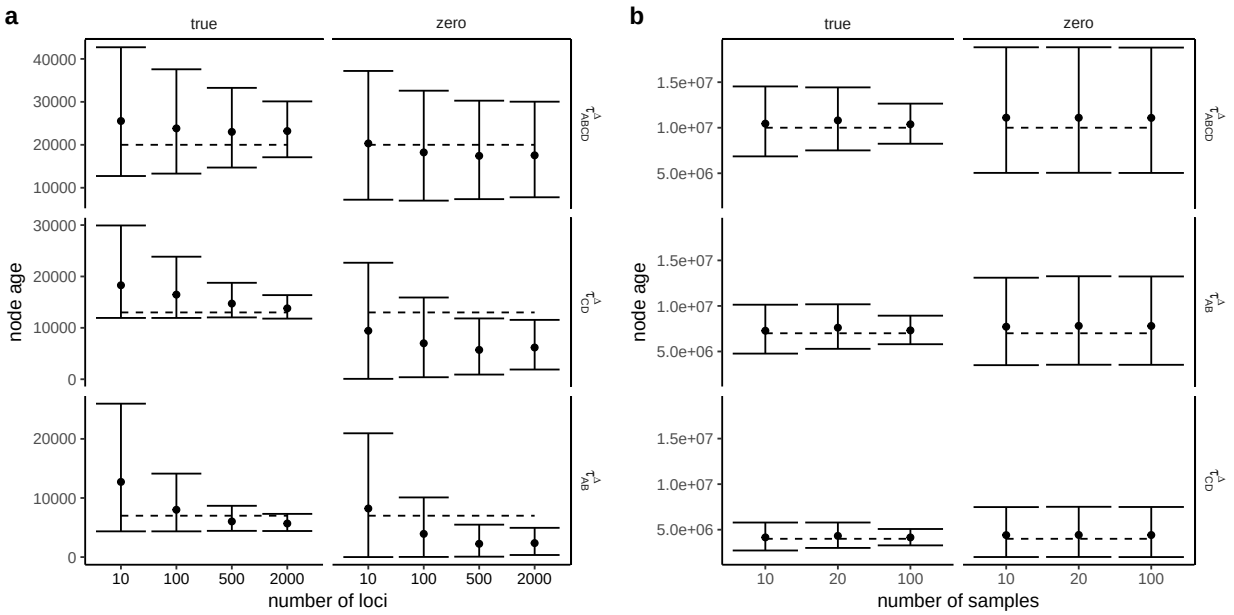


FIGURE 3.10. Average posterior means and 95% CIs of divergence times in years over 20 replicate datasets simulated (**a**) under the model of figure 3.4b with recent population divergences and (**b**) under the species tree of figure 3.4a. In each panel, the left column is for results when the sample dates are accommodated in the method, while the right column is for results when all sample dates are set to zero. In both (a) and (b), there are two extinct species (A and C) with sample dates from U(5000, 50,000) years before present and with $\theta = 0.0001$. The dashed lines show the true parameter values.

improved, with smaller credible sets and less bias (Fig. 3.10a). With less data, estimates of τ were upwardly biased, apparently due to the influence of the prior. With 2000 loci, the coverages for τ_{ABCD}, τ_{AB}, τ_{CD} were all 100% and for μ the coverage was 88.9%. The mutation rate was biased downward with smaller amounts of data, likely due to the interaction of the prior and the sample ages. The bias decreased as the amount of data increased (Fig. 3.15). Of the θ parameters, only the root population size was estimated with increased precision as the amount of data increased (Fig. 3.15). This is likely due to the fact that few lineages are expected to coalesce in contemporary populations due to the young divergence times relative to the effective population size (most will coalesce in the root population), so there is little information about contemporary θs.

3.3.3.2. *Biases when Ancient Samples were Treated as Contemporary.* When the samples were treated as contemporary, population divergence times were underestimated (Fig. 3.10a). This effect

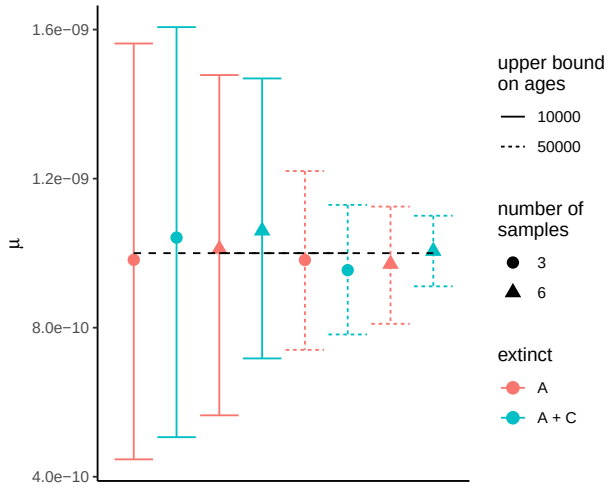


FIGURE 3.11. Average posterior means and 95% HPD CIs (bars) of the mutation rate over 20 replicate datasets, each of 2000 loci, simulated under the model of figure 3.4a with $\theta = 0.0001$. Solid lines are for sample dates between 5,000 to 10,000 ybp while dashed lines are for sample dates between 5,000 and 50,000 ybp. Either species A (red) or both A and C (teal) are extinct, and from each extinct species either 3 (circle) or 6 (triangle) samples are taken. The dashed line shows the true parameter value.

was more pronounced for τ_{AB} and τ_{CD} than τ_{ABCD}; the credible sets for these parameters became smaller and the bias became larger as the number of loci increased.

3.3.4. Analysis of Genomic Data from Elephants and Mammoths.

3.3.4.1. *Mitochondrial Dataset.* The posterior mean divergence time estimate for the two African elephants of 29 KY was extremely recent and the posterior mean divergence time between the Eurasian and African elephants of 1.8 MY was much smaller than previous estimates of 7.6 MY (Fig. 3.17). The mean of the posterior distribution of the mutation rate was higher than the mean of the prior. The mean transition transversion ratio, κ, was 46, which is at least an order of magnitude larger than typical empirical datasets for mammals, likely due to DNA degradation.

3.3.4.2. *Nuclear Dataset.* The estimates of the τs and μ were very similar for all analyses, independent of whether the mastodon sample was included in the analysis and of the sample ages used for the mastodon (Fig. 3.17). The divergence between the African elephants, Asian elephant and mammoth, African and Eurasian elephants, and mastodon was estimated to be 3.0 (0.7-6.4) MY, 2.7 (0.6-5.7) MY, 5.4 (1.6-11.3) MY, and 26.9 (7.8-55.7) MY, respectively, for the dating of the

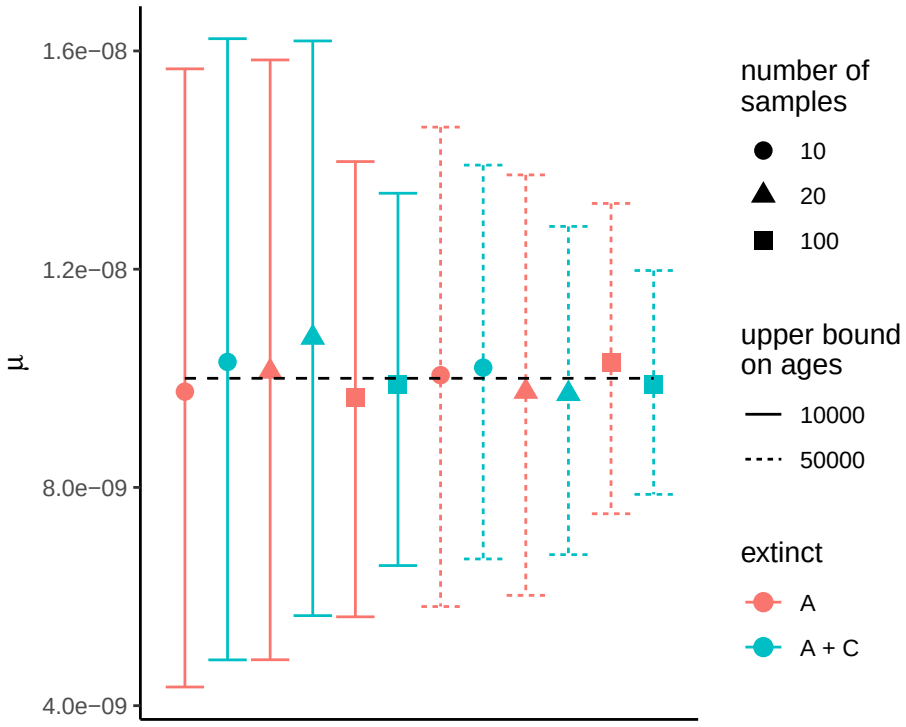


FIGURE 3.12. Average posterior means and 95% HPD CIs (bars) of the mutation rate over 20 replicate datasets, simulated under the model of figure 1a with $\theta = 0.00025$. Solid lines are for sample dates between 5,000 to 10,000 ybp while dashed lines are for sample dates between 5,000 and 50,000 ybp. Either species A (red) or both A and C (teal) are extinct, and from each extinct species either 10 (circle), 20 (triangle), or 100 (square) samples are taken. The dashed line shows the true values of the μ.

mastodon at 90 KY. The credible sets were large for $\tau^{\triangle}$, reflecting the limited information about μ available from these data. The estimates were broadly concordant with results from previous studies when analyzing either the nuclear or mitochondrial DNA, though the point estimates of the divergence times tend to be slightly more recent.

3.4. Discussion

Ancient DNA data provide a new way to study historical populations and their relationships to contemporary populations. However, the processes that generate aDNA data do not fit the model assumptions commonly used in aDNA analyses. Here, a new MSC model with tip dating

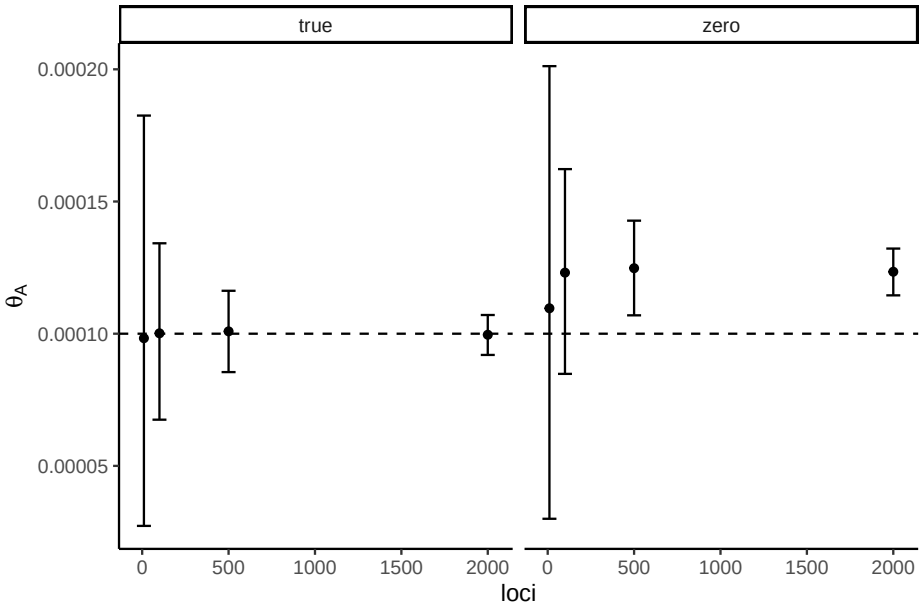


FIGURE 3.13. Average posterior means and 95% HPD CIs (bars), over 40 replicate nuclear datasets, of θ_A when sample dates were set to their true values (left) or zero (right). The datasets had 6 samples in each extinct species and the upper bound on the sample dates equal to 50,000 ybp. The dashed lines show the true values of the θ_A.

was developed to incorporate the sample ages into population genomic data analysis for multiple species and implemented in BPP.

The simulation study demonstrates that the new method can accurately and precisely estimate speciation times in years before present for a variety of data types, including nuclear and mitochondrial sequences, and for population histories with divergence times ranging from several thousand to several million years. In particular, with older samples, more loci, more samples, and more extinct species, the confidence intervals for the divergence times become smaller. The simulation study only explored samples up to 50,000 years old, which is approximately the oldest age current radiocarbon dating technology can date (Hajdas *et al.*, 2021). The ability to use older dated samples in future may improve estimates. While the simulation study only used up to 2000 loci, the trend suggests that more loci could lead to even greater improvements in the estimates.

The ability of the method to infer times in years is based on the sampling of genetic data through time. This provides a means to separately estimate the mutation rate and time and thus to convert branch lengths from expected numbers of mutations to years. Many methods used

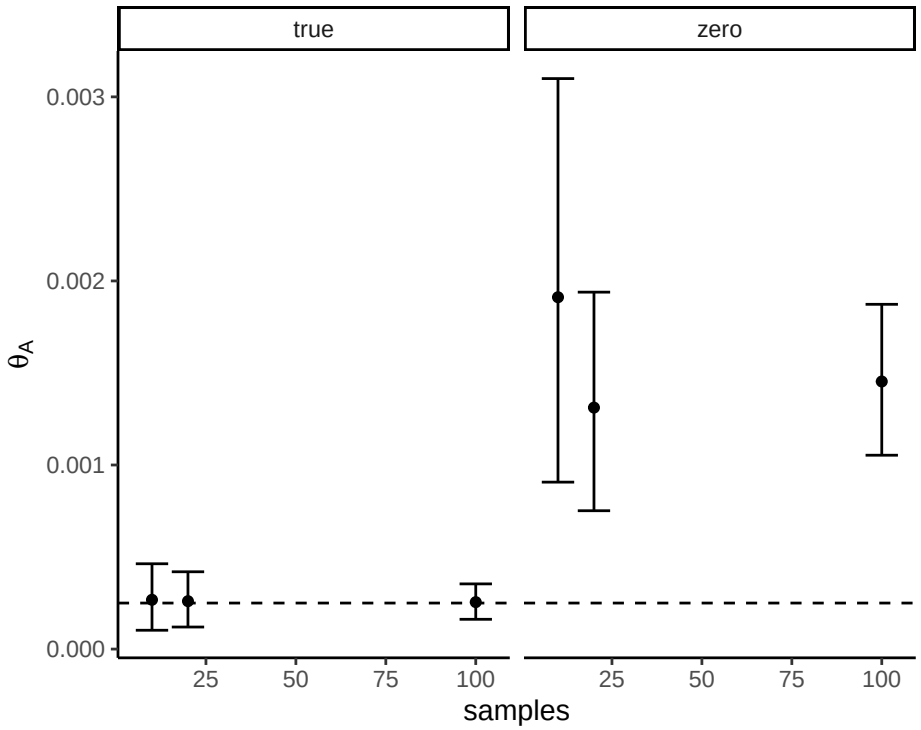


FIGURE 3.14. Average posterior means and 95% HPD CIs (bars), over 40 replicate mitochondrial datasets, of θ_A when sample dates were set to their true values (left) or zero (right). The datasets had the upper bound on the sample dates equal to 50,000 ybp. The dashed lines show the true values of the θ_A.

with aDNA assume a particular mutation rate, which makes the results highly sensitive to that parameter choice. As a Bayesian method, BPP naturally accommodates uncertainty, allowing the prior variance to be chosen to reflect the uncertainty in mutation rate. Our simulation showed that even with a low mutation rate, reliable estimation of the mutation rate and absolute divergence times is possible when a large number of loci is used.

The simulation study also demonstrated the detrimental effects that ignoring sample dates can have on inference. In all population histories explored in this simulation study, mutation scaled population sizes (θ) of populations with aDNA were overestimated and divergence time in years had wide credible intervals when ages were ignored. The large credible intervals for divergence times were driven by the uncertainty in mutation rate. Without sample dates, the posterior distribution of μ is the same as the prior distribution, reflecting the lack of separate information about rate and time. For recent population divergences, we observed that the divergence times were underestimated

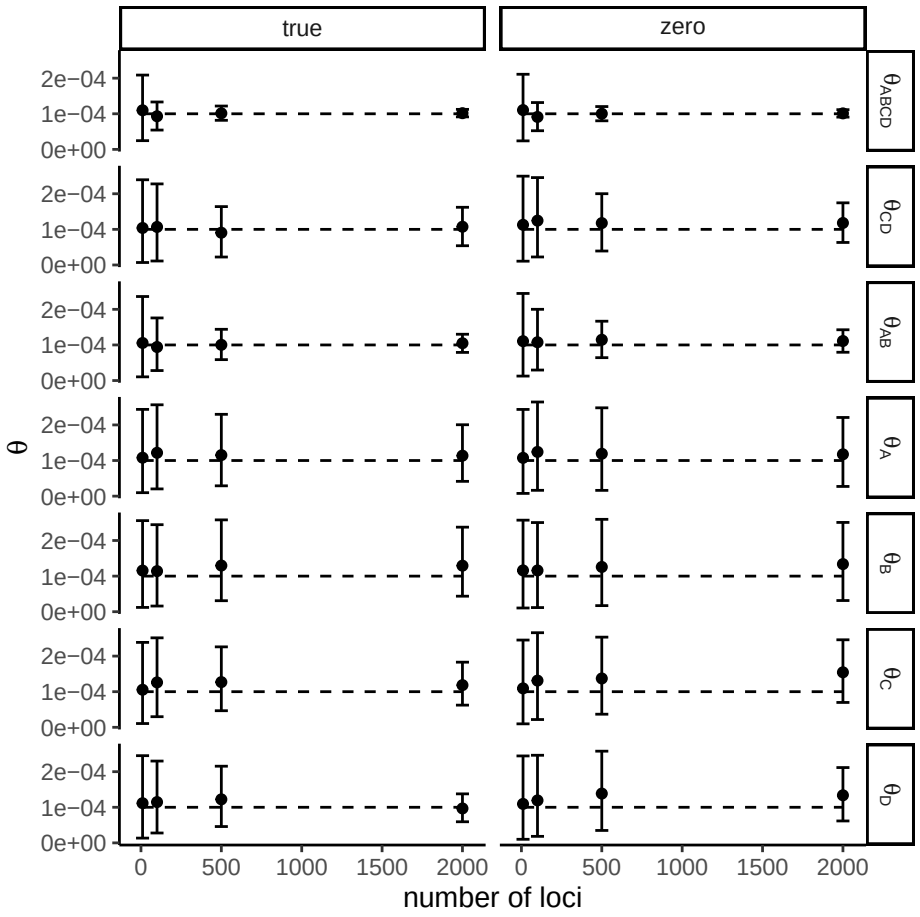


FIGURE 3.15. Average posterior mean and 95% HPD CIs (bars) for θ across 20 replicate simulations for the recent population divergence analysis. The left and right plots show the inferences when the sample ages are set to their true values and zero, respectively. The dashed lines show the true value of θ.

when ancient samples were incorrectly treated as contemporary. This reflects the effects of "missing" mutations between the present time (time zero) and the sample time when using an incorrect model. This effect was not observed for simulations that used extinct species.

The method was developed with the intention of analyzing non-coding sequences or coding sequence which evolve at a similar rate. While coding sequences can evolve at drastically different rates, most of the genome evolves at a similar rate, at least in mammals (Hodgkinson and Eyre-Walker, 2011). The method assumes that the species tree is known, there is no migration between species, and sequence evolution follows a strict clock. The latest version of BPP relaxes these assumptions (Flouri *et al.*, 2018, 2020, 2023), but does not include tip dating. Future work should

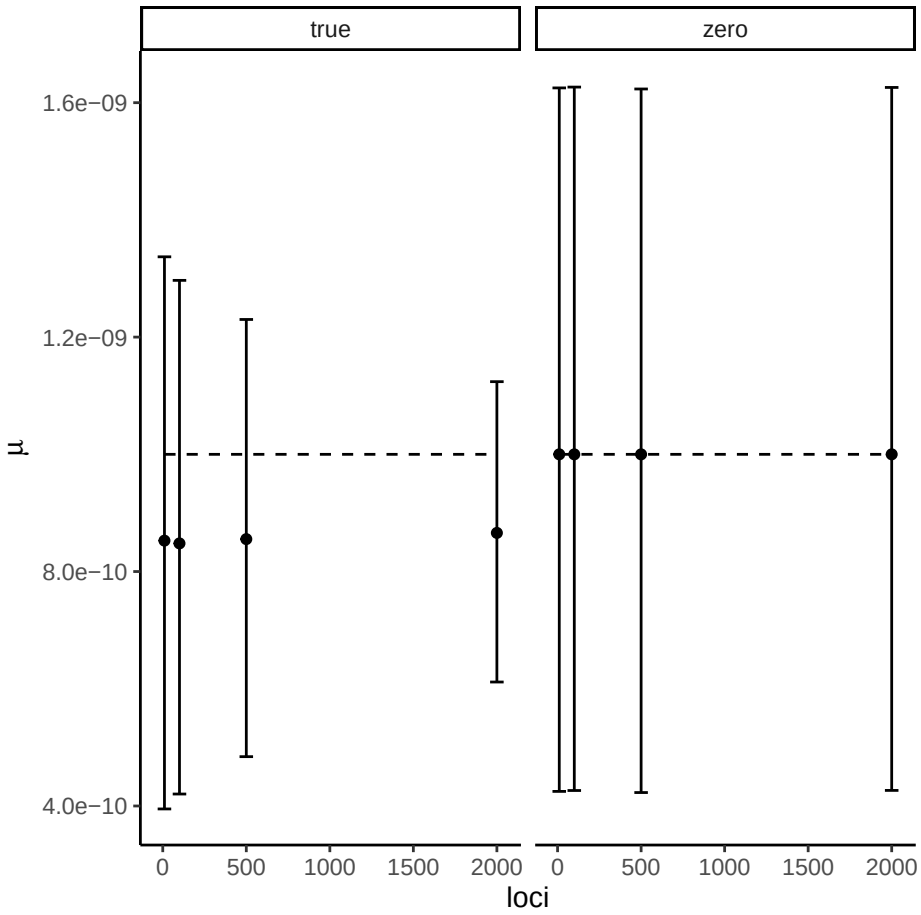
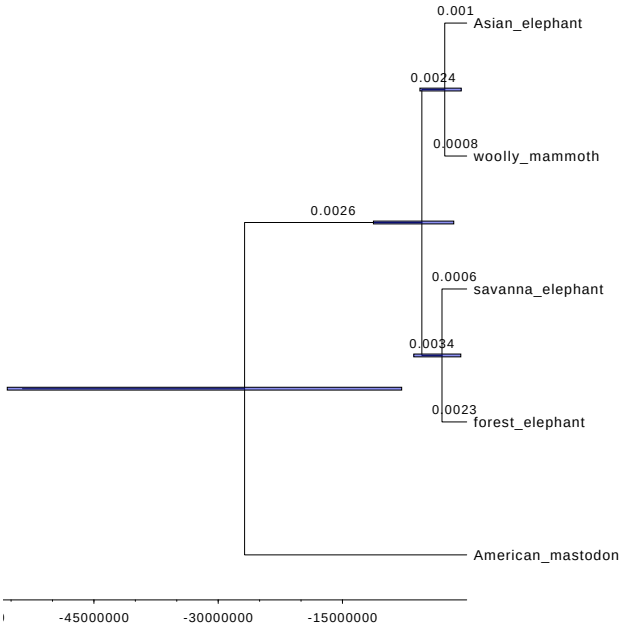


FIGURE 3.16. Each point shows the average of the posterior mean mutation rate and mean 95% credible set averaged across inferences for 20 replicate datasets when the samples ages are set to their true value (left) or zero (right). For all datasets, there were 2000 loci and θ is equal to 0.0001. The dashed line shows the true value of μ.

merge these models into the program with tip dating. BPP also assumes every sample has a known age, in contrast to programs such as BEAST which allows uncertain ages. Adding unknown sample dates for aDNA to BPP would naturally accommodate the use of data without known sample dates, such as the mastodon data used in this study.

An alternative to tip dating when calibrating a molecular phylogeny is to use fossil calibrations. With aDNA, tip dating can be combined with fossils to estimate a time scaled phylogeny, which is currently possible in BEAST. However, placing fossils on the phylogenetic tree is often difficult and error prone; aDNA samples have the advantage that they can provide calibrations and be positioned



Analysis	mastodon age	$\tau^{\triangle}_{Forest,Savannah}$	$\tau^{\triangle}_{Mammoth,Asian}$	$\tau^{\triangle}_{Mammoth,African,Asian}$	$\tau^{\triangle}_{Mastodon,elephant/mammoth}$	$\mu \times 10^{-9}$
mt	NA	29 (5.4 - 50) KY	1.5 (0.7 - 2.2) MY	1.8 (1.4 - 2.4) MY	NA	15 (11 - 18)
nuclear	50 KY	3.0 (0.7 - 6.4) MY	2.7 (0.6 - 5.7) MY	5.4 (1.6 - 11.3) MY	26.9 (7.8 - 55.7) MY	0.51 (0.12 -0.95)
nuclear	90 KY	3.0 (0.7 - 6.4) MY	2.7 (0.7 - 5.6) MY	5.4 (1.6 - 11.3) MY	26.8 (7.9 - 55.5) MY	0.51 (0.12 -0.96)
nuclear	130 KY	3.0 (0.7 - 6.4) MY	2.7 (0.7 - 5.7) MY	5.4 (1.6 - 11.2) MY	26.6 (7.7 - 55.0) MY	0.51 (0.12 -0.96)
nuclear	NA	3.0 (0.7 - 6.5) MY	2.7 (0.7 - 5.6) MY	5.0 (1.4 - 10.5) MY	NA	0.51 (0.12 - 0.96)
mt (Rohland 2007)	NA	NA	6.7 (5.8-7.7) MY	7.6 (6.6-8.8) MY	26 (24-28) MY	4.2 (3.6 -4.9)
nuclear (Rohland 2010)	NA	(2.6-5.6) MY	(2.5-5.4) MY	(4.2 - 9.0) MY	(34-72) MY	

FIGURE 3.17. The tree shows the nuclear estimates with the mastodon age as 90 KY, which match the table. The branches are labeled with the mean effective population sizes. The table shows the analysis from this study with the mean estimate and the 95% HPD credible set. The analysis from Rohland 2007 and 2010 show the 95% and 90% credible intervals, respectively.

on the tree through the use of sequence data alone rather than using sparse morphological characters as with fossils. Since fossils provide additional information, a combined approach may allow for more accurate estimation of divergence times, but only if fossils can be accurately placed. BPP does not currently accommodate fossil calibrations. Incorporating fossil calibrations in BPP is another possible area for future work.

The new method had convergence issues, particularly in analysis of large datasets (e.g., with 2000 loci). Ancestral population sizes often did not converge when the rest of the parameters did converge. In a more limited set of simulations, the root age in expected number of mutations

also had convergence issues. The lack of convergence appears to be related to poor mixing of the chain in the MCMC and may be improved with MCMC proposals that change multiple correlated parameters.

The mitochondrial mammoth and elephant datasets produced younger estimated divergence times, by comparison with previous estimates, when analyzed with our new method. The very young divergence time between African elephants may reflect recent migration (reviewed in Roca, 2019). The other divergence times are also younger than the nuclear analysis and other analyses. The estimate of κ, the transition transversion rate ratio, is extremely large, about an order of magnitude higher than typical values. This is likely a result of DNA degradation, which causes excessive post-mortem C to T changes (or G to A changes on the other strand), resulting in very high transition rates. The elevated κ combined with the relatively high mutation rate estimate suggests the dataset contained degraded sequences which inflated mutation rate estimates and resulted in estimation of young divergence times. aDNA research, including the dataset used from van der Valk *et al.* (2021), typically extensively characterizes evidence for DNA degradation and attempts to remove degraded sequences. However, our results suggest this approach may be insufficient to remove the impact of degradation and highlights the need to systematically assess the impact of DNA degradation in downstream analysis. DNA degradation depends on the environmental conditions of the sample and does not occur in a clock-like fashion (Mitchell *et al.*, 2005). While DNA degradation has been modeled, it has been in the context of studying modern DNA contamination and potentially removing contaminant sequence reads (Al-Asadi *et al.*, 2019; Peyrégne and Peter, 2020; Renaud *et al.*, 2015; Skoglund *et al.*, 2014). Developing models of DNA degradation which can be applied to alignments may improve estimates from inference methods that use aDNA.

The estimates of divergence times with the elephant and mammoth nuclear dataset were broadly consistent with previous estimates using fossil calibrations. The large credible intervals reflect the limited amount of information about μ in the data. The simulation study suggests that more ancient samples and more loci would improve the precision of the estimates of $\tau^{\triangle}$s and μ.

The age of the mastodon sample did not meaningfully impact the results. This may be due to the relatively small number of loci and the short sequence lengths. This suggests that with a

limited amount of data, uncertainty in sample dates have less impact on the results than uncertainty in other model parameters. Moreover, existing analyses using small amounts of data with uncertain sample dates may report reasonable results. However, the simulations show that incorrect sample dates negatively affect inference as the amount of data increases. As analyses of large genomic datasets including aDNA become more commonplace, researchers should use methods which explicitly account for sample dates, even with relatively young aDNA.

3.5. Funding

This work was supported by the National Science Foundation Graduate Research Fellowship Program (grant no. 2036201) to A.A.N., National Institutes of Health Grant (GM123306) to Bruce Rannala., and Biotechnology and Biological Sciences Research Council (BBSRC) grants (BB/T003502/1, BB/X007553/1, BB/R01356X/1) to Ziheng Yang.

APPENDIX A

Combining Posteriors Example

A.1. Sample from a Bivariate Normal PDF

Suppose that we have samples $Y = y_1, \ldots, y_a$ and $X = x_1, \ldots, y_a$ from a bivariate normal density with means $\mu_y = \mu_x = \mu$, variances $\sigma_x^2 = \sigma_y^2 = 1$ and correlation parameter ρ. Our goal will be to generate the posterior density of μ by combining posterior densities for x and y. This is given as an example to compare the posterior density of μ using analytical results and an approximation that can be used in cases where we cannot directly estimate the posterior probability density of μ given X and Y. We will treat the variables Y and X as independent in our inference procedure, though in reality ρ may be non-zero. For simplicity, we use a normal prior density for μ, which is a conjugate prior for the normal density and so the posterior is also normal. Suppose that $Y \sim \mathcal{N}(\mu, 1)$ and $X \sim \mathcal{N}(\mu, 1)$. Let the prior for Y be $f_y(\mu) \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and the prior for X be $f_x(\mu) \sim \mathcal{N}(\mu_2, \sigma_2^2)$. The "preferred prior" for use in generating the posterior based on both X and Y is $f_p(\mu) \sim \mathcal{N}(\mu_p, \sigma_p^2)$. The posteriors are then

$$f(\mu|Y) \sim \mathcal{N}\left(\frac{\frac{\mu_1}{\sigma_1^2} + a\bar{y}}{\frac{1}{\sigma_1^2} + a}, \frac{\sigma_1^2}{1 + a\sigma_1^2}\right)$$

and

$$f(\mu|X) \sim \mathcal{N}\left(\frac{\frac{\mu_2}{\sigma_2^2} + a\bar{x}}{\frac{1}{\sigma_2^2} + a}, \frac{\sigma_2^2}{1 + a\sigma_2^2}\right).$$

The approximation of the posterior of μ, given X and Y, is then

$$f(\mu|X, Y) = \frac{f(\mu|X)f(\mu|Y)}{f_y(\mu)f_x(\mu)} \times f_p(\mu) \times \frac{C_x C_y}{C_{xy}}. \tag{A.1}$$

The true posterior is know in this case when $\rho = 0$. Let $Z = X \cup Y$ and $n = 2a$, then

$$f(\mu|X, Y) \sim \mathcal{N}\left(\frac{\frac{\mu_p}{\sigma_p^2} + n\bar{z}}{\frac{1}{\sigma_p^2} + n}, \frac{\sigma_p^2}{1 + n\sigma_p^2}\right). \tag{A.2}$$

This simple case can be used to test methods for inferring the posterior from combined samples. Rather than doing MCMC, instead simply sample iid random variables from $f(\mu|Y)$, $f(\mu|X)$, $f_y(\mu)$, and $f_x(\mu)$ and use kernel density estimation to infer the density functions for each. Then apply equation A.1 to estimate the posterior. The accuracy of the estimate can be determined by comparison with results from equation A.2. For example, curves could be plotted for the true density versus the approximation. The approximate density will need to be renormalized so that it integrates to 1. The constant, C, to multiply values by to normalize could be estimated as

$$\frac{1}{C} = \int \frac{f(\mu|X)f(\mu|Y)}{f_y(\mu)f_x(\mu)} \times f_p(\mu)d\mu.$$

To illustrate this method of combing posteriors and show that it produces correct results in this simple case, we let $\mu = 0$, $\mu_1 = 1$, $\mu_2 = -1$, and $\mu_p = 0.5$, and we let $\sigma_1^2 = 1$, $\sigma_2^2 = 1.1$, and $\sigma_p^2 = 0.5$. 10 samples from both X and Y were drawn. Then, 100,000 samples were drawn from both of their prior distributions and posterior distributions. In HIVtree, the samples are drawn using MCMC, but here we have analytical results to use in sampling. Then, KDE was used to estimate the prior and posterior distributions for both X and Y (Fig. A.1). The analytical distribution is in good agreement with the distribution found using KDE and equation A.1 (Fig. A.2).

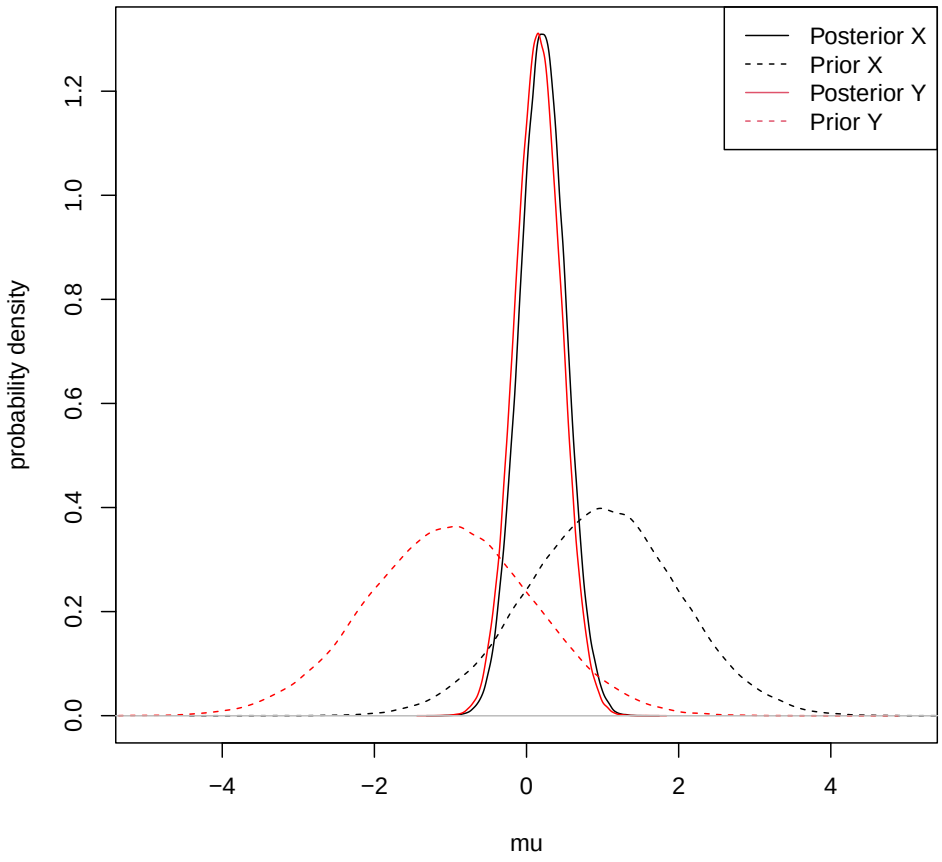


FIGURE A.1. The prior and posterior distribution for X and Y estimated with KDE.

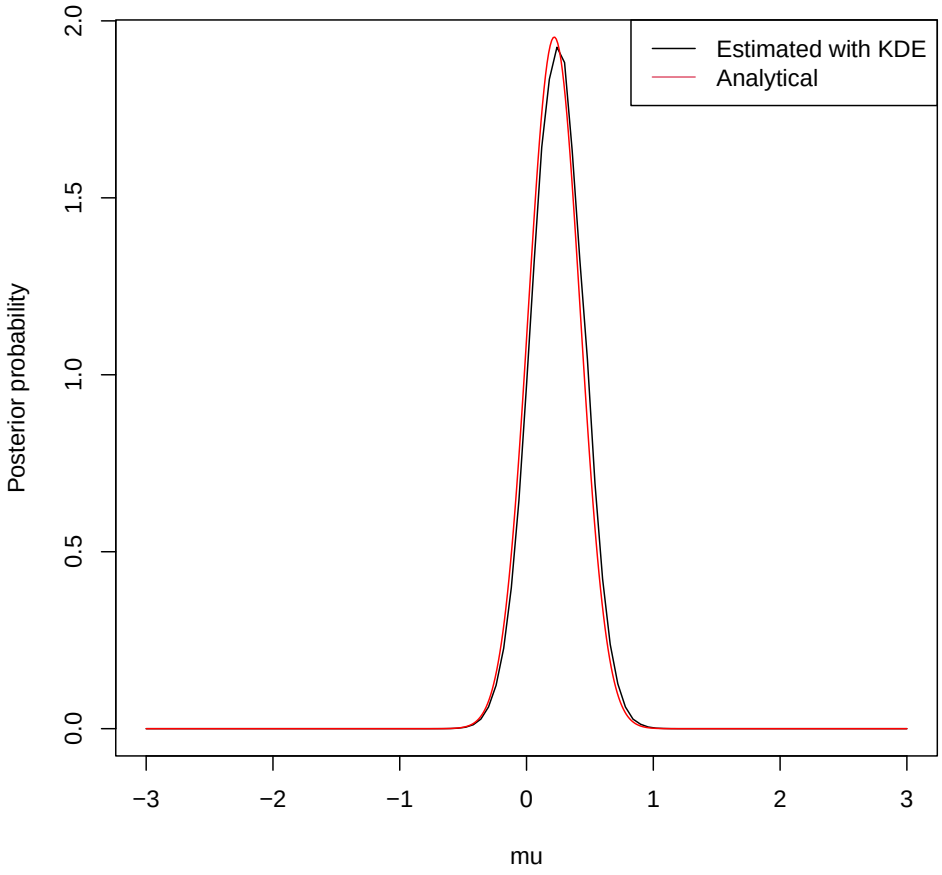


FIGURE A.2. The analytical posterior probability distribution for mu is compared with the distribution estimated using KDE. Their distributions are a close match.

Bibliography

Abrahams, M.-R., Joseph, S. B., Garrett, N., Tyers, L., Moeser, M., Archin, N., Council, O. D., Matten, D., Zhou, S., Doolabh, D., Anthony, C., Goonetilleke, N., Karim, S. A., Margolis, D. M., Pond, S. K., Williamson, C., and Swanstrom, R. 2019. The replication-competent HIV-1 latent reservoir is primarily established near the time of therapy initiation. *Science Translational Medicine*, 11(513): eaaw5589.

Al-Asadi, H., Dey, K. K., Novembre, J., and Stephens, M. 2019. Inference and visualization of DNA damage patterns using a grade of membership model. *Bioinformatics*, 35(8): 1292–1298.

Angelis, K. and Dos Reis, M. 2015. The impact of ancestral population size and incomplete lineage sorting on Bayesian estimation of species divergence times. *Current Zoology*, 61(5): 874–885.

Beaulieu, J. M. and O'Meara, B. C. 2016. Detecting hidden diversification shifts in models of trait-dependent speciation and extinction. *Systematic Biology*, 65(4): 583–601.

Bedford, T., Greninger, A. L., Roychoudhury, P., Starita, L. M., Famulare, M., Huang, M.-L., Nalla, A., Pepper, G., Reinhardt, A., Xie, H., Shrestha, L., Nguyen, T. N., Adler, A., Brandstetter, E., Cho, S., Giroux, D., Han, P. D., Fay, K., Frazar, C. D., Ilcisin, M., Lacombe, K., Lee, J., Kiavand, A., Richardson, M., Sibley, T. R., Truong, M., Wolf, C. R., Nickerson, D. A., Rieder, M. J., Englund, J. A., The Seattle Flu Study Investigators, Hadfield, J., Hodcroft, E. B., Huddleston, J., Moncla, L. H., Müller, N. F., Neher, R. A., Deng, X., Gu, W., Federman, S., Chiu, C., Duchin, J. S., Gautom, R., Melly, G., Hiatt, B., Dykema, P., Lindquist, S., Queen, K., Tao, Y., Uehara, A., Tong, S., MacCannell, D., Armstrong, G. L., Baird, G. S., Chu, H. Y., Shendure, J., and Jerome, K. R. 2020. Cryptic transmission of SARS-CoV-2 in Washington state. *Science*, 370(6516): 571–575.

Boore, J. L. 1999. Animal mitochondrial genomes. *Nucleic Acids Research*, 27(8): 1767–1780.

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., De Maio, N., Matschiner, M., Mendes, F. K., Müller, N. F.,

Ogilvie, H. A., du Plessis, L., Popinga, A., Rambaut, A., Rasmussen, D., Siveroni, I., Suchard, M. A., Wu, C.-H., Xie, D., Zhang, C., Stadler, T., and Drummond, A. J. 2019. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Computational Biology*, 15(4): e1006650.

Boussau, B. and Scornavacca, C. 2020. Reconciling Gene trees with Species Trees. In C. Scornavacca, F. Delsuc, and N. Galtier, editors, *Phylogenetics in the Genomic Era*, pages 3.2:1–3.2:23. No commercial publisher — Authors open access book.

Brodin, J., Zanini, F., Thebo, L., Lanz, C., Bratt, G., Neher, R. A., and Albert, J. 2016. Establishment and stability of the latent HIV-1 DNA reservoir. *Elife*, 5: e18889.

Bromham, L. and Penny, D. 2003. The modern molecular clock. *Nature Reviews Genetics*, 4(3): 216–224.

Bruner, K. M., Murray, A. J., Pollack, R. A., Soliman, M. G., Laskey, S. B., Capoferri, A. A., Lai, J., Strain, M. C., Lada, S. M., Hoh, R., Ho, Y.-C., Richman, D. D., Deeks, S. G., Siliciano, J. D., and Siliciano, R. F. 2016. Defective proviruses rapidly accumulate during acute HIV-1 infection. *Nature Medicine*, 22(9): 1043–1049.

Casella, G. and Berger, R. 2002. *Statistical Inference*. Duxbury advanced series in statistics and decision sciences. Thomson Learning.

Chang, D., Knapp, M., Enk, J., Lippold, S., Kircher, M., Lister, A., MacPhee, R. D. E., Widga, C., Czechowski, P., Sommer, R., Hodges, E., Stümpel, N., Barnes, I., Dalén, L., Derevianko, A., Germonpré, M., Hillebrand-Voiculescu, A., Constantin, S., Kuznetsova, T., Mol, D., Rathgeber, T., Rosendahl, W., Tikhonov, A. N., Willerslev, E., Hannon, G., Lalueza-Fox, C., Joger, U., Poinar, H., Hofreiter, M., and Shapiro, B. 2017. The evolutionary and phylogeographic history of woolly mammoths: a comprehensive mitogenomic analysis. *Scientific Reports*, 7(1): 44585.

Chun, T.-W., Carruth, L., Finzi, D., Shen, X., DiGiuseppe, J. A., Taylor, H., Hermankova, M., Chadwick, K., Margolick, J., Quinn, T. C., Kuo, Y.-H., Brookmeyer, R., Zeiger, M. A., Barditch-Crovo, P., and Siliciano, R. F. 1997. Quantification of latent tissue reservoirs and total body viral load in HIV-1 infection. *Nature*, 387(6629): 183–188.

Chun, T.-W., Engel, D., Berrey, M. M., Shea, T., Corey, L., and Fauci, A. S. 1998. Early establishment of a pool of latently infected, resting CD4+ T cells during primary HIV-1 infection.

Proceedings of the National Academy of Sciences, 95(15): 8869–8873.

Davey, R. T., Bhat, N., Yoder, C., Chun, T.-W., Metcalf, J. A., Dewar, R., Natarajan, V., Lempicki, R. A., Adelsberger, J. W., Miller, K. D., Kovacs, J. A., Polis, M. A., Walker, R. E., Falloon, J., Masur, H., Gee, D., Baseler, M., Dimitrov, D. S., Fauci, A. S., and Lane, H. C. 1999. HIV-1 and T cell dynamics after interruption of highly active antiretroviral therapy (HAART) in patients with a history of sustained viral suppression. *Proceedings of the National Academy of Sciences*, 96(26): 15109–15114.

Deeks, S. G., Overbaugh, J., Phillips, A., and Buchbinder, S. 2015. HIV infection. *Nature Reviews Disease Primers*, 1(1): 1–22.

Degnan, J. H. and Rosenberg, N. A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution*, 24(6): 332–340.

Dellicour, S., Lequime, S., Vrancken, B., Gill, M. S., Bastide, P., Gangavarapu, K., Matteson, N. L., Tan, Y., du Plessis, L., Fisher, A. A., Nelson, M. I., Gilbert, M., Suchard, M. A., Andersen, K. G., Grubaugh, N. D., Pybus, O. G., and Lemey, P. 2020. Epidemiological hypothesis testing using a phylogeographic and phylodynamic framework. *Nature Communications*, 11(1): 5620.

Di Giallonardo, F., Duchene, S., Puglia, I., Curini, V., Profeta, F., Cammà, C., Marcacci, M., Calistri, P., Holmes, E. C., and Lorusso, A. 2020. Genomic epidemiology of the first wave of SARS-CoV-2 in Italy. *Viruses*, 12(12).

Dinoso, J. B., Kim, S. Y., Wiegand, A. M., Palmer, S. E., Gange, S. J., Cranmer, L., O'Shea, A., Callender, M., Spivak, A., Brennan, T., Kearney, M. F., Proschan, M. A., Mican, J. M., Rehm, C. A., Coffin, J. M., Mellors, J. W., Siliciano, R. F., and Maldarelli, F. 2009. Treatment intensification does not reduce residual HIV-1 viremia in patients on highly active antiretroviral therapy. *Proceedings of the National Academy of Sciences*, 106(23): 9403–9408.

Douglas, J., Jiménez-Silva, C. L., and Bouckaert, R. 2022. StarBeast3: Adaptive parallelized Bayesian inference under the multispecies coalescent. *Systematic Biology*, 71(4): 901–916.

Drummond, A. J. and Rambaut, A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, 7(1): 1–8.

Drummond, A. J., Nicholls, G. K., Rodrigo, A. G., and Solomon, W. 2002. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence

data. *Genetics*, 161(3): 1307–1320.

Drummond, A. J., Pybus, O. G., Rambaut, A., Forsberg, R., and Rodrigo, A. G. 2003. Measurably evolving populations. *Trends in Ecology & Evolution*, 18(9): 481–488.

Drummond, A. J., Rambaut, A., Shapiro, B., and Pybus, O. G. 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution*, 22(5): 1185–1192.

Dufour, C., Gantner, P., Fromentin, R., and Chomont, N. 2020. The multifaceted nature of HIV latency. *Journal of Clinical Investigation*, 130(7): 3381–3390.

Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5): 1792–1797.

Feller, W. 1939. Die grundlagen der volterraschen theorie des kampfes ums dasein in wahrscheinlichkeitstheoretischer behandlung. *Acta Biotheoretica*, 5(1): 11–40.

Felsenstein, J. 1985. Phylogenies and the comparative method. *The American Naturalist*, 125(1): 1–15.

Felsenstein, J. and Kishino, H. 1993. Is there something wrong with the bootstrap on phylogenies? a reply to Hillis and Bull. *Systematic Biology*, 42(2): 193–200.

FitzJohn, R. G. 2010. Quantitative Traits and Diversification. *Systematic Biology*, 59(6): 619–633.

FitzJohn, R. G. 2012. Diversitree: comparative phylogenetic analyses of diversification in R. *Methods in Ecology and Evolution*, 3(6): 1084–1092.

Flouri, T., Jiao, X., Rannala, B., and Yang, Z. 2018. Species tree inference with BPP using genomic sequences and the multispecies coalescent. *Molecular Biology and Evolution*, 35(10): 2585–2593.

Flouri, T., Jiao, X., Rannala, B., and Yang, Z. 2020. A Bayesian implementation of the multispecies coalescent model with introgression for phylogenomic analysis. *Molecular Biology and Evolution*, 37(4): 1211–1223.

Flouri, T., Huang, J., Jiao, X., Kapli, P., Rannala, B., and Yang, Z. 2022. Bayesian phylogenetic inference using relaxed-clocks and the multispecies coalescent. *Molecular Biology and Evolution*, 39(8): msac161.

Flouri, T., Jiao, X., Huang, J., Rannala, B., and Yang, Z. 2023. Efficient Bayesian inference under the multispecies coalescent with migration. *Proceedings of the National Academy of Sciences*,

120(44): e2310708120.

Forest, F. 2009. Calibrating the tree of life: fossils, molecules and evolutionary timescales. *Annals of Botany*, 104(5): 789–794.

Fortes, G. G., Grandal-d'Anglade, A., Kolbe, B., Fernandes, D., Meleg, I. N., García-Vázquez, A., Pinto-Llona, A. C., Constantin, S., de Torres, T. J., Ortiz, J. E., Frischauf, C., Rabeder, G., Hofreiter, M., and Barlow, A. 2016. Ancient DNA reveals differences in behaviour and sociality between brown bears and extinct cave bears. *Molecular Ecology*, 25(19): 4907–4918.

Gao, J., May, M. R., Rannala, B., and Moore, B. R. 2022. New phylogenetic models incorporating interval-specific dispersal dynamics improve inference of disease spread. *Molecular Biology and Evolution*, 39(8): msac159.

Gao, J., May, M. R., Rannala, B., and Moore, B. R. 2023. Model misspecification misleads inference of the spatial dynamics of disease outbreaks. *Proceedings of the National Academy of Sciences*, 120(11): e2213913120.

Gavryushkina, A., Heath, T. A., Ksepka, D. T., Stadler, T., Welch, D., and Drummond, A. J. 2016. Bayesian total-evidence dating reveals the recent crown radiation of penguins. *Systematic Biology*, 66(1): 57–73.

Geyer, C. J. 1992. Practical Markov chain Monte Carlo. *Statistical Science*, 7(4): 473–483.

Gill, M. S., Lemey, P., Faria, N. R., Rambaut, A., Shapiro, B., and Suchard, M. A. 2012. Improving Bayesian population dynamics inference: A coalescent-based model for multiple loci. *Molecular Biology and Evolution*, 30(3): 713–724.

Gillespie, J. H. and Langley, C. H. 1979. Are evolutionary rates really variable? *Journal of Molecular Evolution*, 13(1): 27–34.

Goldberg, E. E., Kohn, J. R., Lande, R., Robertson, K. A., Smith, S. A., and Igić, B. 2010. Species selection maintains self-incompatibility. *Science*, 330(6003): 493–495.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H., Hansen, N. F., Durand, E. Y., Malaspinas, A. S., Jensen, J. D., Marques-Bonet, T., Alkan, C., Prufer, K., Meyer, M., Burbano, H. A., Good, J. M., Schultz, R., Aximu-Petri, A., Butthof, A., Hober, B., Hoffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E. S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic,

D., Kucan, Z., Gusic, I., Doronichev, V. B., Golovanova, L. V., Lalueza-Fox, C., de la Rasilla, M., Fortea, J., Rosas, A., Schmitz, R. W., Johnson, P. L., Eichler, E. E., Falush, D., Birney, E., Mullikin, J. C., Slatkin, M., Nielsen, R., Kelso, J., Lachmann, M., Reich, D., and Paabo, S. 2010. A draft sequence of the neandertal genome. *Science*, 328: 710–722.

Griffiths, R. C. and Tavare, S. 1994. Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 344(1310): 403–410.

Hajdas, I., Ascough, P., Garnett, M. H., Fallon, S. J., Pearson, C. L., Quarta, G., Spalding, K. L., Yamaguchi, H., and Yoneda, M. 2021. Radiocarbon dating. *Nature Reviews Methods Primers*, 1(1): 62.

Harvey, M. G., Bravo, G. A., Claramunt, S., Cuervo, A. M., Derryberry, G. E., Battilana, J., Seeholzer, G. F., McKay, J. S., O'Meara, B. C., Faircloth, B. C., Edwards, S. V., Pérez-Emán, J., Moyle, R. G., Sheldon, F. H., Aleixo, A., Smith, B. T., Chesser, R. T., Silveira, L. F., Cracraft, J., Brumfield, R. T., and Derryberry, E. P. 2020. The evolution of a tropical biodiversity hotspot. *Science*, 370(6522): 1343–1348.

Hasegawa, M., Kishino, H., and Yano, T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22: 160–174.

Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1): 97–109.

Hatano, H., Hayes, T. L., Dahl, V., Sinclair, E., Lee, T.-H., Hoh, R., Lampiris, H., Hunt, P. W., Palmer, S., McCune, J. M., Martin, J. N., Busch, M. P., Shacklett, B. L., and Deeks, S. G. 2011. A randomized, controlled trial of raltegravir intensification in antiretroviral-treated, HIV-infected patients with a suboptimal CD4+ T cell response. *The Journal of Infectious Diseases*, 203(7): 960–968.

Heath, T. A., Huelsenbeck, J. P., and Stadler, T. 2014. The fossilized birth–death process for coherent calibration of divergence-time estimates. *Proceedings of the National Academy of Sciences*, 111(29): E2957–E2966.

Hill, A. L., Rosenbloom, D. I. S., Fu, F., Nowak, M. A., and Siliciano, R. F. 2014. Predicting the outcomes of treatment to eradicate the latent reservoir for HIV-1. *Proceedings of the National*

Academy of Sciences, 111(37): 13475–13480.

Hillis, D. M. and Bull, J. J. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology*, 42(2): 182–192.

Ho, D. D., Neumann, A. U., Perelson, A. S., Chen, W., Leonard, J. M., and Markowitz, M. 1995. Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. *Nature*, 373(6510): 123–126.

Ho, S. Y. 2020. The molecular clock and evolutionary rates across the tree of life. *The Molecular Evolutionary Clock: Theory and Practice*, pages 3–23.

Hodgkinson, A. and Eyre-Walker, A. 2011. Variation in the mutation rate across mammalian genomes. *Nature Reviews Genetics*, 12(11): 756–766.

Horsburgh, K. A., Gosling, A. L., Cochrane, E. E., Kirch, P. V., Swift, J. A., and McCoy, M. D. 2022. Origins of Polynesian pigs revealed by mitochondrial whole genome ancient DNA. *Animals*, 12(18): 2469.

Hudson, R. 1990. Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology*, 7(1): 44.

Huelsenbeck, J. P. and Rannala, B. 2004. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. *Systematic Biology*, 53(6): 904–913.

Huey, R. B., Garland, T., and Turelli, M. 2019. Revisiting a key innovation in evolutionary biology: Felsenstein's "Phylogenies and the comparative method". *The American Naturalist*, 193(6): 755–772.

Jones, B. R. and Joy, J. B. 2020. Simulating within host human immunodeficiency virus 1 genome evolution in the persistent reservoir. *Virus Evolution*, 6(veaa089).

Jones, B. R. and Joy, J. B. 2023. Inferring human immunodeficiency virus 1 proviral integration dates with Bayesian inference. *Molecular Biology and Evolution*, 40(8): msad156.

Jones, B. R. and Poon, A. F. Y. 2017. node.dating: dating ancestors in phylogenetic trees in R. *Bioinformatics*, 33(6): 932–934.

Jones, B. R., Kinloch, N. N., Horacsek, J., Ganase, B., Harris, M., Harrigan, P. R., Jones, R. B., Brockman, M. A., Joy, J. B., Poon, A. F. Y., and Brumme, Z. L. 2018. Phylogenetic approach

to recover integration dates of latent HIV sequences within-host. *Proceedings of the National Academy of Sciences*, 115(38): E8958–E8967.

Jukes, T. and Cantor, C. 1969. *Evolution of Protein Molecules*, pages 21–123. Academic Press, New York.

Kaessmann, H., Wiebe, V., Weiss, G., and Pääbo, S. 2001. Great ape DNA sequences reveal a reduced diversity and an expansion in humans. *Nature Genetics*, 27(2): 155–156.

Kendall, D. G. 1950. An artificial realization of a simple "birth-and-death" process. *Journal of the Royal Statistical Society. Series B (Methodological)*, 12(1): 116–119.

Kingman, J. 1982a. The coalescent. *Stochastic Processes and their Applications*, 13(3): 235–248.

Kingman, J. F. C. 1982b. On the genealogy of large populations. *Journal of Applied Probability*, 19(A): 27–43.

Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., and Stamatakis, A. 2019. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, 35(21): 4453–4455.

Kühnert, D., Wu, C.-H., and Drummond, A. J. 2011. Phylogenetic and epidemic modeling of rapidly evolving infectious diseases. *Infection, Genetics and Evolution*, 11(8): 1825–1841.

Lemey, P., Rambaut, A., Drummond, A. J., and Suchard, M. A. 2009. Bayesian phylogeography finds its roots. *PLoS Computational Biology*, 5(9): e1000520.

Lewis, P. O., Holder, M. T., and Holsinger, K. E. 2005. Polytomies and Bayesian phylogenetic inference. *Systematic Biology*, 54(2): 241–253.

Li, H. and Durbin, R. 2011. Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357): 493–496.

Li, W.-H., Tanimura, M., and Sharp, P. M. 1988. Rates and dates of divergence between AIDS virus nucleotide sequences. *Molecular Biology and Evolution*, 5(4): 313–330.

Lintusaari, J., Gutmann, M. U., Dutta, R., Kaski, S., and Corander, J. 2016. Fundamentals and recent developments in approximate Bayesian computation. *Systematic Biology*, 66(1): e66–e82.

Liu, Y., McNevin, J., Cao, J., Zhao, H., Genowati, I., Wong, K., McLaughlin, S., McSweyn, M. D., Diem, K., Stevens, C. E., Maenza, J., He, H., Nickle, D. C., Shriner, D., Holte, S. E., Collier, A. C., Corey, L., McElrath, M. J., and Mullins, J. I. 2006. Selection on the human

immunodeficiency virus type 1 proteome following primary infection. *Journal of Virology*, 80(19): 9519–9529.

Liu, Y., McNevin, J., Zhao, H., Tebit, D. M., Troyer, R. M., McSweyn, M., Ghosh, A. K., Shriner, D., Arts, E. J., McElrath, M. J., and Mullins, J. I. 2007. Evolution of human immunodeficiency virus type 1 cytotoxic T-lymphocyte epitopes: fitness-balanced escape. *Journal of Virology*, 81(22): 12179–12188.

Liu, Y., McNevin, J. P., Holte, S., McElrath, M. J., and Mullins, J. I. 2011. Dynamics of viral evolution and CTL responses in HIV-1 infection. *PLoS One*, 6(1): e15639.

Lord, E., Dussex, N., Kierczak, M., Díez-del Molino, D., Ryder, O. A., Stanton, D. W. G., Gilbert, M. T. P., Sánchez-Barreiro, F., Zhang, G., Sinding, M.-H. S., Lorenzen, E. D., Willerslev, E., Protopopov, A., Shidlovskiy, F., Fedorov, S., Bocherens, H., Nathan, S. K. S. S., Goossens, B., van der Plicht, J., Chan, Y. L., Prost, S., Potapova, O., Kirillova, I., Lister, A. M., Heintzman, P. D., Kapp, J. D., Shapiro, B., Vartanyan, S., Götherström, A., and Dalén, L. 2020. Pre-extinction demographic stability and genomic signatures of adaptation in the woolly rhinoceros. *Current Biology*, 30(19): 3871–3879.e7.

Lu, J., Du Plessis, L., Liu, Z., Hill, V., Kang, M., Lin, H., Sun, J., François, S., Kraemer, M. U., Faria, N. R., McCrone, J. T., Peng, J., Xiong, Q., Yuan, R., Zeng, L., Zhou, P., Liang, C., Yi, L., Liu, J., Xiao, J., Hu, J., Liu, T., Ma, W., Li, W., Su, J., Zheng, H., Peng, B., Fang, S., Su, W., Li, K., Sun, R., Bai, R., Tang, X., Liang, M., Quick, J., Song, T., Rambaut, A., Loman, N., Raghwani, J., Pybus, O. G., and Ke, C. 2020. Genomic epidemiology of SARS-CoV-2 in Guangdong province, China. *Cell*, 181(5): 997–1003.

Luo, R., Piovoso, M. J., Martinez-Picado, J., and Zurakowski, R. 2012. HIV model parameter estimates from interruption trial data including drug efficacy and reservoir dynamics. *PLoS ONE*, 7(7): e40198.

MacDonald, G. M., Beilman, D. W., Kuzmin, Y. V., Orlova, L. A., Kremenetski, K. V., Shapiro, B., Wayne, R. K., and Van Valkenburgh, B. 2012. Pattern of extinction of the woolly mammoth in Beringia. *Nature Communications*, 3(1): 893.

Maddison, W. P. 1997. Gene trees in species trees. *Systematic Biology*, 46(3): 523–536.

Maddison, W. P., Midford, P. E., and Otto, S. P. 2007. Estimating a binary character's effect on speciation and extinction. *Systematic Biology*, 56(5): 701–710.

Mailund, T., Halager, A. E., Westergaard, M., Dutheil, J. Y., Munch, K., Andersen, L. N., Lunter, G., Prüfer, K., Scally, A., Hobolth, A., and Schierup, M. H. 2012. A new isolation with migration model along complete genomes infers very different divergence processes among closely related great ape species. *PLoS Genetics*, 8(12): e1003125.

Martins, E. P. 1996. *Phylogenies and the Comparative Method in Animal Behavior*, chapter 2. Oxford University Press, USA.

May, M. R., Höhna, S., and Moore, B. R. 2016. A Bayesian approach for detecting the impact of mass-extinction events on molecular phylogenies when rates of lineage diversification may vary. *Methods in Ecology and Evolution*, 7(8): 947–959.

McMahon, D., Jones, J., Wiegand, A., Gange, S. J., Kearney, M., Palmer, S., McNulty, S., Metcalf, J. A., Acosta, E., Rehm, C., Coffin, J. M., Mellors, J. W., and Maldarelli, F. 2010. Short-course raltegravir intensification does not reduce persistent low-level viremia in patients with HIV-1 suppression during receipt of combination antiretroviral therapy. *Clinical Infectious Diseases*, 50(6): 912–919.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. 1953. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6): 1087–1092.

Miller, W., Schuster, S. C., Welch, A. J., Ratan, A., Bedoya-Reina, O. C., Zhao, F., Kim, H. L., Burhans, R. C., Drautz, D. I., Wittekindt, N. E., Tomsho, L. P., Ibarra-Laclette, E., Herrera-Estrella, L., Peacock, E., Farley, S., Sage, G. K., Rode, K., Obbard, M., Montiel, R., Bachmann, L., Ingólfsson, O., Aars, J., Mailund, T., Wiig, Ø., Talbot, S. L., and Lindqvist, C. 2012. Polar and brown bear genomes reveal ancient admixture and demographic footprints of past climate change. *Proceedings of the National Academy of Sciences*, 109(36): E2382–E2390.

Minin, V. N., Bloomquist, E. W., and Suchard, M. A. 2008. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Molecular Biology and Evolution*, 25(7): 1459–1471.

Mitchell, D., Willerslev, E., and Hansen, A. 2005. Damage and repair of ancient DNA. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 571(1): 265–276.

Moore, B. R., Höhna, S., May, M. R., Rannala, B., and Huelsenbeck, J. P. 2016. Critically evaluating the theory and performance of Bayesian analysis of macroevolutionary mixtures. *Proceedings of the National Academy of Sciences*, 113(34): 9569–9574.

Moss, J. and Tveten, M. 2019. kdensity: An R package for kernel density estimation with parametric starts and asymmetric kernels. *Journal of Open Source Software*, 4(42): 1566.

Nascimento, F. F., Reis, M. d., and Yang, Z. 2017. A biologist's guide to Bayesian phylogenetic analysis. *Nature Ecology & Evolution*, 1(10): 1446–1454.

Nee, S., May, R. M., and Harvey, P. H. 1994. The reconstructed evolutionary process. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 344(1309): 305–311.

Nielsen, R., Akey, J. M., Jakobsson, M., Pritchard, J. K., Tishkoff, S., and Willerslev, E. 2017. Tracing the peopling of the world through genomics. *Nature*, 541(7637): 302–310.

Nowak, M. A. and Bangham, C. R. M. 1996. Population dynamics of immune responses to persistent viruses. *Science*, 272(5258): 74–79.

Ogilvie, H. A., Bouckaert, R. R., and Drummond, A. J. 2017. StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. *Molecular Biology and Evolution*, 34(8): 2101–2114.

Orlando, L., Allaby, R., Skoglund, P., Der Sarkissian, C., Stockhammer, P. W., Ávila Arcos, M. C., Fu, Q., Krause, J., Willerslev, E., Stone, A. C., and Warinner, C. 2021. Ancient DNA analysis. *Nature Reviews Methods Primers*, 1(1): 1–26.

Pagel, M. 1994. Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 255(1342): 37–45.

Palkopoulou, E., Mallick, S., Skoglund, P., Enk, J., Rohland, N., Li, H., Omrak, A., Vartanyan, S., Poinar, H., Götherström, A., Reich, D., and Dalén, L. 2015. Complete genomes reveal signatures of demographic and genetic declines in the woolly mammoth. *Current Biology*, 25(10): 1395–1400.

Palkopoulou, E., Lipson, M., Mallick, S., Nielsen, S., Rohland, N., Baleka, S., Karpinski, E., Ivancevic, A. M., To, T.-H., Kortschak, R. D., Raison, J. M., Qu, Z., Chin, T.-J., Alt, K. W., Claesson, S., Dalén, L., MacPhee, R. D. E., Meller, H., Roca, A. L., Ryder, O. A., Heiman, D., Young, S., Breen, M., Williams, C., Aken, B. L., Ruffier, M., Karlsson, E., Johnson, J., Di Palma, F., Alfoldi, J., Adelson, D. L., Mailund, T., Munch, K., Lindblad-Toh, K., Hofreiter, M., Poinar, H., and Reich, D. 2018. A comprehensive genomic history of extinct and living elephants. *Proceedings of the National Academy of Sciences*, 115(11): E2566–E2574.

Pankau, M. D., Reeves, D. B., Harkins, E., Ronen, K., Jaoko, W., Mandaliya, K., Graham, S. M., McClelland, R. S., Iv, F. A. M., Schiffer, J. T., Overbaugh, J., and Lehman, D. A. 2020. Dynamics of HIV DNA reservoir seeding in a cohort of superinfected Kenyan women. *PLOS Pathogens*, 16(2): e1008286.

Paradis, E. 2014. *An Introduction to the Phylogenetic Comparative Method*, pages 3–18. Springer Berlin Heidelberg, Berlin, Heidelberg.

Paradis, E., Claude, J., and Strimmer, K. 2004. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20(2): 289–290.

Parag, K. V., du Plessis, L., and Pybus, O. G. 2020. Jointly inferring the dynamics of population size and sampling intensity from molecular sequences. *Molecular Biology and Evolution*, 37(8): 2414–2429.

Pavia, J. M. 2015. Testing goodness-of-fit with the kernel density estimator: GoFKernel. *Journal of Statistical Software*, 66: 1–27.

Peluso, M. J., Bacchetti, P., Ritter, K. D., Beg, S., Lai, J., Martin, J. N., Hunt, P. W., Henrich, T. J., Siliciano, J. D., Siliciano, R. F., Laird, G. M., and Deeks, S. G. 2020. Differential decay of intact and defective proviral DNA in HIV-1–infected individuals on suppressive antiretroviral therapy. *JCI Insight*, 5(4): e132997.

Perelson, A. S. and Ribeiro, R. M. 2013. Modeling the within-host dynamics of HIV infection. *BMC Biology*, 11(1): 96.

Peyrégne, S. and Peter, B. M. 2020. AuthentiCT: a model of ancient DNA damage to estimate the proportion of present-day DNA contamination. *Genome Biology*, 21(1): 1–16.

Phillips, A. N. 1996. Reduction of HIV concentration during acute infection: Independence from a specific immune response. *Science*, 271(5248): 497–499.

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

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

Pyron, R. A. 2011. Divergence time estimation using fossils as terminal taxa and the origins of Lissamphibia. *Systematic Biology*, 60(4): 466–481.

Rabosky, D. L. 2014. Automatic detection of key innovations, rate shifts, and diversity-dependence on phylogenetic trees. *PLoS ONE*, 9(2): e89543.

Rabosky, D. L. and Goldberg, E. E. 2015. Model inadequacy and mistaken inferences of trait-dependent speciation. *Systematic Biology*, 64(2): 340–355.

Rambaut, A. 2000. Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics*, 16(4): 395–399.

Ramos-Madrigal, J., Smith, B. D., Moreno-Mayar, J. V., Gopalakrishnan, S., Ross-Ibarra, J., Gilbert, M. T. P., and Wales, N. 2016. Genome sequence of a 5,310-year-old maize cob provides insights into the early stages of maize domestication. *Current Biology*, 26(23): 3195–3201.

Rannala, B. 2016. Conceptual issues in Bayesian divergence time estimation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1699): 20150134.

Rannala, B. and Yang, Z. 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *Journal of Molecular Evolution*, 43: 304–311.

Rannala, B. and Yang, Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*, 164(4): 1645–1656.

Rannala, B. and Yang, Z. 2017. Efficient Bayesian species tree inference under the multispecies coalescent. *Systematic Biology*, 66: 823–842.

Rannala, B., Zhu, T., and Yang, Z. 2011. Tail paradox, partial identifiability, and influential priors in Bayesian branch length inference. *Molecular Biology and Evolution*, 29(1): 325–335.

Rasmussen, M., Li, Y., Lindgreen, S., Pedersen, J. S., Albrechtsen, A., Moltke, I., Metspalu, M., Metspalu, E., Kivisild, T., Gupta, R., Bertalan, M., Nielsen, K., Gilbert, M. T. P., Wang, Y.,

Raghavan, M., Campos, P. F., Kamp, H. M., Wilson, A. S., Gledhill, A., Tridico, S., Bunce, M., Lorenzen, E. D., Binladen, J., Guo, X., Zhao, J., Zhang, X., Zhang, H., Li, Z., Chen, M., Orlando, L., Kristiansen, K., Bak, M., Tommerup, N., Bendixen, C., Pierre, T. L., Grønnow, B., Meldgaard, M., Andreasen, C., Fedorova, S. A., Osipova, L. P., Higham, T. F. G., Ramsey, C. B., Hansen, T. v. O., Nielsen, F. C., Crawford, M. H., Brunak, S., Sicheritz-Pontén, T., Villems, R., Nielsen, R., Krogh, A., Wang, J., and Willerslev, E. 2010. Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature*, 463(7282): 757–762.

Renaud, G., Slon, V., Duggan, A. T., and Kelso, J. 2015. Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA. *Genome Biology*, 16(1): 1–18.

Rieux, A. and Balloux, F. 2016. Inferences from tip-calibrated phylogenies: a review and a practical guide. *Molecular Ecology*, 25(9): 1911–1924.

Robert, C. P. and Casella, G. 2004. *Monte Carlo Statistical Methods*. Springer New York, NY.

Roca, A. L. 2019. African elephant genetics: enigmas and anomalies. *Journal of Genetics*, 98(3): 83.

Rodrigo, A. G. and Felsenstein, J. 1999. *The Evolution of HIV*, chapter Coalescent Approaches to HIV Population Genetics, pages 233–267. The John Hopkins University Press.

Rohland, N., Malaspinas, A.-S., Pollack, J. L., Slatkin, M., Matheus, P., and Hofreiter, M. 2007. Proboscidean mitogenomics: Chronology and mode of elephant evolution using mastodon as outgroup. *PLoS Biology*, 5(8): e207.

Rohland, N., Reich, D., Mallick, S., Meyer, M., Green, R. E., Georgiadis, N. J., Roca, A. L., and Hofreiter, M. 2010. Genomic DNA sequences from mastodon and woolly mammoth reveal deep speciation of forest and savanna elephants. *PLoS Biology*, 8(12): e1000564.

Römpler, H., Rohland, N., Lalueza-Fox, C., Willerslev, E., Kuznetsova, T., Rabeder, G., Bertranpetit, J., Schöneberg, T., and Hofreiter, M. 2006. Nuclear gene indicates coat-color polymorphism in mammoths. *Science*, 313(5783): 62–62.

Ronquist, F. 2004. Bayesian inference of character evolution. *Trends in Ecology & Evolution*, 19(9): 475–481.

Ross, S. M. 1997. *Simulation*. Academic Press, Inc., 2nd edition.

Rozewicki, J., Li, S., Amada, K. M., Standley, D. M., and Katoh, K. 2019. MAFFT-DASH: integrated protein sequence and structural alignment. *Nucleic Acids Research*, 47(W1): W5–W10.

Sanderson, M. J. 2002. Estimating absolute rates of molecular evolution and divergence times: A penalized likelihood approach. *Molecular Biology and Evolution*, 19(1): 101–109.

Shapiro, B., Drummond, A. J., Rambaut, A., Wilson, M. C., Matheus, P. E., Sher, A. V., Pybus, O. G., Gilbert, M. T. P., Barnes, I., Binladen, J., Willerslev, E., Hansen, A. J., Baryshnikov, G. F., Burns, J. A., Davydov, S., Driver, J. C., Froese, D. G., Harington, C. R., Keddie, G., Kosintsev, P., Kunz, M. L., Martin, L. D., Stephenson, R. O., Storer, J., Tedford, R., Zimov, S., and Cooper, A. 2004. Rise and fall of the Beringian steppe bison. *Science*, 306(5701): 1561–1565.

Siliciano, J. D., Kajdas, J., Finzi, D., Quinn, T. C., Chadwick, K., Margolick, J. B., Kovacs, C., Gange, S. J., and Siliciano, R. F. 2003. Long-term follow-up studies confirm the stability of the latent reservoir for HIV-1 in resting CD4 + T cells. *Nature Medicine*, 9(6): 727–728.

Siliciano, R. F. and Greene, W. C. 2011. HIV latency. *Cold Spring Harbor Perspectives in Medicine*, 1(1): a007096.

Skoglund, P., Northoff, B. H., Shunkov, M. V., Derevianko, A. P., Pääbo, S., Krause, J., and Jakobsson, M. 2014. Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. *Proceedings of the National Academy of Sciences*, 111(6): 2229–2234.

Slatkin, M. and Hudson, R. R. 1991. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics*, 129(2): 555–562.

Soetaert, K., Petzoldt, T., and Setzer, R. W. 2010. Solving differential equations in R : Package deSolve. *Journal of Statistical Software*, 33(9).

Soubrier, J., Gower, G., Chen, K., Richards, S. M., Llamas, B., Mitchell, K. J., Ho, S. Y. W., Kosintsev, P., Lee, M. S. Y., Baryshnikov, G., Bollongino, R., Bover, P., Burger, J., Chivall, D., Crégut-Bonnoure, E., Decker, J. E., Doronichev, V. B., Douka, K., Fordham, D. A., Fontana, F., Fritz, C., Glimmerveen, J., Golovanova, L. V., Groves, C., Guerreschi, A., Haak, W., Higham, T., Hofman-Kamińska, E., Immel, A., Julien, M.-A., Krause, J., Krotova, O., Langbein, F., Larson, G., Rohrlach, A., Scheu, A., Schnabel, R. D., Taylor, J. F., Tokarska, M., Tosello, G., van der Plicht, J., van Loenen, A., Vigne, J.-D., Wooley, O., Orlando, L., Kowalczyk, R., Shapiro, B.,

and Cooper, A. 2016. Early cave art and ancient DNA record the origin of European bison. *Nature Communications*, 7(1): 13158.

Stadler, T. and Yang, Z. 2013. Dating phylogenies with sequentially sampled tips. *Systematic Biology*, 62(5): 674–688.

Stafford, M. A., Corey, L., Cao, Y., Daar, E. S., Ho, D. D., and Perelson, A. S. 2000. Modeling plasma virus concentration during primary HIV infection. *Journal of Theoretical Biology*, 203(3): 285–301.

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., and Rambaut, A. 2018. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution*, 4(1): vey016.

Suzuki, Y., Glazko, G. V., and Nei, M. 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. *Proceedings of the National Academy of Sciences*, 99(25): 16138–16143.

Tavaré, S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences*, 17(2): 57–86.

Thorne, J. L., Kishino, H., and Painter, I. S. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Molecular Biology and Evolution*, 15(12): 1647–1657.

To, T.-H., Jung, M., Lycett, S., and Gascuel, O. 2016. Fast dating using least-squares criteria and algorithms. *Systematic Biology*, 65(1): 82–97.

van der Valk, T., Pečnerová, P., Díez-del Molino, D., Bergström, A., Oppenheimer, J., Hartmann, S., Xenikoudakis, G., Thomas, J. A., Dehasque, M., Sağlıcan, E., Fidan, F. R., Barnes, I., Liu, S., Somel, M., Heintzman, P. D., Nikolskiy, P., Shapiro, B., Skoglund, P., Hofreiter, M., Lister, A. M., Götherström, A., and Dalén, L. 2021. Million-year-old DNA sheds light on the genomic history of mammoths. *Nature*, 591(7849): 265–269.

Varga, T., Krizsán, K., Földi, C., Dima, B., Sánchez-García, M., Sánchez-Ramírez, S., Szöllősi, G. J., Szarkándi, J. G., Papp, V., Albert, L., Andreopoulos, W., Angelini, C., Antonín, V., Barry, K. W., Bougher, N. L., Buchanan, P., Buyck, B., Bense, V., Catcheside, P., Chovatia, M., Cooper, J., Dämon, W., Desjardin, D., Finy, P., Geml, J., Haridas, S., Hughes, K., Justo, A., Karasiński, D., Kautmanova, I., Kiss, B., Kocsubé, S., Kotiranta, H., LaButti, K. M., Lechner, B. E., Liimatainen, K., Lipzen, A., Lukács, Z., Mihaltcheva, S., Morgado, L. N., Niskanen, T.,

Noordeloos, M. E., Ohm, R. A., Ortiz-Santana, B., Ovrebo, C., Rácz, N., Riley, R., Savchenko, A., Shiryaev, A., Soop, K., Spirin, V., Szebenyi, C., Tomšovský, M., Tulloss, R. E., Uehling, J., Grigoriev, I. V., Vágvölgyi, C., Papp, T., Martin, F. M., Miettinen, O., Hibbett, D. S., and Nagy, L. G. 2019. Megaphylogeny resolves global patterns of mushroom evolution. *Nature Ecology & Evolution*, 3(4): 668–678.

Verhofstede, C., Noë, A., Demecheleer, E., De Cabooter, N., Van Wanzeele, F., Van Der Gucht, B., Vogelaers, D., and Plum, J. 2004. Drug-resistant variants that evolve during nonsuppressive therapy persist in HIV-1–infected peripheral blood mononuclear cells after long-term highly active antiretroviral therapy. *Journal of Acquired Immune Deficiency Syndromes*, 35(5): 473–483.

Volz, E. M., Koelle, K., and Bedford, T. 2013. Viral phylodynamics. *PLOS Computational Biology*, 9(3): e1002947.

Wei, X., Ghosh, S. K., Taylor, M. E., Johnson, V. A., Emini, E. A., Deutsch, P., Lifson, J. D., Bonhoeffer, S., Nowak, M. A., Hahn, B. H., Saag, M. S., and Shaw, G. M. 1995. Viral dynamics in human immunodeficiency virus type 1 infection. *Nature*, 373(6510): 117–122.

Whitney, J. B., Hill, A. L., Sanisetty, S., Penaloza-MacMaster, P., Liu, J., Shetty, M., Parenteau, L., Cabral, C., Shields, J., Blackmore, S., Smith, J. Y., Brinkman, A. L., Peter, L. E., Mathew, S. I., Smith, K. M., Borducchi, E. N., Rosenbloom, D. I. S., Lewis, M. G., Hattersley, J., Li, B., Hesselgesser, J., Geleziunas, R., Robb, M. L., Kim, J. H., Michael, N. L., and Barouch, D. H. 2014. Rapid seeding of the viral reservoir prior to SIV viraemia in rhesus monkeys. *Nature*, 512(7512): 74–77.

Worobey, M., Gemmel, M., Teuwen, D. E., Haselkorn, T., Kunstman, K., Bunce, M., Muyembe, J.-J., Kabongo, J.-M. M., Kalengayi, R. M., Van Marck, E., Gilbert, M. T. P., and Wolinsky, S. M. 2008. Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature*, 455(7213): 661–664.

Yang, Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution*, 10(6): 1396–1401.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution*, 39: 306–314.

Yang, Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24(8): 1586–1591.

Yang, Z. 2014. *Molecular Evolution: A Statistical Approach*. Oxford University Press, Oxford, England.

Yang, Z. and Rannala, B. 2005. Branch-length prior influences Bayesian posterior probability of phylogeny. *Systematic Biology*, 54(3): 455–470.

Yang, Z. and Rannala, B. 2006. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Molecular Biology and Evolution*, 23(1): 212–226.

Yang, Z. and Rannala, B. 2012. Molecular phylogenetics: principles and practice. *Nature Reviews Genetics*, 13(5): 303–314.

Yang, Z. and Rodríguez, C. E. 2013. Searching for efficient Markov chain Monte Carlo proposal kernels. *Proceedings of the National Academy of Sciences*, 110(48): 19307–19312.

Yang, Z. and Zhu, T. 2018. Bayesian selection of misspecified models is overconfident and may cause spurious posterior probabilities for phylogenetic trees. *Proceedings of the National Academy of Sciences*, 115(8): 1854–1859.

Yule, G. U. 1925. II.—A mathematical theory of evolution, based on the conclusions of Dr. J.C. Willis, F.R.S. *Philosophical Transactions of the Royal Society of London. Series B*, 213(402-410): 21–87.

Zhang, C., Stadler, T., Klopfstein, S., Heath, T. A., and Ronquist, F. 2015. Total-evidence dating under the fossilized birth–death process. *Systematic Biology*, 65(2): 228–249.