


Joint Inference of Migration and Reassortment Patterns for Viruses with Segmented Genomes

Ugnė Stolz ^{1,2,*} Tanja Stadler,^{1,2} Nicola F. Müller,^{1,2,3,†} and Timothy G. Vaughan ^{1,2,*†}

¹Department of Biosystems Science and Engineering, ETH Zürich, Basel, Switzerland

²Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland

³Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA

[†]These authors contributed equally to this work.

*Corresponding authors: E-mails: ugne.stolz@bsse.ethz.ch; timothy.vaughan@bsse.ethz.ch.

Associate editor: Keith Crandall

Abstract

The structured coalescent allows inferring migration patterns between viral subpopulations from genetic sequence data. However, these analyses typically assume that no genetic recombination process impacted the sequence evolution of pathogens. For segmented viruses, such as influenza, that can undergo reassortment this assumption is broken. Reassortment reshuffles the segments of different parent lineages upon a coinfection event, which means that the shared history of viruses has to be represented by a network instead of a tree. Therefore, full genome analyses of such viruses are complex or even impossible. Although this problem has been addressed for unstructured populations, it is still impossible to account for population structure, such as induced by different host populations, whereas also accounting for reassortment. We address this by extending the structured coalescent to account for reassortment and present a framework for investigating possible ties between reassortment and migration (host jump) events. This method can accurately estimate subpopulation dependent effective populations sizes, reassortment, and migration rates from simulated data. Additionally, we apply the new model to avian influenza A/H5N1 sequences, sampled from two avian host types, Anseriformes and Galliformes. We contrast our results with a structured coalescent without reassortment inference, which assumes independently evolving segments. This reveals that taking into account segment reassortment and using sequencing data from several viral segments for joint phylodynamic inference leads to different estimates for effective population sizes, migration, and clock rates. This new model is implemented as the Structured Coalescent with Reassortment package for BEAST 2.5 and is available at <https://github.com/jugne/SCORE>.

Key words: phylodynamics, phylogenetics, structured populations, reassortment, migration.

Introduction

Influenza viruses are continuously evolving, escaping host immunity, or switching host species. Additionally, influenza diversity is promoted by intermediate mammalian hosts, such as swine, which can serve as a mixing vessel for human, avian, and its own influenza strains, with subsequent spillover back to the human population (Khiabani et al. 2009; Ma et al. 2009). This pattern manifests through *reassortment*—a form of recombination in segmented viruses—where different strains infect a single host cell and exchange genetic segments upon replication (McDonald et al. 2016).

Like other forms of recombination, reassortment poses challenges for phylodynamic inference methods that seek to infer population dynamics from influenza virus sequences. This is because, without an explicit model for the reassortment process, the individual genomic segments of influenza virus genomes must be analyzed in isolation to avoid making incorrect assumptions about the degree to which ancestry is shared between segments.

To address this, we recently introduced the coalescent with reassortment model (motivated by the coalescent with recombination model of Hudson [1983]), which explicitly accounts for reassortment between influenza genome segments (Müller et al. 2020). Together with a Markov chain Monte Carlo (MCMC) algorithm for sampling from the posterior distribution of reassortment networks, this allows genetic data from all genome segments to be incorporated into a single Bayesian phylodynamic analysis, even in the presence of substantial reassortment. The model ignores other possible kinds of recombination (such as template-switching) and treats viral segments as distinct molecules comprising the RNA of the virus (McDonald et al. 2016). This means that segment reassortment events have clear boundaries on the genome.

Here, we extend this model in order to account for the population structure that is introduced by having different subpopulations. These subpopulations or types can, for example, be different host species or coarse-grained geographic

locations. To model this, we extend the structured coalescent (Hudson 1990; Notohara 1990) to account for reassortment (Müller et al. 2020). We develop both exact and approximate approaches (Müller et al. 2017) for sampling from the posterior distribution of reassortment networks, reassortment rates, effective population sizes, and migration rates under the combined model: the structured coalescent with reassortment (SCoRe). In order to gain information about when and where migration events occurred, we also implement a stochastic mapping technique (Nielsen 2002; Huelsenbeck et al. 2003) that allows us to retrieve explicit migration events on the reassortment networks.

Using simulated data, we first demonstrate that SCoRe is able to correctly estimate the effective population sizes, reassortment, and migration rates. We then show how assuming independently evolving viral segments can bias the inference of effective population sizes and migration rates. Next, we apply the SCoRe to an avian influenza A/H5N1 data set with isolates from two bird orders: Anseriformes and Galliformes. This data set comprises sequences of influenza segments Haemagglutinin (HA) and Neuraminidase (NA) sampled in 2008–2016 from the highly pathogenic avian influenza (HPAI) virus A/H5N1 Gs/Gd lineage first detected in China in 1996. The majority of the sequences are from Asia (70.4%) and Africa (29.4%), with a small proportion from Europe (0.2%) (see Materials and Methods for subsampling procedure and [supplementary table S1, Supplementary Material](#) online for breakdown by country and clade).

Anseriformes are water-fowl birds that have been previously identified as a reservoir for avian influenza viruses (Webster et al. 2006; Kim et al. 2009). Migratory patterns of wild Anseriformes facilitate the spread of influenza between different locations, whereas domesticated Anseriformes (primarily ducks) develop less pathogenic infection and play a significant role in transmitting to other domestic poultry (Li et al. 2004; Hulse-Post et al. 2005). Such spillover from Anseriformes to Galliformes (ground-feeding, mostly domesticated birds) in turn can cause further transmission to non-avian livestock as well as humans (Kaplan and Webby 2013). Our analysis is able to recover this transmission pattern and shows different estimates of model parameters than the ones obtained when assuming independently evolving segments. We also investigate the possible correlation between reassortment and migration (host jump) events, setting a framework for such studies on different influenza strains and host types.

New Approaches

Previously, performing model-based Bayesian phylogeographic inference of the migration patterns of segmented viruses required either 1) assuming that reassortment is frequent and thus that segments evolved according to independent phylogenies; 2) assuming that reassortment is rare and thus that all segments share a single phylogeny; or 3) limiting the analysis to a single segment. The first two possibilities lead to biased inference when broken, whereas the third discards data from other segments, reducing the statistical power. To address this, we introduce a SCoRe model that allows multiple segments to be analyzed jointly without biasing

inferences. This model combines the structured coalescent (Hudson 1990; Notohara 1990) and the coalescent with reassortment (Müller et al. 2020). The SCoRe models a backwards in time process where lineages can coalesce and reassort within and migrate between different subpopulations.

To allow for efficient inference under this model, we extend the marginal approximation of the structured coalescent (MASCOT, Müller et al. [2017, 2018]) to account for reassortment events. MASCOT ignores the correlations between lineage states (subpopulations) by assuming that they are pairwise independent, given the tree, and integrates over all possible migration histories by calculating the marginal probability of a lineage being in any possible state from present to past. In other words, instead of inferring the discrete state of each lineage at any point in time, we compute its probability of being in any state at any point in time. This is done by numerically solving a set of differential equations that describe how the probability of any lineage being in any state changes over time (see Materials and Methods for details). To account for reassortment events, we extend these differential equations from trees to networks. This allows us to compute the probability of observing a network under the SCoRe given a set of effective population sizes, migration rates, and reassortment rates, as well as sampling locations of the individual tips. We then use a recently developed MCMC algorithm for sampling reassortment networks (Müller et al. 2020) that allows us to infer a posterior distribution of these networks, the embedding of segments trees within those networks, migration histories, and other associated parameters.

Integrating over the ancestral migration histories improves the efficiency of the inference algorithm, but having explicit migration events can often be useful to test hypotheses. This can include whether migration events, that refer to host jump events when the structure considered is host types, are correlated with reassortment events. Our MASCOT extension, on its own, does not infer the individual migration events. Thus, we also introduce a stochastic mapping algorithm to impute these over the network. Once the mapping is complete, each network sampled from the posterior distribution is annotated with a possible sequence of reassortment, coalescent, and migration events and can therefore be used to investigate possible relationships and correlations between these events.

An early version of this method was part of Master's thesis of the first author (Jankauskaite 2019).

Results

Implementation Validation and True Network Parameter Estimation

We first ensure that our implementation of the exact variant of SCoRe (SCoRe-exact) samples from the true distribution of SCoRe. Additionally, we show that the implementation of the approximate version of SCoRe does not qualitatively distort shapes and summary statistics of these distributions. To do this, we used our MCMC algorithm to produce ensembles of reassortment networks under 1) SCoRe-exact and 2) SCoRe.

We then used direct simulation (Gillespie 1976) to produce one-third set of networks simulated under the same set of parameter values. Next, we compared the frequency distributions of network height, length, and reassortment node count from each of these ensembles (supplementary figs. S16 and S17, Supplementary Material online). To measure the difference of distributions more precisely, we calculated the Kolmogorov–Smirnov (KS) statistic as a function of the iteration count. In the exact case, the KS differences asymptote to around 0.01–0.003, and from this we conclude that distributions of network statistics match with high precision. For the approximation, the KS differences are slightly larger (<0.1), but the distributions maintain similar shapes and mean values. For the analyses below, we only apply the approximate version of SCoRe, as it is substantially faster, allowing us to apply it to considerably larger data sets (see Materials and Methods).

To demonstrate that SCoRe allows us to estimate the model parameters correctly, we considered two different well-calibrated simulation studies. First, we show that SCoRe is able to recover effective population sizes, reassortment, and migration rates when the true network is known and fixed (supplementary fig. S11, Supplementary Material online). We then show the ability of SCoRe to jointly infer the reassortment network and population parameters when the evolutionary rate of the individual segments is either high, low, or mixed for two different migration priors (see Materials and Methods). Figure 1 shows the results for high evolutionary rate (5×10^{-3} substitutions per site and year) and an exponential migration rate prior. Supplementary figures S12–S14, Supplementary Material online show the results for remaining combinations of migration rate priors and clock rates. Overall, between 91% and 98% of true parameter values fell within the 95% HPD interval (supplementary table S2, Supplementary Material online).

Joint Inference from Viral Segments Reduces the Relative Error of Model Parameters

We next compared the relative error of effective population size and migration rate posterior distributions inferred under SCoRe when the true reassortment rate is known to the distributions obtained assuming independent genomic segments under the structured coalescent model that does not include reassortment (MASCOT package, Müller et al. [2018]) from simulated sequences. The true effective population sizes, migration, and reassortment rates were randomly drawn from their respective known prior distributions. We obtained 100 such sets of parameter values and simulated sequences for four segments and two subpopulations repeatedly with low (5×10^{-4} substitutions per site and year) and high (5×10^{-3} substitutions per site and year) clock rates (see Materials and Methods for more details).

We drew the true rate values from the prior distributions of the effective population sizes and migration rates. Additionally, we fixed the reassortment rates to the true values when using SCoRe. Both methods had similar accuracy for low clock rates, with the median relative error being slightly smaller for SCoRe compared with MASCOT. In the

case of high clock rates, the difference was more pronounced, with the median relative error being up to 5.4% smaller for migration rates for SCoRe compared with MASCOT (supplementary fig. S15, Supplementary Material online).

Network and Its Parameter Inference for Avian Influenza A/H5N1

We next assembled a data set using genetic sequences of influenza A/H5N1 viruses sampled between 2008 and 2016. We then grouped the sequences into two host types, based on whether they were isolated from Anseriformes or Galliformes (see Materials and Methods for details). For each of the samples we used the genetic sequences of the two surface proteins HA and NA, for further analyses. Because we only sample influenza A subtype H5N1, we cannot make insights into HA and NA subtype labels at any particular reassortment event. However, by inspecting the structure of a network and embedded segment trees, we can evaluate the overall segment evolution and reassortment accumulation over time. We then randomly subsampled this data set into ten smaller subsets, each containing 200 sequences and ran three independent analyses for each subset under SCoRe and MASCOT in BEAST 2.5 (Bouckaert et al. 2018), using parallel tempering (Altekar et al. 2004; Müller and Bouckaert 2020). For each segment, we allowed for different evolutionary rates on the first two and the third codon position evolving under an HKY+ Γ_4 (Hasegawa et al. 1985; Yang 1994) substitution model.

The inferred posterior distributions for migration rates are bimodal (fig. 2; supplementary figs. S1 and S2, “Unfiltered,” Supplementary Material online) and the inferred bird order at the root node of the segment trees (MASCOT) and the network (SCoRe) varies with the two different modes. Furthermore, the maximum clade credibility (MCC) and maximum posterior networks show that the inferred root state strongly correlates with the majority of the sampled A/H5N1 evolution occurring in the same state (supplementary figs. S4 and S5, Supplementary Material online). Ewing et al. (2004) discuss such bimodality in migration rates for MCMC structured tree inference methods and suggest using additional prior knowledge to weight one of the migration directions. In this case, it has been shown that migratory water-fowl birds (Anseriformes) are the main natural influenza A/H5N1 reservoir (Olsen et al. 2006; Webster et al. 2006; Kim et al. 2009; Trovão et al. 2015). Furthermore, H5N1 is considered to be less pathogenic to domesticated ducks (also Anseriformes) than other poultry, thus prolonging the disease shedding period and promoting introduction into the domesticated bird populations (Li et al. 2004; Hulse-Post et al. 2005). To obtain a conditioned posterior distribution, we conditioned on the root node type. We filter the ensembles produced by MCMC, retaining only those samples in which both segment roots are associated with Anseriformes (see Materials and Methods). Figure 2 and supplementary figures S1 and S2, Supplementary Material online show that conditioning completely removed bimodality of estimated migration rate distributions for most subsets with SCoRe and all subsets with MASCOT. We rely on filtered posterior

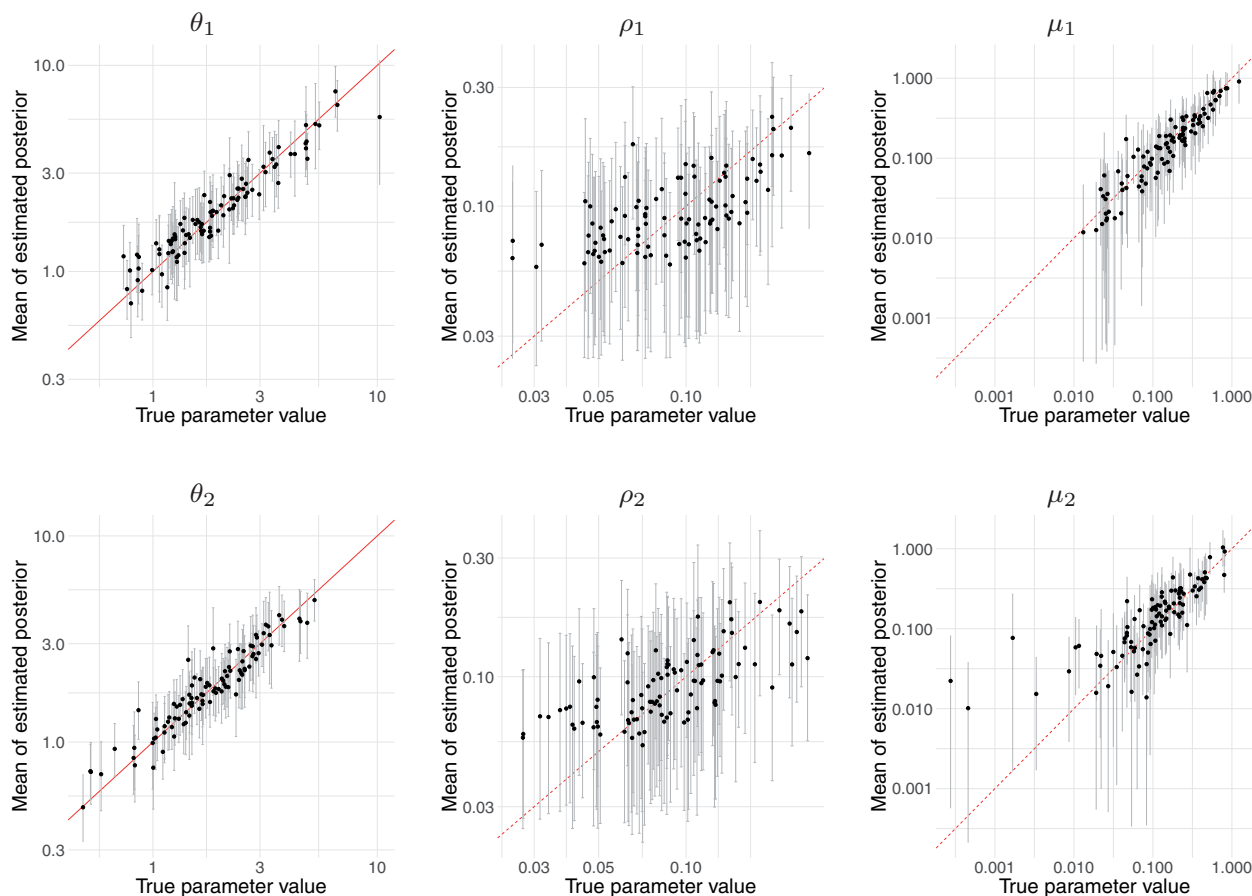


FIG. 1. Inference of effective population size (θ), reassortment (ρ), and migration (μ) rates from 100 simulated genetic sequence data for two types with exponential migration rate prior and high clock rate (5×10^{-3} substitutions per site and year for all four segments). First row of is for type 1 and second—for type 2 parameters. True (x -axis) versus estimated (y -axis) effective population sizes. Gray bars are 95% confidence intervals, red marks the $x = y$ curve.

distributions in the further analyses and present equivalent figures obtained before filtering for consistency.

With this conditioning in place, both SCoRe and MASCOT recover the expected trajectory of the virus, with high backwards in time migration from ground-feeding birds (Galliformes) to water-fowl (Anseriformes) (see [supplementary fig. S3, Supplementary Material](#) online). Backwards migration rate from Anseriformes to Galliformes is, in comparison, minuscule. The root of the MCC network as well as majority of the network length is in the Anseriformes for all ten subsets ([fig. 3; supplementary fig. S6, Supplementary Material](#) online).

Correlation between Reassortment and Migration Events in Avian Influenza A/H5N1

Next, we investigated whether there is an association between reassortment events and host jump (migration) events. To do so, we define a short-time window immediately before each host jump event in the network. We then compute the empirical rate of reassortment events within these windows (“on-window”) and compare these to the corresponding rates for all parts of the network that fall outside of these windows (“off-window”). We computed these rates for all networks sampled from the posterior distribution of

reassortment networks and then computed the difference between the rates inside and outside of the window. If this difference is above zero, the rate of reassortment is higher within the windows, that is, before host jump events. If the difference is below zero, this suggests lower rates of reassortment before host jump events.

Additionally, we computed the same difference for networks simulated under the SCoRe and the inferred parameters, that is, for posterior predictive simulations. Because a fitness effect of reassortment was shown previously from human influenza viruses ([Müller et al. 2020](#)), we also tried to disentangle general fitness benefits of reassortment from its association with host jumps. To do so, we split all networks into “fit” and “unfit” based on how long the descendants of an edge persist into the future. If they persist for more than 2 years, we classify an edge as fit and as unfit otherwise. For edges in the two classes, we again compute the difference between reassortment rates “on-” and “off-windows” before the migration events as described above.

As shown in [figure 4](#), we see a slight increase in reassortment rates before host hump events for all ten subsets. In all cases, however, the 95% highest posterior density (HPD) interval still includes no increase in rates of reassortment before host jump events. For most subsets, we also find a slight increase in rates of reassortment before host jump events

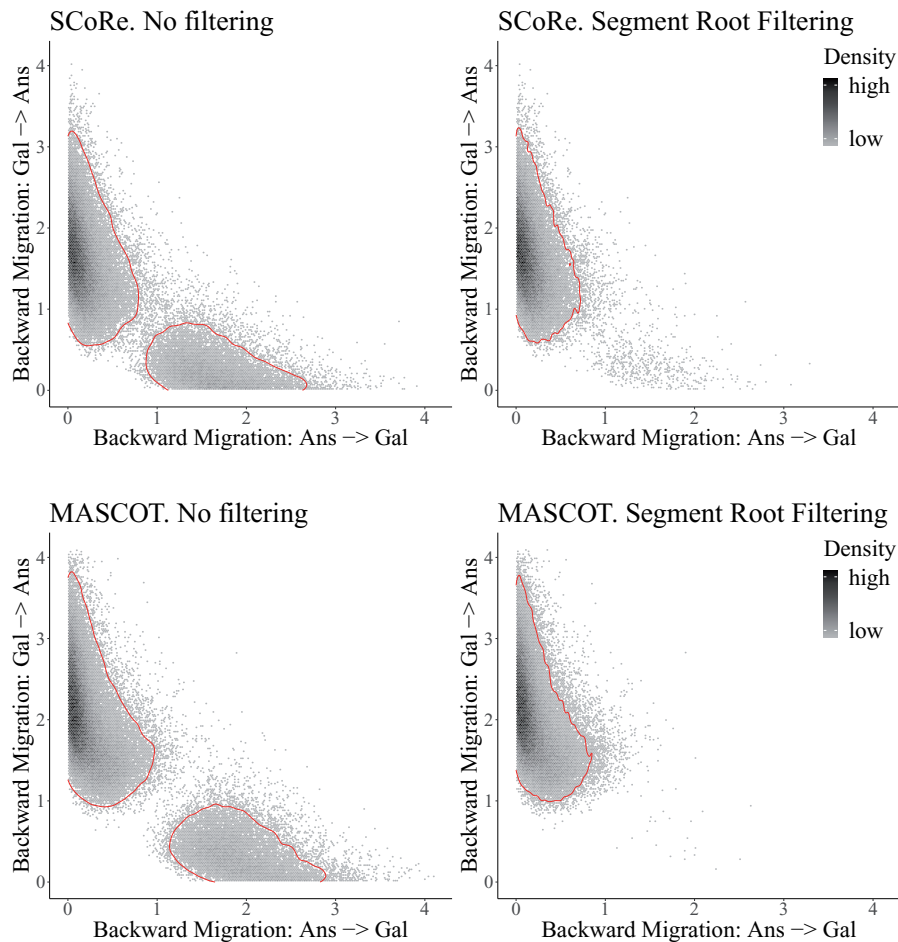


FIG. 2. Two-dimensional density of backward in time migration rate posterior estimates for all ten subsets before and after segment root filtering for MASCOT and SCoRe. Red contour line marks the 95% HPD area for the estimated 2D density.

when looking only at fit versus unfit parts of the networks. These estimates, however, are very uncertain and their 95% HPD intervals include no association at all, as well. Additionally, we looked at what happens when we change the size of the window around host jump events, as well as the definition of what constitutes a fit and unfit edge (see supplementary figs. S7–S10, [Supplementary Material](#) online for different window sizes and fitness distances), which shows largely consistent results.

Discussion

The approach presented here enables inference of reassortment networks from viruses with segmented genomes while accounting for population structure. This is done by expanding the structured coalescent approach, which was previously described, validated and compared with the competing methods in [Müller et al. \(2017, 2018\)](#), to the SCoRe. The primary goal of this extension is to allow for joint estimation of type trait evolution and reassortment in a unified framework and without a loss in accuracy, and not to improve on the computational speed of the existing inference methods for structured coalescent.

Using simulation studies, we showed that, even though this approach is approximate, our method can reliably recover

effective population sizes, reassortment, and migration rates. Furthermore, we showed that by extending MASCOT and jointly modeling coalescent, migration, and reassortment processes, SCoRe does not suffer loss in accuracy and might improve on it. The difference between the two approaches could, however, be more dramatic when considering more segments.

We find that both of these approaches—accounting for reassortment on the one hand, and assuming segments evolve independently on the other—recover similar transmission dynamics for a data set of HA and NA avian influenza A/H5N1 sequences with two host types. However, the assumption of independent segments leads to quantitatively different estimates for the evolutionary rate and backwards migration rates than those obtained by SCoRe. This is consistent with previous findings for the unstructured case ([Müller et al. 2020](#)), where assuming independently evolving segments also lead to higher evolutionary rate estimates. Additionally, both models showed a strong bimodality in the inferred migration rate distributions. This may be due to insufficient data, the omission of intermediate host types from the model, or falsely assuming constant effective population sizes and rates of migration. The first two possibilities could be addressed by using a larger amount of sequences and distinguishing between more host types or geographic

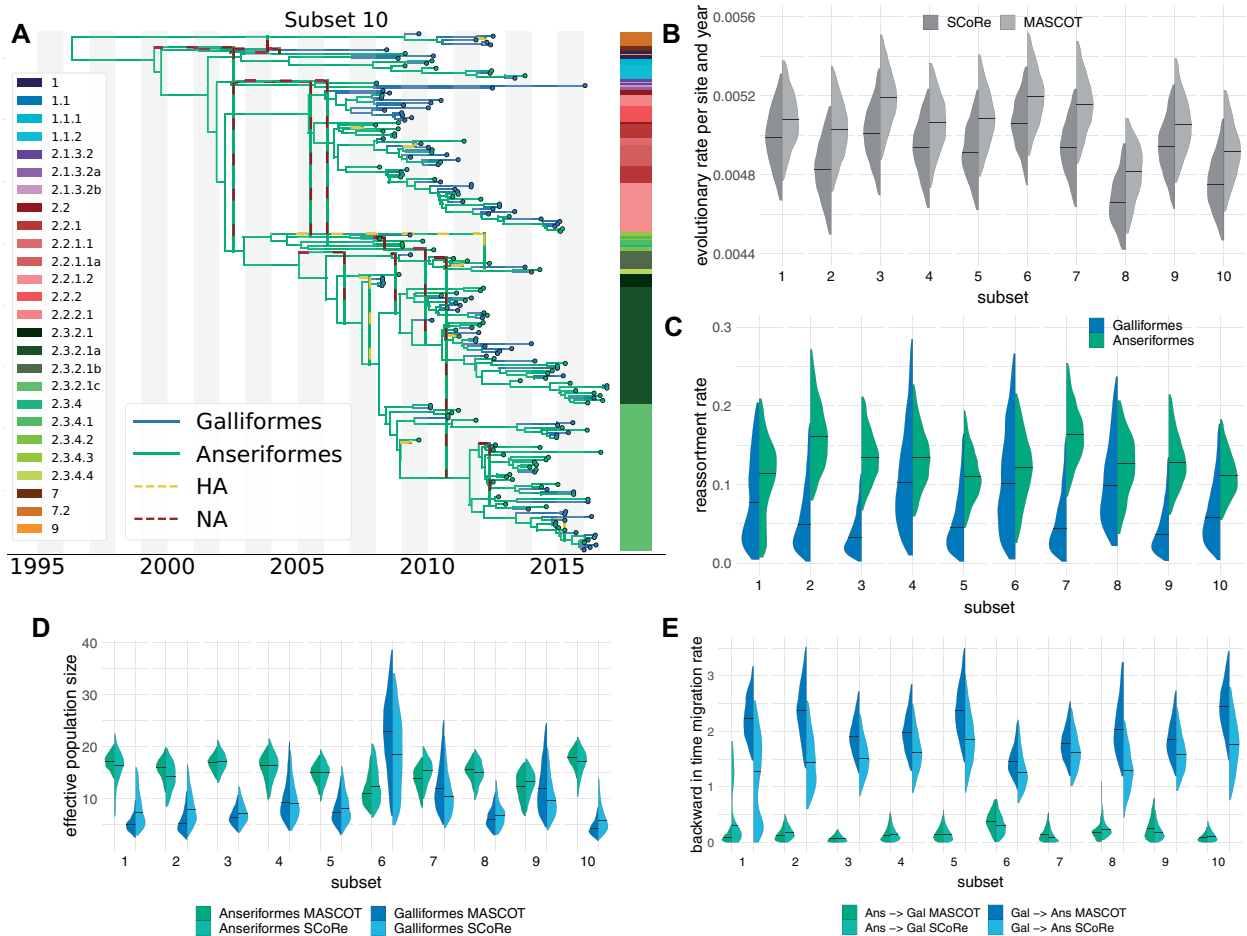


Fig. 3. MCC network and posterior parameter estimates. Segment root type filtering applied. The distributions are compared for both host types (Anseriformes and Galliformes) and for MASCOT and SCoRe packages. (A) An example MCC network for random subset ten of A/H5N1. The color column to the right of the network shows to which HPAI A/H5 clade samples belong to [Smith et al. \(2015\)](#). Dashed network lines show segment movement at reassortment events. (B) Comparison of inferred clock rates per site and year. SCoRe obtains lower estimates than MASCOT. (C) Comparison of reassortment rates per lineage and year inferred by SCoRe for two bird orders. (D) Comparison of effective population sizes. (E) Comparison of inferred backwards in time migration rates per lineage and year for Anseriformes (Ans) and Galliformes (Gal).

locations. The latter one requires extensions to the model, for example, by allowing for effective population sizes to change over time ([Drummond et al. 2005](#)).

We have also investigated possible associations between reassortment and host jump events. Although we find some evidence that reassortment rates are elevated prior to host jumps, the credible intervals did not exclude the possibility of not elevated rates. Using a data set that spans a longer time window could potentially help increase the precision of these estimates. Additionally, the role of reassortment in host switching may be stronger for more distantly related species. Although we do not find conclusive evidence for the role of reassortment in host switching for the analyzed data set, we present a framework to investigate such ties in the future.

Materials and Methods

The SCoRe

Here, we extend the coalescent with reassortment ([Müller et al. 2020](#)) process to allow for the population structure ([fig. 5](#)). Each lineage L of a network G , described by this

process, carries a full set of genomic segments, a subset of which $\mathcal{C}(L)$ are ancestral to the samples. In addition, each network lineage is in a particular type (member of a subpopulation) and migration between types is allowed. At any given time, we define a lineage L by its type l , which takes values from a type set $\{1, 2, \dots, m\}$ and ancestral segments $\mathcal{C}(L)$ it carries: $L = [l, \mathcal{C}(L)]$. Given n coexisting lineages and m types, there are m^n possible network configurations $\mathcal{K} := \{L_i = [l_i, \mathcal{C}(L_i)] | l_i \in \{1, 2, \dots, m\}\}$, w.r.t. the type of each lineage. We model the generation of SCoRe networks as a continuous backward in time Markov chain, that generates the network configuration \mathcal{K} (a state in the Markov chain) by *coalescent*, *migration*, *reassortment*, and *sampling* events.

Under a standard coalescent model, the probability per unit of time that two coexisting lineages have a common ancestor is equal to an inverse of effective population size N_e . We only allow for coalescent events between lineages L and L' if they are in the same type a . Therefore, the pairwise coalescent rate of type a , λ_a , is the inverse effective population size N_{e_a} of each subpopulation. Immediately after a

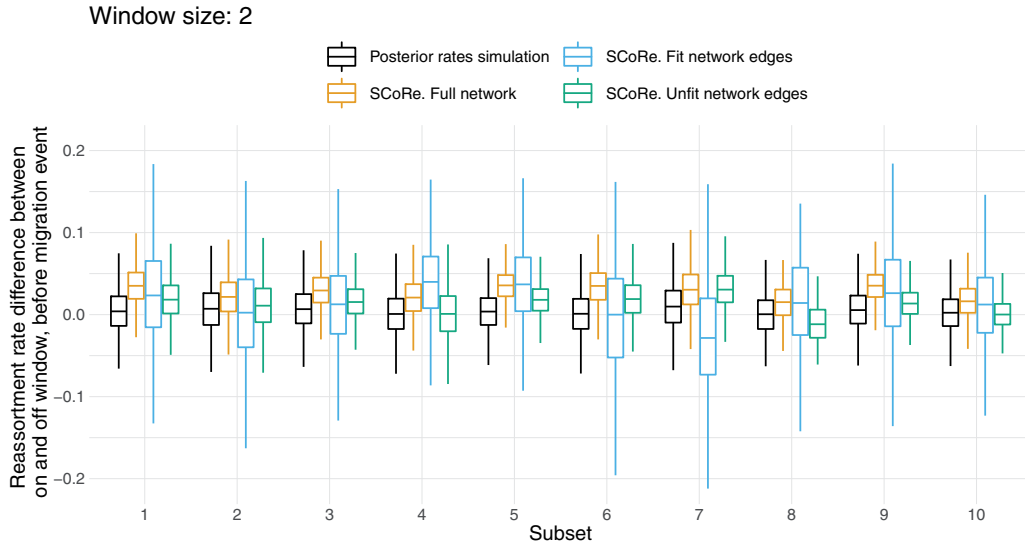


Fig. 4. Reassortment rates near host jump events compared with elsewhere. Window size: 2 years. Lineage is fit if it has descendants 2 year or more in the future. For each run, we compare distributions when (A) simulating (no sequencing data) with the posterior rates obtained by the inference or (B) inferring from the sequencing data under the SCoRe model. Because there is a previously defined increase in reassortment due to fitness, which is informed by the data, we show the reassortment rate difference for full network and separately for fit and unfit network edges.

coalescent event, the parental lineage L_p carries ancestral segments of both child lineages (Müller et al. 2020) and inherits their type $L_p = [a, \mathcal{C}(L) \cup \mathcal{C}(L')]$. If $k_a(\mathcal{K})$ is the number of lineages in type a for some configuration \mathcal{K} , then the total coalescent rate is the sum over all possible types:

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

Migration event at rate μ , measured per lineage and time unit, involves a single lineage and changes the type of that lineage, but not which segments it carries. That is, a migration event from type a to b on a lineage L changes it from $L = [a, \mathcal{C}(L)]$ to $L' = [b, \mathcal{C}(L)]$. The rate of migration event from a to b is given by μ_{ab} , with $\mu_{aa} = 0$. The total migration rate for some configuration \mathcal{K} is the sum over all n lineages to change their current type l_i into any other:

$$\mathcal{M} = \sum_{i=1}^n \sum_{a=1}^m \mu_{l_i, a}. \quad (2)$$

Upon a reassortment event occurring, we observe a lineage L and its two parents L_{p1} and L_{p2} . Each segment of the child lineage is being assigned to one of the two parental lineages L_{p1} with probability p or L_{p2} with probability $1 - p$. Here, we assume that this probability is equal for both parents: $p = 1/2$, though this assumption could be relaxed. A reassortment event is observable, if at least one ancestral segment originated from a different parent than all other segments: $\mathcal{C}(L) \cap \mathcal{C}(L_{p1}) \neq \emptyset$, $\mathcal{C}(L) \cap \mathcal{C}(L_{p2}) \neq \emptyset$ and $\mathcal{C}(L_{p1}) \cap \mathcal{C}(L_{p2}) = \emptyset$. This occurs at a probability $(1 - 2 \times (\frac{1}{2})^{|\mathcal{C}(L)|}) = (1 - (\frac{1}{2})^{|\mathcal{C}(L)|-1})$, where $|\mathcal{C}(L)|$ is the number of ancestral segments carried by lineage L . The rate of reassortment ρ is given in units per lineage and unit of time. If ρ_a is the reassortment rate at any lineage in type a

and $P_t(L_j = a | \mathcal{K}, G)$ is an indicator probability of lineage L_j being in the type a at time t , given the configuration \mathcal{K} and network G . Then, we can write total rate of observed reassortment as:

$$\mathcal{R} = \sum_{a=1}^m \rho_a \sum_{j=1}^n P_t(L_j = a | \mathcal{K}, G) \left(1 - \left(\frac{1}{2} \right)^{c(L_j)-1} \right). \quad (3)$$

As is standard in coalescent models, we condition on sampling events.

Target Posterior Probability

In order to perform MCMC sampling, we describe the target posterior distribution of networks using the Bayes theorem.

$$P(G, M, \Lambda, R, \gamma | \Sigma, A) \propto P(A | G, \gamma) P(G | \Sigma, M, \Lambda, R) P(M, \Lambda, R, \gamma). \quad (4)$$

Here, A is the set of multiple sequence alignments for each segment and Σ is the types of each sample. M , Λ , and R are, respectively, sets of migration, coalescent, and reassortment rates of the network. Finally, γ is a set of substitution model parameters.

The probability of a multiple sequence alignment (the data) given the network and substitution rate parameters $P(A | G, \gamma)$ can be factored into a sum of probabilities given the segment trees and calculated by the Felsenstein pruning algorithm (Felsenstein 1981; Müller et al. 2020). We set the joint probability of network parameters $P(M, \Lambda, R, \gamma)$ to be equal to the multiplication of their independent prior probabilities. Next, we explain how to obtain the probability of the SCoRe network $P(G | \Sigma, M, \Lambda, R)$ and use MASCOT approximation (Müller et al. 2018) to make its calculation feasible.

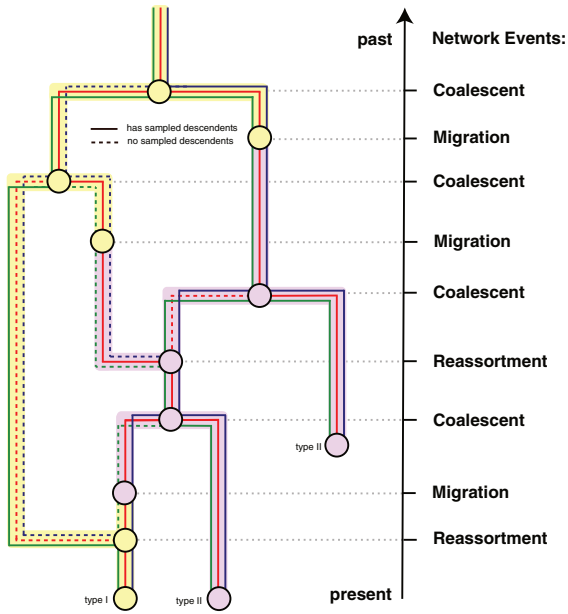


Fig. 5. Example SCoRe network. Network has two types (subpopulations) and three samples: one of type I (light yellow) and two of type II (light purple). It tracks the evolution of three genetic segments, denoted by green, red, and blue and, besides sampling, can have three kinds of events: reassortment, migration, and coalescent. Segments that are not ancestral to our samples are shown in dashed lines. Time increases into the past.

The SCoRe as a Network Prior

As described by Müller et al. (2017), we seek to marginalize over all possible migration histories H to obtain the probability of a network G .

$$P(G|\Sigma, M, \Lambda, R) = \int_H P(G, H|\Sigma, M, \Lambda, R) dH. \quad (5)$$

Equivalent to the tree case (Müller et al. 2017), the above equation for networks can be described as:

$$P(G|\Sigma, M, \Lambda, R) = \sum_{a=1}^m P_{t_{mrca}}(L_{root} = [a, \mathcal{C}(L_{root})], G) \quad (6)$$

where $P_t(\mathcal{K}, G)$ is a joint probability of the network G before the time t and the network configuration \mathcal{K} at this time. When the time t_{mrca} of the network root is reached, there is only one lineage and the probability reduces to $P_{t_{mrca}}(L_{root} = [a, \mathcal{C}(L_{root})], G)$. The ancestral segment set of the root lineage $\mathcal{C}(L_{root})$ will always contain all segments ancestral to the samples.

Approximation of the SCoRe

To obtain $P_{t_{mrca}}$, we numerically integrate P_t until $t = t_{mrca}$, where $t = 0$ is the time of the most recent sample and time flows from present to past. P_t can be factored into the contribution of network events (coalescent and reassortment) and the intervals between them. Note that there is no contribution of migration events because we marginalize over all possible migration histories. We have derived the exact equations for $P_t(\mathcal{K}, G)$ in terms of network events and intervals contributions (see Supplementary Material). However, in order to evaluate them and account for all possible network type configurations

\mathcal{K} , we need to numerically solve m^n differential equations, which becomes intractable for larger or highly structured data sets. As in Müller et al. (2018), we assume that lineages and their states are pairwise independent, given the tree.

$$P_t(L_i = [l_i, \mathcal{C}(L_i)], L_j = [l_j, \mathcal{C}(L_j)]|G) \stackrel{\text{MASCOT}}{=} P_t(L_i = [l_i, \mathcal{C}(L_i)]|G) P_t(L_j = [l_j, \mathcal{C}(L_j)]|G).$$

This approximation eliminates the need to jointly account for all possible lineage type configurations. Thus, it reduces the number of equations from m^n to $m \times n$. Next, we show how to extend this approach to the SCoRe and therefore from trees to networks.

The previously derived ODEs needed to calculate structured coalescent tree prior (Müller et al. 2017, 2018) involves migration and coalescent terms that are equivalent to the structured network case. For completeness, we restate them and include the necessary reassortment terms in order to obtain the approximate structured network prior. To evaluate the change in a marginal lineage type probability $P_t(L_i = [l_i, \mathcal{C}(L_i)], G)$ within a network interval, we obtain its derivative over time. Furthermore, to avoid numerical instability caused by vanishingly small values (Müller et al. 2018), we calculate $P_t(L_i = [l_i, \mathcal{C}(L_i)]|G) = P_t(L_i = [l_i, \mathcal{C}(L_i)], G) / P_t(G)$ instead. The derivative of which is (see Supplementary Material for derivation):

$$\begin{aligned} \frac{d}{dt} P_t(L_i = [l_i, \mathcal{C}(L_i)]|G) &= \sum_{a=1}^m (\mu_{ai} P_t(L_i = [a, \mathcal{C}(L_i)]|G) - \mu_{ia} P_t(L_i = [l_i, \mathcal{C}(L_i)]|G)) \\ &+ P_t(L_i = [l_i, \mathcal{C}(L_i)]|G) \sum_{a=1}^m \lambda_a P_t(L_i = [a, \mathcal{C}(L_i)]|G) \sum_{k \neq i}^n P_t(L_k = [a, \mathcal{C}(L_k)]|G) \\ &- P_t(L_i = [l_i, \mathcal{C}(L_i)]|G) \lambda_i \sum_{k \neq i}^n P_t(L_k = [l_i, \mathcal{C}(L_i)]|G) \\ &+ P_t(L_i = [l_i, \mathcal{C}(L_i)]|G) \sum_{a=1}^m \rho_a \left(1 - \left(\frac{1}{2}\right)^{|\mathcal{C}(L_i)|-1}\right) P_t(L_i = [a, \mathcal{C}(L_i)]|G) \\ &- P_t(L_i = [l_i, \mathcal{C}(L_i)]|G) \rho_i \left(1 - \left(\frac{1}{2}\right)^{|\mathcal{C}(L_i)|-1}\right). \end{aligned} \quad (7)$$

The migration, coalescent, and reassortment rates (μ , λ , ρ) are as described above.

Similarly, we add the reassortment term to the previously derived differential equation (Müller et al. 2018) to compute the probability for a network history G up to time t (see Supplementary Material for derivation)

$$\begin{aligned} \frac{d}{dt} P_t(G) &= -P_t(G) \sum_{a=1}^m \frac{\lambda_a}{2} \sum_{j=1}^n \sum_{k \neq j}^n P_t(L_j = [a, \mathcal{C}(L_j)]|G) P_t(L_k = [a, \mathcal{C}(L_k)]|G) \\ &- P_t(G) \sum_{a=1}^m \rho_a \sum_{j=1}^n \left(1 - \left(\frac{1}{2}\right)^{|\mathcal{C}(L_j)|-1}\right) P_t(L_j = [a, \mathcal{C}(L_j)]|G). \end{aligned} \quad (8)$$

Finally, the above probability will be modified at each coalescent or reassortment event. As in Müller et al. (2018), we wrote for the coalescent event between two lineages i and j :

$$p_t^{\text{after}}(G) = p_t^{\text{before}}(G) \sum_{a=1}^m \lambda_a P_t(L_i = [a, \mathcal{C}(L_i)] | G) P_t(L_j = [a, \mathcal{C}(L_j)] | G). \quad (9)$$

If there is a reassortment event on lineage i , we write:

$$p_t^{\text{after}}(G) = p_t^{\text{before}}(G) \sum_{a=1}^m \rho_a P_t(L_i = [a, \mathcal{C}(L_i)] | G) \left(1 - \left(\frac{1}{2}\right)^{C(L_i)-1}\right). \quad (10)$$

Numerical Integration

To numerically integrate the above differential equation, we employ second-order Taylor approximation with the numerical integration step size estimated by the third derivative. More details can be found in Müller et al. (2018), and the necessary derivatives are given in the [Supplementary Material](#).

Sampling Lineage Types

We use a stochastic mapping technique (Nielsen 2002; Huelsenbeck et al. 2003) to sample the migration (type change) history on top of a network. That is, we first perform backwards in time integration, as described above, and obtain the root node type probability distribution $\pi = (\pi_1, \dots, \pi_m)$. Then, we sample a type π_{mrcA} for the root node from π and simulate migration as a continuous time-inhomogeneous Markov process along the network branches forward in time. The times and types of the endpoints (network leaves) are assumed to be known. We add an additional constraint that both parents of a reassortment event must be in the same type immediately prior to the reassortment event. If either of the two last conditions is not met, the mapped history is rejected. Briefly, the forward simulation algorithm on a network interval between time t_0 and time t_1 is:

- Let Q be a time-dependent generator matrix with off-diagonal entries $q_{ab} > 0$ and diagonal entries $q_{aa} = -\sum_{a \neq b} q_{ab}$.
- For each coexisting lineage L_i with current type l_i draw an exponential waiting time $\tau_i \sim \text{Exponential}(Q^i)$.
- If $\min_i \tau_i > t_1$, all lineages retain their types. If the node at time t_1 is a reassortment node, we check that its parent lineages have the same simulated type at time t_1 . If not, we reject the simulation and start from the root node again. Otherwise, continue with time $t_0 := t_1$ and t_1 reset as the end of the next network interval.
- If $\tau_x = \min_i \tau_i < t_1$ simulate the new type y for a lineage L_x from a distribution with probabilities $-q_{l_x, y}^x / q_{l_x, l_x}^x$, such that L_x configuration changes from $[l_x, \mathcal{C}(L_x)]$ to $[y, \mathcal{C}(L_x)]$. Set $t_0 = \tau_x$ and repeat from the first step.

The equations needed to obtain the generator matrix Q are given in the [Supplementary Material](#). We implemented this stochastic mapping for both BEAST2 packages, SCoRe and MASCOT.

Root State Conditioning for the Structured Segment Trees and Networks

First, we will discuss the filtering of the posterior distribution for segment trees obtained by MASCOT. Let $\{a, g\}$ be the two possible types of the segment root node and p_a^j — a probability of segment tree i having the root node in type a . In order to filter the posterior tree distribution, for each pair of segment trees we sample a combination of root types (aa, ga, ag, gg) given by the probabilities $(p_a^j p_a^j, p_g^j p_a^j, p_a^j p_g^j, p_g^j p_g^j)$ and accept the trees if sample is equal to aa . Note that the required root type probabilities are already calculated by MASCOT.

In order to filter a posterior distribution of structured networks obtained by SCoRe, we first obtain the network nodes corresponding to the root of each segment tree. Then, we accept such a network if both nodes were mapped to desired type by the stochastic mapping algorithm (see Sampling Lineage Types).

Implementation

We implemented SCoRe as a BEAST 2.5 (Bouckaert et al. 2018) package that depends on the two packages CoalRe (Müller et al. 2020) and MASCOT (Müller et al. 2018). For the MCMC inference, we largely rely on the operators of CoalRe to propose new reassortment networks. However, we have to adjust parameter values supplied to the operator which resimulates unstructured network above the most recent common ancestor of all segment trees as this section of the network is not informed by the sequencing data (see “Gibbs operator” in Müller et al. [2020] for more details). In the unstructured setting, we resimulate with the most recent update of the parameter values. The network proposed by this operator would always be accepted as its Hastings ratio is the inverse ratio for the density of the current and proposed networks.

The source code can be found here: <https://github.com/jugne/SCORE> and includes tools to obtain summarized MCC networks and investigate reassortment and migration correlations. The MCC networks can be visualized with <https://icytree.org> (Vaughan 2017) or using baltic package (<https://github.com/evogytis/baltic>): <https://github.com/jugne/score-paper-material.git>. A tutorial on how to install and use SCoRe can be found here: <https://github.com/jugne/SCORE-tutorial>.

Simulation Study Setup

We have run two well-calibrated simulation studies: 1) assuming known fixed structured coalescent network and inferring its respective effective population sizes, migration, and reassortment rates; 2) inferring both—the network and its parameters. Genetic sequences for each segment were simulated according to JC69 (Jukes and Cantor 1969) substitution model. In the fixed network setting, we simulated 1,000 networks with 100 taxa, each carrying four segments, and being in one out of two types. For every simulation,

effective population sizes, reassortment, and migration rates are randomly drawn from log-normal distributions, $N_{e_a} \sim \text{LogNormal}(m = 5, \sigma^2 = 0.25)$, $\rho_a \sim \text{LogNormal}(m = 0.1, \sigma^2 = 0.25)$, $\mu_{ab} \sim \text{LogNormal}(m = 0.2, \sigma^2 = 0.25)$, for any type $a, b \in \{1, 2\}$ and $a \neq b$. Note that here m denotes the mean of a real variable. The mean for the natural logarithm of this variable can be calculated as $\ln(m) - (\sigma^2/2)$. Then, we inferred the parameter values, given the simulated networks, using the above parameter distributions as priors.

For the joint network and parameters inference, we simulated 100 networks and embedding of the segment trees for 100 taxa with four segments. The types and sampling times were drawn as described above. Clock rates were set to either high (5×10^{-3} substitutions per site and year), low (5×10^{-4} substitutions per site and year), or mixed (two segments with high and two with low). The prior distribution of a reassortment rate was the same as above, whereas we set a mean of two for the log-normal distribution of effective population sizes. For migration rates, we studied cases where the prior distributions were the same as detailed above and where it was the exponential distribution with a mean of 0.2. The different migration prior was studied because we noticed it to be a more natural choice when applying the model to the real seasonal influenza data set. Each inference was run for 48 h, and we used only those runs for which the effective sample size of posterior probability was higher than 100.

Finally, we investigated the relative error in parameter estimates obtained by SCoRe and MASCOT. We again simulated 100 networks with 100 taxa and four segments, each simulation repeated with high and low clock rates. The parameters were drawn from the following distributions: $N_{e_a} \sim \text{LogNormal}(m = 2, \sigma^2 = 0.25)$, $\rho_a \sim \text{LogNormal}(m = 0.5, \sigma^2 = 0.25)$, $\mu_{ab} \sim \text{LogNormal}(m = 0.2, \sigma^2 = 0.25)$, for any type $a, b \in \{1, 2\}$ and $a \neq b$. Higher reassortment rates were chosen to model a case where reassortment highly influences the evolution of a virus. We used the same parameter priors for both methods and additionally set the reassortment rate to the true value in SCoRe. The relative error was calculated as $|\frac{p_{\text{true}} - p_{\text{estimated}}}{p_{\text{true}}}|$, where p_{true} is the true parameter value and $p_{\text{estimated}}$ is the median parameter estimate obtained by SCoRe or MASCOT.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank the anonymous reviewers for their insightful feedback on the manuscript. We thank Claire Guinat and Sophie Seidel for valuable comments and discussions. We also thank the authors, originating and submitting laboratories who generously contributed influenza A/H5N1 sequence data to GISAID's EpiFlu Database (Shu and McCauley 2017) and the Influenza Research Database (Zhang et al. 2017). N.F.M. is funded by the Swiss National Science Foundation

(P2EZP3191891). U.S., T.G.V., and T.S. thank ETH Zürich for funding.

Data Availability

We gathered a data set of 1,410 pairs for avian influenza A/H5N1 segment HA and NA sequences for Galliformes (822) and Anseriformes (588) with complete sampling dates between 2008 and 2016 from GISAID (Shu and McCauley 2017; <https://www.gisaid.org/>) and the Influenza Research Database (Zhang et al. 2017; <http://www.fludb.org>) (see supplementary file `gisaid_and_ird_acknowledgement_table_clades.csv`, Supplementary Material online). Then, we iterate over all sampling times, and pool sequences that are 1) from the same geographic location, 2) of the same bird order, and 3) are within 30 days distance. There were 123 such location-order-date pools for Anseriformes and 159 for Galliformes. Then we randomly sampled one sequence per pool and discarded the remaining sequences in this pool. Finally, we further reduce this subsample to 100 segment sequence pairs for Anseriformes and Galliformes. The last step is done by weighted subsampling with regard to date and location. We repeat this procedure ten times, thus obtaining ten random subsamples of our data. Given random sampling from the location-order-date pools and further random reduction, ten subsampled data sets may overlap but are not identical. The BEAST2 XML files, which include sequence names, can be found on <https://github.com/jugne/score-paper-material.git>.

References

- Altekar G, Dwarkadas S, Huelsenbeck JP, Ronquist F. 2004. Parallel metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics* 20(3):407–415.
- Bouckaert R, Vaughan TG, Barido-Sottani J, and Duchene S, Fourment M, Gavryushkina A, Heled J, Jones G, Kuhnert D, De Maio N, et al. 2018. Beast 2.5: an advanced software platform for Bayesian evolutionary analysis. *BioRxiv*, 474296.
- Drummond AJ, Rambaut A, Shapiro B, Pybus OG. 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol.* 22(5):1185–1192.
- Ewing G, Nicholls G, Rodrigo A. 2004. Using temporally spaced sequences to simultaneously estimate migration rates, mutation rate and population sizes in measurably evolving populations. *Genetics* 168(4):2407–2420.
- Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol.* 17(6):368–376.
- Gillespie DT. 1976. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J Comp Phys.* 22(4):403–434.
- Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol.* 22(2):160–174.
- Hudson RR. 1983. Properties of a neutral allele model with intragenic recombination. *Theor Popul Biol.* 23(2):183–201.
- Hudson RR. 1990. Gene genealogies and the coalescent process. *Oxf Surv Evol Biol.* 7(1):44.
- Huelsenbeck JP, Nielsen R, Bollback JP. 2003. Stochastic mapping of morphological characters. *Syst Biol.* 52(2):131–158.
- Hulse-Post DJ, Sturm-Ramirez KM, Humberd J, Seiler P, Govorkova E, Krauss S, Scholtissek C, Puthavathana P, Buranathai C, Nguyen T, et al. 2005. Role of domestic ducks in the propagation and biological

- evolution of highly pathogenic H5N1 influenza viruses in Asia. *Proc Natl Acad Sci U S A*. 102(30):10682–10687.
- Jankauskaite U. 2019. Modelling viral reassortment in structured populations [master's thesis]. ETH Zurich.
- Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In: HN Munro, editor. *Mammalian protein metabolism*. New York: Academic Press. p. 21–132.
- Kaplan BS, Webby RJ. 2013. The avian and mammalian host range of highly pathogenic avian H5N1 influenza. *Virus Res*. 178(1):3–11.
- Khiabani H, Trifonov V, Rabadan R. 2009. Reassortment patterns in swine influenza viruses. *PLoS One* 4(10):e7366.
- Kim J, Negovetich NJ, Forrest HL, Webster RG. 2009. Ducks: the “trojan horses” of H5N1 influenza. *Influenza Other Respir Viruses* 3(4):121–128.
- Li KS, Guan Y, Wang J, Smith GJD, Xu K, Duan L, Rahardjo AP, Puthavathana P, Buranathai C, Nguyen TD, et al. 2004. Genesis of a highly pathogenic and potentially pandemic H5N1 influenza virus in Eastern Asia. *Nature* 430(6996):209–213.
- Ma W, Kahn RE, Richt JA. 2009. The pig as a mixing vessel for influenza viruses: human and veterinary implications. *J Mol Genet Med*. 03(01):158.
- McDonald SM, Nelson MI, Turner PE, Patton JT. 2016. Reassortment in segmented RNA viruses: mechanisms and outcomes. *Nat Rev Microbiol*. 14(7):448–460.
- Müller NF, Bouckaert R. 2020. Adaptive metropolis-coupled MCMC for beast 2. *PeerJ*. 8:e9473.
- Müller NF, Rasmussen DA, Stadler T. 2017. The structured coalescent and its approximations. *Mol Biol Evol*. 34(11):2970–2981.
- Müller NF, Rasmussen DA, Stadler T. 2018. Mascot: parameter and state inference under the marginal structured coalescent approximation. *Bioinformatics* 34(22):3843–3848.
- Müller NF, Stolz U, Dudas G, Stadler T, Vaughan TG. 2020. Bayesian inference of reassortment networks reveals fitness benefits of reassortment in human influenza viruses. *Proc Natl Acad Sci U S A*. 117(29):17104–17111.
- Nielsen R. 2002. Mapping mutations on phylogenies. *Syst Biol*. 51(5):729–739.
- Notohara M. 1990. The coalescent and the genealogical process in geographically structured population. *J Math Biol*. 29(1):59–75.
- Olsen B, Munster VJ, Wallensten A, Waldenström J, Osterhaus AD, Fouchier RA. 2006. Global patterns of influenza a virus in wild birds. *Science* 312(5772):384–388.
- Shu Y, McCauley J. 2017. GISAID: global initiative on sharing all influenza data—from vision to reality. *Eurosurveillance* 22(13):30494.
- Smith GJD, Donis RO; World Health Organization/World Organisation for Animal Health/Food and Agriculture Organization (WHO/OIE/FAO) H5 Evolution Working Group. 2015. Nomenclature updates resulting from the evolution of avian influenza A(H5) virus clades 2.1.3.2a, 2.2.1, and 2.3.4 during 2013–2014. *Influenza Other Respir Viruses* 9(5):271–276.
- Trovão NS, Suchard MA, Baele G, Gilbert M, Lemey P. 2015. Bayesian inference reveals host-specific contributions to the epidemic expansion of influenza a H5N1. *Mol Biol Evol*. 32(12):3264–3275.
- Vaughan TG. 2017. IcyTree: rapid browser-based visualization for phylogenetic trees and networks. *Bioinformatics* 33(15):2392–2394.
- Webster RG, Peiris M, Chen H, Guan Y. 2006. H5N1 outbreaks and enzootic influenza. *Biodiversity* 7(1):51–55.
- Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol*. 39(3):306–314.
- Zhang Y, Aevermann BD, Anderson TK, Burke DF, Dauphin G, Gu Z, He S, Kumar S, Larsen CN, Lee AJ, et al. 2017. Influenza research database: an integrated bioinformatics resource for influenza virus research. *Nucleic Acids Res*. 45(D1):D466–D474.