Accounting for sampling bias in ancestral state reconstruction

by

Yexuan Song

B.Sc., University of Toronto, 2019

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
Department of Mathematics
Faculty of Science

© Yexuan Song 2022
SIMON FRASER UNIVERSITY
Summer 2022

Copyright in this work is held by the author. Please ensure that any reproduction
or re-use is done in accordance with the relevant national copyright legislation.

Declaration of Committee

Name:	**Yexuan Song**
Degree:	**Master of Science**
Thesis title:	**Accounting for sampling bias in ancestral state reconstruction**
Committee:	**Chair:** Ben Ashby Associate Professor, Mathematics
	Caroline Colijn Co-supervisor Professor, Mathematics
	Ailene MacPherson Co-supervisor Assistant Professor, Mathematics
	Paul Tupper Committee Member Professor, Mathematics
	Cedric Chauve Examiner Professor, Mathematics

Abstract

Viral transmission plays an essential role in our understanding of and response to infectious diseases, for example, by informing policy decisions about transportation and borders. Phylogenetic methods take advantage of the evolutionary relationships between the genome sequences of viruses to infer geographical locations of unobserved ancestors from sampled data. Here I introduce a new approach to examine the inference of ancestral locations and predict the geographic movement of viral lineages from the known locations of samples. In contrast to existing methods, my method accounts for differences in sampling policies among areas, to avoid biased inference of the ancestral locations. I begin by summarizing existing methods for ancestral state reconstruction. I then introduce an ancestral state reconstruction method that accounts for variation in sampling rate among locations and compare it to the classic Maximum Likelihood method. I show that my method infers ancestral states for small trees more accurately than this classic approach.

Keywords: Zoonosis,Infectious disease, Transmission , Phylogeography

Dedication

For my grandmother

Acknowledgements

Caroline Colijn and Ailene MacPherson have been ideal teachers, mentors and supervisors, offering great advice and encouragement. Caroline helped me set up the topic and gave me many good references to start my thesis. She guides me through this journey with valuable thoughts and comments. Ailene helps me with details of my methods and codings. It would take me 'a hundred years' to finish my work. They are always available when I ask for help. I am proud to say they are the best supervisors I have ever had during my academic career. It will be my pleasure to work with them again in my Ph.D.

I want to thank the MAGPIE members Amy Langdon, Aniket Mane, Ben Sobkowiak, Elisha Are, Jaskaran Oberoi, Kurnia Susvitasari, Lisa McQuarrie, Madi Yerlanov, Nicola Mulberry, Niloufar Abhari, Omid Geysar, Pengyu Liu, Pouya Haghmaram and Vijay Naidu. They have collaborated with me in the past or given me valuable feedback in the group meetings and in general. I would also like to thank faculty members Paul Tupper, Lloyd Elliott, Cedric Chauve, Ben Ashby and Sandy Rutherford. They gave me insightful lectures or presentations.

I would also like to thank my friends Jingzhou Na, Pengyu Liu, Xin Wei and Xinyu Zhang, with who I spent most of my time and supported my thesis defence. Thanks to them, I have had an incredible year in this miserable time. Thanks to them, I can see Vancouver's stunning view rather than be stuck on Burnaby Mountain.

Last but not least, I want to thank my parents, Shijun Song and Yong Wang. They support me emotionally and financially to complete my bachelor's degree and master's degree.

Lastly, I would like to quote a sentence from one of my favourite YouTubers Alexander Technobalde who passed away last month: “That one day, we’ll look back at where we started and be amazed by how far we’ve come.”

Table of Contents

List of Figures

Chapter 1

Introduction

A phylogenetic tree is an acyclic graph where leaves (tips) correspond to organisms or samples with labels, and internal nodes are unobserved and inferred. It represents the evolutionary relationships among biological species or the history of viral transmission among hosts in the epidemiological case that we focus on here. Phylogenetic trees can be reconstructed from the evolutionary relationships between sampled viral genome sequences. In addition to these sequences, sampled viruses may be characterized by many discrete or continuous "character states," such as their current location or the species of host they infect. Ancestral state reconstruction (ASR) is a fundamental method in phylogenetics for identifying the character states of evolutionary ancestors. Ancestral state reconstruction methods use the observed character states at each tip (sampled lineage) in a phylogenetic tree to infer ancestral states consistent with the genome-wide evolutionary history.

Ancestral state reconstruction has critical applications in both micro-and macro-evolution. For instance, ASR is used to study whether plant chemical defences are becoming stronger over time to protect them from herbivores in the same environment and to understand co-evolution.[1] Nextstrain, an open-source website tracking pathogen evolution, applies the ancestral state reconstruction, in the form of the commonly used Maximum Likelihood method first introduced by Pagel [2] discussed in more detail below, to infer viral transmissions of the SARS-CoV-2 pandemic [3] and to identify key introduction events among countries and other pathogens such as influenza and Tuberculosis. This application of ASR is an example of the much-border application of ASR to the field of phylogeography. The aim of phylogeography in the epidemiological context is to describe the role of geographi-

cal space on viral transmission, including the inference of viral introductions into distinct geographical regions and to determine the origin of outbreaks.

Several ancestral state reconstruction methods have been developed over the past few decades [4]. These methods can be broadly summarized into three categories: Maximum Parsimony, Maximum Likelihood and Bayesian methods. The Maximum Parsimony method minimizes the number of character state changes that explain the states (geographical locations) observed at the tips of phylogenetic trees. The Maximum Likelihood method attempts to find one set of ancestral states that optimize the probability of observing the sampled locations given the phylogeny. The Bayesian method infers each ancestor's full "posterior" probability distribution for a given set of sequences on fixed trees [5]. Regardless of the method used, one fundamental issue of ancestral state reconstruction that has yet to be addressed is the unequal sampling rates among locations. Intuitively if sampling bias is unaccounted for, the more samples that are collected from a given location, the more likely it will be that the inferred internal states will be from that location. Oversampling a region will artificially increase its representation in the sample and cause us to infer a more significant movement into that region (i.e. transmission to that species in the multi-species cases) and a higher probability of remaining in that region (i.e. higher transmission rate within that species). As an illustration, consider Figure 1.1 showing the phylogeny from Ebola sequences sampled through time from both humans and bats. Given the over-representation of human samples, current ASR methods will infer the root state as human and conclude that humans transmitted the virus to bats [6]. However, it is known that Ebola is a zoonotic virus that spills over into humans from bats [7]. Due to the lack of viral sequences collected from the bat, the ancestral state reconstruction result will be misleading.

It is worth mentioning that in my thesis, I will only refer to the above methods in the ancestral state reconstruction realm. Even though Maximum Parsimony, Maximum Likelihood and Bayesian methods also refer to phylogenetic tree reconstruction, I am only focusing on the ancestral state reconstruction methods given a fixed "known" phylogenetic tree.

Extensive works have been done on characterizing the impact of sampling bias on phylogeographical analysis. De Maio [6] compared different ASR methods (Discrete Trait Analysis and MultiType Tree) using an Ebola data set. Despite the higher prevalence of Ebola in bat's, the data set consists of 78 samples from human patients and 7 samples from bats. They showed that current methods could give unintuitive results based on the settings: the existing phylogeography methods would conclude that unseen human-to-human transmissions cause the spreading due to sampling bias.

Magee & Scotch [8] also studied the effect of different down-sampling schemes in a Bayesian framework on the reconstruction of the root state. They proposed two different sampling schemes: (1) randomly downsample a certain percentage of the data irregadless of location and (2) downsample a different percentage in each location. They found that the correct root state can be accurately reconstructed even with a limited amount of data (downsampled to 25%-50% of the data, regardless of downsampling schemes.); they concluded that including most of the data (more than 90%) did not necessarily improve the root reconstruction. Furthermore, if only 10% of the data were sampled, bias sampling would result in lower reconstruction accuracy than selected 10% of the data from each region. Together these results suggest that the reconstruction accuracy is more sensitive to sampling bias than the small sample size, and hence the collection of larger, but yet biased, data sets are unlikely to improve reconstruction accuracy.

To complement these previous findings, my coauthors and I quantify the effect of sampling bias on multiple aspects of ASR (internal node states, root states, and migration events) using a simulation approach [9]. We demonstrated that biased downsampling could result in misleading inferences about the movement of viral lineages between locations. I used available Ebola sequences [10]. Below is a summary of my work as it pertains to the motivation and aims of this thesis. The data consisted of 262 tips collected from 4 counties: Guinea, Liberia, Mali and Sierra Leone. The time-scaled tree is generated using BEAST1.0 [11]. I examined two subsampling schemes: (1): removal of 80% recent migrants (individuals whose location differed from the inferred state of their immediate ancestor when the full data set was used) and (2): removal of 80% samples at a specific location (Guinea or Sierra

Leone). In this tree (Figure 1.2 panel A), the rate of movement among locations is low. The tree (shown in Figure 1.2 panel A) exhibits monomorphic clades: nodes in the same location are grouped. As a result of this low movement rate, the reconstruction accuracy is high across downsampling schemes. However, downsampling results in inaccurate state reconstruction at several individual nodes or groups of nodes relative to that found with the complete data set (red boxes indicated in Figure 1.3 and Figure 1.4). These cases illustrate that cross-jurisdiction disease introduction events can be inaccurately reconstructed due to sampling bias. For example, the red box in Figure 1.3 indicates that the virus is introduced from Liberia to Guinea, which is incorrect in the "true tree" containing the full data set in which transmission is inferred to be from Sierra Leone to Guinea. Downsampling recent migrant tips have similar but less drastic effects in this case. The ancestral state reconstruction is accurate except at the internal nodes in the red box in Figure 1.2; after downsampling, the introduction appears to be from Sierra Leone to Guinea, whereas the introduction event in the true tree is from Liberia to Guinea.

Maddison et al. [12] introduced the binary state speciation and extinction (BiSSE) model that estimates speciation, extinction and transition rates, assuming phylogenetic trees are generated under the birth-death process. Recent work by Freyman and Hohana [13] used the idea of BiSSE and "Stochastic Mapping" methods to accurately reconstruct ancestral states and the location of evolutionary transitions accounting for state-dependent speciation and extinction. Fitzjohn [14] extended the BiSSE model by considering more than two states and, importantly for us here, including the possibility of unequal sampling rates among types.

Notice that the aims of phylogeographic analysis are two-fold: 1) to estimate the transition (aka migration) rate between states and 2) to estimate the internal states. In the thesis, I am infer ancestral states of the internal nodes of a fixed tree and given the fixed migration rates between states.

In the introduction, I will summarize all the well-known methods that estimate ancestral states and the methods that my idea comes from. I will give some illustrative examples using these methods and discuss their advantages and weaknesses. In the next chapter, I

will describe my approach in detail: I will combine the methods described in [12] and [13] to account for sampling bias. In the third chapter, I will present preliminary results and compare accuracy to the classical Maximum Likelihood method. Finally, I will discuss the limitations of my approach and propose further directions.

1.1 Maximum Parsimony

The Maximum Parsimony method for ancestral state reconstruction aims to infer internal states that require the minimum number of state changes in the entire phylogeny. Edwards and Cavalli-Sforza first introduced the idea of "minimum evolution" (later called parsimony) in 1963 [15, 16] based on Darwin's theory of evolution (similar species are closely related). They proposed that the most likely evolutionary tree is the one with a minimum amount of evolution if there is no evolutionary history or fossil record. Later Farries [17, 18] and Fitch [19] formalized the idea of Maximum Parsimony methods using mathematical equations.

There exists several heuristic methods for Parsimony reconstruction of ancestral states of fixed trees [18,20,21] developed in the 1970's and 80's. More recently Miklos Csuros proposed an alternative dynamic programming method for optimizing internal states. Here, however, I focus on one of the former methods presented by Swofford and Maddison [21] that remains widely used and because this algorithm exemplifies the methodology employed later in this thesis of applying both post-traversal and pre-traversal algorithms to infer ancestral states is later widely used in the field of ancestral state reconstruction, which I will describe in the later sections. Consider a fixed tree with known tip states. The method first applies a post-traversal algorithm to find the set of "downpass" states, the downpass states of internal nodes represent the probable states of internal nodes as determined by (only) the states of their direct descendants. Next, a pre-traversal algorithm, which moves from root to tip, is applied to find the uppass set that each internal nodes inherits directly from their immediate ancestors. Finally, the algorithm uses both downpass and uppass sets to infer the final state of the internal node. The downpass and uppass sets can be calculated by the following rules:

1. We begin by performing the post-traversal algorithm. To find the downpass state for node N_d, consider its child nodes' downpass states $M1_d$ and $M2_d$ (tip states are

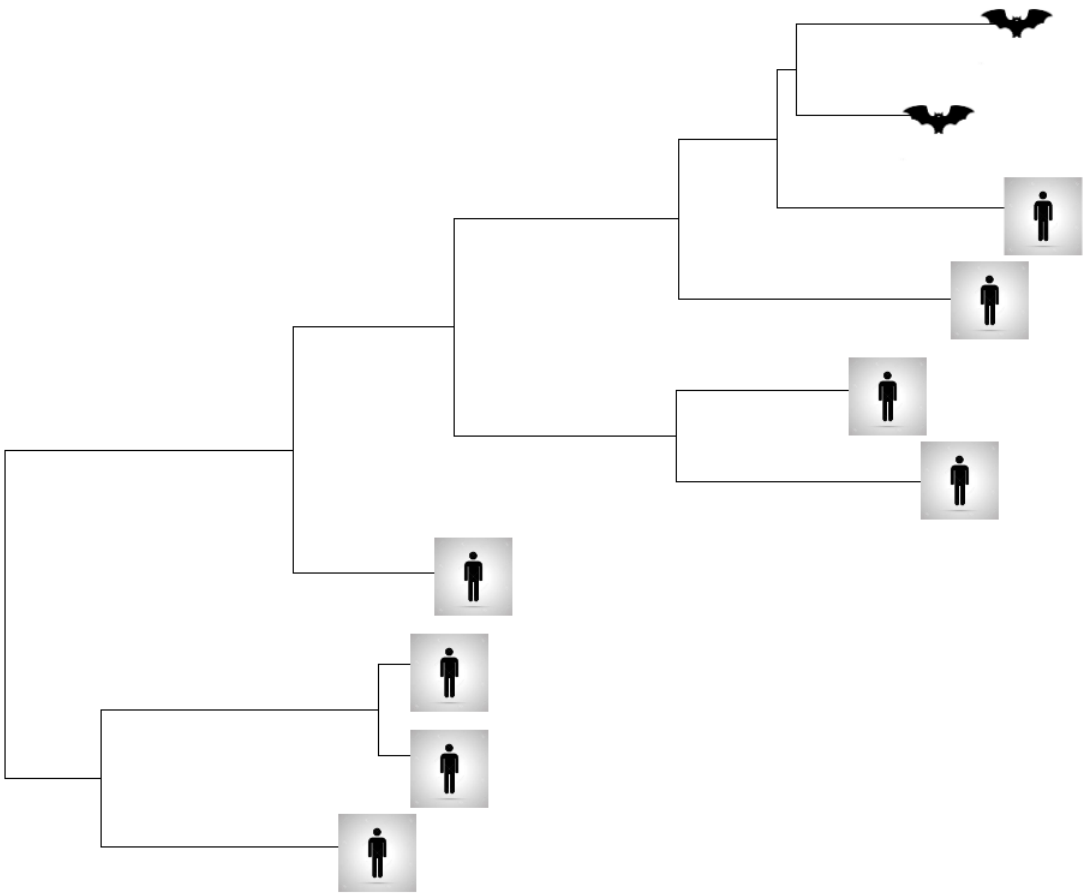

Figure 1.1: **Misleading ancestral reconstruction due to sampling bias.** Suppose Ebola viral sequences are sampled through time among humans and bats, and if humans have much more samples than bats, it is more likely that the most common ancestor is inferred as human.

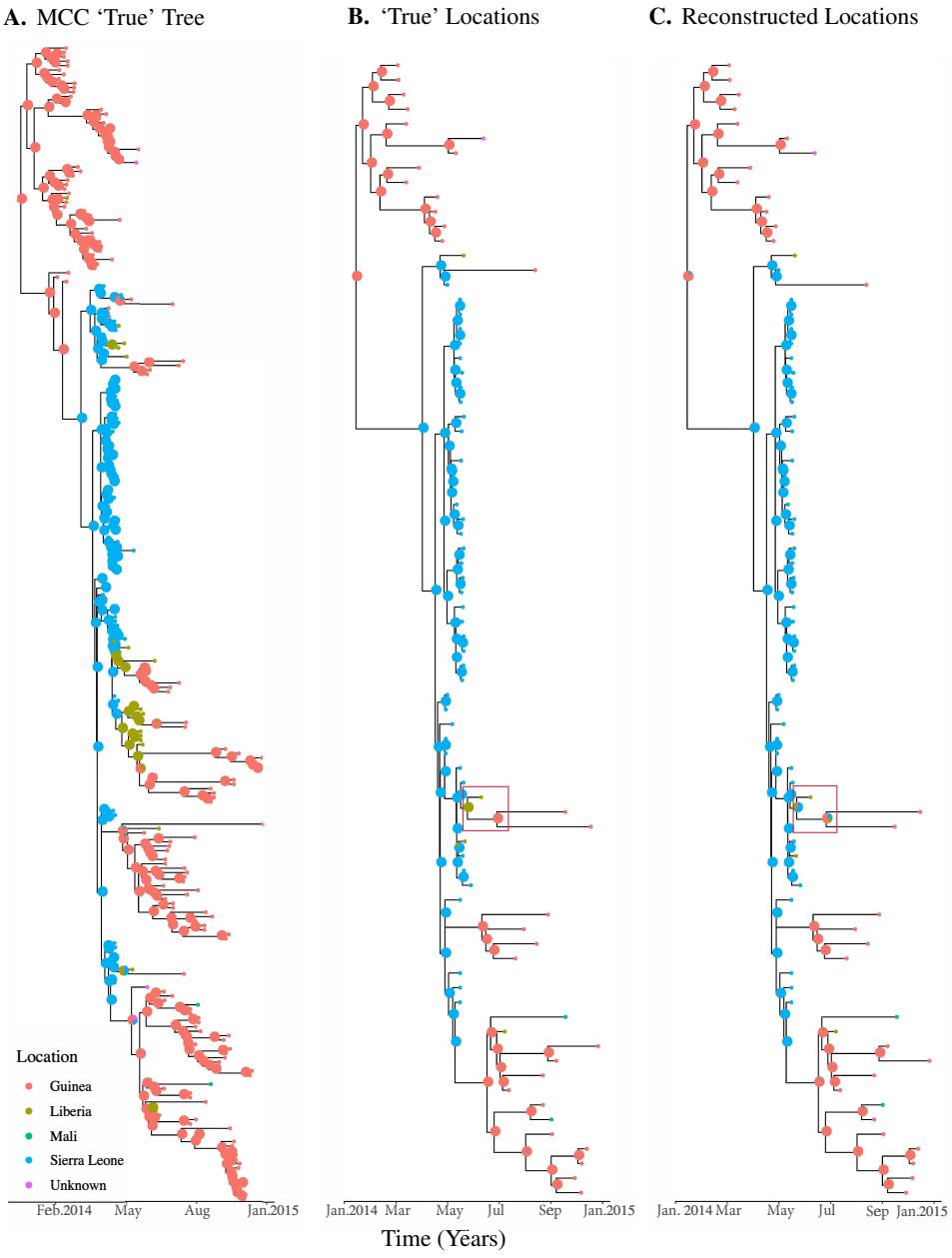

Figure 1.2: **Ancestral state reconstruction of Ebolavirus trees with recent migrants downsampled.**. Panel A: True tree with the correct internal states (Panel A). Downsampled trees were obtained by dropping 80% of recent migrants with either the true ancestral locations (Panel B) or ancestral locations reconstructed after downsampling migrants (Panel C). 5

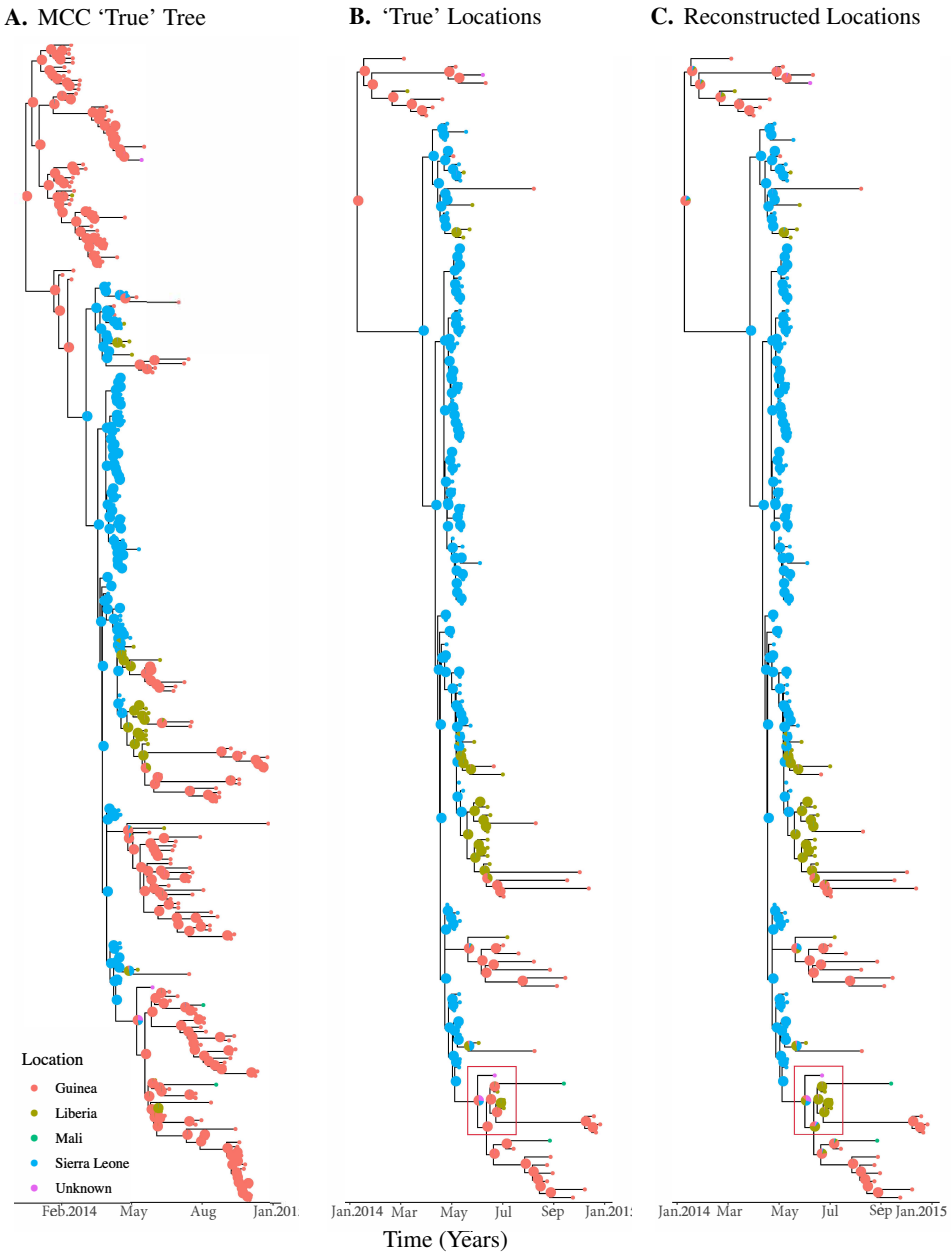

Figure 1.3: **Ancestral state reconstruction of Ebola virus trees with Guinea downsampled**. Panel A: True tree with the correct internal states (Panel A). Downsampled tree was obtained by dropping 80% of Guinea samples with true ancestral locations (Panel B). After dropping Guinea samples, the downsampled tree with reconstructed ancestral location (Panel C). 15

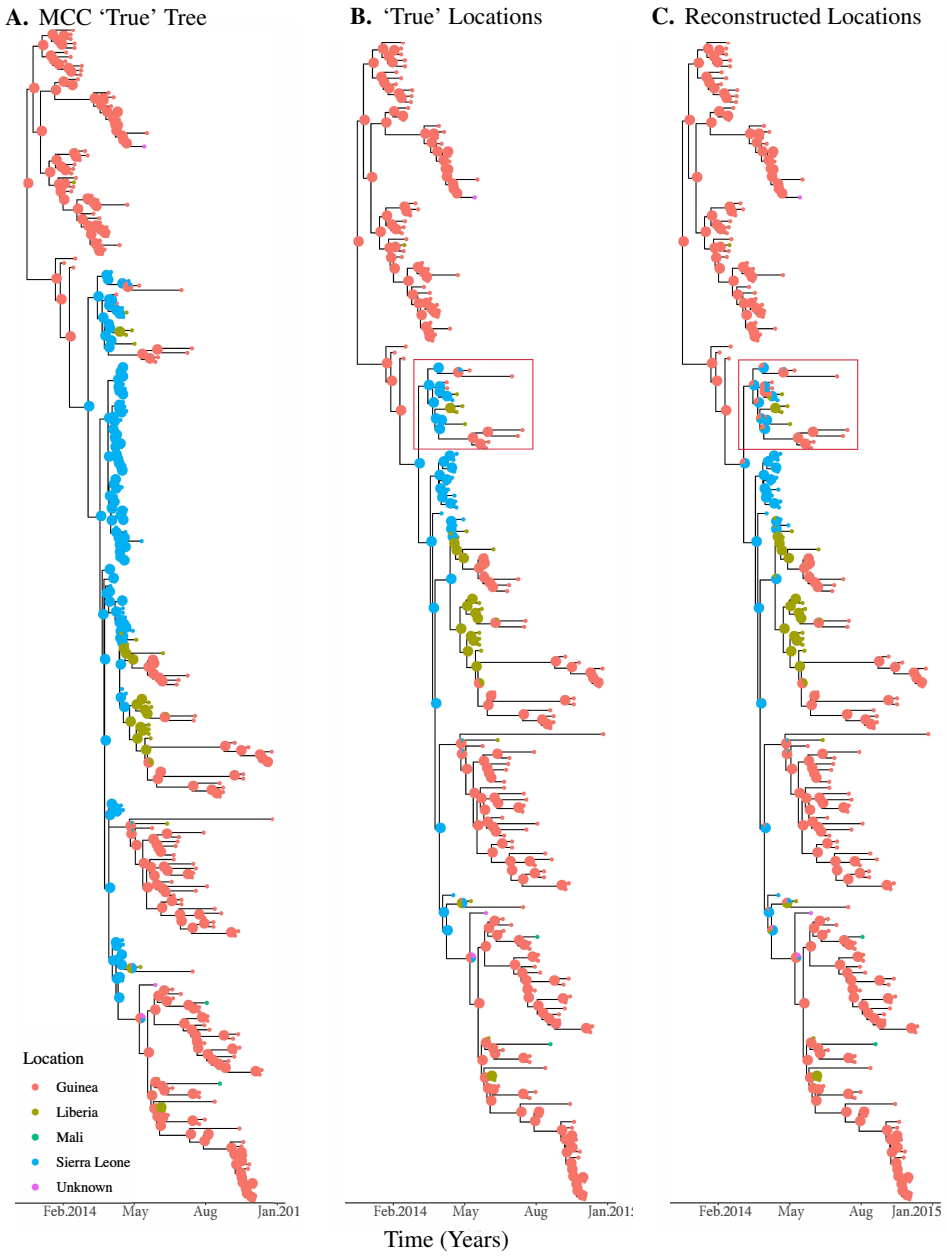

Figure 1.4: **Ancestral state reconstruction of Ebola virus trees with Sierra Leone downsampled**. Panel A: True tree with the correct internal states (Panel A). Downsampled tree was obtained by dropping 80% of Sierra Leone samples with true ancestral locations (Panel B). After dropping Sierra Leone samples, the downsampled tree with reconstructed ancestral location (Panel C). 14

considered as downpass states). If no states are shared in common ($M1_d \cap M2_d = \varnothing$), assign the union to N_d ($N_d = M1_d \cup M2_d$). The downpass state for N_d is an ambiguous state, it can either be in state $M1_d$ or $M2_d$. If child nodes share any states in common ($M1_d \cap M2_d \neq \varnothing$), assign the set of shared states to N ($N_d = M1_d \cap M2_d$).

2. We assign the downpass state as the final state for the root.

3. Next, we perform a pre-traversal algorithm. To find the uppass state for node N_u, consider the downpass state of parent node P_d and node that shared most common ancestor with N: S_d. The same rules apply. If no states are shared in common ($P_d \cap S_d = \varnothing$), assign the union to N_u ($N_u = P_d \cup S_d$). If nodes share any states in common ($P_d \cap S_d \neq \varnothing$), assign the set of shared states to N_u ($N_u = P_d \cap S_d$).

4. Finally, to determine the final state of N, consider that node's uppass state N_u and the downpass states of its two child nodes $M1_d$ and $M2_d$. Choose the state that has the majority number in all three sets. If none is the majority, it remains ambiguous. (meaning the number of state changes remains minimum regardless of this internal state.)

Figure 1.5 is an example of using the Maximum Parsimony method to reconstruct ancestral states. All the internal states are assigned as white. Notice that this is consistent with the Maximum Parsimony method: the best reconstruction is the one with the minimum amount of state change. Here there is only one state change from white to black (node D to tip B), which is the minimum state change of the phylogenetic tree. Due to the simplicity of the idea that we want to reconstruct the internal states requiring the minimum amount of state change, using Maximum Parsimony for ancestral state reconstruction can be misleading. Felsenstein used examples to show that Maximum Parsimony methods are not always statistically consistent: even with sufficient samples, the Maximum Parsimony method does not necessarily reconstruct a phylogenetic trees with high likelihood [20] as calculated via a maximum likelihood approach (see below). Figure 1.5 demonstrates a second problem with the Maximum Parsimony method. If we apply the Maximum Parsimony method to this tree, all the internal states are inferred to be white since the most parsimonious tree

is the one that has only a single state change on the branch from D to B. However, if sampling rates are different between the black and white states, or it is known that the root state is black, the Maximum Parsimony method will give misleading results. Given these limitations of Maximum Parsimony, alternative methods, such as the Maximum Likelihood method presented in the next section, are often preferred.

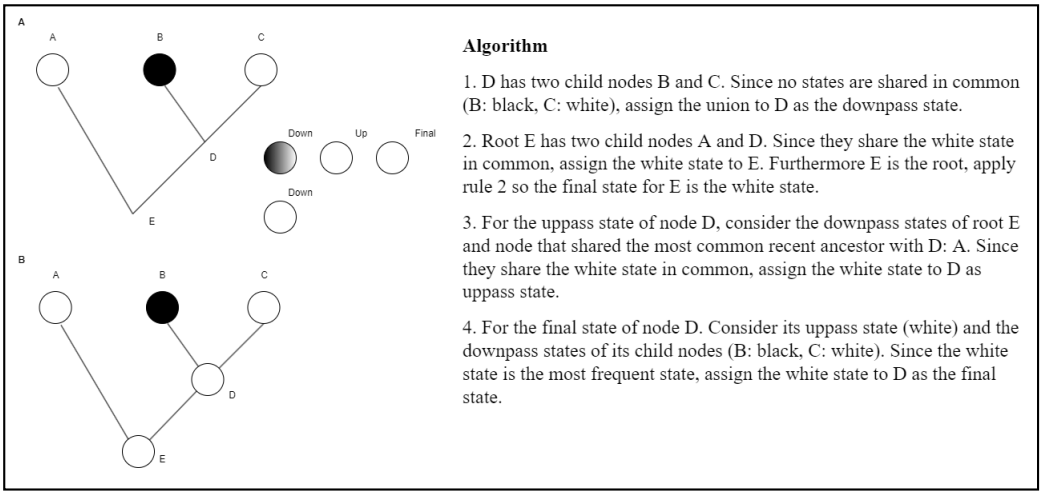

Figure 1.5: **Reconstruct ancestral states using the Maximum Parsimony method.** The final states for D and E are both white. The minimum number of state changes is one.

1.2 Maximum Likelihood

The term "state" is used in many areas of mathematics. A Continuous-time Markov Chain (CTMC) is a stochastic process in which the state of the system changes through time according to an a series of events described by the rate matrix Q. Examples of other CTMC in biologoy include molecular evolution models where the state refers to the four bases in DNA (A, T, C, G) or RNA (A, U, C, G). In the model I am focusing on in the thesis, the state often refers to discrete locations, and I model changes in location on the tree.

Maximum Likelihood methods (ML) reconstruct the internal states of phylogenetic trees to find internal states that maximize the probability of observing the tip states given the phylogenetic tree topology and a specific evolutionary model; in this case, the evolutionary model refers to the transition rates between the character states. ML methods for ances-

tral state reconstruction used character states at the tips on a fixed tree to infer ancestral character states. Maximum Likelihood methods were applied to genetic sequence evolution by Felsenstein in 1981 [21] who developed a dynamic algorithm (referred to as the Felsenstein pruning algorithm) to estimate evolutionary trees. Pagel and Hadfield proposed the first Maximum Likelihood method for ancestral state reconstruction for discrete character evolution. [2, 22, 3].

Unlike Maximum Parsimony methods for ancestral state reconstruction, Maximum Likelihood methods consider the stochastic evolutionary process, specified by the rate of evolutionary events occurring along the length of each branch in the tree. Pagel introduced a still widely used Maximum Likelihood method for discrete character states, hereafter referred to as the 'classic Maximum Likelihood method'. He used matrix exponentiation to develop a quick algorithm for estimating transition probabilities along the branches of the tree and the likelihood of ancestral states. One weakness of the method is that Pagel assumed that the topology of the tree that results from speciation and extinction were independent of the state of the lineages. However, speciation rates are usually different among species or viral variants, as mentioned above regarding the BiSSE model. Various Maximum Likelihood methods have been developed recently, aiming for better parameter estimation and ancestral state reconstruction accounting for state-dependent speciation and extinction rates. For instance, Nielsen [23] developed a Stochastic Mapping method that allowed the inference of not only ancestral states at internal nodes but also the location of state changes along the tree branches. Later Nielsen extended the idea to the mapping of morphological characters [23]. And importantly for us here, Freyman & Hohna recently extended the Stochastic Character Mapping method to allow for state-dependent speciation and extinction and for faster running time and infinite-state substitution processes [24, 13]. Freyman and Hohna's method is grounded in the State Speciation and Extinction family (SSE) of methods first introduced by Maddison et al.[12]. The model assumes that the tree is complete, ultrametric, rooted and only consists of binary states. Later they extended the idea to multi-state (MUSSE) trees with unequal sampling rates and incomplete trees. [14]

Here I will focus on the classic Maximum Likelihood method that Pagel fully describes for discrete character state reconstruction [22]. I use the Maximum Likelihood method proposed by Pagel because of its implementation in the widely used *ape* package in *R* for discrete state ancestral state reconstruction. For this same reason, I will use this classic method to quantify the utility of my new approach to compute reconstruction accuracy as we did in the study as mentioned above [9].

The classic Maximum Likelihood method can be summarized into two steps. First, we want to calculate the transition probability between nodes by considering the instantaneous rate of change between different states using the Markov chain. The second step is to calculate the likelihood of the tree given a specific set of internal nodes. The method the defines the state of each internal node as the state that maximizes the likelihood.

Markov Process

Let $P(t)$ be the transition probability matrix such that the element $P_{ij}(t)$ is the probability of going from state i to state j in time t, including the probability of visiting other states in between. Let $P(t)$ be the transition matrix such that each entry is $P_{ij}(t)$. This transition probability can be derived from a CTMC with constant instantaneous transition rates q_{ij} from state i to state j. Figure 1.6 shows a binary tree (state 0 and 1) with two tips and one root. Then the probability that a state changes from 0 to 1 over a time interval $t + dt$ is

$$P_{01}(t + dt) = P_{00}(t)q_{01}dt + P_{01}(t)(1 - q_{10})dt \tag{1.1}$$

$P_{00}(t)q_{01}dt$ is the probability that state 0 does not change in time interval t but does change to state 1 in the infinitesimal time interval dt. $P_{01}(t)(1 - q_{10})dt$ is the probability that state 0 changes in the first time interval t and does not change in the time interval dt. Rewriting equation 1.1 in matrix form:

$$
\begin{aligned}
P(t+dt) &= \begin{pmatrix} 1 - P_{01}(t+dt) & P_{01}(t+dt) \\ P_{10}(t+dt) & 1 - P_{10}(t+dt) \end{pmatrix} \\
&= \begin{pmatrix} P_{00}(t) & P_{01}(t) \\ P_{10}(t) & P_{11}(t) \end{pmatrix} \begin{pmatrix} (1-q_{01})dt & q_{01}dt \\ q_{10}dt & (1-q_{10})dt \end{pmatrix} \\
&= P(t)(I + Qdt)
\end{aligned} \tag{1.2}
$$

We can obtain a differential equation for $P(t)$ by taking the derivative:

$$
\frac{dP(t)}{dt} = \frac{P(dt+t) - P(t)}{dt} = P(t)Q \tag{1.3}
$$

We can solve the differential equation:

$$
P(t) = e^{Qt} \cdot c \tag{1.4}
$$

The initial condition for the differential equation $\frac{dP(t)}{dt}$ is $P(0) = I$ (state doesn't change when time equals to zero). Hence we obtain the transition probability between states given a certain amount of time t: $P(t) = e^{(Qt)}$.

In the simplest case, we apply CTMC on a binary state space (states 0 and 1). Then the rate matrix Q is a two by two matrix with $\alpha, \beta > 0$:

$$
Q = \begin{pmatrix} -\alpha & \alpha \\ \beta & -\beta \end{pmatrix} \tag{1.5}
$$

Since $P(t) = exp(Qt)$ is matrix exponential, we can diagonalize Q to get the expression of $P(t)$ using only α and β. Using linear algebra, we can find that $\lambda_1 = 0$ and $\lambda_2 = -\alpha - \beta$ are the eigenvalues and $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ and $\begin{pmatrix} \alpha \\ -\beta \end{pmatrix}$ are the eigenvectors, respectively. Then we can

diagonalize $Q(t)$ to get $P(t)$:

$$\begin{aligned} P(t) &= Me^{\Lambda(t)}M^{-1} \\ &= \begin{pmatrix} 1 & \alpha \\ 1 & -\beta \end{pmatrix} e^{\Lambda t} \begin{pmatrix} \frac{\beta}{\alpha+\beta} & \frac{\alpha}{\alpha+\beta} \\ \frac{1}{\alpha+\beta} & \frac{-1}{\alpha+\beta} \end{pmatrix} \\ &= \begin{pmatrix} 1 & \alpha \\ 1 & -\beta \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & e^{-(\alpha+\beta)} \end{pmatrix} \begin{pmatrix} \frac{\beta}{\alpha+\beta} & \frac{\alpha}{\alpha+\beta} \\ \frac{1}{\alpha+\beta} & \frac{-1}{\alpha+\beta} \end{pmatrix} \\ &= \begin{pmatrix} \frac{\beta}{\alpha+\beta} + \frac{\alpha}{\alpha+\beta}e^{-(\alpha+\beta)t} & \frac{\alpha}{\alpha+\beta} - \frac{\alpha}{\alpha+\beta}e^{-(\alpha+\beta)t} \\ \frac{\beta}{\alpha+\beta} - \frac{\beta}{\alpha+\beta}e^{-(\alpha+\beta)t} & \frac{\alpha}{\alpha+\beta} + \frac{\beta}{\alpha+\beta}e^{-(\alpha+\beta)t} \end{pmatrix} \end{aligned} \tag{1.6}$$

Maximum Likelihood of the ancestral states

Using above results we can calculate the tree's overall likelihood and the ancestral states' likelihood. Let T be the phylogenetic tree, $a \in \{0, 1\}$ be the ancestral states, d be the data observed at the tips $d \in \{0, 1\}$, $w(a)$ be the prior distribution of the root state (i.e. our prior expectation that the root has state a). Then the likelihood of the model m given the tree T will be the sum of the likelihood of the root being in state 0 and the root being in state 1. Expanding the sum gives:

$$\begin{aligned} L(m|T) &= \sum_{a=0}^{1} w(a)P(d|m, a) \\ &= \sum_{a=0}^{1} w(a)(P_{a0}(t)P_{a1}(t)) \\ &= w(0)(P_{00}(t)P_{01}(t)) + w(1)(P_{10}(t)P_{11}(t)) \end{aligned} \tag{1.7}$$

To find the likelihood of a particular state i ($i \in (0, 1)$) of an ancestral state a, we can assign state i to the ancestor node (here the root) and calculate the likelihood:

$$L(a = i, m|T) = w(i)(P_{i0}(t)P_{i1}(t)) \tag{1.8}$$

Rather than using the estimated rates α and β from the maximum likelihood of the model for finding the ancestral state of an internal node, we want to re-estimate the rates

α and β after fixing an internal node (or the root) to a particular state i [22]. The reason is that assigning a state i to an internal node implies a new set of rate parameters (α and β). By assigning a state to an internal node, we have additional information. Therefore, we need to re-estimate rates, and the sum of two likelihoods will not necessarily equal $L(m|T)$ in equation (1.7).

Here I discuss three different cases. (1) The simplest example where the branch lengths and prior probabilities of the root state are the same ($t_1 = t_2 = 1$ and $w(0) = w(1) = 0.5$). (2) The branch lengths are the same but the prior probabilities of the root state are different ($t_1 = t_2 = 1$ and $w(0) = \frac{1}{5}$, $w(1) = \frac{4}{5}$). (3) The branch lengths are different but the prior probabilities of the root state are the same ($t_1 = 1$, $t_2 = 3$ and $w(0) = w(1) = 0.5$).

For the simplest case, if the branch length equals 1 for both lineages ($t_1 = t_2 = 1$), we can calculate the likelihood of the model given the tree as follows. Without information about the root state, both states (0 and 1) are equally likely to be the root state. Hence $w(0) = w(1) = 0.5$ Let α and β be the transition rates described above, then:

$$\begin{aligned} L(m|T) &= w(0)(P_{00}(1)P_{01}(1)) + w(1)(P_{10}(1)P_{11}(1)) \\ &= \frac{1}{2}(\frac{\beta}{\alpha+\beta} + \frac{\alpha}{\alpha+\beta}e^{-(\alpha+\beta)})(\frac{\alpha}{\alpha+\beta} - \frac{\alpha}{\alpha+\beta}e^{-(\alpha+\beta)}) \\ &+ \frac{1}{2}(\frac{\beta}{\alpha+\beta} - \frac{\beta}{\alpha+\beta}e^{-(\alpha+\beta)})(\frac{\alpha}{\alpha+\beta} + \frac{\beta}{\alpha+\beta}e^{-(\alpha+\beta)}) \end{aligned} \tag{1.9}$$

In order to maximize the model, we need to find values of α and β such that the above equation yields the largest value. In this example, I used mathematical software (Mathematica) to find the maximum likelihood of the tree and maximum estimates $\hat{\alpha}$ and $\hat{\beta}$. $\hat{\alpha} = \hat{\beta} = 3.32$) yield the Maximum Likelihood solution $L(m|T, \hat{\alpha}, \hat{\beta}) = 0.25$.

Furthermore, we can calculate the likelihood of ancestral states using equation (1.8):

$$\begin{aligned} L(a = 0, m|T) &= \frac{1}{4} \\ L(a = 1, m|T) &= \frac{1}{4} \end{aligned} \tag{1.10}$$

Since $L(a = 0) = L(a = 1) = \frac{1}{4}$, we can conclude that the Maximum Likelihood method for ancestral state reconstruction does not favor any state in this example. The result gives

an ambiguous state. We expect this result since only two tips, and the branch lengths are equal; we cannot distinguish the root state.

For the second case, let the prior probability be different between state 0 and 1 ($w(0) = \frac{1}{5}$, $w(1) = \frac{4}{5}$), and keep the branch lengths the same. We can calculate the likelihood of the evolutionary model using equation (1.7):

$$\begin{aligned} L(m|T) &= w(0)(P_{00}(1)P_{01}(1)) + w(1)(P_{10}(1)P_{11}(1)) \\ &= \frac{1}{5}(\frac{\beta}{\alpha+\beta} + \frac{\alpha}{\alpha+\beta}e^{-(\alpha+\beta)})(\frac{\alpha}{\alpha+\beta} - \frac{\alpha}{\alpha+\beta}e^{-(\alpha+\beta)}) \\ &+ \frac{4}{5}(\frac{\beta}{\alpha+\beta} - \frac{\beta}{\alpha+\beta}e^{-(\alpha+\beta)})(\frac{\alpha}{\alpha+\beta} + \frac{\beta}{\alpha+\beta}e^{-(\alpha+\beta)}) \end{aligned} \tag{1.11}$$

We can find the maximum estimates $\hat{\alpha}$ and $\hat{\beta}$: $\hat{\alpha} = \hat{\beta} = 34$ and the likelihood of the tree $L(m|T, \hat{\alpha}, \hat{\beta}) = 0.25$. And the likelihood of ancestral states:

$$\begin{aligned} L(a = 0, m|T) &= 0.05 \\ L(a = 1, m|T) &= 0.2 \end{aligned} \tag{1.12}$$

The likelihood of the model is not changing compared to the simplest scenario. However, the likelihood of the ancestral state changes if we unbalance the prior distribution of the root state. The difference is the same as the prior difference.

For the last case, let the branch lengths be different ($t_1 = 1, t_2 = 3$), and keep the prior distribution the same. We can calculate the likelihood of the evolutionary model using equation (1.7):

$$\begin{aligned} L(m|T) &= w(0)(P_{00}(1)P_{01}(1)) + w(1)(P_{10}(1)P_{11}(1)) \\ &= \frac{1}{2}(\frac{\beta}{\alpha+\beta} + \frac{\alpha}{\alpha+\beta}e^{-(\alpha+\beta)})(\frac{\alpha}{\alpha+\beta} - \frac{\alpha}{\alpha+\beta}e^{-3(\alpha+\beta)}) \\ &+ \frac{1}{2}(\frac{\beta}{\alpha+\beta} - \frac{\beta}{\alpha+\beta}e^{-(\alpha+\beta)})(\frac{\alpha}{\alpha+\beta} + \frac{\beta}{\alpha+\beta}e^{-3(\alpha+\beta)}) \end{aligned} \tag{1.13}$$

We can find the maximum estimates $\hat{\alpha}$ and $\hat{\beta} : \hat{\alpha} = 0.587, \quad \hat{\beta} = 0.311$ and the likelihood of the tree $L(m|T, \hat{\alpha}, \hat{\beta}) = 0.256$. And the likelihood of ancestral states:

$$\begin{aligned} L(a = 0, m|T) &= 0.236 \\ L(a = 1, m|T) &= 0.125 \end{aligned} \tag{1.14}$$

The likelihood of the model is different from the other two cases. Furthermore, the likelihood of that the root is in state 0 is almost twice as high as the likelihood that the root is in state 1. The intuition is that if the branch length t_2 is larger than t_1, there is a greater chance to change along the branch. Then the state of that branch is less likely to reflect the root.

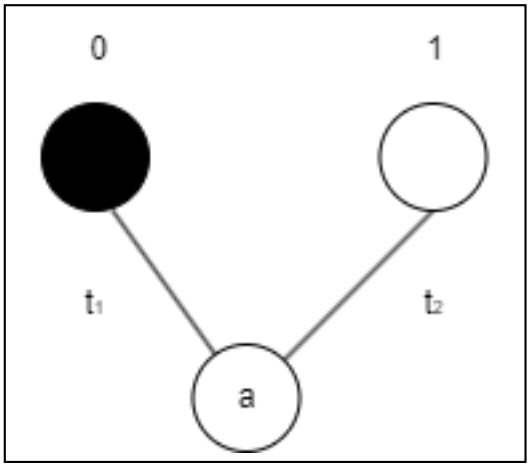

Figure 1.6: Maximum Likelihood method on a small tree. The tree consists of two tips (one from state 0 and one from state 1) and branch lengths t_1 and t_2.

1.3 Bayesian Methods for ancestral state reconstruction

Bayesian methods provide an interesting contrast to the fixed-tree models we have focused on above. These methods not only reconstruct ancestral states of a single fixed tree but average over uncertainty in the reconstructing one of the phylogenetic tree. Although Bayesian methods can be used on a fixed tree to infer ancestral states as suggested by [5], here I will

discuss start of the art method for Bayesian phylogeography method introduced by Lemey which accounts for phylogenetic uncertainty.

Lemey [25] introduced the Bayesian framework for phylogeographical inference on rooted and time-scaled phylogenies with discrete states. Suppose geography can be partitioned into a finite number of sites $\{S_k\}$(geographical locations). $x = (x_1, ..., x_N)$ at the tips of the phylogeny F, we record locations x. Similar to the Maximum Likelihood method, let $\Lambda = \{\lambda_{jk}\}$ be a K by K infinitesimal rate matrix, then we can compute the finite transition probability: $\{P_{jk}(t)\} = P(t) = e^{\Lambda t}$. GTR model often refers to the evolution of the DNA or RNA sequence data, not of the geographical locations. If we consider the GTR model (general time-reversible form) for molecular evolution [26], $k(k-1)$ transitions have non-negligible probability. However, if we consider the GTR model for locations, we would expect that many infinitesimal rates are around zero. Many zero rates would raise the issue that many degrees of freedom fit limited data, leading to high variance estimates. Not only for Λ but also inferences of the unobserved ancestral locations and the root x_{root}. Lemey [25] circumvents this sparse data issue by using BSSVS to select a parsimonious parameterization of Λ. BSSVS is a linear regression framework that uses potential predictors $x_1, ..., x_p$ and asks which is associated linearly with outcome Y.

$$Y = [x_1, ..., x_p]\beta + \epsilon \tag{1.15}$$

If β_p differs from 0, then x_p helps to predict Y. BSSVS enables simultaneously to determine which infinitesimal rates are 0 depending on the evidence of the data and efficiently infer the ancestral locations: consider a random graph in which each of the $\frac{k(k-1)}{2}$ edges either exist or not exist in G. Let δ_{jk} be the indicator that an edge exists connecting two locations. Rate Λ plays a similar role to the regression coefficient in BSSVS.

Later Lemey [27] also considered the Bayesian method on continuous traits. Other methods use structured coalescent theory. De Maio [28] treated lineage states probabilistically instead of using MCMC-based sampling. The probability of each lineage being in each state is calculated using a set of previously described differential equations, and such an approach allows the analysis of a large data set. Müller [29] derived an exact numerical solution of

the structured coalescent with discrete states to clarify the assumptions used in De Maio's approach and develop a more refined approximation to the structured coalescent. Bayesian methods for tree reconstruction and ancestral state reconstruction of viral genetics are popular given the sequences, and evolutionary model [30, 31].

1.4 Binary State Speciation and Extinction model (BiSSE)

Binary state speciation and extinction model (BiSSE) was introduced by Maddison et al. [12] in 2007 to solve the aforementioned problems in classic Maximum Likelihood methods. Character state reconstruction methods were based on a simple transition model [22], which did not consider state speciation and extinction. This question can be solved using sister clade analysis [32]. However, Barraclough and Nee [33] showed that sister clade analysis could not distinguish differential speciation from differential extinction. BiSSE solves this problem by allowing speciation and extinction rates to depend on the character state allows us to explore how particular states of interest shape diversification and biodiversity patterns. BiSSE calculates the probability of phylogenetic trees and observed states using a binary state model with six parameters: instantaneous rate of speciation and extinction (for both states 0 and 1) and instantaneous rate of character state (location) change (0 to 1 and 1 to 0). BiSSE model motivates Freyman and Hohna's work [13] on which my I approach is based.

The BiSSE model assumes that the complete rooted ultrametric tree with branch lengths and state at tips is known. The general approach for the BiSSE model is to derive a set of ODEs that calculate the probability that a lineage would evolve into the clade that is observed through time: $\frac{dD_{N,i}(t)}{dt}$, where N is the clade from a lineage and i is the state of that lineage at time t. Computing the likelihoods, one needs to derive a set of ODEs. BiSSE initializes the procedure by starting at the tips and working towards the root. When dealing with branches at internal nodes, we multiply the probabilities of the decedent nodes at state i and the instantaneous speciation rate to get the initial value for the ODEs. Continue the algorithm to the tree's root, which results in a set of k probabilities at the root representing

probabilities of observing the phylogeny given the root being in each state. The overall probability is the weights average of the k probabilities at the root: $\sum D_{R,i}p_i$.

To take care of probability of extinction, $\frac{dD_{N,i}(t)}{dt}$ must consider lineages that arise along the branch in the tree but go extinct before the present, which requiring a second set of ODEs: the probability of extinction $\frac{dE_i(t)}{dt}$.

While a lineage has state 0, the speciation rate is λ_0, the extinction rate is μ_0, and the transition rate to state 1 is q_{01}. Similarly, while a lineage has state 1, one can derive the speciation rate, extinction rate and transition rate to state 0 to be λ_1, μ_1 and q_{10}. The derivation for $D_{N,i}(t)$ and $E_i(t)$ are as follows:

$D_{N,i}(t)$ is the probability of observing lineage N at state i descending from the branch at time t. To compute the probability at an earlier time $D_{N,i}(t+\Delta t)$, consider all the events that can happen in a tiny interval Δt. We have assumed that Δt is small, and it is implausible that two or more events could occur in the interval. There are four possible events: (1) nothing happens; (2) a transition event from state i to state j occurs; (3) a speciation event occurs, the right descendant goes extinct before the present; (4) a speciation event occurs, left descendant goes extinct before the present. We can compute the difference equation for $D_{N,i}(t+\Delta t)$:

$$\begin{aligned} D_{N,i}(t+\Delta t) \approx\ & D_{N,i}(t) \\ & + \underbrace{(1-\mu_i\Delta t)(1-q_{ij}\Delta t)(1-\lambda_i\Delta t)D_{N,i}(t)}_{\text{nothing happens}} \\ & + \underbrace{(1-\mu_i\Delta t)(q_{ij}\Delta t)(1-\lambda_i\Delta t)D_{N,j}(t)}_{\text{state change}} \\ & + \underbrace{(\mu_i\Delta t)(1-q_{ij}\Delta t)(1-\lambda_i\Delta t)0}_{\text{death event}} \\ & + \underbrace{2(1-\mu_i\Delta t)(1-qij\Delta t)(\lambda_i\Delta t)E_i(t)D_{N,i}(t)}_{\text{speciation event}} \end{aligned} \tag{1.16}$$

By the definition of a derivative, we will obtain the following differential equation for D_{Ni}:

$$\frac{d}{dt}D_{N,i}(t) = -(\lambda+\mu+q_{ij})D_{N,i}(t) + 2\lambda_i E_i(t)D_{N,i}(t) + q_{ij}D_{N,j}(t) \tag{1.17}$$

We can derive $E_i(t)$ similarly, considering the four distinct events that can occur between time t and $t + \Delta t$: (1) The lineage goes extinct. (2) There is no state change nor speciation event, resulting in a single lineage going extinct before the present. (3) There is a state change but no speciation event, resulting in a single lineage going extinct before the present. (4) No state change occurs, but speciation occurs, resulting in both lineages going extinct before the present. We can compute the difference equation for $E_i(t + \Delta t)$:

$$\begin{aligned} E_i(t+\Delta t) \approx E_i(t) & \\ & + \underbrace{\mu_i \Delta t}_{\text{extinction event}} \\ & + \underbrace{[(1-\mu_i\Delta t)(1-q_{ij}\Delta t)(1-\lambda_i\Delta t)E_i(t)}_{\text{no birth event nor state change, lineage goes extinct}} \\ & + \underbrace{(1-\mu_i\Delta t)(q_{ij}\Delta t)(1-\lambda_0\Delta t)E_j(t)}_{\text{state change, lineage goes extinct}} \\ & + \underbrace{(1-\mu_i\Delta t)(1-q_{01}\Delta t)(\lambda_0\Delta t)E_i(t)^2]}_{\text{speciation event, both lineages go extinct}} \end{aligned} \tag{1.18}$$

Hence we will obtain the differential equation for E_i:

$$\frac{d}{dt}E_i(t) = \mu - (\lambda + \mu + q_{ij})E_i(t) + \lambda E_i^2(t) + q_{ij}E_j(t) \tag{1.19}$$

The initial condition for $E_i(t)$ is the probability of state i being extinct at present:

$$E_i(0) = 0 \tag{1.20}$$

The initial condition for $D_{BNi}(t)$ depends on the time t and its position in the phylogenetic tree:

$$D_{N,i}(t) = \begin{cases} 1, \text{if } t = 0 \\ \\ \lambda_i D_{M,i}(t) D_{N,i}(t), \text{if } t \neq 0 \text{ and its an internal node} \end{cases} \tag{1.21}$$

Once we find all the D values at the root, we can calculate the overall likelihood of the tree: $L(T, S|M) = \sum D_{R,i} p_i$, where p_i is the prior probability of the root state.

1.5 Stochastic Character Mapping Method

Unlike most ancestral state reconstruction methods that only reconstruct ancestral character states at internal nodes, Stochastic Character Mapping infers the full evolutionary history of ancestral states along the branches of a phylogeny, including the location of state changes as well as the states at the internal nodes. Nielsen [23] first introduced the Stochastic Character Mapping method using a rejection sampling approach. However, the method has two significant limitations: first, the rejection sampling approach is inefficient for many states, and second, unlike the BiSSE model, it assumes that speciation and extinction are independent of the character state. Freyman & Hohna [13] extend the idea in Nielsen, developing an algorithm that does not rely on rejection sampling approaches and allows for state-dependent diversification.

Nielsen's standard Stochastic Character Mapping consists of three steps: First, the probability of the character states at each node is calculated from tip to root using Felsenstein's pruning algorithm [21]. The calculation of these state probabilities involves transition probabilities which are computed using matrix exponentiation described above. Following the tip to root (post-traversal) calculation, the states at each internal node are then sampled from root to tip and the states chosen are used to correct the probability at subsequent nodes analogous to the pre-traversal algorithm described for maximum parsimony above. Finally, given the states at all internal nodes, character histories are simulated using rejection sampling methods for each tree branch. Consider a lineage that consists of two nodes: an ancestral node assigned to state A and a descendent node assigned to state B. We can simply simulate a realization of the CTMC chain defined by the transition model from the ancestral state A. A new simulation is performed if the final state is not state B. The simulation is repeated until a path of the chain that last visits state B is found. The simulation is completed by estimating the waiting time between each transition. If the total simulation time is larger than the branch length of the lineage, the simulation is complete, and we get all the transition histories along the branch. Otherwise, a new transition is mapped on the tree.

Freyman & Hohna extended this Stochastic Character Mapping algorithm as follows: they begin similarly by using post-traversal and pre-traversal algorithms to calculate ancestral states. However, they discretized the branches of the tree into small time intervals to find the entire transition history avoiding the rejection sampling algorithm.

To derive a set of differential equations backwards in time, Freyman & Hohna generalized the model by allowing cladogenetic events, where descendent lineages can inherit different states (i.e., a state transition occurs at the moment of speciation). Similar to BiSSE, we want to calculate two differential equations: one for the function $D_{N,i}(t)$ and the other for $E_i(t)$. Where $D_{N,i}(t)$ is the probability that a lineage in state i at time t evolves into the observed clade N, and $E_i(t)$ is the probability that a lineage in state i at time t goes extinct before being observed or sampled.

For $D_{N,i}(t)$ backward in time, there are four possible events: (1) nothing happens; (2) a transition event from state i to state j occurs; (3) a speciation event occurs, right descendant goes extinct before the present; (4) a speciation event occurs, left descendant goes extinct before the present. We can compute the difference equation for $D_{N,i}(t + \Delta t)$:

$$
\begin{aligned}
D_{N,i}(t+\Delta t) = D_{N,i}(t) & \\
& + [\underbrace{-(\sum_j \sum_k \lambda_{ijk} + \sum_{j \neq i} Q_{ij} + \mu_i) D_{N,i}(t)}_{\text{nothing happens}} \\
& + \underbrace{\sum_{j \neq i} Q_{ij} D_{N,j}(t)}_{\text{state change}} \\
& + \underbrace{\sum_j \sum_k \lambda_{ijk} (D_{N,k}(t) E_j(t) + D_{N,j}(t) E_k(t))]\Delta t}_{\text{speciation event, one of the lineages go extinct}} \\
& + O(\Delta t^2)
\end{aligned}
\tag{1.22}
$$

Similarly, we can consider five possible extinction events: (1) The lineage goes extinct. (2) There is no state change nor speciation event, resulting in a single lineage going extinct before the present. (3) There is a state change but no speciation event, resulting in a single lineage going extinct before the present. (4) There is a speciation event, giving birth to

a left descendent lineage in state j and a right descendent lineage in state k, and both lineages go extinct before the present. (5) There is a speciation event, giving birth to a left descendent lineage in state k and a right descendent lineage in state j, and both lineages go extinct before the present. The difference from the BiSSE model is the extension by allowing cladogenetic events that result in events four and five being treated distinctively. We can compute the difference equation for $E_i(t + \Delta t)$:

$$\begin{aligned} E_i(t+\Delta t) = E_i(t) & \\ & + \underbrace{[\mu_i \Delta t}_{\text{lineage goes extinct within the interval } \Delta t} \\ & - \underbrace{((\sum_j \sum_k \lambda_{ijk} + \sum_{j \neq i} Q_{ij} + \mu_i) E_i(t)) \Delta t}_{\text{lineage eventually goes extinct}} \\ & + \underbrace{\sum_{j \neq i} Q_{ij} E_j(t) \Delta t}_{\text{state change and eventually goes extinct}} \\ & + \underbrace{\sum_j \sum_k \lambda_{ijk} E_j(t) E_k(t) \Delta t]}_{\text{speciation event and eventually all lineages go extinct}} \\ & + O(\Delta t^2) \end{aligned} \tag{1.23}$$

The initial conditions are the same as described in the BiSSE model.

For the set of differential equations forward in time, we want to compute $D(t - \Delta t)$ and $E(t - \Delta t)$. For $D_{N,j}(t)$ forward in time, there are four possible events: (1) nothing happens. (2) with probability $D_{N,j}(t)$, the lineage was in state j, and then a state changed to state i. (3) with probability $D_{N,j}(t)$ the lineage was in state j, and then speciation event occurs, giving birth to a left descendent lineage in state i and a right descendent lineage in state k but goes extinct before the present. (4) with probability $D_{N,j}(t)$ the lineage was in state j and then speciation event occurs, giving birth to a left descendent lineage in state k and a right descendent lineage in state i but goes extinct before the present. We can write the

difference equation for $D_k(t - \Delta t)$:

$$\begin{aligned} D_{N,i}(t - \Delta t) = D_{N,i}(t) + \\ [-(\sum_j \sum_k \lambda_{ijk} + \sum_{j \neq i} Q_{ij} + \mu_i) D_{N,i}(t) \\ + \sum_{j \neq i} Q_{ji} D_{N,j}(t) \\ + \sum_j \sum_k \lambda_{jik} D_{N,j}(t) E_k(t - \Delta t) \\ + \sum_j \sum_k \lambda_{jki} D_{N,j}(t) E_k(t - \Delta t)] \Delta t \\ + O(\Delta t^2) \end{aligned} \tag{1.24}$$

However, we don't know $E_k(t - \Delta t)$, so instead we want to approximate the equation by using $E_k(t)$ for $E_k(t + \Delta t)$, this will give the following approximation:

$$\begin{aligned} D_{N,i}(t - \Delta t) \approx D_{N,i}(t) + \\ [-(\sum_j \sum_k \lambda_{ijk} + \sum_{j \neq i} Q_{ij} + \mu_i) D_{N,i}(t) \\ + \sum_{j \neq i} Q_{ji} D_{N,j}(t) \\ + \sum_j \sum_k \lambda_{jik} D_{N,j}(t) E_k(t) \\ + \sum_j \sum_k \lambda_{jki} D_{N,j}(t) E_k(t)] \Delta t \end{aligned} \tag{1.25}$$

For $E_i(t)$ forward in time again, there are five possible events: (1) lineage goes extinct. (2) nothing happens in the interval Δt, but the lineage goes extinct before the present. (3) state changes but goes extinct before the present. (4) lineage speciates, giving birth to the left descendent lineage in state k and the right descendent lineage in state j, and both lineages go extinct before the present. (5) lineage speciates, giving birth to the left descendent lineage in state j and the right descendent lineage in state k, and both lineages

go extinct before the present. We can write the difference equation for $E_i(t-\Delta t)$:

$$\begin{aligned} E_i(t-\Delta t) = E_i(t) - \\ [\mu_i \\ - (\sum_j \sum_k \lambda_{ijk} + \sum_{j\neq i} Q_{ij} + \mu_i) E_i(t) \\ + \sum_{j\neq i} Q_{ij} E_j(t-\Delta t) \\ + \sum_j \sum_k \lambda_{ijk} E_j(t-\Delta t) E_k(t-\Delta t)]\Delta t \end{aligned} \tag{1.26}$$

Again we do not know $E_j(t-\Delta t)$, we approximate $E_j(t-\Delta t)$ by $E_j(t)$, this will give the following approximation:

$$\begin{aligned} E_i(t-\Delta t) = E_i(t) - \\ [\mu_i \\ - (\sum_j \sum_k \lambda_{ijk} + \sum_{j\neq i} Q_{ij} + \mu_i) E_i(t) \\ + \sum_{j\neq i} Q_{ij} E_j(t) \\ + \sum_j \sum_k \lambda_{ijk} E_j(t) E_k(t)]\Delta t \end{aligned} \tag{1.27}$$

In order to avoid using the rejection sampling algorithm, the branches are discretized into small time intervals. Using the post-traversal algorithm, we can calculate $D_{N,i}$ in each small time interval. Assign $L_N(t) = D_N(t)$ as the backward states probabilities for this time interval. Once the state probabilities are calculated along all time intervals and root, we calculate the root states using root frequencies: $p_i = \frac{\pi_i D_{R,i}(t)}{\sum \pi_i D_{R,i}(t)}$, where π_i is the prior distribution of the character state of the root. Then states are drawn from the root to the tip for every short time interval saved in the first step. In order to calculate the ancestral state characters for each time interval, we multiply the probability of $D_N(t)$ from the forward algorithm by $L_N(t)$: $D_N(t) \cdot L_N(t)$. Since we are calculating the ancestral states on every

short time interval to find the full transition history, the rejection sampling algorithm is unnecessary.

I have summarized three major categories of ancestral state reconstruction method (Maximum Parsimony, Maximum Likelihood and Bayesian) and two methods that are important to my approach accounting for sampling bias (BiSSE and Stochastic Character Mapping). In the following chapters I will introduce ancestral state reconstruction method that accounting for sampling bias.

Chapter 2

Methods

2.1 Inference accounting for sampling bias

I begin by generalizing the ancestral state reconstruction method proposed by Freyman and Hohna [13] to allow for the sampling of lineages through time and state-dependent sampling. I consider a rooted binary phylogenetic tree with the known tree topology and character states at the tips. I assume binary states, but extending the method to multi-character states is straightforward. As we are primarily interested in the effect of state-dependent sampling (i.e., sampling bias) on ancestral state reconstruction, we will focus on the case of a neutral binary character. Specifically, the character has no impact on the speciation or extinction rate. The method provided is easily extended to the non-neutral case, as illustrated by the original development of the model [13].

I assume that the tree results from a birth-death Markov process with sampling through time, as in the case of sampling pathogen sequences during an ongoing epidemic. Lineages speciate at rate λ, go extinct rate μ and sampled at maximal rate ψ. To impose state-dependent sampling, let f_i be the probability of sampling at state i such that lineages of type i are sampled at rate $f_i * \psi$ ((note that $\sum_i f_i = 1$)). Finally, lineages transition between states at a given rate such that q_{ij} is the transition rate from state i to state j. I extend the method proposed by [13] by considering sampling rate ψ and the probability of sampling at state i.

As with the non-neutral characters considered by Freyman and Hohna, state-dependent sampling will impact both the state of the sampled nodes and the tree's topology. I first

perform a post-traversal algorithm to account for tip states and topology on ancestral state reconstruction. Moving from the tips to the root, we calculate the probability of observing the descendants of each internal node. Once reaching the root, we perform a pre-traversal algorithm, moving from the root towards the tips and calculating the probability of observing each internal node given its ancestors.

Let $D_{BNi}(t)$ be the probability that a lineage N in state i at time t gives rise to the observed descendants. Furthermore, let E_i be the probability that a lineage of type i at time t has no observed decedents between time t and the present day. In contrast, measuring t forward in time from the root ($t = 0$) to the present day ($t = T$), let D_{FNi} be the probability that a lineage N in state i at time t arose from its ancestors. Finally, we represent the ancestral state reconstruction of node N with the probability A_{Ni}, the probability that node N was in state i given both the observed tip states and tree topology.

Backwards in time differential equation

To derive the initial value problem for D_{BNi}, there are five different events (Figure 2.1, top panel. (1) There is no state change or speciation. (2) There is a stage change but no speciation. (3) No state change occurs, but speciation occurs, giving birth to the left and right lineage. Only the left lineage survives or gets sampled. (4) No state change occurs, but speciation occurs, giving birth to the left and right lineage. Only the right lineage survives or gets sampled. (5) The lineage gets sampled. Then the sum of these five probabilities will equal to D_{BNi}:

$$\begin{aligned} D_{BNi}(t+\Delta t) \approx\ & D_{BNi}(t) \\ & + (1-\mu\Delta t)(1-\psi\Delta t)\,[\underbrace{(1-q_{ij}\Delta t)(1-\lambda\Delta t)D_{BNi}(t)}_{\text{nothing happens}} \\ & + \underbrace{(q_{ij}\Delta t)(1-\lambda\Delta t)D_{BNj}(t)}_{\text{state change}} \\ & + \quad \underbrace{2(1-q_{ij}\Delta t)(\lambda\Delta t)E_i(t)D_{BNi}]}_{\text{speciation event, one of the lineages go extinct}} \\ & + (\mu\Delta t)\cdot 0 + (\psi\Delta t)\cdot 0 + O(\Delta t^2) \end{aligned} \tag{2.1}$$

By the definition of a derivative, we will obtain the following differential equation for D_{BNi}:

$$\frac{d}{dt}D_{BNi}(t) = -(\lambda + \mu + \psi + q_{ij})D_{BNi}(t) + 2\lambda E_i D_{BNi}(t) + q_{ij}D_{BNj}$$

We can derive the initial value problem for the probability E_i similarly, considering the five different events (Figure 2.1, bottom panel) that can occur between time t and $t + \Delta t$: (1) The lineage goes extinct. (2) There is no state change or speciation event. (3) There is a state change but no speciation event. (4) There is no state change, but speciation occurs. (5) The lineage goes extinct, not being observed. Then:

$$\begin{aligned} E_i(t + \Delta t) \approx\ & E_i(t) \\ & + \underbrace{\mu\Delta t}_{\text{lineage goes extinct within the interval } \Delta t} \\ & + \underbrace{(1 - \mu\Delta t)(1 - q_{ij}\Delta t)(1 - \lambda\Delta t)(1 - \psi\Delta t)E_i(t)}_{\text{lineage eventually goes extinct}} \\ & + \underbrace{(1 - \mu\Delta t)(q_{ij}\Delta t)(1 - \lambda\Delta t)(1 - \psi\Delta t)E_j(t)}_{\text{state change}} \\ & + \underbrace{(1 - \mu\Delta t)(1 - q_{ij}\Delta t)(\lambda\Delta t)(1 - \psi\Delta t)E_i^2(t)}_{\text{speciation event, both lineages go extinct}} \\ & + O(\Delta t^2) \end{aligned} \tag{2.2}$$

Hence we will obtain the differential equation for E_i:

$$\frac{d}{dt}E_i(t) = \mu - (\lambda + \mu + \psi + q_{ij})E_i(t) + \lambda E_i^2(t) + q_{ij}E_j(t)$$

Initial conditions

The initial condition for $E_i(t)$ is the probability an extant lineage is unsampled (since we are not forcing to sample all the lineage or any percentage of the lineage at present day):

$$E_i(0) = 1$$

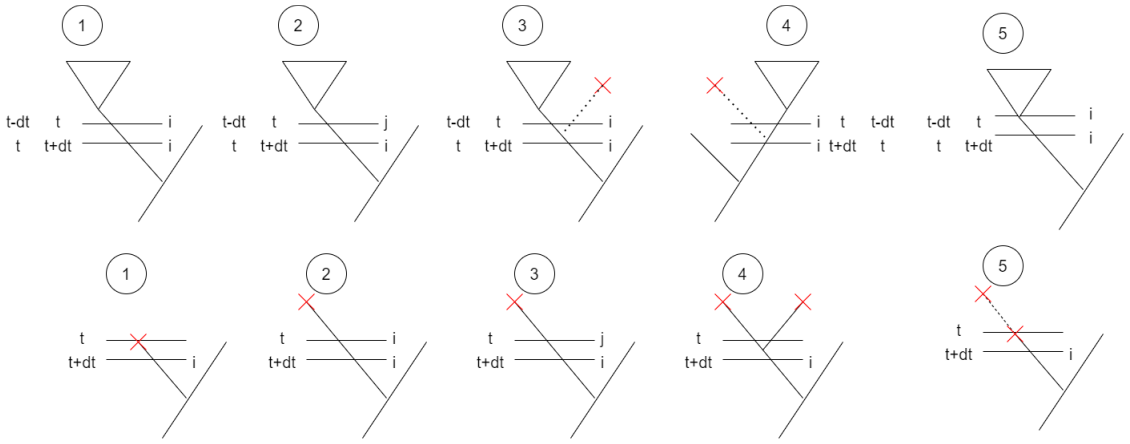

Figure 2.1: **Different events that can occur in a short time interval** Δt Top panel: speciation events. (1) There is no state change or speciation. (2) There is a stage change but no speciation. (3) No state change, but speciation occurs, giving birth to left and right lineages. Only left the lineage survives or gets sampled. (4) No state change, but speciation occurs, giving birth to left and right lineages. Only the right lineage survives or gets sampled. (5) The lineage goes extinct, with zero probability of being observed. Bottom panel: extinction events. (1) The lineage goes extinct. (2) There is no state change or speciation event. (3) There is a state change but no speciation event. (4) There is no state change, but speciation occurs. (5) lineage goes extinct, not being observed.

The initial condition for $D_{BNi}(t)$ depends on the time t and its position in the phylogenetic tree:

$$D_{BNi}(t) = \begin{cases} \psi f_i, \text{if its a tip} \\ \lambda D_{BMi}(t) D_{BSi}(t), \text{if its an internal node} \end{cases} \tag{2.3}$$

Where M and S are the descendants of internal node N.

Forwards in time differential equation

We can similarly derive the forward-in-time differential equations. We are considering the events between time t and time $t + \Delta t$, where t is measured from the roots to the tips.

$$\begin{aligned} D_{FNi}(t - \Delta t) = D_{FNi}(t) \\ - [(-\Delta t(\lambda + \mu + \psi + q_{ij}) D_{FNi}(t) \\ + 2\lambda \Delta t D_{FNi}(t) E_i(t) + q_{ij} \Delta t D_{FNj}(t) + \mu \Delta t \cdot 0 + \psi \Delta t \cdot 0 + O(\Delta t^2))] \end{aligned} \tag{2.4}$$

Hence we will obtain the differential equation for $D_{FNi}(t)$:

$$\frac{d}{dt}D_{FNi}(t) = -(-(\lambda + \mu + \psi + q_{ij})D_{FNi}(t) + 2\lambda D_{FNi}(t)E_i(t) + q_{ij}D_{FNj}(t))$$

The initial condition for the forward differential equation is the ancestral probability of state i:

$$D_{FNi}(t) = A_{Ni}$$

The ancestral state reconstruction probability of the root A_{Rooti} for each state i is weighted by the prior probability of state i and D_{BNi}:

$$A_{Rooti} = \frac{D_{BNi}\pi_i}{\sum_j D_{BNi}\pi_j}$$

The ancestral state reconstruction probability of state i for internal node N is the forward probability D_{FNi} times backward probability D_{BNi}:

$$A_{Ni} = D_{BNi} \cdot D_{FNi}$$

The reconstruction ancestral state i is the one with the highest A_{Ni} value ($max_i A_{Ni}$). However, there are two computational problems when I try to implement my algorithm. First, I use a recursive algorithm to find backward probabilities D_{BNi}. However, if there are more than twenty tips, my recursive algorithm takes too long to execute. The long running time is because of finding D_{BNi} for an internal state N. We need D_{BMi} and D_{BSi} from its descendent states. The recursion re-iterates D_{BMi} and D_{BSi} again, which has already been calculated. I implemented this way because, in order to find D_{BNi} if N is not a tip, we need to know D_{BMi} and D_{BSi} for the initial condition. However, it is hard to track these values. Therefore as the recursive algorithm moves along the tree to the root, the running time for finding a single D_{BNi} gets longer and longer. Second, tracking the initial condition for the forward differential equation is difficult. Since the initial condition for the forwards in time equation D_{FNi} is the probability of the ancestral state A_{Mi}, my current algorithm cannot track which internal state M is the ancestor of the state N that I am trying to calculate

D_{FNi}. These two problems cause me not to be able to generate more trees with more tips than the current examples to make general conclusions.

Here I test my approach using simulated phylogenetic trees. The phylogenetic trees are generated using the birth-death-sampling model. Once the proper tree is generated, the actual internal node states are masked, so only the typologies and tips of the trees are used. The accuracy of my method and the classic Maximum Likelihood method is calculated.

2.2 Simulation Approach

We wrote a birth-death-sampling simulation tool, a continuous-time Markov process that describes the viral transmission in a population through time. We are simulating a birth-death-sampling process with two locations (location 0 and 1), and there is a transition rate ($q_{01} = q_{10}$) from one location to another. Let N be the number of living individuals at the current time. A birth event increases the number of living individuals by one ($N \to N + 1$) at rate λ. A death event decreases the number of living individuals by one ($N \to N - 1$) at the rate μ. A sampling event at rate ψ does not change the number of living individuals but indicates at what time t the individual gets sampled. This node will appear in the phylogeny. We will record the location and branch length of the root whenever an individual is sampled. To emphasize sampling bias, in the simulation, I assume that transition rates, birth rates and death rates are the same in both locations ($q_{01} = q_{10}, \lambda_0 = \lambda_1, \mu_0 = \mu_1$). The sampling rates are different between locations ($\psi_0 \neq \psi_1$). The simulation model will generate a shared distance matrix. If we generate k tips after a simulation, then the shared distance matrix M is a k by k symmetric matrix where the diagonal entry M_{ii} is the distance from tip i to the root, and the off-diagonal entry M_{ij} is the common branch length shared with tip i and tip j. Hence it is the distance from their most common ancestor (MRCA) to the root. Using the shared distance matrix, we can draw the phylogenetic tree by converting the matrix to Newick format.

Newick format represents a tree using branch lengths which do not measure how far apart between tips, but it is easier to visualize and plot the tree than the shared distance matrix. The Maximum Likelihood method in the *ape* package in *R* can only use Newick format

trees as inputs, whereas my approach requires using shared distance matrices as inputs. Since my method was originally developed on the shared distance matrix. An encoding is needed to convert the distance matrix format to the Newick format in order to compare my method to the classic Maximum Likelihood method. I implemented a recursive algorithm to convert the format. (1) If dim(M) = 1, we can easily convert to the Newick format (by adding a ":"). (2) If dim(M) = 2, this indicates that these two nodes are coming from the same branch (have the same MRCA), we can group them in Newick format: "(:,:)". (3) If dim(M) > 2, break the shared distance matrix into two smaller matrices (M_1, M_2) by subtracting the minimum value of the matrix. At least one row (column) will be zero with all entries except the diagonal element. Then apply the algorithm to both smaller matrices and also add group them together by adding "$(:M_1)$,$:M_2)$".

2.3 Accuracy Calculation

To quantitatively compare my approach to the classic Maximum Likelihood method, I compute the absolute accuracy and relative accuracy of the ancestral state reconstruction using each method as described in [9]. Given the true tree T, let c_i be the location of internal node i in T, where $c_i = 1$ if node i is in location 0 and $c_i = 0$ if its in location 1. Let $\hat{c}_i$ be the likelihood of node i being in location 0 as inferred by the ancestral state reconstruction method (my approach and the classic Maximum Likelihood method). I define the absolute accuracy of a method for node i as $a_i = 1 - |c_i - \hat{c}_i|$ and a_μ as the average over all interval nodes a_i. Whereas the absolute accuracy compares the reconstructed states to the 'true' values, the relative accuracy compares the absolute accuracy of the reconstruction to an "expected accuracy" under a null model. The intuition of this null model is that if the sampling rates are known and they are constant through time, then we would expect the likelihood of character states at each internal node should be close to the sampling rates. We are not using other methods (eg. Maximum Parsimony) as null models since they do not take sampling rates into account. Suppose that the fraction of the tips that are in state 0 is given by f_0 then the expected accuracy of node i under the null model is $e_i = 1 - |\hat{c}_i - f_0|$

and e_μ is the expected accuracy averaged over all internal nodes of the reconstructed tree. I then compute the relative mean reconstruction accuracy of the tree as $r_\mu = a_\mu - e_\mu$.

Chapter 3

Results

I will examine three scenarios for my approach and compare them to the classic Maximum Likelihood method: (1) A conceptual example where there are more human samples than bats. (2) Two simulation examples. One has similar sampling rates in both locations, and one has different sampling rates in different locations. (3) Simulate multiple trees and summarize absolute and relative accuracy as a dot plot.

Figure 3.1 is a conceptual example that I constructed to demonstrate the difference between my approach and the classic Maximum Likelihood method. The tree consists of four tips; I assumed three samples were from humans and one from bats and try to infer ancestral character states using my method and the classic Maximum Likelihood method. This conceptual example does not take branch lengths into account, and that depending on the sampling and evolutionary rates within humans vs bats, the branch length would be different in examples from real data. For the parameters described in the method, I assumed that the probability of sampling humans is twice as high as the probability of sampling bats ($f_{human} = 2f_{bat}$), while other parameters were the same. My method (Figure 3.1, left panel) infers that the root state is bats, and bats transmit the Ebola virus to humans. In contrast, the classic Maximum Likelihood method infers that all internal nodes are humans, and the inaccurate inferences lead to a false conclusion that humans transmit the Ebola virus to bats.

In addition to the conceptual example, I tested my approach using simulated trees described in the method. Figure 3.2 and 3.3 are the ancestral state reconstructions between my approach and the classic Maximum Likelihood method with similar and different sampling

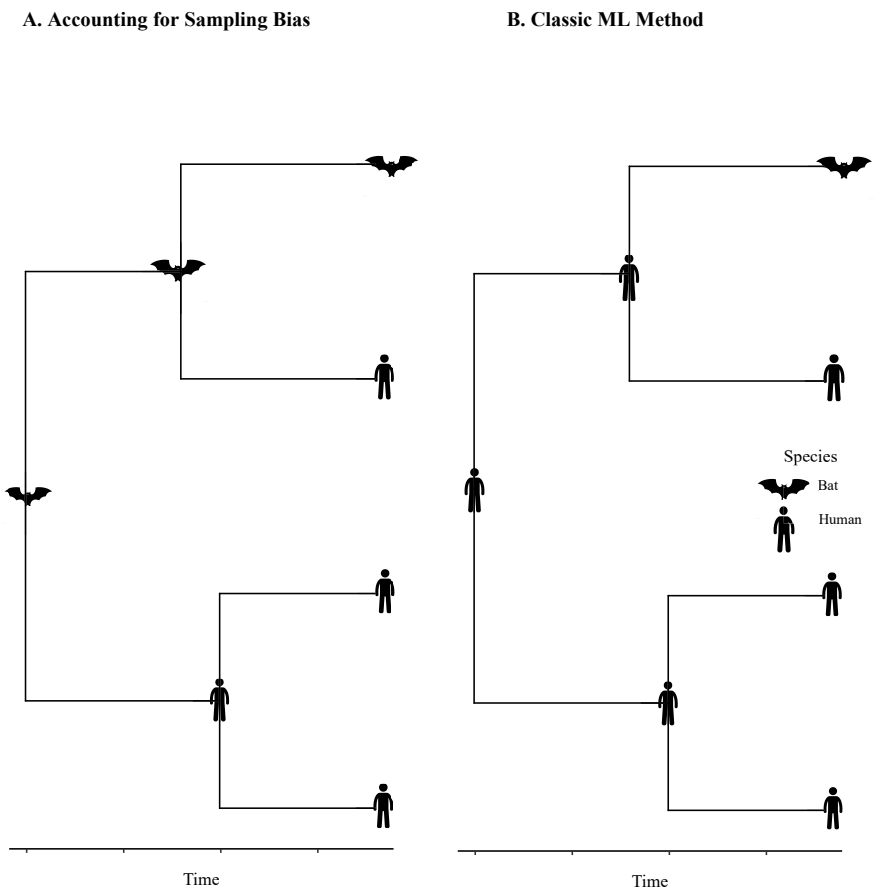

Figure 3.1: **Ancestral state reconstruction of Ebolavirus tree using my approach and Maximum Likelihood** Left panel: ASR accounting for sampling bias. Right panel: ASR using Maximum Likelihood.

rates, respectively. In Figure 3.2, parameters are the same except for the maximal sampling rate ψ_i and state-dependent sampling rate f_i. In specific: $\psi_2 = 1.2 \cdot \psi_1$ and $f_2 = 1.2 \cdot f_1$ (hence location 2 samples 1.2 times more than location 1). My approach and the classic Maximum Likelihood method give similar inferences to the true tree on the left panel.

Furthermore, both methods "correctly" reconstruct all the internal states (with higher than 50% probability). However, my method infers ancestral states more accurately than the classical method. For instance, my method predicts that the probability of the root state at location 0 is 56%, whereas the classical method suggests the probability is 52%.

The classic Maximum Likelihood approach's absolute and relative accuracy are 0.9289 and 0.3579, respectively. My approach's absolute and relative accuracy are 0.9389 and 0.3797, respectively. My method has higher absolute and relative accuracy than the Maximum Likelihood approach.

Figure 3.3 demonstrates that my method has higher reconstruction accuracy than the classic Maximum Likelihood method if the sampling rates among locations are different. Similar to Figure 3.2, parameters are the same except for the maximal sampling rate ψ_i and state-dependent sampling rate f_i. In specific: $\psi_2 = 2 \cdot \psi_1$ and $f_2 = 2 \cdot f_1$ (hence location 2 samples 2 times more than location 1). In Figure 3.3, the classic Maximum Likelihood method incorrectly infers two internal node states as location 0, but they are location 1, whereas my approach correctly reconstructs these internal nodes. The classic Maximum Likelihood approach's absolute and relative accuracy are 0.71109 and 0.04, respectively. My approach's absolute and relative accuracy are 0.9016 and 0.3032, respectively. This suggests that my method has higher reconstruction accuracy when we consider sampling rate than the classic Maximum Likelihood method.

In addition to examples shown in Figure 3.2 and 3.3, six additional trees (refer in Appendix) are generated under two different schemes using simulation to compute absolute and relative reconstruction accuracy: three trees have similar sampling rates in both locations, and three trees have different sampling rates in both locations. Similarly to the previous simulated trees, parameters are the same except for the maximal sampling rate ψ_i and state-dependent sampling rate f_i. For similar sampling rate in both locations, $\psi_2 = 1.2 \cdot \psi_1$ and $f_2 = 1.2 \cdot f_1$. For different sampling rate in both locations, $\psi_2 = 1.7 \cdot \psi_1$ and $f_2 = 1.7 \cdot f_1$. The absolute and relative accuracy are calculated and plotted in Figure 3.4. The mean values for both absolute and relative accuracy are higher for my approach than the classic Maximum Likelihood method. For the absolute accuracy on the left panel in Figure 3.4, the absolute accuracy rate for similar sampling rates is similar for both schemes, whereas, for different sampling rates, they are quite different. The mean accuracy rate for the classic Maximum Likelihood approach is around 0.65, and for my approach is around 0.75. On the other hand, the relative accuracy on the right panel has different mean relative accuracy between

my approach and the classic approach on both similar and different sampling rates. For similar sampling rates, the mean relative accuracy for my approach is 0.4 and for the classic approach is 0.1. For different sampling rates, the mean relative accuracy for my approach is 0.3, whereas for the classic approach is near 0.

My approach has higher absolute and relative accuracy in most trees than the classic Maximum Likelihood method. However, one tree from a similar sampling rate scheme has lower absolute and relative accuracy using my approach than the classic method.

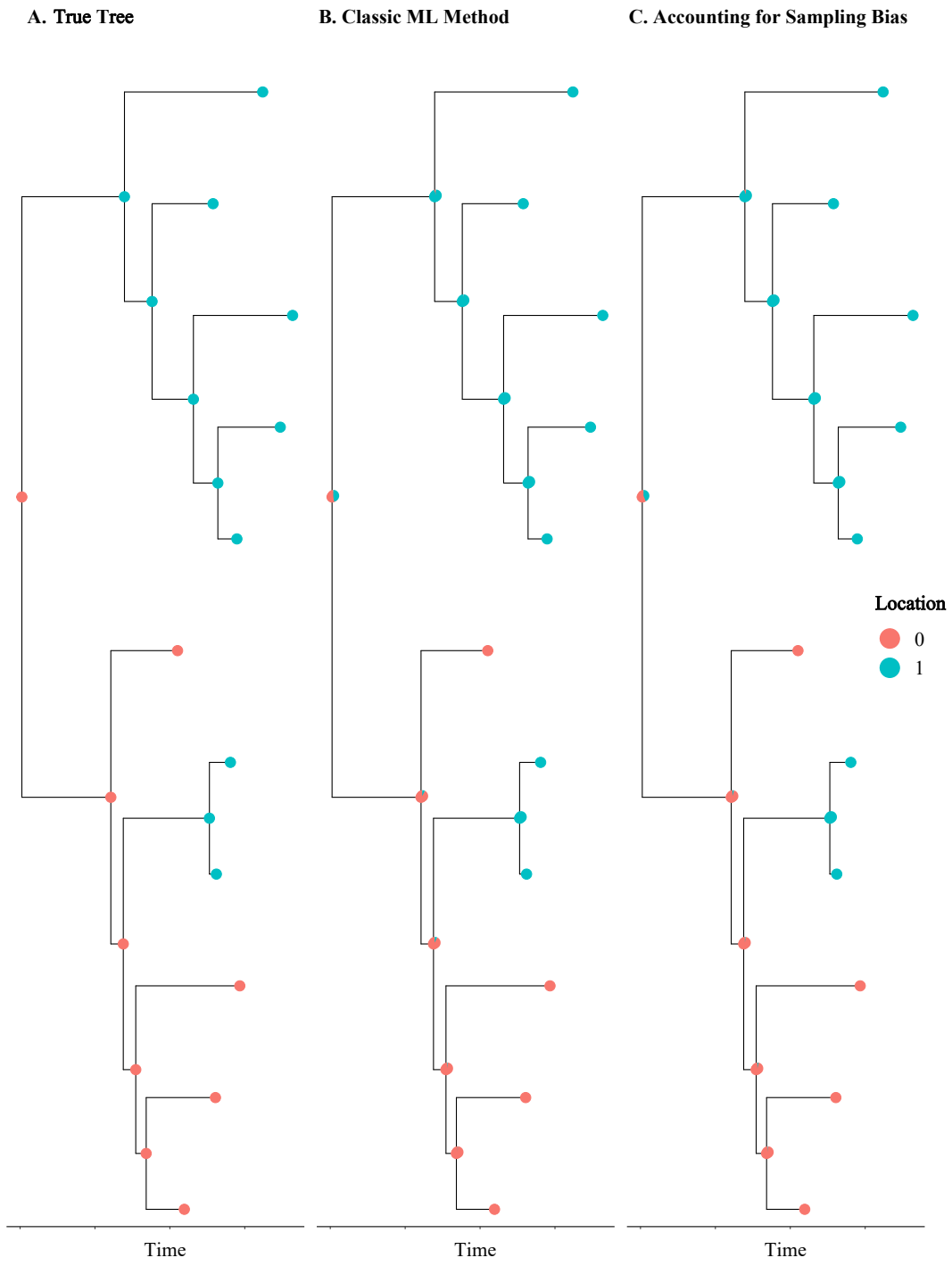

Figure 3.2: **Ancestral state reconstruction of the simulated tree using my approach and the classic Maximum Likelihood with similar sampling rate** Left panel: True tree with the correct internal states. Middle panel: Ancestral state reconstruction using the classic Maximum Likelihood method. Left panel: Ancestral state reconstruction using my approach.

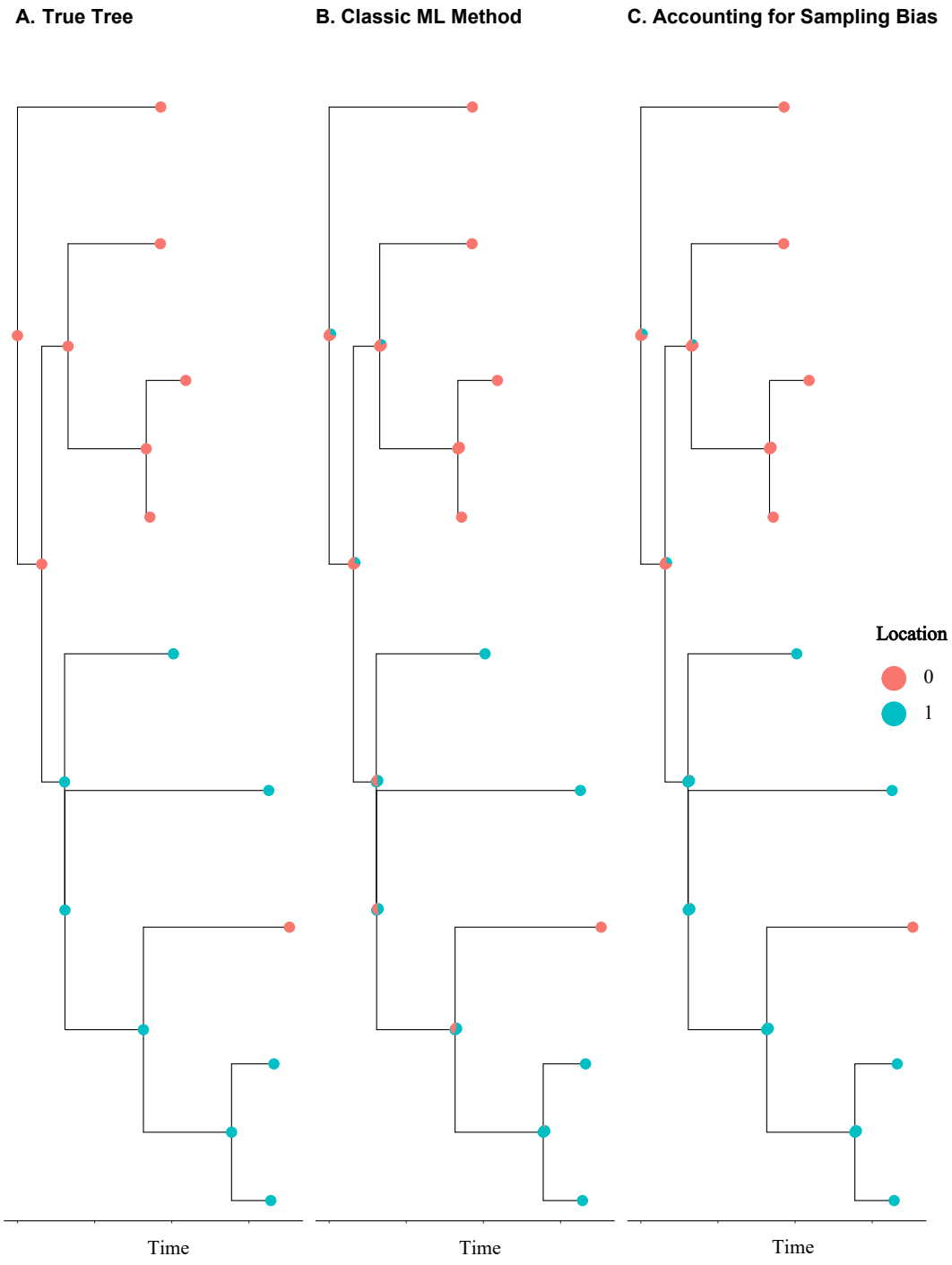

Figure 3.3: **Ancestral state reconstruction of the simulated tree using my approach and the classic Maximum Likelihood with different sampling rates** Left panel: True tree with the correct internal states. Middle panel: Ancestral state reconstruction using the classic Maximum Likelihood method. Left panel: Ancestral state reconstruction using my approach.

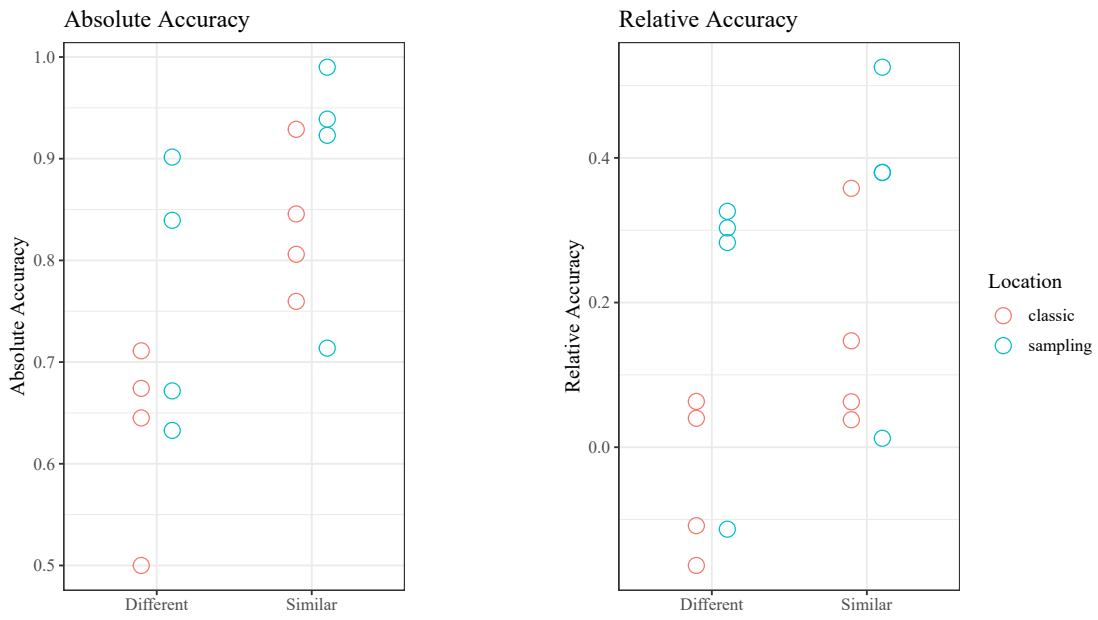

Figure 3.4: **Absolute accuracy and relative accuracy box plot** Left panel: Absolute accuracy plot for my approach (blue) and the classic Maximum Likelihood approach (red). Left panel: Relative accuracy plot for my approach (blue) and the classic Maximum Likelihood approach (red)

Chapter 4

Discussion

Here I proposed a new method that accounts for sampling bias when inferring ancestral states on a fixed simulated phylogenetic tree and then presents a comparison of the results with the classic Maximum Likelihood method. I have shown that if the sampling rates are similar in both locations, the reconstruction accuracy is the same for my method, where sampling bias is taken into account, and the classic Maximum Likelihood method. Furthermore, if sampling bias exists (one location is sampled more than the other), my approach correctly identifies the true internal states with a higher probability. In contrast, the classic approach favours the up-weighted location, causing inaccurate ancestral state reconstruction. Here I demonstrate that my approach has higher ancestral state reconstruction accuracy than the classic method for small phylogenetic trees and has the potential to more accurately reconstruct ancestral states in large trees with more than 500 tips) with sampling bias.

Ancestral state reconstruction that accounts for sampling bias can be applied in many fields of biology. In viral genetics, we can reconstruct the viral transmission among locations more accurately and identify the key introduction event. For example, over ten million public COVID-19 sequences are available on GISAID [34]. However, computational time to generate trees using sequences is expensive, so it is impossible to include all the sequences and reconstruct ancestral states. A concession must be made to generate a tree with a limited number of viral sequences. My method will become helpful in ancestral state reconstruction after the tree has been reconstructed. Developed in a likelihood framework, future implementations of my method should be applicable for inferring ancestral states for large data sets such as this, but the results here also demonstrate that it is informative when the

sequences are limited. Different sampling and viral sequencing policies between areas can cause bias in phylogeographical inference, including the inference of ancestral locations. For example, undersampling virus genomes can lead to underestimating the number of introductions and distorting overall rates and trends in viral transmission. My method can account for sampling bias and give accurate reconstruction. In the macro-evolutionary context, my method can more accurately reconstruct the evolutionary history of focal characters (e.g., reproductive system) shaped by sampling biases central to addressing long-standing evolutionary questions.

My model has some limitations. First, I only simulated the evolution of a binary character with the same speciation and extinction rate. My method can, however, be easily generalized to the multi-location case with different speciation and extinction rates. I predict that the results will be consistent with the two-state case; locations with lower sampling rates will be overrepresented in the classic approach and accurately modelled using my approach. Second, I only compare my method with a single alternative inference approach, Pagel's Maximum Likelihood approach, leaving out comparisons with the other approaches highlighted in the introduction. My future work will compare my model's accuracy with other methods, such as the Bayesian methods [31], where the sampling bias issue is also considered or described. Third, my method has only been applied to simulated trees. In the future, I will apply my method to real data sets. For instance, the Ebola data set described in the introduction where three major locations can be considered (Sierra Leone, Guinea and Liberia). In addition, I will also apply the method to SARS-CoV-2 trees: Figure A.7 is a phylogenetic tree consisting of a data set collected from Peru cities over the past year (sequences are available on Nextstrain [3], a public data set tracking real-time pathogen evolution). The phylogenetic tree is obtained by using BEAST [35, 36].

Furthermore, I only tested my results using trees with fewer than 20 tips because the algorithm implemented was computationally expensive due to one recursive algorithm. Specifically, it was challenging to implement an efficient algorithm that tracks the probability at nodes in the forward direction (pre-traversal algorithm). I will explore other people's work, such as Freyman and Hohna [13], who used a similar method. I will also attempt the dynamic

algorithm described by Felsenstein [21] to reduce the running time. I will also attempt to extend my method by making sampling rates time-dependent. In my current model, sampling rates are constant throughout time. However, the area may have different data collection and sequencing protocols. For instance, for SARS-CoV-2, different countries have different restriction policies and testing protocols over time. Different sampling rates over time will highly affect ancestral state reconstruction accuracy.

In conclusion, my method has better ancestral state reconstruction accuracy than the classical Maximum Likelihood method on small trees and can be applied to real-world problems to discover the transmission of outbreaks and disease dynamics.

Bibliography

[1] Douglas J. Futuyma and Anurag A. Agrawal. Macroevolution and the biological diversity of plants and herbivores. *PANS*, 106:18054–18061, 2009.

[2] Pagel Mark. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Biological Sciences*, 255:37–45, 1999.

[3] James Hadfield Colin Megill Sidney M Bell John Huddleston Barney Potter Charlton Callender Pavel Sagulenko Trevor Bedford and Richard A Neher. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, 34:4121–4123, 2018.

[4] Richard H. Liang Rosemary M. McCloskey T. Nguyen Art F. Y. Poon Jeffrey B. Joy. Ancestral reconstruction. *PLoS Comput Biol*, 12:e1004763, 2016.

[5] Nei M Yang Z, Kumar S. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics*, 141:1641–1650, 1995.

[6] O'Reilly KM De Maio N Wu C-H and Wilson D. New routes to phylogeography: A bayesian structured coalescent approximation. *PLoS Genet*, 11, 2015.

[7] Rewar S and Mirdha D. Transmission of ebola virus disease: an overview. *Ann Glob Health*, 80:444–451, 2014.

[8] Magee D and Scotch M. The effects of random taxa sampling schemes in bayesian virus phylogeography. *Infect Genet Evol*, 64:225–230, 2018.

[9] Pengyu Liu Yexuan Song Caroline Colijn and Ailene MacPherson. The impact of sampling bias on viral phylogeographic reconstruction. 2022.

[10] Carroll M. Matthews D. Hiscox J. et al. Temporal and spatial analysis of the 2014–2015 ebola virus outbreak in west africa. *Nature*, 524:97–101, 2015.

[11] Suchard MA Lemey P Baele G Ayres DL Drummond AJ Rambaut A. Bayesian phylogenetic and phylodynamic data integration using beast 1.10. *Virus Evolution*, 4:vey016, 2018.

[12] Wayne P. Maddison Peter E. Midford and Sarah P. Otto. Estimating a binary character's effect on speciation and extinction. *Systematic Biology*, 56:701–710, 2007.

[13] Freyman and Höhna S. Stochastic character mapping of state-dependent diversification reveals the tempo of evolutionary decline in self-compatible onagraceae lineages. *Systematic Biology*, 68:505–519, 2019.

[14] FitzJohn R.G. Diversitree: comparative phylogenetic analyses of diversification in r. *Methods in Ecology and Evolution*, 3:1084–1092, 2012.

[15] Edwards AW. Statistical methods for evolutionary trees. *Genetics*, 183:5–12, 2009.

[16] Cavalli-Sforza LL and Edwards AW. The reconstruction of evolution. *Ann. Hum. Genet. London*, 27:105, 1963.

[17] Farris James S. Estimation of conservatism of characters by constancy within biological populations. *Evolution*, 20:587–591, 1966.

[18] Farris James S. Methods for computing wagner trees. *Systematic Zoology*, 19:83–92, 1970.

[19] Fitch Walter M. Toward defining the course of evolution: Minimum change for a specific tree topology. *Systematic Zoology*, 20:406–416, 1971.

[20] Felsenstein Joseph. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology*, 27:401–410, 1987.

[21] Felsenstein Joseph. Evolutionary trees from dna sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–376, 1981.

[22] Pagel Mark. The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. *Systematic Biology*, 48:612–622, 1999.

[23] Nielsen Rasmus. Mapping mutations on phylogenies. *Systematic Biology*, 51:729–739, 2002.

[24] Nielsen Rasmus. Huelsenbeck John P. and Bollback Jonathan P. Stochastic mapping of morphological characters. *Systematic Biology*, 52:131–158, 2003.

[25] Lemey P Rambaut A Drummond AJ Suchard MA. Bayesian phylogeography finds its roots. *PLoS Comput Biol*, 25, 2009.

[26] Tavaré S. *Some probabilistic and statistical problems in the analysis of DNA sequences.* 1986.

[27] Lemey P Rambaut A Welch JJ Suchard MA. Phylogeography takes a relaxed random walk in continuous space and time. *Mol Biol Evol*, 8:1877–1885, 2010.

[28] De Maio N Wu CH O'Reilly KM Wilson D. New routes to phylogeography: A bayesian structured coalescent approximation. *PLoS Genet*, 11, 2015.

[29] Müller NF Rasmussen DA Stadler T. The structured coalescent and its approximations. *Mol Biol Evol*, 34:2970–2981, 2017.

[30] Lemey P Rambaut A Bedford T Faria N Bielejec F Baele G et al. Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza h3n2. *PLoS Pathog*, 10, 2014.

[31] Lemey P. Hong S.L. Hill V. et al. Accommodating individual travel history and unsampled diversity in bayesian phylogeographic inference of sars-cov-2. *Nat Commun*, 11, 2020.

[32] Mitter Charles et al. The phylogenetic study of adaptive zones: Has phytophagy promoted insect diversification? *The American Naturalist*, 132:107–128, 1988.

[33] Barraclough TG Nee S. Phylogenetics and speciation. *Trends Ecol Evol*, 16:391–399, 2001.

[34] I. Lipman D. et al. Bogner, P. Capua. A global initiative on sharing avian flu data. *Nature*, 442, 2006.

[35] Bouckaert R. Vaughan T.G. Barido-Sottani J. Duchêne S. Fourment M. Gavryushkina A. et al. Beast 2.5: An advanced software platform for bayesian evolutionary analysis. *PLoS computational biology*, 15, 2019.

[36] Romero Pedro Eduardo and Camila Castillo-Vilcahuaman. Data mining of dna sequences submitted by peruvian institutions to public genetic databases. *Revista Peruana De Biología*, 28, 2021.

Appendix A

Figure

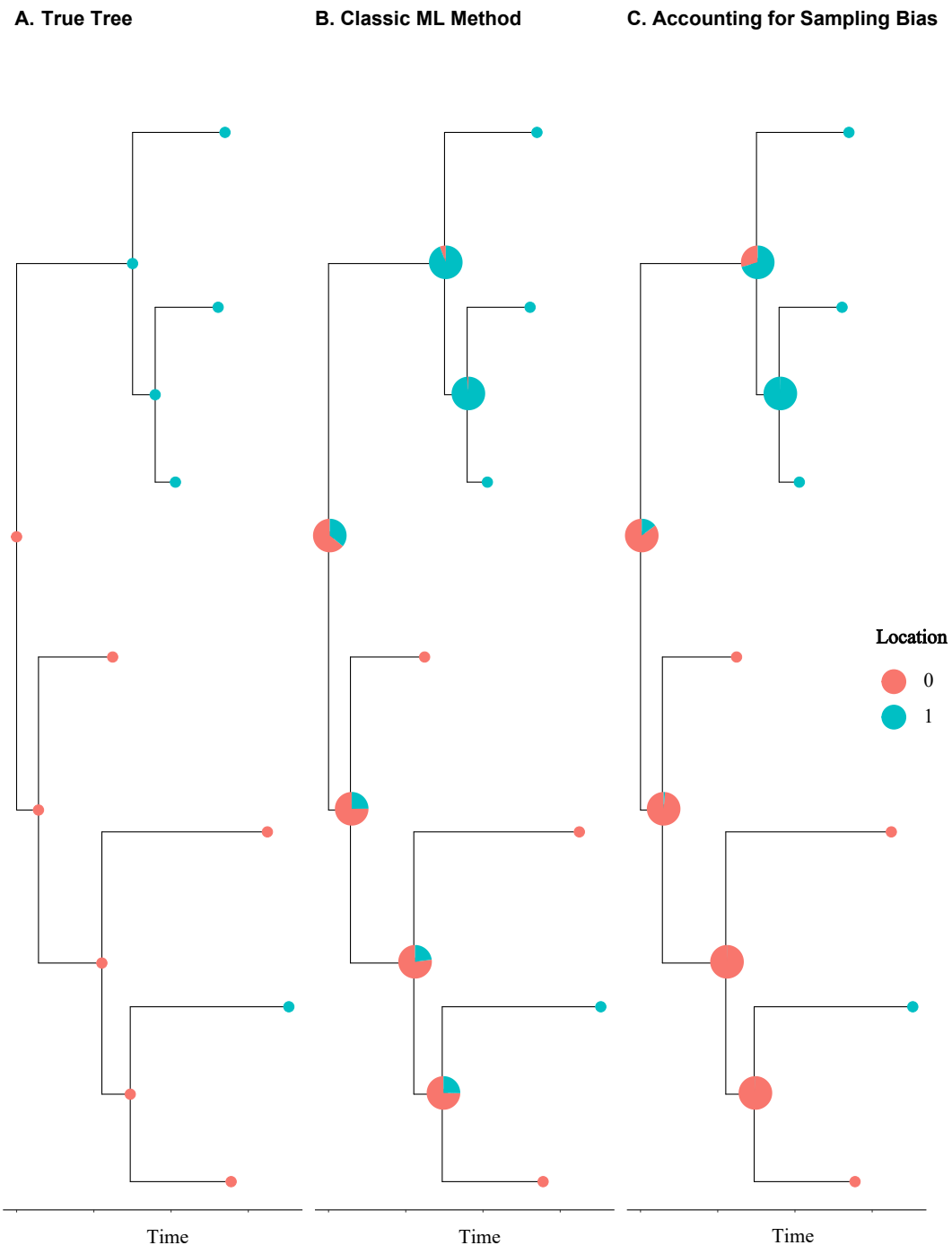

Figure A.1: This is the first of 3 simulated trees with a similar rate to which we tested the model. Left panel: True tree with the correct internal states. Middle panel: Ancestral state reconstruction using the classic Maximum Likelihood method. Right panel: Ancestral state reconstruction using my approach accounting for sampling bias. Similar to Figure 3.2, $\psi_2 = 1.2\psi_1$ and $f_2 = 1.2f_1$. The colour indicates different locations. The pie chart indicates the probability of the ancestral states.

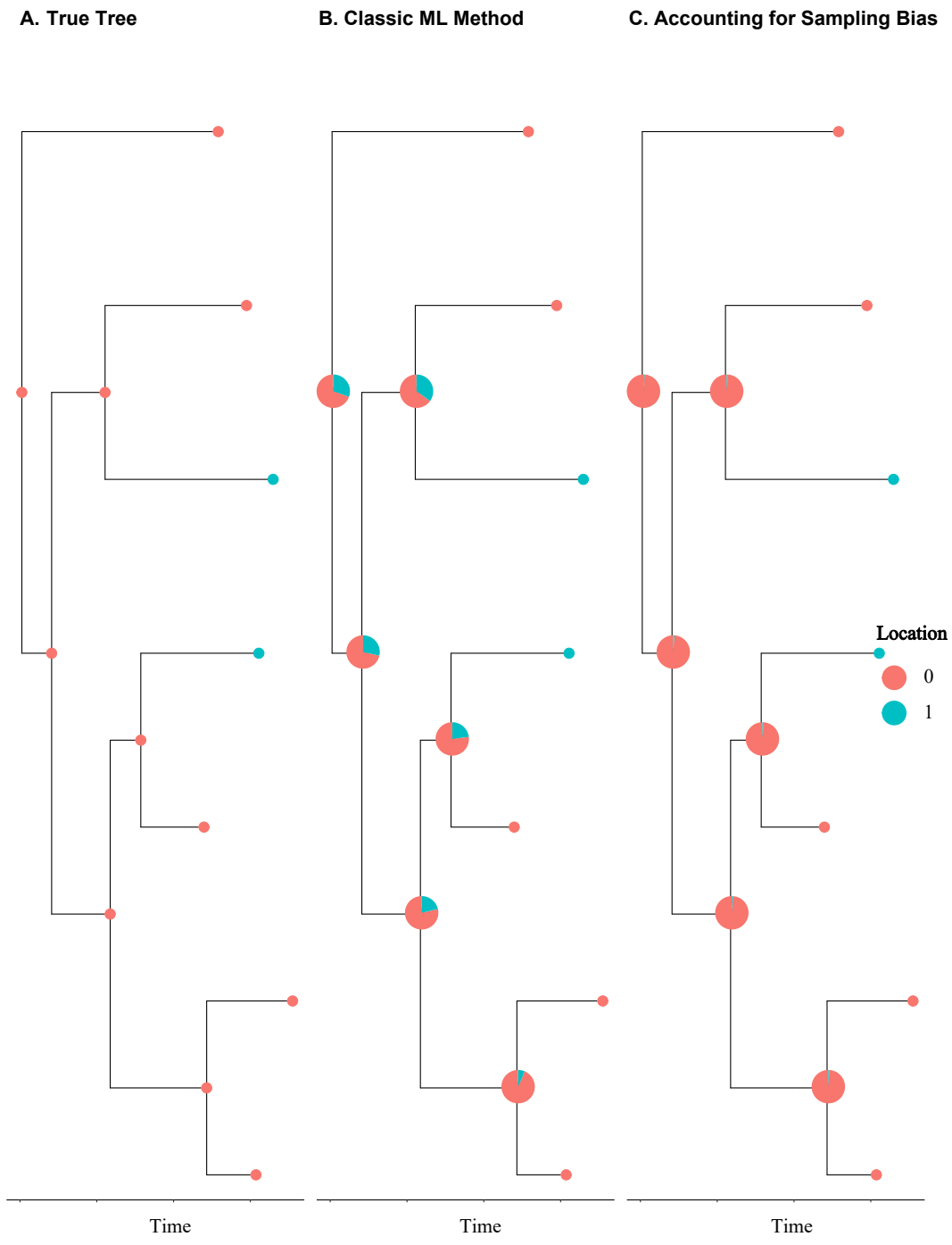

Figure A.2: This is the second of 3 simulated trees with a similar rate in which we tested the model. Left panel: True tree with the correct internal states. Middle panel: Ancestral state reconstruction using the classic Maximum Likelihood method. Right panel: Ancestral state reconstruction using my approach accounting for sampling bias. Similar to Figure 3.2, $\psi_2 = 1.2\psi_1$ and $f_2 = 1.2f_1$. The colour indicates different locations. The pie chart indicates the probability of the ancestral states.

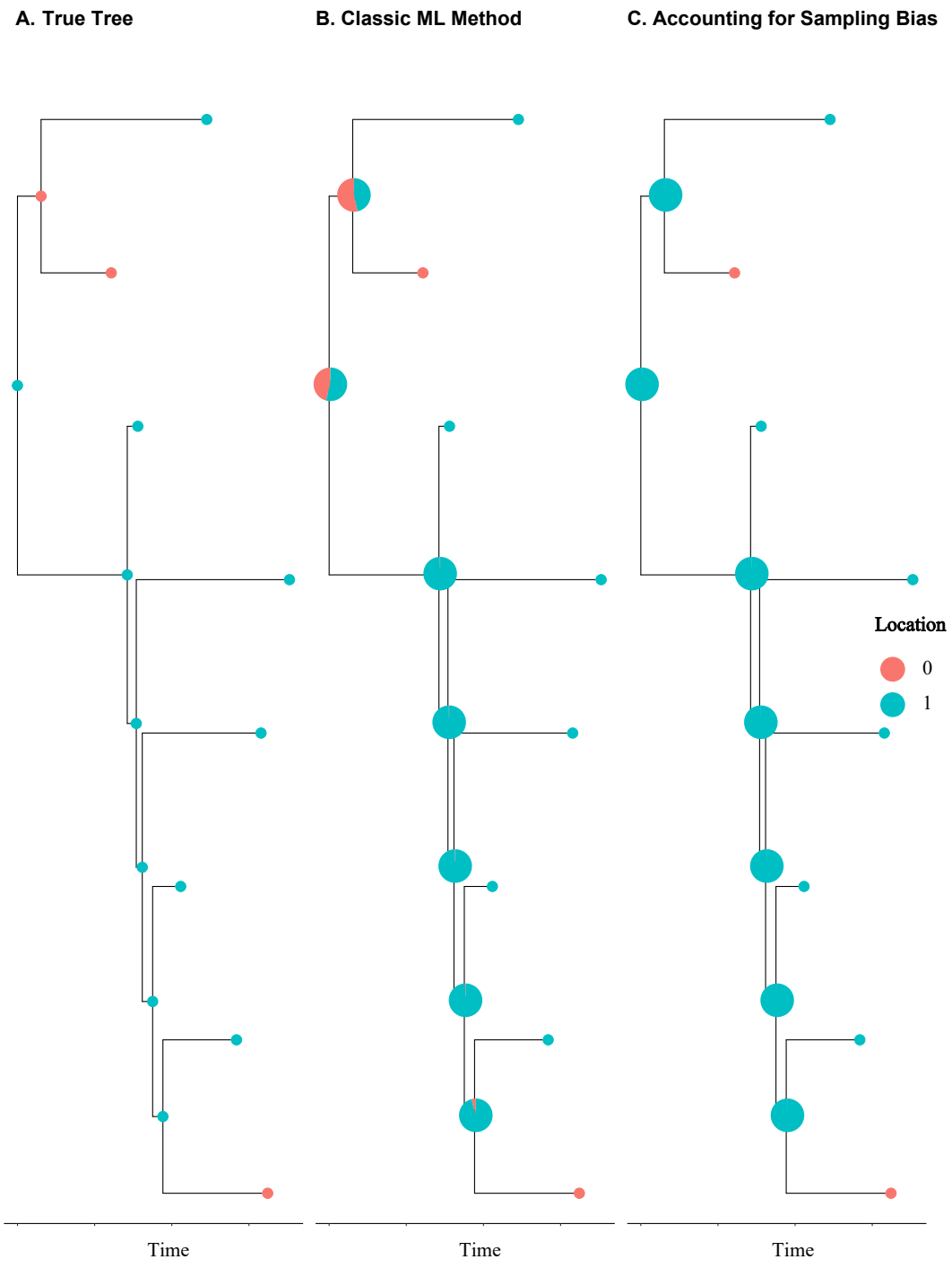

Figure A.3: This is the third of 3 simulated trees with a similar rate in which we tested the model. Left panel: True tree with the correct internal states. Middle panel: Ancestral state reconstruction using the classic Maximum Likelihood method. Right panel: Ancestral state reconstruction using my approach accounting for sampling bias. Similar to Figure 3.2, $\psi_2 = 1.2\psi_1$ and $f_2 = 1.2f_1$. The colour indicates different locations. The pie chart indicates the probability of the ancestral states.

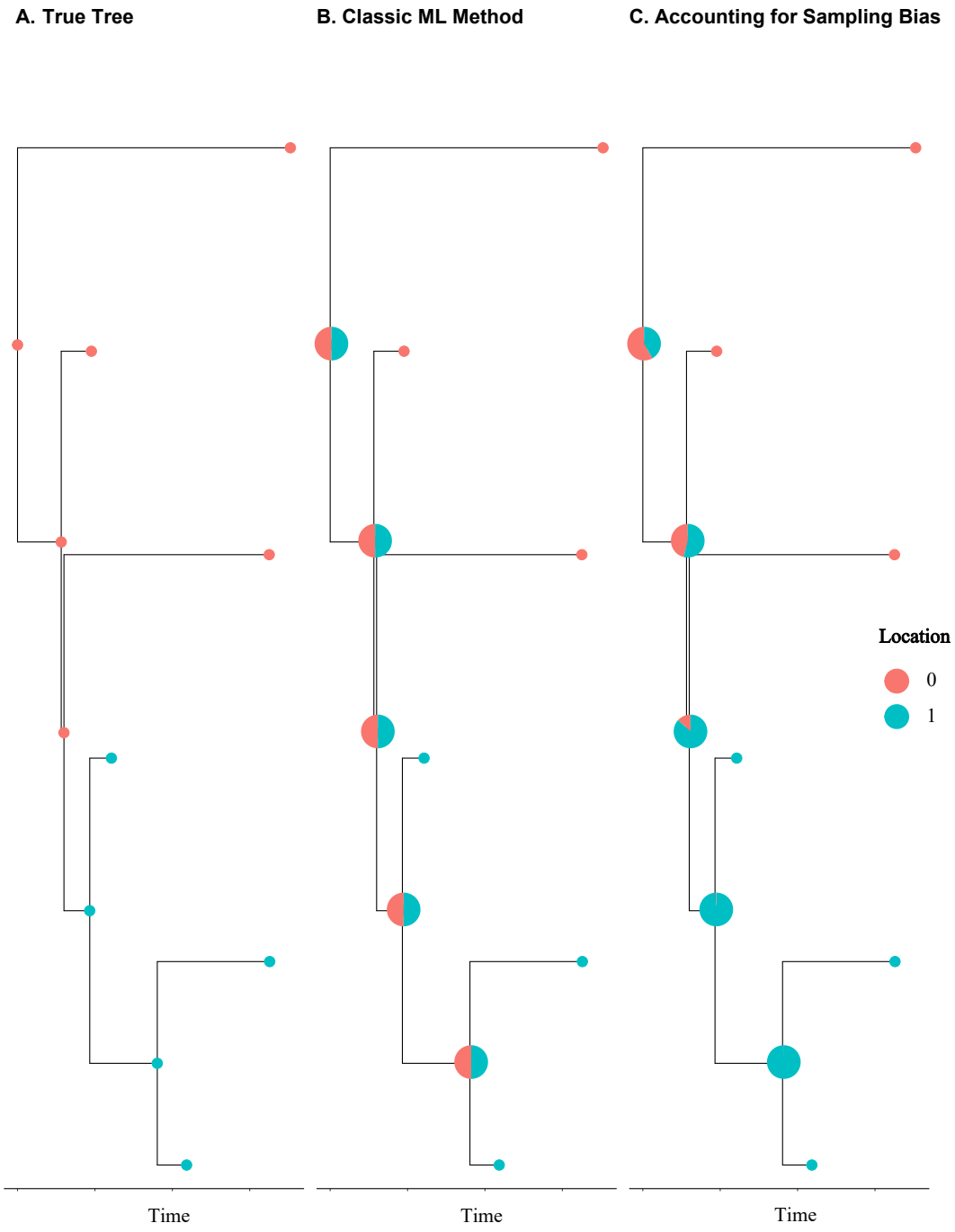

Figure A.4: This is the first of 3 simulated trees with different rates in which we tested the model. Left panel: True tree with the correct internal states. Middle panel: Ancestral state reconstruction using the classic Maximum Likelihood method. Right panel: Ancestral state reconstruction using my approach accounting for sampling bias. Similar to Figure 3.3, $\psi_2 = 1.7\psi_1$ and $f_2 = 1.7f_1$. The colour indicates different locations. The pie chart indicates the probability of the ancestral states.

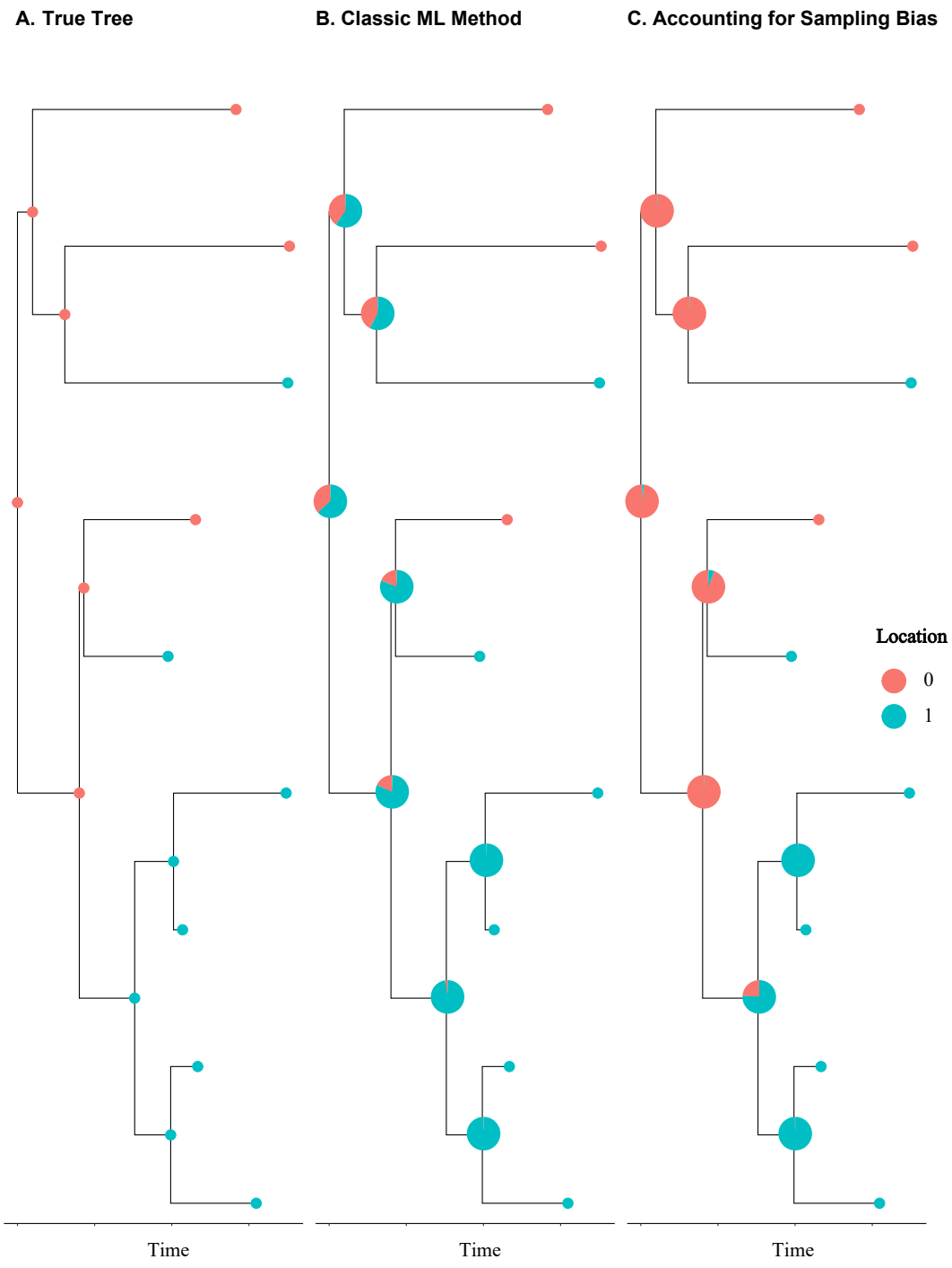

Figure A.5: This is the second of 3 simulated trees with different rates in which we tested the model. Left panel: True tree with the correct internal states. Middle panel: Ancestral state reconstruction using the classic Maximum Likelihood method. Right panel: Ancestral state reconstruction using my approach accounting for sampling bias. Similar to Figure 3.3, $\psi_2 = 1.7\psi_1$ and $f_2 = 1.7f_1$. The colour indicates different locations. The pie chart indicates the probability of the ancestral states.

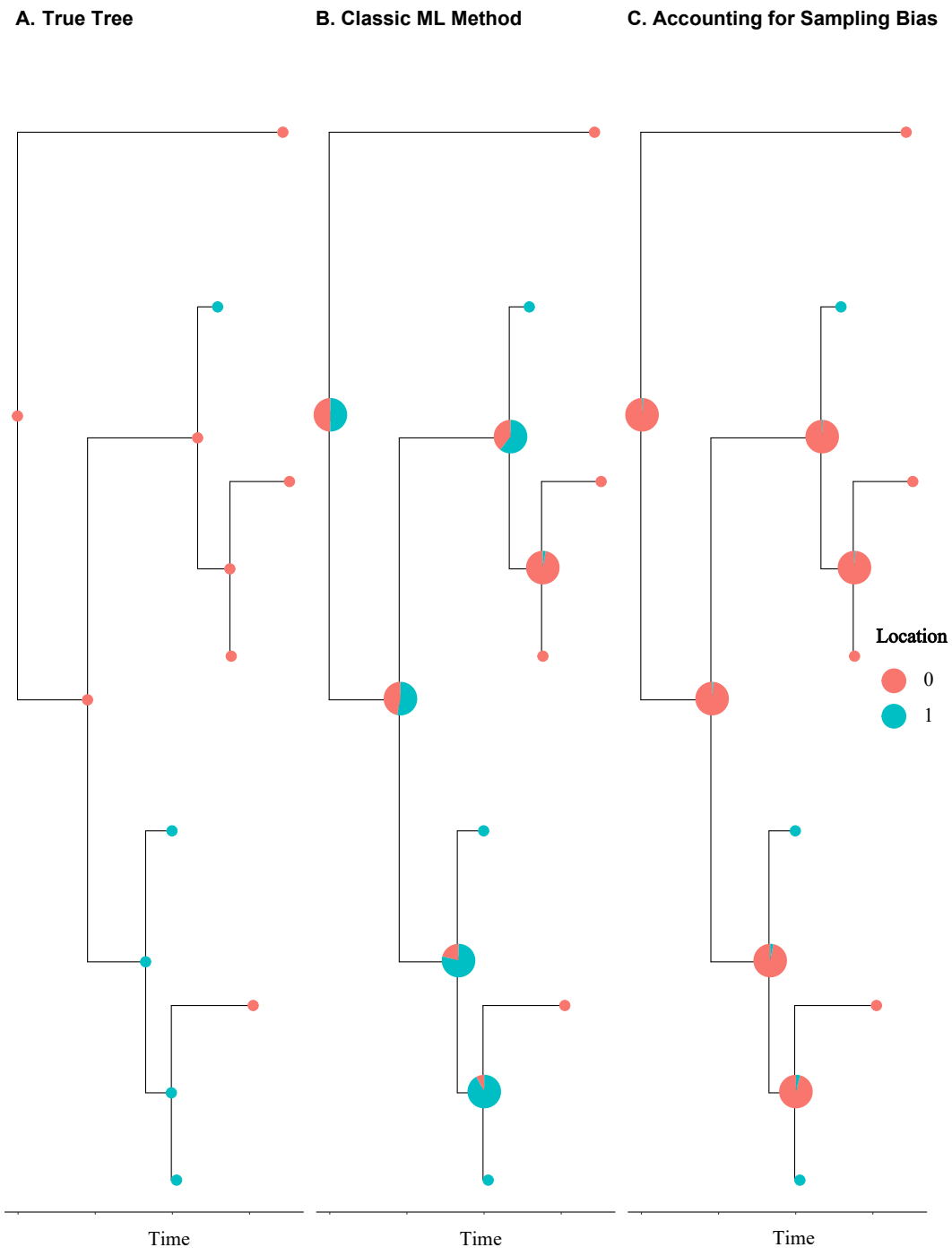

Figure A.6: This is the third of 3 simulated trees with different rates in which we tested the model. Left panel: True tree with the correct internal states. Middle panel: Ancestral state reconstruction using the classic Maximum Likelihood method. Right panel: Ancestral state reconstruction using my approach accounting for sampling bias. Similar to Figure 3.3, $\psi_2 = 1.7\psi_1$ and $f_2 = 1.7f_1$. The colour indicates different locations. The pie chart indicates the probability of the ancestral states.

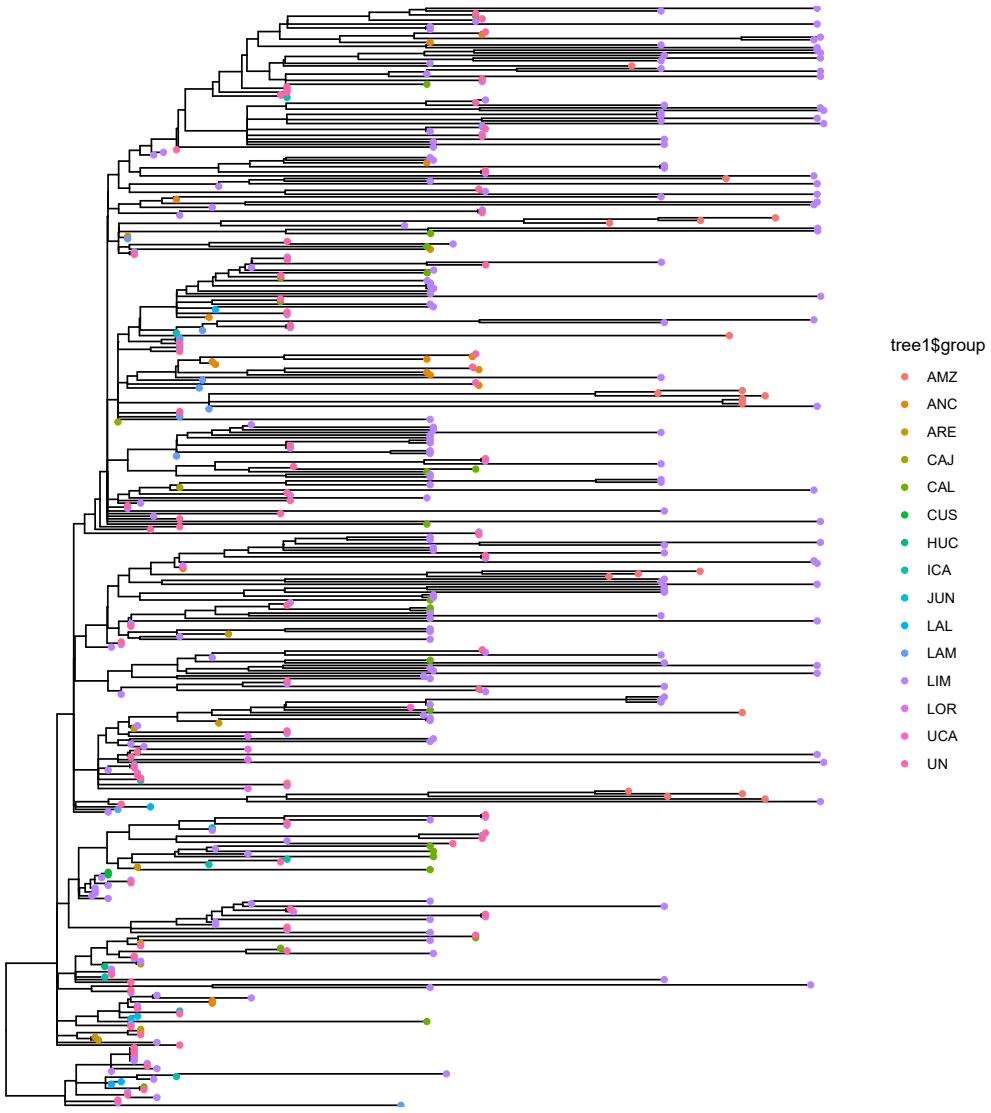

Figure A.7: **Peru time-scaled tree from Nextstarin.**