Bioinformatics, 34(22), 2018, 3843–3848
doi: 10.1093/bioinformatics/bty406
Advance Access Publication Date: 22 May 2018
Original Paper

OXFORD

Phylogenetics

MASCOT: parameter and state inference under the marginal structured coalescent approximation

Nicola F. Müller[1,2,*], David Rasmussen[1,2,3,4] and Tanja Stadler[1,2,*]

[1]Department of Biosystems Science and Engineering, ETH Zurich, 4058 Basel, Switzerland, [2]Swiss Institute of Bioinformatics (SIB), 1015 Lausanne, Switzerland, [3]Department of Entomology and Plant Pathology and [4]Bioinformatics Research Center, North Carolina State University, Raleigh, NC 27695-7566, USA

*To whom correspondence should be addressed.

Associate Editor: Janet Kelso

Received on September 15, 2017; revised on May 9, 2018; editorial decision on May 14, 2018; accepted on May 16, 2018

Abstract

Motivation: The structured coalescent is widely applied to study demography within and migration between sub-populations from genetic sequence data. Current methods are either exact but too computationally inefficient to analyse large datasets with many sub-populations, or make strong approximations leading to severe biases in inference. We recently introduced an approximation based on weaker assumptions to the structured coalescent enabling the analysis of larger datasets with many different states. We showed that our approximation provides unbiased migration rate and population size estimates across a wide parameter range.

Results: We extend this approach by providing a new algorithm to calculate the probability of the state of internal nodes that includes the information from the full phylogenetic tree. We show that this algorithm is able to increase the probability attributed to the true sub-population of a node. Furthermore we use improved integration techniques, such that our method is now able to analyse larger datasets, including a H3N2 dataset with 433 sequences sampled from five different locations.

Availability and implementation: The presented methods are part of the BEAST2 package MASCOT, the Marginal Approximation of the Structured COalescenT. This package can be downloaded via the BEAUti package manager. The source code is available at https://github.com/nicfel/Mascot.git.

Contact: nicola.mueller@bsse.ethz.ch or tanja.stadler@bsse.ethz.ch

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Phylogenies contain information regarding the history of a population and can be used to quantify demographic parameters. This has been widely done to study the spread of pathogens (Pybus *et al.*, 2001; Russell *et al.*, 2008), the speciation dynamics of extant species or the migration pattern of humans to name but a few. Forwards in time birth-death and backwards in time-coalescent models allow us to elucidate population dynamics from trees by calculating the probability of a phylogeny T given a set of demographic parameters Θ. To do so they classically rely on the assumption of well mixed populations, meaning that the rate at which any two pairs of lineages coalesce is the same. In most empirical applications this assumption of well mixed populations is however violated.

To address this model violation, so-called structured methods have been developed that consider birth-death processes in heterogeneous populations (Stadler and Bonhoeffer, 2013). In the backward in-time coalescent framework, the structured coalescent (Hudson, 1990; Notohara, 1990; Takahata, 1988) describes a coalescent process in sub-populations between which individuals can migrate. Such coalescent methods however typically 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). Inferring lineage states is computationally expensive, as it normally requires MCMC-based sampling, and limits the complexity of scenarios that can be analysed. As the number of different states is increased, convergence of the MCMC chains becomes a severe issue for inference under the structured coalescent when the sampling of migration histories is needed (De Maio *et al.*, 2015; Vaughan *et al.*, 2014). This essentially limits the number of different states that can be accounted for to three or four.

We addressed this limitation recently by introducing a new approximation of the structured coalescent that avoids this MCMC sampling of lineage states by integrating over all possible migration histories using a set of ordinary differential equations (Müller *et al.*, 2017). In contrast to previous approximations that treat the movement of one lineage completely independently of all other lineages (De Maio *et al.*, 2015; Volz, 2012), we explicitly include information about the location of other lineages and their probability of coalescing when modelling the movement of a lineage. This avoids the strong biases resulting from this independence assumption (Müller *et al.*, 2017). We showed that this approximation is able to infer coalescent and migration rates well in various scenarios. However, this approach currently lacks the possibility to estimate the ancestral state of any internal nodes except the root.

Here, we introduce a new algorithm (Fig. 1) to calculate the probability of internal nodes being in any state that incorporates information from the entire tree using a forwards/backwards approach (Pearl, 1982). We additionally make improvements of the current BEAST2 (Bouckaert *et al.*, 2014) implementation of Müller *et al.* (2017) in terms of calculation speed, allowing us to analyse datasets with more states and lineages. We show first on simulated datasets how this new implementation performs in inferring migration rates and effective population sizes in high dimensional parameter space. Next, we show how our new algorithm can dramatically improve ancestral state inference. We then apply our new approach to a geographically distributed samples of human Influenza A/H3N2 virus to demonstrate its applicability to large datasets.

2 Materials and methods

2.1 The approximate structured coalescent

In Müller *et al.* (2017), we introduced a new approximation to the structured coalescent that integrates over every possible migration history and avoids the sampling of lineage states. This is done by calculating the marginal probability of a lineage i being in any of m possible states, jointly with the probability of having observed the coalescent history T from the present backwards in time until time point t in the tree, with time 0 being the time of the most recent sample with time increasing into the past. To do so, we need to make the following approximation:

$$P_t(L_i = l_i, L_j = l_j, L_k = l_k|T) \underset{MASCO}{=} P_t(L_i = l_i|T)P_t(L_j = l_j|T)P_t(L_k = l_k|T)$$

In other words, we assume that lineages i, j and k and their states l_i, l_j and l_k are pairwise independent.

2.2 The probability of a lineage being in a state

As described in Müller *et al.* (2017), we seek to calculate the probability of every lineage being in any state jointly with the probability of having observed the coalescent history T up to time t.

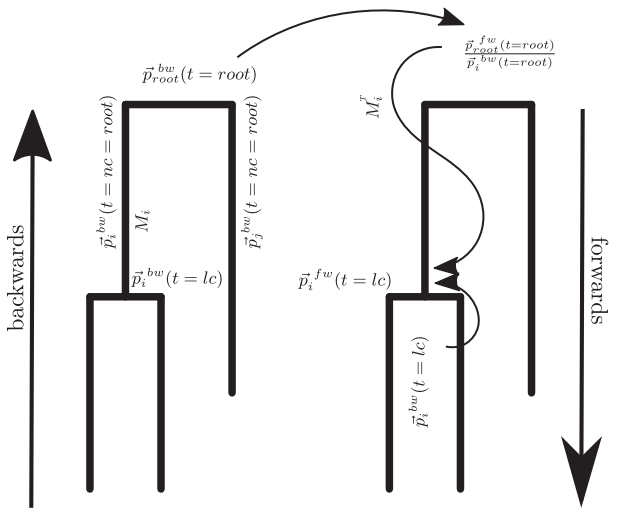

Fig. 1. Flow of information using the backwards/forwards algorithm. Going backwards in the tree, we calculate the probability of each node being in any state that includes information up to time t. The vector $\vec{p}_i^{\,bw}(t = lc)$ has the entries $P_{t=lc}(L_i = a|T)$ in position a. At the root, the backwards probabilities $\vec{p}_{root}^{\,bw}(t = \text{root})$ are equal to the forwards probabilities $\vec{p}_{root}^{\,fw}(t = \text{root})$. To calculate the downwards probabilities $\vec{p}_i^{\,fw}(t = lc)$, we use the information from all the other parts of the tree and the transition matrix M_i and the backwards probabilities $\vec{p}_i^{\,bw}(t = lc)$

We previously denoted this probability as $P_t(L_i = l_i, T)$. Calculating these terms over time for increasing t leads to ever smaller values, eventually causing numerical issues. To avoid this, we can calculate $P_t(L_i = l_i|T) = P_t(L_i = l_i, T)/P_t(T)$ instead. The expression for $(d/dt)P_t(L_i = l_i|T)$ can be directly derived from $(d/dt)P_t(L_i = l_i, T)$ (see Supplementary Material) and can be written as:

$$\begin{aligned}\frac{d}{dt}P_t(L_i = l_i|T) = &\sum_{a=1}^{m}\left(\mu_{al_i}P_t(L_i = a|T) - \mu_{l_ia}P_t(L_i = l_i|T)\right)\\ &+P_t(L_i = l_i|T)\sum_{a=1}^{m}\lambda_a P_t(L_i = a|T)\sum_{\substack{k=1\\k\neq i}}^{n}P_t(L_k = a|T)\\ &-P_t(L_i = l_i|T)\lambda_{l_i}\sum_{\substack{k=1\\k\neq i}}^{n}P_t(L_k = l_i|T)\end{aligned} \quad (1)$$

with μ_{al_i} denoting the backwards in time rate at which lineages migrate from state a to state l_i and λ_a denoting the rate of coalescence in state a. To calculate $P_t(T)$, i.e. the probability of having observed the coalescent history T up to time t, the following differential equation has to be solved (see Supplementary Material for derivation):

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

At coalescent events between lineages i and j, we update $P_t(T)$ by multiplication with the probability of the coalescent event:

$$P_t^{\text{after}}(T) = P_t^{\text{before}}(T)\left(\sum_{a=1}^{m}\lambda_a P_t(L_i = a|T)P_t(L_j = a|T)\right) \quad (3)$$

Integrating these equations from the present to the root of a phylogeny, allows us to calculate $P_{\text{root}}(T)$, that is the probability density of a phylogeny T under the MASCO approximations of the structured coalescent.

2.3 The probability of a node being in a state given the whole phylogeny

2.3.1 Backwards calculation of node states conditional on the Sub-trees

Integrating Equation (1) allows us to calculate the probability of each lineage being in any state given the coalescent history between the lineage and the present. However, for applications, it is much more interesting to calculate the probability of each lineage at time t being in any state given the whole phylogenetic tree between the time of root and the present. At coalescent events between lineage i and j at time t, the probability of the parent lineage P at time t being in state a can be calculated as follows:

$$P_t^{\mathrm{bw}}(L_{\mathrm{p}}=a|T)=\frac{\lambda_a P_t^{\mathrm{bw}}(L_i=a|T)P_t^{\mathrm{bw}}(L_j=a|T)}{\sum_{b=1}^{m}\lambda_b P_t^{\mathrm{bw}}(L_i=b|T)P_t^{\mathrm{bw}}(L_j=b|T)}, \quad (4)$$

with $P_t^{\mathrm{bw}}(L_i=a|T)$ denoting the probability of the daughter lineage i being in state a just before the coalescent event at time t calculated in the backwards step using Equation (1). Since $P_t^{\mathrm{bw}}(L_{\mathrm{p}}=a|T)$ denotes the probability of the parent node of lineages i and j calculated in the backwards step at time t, it includes only information up to the time of coalescence and does not include information from the full phylogeny. We introduce the label 'bw' to differentiate from the forward probabilities introduced below. To additionally incorporate information from the phylogeny between the time of the root and time t of the coalescent event, one has to deploy a 'backwards/forwards' approach that is related to Pearl (1982).

For convenience, we now change to vector notation. We define $\vec{p}_{\mathrm{p}}^{\,\mathrm{bw}}(t)$ as the vector for the parent lineage p with entries $P_t^{\mathrm{bw}}(L_{\mathrm{p}}=a|T)$ in position a that only includes information from time 0 up to time t. $\vec{p}_i^{\,\mathrm{bw}}(t)$ is the vector with entries $P_t^{\mathrm{bw}}(L_i=a|T)$.

2.3.2 Calculation of transition probabilities

Going through the tree backwards in time, we also seek to calculate the probability that, given lineage i was in state a at the last coalescent event lc where lineage i was the parent and given the tree T between 0 and time t, the lineage i is in state b at the time of the next coalescent event nc where lineage i is one of the two daughter lineages. We denote this probability as $P_{t=nc}(L_i=b|L_{i(t=lc)}=a)$. The matrix M_i with entries $P_{t=nc}(L_i=b|L_{i(t=lc)}=a)$ in positions (a, b) now denotes the matrix for which the following equation holds:

$$\vec{p}_i^{\,\mathrm{bw}}(t=nc)=\vec{p}_i^{\,\mathrm{bw}}(t=lc)M_i, \quad (5)$$

with $\vec{p}_{t=nc}^{\,b}$ being the vector with the state probabilities of lineage i just before its next coalescent event nc. To calculate the entries of the matrix M_i, we solve the following differential equation between the two coalescent events lc and nc involving lineage i:

$$\begin{aligned}
&\frac{d}{dt}P_t(L_i=b|L_{i(t=lc)}=a)\\
&=\sum_{c=1}^{m}\Big(\mu_{cb}P_t(L_i=c|L_{i(t=lc)}=a)-\mu_{bc}P_t(L_i=b|L_{i(t=lc)}=a)\Big)\\
&+P_t(L_i=b|L_{i(t=lc)}=a)\\
&\times\sum_{c=1}^{m}\lambda_c P_t(L_i=c|L_{i(t=lc)}=a)\sum_{\substack{k=1\\k\neq i}}^{n}P_t(L_k=c|T)\\
&-P_t(L_i=b|L_{i(t=lc)}=a)\lambda_{l_i}\sum_{\substack{k=1\\k\neq i}}^{n}P_t(L_k=b|T),
\end{aligned} \quad (6)$$

The entries in positions (a, b) of the matrix M_i are then the solution of the above differential equation at time nc with initial values 1 if $b=a$ and 0 otherwise. In other words, Equation (6) describes the same equation as (1) with lineage i starting in a [rather than in any state as in Equation (1)], assuming that all other lineages, other than i, evolve according to Equation (1).

2.3.3 Forwards calculation of node states including all information in the phylogeny

Based on Section 2.2, we know the probabilities of every internal node being in any state. Based on Section 2.3.2, we know how these probabilities change between coalescent events. Going backwards, we calculate $P_t^{\mathrm{bw}}(L_i=a|T)$, which only includes information up to time t. At the root however $P_t^{\mathrm{bw}}(L_{\mathrm{root}}=a|T)$ includes information from the full phylogenetic tree from time 0 up to the time of the root. We hence write $P_t^{\mathrm{bw}}(L_{\mathrm{root}}=a|T)=P_t^{\mathrm{fw}}(L_{\mathrm{root}}=a|T)$, that is the forwards probability of the root being in any state. The forwards probabilities denote the probability of a lineage being in a state that includes information from the full phylogenetic tree. We use the forwards probability $P_t^{\mathrm{fw}}(L_{\mathrm{root}}=a|T)$ at the root as a starting point to calculate $P_t^{fw}(L_i=a|T)$ for every lineage i. From the root, we proceed forwards in the tree to calculate $P_t^{fw}(L_i=a|T)$ for every internal node at the time t of the coalescent event for which lineage i was the parent lineage. $P_t^{\mathrm{fw}}(L_i=a|T)$ could be calculated at other times as well, we here however focus on the state of nodes. This we do as follows:

$$\vec{p}_i^{\,\mathrm{fw}}(t=lc)=\frac{\left(\frac{\vec{p}_p^{\,\mathrm{fw}}(t=nc)}{\vec{p}_i^{\,\mathrm{bw}}(t=nc)}M_i^{\top}\right)\cdot\vec{p}_i^{\,\mathrm{bw}}(t=lc)}{\|\left(\frac{\vec{p}_p^{\,\mathrm{fw}}(t=nc)}{\vec{p}_i^{\,\mathrm{bw}}(t=nc)}M_i^{\top}\right)\cdot\vec{p}_i^{\,\mathrm{bw}}(t=lc)\|_1} \quad (7)$$

with $\vec{p}_p^{\,\mathrm{fw}}(t=nc)/\vec{p}_i^{\,\mathrm{bw}}(t=nc)$ denoting the element-wise division of $\vec{p}_p^{\,fw}(t=nc)$, the parent lineage of i at the time nc of the coalescent event with $\vec{p}_i^{\,\mathrm{bw}}(t=nc)$, which is the daughter lineage at that time. $\vec{p}_p^{\,\mathrm{fw}}(t=nc)/\vec{p}_i^{\,\mathrm{bw}}(t=nc)$ denotes the information of the state of the parent lineage p that does not come from lineage i. The multiplication with transposed matrix $M_i^{\top}$ then denotes how much these probabilities have changed until the time of the last coalescent event lc, where lineage i was the parent lineage. The element-wise multiplication with $\vec{p}_i^{\,\mathrm{bw}}(t=lc)$ then combines this information with the one from the backwards step. Or in other words, it denotes the updated probability from the forwards and the backwards node state probabilities. After normalization to ensure that $||\vec{p}_i^{\,\mathrm{fw}}(t=lc)||_1=1$, we get the forwards probability of lineage i being in any state at time lc.

2.4 Integration of the differential equations

To integrate Equation (1), we used a second-order Taylor method with third-order step size estimation. This integration technique is similar to the very basic Euler integration, but makes use of the second derivative as well:

$$y_{t+1}=y_t+y'_t\Delta t+\frac{1}{2}y''_t\Delta t^2+O(h) \quad (8)$$

$O(h)$ stands for derivatives higher than second order. The error that is made by only considering the first and second derivative can be calculated as follows:

$$y_{t+1\,\mathrm{est}}-y_{t+1\,\mathrm{true}}=O(h)=\epsilon \quad (9)$$

With $y_{t+1\,\mathrm{est}}$ being the updated term using Equation (8) and $y_{t+1\,\mathrm{true}}$ being the hypothetical true value if all derivatives would be considered. We now assume that the Taylor term of every derivative higher

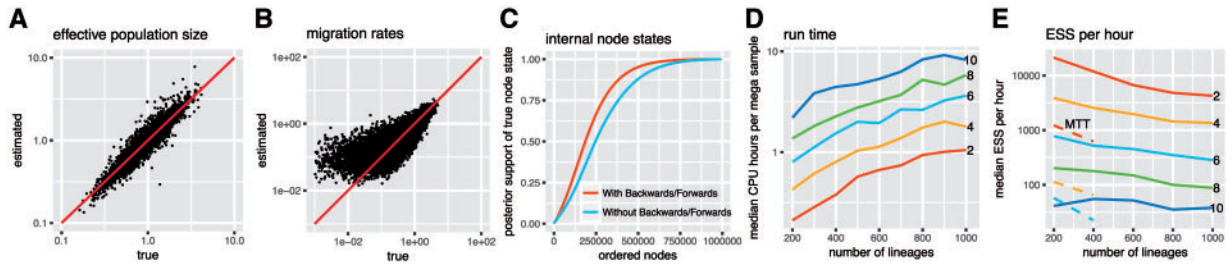

Fig. 2. Inference of effective population sizes, migration rates and node states. (A) Inferred effective population sizes on the *y*-axis versus true effective population sizes on the *x*-axis. The effective population sizes for the tree simulations were sampled from a lognormal (-0.125, 0.5) distribution. The coverage of migration rate estimates was 95.5% and for effective population size estimates 94.9%. (B) Inferred migration rates on the *y*-axis versus true migration rates on the *x*-axis. The migration rates between states were sampled from an exponential distribution with mean = 0.5. (C) Inferred node states using MASCOT with and without the backwards/forwards algorithm. (D) Median CPU time per mega sample depending on the number of lineages and the number of different states. The CPU time was taken from 100 replicates of the simulation scenario used. (E) Median posterior ESS per hour from 100 replicates from MASCOT and MultiTypeTree (dashed lines). The different colours indicate the different number of states. Dashed lines show median ESS per hour values for MultiTypeTree

than the third are zero. The error ϵ we introduce at every step can therefore be approximated as:

$$\epsilon \approx \frac{1}{6} y_t''' \Delta t^3 \tag{10}$$

For every integration step, we now choose Δt such that the absolute value of ϵ is smaller than a specified value. While we calculate the second derivative exactly (see Supplementary Material), we approximate the third derivative, assuming that the sum of probability mass in each state and that the sum of the derivatives of lineage i coalescing in any state is constant (see Supplementary Material). We then use this to update Equation (1) between sampling and coalescent events. At the root of the tree, the probability of the tree under the approximate structured coalescent is then calculated by solving Equation (3) at the root. This is covered in more depth in Müller *et al.* (2017).

2.5 Software

The method above is implemented into our BEAST 2 package MASCOT (Marginal Approximation of the Structured COalsescenT) and the analyses were done using version 1.0.0 and BEAST v2.5.0 (Bouckaert *et al.*, 2014). 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 2.4.7. MultiTypeTree analyses were performed using version 6.3.1 (Vaughan *et al.*, 2014) and BEAST 2.4.7. Script generation and post-processing were performed in Matlab R2015b. Plotting was done in R 3.2.3 using ggplot2 (Wickham, 2009) and igraph 1.1.2 (Csardi and Nepusz, 2006). 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 (ESS) for MCMC runs were calculated for the posterior probability after a burn-in of 10% using coda 0.18-1 (Plummer *et al.*, 2006). Parameter and state inference of the simulated data were only used if the posterior had an ESS of at least 100 and discarded elsewise.

2.6 Data availability

The source code for MASCOT is available at https://github.com/nicfel/Mascot.git. All scripts for performing the simulations and analyses presented in this article are available at https://github.com/nicfel/Mascot-Material.git, including the MASCOT xml file of the H3N2 analysis. Output files from these analyses, which are not on the github folder, are available upon request from the authors. A tutorial is available through the Taming the BEAST project (Barido-Sottani *et al.*, 2017) on how to use MASCOT and its BEAUti interface is available at https://github.com/nicfel/Mascot-Tutorial.git.

3 Results

3.1 Inference of migration rates, effective population sizes and internal node states

First, we tested how well effective population sizes and migration rates are inferred using MASCOT. We simulated 1000 trees with MASTER (Vaughan and Drummond, 2013) using randomly sampled effective population sizes from LogNormal Distribution($\mu = -0.125$, $\sigma = 0.5$) and migration rates from an exponential distribution with mean = 0.5. We used 1000 tips and 6 different states. In order to have scenarios of under- and over-sampling of states, we randomly sampled the number of tips in each state. The number of samples per state was randomly drawn to be a value between 20 and 1000 in increments of 20, conditional on the overall number of samples in each state being exactly 1000. We then inferred the effective population size of every state and the migration rates between each state using MASCOT from fixed phylogenies. The results of these simulations are summarized in Figure 2.

Both effective population sizes and migration rates are inferred well. Population size estimates are however much more precise than estimates of migration rates, see Figure 2. This is expected since there are typically many fewer migration events in a phylogeny than coalescent events. Additionally, the number of migration rate parameters estimated (30) is much larger than the number of effective population size parameters (6). The estimates are well correlated with the truth, only at lower migration rates do estimates become worse. This is also to be expected since a low migration rate automatically means less events which will put the estimates closer to the prior (exponential with mean 1). The coverage is 95% for both migration rate estimates and effective population size estimates. We further inferred the state of each internal node with and without the backwards/forwards algorithm. Using the backwards/forwards algorithm reduces the probability mass that is attributed to the wrong node states in this scenario (Fig. 2C).

To estimate how the CPU time varies with the number of lineages and states, we used the same framework but with varying numbers of states and lineages. The CPU time per million samples depends approximately linearly on the number of sampled sequences. With an increasing number of states, the calculation time per million sample increases approximately quadratically.

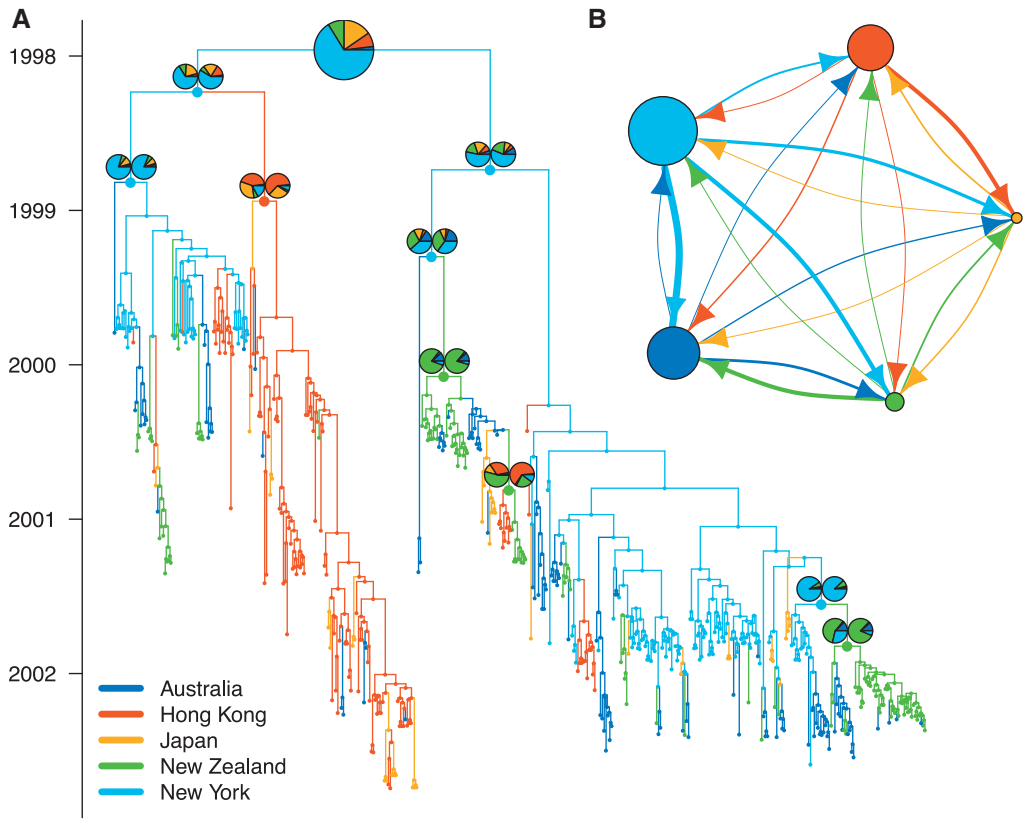

Fig. 3. MASCOT analysis of globally sampled Influenza A/H3N2 viruses. (A) Here we show the maximum clade credibility tree inferred from H3N2 sequences from Australia, Hong Kong, New York, New Zealand and Japan. The colour of each branch indicates the most likely state of its daughter node. The pie charts indicate the probability of chosen nodes being in any of the possible states. The left pie chart is the probability inferred using the backwards/forwards algorithm and the right pie chart without using the backwards/forwards algorithm. Since at the root, these probabilities are the same, only one chart is shown. The node heights are the median node heights. (B) The median inferred immigration rates as indicated by the width of the arrow. The wider an arrow into a location, the more likely it is that a lineage in the destination originated from the source location of that arrow. The different source locations are denoted by the colours of the arrows. A wider arrow from light blue to green than from red to green shows that lineages in New Zealand are more likely to have originated from New York than from Hong Kong. The dot sizes are proportional to the median inferred effective population sizes of that state

We next compared the convergence properties of MASCOT with MultiTypeTree version 6.3.1 (Vaughan *et al.*, 2014), which is based on the exact structured coalescent model without approximation, but requires MCMC sampling of lineage states. To do so, we compared the ESS per hour to MultiTypeTree (Vaughan *et al.*, 2014) using fixed trees. MASCOT shows much higher ESS values per hour in each scenario, demonstrating the drastically improved convergence properties originating from not having to sample migration histories. While these estimates can vary based on the parameters under which simulations were performed and based on the MCMC operator setup used, they show the benefits of integrating over migration histories.

3.2 Application to H3N2

We then applied MASCOT to 433 Influenza A/H3N2 sequences sampled between 2000 and 2003 from Australia, Hong Kong, New York, New Zealand and Japan. We ran five independent chains each for 120 Million iterations using an HKY + Γ_4 site model with a fixed clock rate of $5 * 10^{-3}$ substitutions per site and year. We fixed the clock rate due to a lack of temporal information from the sequences collected for this short amount of time. We then inferred the phylogenetic tree as well as the effective population sizes of every location, the migration rates between them, as well as the additional parameters from the HKY + Γ_4 model.

Figure 3 shows the maximum clade credibility tree with the different colours indicating the maximum posterior location estimate of each node. The pie charts indicate the probability of the marked nodes being in any possible location inferred with and without the backwards/forwards algorithm. These probabilities are the average over all the node state probabilities for each tree in the posterior containing that clade. We inferred New York to be a source location mainly for strains in Australia and New Zealand. Strains from Japan were inferred to originate mainly from Hong Kong and New York. The root of the phylogeny was inferred to be most likely in New York. The lack of samples near the root however makes the inference of its location unreliable.

4 Discussion

We provide a new algorithm to calculate the state of any node in a phylogeny under the marginal approximation of the structured coalescent (Müller *et al.*, 2017). This algorithm entirely avoids the sampling of migration histories. Additionally, we improve the calculation time of our previously introduced approximation to allow for the analysis of phylogenies with more samples and more states.

We have shown on simulated data that our approach is able to infer migration rates and effective population sizes reliably even when many different states (6) are present. This is a case where exact

methods that sample migration histories are currently not able to reach convergence. Even though MASCOT is an approximation, we reach a coverage of 95% for migration rates and effective population size estimates.

We also showed on simulated data that adding a backwards/forwards approach for the calculation of node states improves the inference of internal nodes. We use the backwards/forwards to calculate the state of every internal nodes in a way that is consistent with the complete phylogeny, which is not given by the backwards step alone. To estimate the probability of a node being in any possible state given a set of parameters we therefore do not need to average over many MCMC samples of migration histories. Whereas for some nodes the difference between with and without backwards/forwards is small, it is especially large for nodes where the difference in where the node is inferred to be compared with the parent node is large. This is due to conflicting information of the state of a node from the backwards and forwards step. This can for example be the result of a sampling event that adds information of where a lineage was further in the past. Future extensions could include an explicit sampling of migration histories or of the number of state changes. To do so, an algorithm similar to that of Minin and Suchard (2007, 2008) could be deployed, but lineage states would need to be sampled in a way that is probabilistically consistent with our equations for the forward line state probabilities. Finally, we applied MASCOT to a globally sampled H3N2 dataset where we inferred the phylogenetic tree and associated parameters. Our approach is able to reach convergence, even when a large number of sequences and different locations is present. The calculation time still causes challenges in the analysis of very large datasets. These could be circumvented by a further approximation of $\sum_{k=1}^{n} \approx \sum_{k=1}^{n}$ in Equation (1) when many lineages are present. This would allow every lineage to have the same transition probabilities and would therefore reduce the number of ODEs that have to be solved.

MASCOT still requires all migration rates and effective population sizes to be inferred. Especially the number of migration rates (states $*$ (states $-$ 1)) can become problematic relatively fast. Future additions could however reduce the parameter space by for example deploying Bayesian Search Variable Selection (Lemey *et al.*, 2009) or by making use of generalized linear models (Lemey *et al.*, 2014) to describe migration rates as a combination of different covariates and hence only require the parameters of the GLM model to be inferred.

Acknowledgements

We thank Remco Bouckaert and Tim Vaughan for helpful comments on the implementation of MASCOT. We also thank three anonymous reviewers for the helpful comments on the manuscript.

Funding

N.M. and T.S. are funded in part by the Swiss National Science Foundation (SNF; grant number CR32I3_166258). D.R. is funded by the ETH Zürich Postdoctoral Fellowship Program and the Marie Curie Actions for People COFUND Program. T.S. is supported in part by the European Research Council under the Seventh Framework Programme of the European Commission (PhyPD: grant agreement number 335529).

Conflict of Interest: none declared.

References

Barido-Sottani,J. *et al.* (2017) Taming the BEAST—A community teaching material resource for BEAST 2. *Systemat. Biol*, **67**, 170–174.

Beerli,P. and Felsenstein,J. (2001) Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proc. Natl. Acad. Sci. USA*, **98**, 4563–4568.

Bouckaert,R. *et al.* (2014) BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.*, **10**, e1003537.

Csardi,G. and Nepusz,T. (2006) The igraph software package for complex network research. *InterJournal* ***Complex Syst.***, **1695**, 1–9.

De Maio,N. *et al.* (2015) New routes to phylogeography: a Bayesian structured coalescent approximation. *PLoS Genet.*, **11**, e1005421.

Ewing,G. *et al.* (2004) Using temporally spaced sequences to simultaneously estimate migration rates, mutation rate and population sizes in measurably evolving populations. *Genetics*, **168**, 2407–2420.

Hudson,R.R. (1990) Gene genealogies and the coalescent process. *Oxf. Surv. Evol. Biol.*, **7**, 44.

Lemey,P. *et al.* (2009) Bayesian phylogeography finds its roots. *PLoS Comput. Biol.*, **5**, e1000520.

Lemey,P. *et al.* (2014) Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. *PLoS Pathogens*, **10**, e1003932.

Minin,V.N. and Suchard,M.A. (2007) Counting labeled transitions in continuous-time Markov models of evolution. *J. Math. Biol.*, **56**, 391–412.

Minin,V.N. and Suchard,M.A. (2008) Fast, accurate and simulation-free stochastic mapping. *Philos.Trans. R. Soc. B Biol. Sci.*, **363**, 3985–3995.

Müller,N.F. *et al.* (2017) The structured coalescent and its approximations. *Mol. Biol. Evol.*, **34**, 2970–2981.

Notohara,M. (1990) The coalescent and the genealogical process in geographically structured population. *J. Math. Biol.*, **29**, 59–75.

Paradis,E. *et al.* (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**, 289–290.

Pearl,J. (1982) *Reverend Bayes on Inference Engines: A Distributed Hierarchical Approach*. University of California, Los Angeles.

Plummer,M. *et al.* (2006) Coda: convergence diagnosis and output analysis for mcmc. *R News*, **6**, 7–11.

Pybus,O.G. *et al.* (2001) The epidemic behavior of the hepatitis c virus. *Science*, **292**, 2323–2325.

Revell,L.J. (2012) phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.*, **3**, 217–223.

Russell,C.A. *et al.* (2008) The global circulation of seasonal influenza A (H3N2) viruses. *Science*, **320**, 340–346.

Stadler,T. and Bonhoeffer,S. (2013) Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods. *Philos. Trans. R. Soc. B Biol. Sci.*, **368**, 20120198–20120198.

Takahata,N. (1988) The coalescent in two partially isolated diffusion populations. *Genet. Res.*, **52**, 213–222.

Vaughan,T.G. and Drummond,A.J. (2013) A stochastic simulator of birth-death master equations with application to phylodynamics. *Mol. Biol. Evol.*, **30**, 1480–1493.

Vaughan,T.G. *et al.* (2014) Efficient Bayesian inference under the structured coalescent. *Bioinformatics*, **30**, 2272–2279.

Volz,E.M. (2012) Complex population dynamics and the coalescent under neutrality. *Genetics*, **190**, 187–201.

Wickham,H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York.