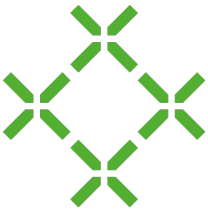
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich



D-BSSE
Department of Biosystems
Science and Engineering

Modelling Viral Reassortment in Structured Populations

Master Thesis

Ugnė Jankauskaitė

jugne@student.ethz.ch

Department of Biosystems Science and Engineering
Computational Evolution Group
ETH Zürich

Supervisors:
Prof. Dr. Tanja Stadler, Dr. Timothy Vaughan, Nicola Müller

June 28, 2019

Abstract

Migration dynamics of viral sub-populations are usually analysed by applying the structured coalescent theory to genetic sequence data. This is, however, not trivial for reassorting viruses that have segmented genomes. Recently, a coalescent with reassortment approach was developed, allowing to use the full segmented genome for phylodynamic analysis. We combine the two techniques into the structured coalescent with reassortment model for exact and approximate inference. We show that this method can accurately estimate sub-population dependent effective populations sizes, reassortment and migration rates. Additionally, we apply the new model on a seasonal influenza A/H3N2 dataset with 150 genomic sequences for four viral segments, sampled at three distinct locations. We contrast our results with a structured coalescent without reassortment inference conducted using genetic sequences of one out of four segments. This revealed 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 evolutionary rates, and the height of a segment tree root node.

The discussed model is implemented as Structured COalescent with REassortment (SCORE) package for BEAST 2 and its source code available at https://github.com/jugne/SCORE.

Contents

Chapter 1

Introduction

Seasonal influenza is a challenge for healthcare systems worldwide. Influenza-induced respiratory complications are estimated to cause 290 000 – 650 000 deaths each year [1]. Additionally, to the seasonal strains, influenza pandemics pose a risk to the population. The most famous example being the Spanish flu in 1918 [2] and the most recent Swine flu pandemic in 2009 [3, 4]. To develop better prevention strategies, it is crucial to understand the sources of yearly influenza variation and transmission patterns.

Influenza is part of the *Orthomyxoviridae* family and therefore has a genome organised in several segments. When a virus enters a host cell, each segment is replicated before the full viral genome is re-assembled. In case of co-infection, segments that are repackaged together can originate from different viral strains. This process – known as *reassortment* – is a substantial source of influenza diversity [5] and increased transmission [6]. Additionally, reassortment events have been described to be at the beginning of

Until recently, no model-based approach to study reassortment existed, and mainly heuristic approaches were used to do so. One approach, for example, involved inference of individual segment trees, with lineages labelled according to a specific bifurcation event [9]. Recently, a method to jointly model coalescent with reassortment was introduced [10]. It relies on Markov Chain Monte Carlo (MCMC) sampling of the phylogenetic network. All segment trees are embedded within the network, and each lineage carries a set of segments. Reassortment events on the network modify this set over time such that not all segments carried by a lineage are ancestral to the sampled sequences. When contrasted to the coalescent without reassortment, which assumes that each segment evolved independently, the new method provides a better estimation of effective population size and clock rate as well as more certain inference of network node hights [10].

To date, only simple population models have been used. In particular, there is no way to account for the population structure. While many criteria of division may be applied, by "structure" we mean spacial separation of a population. There are several approximations [11, 12, 13, 14] addressing it within the backward in time structured coalescent framework [15, 16, 17]. However, these approaches

typically require to infer the state of each phylogenetic tree lineage over time [11, 12]. It is usually done by MCMC sampling, which rapidly increases in computational complexity with the growing number of states. In [13, 14], a different method is suggested, where each lineage is assigned a probability of being in a state. These probabilities are calculated when marginalizing over all possible migration histories in the tree and require to solve a set of ordinary differential equations (ODEs). The latter approach scales better with the number of states. However, it assumes that the state of a lineage is independent of the states of the other tree lineages [14, 18].

The marginal approximation of the structured coalescent [18], on the other hand, allows to jointly infer all possible configurations for lineage states. While this method is able to reduce biases introduced by the lineage state independence assumption, it is not computationally efficient. To address this, a marginal approximation of structured coalescent (MASCOT) was developed [18, 19]. It retains information about the other lineages when calculating the state probabilities for a lineage and can recover a structured coalescent tree with its parameters more accurately than approximate approaches of [13, 14].

While separate methods exist to model structured coalescent or reassortment in unstructured coalescent case, currently, there is no single model capable of using the full segmented genomes to model structured coalescent with reassortment. To address this gap, we here introduce a novel approach by enhancing the marginal approximation of structured coalescent to include reassortment dynamics. To do so, we first derive ordinary differential equations to account for ancestral segments and reassortment events in each lineage similar to [18, 19].

Next, using simulated data, we show that our here introduced method accurately infers state dependent reassortment and migration rates, and effective population sizes. To demonstrate the usability, we compare the run of our here introduced method against MASCOT on influenza H3N2 dataset with samples from the USA, Singapore, Hong Kong and New Zeland.

Chapter 2

Materials and Methods

2.1 The Structured Coalescent with Reassortment

Coalescent theory models the composition of a population. Going backwards in time, it gives the probability per unit time of two individuals from the same generation having a common ancestor. This probability density is called *coalescent rate*, λ and, under the standard Wright-Fisher model, is an inverse of effective population size N_e. Additionally, in our structured coalescent network model, rates at which *reassortment* and *migration* events happen are denoted as ρ and μ_{ab}, where index ab means that a network lineage is changing states from a to b. Each network lineage l carries all genomic segments, a subset of which is directly ancestral to the sampled data and is denoted $C(l)$.

Based on reassortment and coalescent event rates and conditioning on known sampling events, we model the generation of networks using a backward in time continuous Markov process [10]. We use MCMC sampling to obtain the most probable network from the target posterior distribution, where migration rates are used to marginalize over possible migration histories when calculating the network prior probability.

We first discuss the network events and their total rates. Then characterize the target posterior probability of the network and explain how it is obtained in an exact or approximated case. Last, we detail the numerical integration algorithm used to solve the ODEs and the model implementation within the BEAST 2.5 [20].

2.1.1 Network Events

To account for reassortment events, we extended the definitions of network events found in [18]. Given n coexisting lineages and m states, we can label each lineage and its state by L_i and l_i, where $i \in 1, ..., n, l_i \in 1, ..., m$. Additionally, let S_i be the variable for the number of ancestral segments carried by the lineage i, $s_i := |C(L_i)|$. Then, there are m^n possible configurations of coexisting lineages:

$\mathcal{K} := (L_1 = l_1, ..., L_i = l_i, ..., L_n = l_n, S_1 = s_1, ..., S_i = s_i, ...S_n = s_n)$. Under our model, configuration $\mathcal{K}$ may change if either coalescent, migration or reassortment event happens with time going backwards. We extend the definitions of coalescent and migration events [18] by making explicit how they change the ancestral segments carried by a lineage.

A coalescent event between lineages i and j, that are of the same state a and carry q and p ancestral segments respectively, changes the configuration as follows:

$$\begin{aligned}(L_1 =&l_1, ..., L_{i-1} = l_{i-1}, \boldsymbol{L_i = a}, L_{i+1} = l_{i+1}, ..., L_{j-1} = l_{j-1}, \boldsymbol{L_j = a}, L_{j+1} = l_{j+1}, ..., L_n = l_n,\\ S_1 =&s_1, ..., S_{i-1} = s_{i-1}, \boldsymbol{S_i = q}, S_{i+1} = s_{i+1}, ..., S_{j-1} = s_{j-1}, \boldsymbol{S_i = p}, S_{j+1} = s_{j+1}, ..., S_n = s_n)\end{aligned}$$

$$\textit{coalescent event}$$

$$\longrightarrow$$

$$\begin{aligned}(L_1 =&l_1, ..., L_{i-1} = l_{i-1}, \boldsymbol{L_i = a}, L_{i+1} = l_{i+1}, ..., L_{j-1} = l_{j-1}, L_{j+1} = l_{j+1}, ..., L_{n-1} = l_n,\\ S_1 =&s_1, ..., S_{i-1} = s_{i-1}, \boldsymbol{S_i = r}, S_{i+1} = s_{i+1}, ..., S_{j-1} = s_{j-1}, S_{j+1} = s_{j+1}, ..., S_{n-1} = s_n)\end{aligned}$$

After the coalescent event, lineages with index bigger than j are relabeled and lineage i denotes the parent of lineages i and j. The new lineage carries all ancestral segments carried by lineages i and j, $r := |C(L_i) \cup C(L_j)|$.

A migration event from state a to b at lineage i, which carries $q = |C(L_i)|$ ancestral segments, changes the configuration as follows:

$$\begin{aligned}(L_1 =&l_1, ..., L_{i-1} = l_{i-1}, \boldsymbol{L_i = a}, L_{i+1} = l_{i+1}, ..., L_n = l_n,\\ S_1 =&s_1, ..., S_{i-1} = s_{i-1}, \boldsymbol{S_i = q}, S_{i+1} = s_{i+1}, ..., S_n = s_n)\end{aligned}$$

$$\textit{migration event}$$

$$\longrightarrow$$

$$\begin{aligned}(L_1 =&l_1, ..., L_{i-1} = l_{i-1}, \boldsymbol{L_i = b}, L_{i+1} = l_{i+1}, ..., L_{n-1} = l_n,\\ S_1 =&s_1, ..., S_{i-1} = s_{i-1}, \boldsymbol{S_i = q}, S_{i+1} = s_{i+1}, ..., S_{n-1} = s_n)\end{aligned}$$

Here, only the state of lineage i changes. The number of ancestral segments, carried by any lineage stays the same.

Lastly, we introduce a *reassortment* event to the model. The reassortment event at lineage i, which is of state a and carries $q = |C(L_i)|$ number of ancestral segments, changes the configuration as follows:

$$\begin{aligned}(L_1 =&l_1, ..., L_{i-1} = l_{i-1}, \boldsymbol{L_i = a}, L_{i+1} = l_{i+1}, L_{i+2} = l_{i+2}, ..., L_n = l_n,\\ S_1 =&s_1, ..., S_{i-1} = s_{i-1}, \boldsymbol{S_i = q}, S_{i+1} = s_{i+1}, S_{i+2} = s_{i+2}...S_n = s_n)\end{aligned}$$

$$\textit{reassortment event}$$

$$\longrightarrow$$

$$\begin{aligned}(L_1 =&l_1, ..., L_{i-1} = l_{i-1}, \boldsymbol{L_i = a}, \boldsymbol{L_{i+1} = a}, L_{i+2} = l_{i+1}..., L_{n+1} = l_n,\\ S_1 =&s_1, ..., S_{i-1} = s_{i-1}, \boldsymbol{S_i = r}, \boldsymbol{S_{i+1} = q - r}, S_{i+2} = s_{i+2}...S_{n+1} = s_n).\end{aligned}$$

Going backwards in time, reassortment event in lineage i results in two new parent lineages i and $i + 1$. Let the ancestral segment count in lineage i before reassortment event be equal to q. Then, after the reassortment event, sum of the ancestral segment counts in the parent lineages i and $i + 1$ has to be equal to q. We denote these counts by r and $q - r$.

2.1.2 Rates of the Network Events

The total rates of coalescent and migration events are described in [18] and remain the same in our model. Denote pairwise coalescent rate of state a by λ_a and number of lineages in this state by $k_a(\mathcal{K})$ for some configuration $\mathcal{K}$. Then total coalescent rate is expressed by summing over the product of coalescent rate and number of possible lineage pairs for all states:

$$\mathcal{C} = \sum_{a=1}^{m} \lambda_a \binom{k_a(\mathcal{K})}{2} \tag{2.1}$$

Total migration rate for a configuration $\mathcal{K}$ is the sum over all migration rates $\mu_{l_i a}$ from state l_i of lineage i to state a for all lineages and all states:

$$\mathcal{M} = \sum_{i=1}^{n} \sum_{a=1}^{m} \mu_{l_i a} \tag{2.2}$$

Next, we introduce the total reassortment rate. Let ρ_a be a reassortment rate at any lineage of state a and s_j the number of ancestral segments carried by the lineage j, as described above. Then, the total reassortment rate is:

$$\mathcal{R} = \sum_{a=1}^{m} \rho_a \sum_{j=1}^{n} P_t(L_j = a|\mathcal{K}, G) \left(1 - \left(\frac{1}{2}\right)^{s_j - 1}\right) \tag{2.3}$$

Here, $P_t(L_j = a|\mathcal{K}, G)$ is an indicator probability of lineage L_j being of state a at time t, gven the configuration $\mathcal{K}$ and coalescent with reassortment history, given by graph G. Namely, it is equal to 1 if $L_j = a$ and 0 otherwise. $\left(1 - \left(\frac{1}{2}\right)^{s_j - 1}\right)$ is the probability of at least one ancestral segment originating from a different parent than all other segments. In other words, it is the probability that reassortment event will be observed.

2.1.3 Target Posterior Probability

Using Bayes theorem we can define the MCMC target posterior distribution as:

$$P(G, M, \Lambda, R, \theta, \gamma|\Sigma, A) \propto P(A|G, \gamma)P(G|\Sigma, M, \Lambda, R)P(\theta, \gamma, M, \Lambda, R,) \tag{2.4}$$

Here, Σ is the sampling states of the network tips, M the set of migration rates, Λ the set of coalescent rates and R the set of reassortment rates. We denote $\theta = \{N_{e_{l_i}}|l_i \in 1, ..., m\}$ a set of effective population sizes for m states, and γ a set of substitution model parameters.

The probability of the sequencing data, given the network, can be obtained by factoring it into a sum of the probabilities given segment trees which in turn

are calculated by the Felsenstein pruning algorithm ([10], [21]). Joint prior probability of the parameters $P(\theta, \gamma, M, \Lambda, R,)$ can be expressed as multiplication of independent prior probabilities $P(\theta)P(\gamma)P(M)P(\Lambda)P(R)$.

In the next few chapters, we explain how to obtain probability $P(G|\Sigma, M, \Lambda, R)$.

2.1.4 The Probability of a Structured Coalescent with Reassortment Network

Following [18], we want to obtain the probability of the structured coalescent with reassortment network G by integrating over all posible migration histories H:

$$P(G|\Sigma, M, \Lambda, R) = \int_H P(G, H|\Sigma, M, \Lambda, R)\, dH \tag{2.5}$$

Next, we extend the definition of the probability density that samples, which are more recent than time t, evolved according to a coalescent with reassortment history, given by graph G when going backward from time 0 to t. We do so by noting the ancestral segment count carried by each lineage:

$$P_t(\mathcal{K}, G) := P_t(L_1 = l_1, ..., L_i = l_i, ..., L_n = l_n, S_1 = s_1, ..., S_i = s_i, ...S_n = s_n, G)$$

This probability is conditioned on Σ, M, Λ and R. Then, as in the case of a tree without reassortment ([18]), the probability density fo the network G is obtained by summing over the probability of the root network lineage states for all possible variants:

$$P(G|\Sigma, M, \Lambda, R) = \sum_{a=1}^{m} P_{t_{mrca}}(L_1 = a, S_1 = s_1 G) \tag{2.6}$$

The time of the root lineage is denoted by t_{mrca}. The number of ancestral segments carried by the root lineage (s_1 in the above equation) is always the total number of possible segments. This probability can be obtained by developing a master equation for P_t and numerically integrating it to solve for the probability at the root. Since this master equation is conditioned on the network, its solution can be decomposed into network *event contributions* (corresponding to nodes in the network) and *interval contributions* (corresponding to the intervals between nodes in the network).

Further, we describe how to obtain these probabilities precisely and later approximate them to speed up the calculations.

2.1.5 Exact Structured Coalescent with Reassortment

We can obtain the exact form of the probabilities in question. However, they are computationally infeasible to compute for larger datasets. Therefore, we use

them solely as a basis for further approximation, as well as to validate our model on a small scale simulated data.

Interval Contribution

As in [18], we can express the probability $P_{t+\Delta t}$, as a function of P_t, assuming no network events occur in the timestep Δt. Taking the limit $\Delta t \to 0$ yields the following master equation, valid within an interval between the network events.

$$\begin{aligned}\frac{\mathrm{d}P_t(\mathcal{K}, G)}{\mathrm{d}t} = &\sum_{i=1}^{n}\sum_{a=1}^{m}\big(\mu_{al_i}P_t(L_i = a, S_i = s_i, \mathcal{K}_{\setminus i}, G) \\ &- \mu_{l_i a}P_t(\mathcal{K}, G)\big) \\ &- \sum_{a=1}^{m}\lambda_a\binom{k_a(\mathcal{K})}{2}P_t(\mathcal{K}, G) \\ &- \sum_{a=1}^{m}\rho_a\sum_{j=1}^{n}P_t(L_j = a|\mathcal{K}, G)\left(1 - \left(\frac{1}{2}\right)^{s_j - 1}\right)P_t(\mathcal{K}, G) \end{aligned} \tag{2.7}$$

Here, $\mathcal{K}_{\setminus i}$ denotes a configuration for all lineages, except lineage i. The derivation of the above equation is equivalent to the case without reassortment (see [18], equation (2)). Derivation, which includes reassortment rates, is given in the supplementary section A.

Event Contribution

Network configuration probabilities at sampling or coalescent events are described in [18]. Here, we restate them when configuration also accounts for changes in the number of ancestral segments, as explained in section 2.1.1.

- Contribution of the **sampling event**

$$\begin{aligned}&P_t^{after}(\mathcal{K}_{\setminus n+1}, L_{n+1} = l_{n+1}, S_{n+1} = s_{n+1}, G) \\ &\quad = P_t(\mathcal{K}_{\setminus n+1}, G)P_t(L_{n+1} = l_{n+1}, S_{n+1} = s_{n+1}|G)\end{aligned}$$

  Here The number of ancestral segments carried by lineage $n+1$ is always the total number of possible segments. Since we condition on sampling events, probability $P_t(L_{n+1} = l_{n+1}|G)$ is equal to 1.

- Contribution of the **coalescent event** between lineages i and j

$$\begin{aligned}&P_t^{after}(\mathcal{K}_{\setminus i}, L_i = a, S_i = r, G) \\ &\quad = P_t(\mathcal{K}_{\substack{\setminus i \\ \setminus j}}, L_i = a, L_j = a, S_i = q, S_j = p, G)\lambda_a\end{aligned}$$

Notation above for numbers of ancestral segments is the same as given in section 2.1.1, coalescent event description.

The reassortment event contribution in non-structured populations is given in [10]. For structured populations, we have to account for the probability of the reassorting lineage to be of state a, given the history of the network G.

- Contribution of the **reassortment event** at a lineage of state a, carrying q ancestral segments:

$$P_t^{after}(\mathcal{K}_{\substack{\setminus i \\ \setminus i+1}}, L_i = a, L_{i+1} = a, S_i = r, S_{i+1} = q - r, G)$$
$$= P_t(\mathcal{K}_{\setminus i}, L_i = a, S_i = q, G)\rho_a \left(\frac{1}{2}\right)^{q-1}$$

  The notation above is the same as given in section 2.1.1, reassortment event description.

2.1.6 Approximation of the Structured Coalescent with Reassortment

As stated in [18], to obtain an interval contribution in the exact structured coalescent case we need to solve m^n differential equations, where m is the number of states and n the number of lineages in this interval. This is not changed when we add reassortment to the model as the number of ancestral segments in a lineage is known and constant between the network events.

The vast number of equations makes the model unfeasible for large or highly structured datasets. *Müller et al.* [18, 19] suggested a marginal approximation for structured coalescent, in which lineages and their states are pairwise independent:

$$P_t(L_i = l_i, L_j = l_j|G)$$
$$\underset{=}{^{MASCO}}$$
$$P_t(L_i = l_i,|G)P_t(L_j = l_j|G)$$

In this section, we show that MASCO approximation can be successfully applied in the structured coalescent with reassortment case.

Interval Contribution

The approximation requires derivation of the master equation for the marginalized lineage state probability. It closely follows the supplemental material of [18] and for the parts considering migration and coalescence is the same. Therefore, we only explicitly show the steps involving reassortment and provide the full combined ODE for interval contribution at the end.

As above, $P_t(L_j = a|\mathcal{K}, G)$ is a indicator probability of lineage j being in the state a, $\mathcal{K}_{\setminus i}$ denotes a configuration for all lineages, except lineage i. Then, the marginal lineage state probability obtained by summing over all possible configurations $\mathcal{K}_{\setminus i}$ while fixing the state for lineage i :

$$P_t(L_i = l_i, G) = \sum_{\mathcal{K}_{\setminus i}} P_t(\mathcal{K}, G) \tag{2.8}$$

We integrate over the possible migration patterns in each interval, but do not allow for the coalescent or reassortment network events. Therefore, number of ancestral segments carried by any lineage does not change withing the interval and we do not explicitly include notation for segment configuration.

Applying the summation to the reassortment part of the previously derived interval contribution equation (2.7, line 4), we get:

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

Here $\mathcal{R}$ is the total reassortment rate (section 2.1.2). We separate cases involving lineage i:

$$\begin{aligned}\sum_{\mathcal{K}\setminus i} \mathcal{R} P_t(\mathcal{K}, G) = &\sum_{a=1}^{m} \rho_a \sum_{\substack{j=1 \\ j\neq i}}^{n} \left(1 - \left(\frac{1}{2}\right)^{s_j - 1}\right) \sum_{\mathcal{K}\setminus i} P_t(L_j = a|\mathcal{K}, G) P_t(\mathcal{K}, G) \\ &+ \sum_{a=1}^{m} \rho_a \left(1 - \left(\frac{1}{2}\right)^{s_i - 1}\right) \sum_{\mathcal{K}\setminus i} P_t(L_i = a|\mathcal{K}, G) P_t(\mathcal{K}, G)\end{aligned} \tag{2.9}$$

Then, we can derive expression for $\sum_{\mathcal{K}\setminus i} P_t(L_j = a|\mathcal{K}, G) P_t(\mathcal{K}, G)$, using MASCO approximation:

$$\begin{aligned}\sum_{\mathcal{K}\setminus i} P_t(L_j = a|\mathcal{K}, G) P_t(\mathcal{K}, G) &= \sum_{\mathcal{K}\setminus i} P_t(L_j = a, \mathcal{K}, G) \\ &= P_t(L_j = a, L_i = l_i, G) \\ &= P_t(L_j = a, L_i = l_i|G) P_t(G) \\ &\overset{\text{MASCO}}{=} P_t(L_j = a|G) P_t(L_i = l_i, G)\end{aligned} \tag{2.10}$$

Using 2.10 and 2.8 we can rewrite equation 2.9 as follows:

$$\begin{aligned}\sum_{\mathcal{K}\setminus i}\mathcal{R}P_t(\mathcal{K},G) =& \sum_{a=1}^{m}\rho_a\sum_{\substack{j=1\\j\neq i}}^{n}\left(1-\left(\frac{1}{2}\right)^{s_j-1}\right)P_t(L_j=a,L_i=l_i|G)P_t(G)\\ &+\sum_{a=1}^{m}\rho_a\left(1-\left(\frac{1}{2}\right)^{s_i-1}\right)P_t(L_i=a,L_i=l_i|G)P_t(G)\\ \overset{\text{MASCO}}{=}& P_t(L_i=l_i,G)\sum_{a=1}^{m}\rho_a\sum_{\substack{j=1\\j\neq i}}^{n}\left(1-\left(\frac{1}{2}\right)^{s_j-1}\right)P_t(L_j=a|G)\\ &+P_t(L_i=l_i,G)\rho_{l_i}\left(1-\left(\frac{1}{2}\right)^{s_i-1}\right). \end{aligned} \tag{2.11}$$

This is the final equation that we need to subtract from the right hand side of the equation (6) in the supplementary material of [18] and obtain ODE for marginal lineage state probability:

$$\begin{aligned}\frac{d}{dt}P_t(L_i=l_i,G) =& \sum_{a=1}^{m}\left(\mu_{al_i}P_t(L_i=a,G)-\mu_{l_ia}P_t(L_i=l_i,G)\right)\\ &-P_t(L_i=l_i,G)\sum_{a=1}^{m}\frac{\lambda_a}{2}\sum_{\substack{j=1\\j\neq i}}^{n}\sum_{\substack{k=1\\k\neq j,i}}^{n}P_t(L_j=a|G)P_t(L_k=a|G)\\ &-P_t(L_i=l_i,G)\lambda_{l_i}\sum_{\substack{k=1\\k\neq i}}^{n}P_t(L_k=l_i,|G)\\ &-P_t(L_i=l_i,G)\sum_{a=1}^{m}\rho_a\sum_{\substack{j=1\\j\neq i}}^{n}\left(1-\left(\frac{1}{2}\right)^{s_j-1}\right)P_t(L_j=a|G)\\ &-P_t(L_i=l_i,G)\rho_{l_i}\left(1-\left(\frac{1}{2}\right)^{s_i-1}\right). \end{aligned} \tag{2.12}$$

Here, the first line denotes the migration events in and out of the state l_i, second and third lines denotes coalescent events that do or do not involve lineage i. The last two lines mark a newly derived part. Namely, the probabilities that the reassortment event takes place at lineage i or any other lineage.

Event Contribution

Using the MASCO approximation, we can also rewrite the event contribution equations. As before, for the case of sampling and coalescent events, we use

the existing expressions from [18]. In the case of the coalescent event, we make an addition accounting for the change in the number of ancestral segments of a lineage. For the sampling event, the number of tracked segments by definition is the total number of possible segments.

- Contribution of the **sampling event**

$$P_t^{after}(L_{n+1} = l_{n+1}, S_{n+1} = s_{n+1}, G) = P_t(L_{n+1} = l_{n+1}, S_{n+1} = s_{n+1}|G)P_t(G)$$

- Contribution of the **coalescent event** between lineages i and j

$$\begin{aligned} &P_t^{after}(L_i = a, S_i = r, G) \\ &\quad = P_t(L_i = a, S_i = q|G)P_t(L_j = a, S_j = p|G)P_t(G)\lambda_a \end{aligned}$$

  Notation above is the same as given in section 2.1.1, coalescent event description.

- Contribution of the **reassortment event** at a lineage L_i in the state a

$$\begin{aligned} &P_t^{after}(L_i = a, L_{i+1} = a, S_i = r, S_{i+1} = q - r, G) \\ &\quad = P_t(L_i = a, S_i = q|G)P_t(G)\rho_a \left(\frac{1}{2}\right)^{q-1} \end{aligned}$$

  where

  - $P_t(G) = \sum_{a=1}^{m} P_t(L_i = a, G)$ – the probability of observing the coalescent with reassortment history G up to time t;
  - $P_t(L_i = a|G) = \frac{P_t(L_i=a,G)}{P_t(G)}$;

  and $P_t(L_i = a, G)$ obtained from equation (2.12).

As described in the [18], for all lineages k not involved in the coalescent event, the probability after the event can be written as

$$P_t^{after}(L_k = a, G) = P_t(L_k = a|G) \sum_{a=1}^{m} P_t^{after}(L_i = a, G)$$

For all lineages k not involved in the reassortment event this probability is

$$P_t^{after}(L_k = a, G) = P_t(L_k = a|G) \sum_{a=1}^{m} P_t^{after}(L_i = a, L_{i+1} = a, G)$$

The number of ancestral segments does not change for the lineages that are not involved in coalescent or reassortment events. Therefore we do not explicitly include the segment notation.

2.2 Numerical Solution

Our goal is to solve the 2.12 ODE for the marginal lineage state probability. However, decreasing values present computational challenges and may result in an inaccurate solution. To counteract this, conditional lineage state probabilities can be used ([19]):

$$P_t(L_i = l_i|G) = \frac{P_t(L_i = l_i, G)}{\sum_{a=1}^{m} P_t(L_i = a, G)} = \frac{P_t(L_i = l_i, G)}{P_t(G)} \tag{2.13}$$

To solve the ODEs, we use a Taylor series approximation, up to the third derivative. This lets us to obtain an approximate value for the solution after a short time interval, based on the current value and its derivatives:

$$y_{t+\Delta t} = y_t + \frac{\mathrm{d}}{\mathrm{d}t} y_t \Delta t + \frac{\mathrm{d}^2}{\mathrm{d}t^2} y_t \frac{\Delta t^2}{2!} + ... + \frac{\mathrm{d}^n}{\mathrm{d}t^n} y_t \frac{\Delta t^n}{n!} + ... \tag{2.14}$$

As more derivatives of increasing order are used, precision improves. If we use the first n derivatives to find an approximate solution, the error $O(h)$ will be of the order of $n + 1$ derivative. Following the procedure of [19], we will consider all derivatives of higher than third order equal to zero and use the Taylor approximation, where two first derivatives are calculated exactly. The third derivative is calculated approximately and used to evaluate the error $O(h) \approx \frac{\mathrm{d}^3}{\mathrm{d}t^3} y_t \frac{\Delta t^3}{3!}$. By selecting a threshold value for $O(h)$, we may control the accuracy and step-size Δt.

In what follows, we obtain the exact expressions of the first and second and approximation of the third derivatives. These derivations follow closely and extend the equations given in the supplementary material of [19].

2.2.1 Derivative of the Conditional Lineage State Probability

The derivative of the conditional lineage state probability 2.13 is

$$\frac{d}{dt}P_t(L_i = l_i|G) = \frac{d}{dt}\frac{P_t(L_i = l_i, G)}{P_t(G)} = \frac{\frac{dt}{dt}P_t(L_i = l_i, G)}{P_t(G)} - \frac{P_t(L_i = l_i|G)\frac{d}{dt}P_t(G)}{P_t(G)} \tag{2.15}$$

First term of right hand side of the equation 2.15 can be obtained dividing the equation (2.12) by $P_t(G)$:

$$\begin{aligned}
\frac{\frac{d}{dt}P_t(L_i = l_i, G)}{P_t(G)} =& \sum_{a=1}^{m}\left(\mu_{al_i}P_t(L_i = a|G) - \mu_{l_i a}P_t(L_i = l_i|G)\right) \\
&-P_t(L_i = l_i|G)\sum_{a=1}^{m}\frac{\lambda_a}{2}\sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{\substack{k=1\\ k\neq j,i}}^{n}P_t(L_j = a|G)P_t(L_k = a|G) \\
&-P_t(L_i = l_i|G)\lambda_{l_i}\sum_{\substack{k=1\\ k\neq i}}^{n}P_t(L_k = l_i, |G) \\
&-P_t(L_i = l_i|G)\sum_{a=1}^{m}\rho_a\sum_{\substack{j=1\\ j\neq i}}^{n}\left(1 - \left(\frac{1}{2}\right)^{s_j - 1}\right)P_t(L_j = a|G) \\
&-P_t(L_i = l_i|G)\rho_{l_i}\left(1 - \left(\frac{1}{2}\right)^{s_i - 1}\right)
\end{aligned} \tag{2.16}$$

Second term of the right hand side of the equation 2.15 can be obtained by summing over the equation 2.12:

$$
\begin{aligned}
\frac{P_t(L_i = l_i|G)\frac{d}{dt}P_t(G)}{P_t(G)} =& \frac{P_t(L_i = l_i|G)\frac{d}{dt}\sum_{a=1}^{m} P_t(L_i = a, G)}{P_t(G)} \\
=& \frac{P_t(L_i = l_i|G)}{P_t(G)} \Bigg[\sum_{a=1}^{m}\sum_{b=1}^{m} \left(\mu_{ba}P_t(L_i = b, G) - \mu_{ab}P_t(L_i = a, G)\right) (= 0) \\
& - \sum_{a=1}^{m} P_t(L_i = l_i, G) \sum_{a=1}^{m} \frac{\lambda_a}{2} \sum_{\substack{j=1 \\ j\neq i}}^{n} \sum_{\substack{k=1 \\ k\neq j,i}}^{n} P_t(L_j = a|G)P_t(L_k = a|G) \\
& - \sum_{a=1}^{m} P_t(L_i = l_i, G)\lambda_{l_i} \sum_{\substack{k=1 \\ k\neq i}}^{n} P_t(L_k = l_i, |G) \\
& - \sum_{a=1}^{m} P_t(L_i = l_i, G) \sum_{a=1}^{m} \rho_a \sum_{\substack{j=1 \\ j\neq i}}^{n} \left(1 - \left(\frac{1}{2}\right)^{s_j - 1}\right) P_t(L_j = a|G) \\
& - \sum_{a=1}^{m} P_t(L_i = l_i, G)\rho_{l_i} \left(1 - \left(\frac{1}{2}\right)^{s_i - 1}\right) \Bigg] \\
=& - \frac{P_t(L_i = l_i|G)}{P_t(G)} \Bigg[P_t(G) \sum_{b=1}^{m} \frac{\lambda_b}{2} \sum_{\substack{j=1 \\ j\neq i}}^{n} \sum_{\substack{k=1 \\ k\neq j,i}}^{n} P_t(L_j = b|G)P_t(L_k = b|G) \\
& + P_t(G) \sum_{a=1}^{m} P_t(L_i = a|G)\lambda_a \sum_{\substack{k=1 \\ k\neq i}}^{n} P_t(L_k = a|G) \\
& + P_t(G) \sum_{b=1}^{m} \rho_b \sum_{\substack{j=1 \\ j\neq i}}^{n} \left(1 - \left(\frac{1}{2}\right)^{s_j - 1}\right) P_t(L_j = b|G) \\
& + P_t(G) \sum_{a=1}^{m} P_t(L_i = a|G)\rho_a \left(1 - \left(\frac{1}{2}\right)^{s_i - 1}\right) \Bigg] \\
=& - P_t(L_i = l_i|G) \sum_{a=1}^{m} \frac{\lambda_a}{2} \sum_{j=1}^{n} \sum_{\substack{k=1 \\ k\neq j}}^{n} P_t(L_j = a|G)P_t(L_k = a|G) \\
& - P_t(L_i = l_i|G) \sum_{a=1}^{m} \rho_a \sum_{j=1}^{n} \left(1 - \left(\frac{1}{2}\right)^{s_j - 1}\right) P_t(L_j = a|G)
\end{aligned}
\tag{2.17}
$$

We subtract equation 2.17 from 2.16 and thus arrive at the final expression for the derivative of conditional lineage state probability under the structured coalescent

with reassortment model:

$$\begin{aligned}\frac{d}{dt}P_t(L_i = l_i|G) = &\sum_{a=1}^{m}(\mu_{al_i}P_t(L_i = a|G) - \mu_{l_ia}P_t(L_i = l_i|G)) \\ &+ P_t(L_i = l_i|G)\sum a = 1^m\lambda_a P_t(L_i = a|G)\sum_{\substack{k=1\\k\neq i}}^{n} P_t(L_k = a|G) \\ &- P_t(L_i = l_i|G)\lambda_{l_i}\sum_{\substack{j=1\\j\neq i}}^{n} P_t(L_k = l_i|G) \\ &+ P_t(L_i = l_i|G)\sum_{a=1}^{m}\rho_a\left(1 - \left(\frac{1}{2}\right)^{s_i-1}\right)P_t(L_i = a|G) \\ &- P_t(L_i = l_i|G)\rho_{l_i}\left(1 - \left(\frac{1}{2}\right)^{s_i-1}\right) \end{aligned} \tag{2.18}$$

The last two lines of the above equation are newly derived and represent reassortment contribution to the network.

Derivative of the Network Probability

It is easy to obtain the master equation for the network probability $P_t(G)$ from the previous results. We know that

$$\frac{d}{dt}P_t(G) = \sum_{a=1}^{m}\frac{d}{dt}P_t(L_i = a, G) = \frac{d}{dt}\sum_{a=1}^{m}P_t(L_i = a, G)$$

To get, the right hand side of the above expression we multiply equation 2.17 by $\frac{P_t(G)}{P_t(L_i=l_i|G)}$:

$$\begin{aligned}\frac{d}{dt}P_t(G) = &- P_t(G)\sum_{a=1}^{m}\frac{\lambda_a}{2}\sum_{j=1}^{n}\sum_{\substack{k=1\\k\neq j}}P_t(L_j = a|G)P_t(L_k = a|G) \\ &- P_t(G)\sum_{a=1}^{m}\rho_a\sum_{j=1}^{n}\left(1 - \left(\frac{1}{2}\right)^{s_j-1}\right)P_t(L_j = a|G)\end{aligned}$$

Here, again the derivation for the first term is described in [19] and the second term, accounting for reassortment contribution is newly derived.

2.2.2 Second Derivative of the Conditional Lineage State Probability

To implement Taylor series solution for ODE as formulated above, we next derive the exact expression for the second derivative of conditional lineage state

probability. Applying product rule to the equation 2.18, we get:

$$\begin{aligned}
\frac{d^2 P_t(L_i = l_i|G)}{dt^2} =& \sum_{a=1}^{m} \left(\mu_{a l_i} \frac{d}{dt} P_t(L_i = a|G) - \mu_{l_i a} \frac{d}{dt} P_t(L_i = l_i|G) \right) \\
&+ \frac{d}{dt} P_t(L_i = l_i|G) \sum_{a=1}^{m} \lambda_a P_t(L_i = a|G) \sum_{\substack{k=1 \\ k \neq i}}^{n} P_t(L_k = a|G) \\
&+ P_t(L_i = l_i|G) \sum_{a=1}^{m} \lambda_a \frac{d}{dt} P_t(L_i = a|G) \sum_{\substack{k=1 \\ k \neq i}}^{n} P_t(L_k = a|G) \\
&+ P_t(L_i = l_i|G) \sum_{a=1}^{m} \lambda_a P_t(L_i = a|G) \sum_{\substack{k=1 \\ k \neq i}}^{n} \frac{d}{dt} P_t(L_k = a|G) \\
&- \frac{d}{dt} P_t(L_i = l_i|G) \lambda_{l_i} \sum_{\substack{j=1 \\ j \neq i}}^{n} P_t(L_k = l_i|G) \\
&- P_t(L_i = l_i|G) \lambda_{l_i} \sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{d}{dt} P_t(L_k = l_i|G) \\
&+ \frac{d}{dt} P_t(L_i = l_i|G) \sum_{a=1}^{m} \rho_a \left(1 - \left(\frac{1}{2} \right)^{s_i - 1} \right) P_t(L_i = a|G) \\
&+ P_t(L_i = l_i|G) \sum_{a=1}^{m} \rho_a \left(1 - \left(\frac{1}{2} \right)^{s_i - 1} \right) \frac{d}{dt} P_t(L_i = a|G) \\
&- \frac{d}{dt} P_t(L_i = l_i|G) \rho_{l_i} \left(1 - \left(\frac{1}{2} \right)^{s_i - 1} \right)
\end{aligned} \tag{2.19}$$

2.2.3 Approximation of the Third Derivative of the Conditional Lineage State Probability

Finally, we need the expression for the third derivative of the conditional lineage state probability. The third derivative is used to determine the step size Δt of integration, but not the value of $P_{t+\Delta t}$. Therefore, we do not need to calculate it with high precision and can use the approximate expression. As detailed in [19], two assumptions are made to obtain the approximation of the third derivative:

- the sum of probability mass in a state over all lineages but lineage i does not change;

- the sum of the derivatives of lineage i coalescing in any state does not change.

Then the third and fourth lines 3 of the second derivative 2.19 are equal to zero. We do not make any additional assumptions than those detailed in [19]. That is, when taking the third derivative, terms accounting for reassortment are exact:

$$\begin{aligned}\frac{d^3P_t(L_i=l_i|G)}{dt^3} \approx &\sum_{a=1}^{m}\left(\mu_{al_i}\frac{d^2}{dt^2}P_t(L_i=a|G)-\mu_{l_ia}\frac{d^2}{dt^2}P_t(L_i=l_i|G)\right)\\
&+\frac{d^2}{dt^2}P_t(L_i=l_i|G)\left(\sum_{a=1}^{m}\lambda_a P_t(L_i=a|G)\sum_{\substack{k=1\\k\neq i}}^{n}P_t(L_k=a|G)\right)\\
&-\frac{d^2}{dt^2}P_t(L_i=l_i|G)\lambda_{l_i}\sum_{\substack{k=1\\k\neq i}}^{n}P_t(L_k=l_i|G)\\
&+\frac{d^2}{dt^2}P_t(L_i=l_i|G)\left(\sum_{a=1}^{m}\rho_a(1-0.5^{s_i-1})P_t(L_i=a|G)\right)\\
&+2\frac{d}{dt}P_t(L_i=l_i|G)\left(\sum_{a=1}^{m}\rho_a(1-0.5^{s_i-1})\frac{d}{dt}P_t(L_i=a|G)\right)\\
&+P_t(L_i=l_i|G)\left(\sum_{a=1}^{m}\rho_a(1-0.5^{s_i-1})\frac{d^2}{dt^2}P_t(L_i=a|G)\right)\\
&-\frac{d^2}{dt^2}P_t(L_i=l_i|G)\rho_{l_i}\left(1-\left(\frac{1}{2}\right)^{s_i-1}\right)\end{aligned} \tag{2.20}$$

These are all expressions needed to apply MASCO approximation for a structured coalescent with reassortment model. Note, that most terms in the first or second derivatives can be calculated once and reused in the expression of higher order derivative, which speeds up the calculations.

2.3 Simulation

To evaluate our model, we implemented a simulator for structured coalescent with reassortment using the Gillespie algorithm (see, for example, [22]). First, we take the sampling times and network event rates as input. Second, set the most recent sample as a current event (node in a network). Third, take the exponential distribution with mean, which is the probability of the event happening in a current state, for each state and each network event. If migration is non-symmetric, take

such distribution for each pair of possible states. From these distributions, draw random waiting times until each possible event happens. Fourth, take the minimum waiting time $t_{min} = min\{\{t\} \cup t_s\}$, where {t} is a set of all values drawn in the previous step and t_s is the known time of the next sampling event. If t_{min} is for one of the network events, we simulate it, changing the network as described in 2.1.1. Otherwise, we add a sampling event to the network. Finally, we set the newly added event as current and iteratively repeat from the third step until there are no more unused samples, and the network has only one extant lineage left. We set this lineage as root.

In this particular case, we need to simulate migration, coalescent, and reassortment events backwards in time. For any state a, the mean of exponential distributions for the event waiting times are:

coalescent	reassortment	migration	
$\lambda_a \binom{k_a(\mathcal{K})}{2}$	$k_a(\mathcal{K})\rho_a$	$k_a(\mathcal{K})\mu_{ab}$,	$b \in 1, ..., m, b \neq a$

Note that here we use constant effective population sizes and therefore constant coalescent rates.

2.4 Implementation

We implemented the package, called SCORE for the BEAST 2.5 [20]. It depends on the CoalRe ([10]) package for MCMC ([19]) sampling of coalescent with reassortment networks and extends the MASCOT package to calculate the structured network likelihood when integrating over all migration histories.

Most MCMC operators of CoalRe for an unstructured network proposal can also be used in our case. However, we have to adjust parameter values supplied to the operator which re-simulates 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 above the root" section in the supplemental material of [10] for more details). In the unstructured setting, we may re-simulate 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.

In the structured case, parameters can be state dependent. Therefore, we have to decide on their values for unstructured simulation. Here, the parameters are effective population size N_e and reassortment rate ρ. Simply averaging over all m states – $N_e = \sum_{a=1}^{m} N_{e_a}$, $\rho = \sum_{a=1}^{m} \rho_a$ – provided high enough acceptance ratio for our simulated and real data analysis.

Currently, SCORE package can be used to simulate and infer structured coalescent with reassortment networks as well as provide the inferred state proba-

bilities of the network root node. The source code for SCORE can be found at https://github.com/jugne/SCORE.

Chapter 3

Results

3.1 Validation

Simulator Validation

First, we investigated whether the structured coalescent with reassortment simulator agrees with known structured coalescent sampler Multi Type Tree (MTT) [11] and coalescent with reassortment sampler CoalRe [10]. In order to compare against MTT, we had set reassortment rate to zero and therefore simulated under the structured coalescent model. Similarly, to comparing against CoalRe, we simulated without accounting for population structure, i.e., $\mu = 0$. Both comparisons show that the model for either of the two variants is implemented correctly in our simulator (see supplementary figure B.1). We trust, that given both reassortment and structure are allowed, the simulator also provides accurate results in the general case.

MCMC Validation

Exact Structured Coalescent with Reassortment

We seek to ensure that the exact version of SCORE samples from the true distribution of structured coalescent with reassortment. First, we sample networks from the prior distribution. Then, we simulate under the same parameter values. Finally, we compare the distributions of network height, length, and reassortment node count obtained by the two runs. To measure the difference of distributions more precisely, we calculate the Kolmogorov-Smirnov (KS) statistic as a function of iteration count. Supplementary figures B.2 and B.3 show the comparison being made in cases of structured coalescent with reassortment and when we do not account for either structure (no sample states, $\mu = 0$) or reassortment ($\rho = 0$). The KS differences asymptote to around $10^{-3.5} - 10^{-4}$, and from this we conclude that distributions of network statistics match with high enough precision.

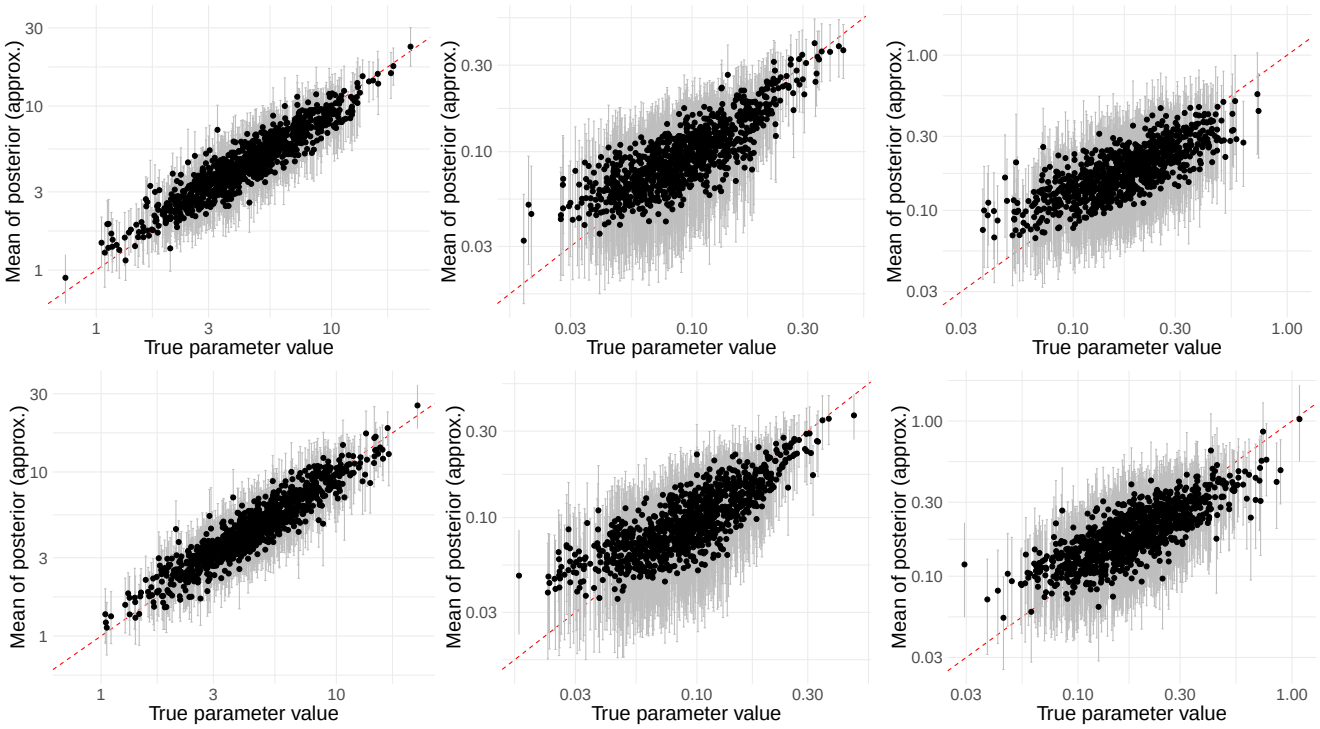


Figure 3.1: Network parameter inference from 1000 simulated genetic sequence data in 2 states. Top row contains plots for state 1, bottom row – state 2. **Left.** True (x-axis) versus estimated (y-axis) effective population sizes. Grey bars are 95% confidence intervals, red marks the x=y curve. **Middle.** True (x-axis) versus estimated (y-axis) reassortment rates. Grey bars are 95% confidence intervals, red marks the x=y curve. **Right.** True (x-axis) versus estimated (y-axis) effective population sizes. Grey bars are 95% confidence intervals, red marks the x=y curve.

Approximate Structured Coalescent with Reassortment

Next, we apply the same test to the approximate version of SCORE. First, we use the source code for the approximate structured coalescent with reassortment on a non-structured dataset. This way, no approximation is needed and sampled, and simulated distributions should match. While here it is unnecessary to use SCORE instead of CoalRe, it allows us to independently verify that parts regarding coalescent and reassortment are implemented correctly (figure B.4, left). The right-hand side of the figure B.4 shows the approximate inference case where the dataset is structured. As expected KS difference is larger, however, the sampled and simulated distributions still maintain highly similar shapes and mean values.

Inference of Effective Population Sizes, Reassortment and Migration Rates

To demonstrate that SCORE allows us to obtain effective population sizes, reassortment, and migration rates with high precision, we considered two different simulation scenarios. First, effective population sizes and rates were inferred from a given fixed network and second, where network and corresponding parameters were jointly inferred from simulated sequence data. In the fixed network setting, we simulated 1000 networks with 100 taxa, each carrying four segments.

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 state $a, b \in \{1, 2\}$ and $a \neq b$.[1] Each taxon was randomly assigned to one of the two states and its sampling time is uniformly drawn from the $[0, 20]$ interval. Then, we inferred the parameter values, given the simulated networks. Using the above parameter distributions as priors.

Figure 3.1 shows that, given the network, SCORE is able to recover the parameter values well. Between 93 and 95 per cent of true parameter values are within the 95% highest posterior density (HPD) interval for the corresponding parameter values, sampled by SCORE (table 3.1, bottom row).

Next, we studied the ability of SCORE to jointly infer the network, effective population sizes, reassortment and migration rates. We simulated 100 networks and embedding of the segment trees for 100 taxa with four segments. The states and sampling times were drawn as described above. The genetic sequence of each segment was simulated according to the JukesCantor (JC69) model with either high (5×10^{-3}), low (5×10^{-4}) or mixed (two segments with high and two with low) evolutionary rates. The prior distribution of a reassortment rate was the same as above, while we set a mean of 2 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 mean 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 dataset. Each inference was run for 48 hours, and we used only those runs for which the effective sample size (ESS) of posterior probability was higher than 100. Overall, between 91 and 98 per cent of true parameter values were within the 95% HPD interval (see table 3.1 and supplementary figures B.5, B.6).

[1] Note that here m denotes the mean of real variable. The mean for the natural logarithm of this variable can be calculated as $ln(m) - \frac{\sigma^2}{2}$.

Migration prior	Clock rate	State 1			State 2		
		N_e	ρ	μ	N_e	ρ	μ
log-normal	high	93.8	91.7	90.7	96.9	93.8	97.9
	low	91.9	91.9	91.9	97.9	91.9	95.9
	mixed	90.9	92.9	92.9	97.9	91.9	95.9
exponential	high	95.9	98.9	96.9	97.9	95.9	92.9
	low	94	97	94	98	97	93
	mixed	91.7	96.9	95.8	97.9	95.8	91.7
Fixed network		94.4	94	93.6	95.5	94.9	94.9

Table 3.1: Percentage of true parameter values falling within 95% HPD interval for simulation study of 2 states. The top six rows shows cases where the network and its parameters were joinly inferred from simulated genetic sequence data for two different migration priors and high (5×10^{-3}), low (5×10^{-4}) or mixed (2 segments with high and 2 with low) clock rates. The bottom row shows results obtained when the true network is known and only its parameters are inferred. True rates and effective population sizes were asymetric in all cases.

3.2 Application to Influenza Virus

We applied both SCORE and MASCOT on an alignment of the seasonal influenza A/H3N2 genomes. We used Influenza Research Database [23] to gather 500 full genomic sequence samples from the USA, New Zealand, Hong Kong and Singapore dated between 2002 and 2005. Then, we combined data from Hong Kong and Singapore into one South East Asia (SEA) set, giving a total of 3 geographic states. Each state was down-sampled to 50 sequences, where SEA consists of 45 Hong Kong and 5 Singapore samples.

We ran three independent analyses under SCORE or MASCOT in BEAST 2.5 [20], using the coupled MCMC [24, 25]. Each chain was run for 5×10^7 iterations under the HKY+Γ_4 evolutionary model. We estimate two evolutionary rates for each segment: one for the first and second codon positions and one for the third codon position.

We inferred effective population sizes, migration and coalescent rates for all three locations as well as the phylogenetic network, segment trees and reassortment rates (SCORE) or single segment phylogenetic tree (MASCOT). The inference under SCORE was conducted using 4 out of 8 genomic segments: HA, NA, PB1 and MP. We used a single segment (HA) alignments for MASCOT runs.

See supplementary chapter C for more details on 150 sequences used.

Structured Coalescent with Reassortment Network

Figure 3.2 **(a)** shows the maximum posterior probability network obtained by SCORE. The inferred reassortment events were mostly confined to SEA. This

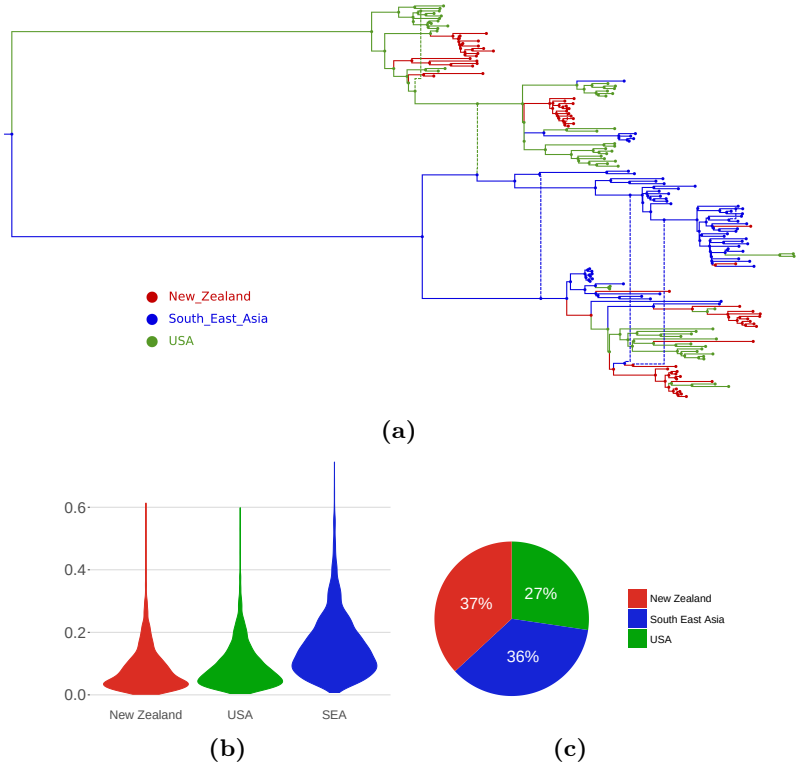


Figure 3.2: Inference of structured coalescent with reassortment network from the seasonal influenza A/H3N2 sequences sampled in the USA, South East Asia and New Zealand under the SCORE prior. **(a)** Maximum posterior probability network colored by sampled regions and inferred migration events. **(b)** Distributions for estimated reassortment rates in each region. **(c)**Obtained probabilities for the network root node to be in either of the three regions.

agrees with reassortment rate distributions (figure 3.2 **(b)**), which suggest almost no reassortment in the USA or New Zealand. Higher reassortment rate in SEA may be driven by larger effective population size (3.3 **(b)**) which indicates greater genetic diversity of influenza. SCORE was not able to produce decisive results on the network root state (figure 3.2 **(c)**). This is expected as the dataset is limited in terms of the geographic locations and number of sequences.

Different Estimates of Model parameters

To demonstrate how accounting for multiple segments influences parameter inference, we compared results obtained by SCORE and MASCOT [19] approximation for structured coalescent without reassortment. MASCOT was run on the same dataset, but only using the segment HA genomic sequences. Figure 3.3**(a)** shows that the posterior distributions for migration rates obtained by SCORE are more certain than those obtained by MASCOT. Mean values were also lower under

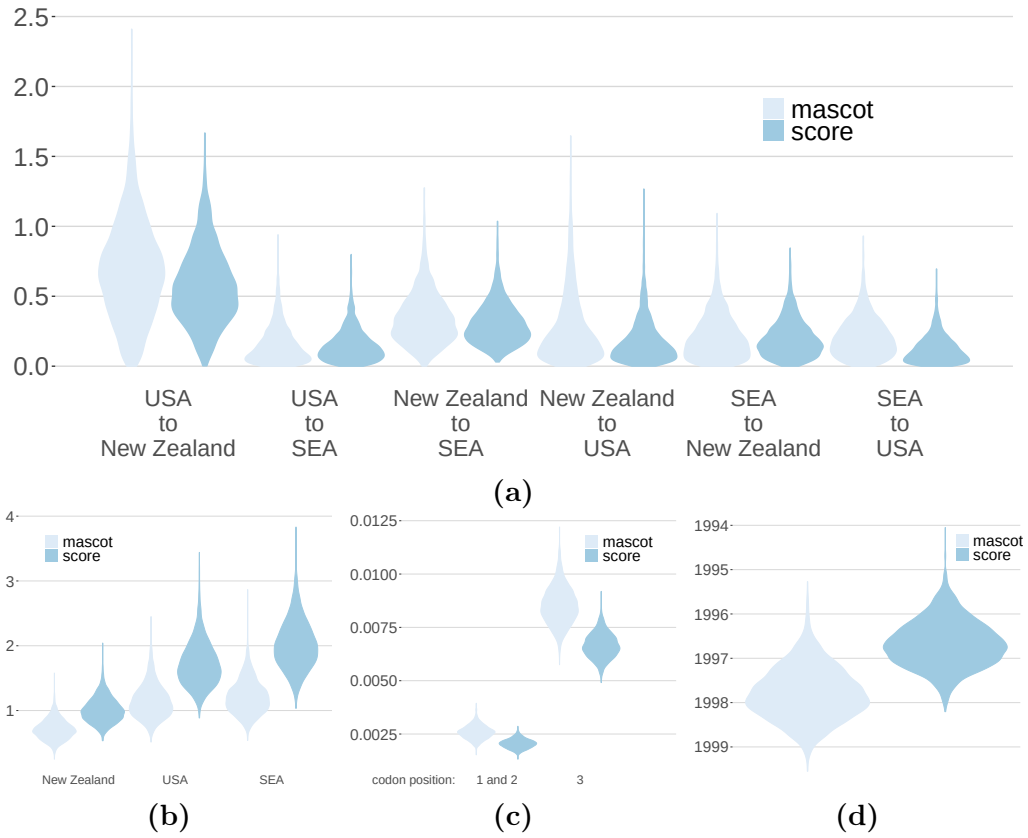


Figure 3.3: Inferred distributions for structured coalescent parameters from the seasonal influenza A/H3N2 sequences sampled in the USA, South East Asia and New Zealand using SCORE and MASCOT priors.**(a)** Asymmetrical migration rates.**(b)** Effective population sizes.**(c)** Evolutionary rates for different codon positions of segment HA. Positions 1 and 2 are combined.**(d)** Distribution of root node time for HA segment tree.

SCORE, with the exception of migration rates from the USA to SEA and SEA to New Zealand. Both methods suggest high migration from the USA to New Zealand. High backwards in time migration rate combined with highest posterior root node probability for New Zealand (figure 3.2(**c**)) suggest that seasonal influenza A/H3N2 originated in Oceania. This is not believed to be true [26, 27, 28] and is likely driven by the lack of samples. Figure 3.3**(b)** shows the comparison for effective population sizes obtained by SCORE and MASCOT. As in the unstructured case [10], N_e is estimated to be higher when we account for reassortment. Both methods attribute the largest effective population size to SEA. This can be driven by combining the sequences from distinct locations: Hong Kong and Singapore. Combining two spatially separated sub-populations into a single state may lead to greater viral diversity within this state and, in turn, to lower coalescent rate (higher effective population size) for this state.

Next, we compare the rate of evolution and tree height of the HA segment.

Lower evolutionary rates were estimated for combined first and second codon positions by both methods (figure 3.3**(c)**). Here models behave as expected since it is believed that the second codon position evolves slower and the third position faster than the first [29]. Similar to results for the migration rate, evolutionary rate estimates inferred by SCORE are more certain, and their mean values are lower than those obtained by MASCOT. When dating the root node of the HA tree, the two methods disagree by little more than one year (figure 3.3**(d)**). The larger tree height, obtained by SCORE, is consistent with the lower inferred evolutionary rates and higher effective population sizes, which lead to longer branches between coalescent events.

Chapter 4

Discusion

We have extended the marginal approximation of the structured coalescent framework [18, 19] to include reassortment dynamics [10]. This includes the derivation of exact ODEs needed to calculate the probability density of structured coalescent with reassortment network and subsequent approximation, assuming the pairwise independence of lineage states. Additionally, we implement the structured coalescent with reassortment simulator.

In our simulation experiments, we have shown that true network statistics distributions are obtained by the exact version of SCORE. For the approximation, between 90 and 98 per cent of true parameter values were within the 95% HPD interval of the respective estimated parameter. Usually, at least 95% of true values are required to be within the 95% HPD interval. However, the lower threshold should be expected here, since SCORE is an approximate method.

Importantly, SCORE allows us to use the full segmented genomes for inference. Even a small scale comparative study shows that joint inference of phylogenetic trees for several segments leads to a different estimation of effective population sizes, migration and evolutionary rates, as well as the height of a segment tree root node. In particular, the assumption of independent evolution of segment tree led to lower effective population sizes and higher evolutionary rates. This agrees to previously described findings for the unstructured coalescent [10].

However, there are limitations to our study. The analysis was done on a small scale dataset in terms of the number of genetic sequences and geographic locations as well as the time span of the samples. We believe that this caused undecisive results on the network root node state as well as inflated backwards in time migration rates from the USA to New Zealand.

In order to gain more trust in our analyses and identify possible model biases, we propose that several simulation studies should be carried out in the future. First, our model assumes a constant population size, which is an idealised case and not true in the early stages of an emerging epidemic. To identify the distortion likely introduced by this assumption, we may use genetic sequences simulated under the increasing effective population sizes for the inference under the

constant population sizes. The known ground truth would allow for an accurate evaluation of the method. Secondly, a comprehensive simulation study comparing inferences obtained by SCORE or MASCOT could verify the observed differences between using a single segment and full genomes for structured coalescent with reassortment research.

Another constraint of SCORE is the inability to assign credible state probabilities to the inner nodes of the network. This can be addressed by implementing forward in time calculations, once the state of the root node is known [19]. In case of structured coalescent without reassortment, backwards/forwards approach led to more improved node state estimates when compared to a backwards-only implementation [19].

Finally, here we discussed using SCORE to coalescent with reassortment in geographically structured populations. However, it can also be applied to investigate reassortment rate dependency on host types. Particularly interesting are the reassortment events between swine, avian and human influenza variants, which have been characterized as part of evolutionary history for novel human influenza strains [30, 31].

Acknowledgements

I am grateful to Prof. Dr. Tanja Stadler, Dr. Timothy Vaughan and Nicola Müller for the invaluable guidance during the past six months. Their teaching and helpful comments broadened my perspective and allowed me to improve research, programming and scientific writing skills. I am also thankful to my family and fiancé without whom I would have not been able to continue my education.

Bibliography

[1] World Health Organisation, "Influenza Fact Sheet," accessed: 2019-06-13. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/influenza-(seasonal)

[2] N. P. Johnson and J. Mueller, "Updating the accounts: global mortality of the 1918-1920" spanish" influenza pandemic," *Bulletin of the History of Medicine*, vol. 76, no. 1, pp. 105–115, 2002.

[3] World Health Organisation, "Pandemic influenza: an evolving challenge," accessed: 2019-06-27. [Online]. Available: https://www.who.int/influenza/pandemic-influenza-an-evolving-challenge/en/

[4] D. Butler, "Swine flu goes global," *Nature*, vol. 458, no. 7242, p. 1082, 2009.

[5] J. Steel and A. C. Lowen, "Influenza A virus reassortment," in *Influenza Pathogenesis and Control-Volume I.* Springer, 2014, pp. 377–401.

[6] C. Li and H. Chen, "Enhancement of influenza virus transmission by gene reassortment," in *Influenza Pathogenesis and Control-Volume I.* Springer, 2014, pp. 185–204.

[7] G. J. Smith, D. Vijaykrishna, J. Bahl, S. J. Lycett, M. Worobey, O. G. Pybus, S. K. Ma, C. L. Cheung, J. Raghwani, S. Bhatt *et al.*, "Origins and evolutionary genomics of the 2009 swine-origin h1n1 influenza a epidemic," *Nature*, vol. 459, no. 7250, p. 1122, 2009.

[8] D. Vijaykrishna, L. Poon, H. Zhu, S. Ma, O. Li, C. Cheung, G. Smith, J. Peiris, and Y. Guan, "Reassortment of pandemic h1n1/2009 influenza a virus in swine," *Science*, vol. 328, no. 5985, pp. 1529–1529, 2010.

[9] G. Dudas, T. Bedford, S. Lycett, and A. Rambaut, "Reassortment between influenza b lineages and the emergence of a coadapted pb1–pb2–ha gene complex," *Molecular biology and evolution*, vol. 32, no. 1, pp. 162–172, 2014.

[10] N. F. Müller, U. Stolz, G. Dudas, T. Stadler, and T. G. Vaughan, "Extensive differences in reassortment rates between human influenza virus subtypes and between pairs of segments revealed by bayesian inference of reassortment networks," *In preparation*, 2019.

[11] T. G. Vaughan, D. Kühnert, A. Popinga, D. Welch, and A. J. Drummond, "Efficient bayesian inference under the structured coalescent," *Bioinformatics*, vol. 30, no. 16, pp. 2272–2279, 2014.

[12] 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*, vol. 98, no. 8, pp. 4563–4568, 2001.

[13] E. M. Volz, "Complex population dynamics and the coalescent under neutrality," *Genetics*, vol. 190, no. 1, pp. 187–201, 2012.

[14] N. De Maio, C.-H. Wu, K. M. O'Reilly, and D. Wilson, "New routes to phylogeography: a bayesian structured coalescent approximation," *PLoS genetics*, vol. 11, no. 8, p. e1005421, 2015.

[15] N. Takahata, "The coalescent in two partially isolated diffusion populations," *Genetics Research*, vol. 52, no. 3, pp. 213–222, 1988.

[16] R. R. Hudson *et al.*, "Gene genealogies and the coalescent process," *Oxford surveys in evolutionary biology*, vol. 7, no. 1, p. 44, 1990.

[17] M. Notohara, "The coalescent and the genealogical process in geographically structured population," *Journal of mathematical biology*, vol. 29, no. 1, pp. 59–75, 1990.

[18] N. F. Müller, D. A. Rasmussen, and T. Stadler, "The structured coalescent and its approximations," *Molecular biology and evolution*, vol. 34, no. 11, pp. 2970–2981, 2017.

[19] N. F. Müller, D. Rasmussen, and T. Stadler, "Mascot: Parameter and state inference under the marginal structured coalescent approximation," *Bioinformatics*, vol. 34, no. 22, pp. 3843–3848, 2018.

[20] R. Bouckaert, T. G. Vaughan, J. Barido-Sottani, S. Duchene, M. Fourment, A. Gavryushkina, J. Heled, G. Jones, D. Kuhnert, N. De Maio *et al.*, "Beast 2.5: An advanced software platform for bayesian evolutionary analysis," *BioRxiv*, p. 474296, 2018.

[21] J. Felsenstein, "Evolutionary trees from dna sequences: a maximum likelihood approach," *Journal of molecular evolution*, vol. 17, no. 6, pp. 368–376, 1981.

[22] D. T. Gillespie, "Stochastic simulation of chemical kinetics," *Annu. Rev. Phys. Chem.*, vol. 58, pp. 35–55, 2007.

[23] Influenza Research Database, accessed: 2019-05-08. [Online]. Available: https://www.fludb.org

[24] G. Altekar, S. Dwarkadas, J. P. Huelsenbeck, and F. Ronquist, "Parallel metropolis coupled markov chain monte carlo for bayesian phylogenetic inference," *Bioinformatics*, vol. 20, no. 3, pp. 407–415, 2004.

[25] N. F. Mueller and R. Bouckaert, "Coupled mcmc in beast 2," *bioRxiv*, p. 603514, 2019.

[26] T. Bedford, S. Riley, I. G. Barr, S. Broor, M. Chadha, N. J. Cox, R. S. Daniels, C. P. Gunasekaran, A. C. Hurt, A. Kelso *et al.*, "Global circulation patterns of seasonal influenza viruses vary with antigenic drift," *Nature*, vol. 523, no. 7559, p. 217, 2015.

[27] C. A. Russell, T. C. Jones, I. G. Barr, N. J. Cox, R. J. Garten, V. Gregory, I. D. Gust, A. W. Hampson, A. J. Hay, A. C. Hurt *et al.*, "Influenza vaccine strain selection and recent studies on the global migration of seasonal influenza viruses," *Vaccine*, vol. 26, pp. D31–D34, 2008.

[28] F. Wen, T. Bedford, and S. Cobey, "Explaining the geographical origins of seasonal influenza a (h3n2)," *Proceedings of the Royal Society B: Biological Sciences*, vol. 283, no. 1838, p. 20161312, 2016.

[29] L. Bofkin and N. Goldman, "Variation in evolutionary processes at different codon positions," *Molecular biology and evolution*, vol. 24, no. 2, pp. 513–521, 2006.

[30] V. Trifonov, H. Khiabanian, and R. Rabadan, "Geographic dependence, surveillance, and origins of the 2009 influenza a (h1n1) virus," *New England journal of medicine*, vol. 361, no. 2, pp. 115–119, 2009.

[31] H. Khiabanian, V. Trifonov, and R. Rabadan, "Reassortment patterns in swine influenza viruses," *PloS one*, vol. 4, no. 10, p. e7366, 2009.

Appendix A

Derivation of the Master Equation

Derivation of master equation in the exact case closely follows [18]. Additionally to the notation in the main text, we denote $\mathcal{K}_{ia}$ a configuration where lineage i is in a state a. All other lineages are of the same state as decribed by $\mathcal{K}$ in the main text. We seek to obtain $P_{t+\Delta t}$ based on P_t for the network interval contribution:

$$\begin{aligned} P_{t+\Delta t}(\mathcal{K}, G) =& P_t(\mathcal{K}, G)(1 - \mathcal{M}\Delta t - \mathcal{C}\Delta t - \mathcal{R}\Delta t) \\ &+ \sum_{i=1}^{n}\sum_{a=1}^{m}(\mu_{al_i}\Delta t P_t(\mathcal{K}_{ia}, G)) + O((\Delta t)^2) \end{aligned}$$

For right hand side, multiplication of the first two terms accounts for not allowing any migration, coalescent or reassortment events that would change the configuration $\mathcal{K}$. The double summation accounts for migration events that would lead to $\mathcal{K}$ from any other possible configuration. The last term stands for higher order possibilities, i.e., seeing more than one event within the interval.

Next, we subtract $P_t(\mathcal{K}, G)$ from both sides and divide them by Δt:

$$\begin{aligned} \frac{P_{t+\Delta t}(\mathcal{K}, G) - P_t(\mathcal{K}, G)}{\Delta t} =& -(\mathcal{M} + \mathcal{C} + \mathcal{R})P_t(\mathcal{K}, G) \\ &+ \sum_{i=1}^{n}\sum_{a=1}^{m}(\mu_{al_i}P_t(\mathcal{K}_{ia}, G)) + O((\Delta t)^2) \end{aligned}$$

Taking the limit $\Delta t \to 0$ yields

$$\begin{aligned} \frac{\mathrm{d}P_t(\mathcal{K}, G)}{\mathrm{d}t} =& -(\mathcal{M} + \mathcal{C} + \mathcal{R})P_t(\mathcal{K}, G) \\ &+ \sum_{i=1}^{n}\sum_{a=1}^{m}(\mu_{al_i}P_t(\mathcal{K}_{ia}, G)) \end{aligned}$$

Expanding the total rate expressions $\mathcal{M}, \mathcal{C}, \mathcal{R}$ from section 2.1.2, we obtain:

$$\begin{aligned}\frac{\mathrm{d}P_t(\mathcal{K}, G)}{\mathrm{d}t} = &- \sum_{i=1}^{n}\sum_{a=1}^{m}(\mu_{al_i} P_t(\mathcal{K}, G)) \\ &- \sum_{a=1}^{m} \lambda_a \binom{k_a(\mathcal{K})}{2} P_t(\mathcal{K}, G) \\ &- \sum_{a=1}^{m} \rho_a \sum_{j=1}^{n} P_t(L_j = a|\mathcal{K}, G) \left(1 - \left(\frac{1}{2}\right)^{s_j - 1}\right) P_t(\mathcal{K}, G) \\ &+ \sum_{i=1}^{n}\sum_{a=1}^{m}(\mu_{al_i} P_t(\mathcal{K}_{ia}, G))\end{aligned}$$

Finally, we rearrange the two migration terms and obtain the final expression for the master equation of the structured coalescent with reassortment network interval contribution:

$$\begin{aligned}\frac{\mathrm{d}P_t(\mathcal{K}, G)}{\mathrm{d}t} = &\sum_{i=1}^{n}\sum_{a=1}^{m}(\mu_{al_i} P_t(\mathcal{K}_{ia}, G) - \mu_{al_i} P_t(\mathcal{K}, G)) \\ &- \sum_{a=1}^{m} \lambda_a \binom{k_a(\mathcal{K})}{2} P_t(\mathcal{K}, G) \\ &- \sum_{a=1}^{m} \rho_a \sum_{j=1}^{n} P_t(L_j = a|\mathcal{K}, G) \left(1 - \left(\frac{1}{2}\right)^{s_j - 1}\right) P_t(\mathcal{K}, G)\end{aligned}$$

Appendix B

Supplementary Figures

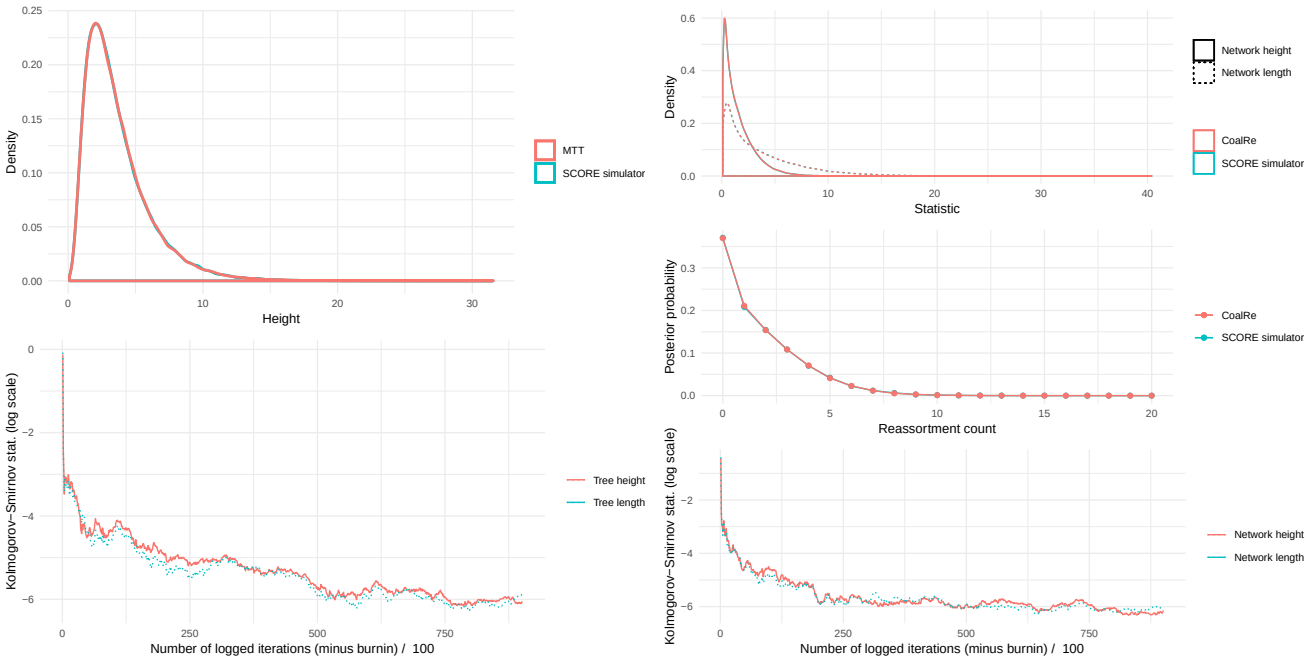


Figure B.1: Comparison of tree statistics for structured coalescent trees sampled under MTT and simulated under SCORE simulator (**left**, 5 taxa of 2 states) or coalescent with reassortment networks sampled under CoalRe and simulated under SCORE simulator (**right**, 2 taxa, each carrying 3 segments).
Left top: Sampled and simulated distributions of tree height and length. **Left bottom**: Difference between distributions quantified as Kolmogorov Smirnov difference.
Right top: Sampled and simulated distributions of network height and length. **Right middle**: Sampled and simulated numbers of reassortment events. **Right bottom**: Difference between distributions quantified as Kolmogorov Smirnov difference.

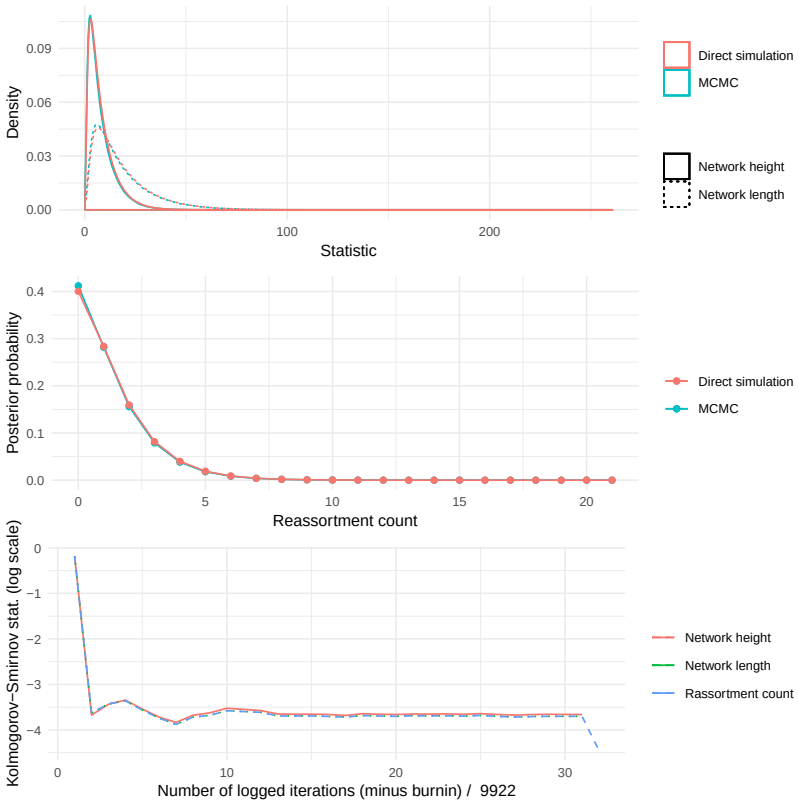


Figure B.2: Comparison of network statistics for sampled and simulated structured coalescent with reassortment for 3 taxa of 2 states, each carrying 3 segments. MCMC sampling done by the exact SCORE variant. **Top**: Sampled and simulated distributions of network height and length. **Middle**: Sampled and simulated numbers of reassortment events. **Bottom**: Difference between distributions quantified as Kolmogorov Smirnov difference.

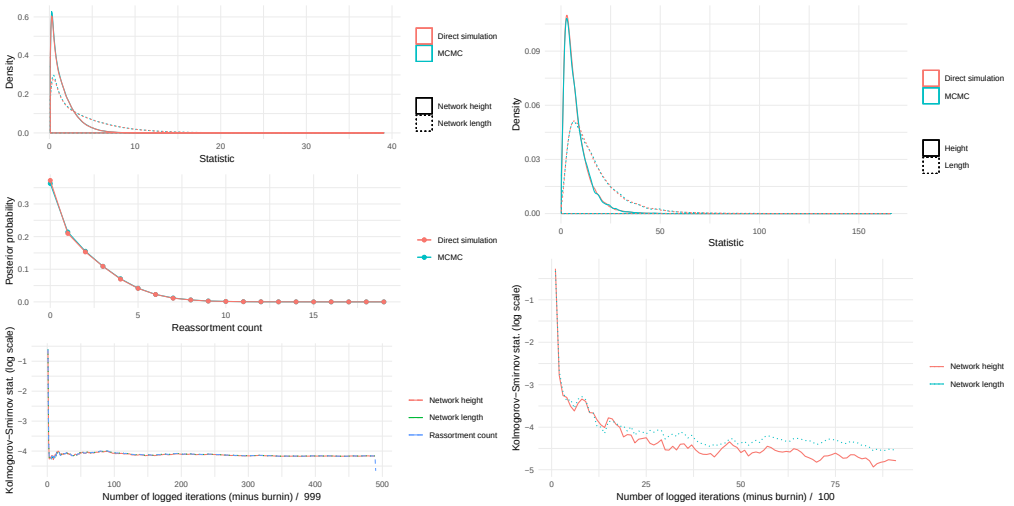


Figure B.3: Comparison of network statistics for sampled and simulated coalescent with reassortment (**left**, 2 taxa, each carrying 3 segments) or structured coalescent without reassortment (**right**, 3 taxa, 2 states). MCMC sampling done by the exact SCORE variant.
Left top: Sampled and simulated distributions of network height and length. **Left middle**: Sampled and simulated numbers of reassortment events. **Left bottom**: Difference between distributions quantified as Kolmogorov Smirnov difference.
Righ top: Sampled and simulated distributions of network height and length. **Right bottom**: Difference between distributions quantified as Kolmogorov Smirnov difference.

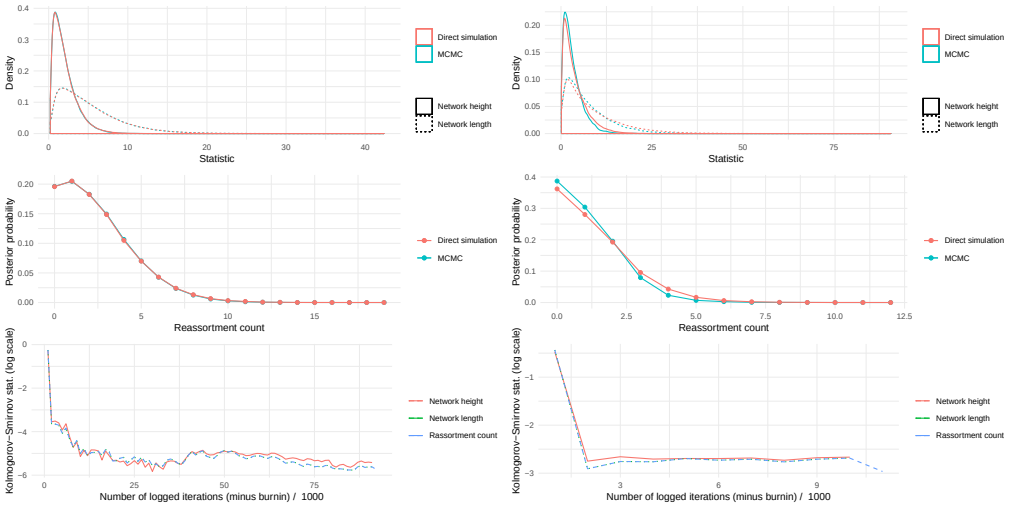


Figure B.4: Comparison of network statistics for sampled and simulated coalescent with reassortment (**left**, 3 taxa, each carrying 3 segments) or structured coalescent with reassortment (**right**, 2 taxa of 2 states, each carrying 2 segments). MCMC sampling done by the approximate SCORE variant. **Top**: Sampled and simulated distributions of network height and length. **Middle**: Sampled and simulated numbers of reassortment events. **Bottom**: Difference between distributions quantified as Kolmogorov Smirnov difference.

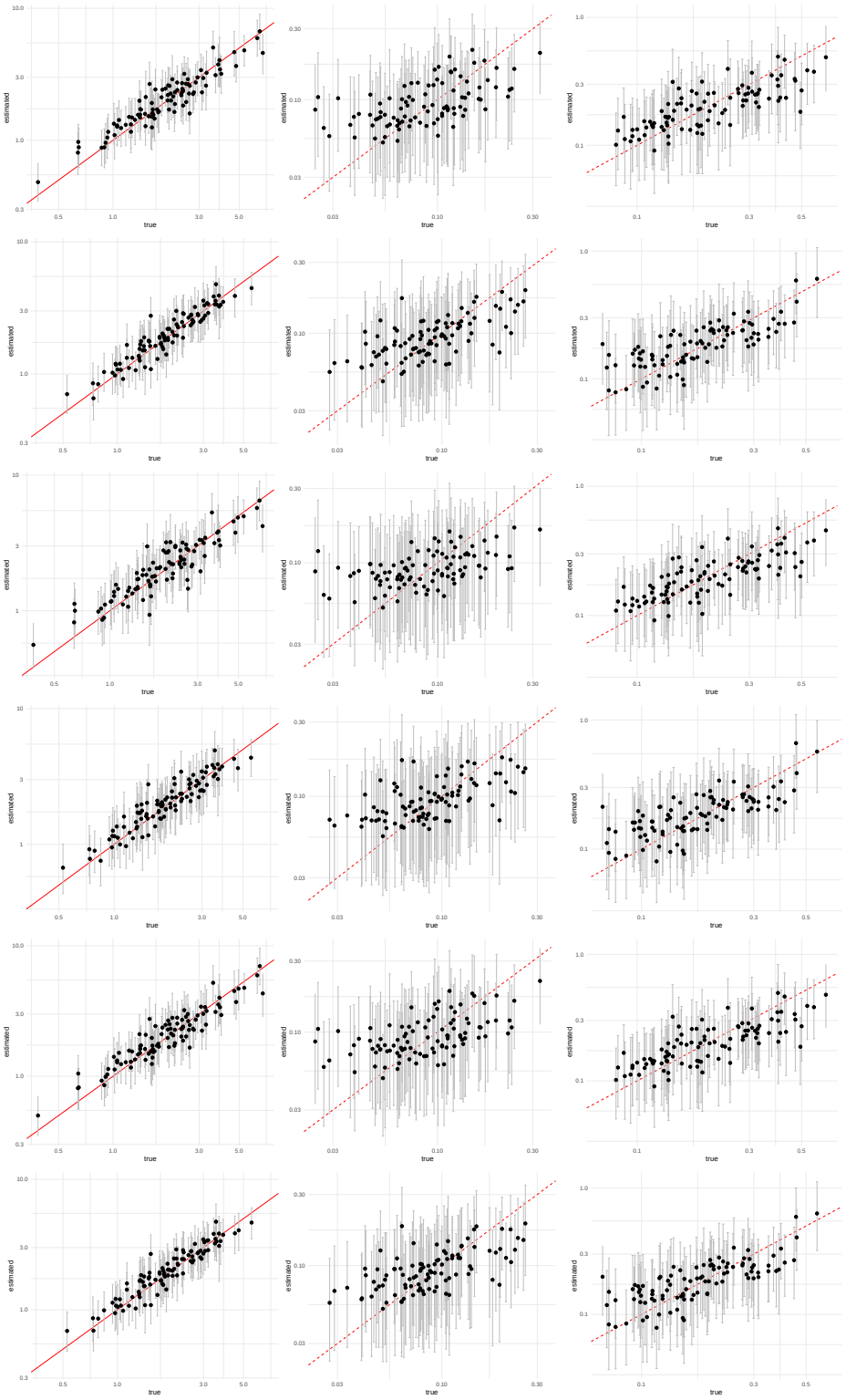


Figure B.5: Network parameter inference from 100 simulated genetic sequence data in 2 states for log-normal migration rate distribution. From top to bottom each row pair shows the high (5×10^{-3}), low (5×10^{-4}) or mixed clock rate case. First row of the pair is for state 1 and second – for state 2. **Left.** True (x-axis) versus estimated (y-axis) effective population sizes. Grey bars are 95% confidence intervals, red marks the x=y curve. **Middle.** True (x-axis) versus estimated (y-axis) reassortment rates. Grey bars are 95% confidence intervals, red marks the x=y curve. **Right.** True (x-axis) versus estimated (y-axis) effective population sizes. Grey bars are 95% confidence intervals, red marks the x=y curve.

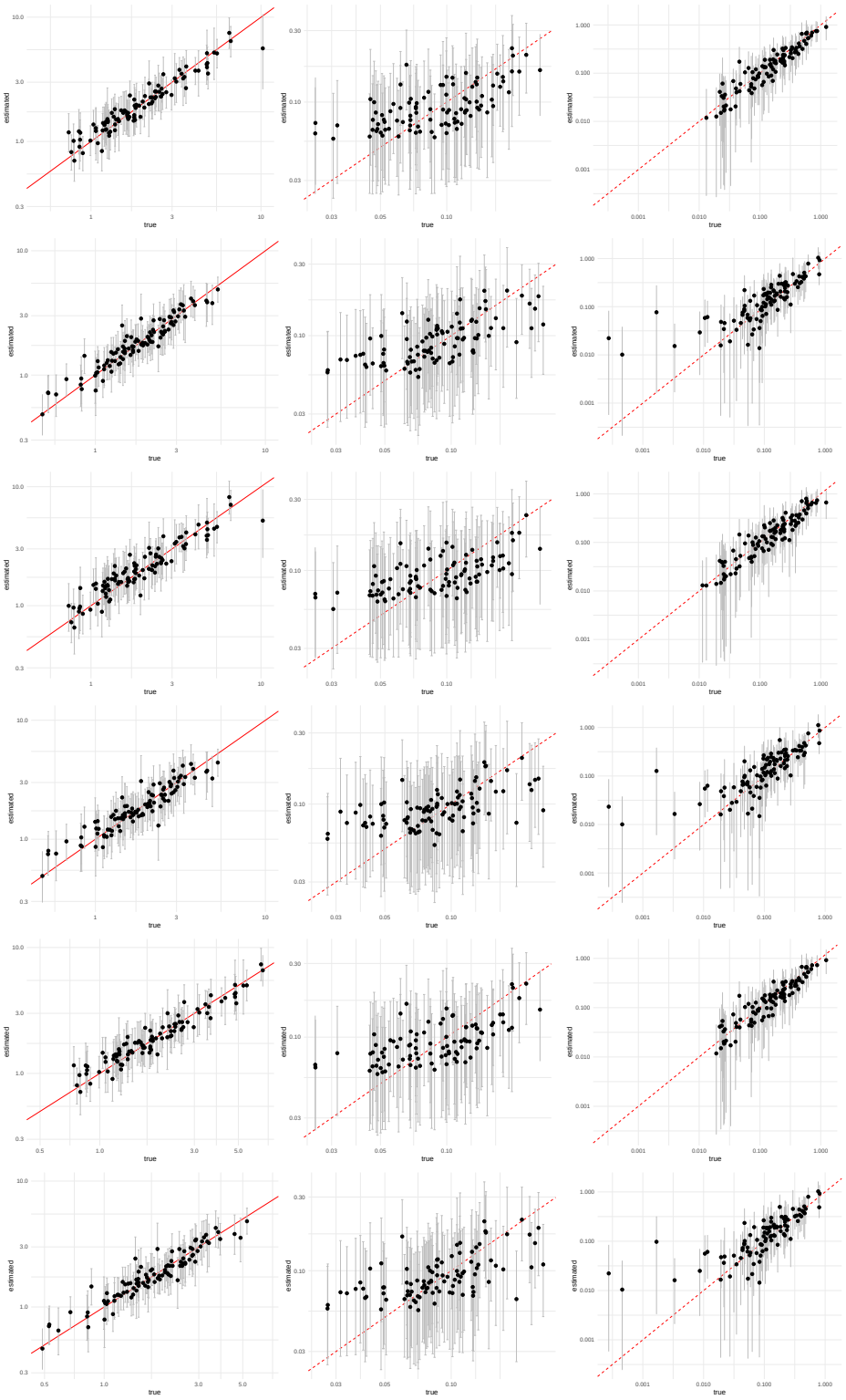


Figure B.6: Network parameter inference from 100 simulated genetic sequence data in 2 states for exponential migration rate distribution. From top to bottom each row pair shows the high (5×10^{-3}), low (5×10^{-4}) or mixed clock rate case. First row of the pair is for state 1 and second – for state 2. **Left.** True (x-axis) versus estimated (y-axis) effective population sizes. Grey bars are 95% confidence intervals, red marks the x=y curve. **Middle.** True (x-axis) versus estimated (y-axis) reassortment rates. Grey bars are 95% confidence intervals, red marks the x=y curve. **Right.** True (x-axis) versus estimated (y-axis) effective population sizes. Grey bars are 95% confidence intervals, red marks the x=y curve.

Appendix C

Data

Below are the names of 150 influenza A/H3N2 sequences used for SCORE and MASCOT analysis. Names are coded as follows: Strain Name|Location|Date.

A/Auckland/608/2002|New_Zealand|2002-06-08
A/Auckland/609/2002|New_Zealand|2002-06-18
A/Auckland/610/2002|New_Zealand|2002-07-29
A/Auckland/611/2002|New_Zealand|2002-09-09
A/Auckland/613/2002|New_Zealand|2002-08-01
A/Auckland/614/2002|New_Zealand|2002-01-14
A/Auckland/621/2005|New_Zealand|2005-01-30
A/California/NHRC0001/2003|USA|2003-11-13
A/California/NHRC0003/2003|USA|2003-12-17
A/California/NHRC0004/2003|USA|2003-12-19
A/California/NHRC0006/2003|USA|2003-12-10
A/Canterbury/01/2002|New_Zealand|2002-04-26
A/Canterbury/02/2002|New_Zealand|2002-04-29
A/Canterbury/104/2004|New_Zealand|2004-09-07
A/Canterbury/105/2004|New_Zealand|2004-09-07
A/Canterbury/108/2004|New_Zealand|2004-09-12
A/Canterbury/12/2004|New_Zealand|2004-08-05
A/Canterbury/124/2005|New_Zealand|2005-07-01
A/Canterbury/129/2005|New_Zealand|2005-07-03
A/Canterbury/16/2005|New_Zealand|2005-05-20
A/Canterbury/22/2002|New_Zealand|2002-05-10
A/Canterbury/232/2005|New_Zealand|2005-07-16
A/Canterbury/24/2005|New_Zealand|2005-05-04
A/Canterbury/303/2004|New_Zealand|2004-09-25
A/Canterbury/386/2003|New_Zealand|2003-07-16
A/Canterbury/395/2003|New_Zealand|2003-07-26
A/Canterbury/405/2003|New_Zealand|2003-07-05
A/Canterbury/67/2005|New_Zealand|2005-06-21
A/Canterbury/68/2002|New_Zealand|2002-07-05
A/Canterbury/72/2002|New_Zealand|2002-07-06
A/Canterbury/75/2002|New_Zealand|2002-07-11
A/Canterbury/80/2002|New_Zealand|2002-07-11
A/Canterbury/81/2002|New_Zealand|2002-07-09
A/Christchurch/15/2004|New_Zealand|2004-08-12
A/Georgia/NHRC0001/2003|USA|2003-12-05
A/Georgia/NHRC0001/2004|USA|2004-01-20
A/Hong_Kong/HKU1/2004|South_East_Asia|2004-02-10
A/Hong_Kong/HKU10/2004|South_East_Asia|2004-03-22
A/Hong_Kong/HKU11/2004|South_East_Asia|2004-03-25
A/Hong_Kong/HKU12/2004|South_East_Asia|2004-03-29
A/Hong_Kong/HKU15/2004|South_East_Asia|2004-04-23
A/Hong_Kong/HKU18/2004|South_East_Asia|2004-05-28
A/Hong_Kong/HKU2/2004|South_East_Asia|2004-02-13
A/Hong_Kong/HKU21/2004|South_East_Asia|2004-06-19
A/Hong_Kong/HKU23/2004|South_East_Asia|2004-07-02
A/Hong_Kong/HKU24/2004|South_East_Asia|2004-07-05
A/Hong_Kong/HKU25/2004|South_East_Asia|2004-07-07
A/Hong_Kong/HKU28/2004|South_East_Asia|2004-07-17
A/Hong_Kong/HKU30/2004|South_East_Asia|2004-07-22
A/Hong_Kong/HKU31/2004|South_East_Asia|2004-07-30
A/Hong_Kong/HKU32/2004|South_East_Asia|2004-08-02
A/Hong_Kong/HKU33/2004|South_East_Asia|2004-08-09
A/Hong_Kong/HKU34/2004|South_East_Asia|2004-08-18
A/Hong_Kong/HKU37/2004|South_East_Asia|2004-09-17
A/Hong_Kong/HKU4/2004|South_East_Asia|2004-03-01
A/Hong_Kong/HKU40/2004|South_East_Asia|2004-11-23
A/Hong_Kong/HKU44/2005|South_East_Asia|2005-02-28
A/Hong_Kong/HKU45/2005|South_East_Asia|2005-03-07
A/Hong_Kong/HKU47/2005|South_East_Asia|2005-03-19
A/Hong_Kong/HKU48/2005|South_East_Asia|2005-04-01
A/Hong_Kong/HKU49/2005|South_East_Asia|2005-04-04
A/Hong_Kong/HKU5/2004|South_East_Asia|2004-03-04
A/Hong_Kong/HKU52/2005|South_East_Asia|2005-04-16
A/Hong_Kong/HKU58/2005|South_East_Asia|2005-05-05
A/Hong_Kong/HKU61/2005|South_East_Asia|2005-05-11
A/Hong_Kong/HKU62/2005|South_East_Asia|2005-05-13

A/Hong_Kong/HKU63/2005|South_East_Asia|2005-05-14
A/Hong_Kong/HKU64/2005|South_East_Asia|2005-05-16
A/Hong_Kong/HKU65/2005|South_East_Asia|2005-05-17
A/Hong_Kong/HKU69/2005|South_East_Asia|2005-05-27
A/Hong_Kong/HKU7/2004|South_East_Asia|2004-03-10
A/Hong_Kong/HKU70/2005|South_East_Asia|2005-05-30
A/Hong_Kong/HKU71/2005|South_East_Asia|2005-05-31
A/Hong_Kong/HKU72/2005|South_East_Asia|2005-06-01
A/Hong_Kong/HKU73/2005|South_East_Asia|2005-06-02
A/Hong_Kong/HKU75/2005|South_East_Asia|2005-06-17
A/Hong_Kong/HKU76/2005|South_East_Asia|2005-06-21
A/Hong_Kong/HKU77/2005|South_East_Asia|2005-06-29
A/Hong_Kong/HKU79/2005|South_East_Asia|2005-07-16
A/Hong_Kong/NHRC0001/2005|South_East_Asia|2005-06-15
A/Illinois/NHRC0001/2005|USA|2005-04-06
A/Illinois/NHRC0002/2005|USA|2005-12-23
A/Missouri/NHRC0002/2004|USA|2004-01-09
A/New_Jersey/NHRC0001/2003|USA|2003-11-26
A/New_Jersey/NHRC0001/2005|USA|2005-02-10
A/New_York/1003/2005|USA|2005-12-20
A/New_York/105/2002|USA|2002-02-05
A/New_York/108/2002|USA|2002-01-31
A/New_York/109/2002|USA|2002-01-30
A/New_York/130/2002|USA|2002-03-11
A/New_York/132/2002|USA|2002-01-27
A/New_York/133/2002|USA|2002-02-06
A/New_York/134/2002|USA|2002-01-04
A/New_York/18/2003|USA|2003-12-01
A/New_York/2/2003|USA|2003-10-27
A/New_York/22/2003|USA|2003-12-02
A/New_York/31/2004|USA|2004-01-05
A/New_York/33/2004|USA|2004-01-09
A/New_York/351/2004|USA|2004-11-06
A/New_York/352/2005|USA|2005-01-04
A/New_York/356/2004|USA|2004-12-01
A/New_York/361/2005|USA|2005-02-02
A/New_York/374/2004|USA|2004-12-27
A/New_York/377/2004|USA|2004-12-20
A/New_York/381/2004|USA|2004-12-28
A/New_York/41/2003|USA|2003-12-08
A/New_York/410/2002|USA|2002-01-15
A/New_York/412/2002|USA|2002-02-01
A/New_York/413/2002|USA|2002-02-08
A/New_York/415/2002|USA|2002-02-22
A/New_York/416/2002|USA|2002-02-28
A/New_York/44/2003|USA|2003-12-08
A/New_York/461/2005|USA|2005-02-02
A/New_York/466/2004|USA|2004-11-27
A/New_York/471/2004|USA|2004-11-08
A/New_York/472/2004|USA|2004-09-27
A/New_York/6/2004|USA|2004-01-06
A/New_York/64/2003|USA|2003-12-26
A/New_York/65/2003|USA|2003-12-29
A/New_York/69/2004|USA|2004-01-06
A/New_York/70/2004|USA|2004-01-06
A/New_York/97/2002|USA|2002-02-13
A/Singapore/NHRC0001/2003|South_East_Asia|2003-09-28
A/Singapore/NHRC0003/2003|South_East_Asia|2003-09-30
A/Singapore/NHRC0004/2003|South_East_Asia|2003-10-01
A/Singapore/NHRC0006/2003|South_East_Asia|2003-10-07
A/Singapore/NHRC0008/2003|South_East_Asia|2003-10-08
A/Singapore/NHRC0009/2003|South_East_Asia|2003-10-09
A/South_Carolina/NHRC0001/2005|USA|2005-02-11
A/Southland/5/2005|New_Zealand|2005-08-08
A/Texas/NHRC0001/2005|USA|2005-02-16
A/Waikato/133/2003|New_Zealand|2003-07-08
A/Waikato/139/2003|New_Zealand|2003-07-24
A/Waikato/147/2003|New_Zealand|2003-07-28
A/Waikato/148/2003|New_Zealand|2003-07-25
A/Waikato/16/2004|New_Zealand|2004-09-09
A/Waikato/21/2003|New_Zealand|2003-06-21
A/Waikato/23/2002|New_Zealand|2002-07-03
A/Waikato/51/2002|New_Zealand|2002-08-12
A/Waikato/56/2004|New_Zealand|2004-09-29
A/Waikato/7/2005|New_Zealand|2005-08-10
A/Waikato/91/2003|New_Zealand|2003-06-28
A/Waikato/94/2003|New_Zealand|2003-06-27
A/Wellington/3/2005|New_Zealand|2005-07-13
A/Wellington/4/2005|New_Zealand|2005-07-12
A/Wellington/47/2003|New_Zealand|2003-07-07
A/Wellington/5/2005|New_Zealand|2005-07-26
A/Wellington/52/2004|New_Zealand|2004-09-23
A/Wellington/64/2004|New_Zealand|2004-10-19
A/Whanganui/128/2004|New_Zealand|2004-09-10