The Structured Coalescent and its Approximations

Nicola F. Müller*†1, David A. Rasmussen*† and Tanja Stadler*†1

*ETH Zurich, Department of Biosystems Science and Engineering, 4058 Basel, Switzerland
†Swiss Institute of Bioinformatics (SIB), Switzerland

[1]Corresponding authors

Abstract: Phylogenetics can be used to elucidate the movement of pathogens between different host populations when the location of samples are considered alongside of pathogen sequence data. Pathogen phylogenies therefore offer insights into the movement of pathogens not available from classic epidemiological data alone. However, current phylogeographic methods to quantify migration patterns from phylogenies have several known shortcomings. In particular, one of the most widely used method treats migration the same as mutation, and as such does not incorporate information about population demography. This may lead to severe biases in estimated migration rates for datasets where sampling is biased across populations. On the other hand, the structured coalescent allows us to coherently model the migration and transmission process, but current implementations struggle with complex datasets due to the need to additionally infer ancestral migration histories. Thus, approximations to the structured coalescent which integrate over all ancestral migration histories have been developed. However, the validity and robustness of these approximations remain unclear. We here provide an exact numerical solution to the structured coalescent that does not require the inference of migration histories. While this solution is computationally unfeasible for large datasets, it clarifies the assumptions of previously developed approximate methods and allows us to provide an improved approximation to the structured coalescent. We have implemented these methods in BEAST2, and we show how our newly described approach outperforms previously described methods in accuracy at comparable computational cost.

1 Introduction

The quantification of pathogen spread in structured host populations using phylogenies enables us to draw conclusions about the sources and origins of infectious diseases. Methods accounting for population structure, also called *phylogeographic* methods, have been used to analyze the global spread of H3N2 (Bedford et al., 2010; Bahl et al., 2011; Lemey et al., 2014; Bedford et al., 2015), the origins of HIV-1 (Faria et al., 2014) and various other diseases (Bourhy et al., 2008; Raghwani et al., 2011).

A range of phylogeographic methods have been proposed. The mugration method (Lemey et al., 2009) treats migration as a continuous time Markov chain, such as used to model mutation, and assumes the migration process to be independent of the tree generating process. This assumption can lead to biases in estimates of migration rates when sampling is biased (De Maio et al., 2015). Other methods, such as the structured coalescent (Takahata, 1988; Hudson, 1990; Notohara, 1990), do not make this independence assumption. In contrast to the mugration-based methods, they require the state (or location) of any ancestral lineage in the phylogeny at any time to be inferred (Beerli and Felsenstein, 2001; Ewing et al., 2004; Vaughan et al., 2014). In other words, they require the ancestral migration history to be inferred. Inferring lineage states is computationally expensive, as it normally requires Markov chain Monte Carlo (MCMC) based sampling, and limits the complexity of scenarios that can be analyzed.

Another approach (Volz, 2012) seeks to marginalize over all possible migration histories by treating lineage states probabilistically instead of using MCMC based sampling. Rather than assigning lineages to particular states, the probability of each lineage being in each state is calculated at all times using a set of previously described differential equations (Volz, 2012). Such a marginalization approach (rather than explicit sampling of states) allows for the analysis of larger datasets (De Maio et al., 2015). While this approach appears to only make the assumption of lineage independence, i.e. that the state or location of one lineage does not depend on any other lineage (De Maio et al., 2015), it remains unclear if there are additional assumptions not being accounted for.

We here derive an exact numerical solution of the structured coalescent, based on the joint probabilities of lineages being in any possible configuration. The derivation of this exact solution clarifies the assumptions required to arrive at the previously described structured coalescent differential equations (Volz, 2012). Clarifying these assumptions allows us to develop a more refined approximation to the structured coalescent. We then show how the different approximations compare in terms of tree, parameter and root state inference under both biased and unbiased sampling conditions. Simulations reveal that our new approximation outperforms previous approximations at comparable computational cost. We then apply these different approximations to a previously described avian influenza virus dataset (Lu et al., 2014) sampled from different regions of North America to show that the choice of method influences the interpretation of data in practice.

2 Materials and Methods

2.1 Principle of the structured coalescent process

The structured coalescent (Takahata, 1988; Hudson, 1990; Notohara, 1990) extends the standard coalescent by allowing lineages to occupy different states. If we consider L_i to be a random variable that denotes the state of lineage i with state space $\{1, ..., m\}$, there are m^n different possible configurations $\mathcal{K}$ of how n lineages can be arranged ($\mathcal{K} = (L_1 = l_1, ..., L_i = l_i, ..., L_n = l_n), l_i \in \{1, ..., m\}$). These configurations can change over time by adding and removing lineages or by lineages changing state. Throughout this paper, we consider time going backwards from present to past, as typically done under the coalescent.

A migration event along one lineage i from state a to state b changes the configuration of lineages as follows:

$$(L_1 = l_1, ..., L_{i-1} = l_{i-1}, L_i = a, L_{i+1} = l_{i+1}, ..., L_n = l_n)$$
$$\xmapsto{migration\ event\ from\ a\ to\ b}$$
$$(L_1 = l_1, ..., L_{i-1} = l_{i-1}, L_i = b, L_{i+1} = l_{i+1}, ..., L_n = l_n)$$

In figure 1, this corresponds to lineage 1 in *blue* changing to *red*.

Configurations can additionally change due to sampling. Sampling events simply add lineages, such as $L_3 = red$ is added in figure 1. Typically, we condition on the sampling events, but one can also introduce a rate for samples being obtained.

Coalescent events remove lineages, changing the configuration as follows:

$$(L_1 = l_1, ..., L_{i-1} = l_{i-1}, L_i = a, L_{i+1} = l_{i+1}, ..., \\ L_{j-1} = l_{j-1}, L_j = a, L_{j+1} = l_{j+1}, ..., L_n = l_n)$$
$$\xmapsto{coalescent\ event}$$
$$(L_1 = l_1, ..., L_{i-1} = l_{i-1}, L_i = a, L_{i+1} = l_{i+1}, ..., \\ L_{j-1} = l_{j-1}, L_j = l_{j+1}, ..., L_{n-1} = l_n)$$

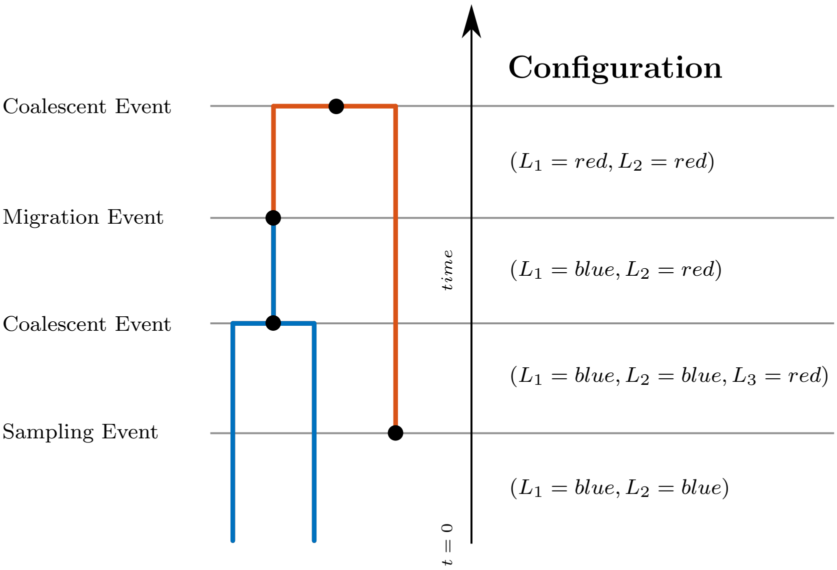

Figure 1: **Events and configurations on an example tree**. Here, we illustrate the possible events and the configurations before and after each event on a simple tree, with time going backwards from present to past. The first two lineages are both in state blue, i.e. the configuration is $(L_1 = blue, L_2 = blue)$. After a lineage in state *red* is sampled, the configuration changes, as given in the figure. A coalescent event in state *blue* then reduces the number of lineages in state *blue* to 1. A migration event then causes lineage L_1 to change state from *blue* to *red*.

The most recent coalescent event in figure 1 for example changes the configuration from $(L_1 = blue, L_2 = blue, L_3 = red)$ to $(L_1 = blue, L_2 = red)$.

The rate at which coalescent events in state a happen can be calculated from the pairwise coalescent rate λ_a in state a and the number of lineages $k_a(\mathcal{K})$ in state a for a given configuration $\mathcal{K}$. The pairwise coalescent rate denotes the rate at which any two lineages in a state coalesce. For a given a configuration $\mathcal{K}$, the total rate $\mathcal{C}$ at which coalescent events between any two lineages in the same state happen is:

$$\mathcal{C} = \sum_{a=1}^{m} \lambda_a \binom{k_a(\mathcal{K})}{2}, \tag{1}$$

where $\binom{k_a(\mathcal{K})}{2}$ is the number of pairs of lineages in state a given configuration $\mathcal{K}$. Under the standard Wright-Fisher model, the pairwise coalescent rates, λ_a, are the inverse of the effective population sizes N_{e_a}.

2.2 Calculating the likelihood for a tree under the structured coalescent

Structured coalescent methods typically use MCMC to integrate over possible lineage state configurations along a tree (Beerli and Felsenstein, 2001; Ewing et al., 2004; Vaughan et al., 2014). This is sometimes referred to as sampling migration histories. Given a migration history, the likelihood for a tree can be calculated under the structured coalescent with given migration and coalescent rates. Here, we want to calculate the marginal likelihood for a tree without sampling those migration histories, but by integrating over all possible migration histories H. Formally, we seek to calculate the following probability:

$$P(T|S, M, \Lambda) = \int_H P(T, H|S, M, \Lambda)dH,$$

with T being the tree, S the sampling states of the tips, M the set of migration rates and Λ the set of coalescent rates.

Let $P_t(L_1 = l_1, ..., L_i = l_i, ..., L_n = l_n, T)$ be the probability density that the samples more recent than time t evolved according to the coalescent history, i.e. the branching pattern, given by our tree T between the present time 0 and time t and that the n lineages at time t, $L_1, \ldots, L_n$, are in states $l_1, \ldots, l_n$. In figure 1, this probability is the joint probability of a configuration at time t with the lineages being either in red or blue, and the probability of the branching pattern being as observed between time t and 0 (ignoring the particular configurations in that time interval).

We aim to calculate P_t for $t = t_{mrca}$, with t_{mrca} being the time of the root of the tree T. At the root of the tree, summing over the probability of the remaining lineage being in any state will yield the likelihood for the tree, $P(T|S, M, \Lambda) = \sum_{a=1}^{m} P_{t_{mrca}}(L_1 = a, T)$.

In order to evaluate P_t at $t = t_{mrca}$, we start at the time of the most recent sample, at $t = 0$, and iteratively calculate $P_{t+\Delta t}$ based on P_t. To calculate P_t, we split the calculation into three parts: time intervals in the tree where no coalescent or sampling events happen, sampling events, and coalescent events. Below, we first consider the interval part of this calculation. Afterwards, we calculate the contribution of coalescent and sampling events.

Interval contribution. For the interval part, we calculate $P_{t+\Delta t}$ based on P_t allowing for no event in time step Δt (second line below), observing a migration event leading to the configuration at $t + \Delta t$ (third line below), or seeing more than one event (i.e. higher order terms which are of order $O((\Delta t)^2)$ leading to the configuration at $t + \Delta t$ (forth line below):

$$\begin{aligned}
&P_{t+\Delta t}(L_1 = l_1, ..., L_i = l_i, ..., L_n = l_n, T) \\
&\qquad = P_t(L_1 = l_1, ..., L_i = l_i, ..., L_n = l_n, T)(1 - \mathcal{M}\Delta t - \mathcal{C}\Delta t) \\
&\qquad + \sum_{i=1}^{n} \sum_{a=1}^{m} \Big(m_{al_i} \Delta t P_t(L_1 = l_1, ..., L_i = a, ..., L_n = l_n, T) \Big) \\
&\qquad\qquad + O((\Delta t)^2)
\end{aligned}$$

Here, $\mathcal{M}$ is the sum of migration rates and $\mathcal{C}$ the sum of coalescent rates for configuration $(L_1 = l_1, ..., L_i = l_i, ..., L_n = l_n)$. The rate m_{al_i} denotes the rate at which migration events from a to l_i happen. Now, when re-arranging and letting $\Delta t \rightarrow 0$, we obtain the differential equation,

$$\begin{aligned}
&\frac{dP_t(L_1 = l_1, ..., L_i = l_i, ..., L_n = l_n, T)}{dt} \\
&\qquad = -(\mathcal{M} + \mathcal{C})P_t(L_1 = l_1, ..., L_i = l_i, ..., L_n = l_n, T) \\
&\qquad\qquad + \sum_{i=1}^{n} \sum_{a=1}^{m} \Big(m_{al_i} P_t(L_1 = l_1, ..., L_i = a, ..., L_n = l_n, T) \Big).
\end{aligned}$$

With explicitly writing $\mathcal{M}$ and $\mathcal{C}$ (using equation 1 for $\mathcal{C}$), we obtain,

$$\begin{aligned}
&\frac{dP_t(L_1 = l_1, ..., L_i = l_i, ..., L_n = l_n, T)}{dt}\\
&\qquad = \sum_{i=1}^{n}\sum_{a=1}^{m}\Big(m_{al_i}P_t(L_1 = l_1, ..., L_i = a, ..., L_n = l_n, T)\\
&\qquad\qquad - m_{l_i a}P_t(L_1 = l_1, ..., L_i = l_i, ..., L_n = l_n, T)\Big)\\
&\qquad - \sum_{a=1}^{m}\lambda_a\binom{k_a}{2}P_t(L_1 = l_1, ..., L_i = l_i, ..., L_n = l_n, T)\\
&\qquad\qquad\qquad\qquad \textit{(interval contribution)} \quad (2)
\end{aligned}$$

With the double summation on the right hand side considering the contribution of migration and the fourth line considering the contribution of coalescence. Further, $k_a = \sum_i^n \delta_{L_i,a}$ where δ is the Kronecker delta with $\delta_{L_i,a} = 1$ for $L_i = a$ and 0 otherwise. Note that in the case of $l_i = a$, the two terms in the migration part cancel each other out and the net migration is 0. This *interval contribution* equation allows us to calculate P_t within intervals by solving the differential equation.

It is important to note that this differential equation shows a direct link between the coalescent process and the probability of a set of lineages being in a configuration. For example, configurations that would favor high coalescent rates among lineages would become less probable over intervals during which no coalescent events occur in the tree.

Sampling event contribution. At every sampling event the state of the sampled lineage is independent of all other lineages in the tree. We can therefore calculate the probability of any configuration at a samping event at time t as follows:

$$\begin{aligned}
&P_t(L_1 = l_1, ..., L_i = l_i, ..., L_{n+1} = l_{n+1}, T)\\
&\qquad = P_t(L_1 = l_1, ..., L_i = l_i, ..., L_n = l_n, T)P_t(L_{n+1} = l_{n+1}, T)\\
&\qquad\qquad\qquad \textit{(sampling event)}
\end{aligned}$$

In scenarios where the sampling state is known to be say a, we have $P_t(L_{n+1} = a, T) = 1$ and $P_t(L_{n+1} = b, T) = 0$ for $b \neq a$. In cases where the sampling state is an inferable parameter or not exactly known, this probability can be between 0 or 1.

Coalescent event contribution. Next, we have to calculate the probability of the new configuration resulting from a coalescent event between lineages i

and j in state a at time t. This probability can be expressed by the following equation:

$$\begin{aligned} &P_t(L_1 = l_1, ..., L_i = a, ..., L_{n-1} = l_n, T) \\ &\qquad = P_t(L_1 = l_1, ..., L_i = a, ..., L_j = a, ..., L_n = l_n, T)\lambda_a \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \textit{(coalescent event)} \end{aligned}$$

Thus based on the three equations, *(interval contribution)*, *(sampling event)*, *(coalescent event)*, we can calculate the likelihood for a tree, $P(T|S, M, \Lambda)$. Below, we refer to this approach as the exact structured coalescent (ESCO). The above approach can also be deployed for coalescent events between states rather than just within states (Volz, 2012). For simplicity, only the case where coalescent events occur between lineages in the same state is discussed here.

2.3 Approximations of the exact structured coalescent

The exact structured coalescent can be approximated by assuming that lineages evolve independent of any configuration $\mathcal{K}$, i.e.:

$$P_t(L_j = a|\mathcal{K}, T) \overset{LISCO}{=} P_t(L_j = a|T)$$

With this assumed independence, we show in the appendix that the *interval contribution* differential equation can be written for each lineage i independent of the other lineages:

$$\frac{dP_t(L_i = l_i, T)}{dt} = \sum_{a=1}^{m} \Big(P_t(L_i = a, T)m_{al_i} - P_t(L_i = l_i, T)m_{l_i a} \Big) - \lambda^i_{l_i}, \quad (3)$$

with

$$\lambda^i_{l_i} = P_t(L_i = l_i, T)\frac{\lambda_{l_i}}{2} \sum_{j \neq i}^{n} \frac{P_t(L_j = l_i, T)}{\sum_{a=1}^{m} P_t(L_j = a, T)}.$$

$\lambda^i_{l_i}$ is the rate at which lineage i in state l_i coalesces with any other lineage in the same state. In contrast to λ_{l_i}, $\lambda^i_{l_i}$ is a rate specific to lineage i, which can be derived from equation 1 (see supplement for the derivation) Below, we refer to this as the lineage independence approximation (LISCO).

Integrating the above differential equation over time is equivalent to calculating the probability that the lineage i is in state l_i and did not coalesce, assuming the lineage independence stated above.

To calculate the probability of lineage i coalescing with lineage j in state a, we proceed as for ESCO, and previously described in (Volz, 2012):

$$P_t(L_i = a, T) = P_t(L_i = a, T)P_t(L_j = a, T)\lambda_a$$

At a sampling event, we simply add a lineage $n+1$ with associated probability $P_t(L_{n+1} = l_{n+1}, T)$ analog to ESCO.

A further approximation can be obtained by ignoring the coalescence term $\lambda^i_{l_i}$ in equation 3, i.e. additionally assuming independence of the lineage states from the coalescent process between events. Thus, we assume,

$$P_t(L_j = a|\mathcal{K}, T) \overset{SISCO}{=} P_t(L_j = a),$$

which directly leads,

$$\frac{dP_t(L_i = l_i)}{dt} = \sum_{a=1}^{m} \Big(P_t(L_i = a)m_{al_i} - P_t(L_i = l_i)m_{l_ia} \Big). \tag{4}$$

In order to obtain $P_t(L, T)$ at the root, we need to calculate $P_t(T)$ which follows $\frac{d}{dt}P_t(T) = -\lambda = -\sum_{a=1}^{m}\sum_{i=1}^{n}\lambda^i_a$ where λ is the total coalescent rate at time t. Using our eq. for $\lambda^i_{l_i}$ together with the SISCO independence assumption, we obtain ,

$$\frac{dP_t(T)}{dt} = -\sum_{a=1}^{m}\frac{\lambda_a}{2}\sum_{i=1}^{n}\sum_{j\neq i}^{n} P_t(L_i = a)P_t(L_j = a).$$

Now, the likelihood for the tree under the structured coalescent with the SISCO approximation is, $P(T|S, M, \Lambda) = P_{t_{mrca}}(T)\sum_{a=1}^{m} P_{t_{mrca}}(L_1 = a) = P_{t_{mrca}}(T)$.

We refer to this as the state independence approximation of the structured coalescent (SISCO). The equations used by SISCO to calculate the state of a lineage over time have been described previously in Volz (2012). While these lineage state probabilities are independent of the coalescent history T between events, they do depend on T at sampling and coalescent events.

To recap the assumptions that are needed to arrive at this equation, we can write the following:

$$P_t(L_j = a|\mathcal{K}, T) \overset{LISCO}{=} P_t(L_j = a|T) \overset{SISCO}{=} P_t(L_j = a)$$

The left hand side is the exact description of the structured coalescent with the probability of lineage i depending on a configuration $\mathcal{K}$ and the coalescent history described by the tree T. LISCO now assumes lineage states being independent from configurations $\mathcal{K}$ and SISCO assumes lineage states being independent of configurations $\mathcal{K}$ and the coalescent history T.

2.4 Implementation

We implemented all three approximations in one common package for BEAST2 (Bouckaert et al., 2014). ESCO and LISCO use a forth order Runge-Kutta solver with fixed step size implemented in the Apache Commons Math library (http://commons.apache.org) to solve equations 2 and 3. SISCO uses matrix exponentiation to solve the lineage state probabilities over time (equation 4). All three structured coalescent methods use pairwise coalescent rates and backwards in time migration rates as described above. In the Results section, we present simulation analyses highlighting the quality of the different structured coalescent approximations.

2.5 Software and Data availability

Simulations were performed using a backwards in time stochastic simulation algorithm of the structured coalescent process using MASTER 5.0.2 (Vaughan and Drummond, 2013) and BEAST 2.4.2 (Bouckaert et al., 2014). Script generation and post-processing were performed in Matlab R2015b. Plotting was done in R 3.2.3 using ggplot2 (Wickham, 2009). Tree plotting and tree height analyses were done using ape 3.4 (Paradis et al., 2004) and phytools 0.5-10 (Revell, 2012). Effective sample sizes for MCMC runs were calculated using coda 0.18-1 (Plummer et al., 2006). All scripts for performing the simulations and analyses presented in this paper as well as the Java source code for the structured coalescent methods are available at https://github.com/nicfel/The-Structured-Coalescent.git. Output files from these analyses are available upon request from the authors.

2.6 Application to Avian Influenza Virus

We applied the different approximations of the structured coalescent to a previously described data set of Avian Influenza Virus H7 hemaglutinen (HA) sequences (Lu et al., 2014), sampled from the bird orders anseriformes, charadriiformes, galliformes and passeriforms in Canada, Mexico and the USA. We used previously aligned sequences from De Maio et al. (2015). The sequences were analyzed in BEAST2 using an HKY+Γ_4 site model. A strict molecular clock model was assumed and the first two and third codon positions were allowed to have different mutation rates. LISCO and SISCO were used as structured coalescent population priors. The dataset was split into 7 different states according to geographic regions in North America (see table S1). Three parallel MCMC chains were run for $2*10^7$ iterations with different initial migration and coalescent rates. After a burnin of 10%, the chains were combined and the probability of the root being in each state was assessed. The combined chain had ESS values above 100 for any inferred probability density or parameter.

3 Results

3.1 Tree height distributions under the structured coalescent and its approximations

The structured coalescent and its approximations describe different probability distributions over trees. To see how these distributions compare, we performed direct backwards-in-time simulations under the structured coalescent using MASTER (Vaughan and Drummond, 2013), analogously to Vaughan et al. (2014). These trees were compared to trees sampled under ESCO, LISCO, SISCO, as well as BASTA (De Maio et al., 2015), a numerical approximation of SISCO. Under these latter four models, trees were sampled from their respective probability distributions using MCMC in BEAST2 (Bouckaert et al., 2014). Since it is difficult to directly compare distributions of trees, we instead compared the distribution of tree heights.

For each of the five scenarios (direct, ESCO, LISCO, SISCO, BASTA), we obtained 8000 trees. We used a model with three different states, sampling one, two and three individuals from each state, respectively. Coalescent rates were different in each state ($\lambda_1 = 1, \lambda_2 = 2, \lambda_3 = 4$) and migration rates were different between states ($m_{1,2} = 0.01, m_{1,3} = 0.02, m_{2,1} = 0.001, m_{2,3} = 0.003, m_{3,1} = 0.01, m_{3,2} = 0.01$).

Figure 2 shows the distribution of tree heights sampled using MCMC and compares them to the distribution of tree heights obtained by directly simulating trees under the structured coalescent. Of the different methods, only the distribution of ESCO is consistent with direct simulation. Assuming lineage independence (LISCO) leads to an underestimation of tree heights. Further assuming lineage states to be independent of the coalescent process (SISCO) results in a greater underestimation of tree heights. BASTA (De Maio et al., 2015), being an approximation of SISCO, underestimates tree heights less than SISCO. SISCO estimating shorter tree heights compared to LISCO can be explained in the following way. Not taking into account how the coalescent process influences lineage states leads to an overestimation of the probability of two lineages being in the same state if no coalescent event is observed by SISCO compared to LISCO. Overestimating the probability of two lineages being in the same state then also leads to an overestimation of the probability of them coalescing. This in turn results in shorter trees

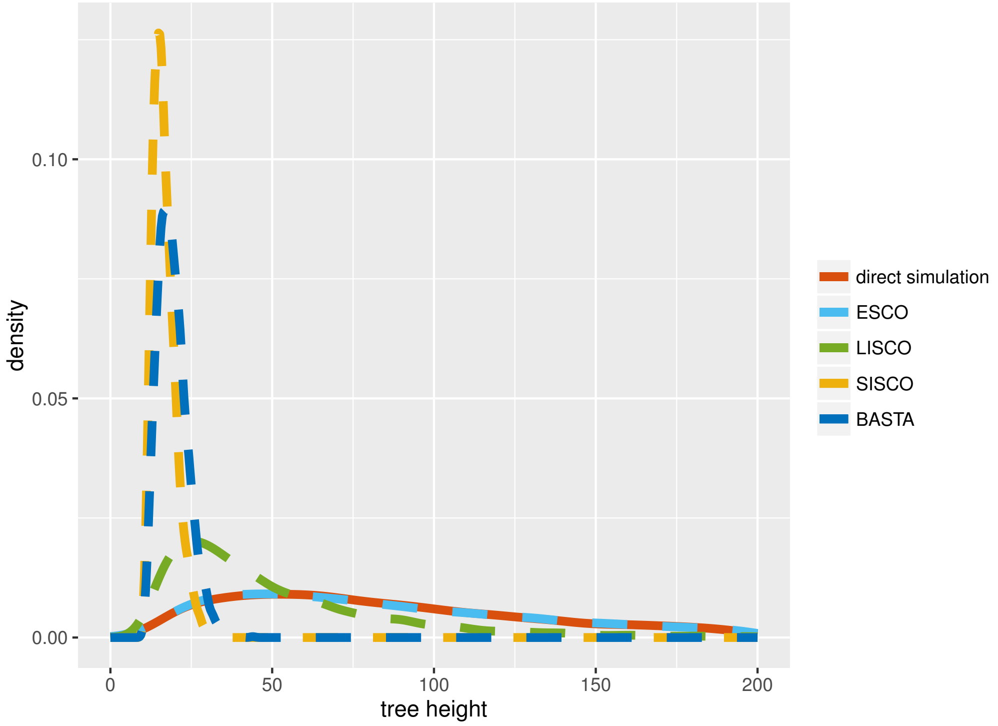

Figure 2: **Comparison of MCMC sampled to simulated tree heights using the different structured coalescent methods**. The trees were sampled using MCMC for 10^6 iterations, storing every 1000^{th} step, after a burnin of 20%.

since lineages are expected to coalesce at a faster rate.

3.2 Root state probabilities

The ancestral state or location of lineages back in time is often of interest for biological questions. For example, in a pathogen phylogeny the root location is informative of the geographic origin of an epidemic. Here we show on one fixed tree how the exact structured coalescent compares in the inference of the root state to its approximations. We additionally inferred the root state using MultiTypeTree (Vaughan et al., 2014), which uses MCMC to sample lineage states and does not rely on approximations, to obtain a reference root state probability (Vaughan et al., 2014). We inferred the probability of the root being in either state for different migration rates in one direction and holding the rate in the other direction constant.

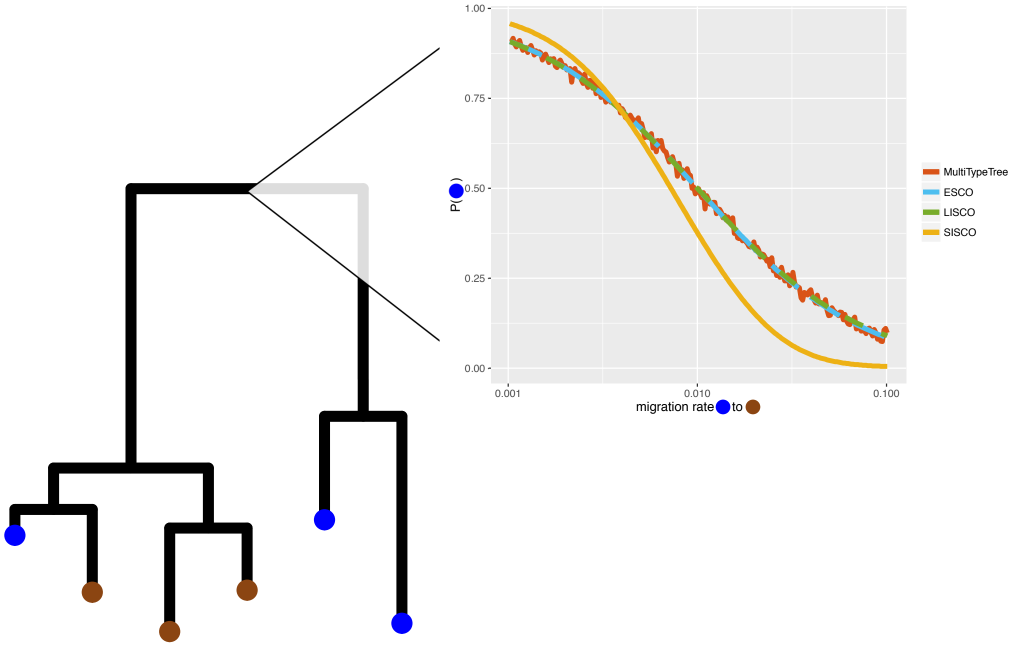

Figure 3: **Inferred location of the root for different migration rates and structured coalescent approaches**. The plot shows the probability of the root being in the blue state (y-axis) depending on the migration rate from blue to brown (x-axis), for the given tree and sampling states. The migration rate from brown to blue was held constant at 0.01.

The exact structured coalescent and the one assuming independence of the lineage states (LISCO) agree well with the inferred posterior mean using MultiTypeTree(Figure 3). The inferred state probabilities using SISCO on the other hand do not, showing that the assumption of independence between the lineage states and the coalescent process not only biases parameter inference but state inference as well.

3.3 Estimation of migration rates

Often, the approximate methods are used to infer population and migration parameters from trees. To show how the inference of the migration rates compares to the true rate, we simulated 1000 trees under the structured coalescent with symmetric migration rates from 10^{-5} to 1 and pairwise coalescent rates of 2 using MASTER. Each tree consisted of 4 contemporaneously sampled leafs from each of the two states. We fixed the coalescent rates to the truth and assumed symmetric migration rates and then inferred the maximum likelihood estimate of the migration rate using the exact structured coalescent (ESCO) and its approximations LISCO and SISCO.

The results are summarized in figure 4. When assuming dependence of the lineage states on the coalescent process (LISCO), the migration rates are slightly underestimated compared to ESCO. The dependence between estimated and simulated migration rates is however linear. Assuming lineage independence and independence of the lineage states and the coalescent process (SISCO) leads to strong biases in estimates of the migration rates. The lower the migration rates are compared to the coalescent rates, the greater the underestimation of the migration rates becomes.

3.4 Estimation of rate asymmetries

In the previous section, we inferred the rate of migration giving the methods the true coalescent rate and the information that the migration rates were the same in both directions. In reality, these rates can greatly vary across states or locations. Coalescent rates for example depend on transmission rates, population sizes, etc. It is therefore important for methods to be able to perform well in situations where rates are asymmetric. Previous work showed that the ability to infer migration rate asymmetries greatly depends on the method used (De Maio et al., 2015). Here we compare inferences of rate asymmetries under LISCO and SISCO. Applying ESCO to the same trees would not be feasible, since the computation time grows with the number of states to the power of the number of lineages. We simulated a total of 2000 trees using MASTER with 100 tips from each of the two different states

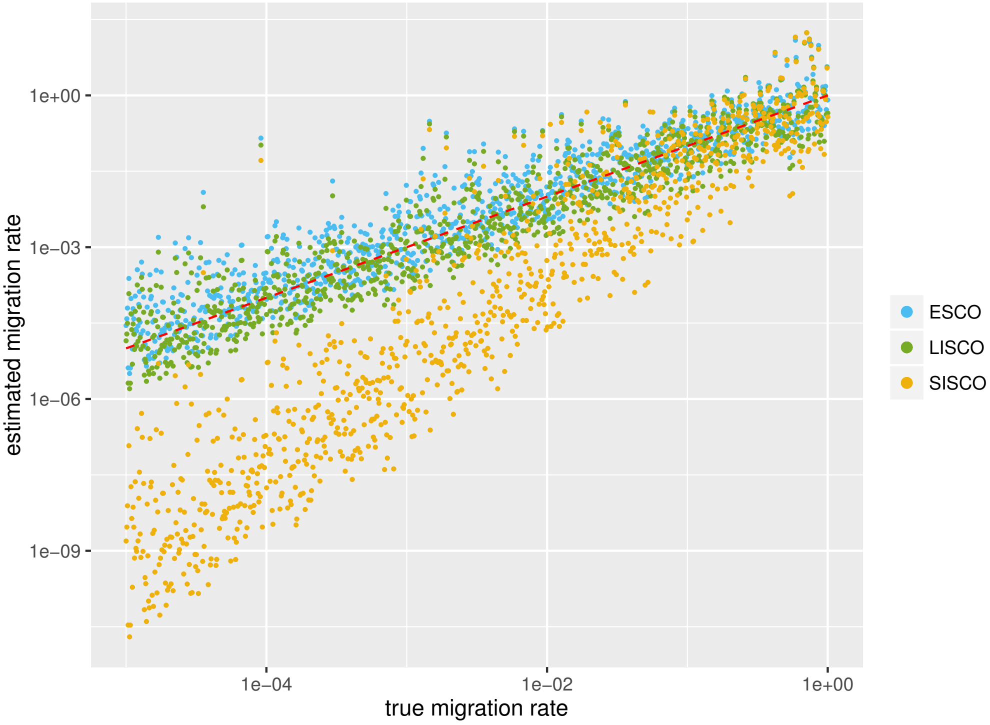

Figure 4: **Maximum likelihood estimates of migration rates using the exact structured coalescent and its approximations**. Here we compare simulated migration rates (x-axis) to the maximum likelihood estimates of the migration rate (y-axis), estimated using the exact structured coalescent ESCO and its approximations LISCO and SISCO. The coalescent rates are fixed to the truth, and the migration rates are assumed to be symmetric. The red line indicates where the true values should lie.

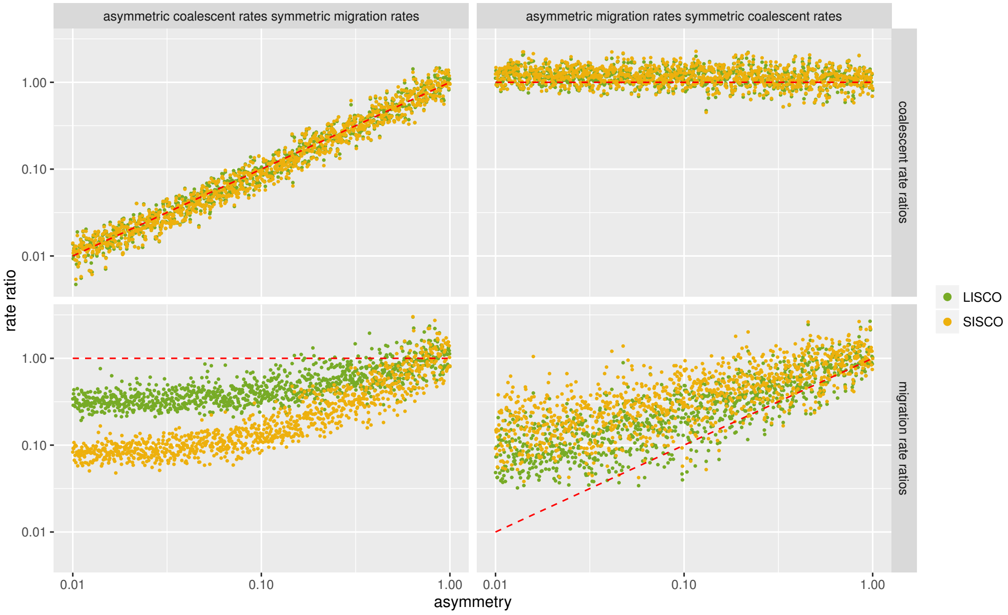

Figure 5: **Inferred asymmetry of migration and coalescent rates**. Here we show the inferred mean coalescent (upper row) and migration (lower row) rate ratios under different conditions. In the first column, the coalescent rate ratios (x-axis) are varied while the migration rates ratios are kept constant. In the second column, the migration rate ratios (x-axis) are varied, while the coalescent rate ratios are kept constant. Both coalescent rates and both migration rates are estimated. The red line indicates where the estimates should lie.

sampled uniformly between times t=0 and t=10. Of these trees, 1000 were simulated with pairwise coalescent rate ratios $\frac{\lambda_1}{\lambda_2}$ from 0.01 to 1, $\lambda_1+\lambda_2=4$ and migration rates in both direction equal to 1. The other 1000 trees were simulated with migration rate ratios from $\frac{m_{12}}{m_{21}}$ from 0.01 to 1, $m_{12}+m_{21} = 2$ and pairwise coalescent rates in both states equal to 2. We then inferred the coalescent and migration rates under all scenarios (i.e. 4 parameters in total), using exponential priors with the mean 2 for the coalescent rates and mean 1 for the migration rates. We then ran three independent MCMC runs with different initial values for each scenario. The 3 runs were then combined to ensure convergence to the same optima.

Figure 5 shows the ratios of mean inferred coalescent and migration rates using LISCO and SISCO. The coalescent rate ratios are accurate using either structured coalescent approximation. When the coalescent rates are symmetric, the migration rate asymmetries are underestimated for increasing asymmetry. However, when taking into account the highest posterior density (HPD) intervals of the estimates, most estimates contain the true rate ratio (see figures S1 and S2). When the coalescent rate ratios are varied while the migration rates are symmetric, both methods are biased. They overestimate migration into the state with the smaller coalescent rate relative to the migration rate in the other direction. SISCO however shows stronger biases then LISCO under any coalescent rate ratio.

3.5 Sampling bias

Previous work showed that the approximate structured coalescent is able to accurately infer migration rates even when sampling fractions are biased, given samples are taken contemporaneously (De Maio et al., 2015). Here we explore the effect of biased sampling fractions in the presence of serial sampling. We compare the exact structured coalescent ESCO to its approximations LISCO and SISCO. We simulated trees under the structured coalescent with 10, 30 or 50 samples from one state and 90, 70 or 50 samples from the other state, sampled uniformly between t=0 and t=25. We then inferred all the migration and coalescent rates using the different structured coalescent methods, using exponential priors with the means being the true rates. Each migration and coalescent rates was inferred seperately between states respectively within states, i.e. we inferred 4 parameters in total.

Figure 6 reveals that ESCO is able to unbiasedly infer the migration rates in both directions, independent of sampling biases or migration rates. The same applies to LISCO, except under very biased sampling and intermediate migration rates. For SISCO however, biased sampling leads to an underestimation of the backwards migration rate into the oversampled state and an overestimation of the rates into the undersampled state for intermediate and high migration rates. At low migration rates, both rates are underestimated.

3.6 Application to Avian Influenza Virus

To show how the inference of the origin of an epidemic varies with the method used, we applied the two approximations of the structured coalescent (LISCO and SISCO) to a previously described avian influenza dataset (Lu et al., 2014; De Maio et al., 2015) to infer the geographic location of the root.

In figure 7 we show the inferred region of the root using LISCO and SISCO. Despite the fact that almost all samples from the central US were collected after 2009 and that samples from the East Coast and the North West fall closer to the root, both methods place a high probability of the

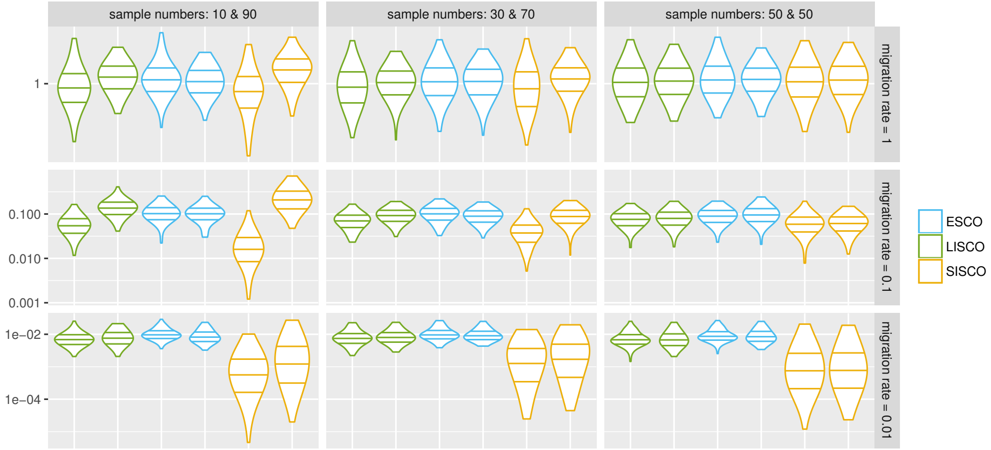

Figure 6: **Inferred migration rates under different sampling conditions**. The plot shows the distribution of mean inferred migration rates using ESCO, LISCO and SISCO. From the left, the first distribution of a color (indicating the different methods) always shows the distribution of mean inferred migration rates from state 1 to state 2. The second distribution from the same color shows the rates from state 2 to 1. From left to right the number of samples from state 1 and state 2 are changed, while from top to bottom the true symmetric migration rates are going from 1 to 0.01. The pairwise coalescent rates were 2 under all conditions. The lines within the violin plots indicate the 25%, 50% and 75% quantiles. The coalescent rates were 2 in both states and the migration rates ranged from 0.01 to 1. The migration rates were always symmetric, i.e. the same in both directions. Each simulation was repeated 100 times and each inference was run with 3 parallel MCMC chains, each with different initial values. An exponential distribution with the mean being the true rate was used on the migration and coalescent rates.

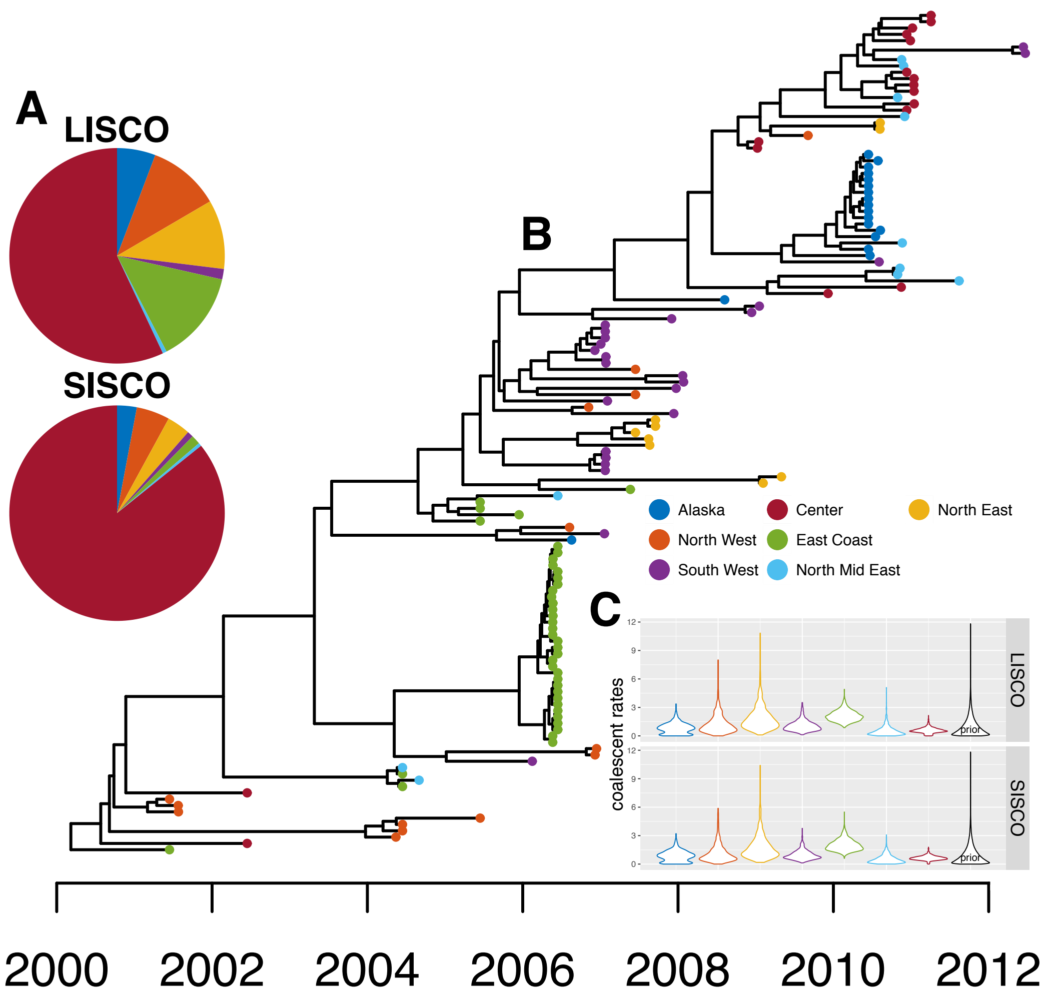

Figure 7: **Inference of the root regions of AIV sampled from different places in North America**. **A** Maximum clade credebility tree inferred from AIV sequences sampled in different regions of the USA, Canada and Mexico using LISCO as a population prior. The node heights represent the mean node heights. The tip color indicate the different sampling regions shown in the legend. **B** Inferred root regions using LISCO (top) and SISCO (bottom). The pie charts show the inferred probability of the root being in either of the different states/regions by LISCO and SISCO. **C** Violin plots of the inferred coalescent rates for the different regions. The black plot distribution is the exponential prior with mean 1. We used this prior for both coalescent and migration rates.

root being in the central US, with over 80% probability under SISCO and 50% probability under LISCO. However, the East Coast and North West are given more probability of being the root location under LISCO than SISCO. We provide a possible explanation to why we observe differences in the inferred root state in the Discussion below.

4 Discussion

The structured coalescent approach described here provides the first exact numerical solution of the structured coalescent process (Takahata, 1988; Hudson, 1990; Notohara, 1990) without the need to sample migration histories, as in previously described approaches (Beerli and Felsenstein, 2001; Ewing et al., 2004; Vaughan et al., 2014). Additionally, we introduce a new approximation that outperforms a previously described approximation (Volz, 2012). This new approximation facilitates a trade-off between speed and accuracy. The increased speed compared to the exact solution originates from the assumption of lineage independence. This assumption leads to better scaling of the computational complexity with the number of states and lineages. We show that the lineage independence assumption allows us to infer migration, coalescent rates and root states well in most scenarios. Additionally assuming independence of the lineages states from the coalescent process as introduced in (Volz, 2012) however leads to major biases in parameter and root state inference. These biases are especially pronounced in our simulations when migration is slow compared to the coalescent rate. This observation can be explained in the following way: The lower migration rates are compared to coalescent rates, the stronger the influence of the coalescent process on the configuration of lineages across states. The assumption of independence of the lineage states from the coalescent process does not allow for the incorporation of this information into the calculation of lineage state probabilities though.

Next, we showed how the approximations of the structured coalescent perform in inferring asymmetric coalescent and migration rates. While coalescent rates are inferred accurately, inference of migration rate ratios is biased when coalescent rates are asymmetric. These biased estimates are more pronounced under SISCO than LISCO. We also showed that under biased sampling, inferences of migration rates are strongly biased under SISCO, but not LISCO.

Both biases can be understood in the following way. A lineage may have a higher probability of coalescing in one state than another either because the pairwise coalescent rate in one state is higher (e.g. due to a smaller effective population size) or because more lineages reside in one state than another (e.g. because of biased sampling). Taking the influence of the coalescent process on lineage states into account, as done under LISCO, reduces the probability of a lineage occupying a state with a high coalescent rate if no coalescent events occur. This reduction is proportional to the probability of that lineage in that state coalescing. In other words, LISCO redistributes via equation 3 the probability mass assigned to each state to reflect a lineage's coalescent history, including the observation that a lineage may have not yet coalesced. As the differential equation for SISCO is missing the coalescent rate (equation 4), SISCO does not redistribute probability mass to reflect the probability that a lineage has not yet coalesced. In order to reduce the probability of lineages coalescing in a state with high rates of coalescence, it overestimates the migration rate out of such states. This overestimation of migration rates out of a state is observable when having asymmetric coalescent rates due to either a higher pairwise coalescent rate within a state or having more lineages in a given state due to biased sampling. Either way, the migration rate out of the state with higher coalescent rate is overestimated and underestimated in the other direction.

Lastly, we applied the different approximations of the structured coalescent to avian influenza virus HA sequences sampled from different orders of birds in North America. We found that the inferred region of the root varies with the method used. SISCO places high confidence in the center of the USA being the root state. LISCO also infers the center to be the most likely root state, however it additionally infers the East Coast, the North West and East to be possible root states.

Asymmetric coalescent rates may offer one explanation why SISCO places more probability on the center being the root location than LISCO and why it excludes all other states from being possible root states. We have shown that asymmetric coalescent rates can bias the inference of migration rates. Asymmetric coalescent rates lead to an overestimation of the migration rate from a state with fast coalescence into a state with slow coalescence and an underestimation of the migration rates in the other direction (recall that we consider backwards in time rates). Our simulations revealed that such biases are much more pronounced in SISCO than in LISCO. Because the coalescent rate in the center is inferred to be low, SISCO puts much more weight on it being the source than LISCO. The opposite appears to occur for the East Coast, which is inferred to have a very high rate of coalescence. LISCO infers the East Coast to be the second most likely source region while it is almost excluded using SISCO.

Although we used the AIV analysis to illustrate how inferences obtained from LISCO and SISCO can differ, the results presented here should be interpreted with caution with regards to any biological implications as we ignored population structure arising between different avian host species. We additionally assumed coalescent and migration rates to be constant over time, potentially further biasing the inference of the root state.

While population dynamics such as changing transmission (i.e. coalescent) and migration rates through time can greatly influence the shape of a phylogeny, we ignored such dynamics in this study. However, compared to mugration type methods (Lemey et al., 2009), the structured coalescent approximation introduced here can be extended in a conceptually straightforward way to allow for dynamic populations (Volz et al., 2009; Volz, 2012). The improved approximation to the structured coalescent introduced here should therefore allow for more accurate quantification of pathogen movement in structured populations with complex population dynamics while still being computationally efficient enough to be applied to large datasets.

5 Acknowledgement

NM and TS were funded in part by a SNF SystemsX grant (TBX). DR is funded by the ETH Zürich Postdoctoral Fellowship Program and the Marie Curie Actions for People COFUND Program. TS is supported in part by the European Research Council under the Seventh Framework Programme of the European Commission (PhyPD: grant agreement number 335529).

References

Trevor Bedford, Sarah Cobey, Peter Beerli, and Mercedes Pascual. Global migration dynamics underlie evolution and persistence of human influenza A (H3N2). *PLoS pathogens*, 6(5):e1000918, may 2010. ISSN 1553-7374. doi: 10.1371/journal.ppat.1000918. URL http://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1000918.

J. Bahl, M. I. Nelson, K. H. Chan, R. Chen, D. Vijaykrishna, R. A. Halpin, T. B. Stockwell, X. Lin, D. E. Wentworth, E. Ghedin, Y. Guan, J. S. M. Peiris, S. Riley, A. Rambaut, E. C. Holmes, and G. J. D. Smith. Temporally structured metapopulation dynamics and persistence of influenza A H3N2 virus in humans. *Proceedings of the National Academy of Sciences*, 108(48):19359–19364, nov 2011. ISSN 0027-8424. doi: 10.1073/pnas.1109314108. URL http://www.pnas.org/content/108/48/19359.abstract.

Philippe Lemey, Andrew Rambaut, Trevor Bedford, Nuno Faria, Filip Bielejec, Guy Baele, Colin A Russell, Derek J Smith, Oliver G Pybus, Dirk Brockmann, and Marc A Suchard. Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. *PLoS pathogens*, 10(2):e1003932, feb 2014. ISSN 1553-7374. doi: 10.1371/journal.ppat.1003932. URL http://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1003932.

Trevor Bedford, Steven Riley, Ian G. Barr, Shobha Broor, Mandeep Chadha, Nancy J. Cox, Rodney S. Daniels, C. Palani Gunasekaran, Aeron C. Hurt, Anne Kelso, Alexander Klimov, Nicola S. Lewis, Xiyan Li, John W. McCauley, Takato Odagiri, Varsha Potdar, Andrew Rambaut, Yuelong Shu, Eugene Skepner, Derek J. Smith, Marc A. Suchard, Masato Tashiro, Dayan Wang, Xiyan Xu, Philippe Lemey, and Colin A. Russell. Global circulation patterns of seasonal influenza viruses vary with antigenic drift. *Nature*, 523 (7559):217–220, jun 2015. ISSN 0028-0836. doi: 10.1038/nature14460. URL http://dx.doi.org/10.1038/nature14460.

Nuno R. Faria, Andrew Rambaut, Marc A. Suchard, Guy Baele, Trevor Bedford, Melissa J. Ward, Andrew J. Tatem, João D. Sousa, Nimalan Arinaminpathy, Jacques Pépin, David Posada, Martine Peeters, Oliver G. Pybus, and Philippe Lemey. The early spread and epidemic ignition of HIV-1 in human populations. *Science*, 346(6205), 2014.

H. Bourhy, J.-M. Reynes, E. J. Dunham, L. Dacheux, F. Larrous, V. T. Q. Huong, G. Xu, J. Yan, M. E. G. Miranda, and E. C. Holmes. The origin and phylogeography of dog rabies virus. *Journal of General Virology*, 89(11):2673–2681, nov 2008. ISSN 0022-1317. doi: 10.1099/vir.0.2008/003913-0. URL http://jgv.microbiologyresearch.org/content/journal/jgv/10.1099/vir.0.2008/003913-0.

Jayna Raghwani, Andrew Rambaut, Edward C. Holmes, Vu Ty Hang, Tran Tinh Hien, Jeremy Farrar, Bridget Wills, Niall J. Lennon, Bruce W. Birren, Matthew R. Henn, and Cameron P. Simmons. Endemic Dengue Associated with the Co-Circulation of Multiple Viral Lineages and Localized Density-Dependent Transmission. *PLoS Pathogens*, 7 (6):e1002064, jun 2011. ISSN 1553-7374. doi: 10.1371/journal.ppat.1002064. URL http://dx.plos.org/10.1371/journal.ppat.1002064.

Philippe Lemey, Andrew Rambaut, Alexei J. Drummond, and Marc a. Suchard. Bayesian phylogeography finds its roots. *PLoS Computational Biology*, 5(9):e1000520, sep 2009.

ISSN 1553734X. doi: 10.1371/journal.pcbi.1000520. URL http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000520.

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, aug 2015. ISSN 1553-7404. doi: 10.1371/journal.pgen.1005421. URL http://www.plosgenetics.org/article/Metrics/info:doi/10.1371/journal.pgen.1005421.

Naoyuki Takahata. The coalescent in two partially isolated diffusion populations. *Genetical research*, 52(3):213–222, dec 1988. ISSN 0016-6723. doi: 10.1017/S0016672300027683. URL http://www.journals.cambridge.org/abstract{_}S0016672300027683.

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

M. Notohara. The coalescent and the genealogical process in geographically structured population. *Journal of Mathematical Biology*, 29(1):59–75, oct 1990. ISSN 0303-6812. doi: 10.1007/BF00173909. URL http://link.springer.com/10.1007/BF00173909.

P Beerli and J Felsenstein. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proceedings of the National Academy of Sciences of the United States of America*, 98 (8):4563–8, apr 2001. ISSN 0027-8424. doi: 10.1073/pnas.081068098. URL http://www.pnas.org/content/98/8/4563.long.

Greg Ewing, Geoff Nicholls, and Allen Rodrigo. Using Temporally Spaced Sequences to Simultaneously Estimate Migration Rates, Mutation Rate and Population Sizes in Measurably Evolving Populations. 2004. doi: 10.1534/genetics.104.030411. URL http://www.genetics.org/content/168/4/2407.

Timothy G Vaughan, Denise Kühnert, Alex Popinga, David Welch, and Alexei J Drummond. Efficient Bayesian inference under the structured coalescent. *Bioinformatics (Oxford, England)*, 30(16):2272–9, aug 2014. ISSN 1367-4811. doi: 10.1093/bioinformatics/btu201. URL http://bioinformatics.oxfordjournals.org/content/early/2014/04/20/bioinformatics.btu201.short.

Erik M Volz. Complex population dynamics and the coalescent under neutrality. *Genetics*, 190(1):187–201, jan 2012. ISSN 1943-2631. doi: 10.1534/genetics.111.134627. URL http://www.genetics.org/content/190/1/187.short.

Lu Lu, Samantha J Lycett, and Andrew J Leigh Brown. Determining the Phylogenetic and Phylogeographic Origin of Highly Pathogenic Avian Influenza (H7N3) in Mexico. *PLoS ONE*, 9(9):e107330, sep 2014. URL http://dx.doi.org/10.1371/journal.pone.0107330.

Remco Bouckaert, Joseph Heled, Denise Kühnert, Tim Vaughan, Chieh-Hsi Wu, Dong Xie, Marc A Suchard, Andrew Rambaut, and Alexei J Drummond. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS computational biology*, 10(4): e1003537, apr 2014. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003537. URL http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003537.

Timothy G Vaughan and Alexei J Drummond. A stochastic simulator of birth-death master equations with application to phylodynamics. *Molecular biology and evolution*, 30(6):1480–93, jun 2013. ISSN 1537-1719. doi: 10.1093/molbev/mst057. URL http://www.ncbi.nlm.nih.gov/pubmed/23505043http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3649681.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009. ISBN 978-0-387-98140-6. URL http://ggplot2.org.

Emmanuel Paradis, Julien Claude, and Korbinian Strimmer. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics (Oxford, England)*, 20(2): 289–90, jan 2004. ISSN 1367-4803. doi: 10.1093/BIOINFORMATICS/BTG412. URL http://www.ncbi.nlm.nih.gov/pubmed/14734327.

Liam J. Revell. phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, 3(2):217–223, apr 2012. ISSN 2041210X. doi: 10.1111/j.2041-210X.2011.00169.x. URL http://doi.wiley.com/10.1111/j.2041-210X.2011.00169.x.

Martyn Plummer, Nicky Best, Kate Cowles, and Karen Vines. CODA: convergence diagnosis and output analysis for MCMC. 2006.

Erik M Volz, Sergei L Kosakovsky Pond, Melissa J Ward, Andrew J Leigh Brown, and Simon D W Frost. Phylodynamics of infectious disease epidemics. *Genetics*, 183(4): 1421–30, dec 2009. ISSN 1943-2631. doi: 10.1534/genetics.109.106021. URL http://www.genetics.org/content/183/4/1421.abstract.