The Structured Coalescent and its Approximations: Supplementary Material

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

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

1Corresponding authors

Derivation of the differential equations for MASCO

We here derive the change in marginal lineage state probability over time, $P_t(L_i = l_i, T)$, from the ESCO interval contribution differential equations governing the change in the probability of a configuration $\mathcal{K}$ and the coalescent history T. We denote with $\mathcal{K} \setminus i$ a configuration for lineages $1, 2, \ldots, i-1, i+1, \ldots, n$. $P_t(L_i = l_i, T)$ denotes the probability of lineage i being in state l_i at time t joint with the probability of the coalescent history T up to time t. We can express $P_t(L_i = l_i, T)$ and its derivative as:

$$\begin{aligned} P(L_i = l_i, T) &= \sum_{\mathcal{K}\setminus i} P_t(\mathcal{K}, T) \\ \frac{d}{dt} P_t(L_i = l_i, T) &= \frac{d}{dt} \sum_{\mathcal{K}\setminus i} P_t(\mathcal{K}, T) = \sum_{\mathcal{K}\setminus i} \frac{d}{dt} P_t(\mathcal{K}, T) \end{aligned}$$

with $\sum_{\mathcal{K}\setminus i}$ denoting the summation over all configurations except fixing the state of lineage i. Using Eqn. 2 from the main text, we can express $\frac{d}{dt} P(L_i = l_i, T)$ as:

$$\frac{d}{dt} P_t(L_i = l_i, T) = \sum_{\mathcal{K}\setminus i} \sum_{j=1}^{n} \sum_{a=1}^{m} \Big(\mu_{a l_j} P_t(\mathcal{K}_{ja}, T) - \mu_{l_j a} P_t(\mathcal{K}_{j l_j}, T) \Big) - \sum_{\mathcal{K}\setminus i} \left(\sum_{a=1}^{m} \lambda_a \binom{k_a(\mathcal{K})}{2} P_t(\mathcal{K}, T) \right)$$

with $\mathcal{K}_{ja}$ being the configuration $\mathcal{K} \setminus j$ and lineage j being in a. We can now split the cases where $j \neq i$ and where $j = i$ to obtain:

$$\begin{aligned} \frac{d}{dt} P_t(L_i = l_i, T) = &\sum_{\mathcal{K}\setminus i} \sum_{\substack{j=1 \\ j\neq i}}^{n} \sum_{a=1}^{m} \Big(\mu_{a l_j} P_t(\mathcal{K}_{ja}, T) - \mu_{l_j a} P_t(\mathcal{K}_{j l_j}, T) \Big) \\ &+ \sum_{\mathcal{K}\setminus i}^{n} \sum_{a=1}^{m} \Big(\mu_{a l_i} P_t(\mathcal{K}_{ia}, T) - \mu_{l_i a} P_t(\mathcal{K}_{i l_i}, T) \Big) \\ &- \sum_{\mathcal{K}\setminus i} \left(\sum_{a=1}^{m} \lambda_a \binom{k_a(\mathcal{K})}{2} P_t(\mathcal{K}, T) \right) \end{aligned}$$

We note that $\sum_{\mathcal{K}\setminus i} \sum_{\substack{j\neq i \\ j=1}}^{n} \sum_{a=1}^{m} \mu_{a l_j} P_t(\mathcal{K}_{ja}, T)$ is the rate to leave via migration any configuration on lineages $1, 2, \ldots, i-1, i+1, \ldots, n$. Further, $\sum_{\mathcal{K}\setminus i} \sum_{j\neq i}^{n} \sum_{a=1}^{m} \mu_{l_j a} P_t(\mathcal{K}_{j l_j}, T)$ is the rate to enter via migration any configuration on lineages $1, 2, \ldots, i-1, i+1, \ldots, n$. Since the net change of probability mass due to migration for lineages other than i has to always be 0, we can write:

$$\frac{d}{dt} P_t(L_i = l_i, T) = \sum_{\mathcal{K}\setminus i}^{n} \sum_{a=1}^{m} \Big(\mu_{a l_i} P_t(\mathcal{K}_{ia}, T) - \mu_{l_i a} P_t(\mathcal{K}_{i l_i}, T) \Big) - \sum_{\mathcal{K}\setminus i} \left(\sum_{a=1}^{m} \lambda_a \binom{k_a(\mathcal{K})}{2} P_t(\mathcal{K}, T) \right)$$

Using the fact that $\sum_{\mathcal{K}\setminus i}^{n} P_t(\mathcal{K}_{ia}, T) = P_t(L_i = a, T)$, we get:

$$\frac{d}{dt} P_t(L_i = l_i, T) = \sum_{a=1}^{m} \Big(\mu_{a l_i} P_t(L_i = a, T) - \mu_{l_i a} P_t(L_i = l_i, T) \Big) - \sum_{\mathcal{K}\setminus i} \left(\sum_{a=1}^{m} \lambda_a \binom{k_a(\mathcal{K})}{2} P_t(\mathcal{K}, T) \right) \quad (1)$$

The first term denotes the reallocation of probability mass due to migration. The second term describes the loss of probability mass due to the coalescent process, with the rate at which coalescent events occur.

Following the main text, $k_a(\mathcal{K})$ denotes the number of lineages in state a for configuration $\mathcal{K}$. We next express $\binom{k_a(\mathcal{K})}{2}$ by using the probabilities of lineages j and k being in state a conditional on the configuration $\mathcal{K}$: we have $k_a(\mathcal{K}) = \sum_{j=1}^{n} P_t(L_j = a|\mathcal{K}, T)$, with $P_t(L_j = a|\mathcal{K}, T) = 1$ if $L_j = a$ in configuration $\mathcal{K}$, and 0 otherwise. We can write the term $\binom{k_a(\mathcal{K})}{2}$, i.e. the number of pairs of lineages that are both in state a, as sums over these conditional probabilities. Imagine a matrix with n rows and n columns, with entries $P_t(L_j = a, L_k = a|\mathcal{K}, T)$ at position (j, k). If we want to get the number of pairs of lineages (j, k) that are both in state a (i.e. $\binom{k_a(\mathcal{K})}{2}$), we can count every element in the upper triangle for which both lineages are in state a (i.e. $\sum_{j=1}^{n}\sum_{k=j+1}^{n} P_t(L_j = a, L_k = a|\mathcal{K}, T)$). We can now write this double summation as $\frac{1}{2}\sum_{j=1}^{n}\sum_{k\neq j}^{n}$, since the matrix is symmetric. This leads to:

$$\sum_{\mathcal{K}\backslash i}\left(\sum_{a=1}^{m}\lambda_a\binom{k_a(\mathcal{K})}{2}P_t(\mathcal{K},T)\right) = \sum_{\mathcal{K}\backslash i}\sum_{a=1}^{m}\frac{\lambda_a}{2}\sum_{j=1}^{n}\sum_{\substack{k=1\\k\neq j}}^{n}P_t(L_j = a, L_k = a|\mathcal{K},T)P_t(\mathcal{K},T) \tag{2}$$

We now write the above term explicitly for cases involving lineage i and cases not involving lineage i and write:

$$\begin{aligned}\sum_{\mathcal{K}\backslash i}\left(\sum_{a=1}^{m}\lambda_a\binom{k_a(\mathcal{K})}{2}P_t(\mathcal{K},T)\right) &= \sum_{a=1}^{m}\frac{\lambda_a}{2}\sum_{\substack{j=1\\j\neq i}}^{n}\sum_{\substack{k=1\\k\neq j,i}}^{n}\sum_{\mathcal{K}\backslash i}P_t(L_j = a, L_k = a|\mathcal{K},T)P_t(\mathcal{K},T)\\ &\quad+\sum_{a=1}^{m}\frac{\lambda_a}{2}\sum_{\substack{k=1\\k\neq i}}^{n}\sum_{\mathcal{K}\backslash i}P_t(L_i = a, L_k = a|\mathcal{K},T)P_t(\mathcal{K},T)\\ &\quad+\sum_{a=1}^{m}\frac{\lambda_a}{2}\sum_{\substack{j=1\\j\neq i}}^{n}\sum_{\mathcal{K}\backslash i}P_t(L_j = a, L_i = a|\mathcal{K},T)P_t(\mathcal{K},T)\\ &= \sum_{a=1}^{m}\frac{\lambda_a}{2}\sum_{\substack{j=1\\j\neq i}}^{n}\sum_{\substack{k=1\\k\neq j,i}}^{n}\sum_{\mathcal{K}\backslash i}P_t(L_j = a, L_k = a|\mathcal{K},T)P_t(\mathcal{K},T)\\ &\quad+\sum_{a=1}^{m}\lambda_a\sum_{\substack{k=1\\k\neq i}}^{n}\sum_{\mathcal{K}\backslash i}P_t(L_i = a, L_k = a|\mathcal{K},T)P_t(\mathcal{K},T)\end{aligned} \tag{3}$$

We now derive an expression for $\sum_{\mathcal{K}\backslash i} P_t(L_j = a, L_k = a|\mathcal{K}, T)P_t(\mathcal{K}, T)$:

$$\begin{aligned}\sum_{\mathcal{K}\backslash i}P_t(L_j = a, L_k = a|\mathcal{K},T)P_t(\mathcal{K},T) &= \sum_{\mathcal{K}\backslash i}P_t(L_j = a, L_k = a, \mathcal{K}, T)\\ &= P_t(L_j = a, L_k = a, L_i = l_i, T)\\ &= P_t(L_j = a, L_k = a, L_i = l_i|T)P_t(T)\end{aligned} \tag{4}$$

Next, we can assume that the state of lineages j, k and i, with $j \neq k \neq i$, are mutually independent, i.e. we assume that:

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

This allows us to simplify equation 4, for $j, k \neq i$, to:

$$P_t(L_j = a, L_k = a, L_i = l_i|T)P_t(T) = P_t(L_j = a|T)P_t(L_k = a|T)P_t(L_i = l_i, T) \tag{5}$$

We can now plug equation 5 into equation 3 to receive:

$$\begin{aligned}
\sum_{\mathcal{K}\setminus i}\left(\sum_{a=1}^{m}\lambda_a\binom{k_a(\mathcal{K})}{2}P_t(\mathcal{K},T)\right) &= P_t(L_i=l_i,T)\sum_{a=1}^{m}\frac{\lambda_a}{2}\sum_{\substack{j=1\\j\neq i}}^{n}\sum_{\substack{k=1\\k\neq j,i}}^{n}P_t(L_j=a|T)P_t(L_k=a|T)\\
&+\sum_{a=1}^{m}\lambda_a\sum_{\substack{k=1\\k\neq i}}^{n}P_t(L_i=a,L_k=a,L_i=l_i|T)P_t(T)\\
&= P_t(L_i=l_i,T)\sum_{a=1}^{m}\frac{\lambda_a}{2}\sum_{\substack{j=1\\j\neq i}}^{n}\sum_{\substack{k=1\\k\neq j,i}}^{n}P_t(L_j=a|T)P_t(L_k=a|T)\\
&+\lambda_{l_i}\sum_{\substack{k=1\\k\neq i}}^{n}P_t(L_i=l_i,L_k=a|T)P_t(T)\\
&= P_t(L_i=l_i,T)\sum_{a=1}^{m}\frac{\lambda_a}{2}\sum_{\substack{j=1\\j\neq i}}^{n}\sum_{\substack{k=1\\k\neq j,i}}^{n}P_t(L_j=a|T)P_t(L_k=a|T)\\
&+P_t(L_i=l_i,T)\lambda_{l_i}\sum_{\substack{k=1\\k\neq i}}^{n}P_t(L_k=l_i|T)
\end{aligned}$$

using the MASCO approximation for the last line. Plugging the above expression for $\sum_{\mathcal{K}\setminus i}\left(\sum_{a=1}^{m}\frac{\lambda_a}{2}\binom{k_a(\mathcal{K})}{2}P_t(\mathcal{K},T)\right)$ into equation 1 yields:

$$\begin{aligned}
\frac{d}{dt}P_t(L_i=l_i,T) = \sum_{a=1}^{m}\Big(\mu_{al_i}P_t(L_i=a,T)-\mu_{l_ia}P_t(L_i=l_i,T)\Big)\\
-P_t(L_i=l_i,T)\sum_{a=1}^{m}\frac{\lambda_a}{2}\sum_{\substack{j=1\\j\neq i}}^{n}\sum_{\substack{k=1\\k\neq j,i}}^{n}P_t(L_j=a|T)P_t(L_k=a|T)\\
-P_t(L_i=l_i,T)\lambda_{l_i}\sum_{\substack{k=1\\k\neq i}}^{n}P_t(L_k=l_i|T) \qquad (6)
\end{aligned}$$

The first term now describes how the marginal probability of lineage i in state l_i changes due to migration. The second line denotes the rate at which coalescent events happen that do not directly involve lineage i and the third line denotes the rate at which coalescent events involving lineage i happen. The reason why we need to consider events that do not involve lineage i is that we seek to calculate the probability of lineages i in state l_i jointly with the full coalescent history T.

Derivation of the differential equations for SISCO

The differential equation describing the change in probability over time using the approximation of state independence can be derived from equation 6. In addition to the MASCO assumption, we now further assume $P_t(L_i = b|T) \overset{SISCO}{=} P_t(L_i = b)$. The differential equations for $P_t(L_i = b, T), i = 1, \ldots, m$ under MASCO become differential equation for $P_t(T)$ and $P_t(L_i = b), i = 1, \ldots, m$ when using the SISCO approximation. We will now first show how to derive $P_t(T)$ and then how to derive $P_t(L_i = b), i = 1, \ldots, m$. Since under SISCO,

$$\sum_{b=1}^{m} P_t(L_i = b, T) = P_t(T) \sum_{b=1}^{m} P_t(L_i = b) = P_t(T),$$

where the second equality follows from $\sum_{b=1}^{m} P_t(L_i = b) = 1$, we obtain,

$$\frac{d}{dt} P_t(T) = \frac{d}{dt} \sum_{b=1}^{m} P_t(L_i = b, T).$$

From the MASCO equation 6 with additionally using the SISCO approximation, we obtain,

$$\begin{align}
\frac{d}{dt} P_t(T) = \frac{d}{dt} \sum_{b=1}^{m} P_t(L_i = b, T) &= \sum_{b=1}^{m} \sum_{a=1}^{m} \Big(\mu_{ab} P_t(L_i = a, T) - \mu_{ba} P_t(L_i = b, T) \Big) \notag \\
&\quad - \sum_{b=1}^{m} P_t(L_i = b, T) \sum_{a=1}^{m} \frac{\lambda_a}{2} \sum_{\substack{j=1 \\ j \neq i}}^{n} \sum_{\substack{k=1 \\ k \neq j,i}}^{n} P_t(L_j = a) P_t(L_k = a) \notag \\
&\quad - \sum_{b=1}^{m} P_t(L_i = b, T) \lambda_b \sum_{\substack{k=1 \\ k \neq i}}^{n} P_t(L_k = b) \notag \\
&= \sum_{b=1}^{m} \sum_{a=1}^{m} \Big(\mu_{ab} P_t(L_i = a, T) - \mu_{ba} P_t(L_i = b, T) \Big) \notag \\
&\quad - P_t(T) \sum_{a=1}^{m} \frac{\lambda_a}{2} \sum_{\substack{j=1 \\ j \neq i}}^{n} \sum_{\substack{k=1 \\ k \neq j,i}}^{n} P_t(L_j = a) P_t(L_k = a) \notag \\
&\quad - P_t(T) \sum_{a=1}^{m} \lambda_a \sum_{\substack{k=1 \\ k \neq i}}^{n} P_t(L_i = a) P_t(L_k = a) \notag \\
&= \sum_{b=1}^{m} \sum_{a=1}^{m} \Big(\mu_{ab} P_t(L_i = a, T) - \mu_{ba} P_t(L_i = b, T) \Big) \notag \\
&\quad - P_t(T) \sum_{a=1}^{m} \frac{\lambda_a}{2} \sum_{j=1}^{n} \sum_{\substack{k=1 \\ k \neq j,i}}^{n} P_t(L_j = a) P_t(L_k = a) \notag \\
&\quad - P_t(T) \sum_{a=1}^{m} \frac{\lambda_a}{2} \sum_{\substack{j=1 \\ j \neq i}}^{n} P_t(L_i = a) P_t(L_j = a) \notag \\
&= \sum_{b=1}^{m} \sum_{a=1}^{m} \Big(\mu_{ab} P_t(L_i = a, T) - \mu_{ba} P_t(L_i = b, T) \Big) (= 0) \notag \\
&\quad - P_t(T) \sum_{a=1}^{m} \frac{\lambda_a}{2} \sum_{j=1}^{n} \sum_{\substack{k=1 \\ k \neq j}}^{n} P_t(L_j = a) P_t(L_k = a) \tag{7} \\
&= -P_t(T) \sum_{a=1}^{m} \frac{\lambda_a}{2} \sum_{j=1}^{n} \sum_{\substack{k=1 \\ k \neq j}}^{n} P_t(L_j = a) P_t(L_k = a) \tag{8}
\end{align}$$

We now derive how lineage move between states, i.e. an expression for the differential equation of $P_t(L_i = b)$. Under the SISCO approximation, we can write:

$$\frac{d}{dt}\sum_{b=1}^{m} P_t(L_i = b, T) = \sum_{b=1}^{m} P_t(L_i = b)\frac{d}{dt}P_t(T) + P_t(T)\sum_{b=1}^{m}\frac{d}{dt}P_t(L_i = b)$$

Since $\sum_{b=1}^{m} P_t(L_i = b) = 1$, the above expression simplifies to:

$$\frac{d}{dt}\sum_{b=1}^{m} P_t(L_i = b, T) \quad = \quad P_t(T)\sum_{b=1}^{m}\frac{d}{dt}P_t(L_i = b) + \frac{d}{dt}P_t(T)$$

and can be transformed to,

$$0 = \sum_{b=1}^{m}\frac{d}{dt}P_t(L_i = b) \quad = \quad \frac{1}{P_t(T)}\left(\frac{d}{dt}\sum_{b=1}^{m}P_t(L_i = b, T) - \frac{d}{dt}P_t(T)\right)$$

The right hand side is equal to $\frac{1}{P_t(T)}(eq\ 7 - eq\ 8)$. Since we aim to get an expression for $\frac{d}{dt}P_t(L_i = b)$, we write:

$$\begin{aligned}\sum_{b=1}^{m}\frac{d}{dt}P_t(L_i = b) \quad &= \quad \frac{1}{P_t(T)}\sum_{b=1}^{m}\sum_{a=1}^{m}\Big(\mu_{ab}P_t(L_i = a, T) - \mu_{ba}P_t(L_i = b, T)\Big)\\ &= \quad \sum_{b=1}^{m}\sum_{a=1}^{m}\Big(\mu_{ab}P_t(L_i = a|T) - \mu_{ba}P_t(L_i = b|T)\Big)\\ &= \quad \sum_{b=1}^{m}\sum_{a=1}^{m}\Big(\mu_{ab}P_t(L_i = a) - \mu_{ba}P_t(L_i = b)\Big)\end{aligned}$$

Hence, we can calculate $\frac{d}{dt}P_t(L_i = b)$ using the following equation,

$$\frac{d}{dt}P_t(L_i = b) = \sum_{a=1}^{m}\Big(\mu_{ab}P_t(L_i = a) - \mu_{ba}P_t(L_i = b)\Big). \tag{9}$$

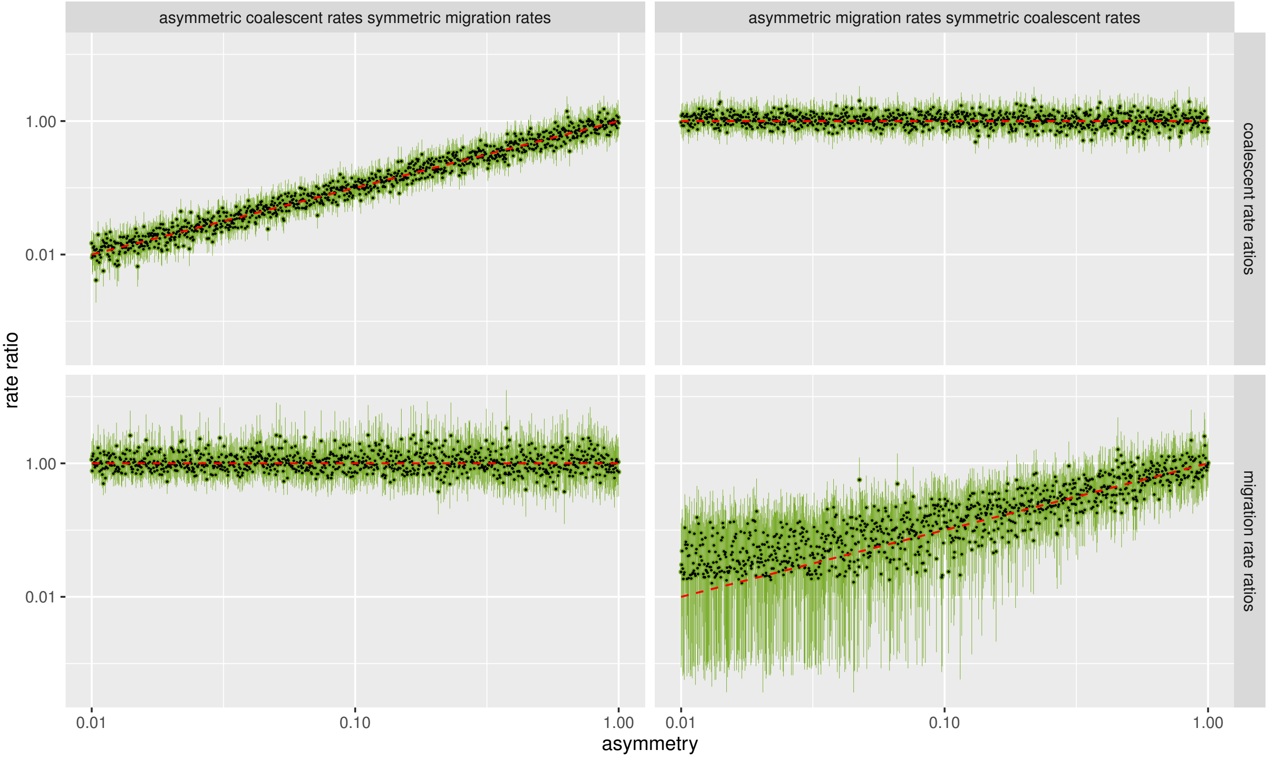

Figure S1: **Inferred asymmetry of migration and coalescent rates with confidence intervals using MASCO**. Here we show the inferred coalescent (upper row) and migration (lower row) rate ratios under different conditions. In the first column, the coalescent rate ratios (x-axis) are varied while the migration rates ratios are kept constant. In the second column, the migration rate ratios (x-axis) are varied, while the coalescent rate ratios are kept constant. The red line indicates where the estimates should lie. The black point represent the mean estimated ratios and the green lines represent the 95% confidence intervals of these estimates

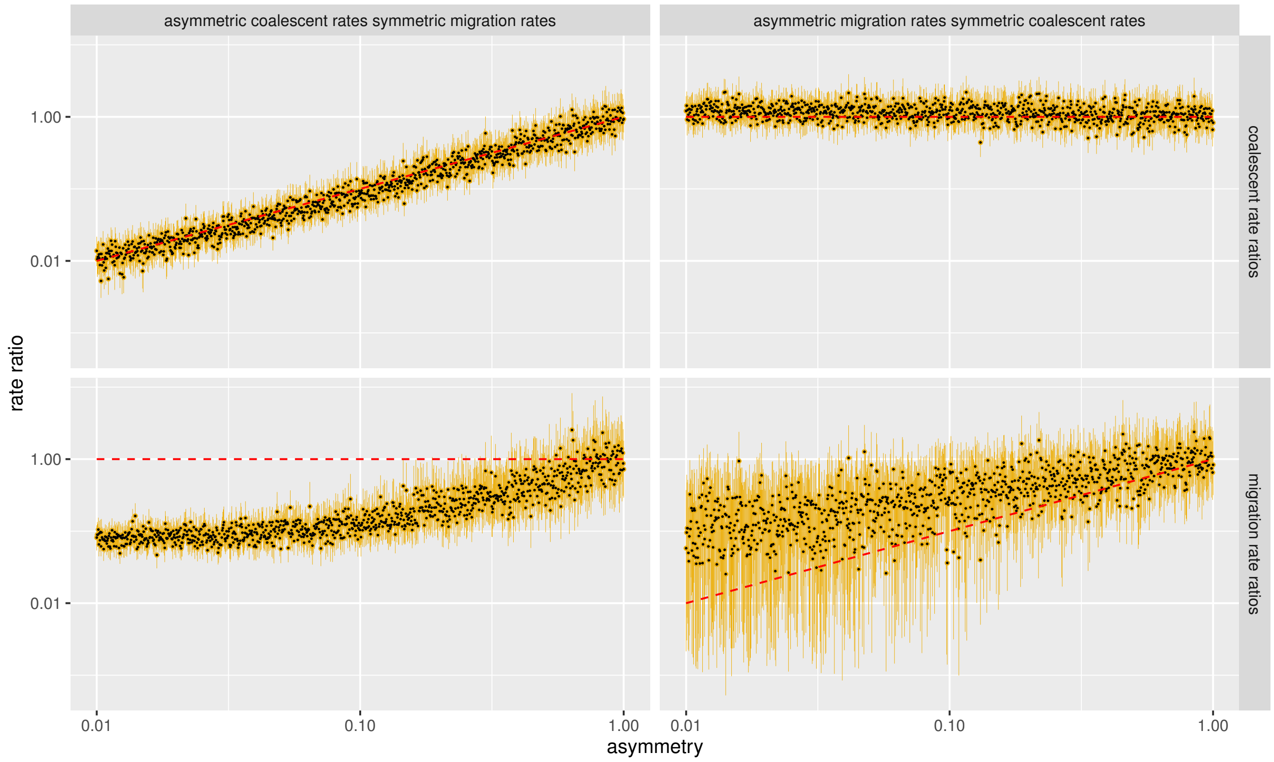

Figure S2: **Inferred asymmetry of migration and coalescent rates with confidence intervals using SISCO**. Here we show the inferred coalescent (upper row) and migration (lower row) rate ratios under different conditions. In the first column, the coalescent rate ratios (x-axis) are varied while the migration rates ratios are kept constant. In the second column, the migration rate ratios (x-axis) are varied, while the coalescent rate ratios are kept constant. The red line indicates where the estimates should lie. The black point represent the mean estimated ratios and the orange lines represent the 95% confidence intervals of these estimates

Table S1: Sample locations and associated regions for the here used AIV sequences

location	region	number of samples
Alaska	Alaska	18
Saskatchewan	North West	1
North Dakota	North West	1
British Columbia	North West	5
Alberta	North West	4
Washington	North West	3
California	South West	21
Mexico	South West	3
Mississippi	Center	7
Missouri	Center	6
Nebraska	Center	2
Texas	Center	2
Delaware	East Coast	7
Delaware Bay	East Coast	14
New Jersey	East Coast	13
Maryland	East Coast	2
Pennsylvania	East Coast	1
North Carolina	East Coast	3
Illinois	North Mid East	3
Ohio	North Mid East	4
Wisconsin	North Mid East	4
New Brunswick	North East	8
Nova Scotia	North East	1