Sampling Aware Ancestral State Inference

Yexuan Song*[a], Ivan Gill[b], Ailene MacPherson[a], Caroline Colijn[a]

[a]*Department of Mathematics, Simon Fraser University, 8888 University Dr., Burnaby, BC, V5A 1S6, Canada*
[b]*Department of Health Sciences, Simon Fraser University, 8888 University Dr., Burnaby, BC, V5A 1S6, Canada*

Abstract

Reconstructing the states of ancestral organisms has long been central to our understanding of the evolution of a wide range of traits. Ancestral state inference tools that account for trait-dependent properties are limited, because of challenges associated with inferring past states in a manner consistent with a phylogenetic tree (and its uncertainty) and with a stochastic process describing how states change over time. In phylogeography, ancestral state inference is used to reconstruct the past locations of viruses, bacteria or other rapidly-evolving organisms, characterizing, for example, how often and when a virus moved among locations, or from one host species to another. However, such reconstructions are sensitive to differences in sampling among different locations or host species, and this can bias the reconstruction of the location of ancestors towards the more widely sampled region/species. Here, we introduce a new method, Sampling Aware Ancestral State Inference (SAASI), which builds on recent advances in state-dependent diversification models and reconstructs ancestral states, and in particular for phylogeographic applications, accounting for sampling differences. Indeed, we find that accounting for sampling changes the inferred historical location of viral lineages and the times of key viral movements. We use simulations to show that with known sampling differences, SAASI infers past viral locations considerably more accurately than standard methods. We apply our method to the spread of the H5N1 virus in the United States in 2024, and explore how robust phylogeographic reconstruction is to differences in sampling and epidemiological rates between wild bird populations, cattle, humans and other species. We find that the key transmission event from wild birds to cattle is estimated to occur later under lower sampling in wild birds (compared to other species) than when sampling is not accounted for. SAASI is rapid and readily scales to trees with 100,000 tips, making it feasible for modern phylogeographic applications.

1. Introduction

Ancestral state inference (sometimes called ancestral state reconstruction) refers to inferring the states of the ancestors of a set of taxa, given data about the states of the taxa. It is used in phylogenetics to infer the likely molecular sequences of ancestral organisms, for example to estimate when and in which lineages a particular polymorphism (for example conferring antibiotic resistance) emerged [1]. Ancestral state inference is also used to infer traits such as a pathogen's host species [2, 3], organisms' geographic locations [4–6], accessory gene presence/absence [7, 8] or morphological or other traits [9, 10]. Phylogeography, in particular, is a major application of ancestral state inference, in which past geographic movements of viruses

Email address: `yexuan_song@sfu.ca` (Yexuan Song*)

or other pathogens are reconstructed. Understanding viral transmission between groups of a heterogeneous host population is important for reconstructing epidemic origins and designing effective control strategies. For example, reconstructing the geographic spread of a virus can inform policy decisions about transportation and borders [11]. Estimates of interspecific transmission of multi-host species movements in zoonotic viruses can help us identify the determinants of cross-species transmission events and, hence, of zoonotic risk.

Ancestral state inference methods use models that describe the process by which traits change over time, or by which the 'state' (which could be the geographic or host location of the taxa) changes. Different traits or states may also be associated with different rates of speciation (branching), extinction (of the relevant lineage) and sampling. The binary state-dependent speciation and extinction (BiSSE) family of models [12], taking this into account, can estimate trait-specific branching and extinction rates given a phylogenetic tree and traits of the sampled taxa. These models include multiple-state (MuSSE) models, hidden Markov models (HiSSE), cladogenesis (ClaSSE) models and more [13–15]. Since trait-dependent speciation and extinction also impact the likelihood of a phylogenetic tree given sequence and trait data, these models inform phylogenetic reconstructions [16] in which state-dependent rates can be estimated alongside phylogenetic trees.

Inferring the states of ancestral organisms poses a problem that is distinct from estimating state-dependent rates. In ancestral state inference, each node in the phylogeny (and sometimes each point in the phylogeny) is associated with a particular state, or with a set of probabilities that the point is in each of the possible states. This may be done with stochastic character-mapping: mapping characters (or traits, or states) on to a phylogeny using a stochastic model [17–19], which is implemented for the case of a molecular evolution model in simmap [20]. Freyman and Höhna [21] developed a stochastic character mapping method for state-dependent models (the BiSSE family). This approach, however, is not yet widely used in viral (or pathogen) phylogeography. In this context, the states (e.g. geographic locations) are usually modelled as changing along a phylogenetic tree according to a continuous-time Markov chain, in a similar manner to how a molecular sequence evolves over time [22–24]: using a continuous-time Markov chain with a given instantaneous rate matrix Q specifying the rates at which the state (location) changes [20, 22]. This is implemented, for example, in the ace function in R's 'ape' package [25]. A feature of viral geographic and/or host species data that is not yet represented in the BiSSE-based/stochastic mapping literature is the heavily-biased sampling of lineages across space or among host species.

Bayesian approaches to phylogeography have also been developed, building on continuous time Markov chains for discrete states [5, 26] and on structured coalescent models [27, 28], among others. Bayesian stochastic search variable selection (BSSVS) [5, 26] allows determination of which transition rates (i.e. which elements of Q) are supported, using a Markov diffusion model. Rates are parameterized as a log-linear function of any number of predictors (these can be phylogenetic distances, range overlap, morphological similarity among species, etc), and predictors may fail to be selected by the model. BSSVS does not, however, account for state-dependent speciation and extinction rates, and it does not allow the phylogeographic reconstructions to depend on the sampling among species or locations.

Sampling differences are known to impact phylogeographic estimates [4, 29, 29–32]. Down-sampling is often used to try to achieve relatively uniform sampling across locations through maximizing spatial or temporal coverage [11, 32–34], using epidemiological data as a reference point (e.g. hospitalizations vs cases) [35], or incorporating information about recent migration events and/or adding 'sequence-free' samples [36, 37]. Down-sampling reduces the amount of

data that can be incorporated, may need to be replicated many times and so is time-consuming, and may not ultimately solve the problem. Structured coalescent models, like Bayesian structured coalescent approximation (BASTA) [27] and Marginal Approximation of the Structured Coalescent (MASCOT) [28] are more robust to sampling, and can accommodate sampling differences through adjusting deme sizes, but they are computationally demanding, still sensitive to unsampled demes, and are not feasible for trees with many thousands of taxa. Sampling may be accounted for in structured coalescent models using a doubly-intractable model [31]. Furthermore, sampling differences may be extreme, as in the case of the recent influenza H5N1 outbreak, in which samples from livestock and human cases are more readily available than samples from wild animal populations. As genomic surveillance expands and we prepare for additional zoonotic spillover events that could bring future large outbreaks or pandemics, it is important to account for sampling and other state-dependent variation in our inferences of viral (and other pathogen) locations.

Here, we develop sampling-aware ancestral state inference (SAASI), building on recent work on stochastic character mapping on fixed phylogenetic trees in state-dependent speciation and extinction (SSE) models [21]. We introduce two core developments: the inclusion of, and consequent adjustment for, sampling differences, and a modification to the stochastic character mapping method for BiSSE-like models in conditioning on the observed tree. Our approach scales readily to very large phylogenetic trees. We test SAASI with simulated data, comparing to ancestral character estimates from the widely used ace function in R [25] and simmap [20]. We then use SAASI to explore robustness to sampling for key host jumps and geographic movements in the recent avian influenza H5N1 outbreak in the United States [2].

2. Methods

We consider a rooted binary phylogenetic tree $\mathcal{T}$ with known tree topology and character states (e.g. location, host species identification) at the tips. We assume that this tree is the result of a state-dependent birth-death-sampling process allowing for both cladogeneic and anagenic state changes. Specifically, a lineage in state i gives birth (e.g., viral transmission in the epidemiological context) to daughter lineages in states j and k at rate λ_{ijk} which define the 3D array Λ, dies (e.g., host recovery) at rate μ_i which define the vector $\vec{\mu}$, and transitions to state j at rate q_{ij} which define the matrix Q (see Table 1 for a list of notation). In addition to these rates, we introduce the state-dependent sampling rate $\psi_i \in \vec{\psi}$ through time, with which viral sequences are collected throughout an ongoing epidemic. Finally, let π_i be the prior probability that the root R of the phylogenetic tree is in state i (defining the vector $\vec{\pi}$).

Here, we propose and implement a method for calculating the probability that at any time τ before the present day, an edge e in the phylogeny was in state $i \in S$, denoted $Y_e(\tau) = i$, accounting for both the observed states at the tips of the tree and the topology of the tree resulting from state-dependent diversification. Using the notation above, we wish to calculate:

$$A_{e,i}(\tau) \equiv \Pr(Y_e(\tau) = i | \mathcal{T}, \theta)$$

where $\theta = \{\Lambda, \vec{\mu}, \vec{\psi}, Q, \vec{\pi}\}$ is the diversification model. Throughout, we define this kind of shorthand notation for each key probability, e.g. $A_{e,i}(\tau)$. To calculate this focal probability, we formalize and extend the stochastic-mapping method presented by [21]. Our method involves first a post-order traversal of the tree (from tips to root) followed by a pre-order traversal (from root to tips) during which the ancestral state probabilities are obtained. To avoid confusion on the direction of time, throughout, we use T to denote the height of the tree, τ to denote time

from the present day to the root of the tree, and t to denote time from the root to the present day such that $\tau = T - t$. In line with SSE models, for functions that are neither more naturally described backward vs. forward in time (e.g., $A_{e,i}(\tau)$), we will use backward-in-time notation. We give an overview, rather than a detailed derivation, of the pre- and post-order traversal approach.

To understand the need for a post/pre-order traversal method, note that, for a focal edge e which is alive at time τ before the present day, the full tree $\mathcal{T}$ can be subdivided into two components: the descendants of edge e which we denote as the subtree $\mathcal{T}_e(\tau)$, and the remainder of the tree, $\mathcal{T}_e^C(\tau)$, which contains all the ancestors of node e as well as their non-e descendants. For ease of writing, we call $\mathcal{T}_e^C(\tau)$ the 'ancestral complement' of edge e. Importantly, as our state model is Markovian (the type of speciation event and state-transition events depend only on the current state of the lineage), $\mathcal{T}_e(\tau)$ and $\mathcal{T}_e^C(\tau)$ are conditionally independent given the state of edge e at time τ, $Y_e(\tau)$. As such, we can decompose the expression for the probability $A_{e,i}(\tau)$ as:

$$A_{e,i}(\tau) = \Pr(Y_e(\tau) = i|\mathcal{T},\theta) = \frac{\Pr(Y_e(\tau) = i|\mathcal{T}_e^C(\tau),\theta)\Pr(\mathcal{T}_e(\tau)|Y_e(\tau) = i,\theta)}{\sum_j \Pr(Y_e(\tau) = j|\mathcal{T}_e^C(\tau),\theta)\Pr(\mathcal{T}_e(\tau)|Y_e(\tau) = j,\theta)} \tag{1}$$

where $\Pr(\mathcal{T}_e(\tau)|Y_e(\tau) = i,\theta)$ is calculated during the post-traversal step and $\Pr(Y_e(\tau) = i|\mathcal{T}_e^C(\tau),\theta)$ during the pre-traversal step described below. (See the Supplementary Materials for a derivation of (1).) Specifically, we can draw from the well-developed tree likelihoods for SSE models to calculate the probability of observing a descendant sub-tree, $\mathcal{T}_e(\tau)$, given that the edge e is present in state i at time τ before the present day:

$$D_{e,i}(\tau) \equiv \Pr(\mathcal{T}_e(\tau)|Y_e(\tau) = i,\theta),$$

a probability that can be computed backward in time. As the probability $\Pr(Y_e(\tau) = i|\mathcal{T}_e^C(\tau),\theta)$ is obtained through a the pre-traversal algorithm (forward-in-time) we rewrite this expression in terms of time t:

$$\tilde{D}_{e,i}(t) \equiv \Pr(Y_e(T-t) = i|\mathcal{T}_e^C(T-t),\theta).$$

Because $\tilde{D}_{e,i}(t)$ is conditioned on $\mathcal{T}^C$, its calculation departs from what is done in previous literature [12, 38]. However, it is again a forward-in-time probability that can be obtained using a system of ordinary differential equations. The initial conditions will depend on the backward-in-time probabilities, hence we implement the post-order algorithm followed by the pre-order algorithm. With this notation, Eq. (1) becomes

$$A_{e,i}(\tau) = \Pr(Y_e(\tau) = i|\mathcal{T},\theta) = \frac{\tilde{D}_{e,i}(t)D_{e,i}(\tau)}{\sum_j \tilde{D}_{e,j}(t)D_{e,j}(\tau)}. \tag{2}$$

To calculate the probability that a node N is in state i, we use the time τ_N when the speciation event occurred, and calculate the corresponding $A_{e,i}(\tau_N)$. In summary, SAASI return a vector of A_i values at each node N, with the sum over states i equal to 1.

2.1. Four-step Algorithm

Step 1: Post-order traversal

Here, we derive an initial value problem backward in time for $D_{e,i}(\tau)$ — the probability that an edge e in state i alive at time τ before the present day gives rise to the observed descendants between time τ and the present. This derivation is standard (see [12, 38, 39] for a review), but we review it here to emphasize differences with the pre-order traversal below (step 3).

We consider the change in $D_{e,i}(\tau)$ over a small interval of time $\Delta\tau$, assuming that at most one event can occur.

$$\begin{aligned} D_{e,i}(\tau+\Delta\tau) = & \underbrace{D_{e,i}(\tau)\prod_{j,k}(1-\lambda_{i,j,k}\Delta\tau)(1-\mu_i\Delta\tau)(1-\psi_i\Delta\tau)\prod_{j\neq i}(1-q_{i,j}\Delta\tau)}_{\text{no event}} \\ & + \underbrace{\sum_{j,k} 2\lambda_{i,j,k}\Delta\tau\prod_{j\neq i}(1-q_{i,j}\Delta\tau)(1-\mu_i\Delta\tau)(1-\psi_i\Delta\tau)D_{e,j}(\tau)E_k(\tau)}_{\text{hidden speciation}} \\ & + \underbrace{\sum_{j\neq i} q_{i,j}\Delta\tau\prod_{j,k}(1-\lambda_{i,j,k}\Delta\tau)(1-\mu_i\Delta\tau)(1-\psi_i\Delta\tau)D_{e,j}(\tau)}_{\text{state change}} + \mathcal{O}(\Delta\tau^2) \end{aligned}$$

where $E_k(\tau)$ is the probability that a lineage in state k at time τ has no observed descendants. Using the definition of the derivative, we obtain:

$$\begin{aligned} \frac{d}{d\tau}D_{e,i}(\tau) = & -\left(\sum_{j,k}\lambda_{i,j,k}+\mu_i+\psi_i+\sum_{j\neq i}q_{i,j}\right)D_{e,i}(\tau) \\ & + \sum_{j,k} 2\lambda_{i,j,k}D_{e,j}(\tau)E_k(\tau) + \sum_{j\neq i} q_{i,j}D_{e,j}(\tau). \end{aligned} \tag{3}$$

An edge e, which by definition does not include the node itself, can originate (at time $\tau_{e,0}$) at one of two types of nodes, either a sampling event or a speciation event where the focal edge gives rise to two descendant edges e_1 and e_2.

$$D_{e,i}(\tau_{e,0}) = \begin{cases} \psi_i & \text{sampling event} \\ \sum_{i,j}\lambda_{i,j,k}D_{e_1,j}(\tau_{e,0}^-)D_{e_2,k}(\tau_{e,0}^-) & \text{speciation event} \end{cases}$$

where the notation $\tau_{e,0}^-$ is used to emphasize that these edges are descendants (closer to the tips) of the focal edge.

To calculate $D_{e,i}(\tau)$, we need to first solve $E_{e,i}(\tau)$ - the probability that an edge e in state i at time τ has no observed descendants between time τ and the present day. This probability can be obtained with the following differential equation (which is derived in a similar manner as above).

$$\frac{d}{d\tau}E_i(\tau) = \mu_i - \left(\sum_{j,k}\lambda_{i,j,k}+\mu_i+\psi_i+\sum_{j\neq i}q_{i,j}\right)E_i(\tau) + \sum_{j,k}\lambda_{i,j,k}E_j(\tau)E_k(\tau) + \sum_{j\neq i}q_{i,j}E_j(\tau) \tag{4}$$

where the initial condition for $E_i(\tau)$ is the probability that the focal edge is unsampled (since

we are not forced to sample all the lineages or any percentage of the lineages at the present day): $E_i(0) = 1$.

Step 2: Root state probabilities

Using the initial value problem (IVP) above to solve for $D_{e,i}(\tau)$ up the whole tree to the root, we can calculate the probability that the root is in state i ($\tilde{D}_{R,i}(0) = \pi_i$):

$$A_{R,i}(T) = \Pr(Y_R = i | \mathcal{T}_R(T), \theta) = \frac{D_{R,i}(T)\pi_i}{\sum_j D_{R,j}(T)\pi_j}$$

Given these root state probabilities, we can then proceed with the pre-order traversal algorithm down the tree to calculate $\tilde{D}_{e,i}(t)$.

Step 3: Pre-order traversal

To derive an IVP for the probability $\tilde{D}_{e,i}(t)$, we use a similar approach as above by considering the change in this probability over a small interval of time Δt. In words, we consider the probability that an edge e is in state i at time $t + \Delta t$ given the state of its immediate ancestor at time t. Unlike above, however, the probability $\tilde{D}_{e,i}(t)$ is conditioned on the observed ancestral complement tree $\mathcal{T}_e^C(T-t)$. Hence along an edge e the only events that we should include are those that reflect a change of state (e.g., a state change or a cladogenic state change where one descendant is unobserved), not those that reflect a change of topology (since we are conditioning on the topology). In consequence, we do not include terms for birth or extinction; conditioning on the observed tree, we know that these have not occurred. This is a substantial contrast to the post-traversal step in which we computed $D_{e,i}(\tau)$, the likelihood of $\mathcal{T}_e$ given that edge e is in state i at time τ.

$$\begin{aligned}\frac{d}{dt}\tilde{D}_{e,i}(t) = &-\left(\underbrace{\sum_{j\neq i,k} 2\lambda_{i,j,k}\tilde{D}_{e,i}(t)E_k(T-t)}_{\text{cladogenic change } i\to j} + \underbrace{\sum_{j\neq i} q_{i,j}\tilde{D}_{e,i}(t)}_{\text{state transition } i\to j}\right) \\ &+\left(\underbrace{\sum_{j\neq i,k} 2\lambda_{j,i,k}\tilde{D}_{e,j}(t)E_k(T-t)}_{\text{cladogenic change } j\to i} + \underbrace{\sum_{j\neq i} q_{j,i}\tilde{D}_{e,j}(t)}_{\text{state transition } j\to i}\right)\end{aligned} \tag{5}$$

Note that the probabilities must sum to one ($\sum_i \tilde{D}_{e,i}(t) = 1$) and hence this equation can be considered in terms of probability flux from one class (e.g., i) to another (e.g., j) and vice versa. If we consider only anagenic state change (i.e. a change of state without an accompanying speciation event), then we can further simplify the equation $\frac{d}{dt}\tilde{D}_{e,i}(t)$ by only considering the state transition events:

$$\frac{d}{dt}\tilde{D}_{e,i}(t) = -\underbrace{\sum_{j\neq i} q_{i,j}\tilde{D}_{e,i}(t)}_{\text{state transition } i\to j} + \underbrace{\sum_{j\neq i} q_{j,i}\tilde{D}_{e,j}(t)}_{\text{state transition } j\to i} \tag{6}$$

Note that this still allows for a rapid transition from one state to another immediately after a speciation event, but does not include speciation events that cause state changes. The method

can readily accommodate these if the relevant λ rates are known. For initial conditions, note that the edge e can originate (forward in time at time $t_{e,0}$) at either the root or at a speciation event. At the root, the ancestral complement $\mathcal{T}_e^C$ consists solely of the root itself and hence its probability is simply the prior on the root state, $\tilde{D}_{e,i}(0) = \pi_R$. At a speciation event, the state of the focal edge e at time t^+ (i.e., immediately following the speciation event) depends on (1) the state of the parent lineage, edge e^-, immediately prior to the speciation event at time t^-, (2) the type of speciation event to occur as given by the probability $\frac{\lambda_{j,i,k}}{\bar{\lambda}}$, and (3) the state of the sister edge e' of the focal edge immediately following the speciation event. Specifically,

$$\tilde{D}_{e,i}(t_{e,0}) = \begin{cases} \pi_i & e \to \text{root.}\ \ t = 0 \\ \frac{\sum_{j,k} 2\lambda_{j,i,k} \tilde{D}_{e,j}(t_{e,0}^-) D_{e',k}(T - t_{e,0})}{\sum_{j,l,k} \lambda_{j,l,k} \tilde{D}_{e,j}(t_{e,0}^-) D_{e',k}(T - t_{e,0})} & \text{internal node} \end{cases} \tag{7}$$

We therefore solve Eq. (5) along each edge, forward in time. Upon reaching an internal node N of the phylogeny, we use Eq. (7) to initialize the ODE on the two edges (say e_1 and e_2) descending from n. At this point, note that the ancestral complement tree, on which we are conditioning, changes from $\mathcal{T}_e^C$ to $\mathcal{T}_{e_1}^C$ or $\mathcal{T}_{e_2}^C$. Both of these include node N.

Step 4: Ancestral State Probabilities

Once $\tilde{D}_{e,i}(t)$ is computed during the pre-traversal step the ancestral state probabilities can be computed via the product outlined above:

$$A_{e,i}(\tau) = \frac{\tilde{D}_{e,i}(T - \tau) D_{e,i}(\tau)}{\sum_j \tilde{D}_{e,j}(T - \tau) D_{e,j}(\tau)} \quad \forall e \neq R \tag{8}$$

At each internal node N we then define the "reconstructed state" as the state with the highest probabilities at those nodes:

$$\hat{A}_N = \arg\max_i A_{e,i}. \tag{9}$$

2.2. Parameter estimation

Since SAASI requires the users to know the speciation, extinction, sampling and transition rates as their inputs, these rates need to be estimated. We estimate the speciation and extinction rates using the maximum likelihood method proposed by [40], where we assume that the sampling rate is known. To estimate the transition rates in the simulations, we use 'ER' model implemented in ace. The 'ER' model assumes that all the transition rates are equal. We note that in principle we could estimate input rates with BiSSE, but there is no readily available implementation allowing for non-ultrametric trees, and [41] reported some limits to BiSSE estimation for larger phylogenies.

We compared the ER model's transition rates to the truth using simulated trees (see below), and found that ace infers the transition rates accurately if there is little to no sampling difference between states. However, if state i is far less sampled than state j, ace overestimates the transition rates from other states j to state i (q_{ji}) and underestimates transition rates from state i to other states j (q_{ij}). The error depends on the sampling ratios. Therefore, we also test how transition rate adjustments affect our inferences, scaling q_{ij} and q_{ji} according to our simulated findings.

Notation	Description	
λ_{ijk}	The rate at which an parent of type i speciates giving rise to a left-hand descendant lineage of type j and right-hand descendant lineage of type k, which defines the 3D array Λ.	
μ_i	Extinction rate of type i, which defines the vector $\vec{\mu}$.	
ψ_i	Sampling rate of type i, which defines the vector $\vec{\Psi}$.	
q_{ij}	Transition rate from type i to type j, which defines the matrix Q.	
$\hat{q}_{ij,c}$	Adjusting transition rate of state i by a factor of c.	
π_i	Prior probability that root is in state i.	
$\mathcal{T}$	Tree topology.	
T	Tree height.	
τ	Time from the present day to the root (backward in time).	
t	Time from the root to the present day (forward in time).	
$\mathcal{T}(\tau)$	The subtree descending from the edge e at time τ.	
$\mathcal{T}_e^c(\tau)$	'Ancestral complement' of the edge e at time τ.	
N	A node in the observed tree consisting of either a sampled tip (a terminal node) or a speciation event (internal node).	
$A_{N_{\text{True}}}$	The true simulated state at node N.	
E_{absolute}	The absolute accuracy.	
$E_{\text{probability}}$	The probability accuracy.	
Quantity	**Description**	**Equation**
$D_{e,i}(\tau)$	The probability that an edge e that is in state i at time τ before the present day gives rise to the observed descendants.	Eq. 3
$E_i(\tau)$	The probability that an edge that is in state i at time τ has no observed descendants.	Eq. 4
$\tilde{D}_{e,i}(t)$	The probability that an edge e is in state i at time t given the state of its immediate ancestor (and the rest of the ancestral complement tree) at time t.	Eq. 5
$A_{e,i}(\tau)$	The probability that an edge e is in state i at time τ.	Eq. 8
$\hat{A}_N$	The "reconstructed state" of the node N.	Eq. 9

Table 1: List of notation, derived quantities and corresponding equations. Throughout, time τ is measured backwards in time from the present day ($\tau = 0$) to the root ($\tau = T$) of the tree. Note that the tree $\mathcal{T}$ includes information on the states at the tips of the tree. Similarly, $\mathcal{T}_e$ includes the tip states of the descendants of edge e and the ancestral complement $\mathcal{T}_e^C$ includes information on the states at all the tips in the tree except those that descend from edge e.

For state i, we simultaneously adjust $\hat{q}_{ij,c}$ and $\hat{q}_{ji,c}$: $\hat{q}_{ij,c} = cq_{ij}^{ace}$ and $\hat{q}_{ji,c} = \frac{1}{c}q_{ji}^{ace}$, where c is the multiplicative factor that adjusts the transition rates and q_{ij}^{ace} is the transition rate estimated from 'ace'. We refer to this transition rate adjustment as "adjusting state i by a factor of c".

2.3. Simulation tests

A. Single demonstration tree We first use an illustrative example in which we compare ancestral state inference using the SAASI versus ace on a simulated tree with binary states. Specifically, we consider a case in which the states differ only in their sampling rates, with State 1 is sampled at one-tenth the rate of State 2. All other diversification rates are identical between the states ($\lambda_1 = \lambda_2 = 1$, $\mu_1 = \mu_2 = 0.045$, $q_{12} = q_{21} = 0.05$ per unit time). We assume that the root is in State 1, reflecting, for example, a scenario where the place of origin of an outbreak has a lower sampling rate than other locations. The resulting tree has 87 tips, of which only 9 (10.3%) are in State 1 (Figure 1, A). We perform SAASI using the true parameters as inputs and ace using the equal rate ('ER') model. We further examine how our accuracy would change if the sampling rates are mis-specified, by varying the sampling rates ψ_1 and ψ_2 (see Figure S2). We consider different sampling ratios $\frac{\psi_1}{\psi_2}$, where the ratio ranges from 0.1 (the true sampling rates) to 2.0 (State 2 is incorrectly assumed to be sampled at one-half the rate of State 1).

B. 100 Simulated trees (SAASI with equal rates) To explore the general performance of SAASI beyond the illustrative simulated example, we simulated 100 trees with binary states using the parameters described above. We perform ancestral state inference for each tree using SAASI (true parameters as inputs), SAASI (estimated parameters as inputs), ace and simmap. We assume equal rates for ace and simmap methods. The parameters are the same as listed in **A**.

C. 100 simulated trees (SAASI with adjusted rates) We also simulated 100 trees with unequal sampling rates and equal transition rates. We then estimated the transition rates using ace, and used the adjusted transition rates to perform SAASI.

2.4. Accuracy comparison

We test the performance by comparing the SAASI reconstruction against the two standard methods: maximum likelihood (ace) and stochastic character mapping (simmap). We use simulated phylogenetic trees for which the true ancestral states are known. These trees are simulated under the birth-death-sampling model, a modification of the framework in [12]. We consider two measures of accuracy: "absolute accuracy" E_{absolute} and "probability accuracy" $E_{\text{probability}}$. Absolute accuracy is the fraction of internal states that are inferred correctly, when each node's inferred state is its highest-probability state (i for which $A_{e,i}(\tau)$ in Eq. (8) is maximized). In contrast, probability accuracy weights each node by the inferred probabilities of the true ancestral state: $\frac{\sum_N A_{N_{\text{true}}}}{|N|}$, where $A_{N_{\text{true}}}$ is the inferred probability of the true state at node N and we sum over internal nodes. For example, if a tree had only one internal node and its true state was 'state 1', and Eq. (8) gave state 1 a probability of 0.51, the absolute accuracy for that tree would be 1, and the probability accuracy would be 0.51.

2.5. Application to avian influenza H5N1

We apply SAASI to a timed phylogeny from the highly pathogenic outbreak of avian influenza A (H5N1; hemagglutinin (HA) segment) in US dairy cattle [2]. The sample collection dates ranged from April 2023 to April 2024, with 104 sequences. We use this phylogeny to illustrate how accounting for sampling differences between wild birds and other species can

affect inferences about the initial spillover event and the subsequent transmissions among host species. Since the influenza virus spread both geographically and across multiple host species, we use SAASI to compare reconstructions of host movements and geographical movements under several plausible sampling adjustments.

This particular phylogeny (Figure S2 in [2]; `mcc-gtrg-ucld-gmrf-colored.pruned.tre`) is a portion of a Bayesian time-calibrated phylogeny. It contains sequences from 6 taxonomic groups (cattle, (wild) mammal, domestic mammal, poultry, human, and wild bird) and 12 sampling locations (California, Indiana, Kansas, Maryland, Michigan, Minnesota, Montana, New Mexico, North Carolina, Ohio, Oklahoma, Texas and Wisconsin). We excluded the single human sample from our analysis. We also combined mammal and domestic mammal samples into one group.

In their study [2], the authors inferred that the H5N1 cattle outbreak was likely due to a single spillover event from wild birds to cattle, followed by transmission within the cattle and subsequent transmission between species. They estimated that the first spillover event occurred on December 9, 2023, with 95% highest posterior density (HPD) between October 12, 2023 and January 26, 2024. Their spatial and host phylogeographic analysis was conducted using BEAST using BSSVS [5, 26].

We estimate the time of the first spillover event from the wild bird population to cattle and determine how robust the time estimate is to the relative sampling. We note that in our context, "sampling" refers to the fraction of infections that are represented in the data (as opposed to the fraction of the population that is sampled, or the number of samples). Accordingly, wild bird infections (wild birds being the normal host population for avian influenza viruses) may have a much lower sampling rate than, for example, infections in domestic cattle.

H5N1 host analysis We consider three scenarios to capture different levels of sampling bias. First, we assume no significant difference in sampling between wild birds and other species. Although this scenario is unlikely, it serves as a natural baseline and reflects the assumption commonly made in phylogeographic methods such as ace or simmap. In the second scenario, we introduce a moderate sampling bias, modelling wild bird infections as being sampled ten times less frequently than H5N1 infections in other species. Finally, the third scenario reflects a severe sampling bias, in which wild bird infections are sampled one hundred times less frequently than infections in other species. Wild bird populations are challenging to sample, and the probability of sampling any given H5N1 infection (compared to that in cattle) is not known. This motivates our wide range of sampling differences.

We then explore the inferred ancestral state of the clade of interest and the reconstructed inter-species transmission events under the different sampling rates.

We adapted the estimated transition rates from [42]. The mean transition rate between non-wild bird species was 2 transitions per year, while the transition rate from wild birds to other species was higher, with a mean of 4 transitions per year. We further assume that the transition from the wild bird to the other species is two times higher (than this mean) and that the transition from other species to the wild bird is two times lower than this mean, informed by our simulation results for how sampling affects estimated transition rates. Hence, we obtain the

following transition rate matrix:

	Cattle	Mammal	Poultry	Wild bird
Cattle	–	2	2	1
Mammal	2	–	2	1
Poultry	2	2	–	1
Wild bird	4	4	4	–

These transition rates are similar to those estimated using the ace symmetric transition rate, which has a mean of 3.68 per year. We assume that the speciation and extinction rates for all species are the same, with estimated rates of 21.1 and 6.8 per year, respectively.

H5N1 geographic analysis We use geographic locations to explore the impact of different geographic sampling rates on inferences about the likely geographic origin of the outbreak, focusing on the geographic location of the transmission event(s) from wild birds to cattle. Given that most of the samples were collected in Texas, we consider two alternative sampling models in the phylogeographic analysis. In the first, baseline model, we assume no significant difference in sampling effort between Texas and other states; that is, all states are modeled as having the same sampling rate. In the second model, we account for sampling bias by assuming that Texas was sampled five times more frequently than other states. Although we do not know the true relative sampling differences between the locations, this analysis will illustrate how sensitive the inferred geographic location of the spillover event is to sampling differences.

In this analysis, because the transition rates are unknown and difficult to estimate, we model that the transition rates are equal between states, and estimate them with 'ace using the 'ER' model.

2.6. Data and Materials

The data and code used in this study can be found in:
https://github.com/yexuansong/Sampling-Aware-Ancestral-State-Inference.git. SAASI is available as an R package at https://github.com/MAGPIE-SFU/saasi. Given the difficulty of specifying a full λ_{ijk} matrix, the current implementation is as in Equation (6) ($\lambda_{ijk} = 0$ if $i \neq j,k$).

3. Results

3.1. Simulation study

If the sampling rate is known (State 1 is sampled at one-tenth the rate of State 2), SAASI correctly infers most of the internal nodes in the simulated trees with $E_{\text{absolute}} = 0.98$ and a probability accuracy of $E_{\text{probability}} = 0.97$. Figure 1 shows the single demonstration tree (simulation test A) and the comparison between SAASI and ace. We find that ace infers many internal nodes incorrectly, with $E_{\text{absolute}} \approx 0.5$ and $E_{\text{probability}} = 0.5$. Furthermore, ace infers that the root state is in State 2, while the true root state is State 1. In this example simulation, if there is no sampling difference between the two states (Figure S1 middle panel), SAASI's reconstruction is very similar to that of ace and simmap.

Figure 2 shows accuracy comparisons for 100 simulated trees (simulation test B). We find that SAASI reliably has higher absolute and probability accuracies than both ace and simmap where sampling differences exist. The median absolute and probability accuracies using SAASI (with known rates) and SAASI with estimated rates are over 0.9 compared to 0.55 using ace

and simmap. We explored the effect of mis-specifying the sampling ratio in SAASI, and found that SAASI's both absolute and probability accuracy decreases as the sampling ratio deviates from the true sampling ratio $\frac{\psi_1}{\psi_2} = 0.1$ to $\frac{\psi_1}{\psi_2} = 2$ (Figure S2). If the sampling ratios are close to the true sampling ratio ($\frac{\psi_1}{\psi_2} = 0.1$ to $\frac{\psi_1}{\psi_2} = 0.5$), both $E_{\text{probability}}$ and $E_{\text{probability}}$ are greater than 0.75. However, as the sampling ratio is mis-specified ($\frac{\psi_1}{\psi_2} > 1$), both $E_{\text{absolute}} \approx 0.5$ and $E_{\text{probability}} = 0.5$, comparable to that of reconstructions using ace.

Figure S3 shows the results of simulation test C, with transition rate adjustments. We find that SAASI has higher accuracies than ace across a range of transition rate adjustments (using estimated transition rates from ace; adjusted transition rates of state 1 by a factor of $2, 3, 5, 10$). The accuracy of our method is not sensitive to the adjustments on transition rates.

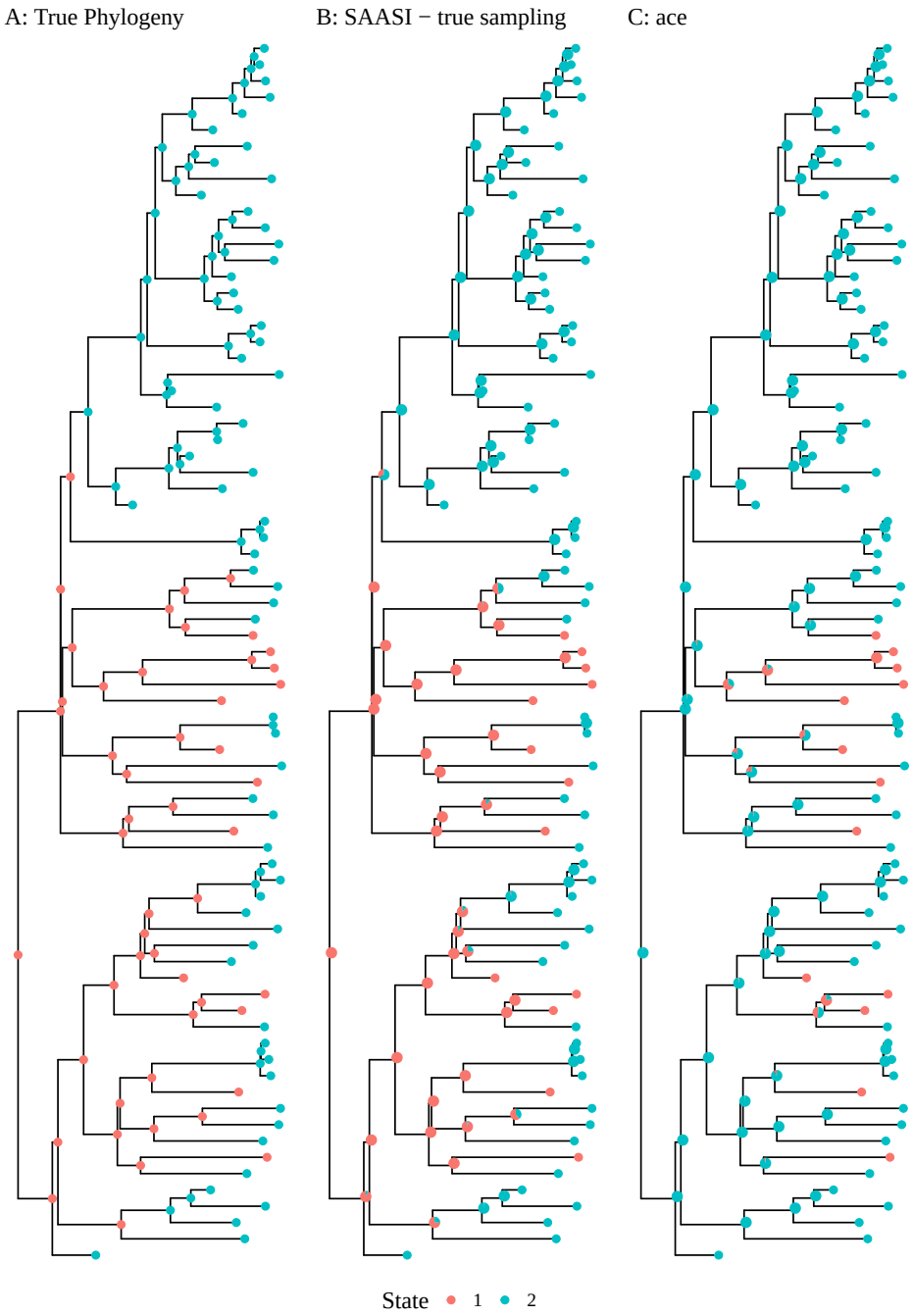

Figure 1: Ancestral state inference using SAASI and ace. A: Simulated tree with known transmission histories; B: SAASI with true parameter values; C: ace with 'ER' transition model. Pie charts indicate the inferred probabilities of being in particular states. The tree was generated with the following parameters: $\lambda_1 = \lambda_2 = 1$, $\mu_1 = \mu_2 = 0.045$, $q_{12} = q_{21} = 0.05$, and $\psi_1 = 0.05, \psi_2 = 0.5$.

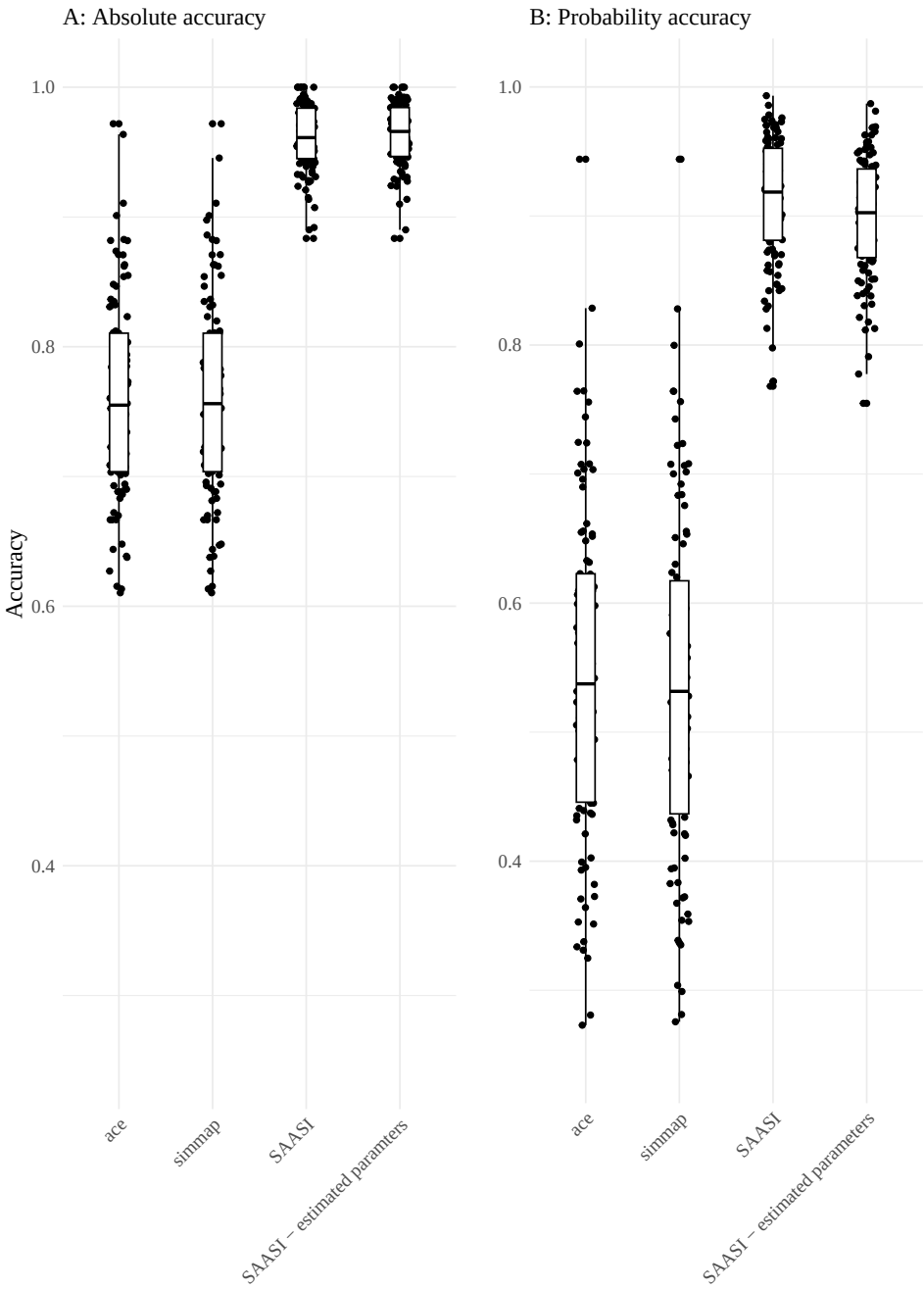

Figure 2: Accuracy of ancestral state inference methods (100 simulated trees in simulation test B) A: Absolute accuracy, defined as the fraction of correctly inferred node states; B: Probability accuracy, accounting for uncertainty in the node inference. We use 'ER' transition model in ace and simmap. The trees were generated using the following parameters: $\lambda_1 = \lambda_2 = 1$, $\mu_1 = \mu_2 = 0.045$, $q_{12} = q_{21} = 0.05$, and $\psi_1 = 0.05, \psi_2 = 0.5$.

3.2. Avian influenza

Figure 3 shows the difference between the three assumptions for wild bird sampling. Adjusting for lower sampling of wild bird infections, we find more than one transmission from wild birds to cattle. There are 10 transmissions from wild birds to cattle when wild birds are at

one-tenth sampling and 12 transmissions when wild birds are at one-one hundredth sampling. These are inferred to have occurred from late January to early February 2024 (Figure 3) B,C, red dashed line), slightly later than the previous estimate [2]. In contrast, if we model no sampling difference between species, both ace and SAASI cannot identify the key transmission events from wild bird to cattle (Figure S4).

We explored the number of movements (or lack of movements) from parent to child pairs of nodes in the phylogeny, and illustrated these with alluvial plots (see Figure 3 panels D–F). This allows us to probe not just where a particular node was located, but to ask about the nature of viral movements between taxonomic groups. We find that cattle are the origin species for most host species jumps, under all sampling models we explored. We find that adjusting for under-sampling of wild bird infections leads inferring that more nodes of the phylogeny are in wild birds, compared to scenarios with more even sampling. This, in turn, means fewer inferred transitions from other populations into wild birds (see Figure 3 D compared to E and F). In particular, without accounting for sampling, standard methods infer transitions from both other mammals and poultry into wild birds (as well as cattle). With sampling adjustment, only cattle infections were re-introduced to wild birds.

We also find more transitions from wild birds into other host species when we account for lower sampling in wild birds, particularly more transitions from wild birds to cattle. These results is also robust to the specific transition rate adjustment (see Figure S5): the key transmission event from wild bird to cattle is estimated to have occurred between late January and early February 2024 (Figure S5 B, C), red dashed line) even with equal assumed transition rates.

Figure 4 shows the results of our H5N1 geographic analysis in which we examined the inferred geographic origin of the clade predominantly in cattle. If we model that there is no sampling difference between jurisdictions (left) compared to modelling a factor of 5 sampling difference between Texas and other states (right) in Figure 4, both models suggest that the first spillover event from the wild bird population to cattle occurred in Texas. If we model that there is no sampling difference between states, SAASI would infer that it is highly likely that a wild bird moved from Texas to New Mexico, leading to the emergence of a predominantly New Mexico clade in March 2024. In contrast, if we model that infections in Texas were sampled at a rate 5 times higher than other states, SAASI would suggest that there was a wild bird movement from New Mexico to Texas between November 2023 and January 2024 (red dashed line). In either case, our analyses show that the virus rapidly spread from Texas to other states in March 2024, including Ohio, Kansas, Michigan, and New Mexico. Overall, the geographic reconstructions are robust to a five times sampling difference between Texas and other states. Phylogeographic analysis in [2] also inferred significant interstate movement, with confirmed transitions from Texas to Kansas, New Mexico, and Michigan. Epidemiological records also documented the movements of infected cattle from a Texas herd to North Carolina and Idaho [43].

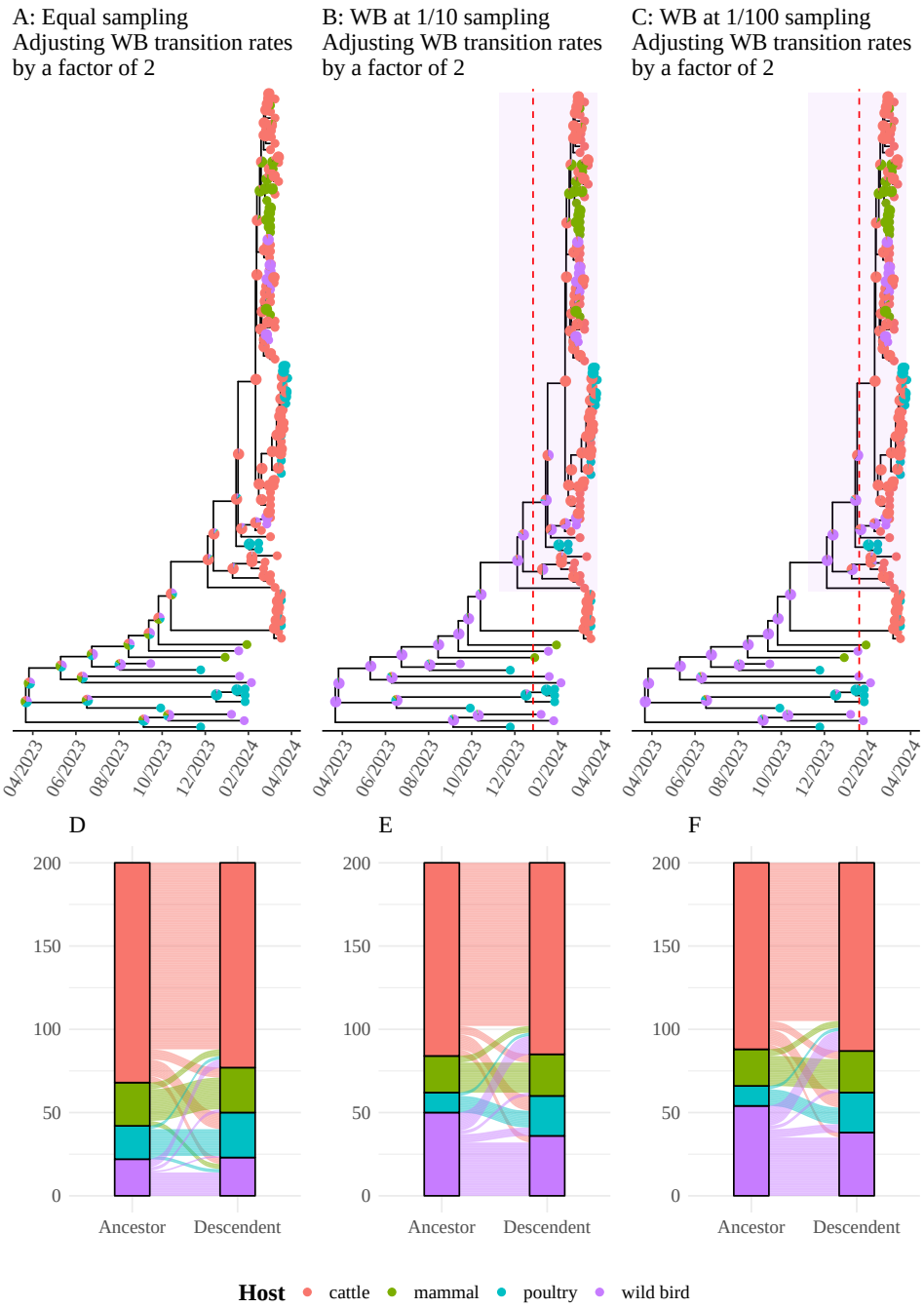

Figure 3: Ancestral state inference of the H5N1 HA segment tree using SAASI under different species-level sampling models. WB: wild birds. A: Inferred species hosts under equal sampling rates; B: wild birds at one-tenth sampling; C: wild birds at one-one hundredth sampling; D: Inferred viral transitions between host species in A; E: Inferred viral transitions between host species in B; F: Inferred viral transitions between host species in C. Pie charts indicate the inferred probabilities of being in particular states. Transition rates of the wild bird population are adjusted by a factor of 2 ($c = 2$). The dashed red line indicates the key transition event from wild bird to cattle. WB refers to the wild bird population.

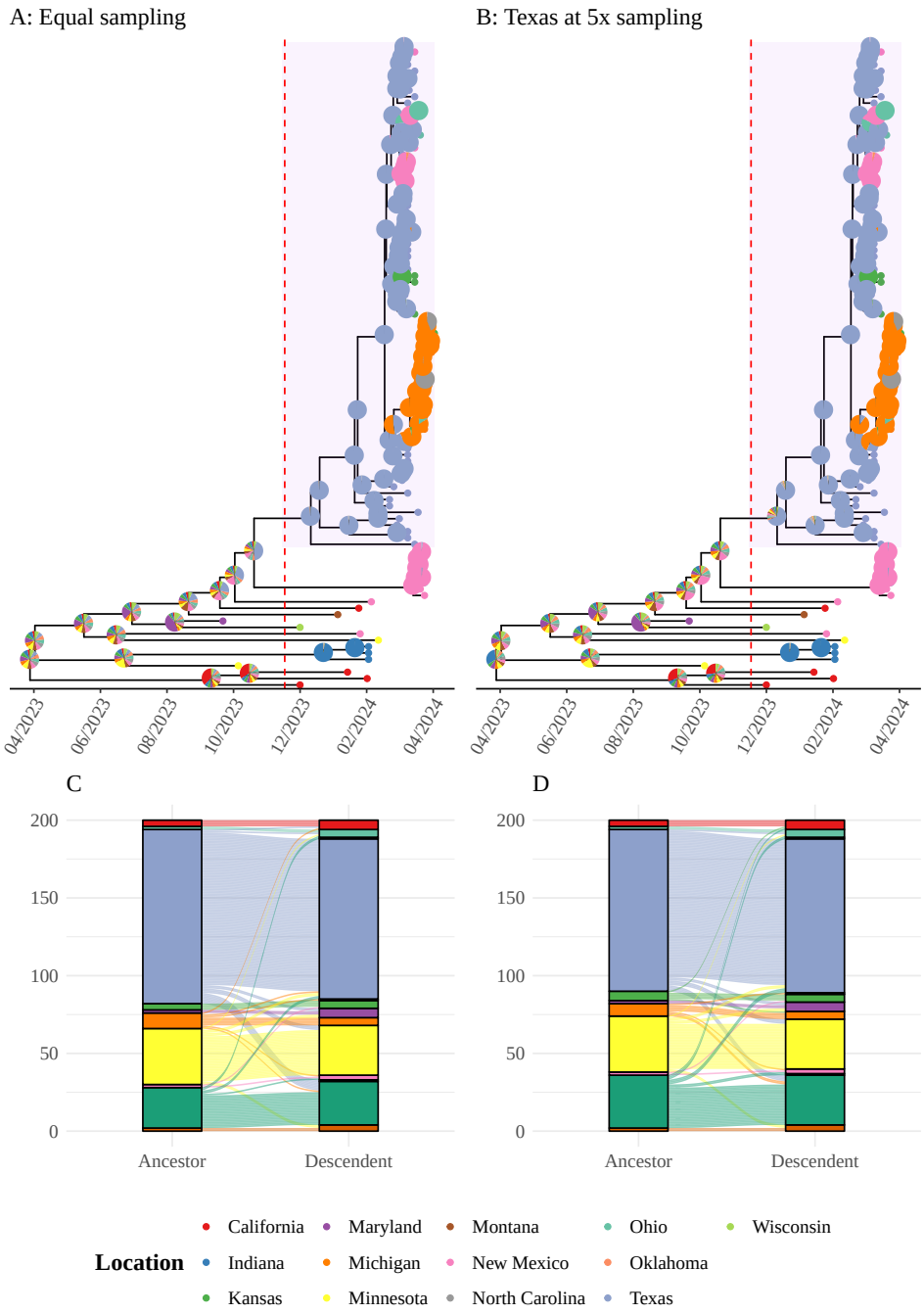

Figure 4: Phylogeographic ancestral state inference using SAASI under different sampling models for Texas and other US states. A: Inferred geographic locations under equal sampling; B: Texas at five times sampling. C: Inferred viral transitions between geographic locations in A; D: Inferred viral transitions between geographic locations in B. Pie charts indicate the inferred probabilities of being in ancestral states. The transition rates between states are modelled as equal and estimated using ace. The dashed red line indicates the estimated time it takes for the virus to move to Texas.

3.3. Scalability to larger trees

SAASI is scalable for trees that contain more than 100,000 tips. We estimate that the relation between the tree size and the running time (in seconds) is linear. For trees with 100,000 tips, SAASI completes in approximately 1 hour (3500 seconds; see Figure S6). For a 10,000-tip tree with two states and a substantial sampling difference (state 1 is sampled 50 times less than state 2), SAASI has absolute accuracy $E_{\text{absolute}} = 0.82$, compared to $E_{\text{absolute}} = 0.52$using ace. We expect the running time to increase if there are more states.

4. Discussion

We have introduced and implemented a fast approach for sampling-aware ancestral inference, SAASI, which is amenable to a range of viral phylogeography applications, including reconstructing the ancestral host species or geographic locations in relatively large phylogenies. SAASI attains a high accuracy, correctly accounting for sampling bias in simulated data, in contrast to standard methods, which are negatively impacted by sampling differences among the states. We applied SAASI to the recent outbreak of avian influenza H5N1, using 104 sequences from six host species in 12 geographic locations [2]. While their approach inferred host species transition events using Bayesian discrete trait analysis, our inference method provides a complementary framework that accounts for sampling biases. We find that SAASI is able to identify the time of transition from wild birds to cattle (from late January to early Feburary 2024), and that the root node for the clade that was predominantly in cattle was in wild birds. In contrast, without accounting for sampling differences, the date of movement out of wild birds would be uncertain. Accounting for sampling differences also impacts the inferred numbers of host species jumps, particularly from and to wild birds, but also among other species. Overall, our H5N1 results highlight the importance of obtaining information about the rate of sampling, and of incorporating sampling differences into ancestral state reconstructions.

SAASI has important limitations. First, the sampling difference under consideration is modelled as known (if it is not known, then SAASI offers an approach to sensitivity analysis). If attempts were made to estimate sampling bias (particularly alongside transition rates, state-dependent branching and death rates, and states of internal nodes), it is likely that these would collectively not be identifiable [44]. For example, having fewer taxa with a trait (and fewer branching events with an apparent lower rate) could be a result of either a lower branching rate or lower sampling intensity. There are other natural trade-offs that would affect the simultaneous estimation of the complete set of parameters. Furthermore, SAASI does not (yet) estimate other relevant parameters; in our application these were estimated separately based on either birth-death models or ace's estimates with appropriate adjustments. These could in principle be estimated in a first pass using one of the BiSSE [12] family of models without ancestral state reconstruction, but at present these require ultrametric trees and so are not suited for longitudinally sampled pathogen datasets.

We have focused on sampling differences, since the fraction of infections that are sampled is likely to vary by location and host species, and this is a recognized challenge for phylogeographic reconstructions [29]. We note that evolutionary parameters, such as the molecular clock, substitution model, as well as the branching and extinction rate, are likely to change from host to host, and rapid adaptation with selection may occur immediately following a host jump. These phenomena are challenging to include in phylogenetic inference. We used a fixed, timed phylogenetic tree because our aim is to account for sampling differences in phylogeographic reconstructions at large scales.

In recent work, Vaughan and Stadler [44] also use the previous work of Freyman and Hohna [21], in their case to develop Bayesian inference of multi-type population trajectories. Their algorithms jointly infer the phylogenetic tree, multi-type birth-death model parameters (our state-dependent branching, death and sampling rates), ancestral node states and type-specific population trajectories. Potentially unidentifiable combinations can be managed in Bayesian analyses (through priors, and through sampling a posterior, which may have some correlations or structure but in any case reflects the estimated uncertainty and is the posterior from which samples are desired). The sampling is done with a combination of particle filtering and Markov Chain Monte Carlo (MCMC) sampling. In contrast to Bayesian phylogenetic approaches, SAASI is intended to operate at very large scales and to provide rapid estimates of the ancestral node states in a way that accounts for known or suspected sampling differences. This focus is motivated in part by the emergence of large-scale genomic surveillance efforts and phylogeographic analyses, in particular for SARS-CoV-2 [33, 34]but also for a wide range of pathogens and organisms including influenza viruses [33, 45, 46] and other pathogens [47, 48] In simulations, SAASI obtains a high accuracy, and will be broadly relevant, as testing and sequencing policies vary from one jurisdiction to another. Despite advances in Bayesian methods, the current state of the art in phylogeography at large scales is to down-sample the data to approximately obtain representative sampling, and re-run analyses using tools like ace [34, 49, 50]. Repeat analyses may be time-consuming, give variable results, and necessitate discarding data in each analysis. This approach also makes strong assumptions, in particular that the traits (here, host species or locations) do not affect the branching, sampling or death rates. SAASI is less restrictive. It offers quick, robust and principled phylogeographic reconstructions that account for sampling bias.

Acknowledgements

This work was supported by grant to Dr. William Hsiao at Simon Fraser University, from the Public Health Agency of Canada (Arrangement: 2223-HQ-000265) Dr. Hsiao's salary was partially supported a Michael Smith Health Research BC Scholar Award. This work is supported by NSERC (CC; IG; YS: CANMOD, the Canadian Network for Modelling Infectious Disease, 560516-2020; AM: CRC-2021-00276 and RGPIN-2022-03113; CC: RGPIN-2019-06624 and), the Canada's 150 Research Chair program, and the Michael Smith Foundation for Health Research Scholar Program. We emphasize our strong appreciation for the authors of Nguyen et al [2], the Flu crew at the U.S. National Animal Disease Center and the authors of the github repository at `https://github.com/flu-crew/dairy-cattle-hpai-2024`, who published their H5N1 repository under a license permitting reuse and publication without restriction (among other permissions). Such data sharing is essential for methods development efforts like ours, and we greatly appreciate it.

References

[1] M. J. Ward, C. L. Gibbons, P. R. McAdam, et al. Time-Scaled Evolutionary Analysis of the Transmission and Antibiotic Resistance Dynamics of Staphylococcus aureus Clonal Complex 398. *Applied and Environmental Microbiology*, 80(23):7275–7282, December 2014. Publisher: American Society for Microbiology.

[2] Thao-Quyen Nguyen, Carl Hutter, Alexey Markin, et al. Emergence and interstate spread of highly pathogenic avian influenza A(H5N1) in dairy cattle, May 2024. Pages: 2024.05.01.591751 Section: New Results.

[3] Cedric C. S. Tan, Lucy van Dorp, and Francois Balloux. The evolutionary drivers and correlates of viral host jumps. *Nature Ecology & Evolution*, 8(5):960–971, May 2024. Publisher: Nature Publishing Group.

[4] Nicola De Maio, Chieh-Hsi Wu, Kathleen M. O'Reilly, and Daniel Wilson. New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation. *PLOS Genetics*, 11(8):e1005421, August 2015. Publisher: Public Library of Science.

[5] Philippe Lemey, Andrew Rambaut, Alexei J. Drummond, and Marc A. Suchard. Bayesian Phylogeography Finds Its Roots. *PLOS Computational Biology*, 5(9):e1000520, September 2009. Publisher: Public Library of Science.

[6] Miles W. Carroll, David A. Matthews, Julian A. Hiscox, et al. Temporal and spatial analysis of the 2014–2015 Ebola virus outbreak in West Africa. *Nature*, 524(7563):97–101, August 2015. Publisher: Nature Publishing Group.

[7] José Maria Gonzalez-Alba, Fernando Baquero, Rafael Cantón, and Juan Carlos Galán. Stratified reconstruction of ancestral Escherichia coli diversification. *BMC Genomics*, 20(1):936, December 2019.

[8] Jason C. Hyun and Bernhard O. Palsson. Reconstruction of the last bacterial common ancestor from 183 pangenomes reveals a versatile ancient core genome. *Genome Biology*, 24(1):183, August 2023.

[9] R. Alexander Pyron, Frank T. Burbrink, and John J. Wiens. A phylogeny and revised classification of Squamata, including 4161 species of lizards and snakes. *BMC Evolutionary Biology*, 13(1):93, April 2013.

[10] Ivan Gomez-Mestre, Robert Alexander Pyron, and John J. Wiens. Phylogenetic analyses reveal unexpected patterns in the evolution of reproductive modes in frogs. *Evolution; International Journal of Organic Evolution*, 66(12):3687–3700, December 2012.

[11] Angela McLaughlin, Vincent Montoya, Rachel L. Miller, Canadian COVID-19 Genomics Network (CanCOGeN) Consortium, Michael Worobey, and Jeffrey B. Joy. Effectiveness of Canadian travel restrictions in reducing burden of SARS-CoV-2 variants of concern, September 2023. ISSN: 2329-4140 Pages: 2023.09.12.23294140.

[12] Wayne P. Maddison, Peter E. Midford, and Sarah P. Otto. Estimating a Binary Character's Effect on Speciation and Extinction. *Systematic Biology*, 56(5):701–710, October 2007.

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

[14] Jeremy M. Beaulieu and Brian C. O'Meara. Detecting Hidden Diversification Shifts in Models of Trait-Dependent Speciation and Extinction. *Systematic Biology*, 65(4):583–601, July 2016.

[15] Emma E. Goldberg and Boris Igić. TEMPO AND MODE IN PLANT BREEDING SYSTEM EVOLUTION. *Evolution*, 66(12):3701–3709, December 2012.

[16] Sebastian Höhna, Michael J. Landis, Tracy A. Heath, Bastien Boussau, Nicolas Lartillot, Brian R. Moore, John P. Huelsenbeck, and Fredrik Ronquist. RevBayes: Bayesian Phylogenetic Inference Using Graphical Models and an Interactive Model-Specification Language. *Systematic Biology*, 65(4):726–736, July 2016.

[17] R. Nielsen and J. P. Huelsenbeck. Detecting positively selected amino acid sites using posterior predictive p-values. In *Biocomputing 2002*, pages 576–588. WORLD SCIENTIFIC, December 2001.

[18] Rasmus Nielsen. Mapping Mutations on Phylogenies. *Systematic Biology*, 51(5):729–739, September 2002.

[19] John P. Huelsenbeck, Rasmus Nielsen, and Jonathan P. Bollback. Stochastic Mapping of Morphological Characters. *Systematic Biology*, 52(2):131–158, April 2003.

[20] Jonathan P. Bollback. SIMMAP: Stochastic character mapping of discrete traits on phylogenies. *BMC Bioinformatics*, 7(1):88, February 2006.

[21] William A Freyman and Sebastian Höhna. Stochastic Character Mapping of State-Dependent Diversification Reveals the Tempo of Evolutionary Decline in Self-Compatible Onagraceae Lineages. *Systematic Biology*, 68(3):505–519, May 2019.

[22] Mark Pagel. Detecting Correlated Evolution on Phylogenies: A General Method for the Comparative Analysis of Discrete Characters. *Proceedings: Biological Sciences*, 255(1342):37–45, 1994. Publisher: Royal Society.

[23] THOMAS H. Jukes and CHARLES R. Cantor. CHAPTER 24 - Evolution of Protein Molecules. In H. N. Munro, editor, *Mammalian Protein Metabolism*, pages 21–132. Academic Press, January 1969.

[24] Masami Hasegawa, Hirohisa Kishino, and Taka-aki Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22(2):160–174, October 1985.

[25] Emmanuel Paradis and Klaus Schliep. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35:526–528, 2019.

[26] Nuno Rodrigues Faria, Marc A Suchard, Andrew Rambaut, Daniel G Streicker, and Philippe Lemey. Simultaneously reconstructing viral cross-species transmission history and identifying the underlying constraints. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 368(1614):20120196, March 2013.

[27] Nicola De Maio, Chieh-Hsi Wu, Kathleen M O'Reilly, and Daniel Wilson. New Routes to Phylogeography:

A Bayesian Structured Coalescent Approximation. *PLoS Genet.*, 11(8):e1005421, August 2015.

[28] Nicola F Müller, David Rasmussen, and Tanja Stadler. MASCOT: parameter and state inference under the marginal structured coalescent approximation. *Bioinformatics*, 34(22):3843–3848, November 2018.

[29] Pengyu Liu, Yexuan Song, Caroline Colijn, and Ailene MacPherson. The impact of sampling bias on viral phylogeographic reconstruction. *PLOS Global Public Health*, 2(9):e0000577, September 2022. Publisher: Public Library of Science.

[30] Daniel Magee, Marc A. Suchard, and Matthew Scotch. Bayesian phylogeography of influenza A/H3N2 for the 2014-15 season in the United States using three frameworks of ancestral state reconstruction. *PLOS Computational Biology*, 13(2):e1005389, February 2017. Publisher: Public Library of Science.

[31] Stéphane Guindon and Nicola De Maio. Accounting for spatial sampling patterns in Bayesian phylogeography. *Proceedings of the National Academy of Sciences*, 118(52):e2105273118, December 2021. Publisher: Proceedings of the National Academy of Sciences.

[32] Maylis Layan, Nicola F Müller, Simon Dellicour, Nicola De Maio, Hervé Bourhy, Simon Cauchemez, and Guy Baele. Impact and mitigation of sampling bias to determine viral spread: Evaluating discrete phylogeography through CTMC modeling and structured coalescent model approximations. *Virus Evol.*, 9(1):vead010, February 2023.

[33] 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(23):4121–4123, December 2018.

[34] Angela McLaughlin, Vincent Montoya, Rachel L Miller, Gideon J Mordecai, Canadian COVID-19 Genomics Network (CanCOGen) Consortium, Michael Worobey, Art FY Poon, and Jeffrey B Joy. Genomic epidemiology of the first two waves of SARS-CoV-2 in Canada. *eLife*, 11:e73896, August 2022. Publisher: eLife Sciences Publications, Ltd.

[35] Simon Dellicour, Samuel L Hong, Bram Vrancken, Antoine Chaillon, Mandev S Gill, Matthew T Maurano, Sitharam Ramaswami, Paul Zappile, Christian Marier, Gordon W Harkins, Guy Baele, Ralf Duerr, and Adriana Heguy. Dispersal dynamics of SARS-CoV-2 lineages during the first epidemic wave in New York City. *PLoS Pathog.*, 17(5):e1009571, May 2021.

[36] Philippe Lemey, Andrew Rambaut, Alexei J Drummond, and Marc A Suchard. Bayesian phylogeography finds its roots. *PLoS Comput. Biol.*, 5(9):e1000520, September 2009.

[37] Antanas Kalkauskas, Umberto Perron, Yuxuan Sun, Nick Goldman, Guy Baele, Stephane Guindon, and Nicola De Maio. Sampling bias and model choice in continuous phylogeography: Getting lost on a random walk. *PLoS Comput. Biol.*, 17(1):e1008561, January 2021.

[38] Ailene MacPherson, Stilianos Louca, Angela McLaughlin, Jeffrey B Joy, and Matthew W Pennell. Unifying Phylogenetic Birth–Death Models in Epidemiology and Macroevolution. *Systematic Biology*, 71(1):172–189, January 2022.

[39] Sean Nee. Birth-Death Models in Macroevolution. *Annual Review of Ecology, Evolution, and Systematics*, 37(Volume 37, 2006):1–17, December 2006. Publisher: Annual Reviews.

[40] Tanja Stadler, Roger Kouyos, Viktor von Wyl, et al. Estimating the Basic Reproductive Number from Viral Sequence Data. *Molecular Biology and Evolution*, 29(1):347–357, January 2012.

[41] Matthew P Davis, Peter E Midford, and Wayne Maddison. Exploring power and parameter estimation of the BiSSE method for analyzing species diversification. *BMC Evol. Biol.*, 13(1):38, February 2013.

[42] Lambodhar Damodaran, Anna Jaeger, and Louise H. Moncla. Intensive transmission in wild, migratory birds drove rapid geographic dissemination and repeated spillovers of H5N1 into agriculture in North America, December 2024. Pages: 2024.12.16.628739 Section: New Results.

[43] USDA, FDA and CDC Share Update on HPAI Detections in Dairy Cattle | Animal and Plant Health Inspection Service.

[44] Timothy G Vaughan and Tanja Stadler. Bayesian phylodynamic inference of multi-type population trajectories using genomic data. *Bioinformatics*, (biorxiv;2024.11.26.625381v1), December 2024.

[45] Yvonne C. F. Su, Justin Bahl, Udayan Joseph, et al. Phylodynamics of H1N1/2009 influenza reveals the transition from host adaptation to immune-driven selection. *Nature Communications*, 6(1):7952, August 2015. Publisher: Nature Publishing Group.

[46] Trevor Bedford, Steven Riley, Ian G. Barr, et al. Global circulation patterns of seasonal influenza viruses vary with antigenic drift. *Nature*, 523(7559):217–220, July 2015. Publisher: Nature Publishing Group.

[47] Nicholas J. Croucher, Jonathan A. Finkelstein, Stephen I. Pelton, et al. Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nature genetics*, 45(6):656–663, June 2013.

[48] Nathan D. Grubaugh, Jason T. Ladner, Moritz U. G. Kraemer, et al. Genomic epidemiology reveals multiple

introductions of Zika virus into the United States. *Nature*, 546(7658):401–405, June 2017.

[49] Idowu B. Olawoye, Paul E. Oluniyi, Judith U. Oguzie, et al. Emergence and spread of two SARS-CoV-2 variants of interest in Nigeria. *Nature Communications*, 14(1):811, February 2023. Publisher: Nature Publishing Group.

[50] Sana Naderi, Peter E Chen, Carmen Lia Murall, Raphael Poujol, Susanne Kraemer, Bradley S Pickering, Selena M Sagan, and B Jesse Shapiro. Zooanthroponotic transmission of SARS-CoV-2 and host-specific viral mutations revealed by genome-wide phylogenetic analysis. *eLife*, 12:e83685, April 2023. Publisher: eLife Sciences Publications, Ltd.

S1. Supplementary Material

S1.1. Derivation of the main decomposition

For clarity, we provide an explicit derivation of our central equation for the $A_{e,i}$,

$$A_{e,i}(\tau) = \Pr(Y_e(\tau) = i|\mathcal{T},\theta) = \frac{\Pr(Y_e(\tau) = i|\mathcal{T}_e^C(\tau),\theta)\Pr(\mathcal{T}_e(\tau)|Y_e(\tau) = i,\theta)}{\sum_j \Pr(Y_e(\tau) = j|\mathcal{T}_e^C(\tau),\theta)\Pr(\mathcal{T}_e(\tau)|Y_e(\tau) = j,\theta)}. \tag{S1}$$

In this derivation, we will suppress τ and θ to simplify notation. For example, we write $\Pr(\mathcal{T}) = \Pr(\mathcal{T}_e, \mathcal{T}_e^C)$, and $A_{e,i} = \Pr(Y_e = i|\mathcal{T})$.

By Bayes' theorem, we have

$$\Pr(Y_e = i|\mathcal{T}) = \frac{\Pr(\mathcal{T}|Y_e = i)\Pr(Y_e = i)}{\Pr(\mathcal{T})}.$$

Since $\mathcal{T}_e$ and $\mathcal{T}_e^C$ are conditionally independent given i (which means, given that edge e is in state i at time τ), the above is

$$Pr(Y_e = i|\mathcal{T}) = \frac{\Pr(\mathcal{T}_e|Y_e = i)\Pr(\mathcal{T}_e^C|Y_e = i)\Pr(Y_e = i))}{\Pr(\mathcal{T})}. \tag{S2}$$

Using Bayes' theorem again, we write

$$\Pr(\mathcal{T}_e^C|Y_e = i) = \Pr(Y_e = i|\mathcal{T}_e^C)\Pr(\mathcal{T}_e^C)/\Pr(Y_e = i)$$

and

$$\Pr(\mathcal{T}) = \Pr(\mathcal{T}_e, \mathcal{T}_e^C) = \Pr(\mathcal{T}_e|\mathcal{T}_e^C)\Pr(\mathcal{T}_e^C).$$

Substituting these into Eq. S2, we have

$$\Pr(Y_e = i|\mathcal{T}) = \frac{\Pr(\mathcal{T}_e|Y_e = i)\Pr(Y_e = i|\mathcal{T}_e^C)}{\Pr(\mathcal{T}_e|\mathcal{T}_e^C)}. \tag{S3}$$

It remains to show that the denominator has the form given in (1) of the main text. Since the state of edge e must be one and only one of the possible ancestral states, we have

$$\begin{aligned}\Pr(\mathcal{T}_e|\mathcal{T}_e^C) &= \sum_j \Pr(\mathcal{T}_e, Y_e = j|\mathcal{T}_e^C)\\ &= \sum_j \Pr(\mathcal{T}_e|\mathcal{T}_e^C, Y_e = j)\Pr(Y_e = j|\mathcal{T}_e^C)\\ &= \sum_j \Pr(\mathcal{T}_e|Y_e = j)\Pr(Y_e = j|\mathcal{T}_e^C)\end{aligned} \tag{S4}$$

where the second line is a conditional probability, and the last line is due to the conditional independence of $\mathcal{T}_e$ and $\mathcal{T}_e^C$ given the state of edge e.

S1.2. Supplementary Figures

Figure S1 gives an example illustrating that SAASI's result is very similar to ace's if we assume equal sampling rates. In other words, while SAASI's underlying mathematical model is quite different from that of ace, the results are very similar if the same assumption is made about sampling. The natural comparison for this figure is Figure 1 in the main text.

Figure S2 shows how the absolute and probability accuracy changes if the sampling ratio is misspecified. The accuracy is sensitive to the sampling rates. Both accuracies remain above 0.75 for sampling ratios $\psi_1/\psi_2 \in [0.15, 1]$. If the sampling rate is strongly mis-specified (with $\psi_1/\psi_2 > 1$ reflecting assuming that state 1 is *more* highly sampled than state 2, rather than less), the reconstruction accuracies decrease to a level similar to the accuracy reconstructed using ace, i.e. just over 0.55.

Figure S3 compares the accuracies obtained using different ASI methods and adjustments to the transition rates, including mis-specification of the transition rates. SAASI with true rates has the highest accuracy, and SAASI's results under various transition rate adjustments are similar to each other. This demonstrates the robustness of SAASI to transition rate misspecification.

Figure S4 compares ancestral state inference using ace and SAASI under the assumption of equal sampling in all hosts. Both inferences cannot identify the key transition events from wild birds to cattle and cannot reliably infer the internal node states before November 2023.

Figure S5 compares ancestral state inference using SAASI under the assumption that wild bird infections are sampled 100 times less than infections in other host species, with equal transition rates. The results are similar to the main text, where we adjusted the transition rates to and from wild birds to other species by a factor of 2.

Figure S6 compares SAASI's running time for different tree sizes (with two states). The running time grows linearly in the number of taxa. Increasing the number of states would also increase the run time.

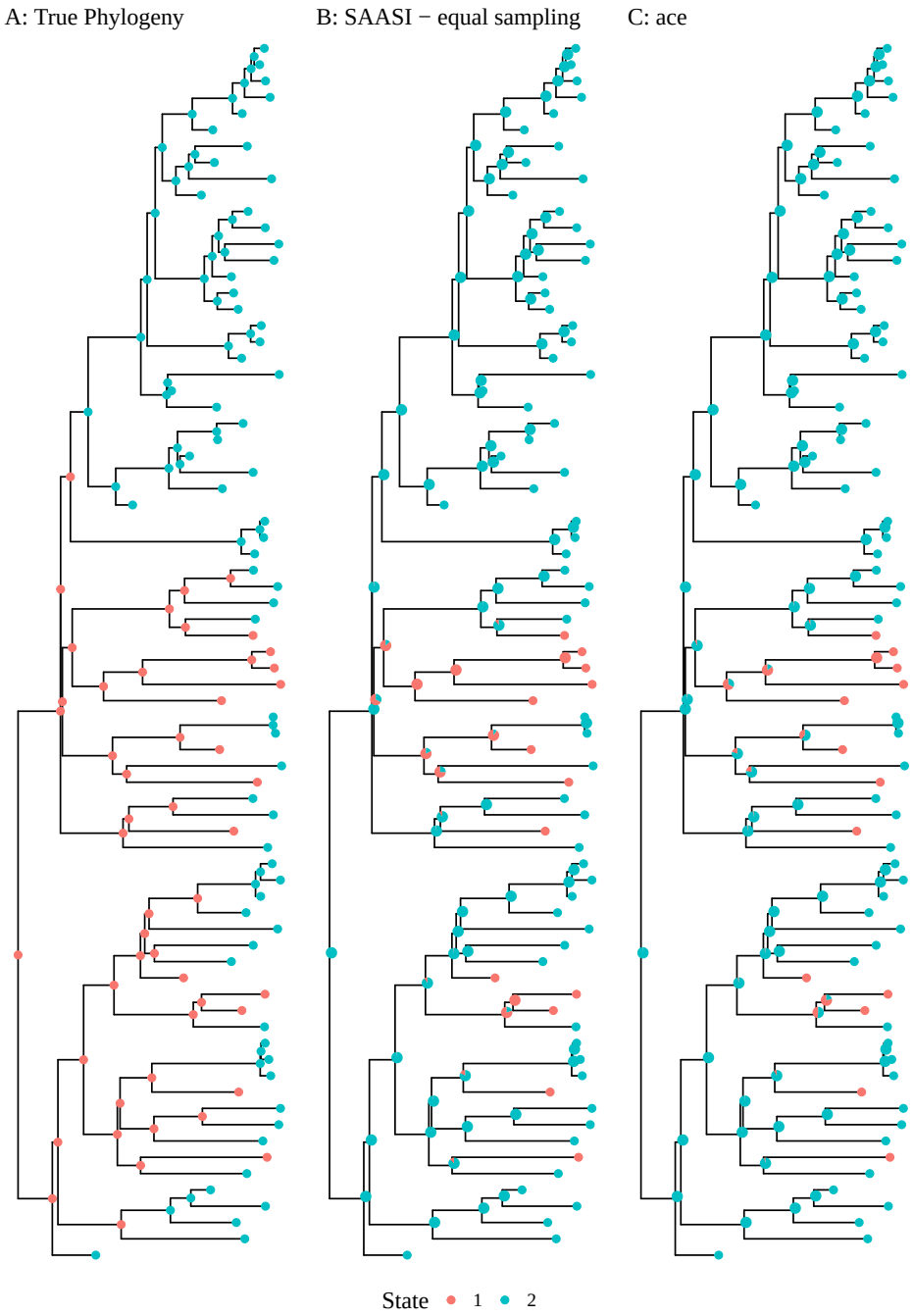

Figure S1: Ancestral state inference using SAASI under and equal sampling model and using ace. A: Simulated tree with known transmission histories; B: SAASI with equal sampling rates ($\psi_1 = \psi_2 = 0.5$); C: ace under equal sampling rate model. The tree is generated using the following parameters: $\lambda_1 = \lambda_2 = 1$, $\mu_1 = \mu_2 = 0.045$, $q_{12} = q_{21} = 0.05$, and $\psi_1 = 0.05$, $\psi_2 = 0.5$.

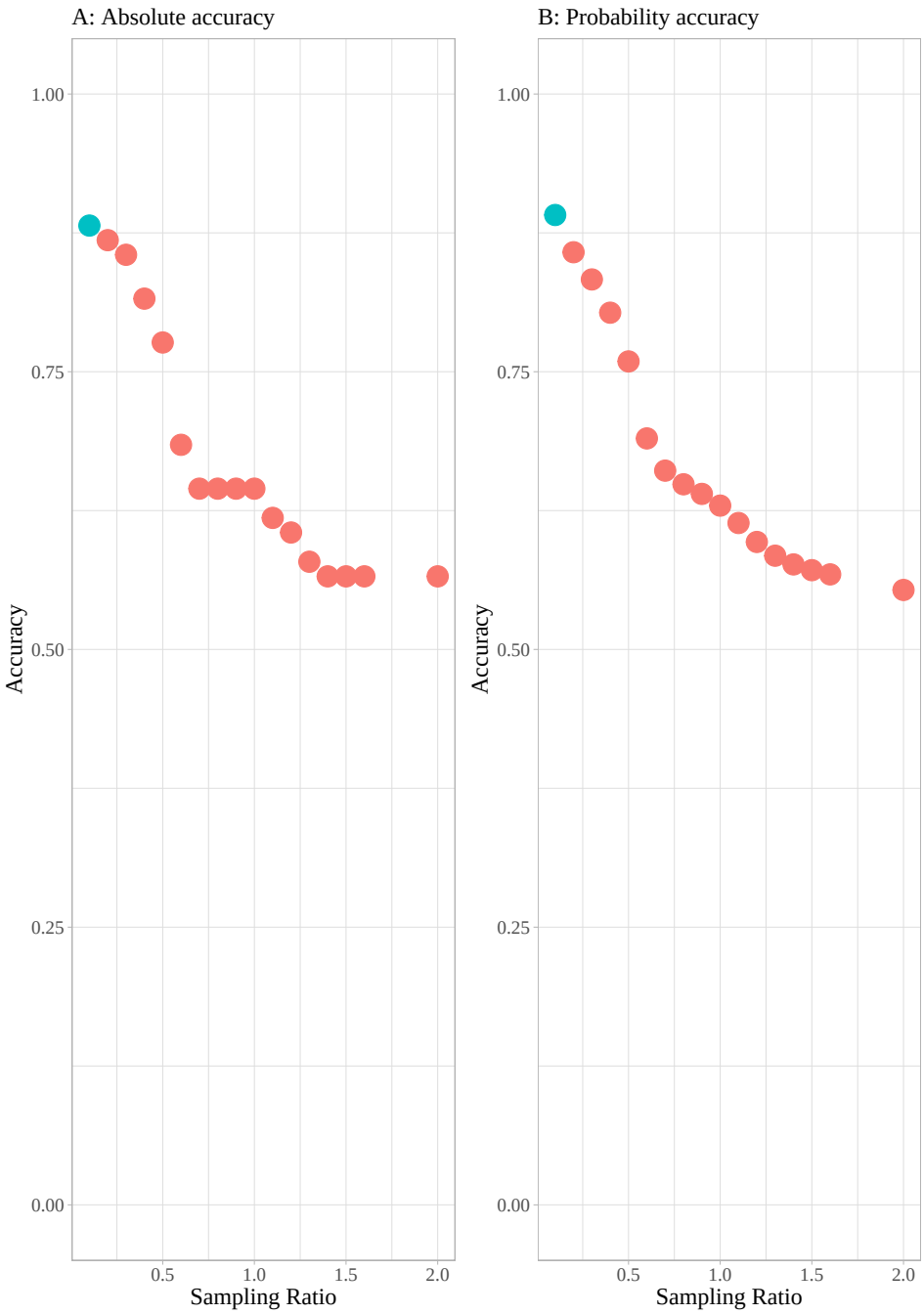

Figure S2: Accuracies of SAASI under varying sampling ratios on a fixed simulated tree. A: Absolute accuracy; B: Probability accuracy. The blue point represents accuracy under the true sampling ratio ($\frac{\psi_2}{\psi_1} = 0.1$). Red points represent accuracy under mis-specified sampling ratios ranging from $\frac{\psi_1}{\psi_2} = 0.15$ to 2.

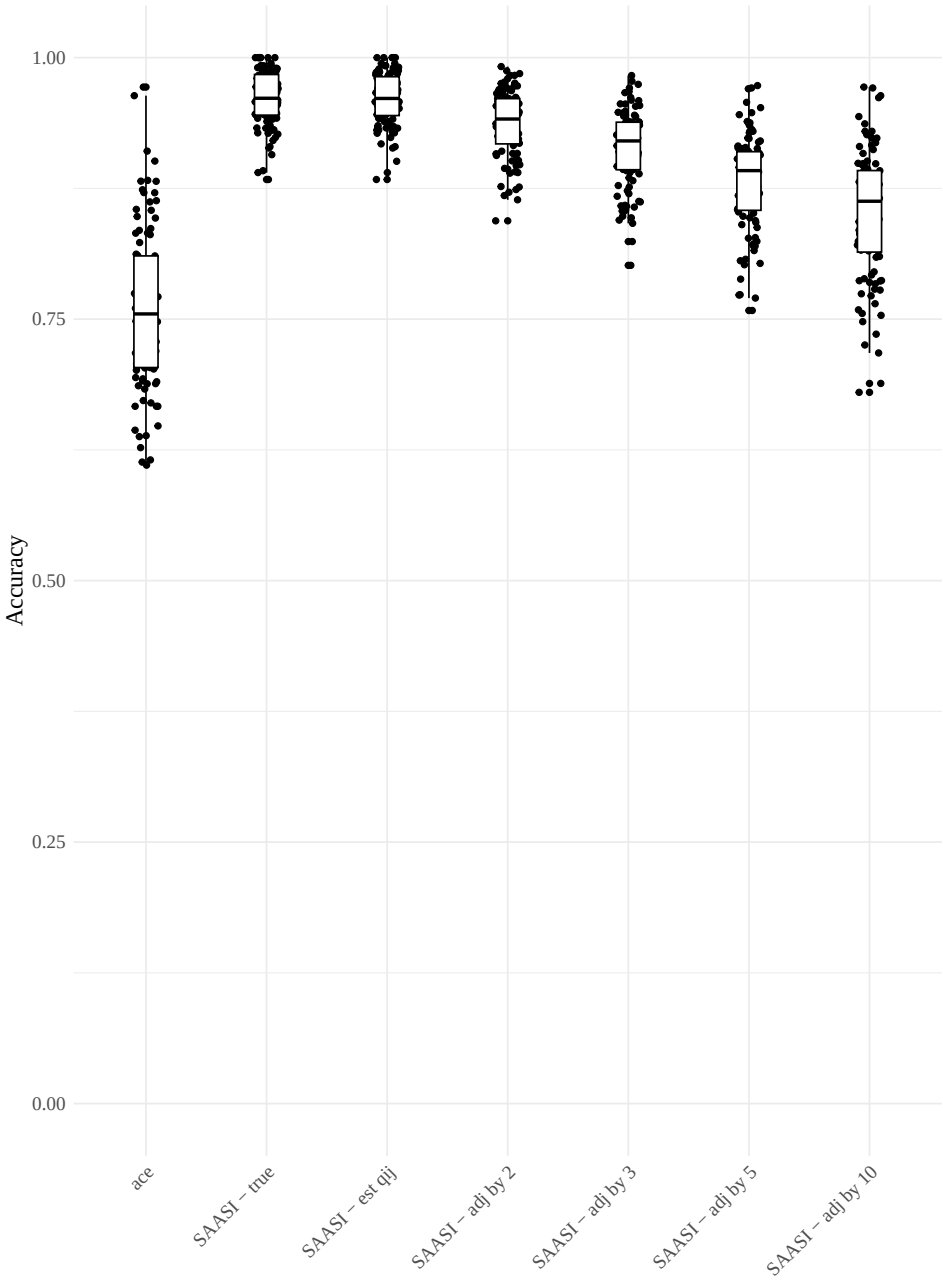

Figure S3: Comparison of absolute accuracies for ancestral state inference using ace and SAASI under various transition rates adjustments. State 1 is sampled 10 times less than the other states (three states in total). Simulations use $q_{ij} = 0.2, \forall i, j$. From left to right: ace; SAASI with true rates; SAASI with estimated transition rates from 'ace' (q_{ij}^{ace}); SAASI with adjusted transition rates of state 1 by a factor of c, ranging from $c = \{2, 3, 5, 10\}$.

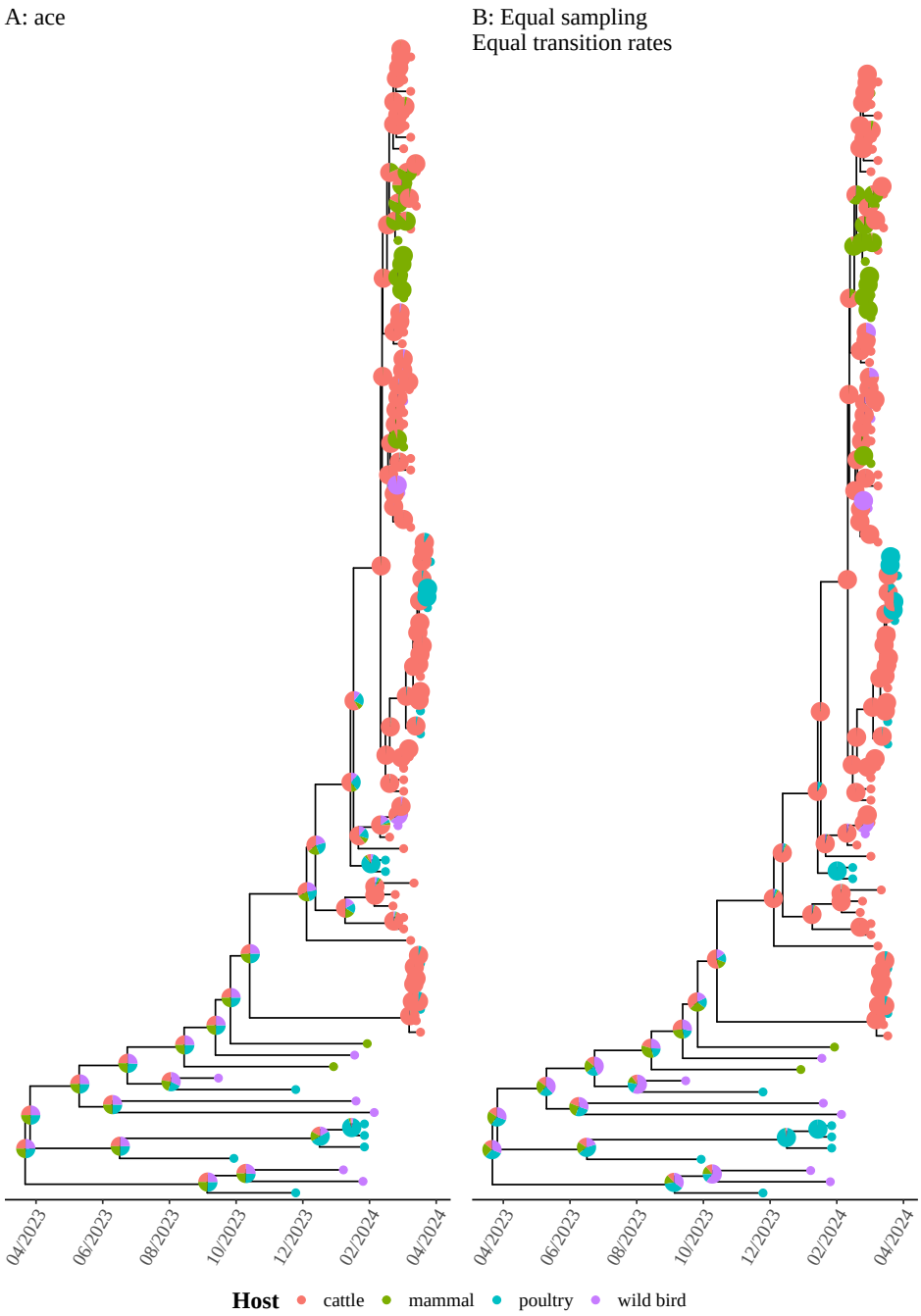

Figure S4: Ancestral state inference of the H5N1 HA segment tree using ace and SAASI, assuming equal sampling across species. A: ace; B: SAASI, equal sampling. SAASI with adjusted transition rates of state 1 by a factor of $c = 2$. Pie charts indicate the inferred probabilities of being in particular states.

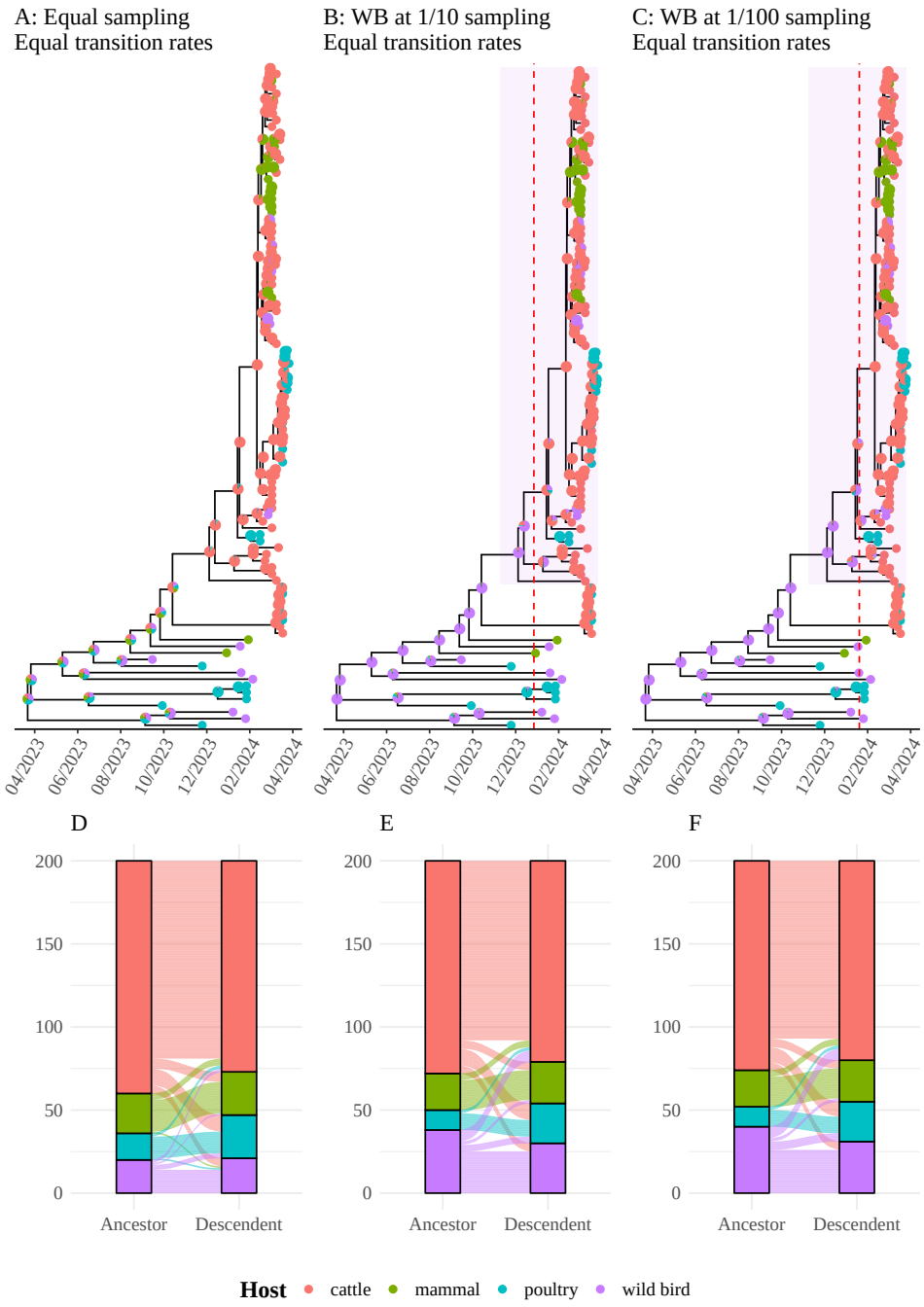

Figure S5: Ancestral state inference of the H5N1 HA segment tree using SAASI under different species-level sampling models. A: Inferred species hosts under equal sampling rates; B: wild birds at one-tenth sampling; C: wild birds at one-one hundredth sampling; D: Inferred viral transitions between host species in A; E: Inferred viral transitions between host species in B; F: Inferred viral transitions between host species in C. Pie charts indicate the inferred probabilities of being in particular states. Transition rates are equal between species. The dashed red line indicates the key transition event from wild bird to cattle. WB refers to the wild bird population.

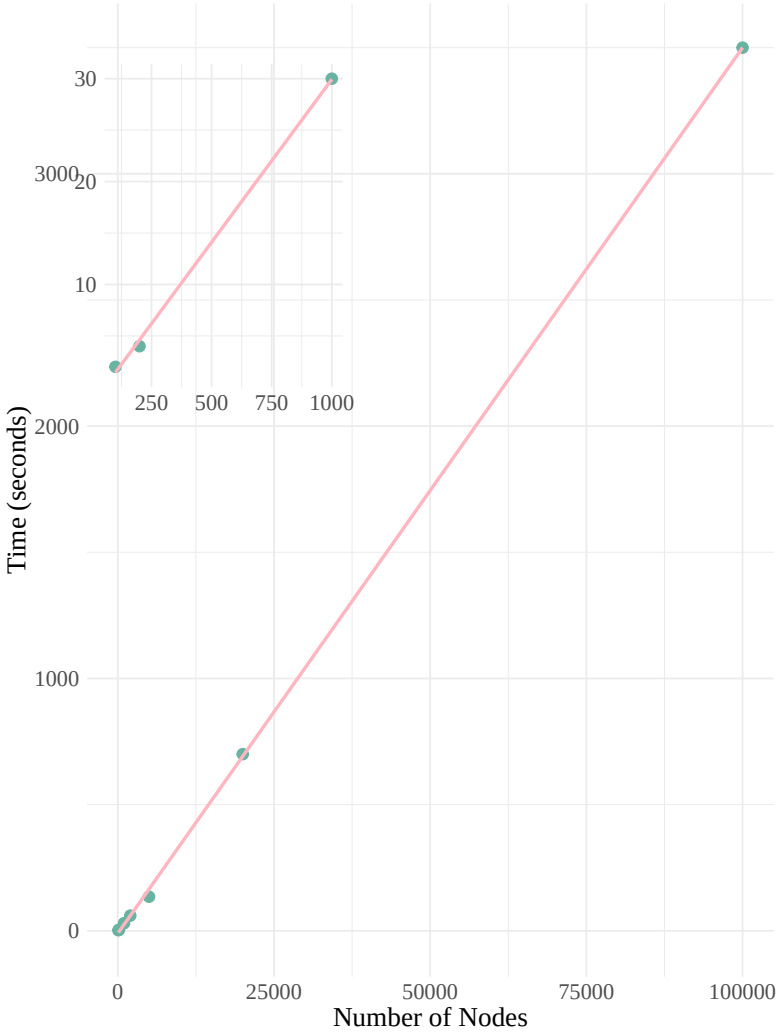

Figure S6: Runtime of SAASI across trees of different sizes. The x-axis represents the number of nodes, and the y-axis represents the times needed using SAASI. The red line shows the line of best fit. The small panel on the top left is a zoomed-in view of small tree sizes.