Estimating Waiting Distances Between Genealogy Changes under a Multi-Species Extension of the Sequentially Markov Coalescent

Patrick F. McKenzie\(^1\) and Deren A. R. Eaton\(^1\*,\)

\(^1\)Department of Ecology, Evolution, and Environmental Biology, Columbia University, New York, NY 10027

*Contact: de2356@columbia.edu

Keywords: Recombination, Phylogeny, SMC, Gene Tree, Species Tree, Concatalescence, ARG

Abstract

Genomes are composed of a mosaic of segments inherited from different ancestors, each separated by past recombination events. Consequently, genealogical relationships among multiple genomes vary spatially across different genomic regions. Expectations for the amount of genealogical variation among unlinked (uncorrelated) genomic regions is well described for either a single population (coalescent) or multiple structured populations (multispecies coalescent). However, the expected similarity among genealogies at linked regions of a genome is less well characterized. Recently, an analytical solution was derived for the expected distribution of waiting distances between changes in genealogical trees spatially across a genome for a single population with constant effective population size. Here we describe a generalization of this result, in terms of the expected distribution of waiting distances between changes in genealogical trees and topologies, for multiple structured populations with branch-specific effective population sizes (i.e., under the multispecies coalescent). Our solutions establish an expectation for genetic linkage in multispecies datasets and provide a new likelihood framework for linking demographic models with local ancestry inference across genomes.

1 Introduction

The multispecies coalescent (MSC) is an extension of the coalescent (Kingman, 1982), a model that describes the distribution of genealogical histories among gene copies from a set of sampled individuals. Whereas the coalescent models a single panmictic population, the MSC includes constraints that prevent samples in different lineages from sharing a most recent genealogical ancestor until prior to a population divergence event that separates them (Maddison, 1997; Maddison & Knowles, 2006). Conceptually, the MSC can be viewed as a piecewise model composed of the standard coalescent applied to each interval of a "species tree" representing the relationships and divergence times among isolated lineages. Genealogies are constrained to be embedded within species trees (Fig. 1a), and the joint likelihood of MSC model parameters can be estimated from the coalescent times among a distribution of sampled genealogies (Degnan & Rosenberg, 2009; Rannala & Yang, 2003). In both the coalescent and MSC models, effective population size \(N_e\) is the key parameter determining the rate of coalescence and can vary across different lineages.

Importantly, both the coalescent and MSC are models of the expected distribution of \textit{unlinked} (uncorrelated) genealogies. By contrast, two linked genealogies that are drawn from nearby regions of a genome
are expected to be more similar than two random draws under these models. This spatial autocorrelation is a consequence of shared ancestry among samples at nearby regions, which decays over time and distance as recombination events reduce their shared ancestry. As a data structure, an ordered series of linked genealogies and the lengths of their associated intervals on a chromosome (Fig. 2) is represented by an ancestral recombination graph (ARG) (Griffiths & Marjoram, 1996), or similarly, a tree-sequence (Kelleher et al., 2016) (hereafter we will refer to it generically as an ARG).

An algorithm to stochastically simulate sequences under the coalescent with recombination was developed early on (Hudson, 1983) and was later extended as a spatial algorithm for stochastically generating full ARGs (Wiuf & Hein, 1999). This process models the difference between sequential genealogies by randomly detaching an edge from a genealogy and sampling a waiting time (based on the population coalescent rate) until it reconnects to the genealogy at a different shared ancestor. Thus, under this spatial model, a set of samples effectively substitutes one ancestor for another on either side of each recombination breakpoint. Implicit to this algorithm is the assumption that recombination occurs at some predictable rate (or rate map) from which the expected waiting distance between recombination events can be modeled as an exponentially distributed random variable (Wiuf & Hein, 1999). Although generating ARGs consistent with a demographic model is relatively simple under this process, inferring an ARG from sequence data – composing a set of recombination breakpoints and local genealogy inferences – remains highly challenging (Brandt et al., 2022). This stems in part from the great complexity of this problem but also reflects limitations of our current models for extracting historical information from linked genome data.

A major advance was achieved through development of the sequentially Markov coalescent (SMC),
a simpler approximation of the coalescent with recombination that restricts the types of recombination events that can occur (McVean & Cardin, 2005). Specifically, an edge that is detached from a genealogy by recombination is allowed only to re-coalesce with ancestral lineages that contributed genetic material to samples in that interval (as opposed to re-coalescing with any ancestral lineage). This greatly reduces the space of possible ARGs without changing the expected distribution of sequential genealogies, and in doing so it enables modeling changes between sequential genealogies as a Markov process (McVean & Cardin, 2005). Under these assumptions a tractable likelihood framework can be developed. Because neither genealogies nor segment lengths can be observed directly, most SMC-based inference methods use a hidden Markov model (HMM) to treat these as hidden states that influence observable changes in sequence data (Spence et al., 2018). Examples of inference tools built on the SMC framework include PSMC (Li & Durbin, 2011) and MSMC (Schiffels & Durbin, 2014) which use pairwise coalescent times between sequential genealogies to infer changes in effective population sizes through time, and ARG-weaver (Hubisz & Siepel, 2020; Rasmussen et al., 2014), which infers ARGs from genome alignments using an SMC-based conditional sampling method.

Marjoram & Wall (2006) described an important extension to the SMC, termed the SMC’, for additionally modeling "invisible" recombination events, in which a detached lineage re-attaches with its own ancestral lineage prior to the time of that lineage’s next coalescent event. This leads to no change between the genealogies in two sequential genomic intervals despite the occurrence of a recombination event between them. The inclusion of such events has been shown to significantly improve inference methods (Wilton et al., 2015). Under the SMC’, a detached lineage can thus re-coalesce with an allowable ancestral lineage over a continuous range of time, leading to one of four possible categorical patterns for the relationship between two sequential genealogies (Fig. 2, Fig. S1): (1) no change (2) shortening of a coalescent time; (3) lengthening of a coalescent time; and (4) a change to the genealogical topology (relationships). These can be grouped more generally into three types: "no-change" (category 1), "tree-change" (categories 2-4), and "topology-change" (category 4). Recently, Deng et al. (2021) derived a set of solutions for the expected waiting distances to each of these three types of outcomes for a single population with constant effective population size. This provided an important advance by establishing a neutral expectation not only for the distance until the next recombination event occurs, but more specifically, for the distance until different categorical types of recombination events occur. Such events leave different detectable signatures in sequence data and extend across different spatial distances of the genome, thus extending the scale over which information from spatial genealogical patterns can be extracted from genomes.

Here, we extend the methods of Deng et al. (2021) to an MSC framework to estimate the expected waiting distances between different types of genealogy changes under a parameterized species tree model. The waiting distances between recombination events that cause topology-change may be of greatest interest, as they leave the most detectable signatures in sequence data and are relevant to expected gene tree distributions that form an important component of many MSC-based methods (Baum, 2007; Degnan & Rosenberg, 2009; Knowles & Kubatko, 2011). The waiting distance between topology-change events may include multiple intervening recombination events of the no-change or tree-change type (Fig. 2). The relative occurrence of events that do not result in topology changes can be especially high in small sample sizes (Wilton et al., 2015), which are common in MSC-type datasets in which samples are partitioned among species tree intervals. Since the partitioning of coalescent events among species tree intervals is
**Figure 2.** An ancestral recombination graph (ARG) is composed of a series of genealogies each spanning non-overlapping intervals of a genome separated by recombination breakpoints. Here, four recombination breakpoints separate the intervals’ start and end positions along a small chromosome alignment. The individual genealogies represent the history of a set of 7 samples constrained by a 4-tip species tree model (as in Fig. 1). Recombination events are indicated by vertical panels. Each shows the SMC’ process by which a subtree (below the red circle) is detached and then re-coalesces (red arrow) with one of the remaining lineages. The former edge (blue dotted) which existed throughout the interval to the left of a panel is replaced by a new edge (red) through the subsequent interval. Vertical bars (white) represent barriers to coalescence between samples in different species tree intervals (MSC model lineages). Four categories of recombination events are shown from left to right, representing different outcomes based on the lineage with which a detached subtree re-coalesces. These are grouped more generally into three *event types*: (1) no-change, (2) tree-change, and (3) topology-change. Every recombination event causes either a no-change or tree-change event, whereas topology-change events are a subset of possible tree-change events. The expected waiting distance until a specific recombination event type occurs can be calculated under the MS-SMC given a starting genealogy, MSC model, and recombination rate.

expected to constrain the types of recombination events that will be observed, the distributions of waiting distances between different types of genealogy changes should be highly dependent on, and thus informative about, the species tree model. We refer to the general framework of embedding the SMC’ in an MSC model as the MS-SMC.

### 2 Approach

#### 2.1 Comparison to Deng et al. (2021)

Our approach is a generalization of the *Deng et al. (2021)* derivation of waiting distances to genealogy changes for a single population of constant size. We modified the single-population model to (1) include barriers to coalescence imposed by a species tree topology, and (2) integrate over changing coalescence rates along paths through multiple species tree intervals with different effective population sizes. In addi-
tion, we have implemented a novel likelihood framework for inferring parameters of an MSC model from observed waiting distances between genealogies.

2.2 MSC model description

Given an MSC model \((S)\) composed of a species tree topology with divergence times \((W)\) in units of generations and constant diploid effective population sizes \((N_e)\) assigned to each branch, an embedded genealogy \((G)\) for any number of sampled gene copies \((k)\) can be generated by randomly sampling coalescent times at which to join two samples into a common ancestor, starting from samples at the present in each interval. Following Kingman (1982), the probability of a coalescent event one generation in the past (reducing the number of samples from \(k\) to \(k-1\)) is given by equation 1. From this, we can model the expected waiting time \((E[t_k])\) until the next coalescence event as an exponentially distributed random variable with rate parameter \(\lambda_k\):

\[
\lambda_k = \mathbb{P}(\text{coal event}|N_e, k) = \frac{k(k-1)}{2N_e}
\]

and:

\[
E[t_k] = \frac{1}{\lambda_k}
\]  

(1)

In a single population model with constant \(N_e\) the expected waiting time between coalescent events increases monotonically after each coalescent event, since the number of remaining samples always decreases. In an MSC model, however, the expected waiting time between coalescence events can increase or decrease through time, as the transition from one population interval to another can be associated with a different \(N_e\) value and an increase in the number of samples.

Based on this generative framework for sampling genealogies, likelihood solutions have been developed to fit coalescent model parameters, such as \(N_e\) in single population models (Kingman, 1982), or multiple \(N_e\) and \(W\) parameters in MSC models (Rannala & Yang, 2003), based on inferred coalescence times. In the latter framework, each species tree branch interval is treated independently, such that the likelihood of a genealogy embedding is calculated from the joint probability of observing each distribution of coalescent waiting times within each species tree branch interval. A key feature of these equations is that when \(k\) lineages are present, we can use the coalescent rate parameters \((\lambda_k)\) to calculate the likelihood of observed waiting times between coalescent events \((t_k)\) in each population interval from an exponential probability density function:

\[
f(t_k; \lambda) = \lambda e^{-\lambda t_k}
\]  

(2)

2.3 MS-SMC model description and notation

Under the SMC’ model, sampling of a linked genealogy requires considering not only demographic model parameters, as we did above, but also an existing genealogy – it is a method for sampling the next genealogy conditional on the previously observed one. If we define the previous genealogy as \(G\), and the sum of its edge lengths as \(L(G)\), then under the assumption of a constant recombination rate through time,
Table 1. A genealogy embedding table for the genealogy and species tree in Figure 1. The probability of coalescence is constant within each discrete interval and is scaled by the number of lineages \((k)\), the effective population size \((n)\), and the interval length \((d)\). Under the MS-SMC, the probability that a detached lineage will re-coalesce on a specific branch of the genealogy (e.g., branch 7 from Figure 1) is calculated using the piecewise constant probabilities from each discrete interval spanning that branch (e.g., rows 1, 6, 7, and 8).

| interval | start \((\sigma)\) | end \((\mu)\) | pop | \(N_e\) \((n)\) | \(N\) edges \((k)\) | coal | length \((d)\) | branches \((b)\) |
|----------|---------------------|-------------|-----|----------------------|-------------------|-------|--------------|------------------|
| 0        | 0                   | \(t_7\)     | A   | \(N_A\)              | 3                 | \(t_7\) | \(t_7 - 0\) | 0,1,2             |
| 1        | \(t_7\)             | \(W_{AB}\)  | A   | \(N_A\)              | 2                 | -     | \(W_{AB} - t_7\) | 2,7              |
| 2        | 0                   | \(t_8\)     | B   | \(N_B\)              | 2                 | \(t_8\) | \(t_8 - 0\) | 3,4              |
| 3        | \(t_8\)             | \(W_{AB}\)  | B   | \(N_B\)              | 1                 | -     | \(W_{AB} - t_8\) | 8               |
| 4        | 0                   | \(W_{ABC}\) | C   | \(N_C\)              | 1                 | -     | \(W_{ABC}\) | 5               |
| 5        | 0                   | \(W_{ABCD}\)| D   | \(N_D\)              | 1                 | -     | -            |                  |
| 6        | \(W_{AB}\)          | \(t_9\)     | AB  | \(N_{AB}\)           | 3                 | \(t_9\) | \(t_9 - W_{AB}\) | 2,7,8            |
| 7        | \(t_9\)             | \(W_{ABC}\)| AB  | \(N_{AB}\)           | 2                 | -     | \(W_{ABC} - t_9\) | 7,9              |
| 8        | \(W_{ABC}\)         | \(t_{10}\)  | ABC | \(N_{ABC}\)          | 3                 | \(t_{10}\) | \(t_{10} - W_{ABC}\) | 5,7,9            |
| 9        | \(t_{10}\)          | \(W_{ABCD}\)| ABC | \(N_{ABC}\)          | 2                 | -     | \(W_{ABCD} - t_{10}\) | 5,10             |
| 10       | \(W_{ABCD}\)        | \(t_{11}\)  | ABCD| \(N_{ABCD}\)         | 3                 | \(t_{11}\) | \(t_{11} - W_{ABCD}\) | 5,6,10           |
| 11       | \(t_{11}\)          | \(t_{12}\)  | ABCD| \(N_{ABCD}\)         | 2                 | \(t_{12}\) | \(t_{12} - t_{11}\) | 10,11            |
| 12       | \(t_{12}\)          | -           | ABCD| \(N_{ABCD}\)         | 1                 | -     | -            | 12               |
that are spanned by a specific branch of an embedded genealogy. As an example, consider branch 7 from Fig. 1. This branch spans four intervals, labeled as rows 1, 6, 7, and 8 in Table 1, and so \( I_7 = \{1, 6, 7, 8\} \). As we will demonstrate below, this and related indexing variables will be used to calculate the probabilities of re-coalescence in different intervals, and on different branches, based on their rates of coalescence. A summary of all variables in our notation is available in table S1.

### 2.4 Deriving probabilities of genealogy changes in the MS-SMC

A recombination event occurring on \( G \) can result in three types of outcomes (Fig. 2). Of these, there is a zero-sum relationship between a no-change and tree-change event, such that one or the other must occur. Therefore, as a first step towards describing probability statements for each of these event types, we focus first on deriving the probability of a no-change event (also termed a tree-unchanged event; Fig. 3a), which is the simplest outcome. Then, from the law of total probability, we also have a result for the probability of recombination resulting in a tree-change event. Finally, to calculate the probability of a topology-change event, we first derive a statement for the probability of a topology-unchanged event (Fig. 3d), which is the union of a no-change event and a subset of tree-change events, where the detached lineage is restricted according to which ancestral lineages it can re-coalesce with. Full detailed derivations of all solutions below are available in the Appendix of the Supplementary Materials.

#### 2.4.1 Probability of a no-change event

We begin by assuming knowledge of when and where recombination takes place, in terms of a recombination event bisecting branch \( b \) at time \( t_r \). For no change to occur to the genealogy, the detached subtree must re-coalesce with its original branch – either in the same interval from which it detached, or in a later interval on the same branch (Fig. 3a). If it connects to any other lineage, this will cause a change to either the tree or topology. Equation 3 describes the probability of a no-change event given a genealogy embedded in a species tree and given the timing and branch on which recombination occurs. The interval in which recombination occurs is labeled \( i \). The first two terms describe the probability that the subtree re-coalesces during interval \( i \) on branch \( b \) (i.e., \( ii \)), while the latter term is the probability of re-coalescing during a later interval on branch \( b \) (i.e., \( ij \)). For this, we define another indexing variable \( J_b(i) = \{j \in I_b \mid j > i\} \), for iterating over the ordered intervals above \( i \) on branch \( b \) (e.g., Fig. 3b).

\[
\Pr(\text{no-change} \mid S, G, b, t_r) = \frac{1}{k_i} + f(i, i) \exp \left\{ \frac{k_i}{2n_i} t_r \right\} + \sum_{j \in J_b(i)} f(i, j) \exp \left\{ \frac{k_i}{2n_i} t_r \right\}
\]  

(3)

This equation is simplified by use of the function \( f(i, j) \) to return the piece-wise constant probabilities of re-coalescence between pairs of intervals. When \( j = i \), this expression involves the probability of coalescing over the remaining length of interval \( i \) above \( t_r \); when \( j > i \) it involves the probability of coalescing in interval \( j \) and not coalescing in interval \( i \) or any other intervals between \( i \) and \( j \). For this latter process, we define another indexing variable, \( Q_b(i, j) = \{q \in I_b \mid j > q > i\} \), for iterating over the ordered intervals above \( i \) and below \( j \) on branch \( b \) (e.g., Fig. 3c). The function \( f(i, j) \) forms the core of the MS-SMC algorithm and will reappear in several later equations.
Figure 3. The probabilities of different recombination event types are calculated using piece-wise constant probabilities of recombination within discrete intervals. (a) The probability of a no-change event involves re-coalescing with the same branch on which recombination occurred, for which intervals are indexed using the variable $I_b$. (b) Indexing variables for calculating tree-change probabilities. The indexing variable $I_b$ records all intervals on branch $b$, and the recombination event for this example occurs in interval $i = 0$. $J_b(i)$ returns all intervals in $I_b$ above interval $i$. (c,f) The function $f(i, j)$ returns the probability that a subtree that detached in interval $i$ will re-coalesce in interval $j$. This involves excluding the probability of re-coalescence in intervals between $i$ and $j$, which are indexed as $Q_b(i, j)$. (d) The probability of a topology-unchanged event involves re-coalescing with the same branch on which recombination occurred, or with its parent or sibling branches. (e) Additional indexing variables return intervals on the parent branch ($I_c$), or branch $b$ above ($M_b$) or below ($L_b$) the time ($t_b^m$) at which it overlaps in the same species tree interval as its sibling branch.

By integrating equation 3 across all times at which recombination could have occurred on branch $b$ (assuming a uniform recombination rate through time) we obtain the probability that recombination anywhere on this branch does not change the tree:

$$P(\text{no-change}|S, G, b) = \frac{1}{t_b^u - t_b^l} \sum_{i \in I_b} \left\{ \frac{1}{k_i} d_i + \frac{2n_i}{k_i} \sum_{j \in I_b} f(i, j) \left( \exp \left\{ \frac{k_i}{2n_i} \mu_i \right\} - \exp \left\{ \frac{k_i}{2n_i} \sigma_i \right\} \right) \right\}$$

Finally, by summing across all branches on the tree while weighting each one by its relative proportion
of edge length, we get the probability that a recombination event occurring anywhere on $G$ will result in a
no-change event.

$$\mathbb{P}(\text{no-change}|S, G) = \sum_{b \in G} \left[ \frac{t_b^u - t_b}{L(G)} \right] \mathbb{P}(\text{no-change}|S, G, b) \quad (6)$$

2.4.2 Waiting distances to no-change and tree-change events

Under the SMC', recombination is modeled as a Poisson point process such that the time between recombi-
nation events is exponentially distributed with rate parameter $\lambda_r$: the product of the per-site per-generation
recombination rate and summed branch lengths of the current genealogy (Wiuf & Hein, 1999) (equation
7). The likelihood of an observed distance ($x$) between recombination events spatially along the genome,
in units of base pairs, can thus be calculated from the exponential probability density function (equation 8).

$$\lambda_r = L(G) \times r \quad (7)$$

$$f(x; \lambda) = \lambda e^{-\lambda x} \quad (8)$$

Having derived the probability that an individual recombination event is of the no-change type, we can
now calculate the rate of no-change type events as a proportion of the rate of all types of recombination
events. Here, waiting distances continue to be exponentially distributed. However, the new rate parameter
$\lambda_n$ is reduced proportionally by the probability that recombination causes no change to the genealogy
(equation 9). Similarly, because a tree-change event is the opposite of a no-change event, its probability is
one minus the probability of no-change (equation 10). This yields rate parameter $\lambda_g$ for the exponential
probability distribution of waiting distances between tree-change events.

$$\lambda_n = L(G) \times r \times \mathbb{P}(\text{no-change}|S, G) \quad (9)$$

$$\lambda_g = L(G) \times r \times (1 - \mathbb{P}(\text{no-change}|S, G)) \quad (10)$$

2.4.3 Probability of topology-change

We next derive an analogous probability distribution for waiting distances between topology-change events.
Similar to our approach for calculating tree-change probabilities as the opposite of those for a no-change
(tree-unchanged) event, here we calculate topology-change probabilities as the opposite of those for a
topology-unchanged event. Topology-unchanged events represent the union of all no-change events and
the subset of possible tree-change events that only affect branch lengths but not the topology. Our ap-
proach for calculating these probabilities follows closely to that of Deng et al. (2021). In order to isolate
re-coalescence events that do not change the topology, we must take into account which specific branches
the detached subtree from branch $b$ re-coalesces with. The relevant branches are its source ($b$), its sibling
($b'$), and its parent ($c$) (Fig. 3d). If the subtree re-coalesces with $b$, no change occurs; if it re-coalesces with
$b'$, the topology remains the same but a coalescent time is shortened; and if it re-coalesces with $c$, the topol-
ogy remains the same but a coalescent time is lengthened. A re-coalescence with any other branch will change the topology.

To index over relevant intervals across the three branches on which re-coalescence can occur, we define several additional variables. The lowest time point in which both $b$ and $b'$ are present and exist within the same species tree interval is labeled $t^m_b$. For a branch $b$ with intervals $\mathcal{I}_b$, the subset of intervals below $t^m_b$ is $\mathcal{L}_b$, and the subset above is $\mathcal{M}_b$. The union of the sets of intervals on branches $b$ and $c$ is $\mathcal{I}_{bc}$ (Fig. 3e).

Once again, we begin by assuming knowledge of the branch on which a recombination event occurs and of that event’s timing. This problem can be broken into two distinct cases: when $t_r$ occurs below $t^m_b$, and when it occurs above $t^m_b$. The former requires integrating over additional intervals on $b$ that are not shared with $b'$. Over all intervals on the three relevant branches, these equations use the function $f(i, j)$ to return the piecewise constant probabilities where recombination occurs in interval $i$ and re-coalesces in interval $j$ (Fig. 3f).

\[
\mathbb{P}(\text{topology-unchanged}|S, G, b, t_r) = \\
\begin{cases} 
\frac{1}{k_i} + \sum_{j \in \mathcal{L}_{bc}} f(i, j) \exp \left( \frac{k_i}{2n_i} t_r \right) + \sum_{j \in \mathcal{M}_b} f(i, j) \exp \left( \frac{k_i}{2n_i} t_r \right), & \text{if } t_r < t^m_b \\
2 \left( \frac{1}{k_i} + \sum_{j \in \mathcal{L}_b} f(i, j) \exp \left( \frac{k_i}{2n_i} t_r \right) \right) + \sum_{j \in \mathcal{I}_c} f(i, j) \exp \left( \frac{k_i}{2n_i} t_r \right), & \text{if } t_r \geq t^m_b 
\end{cases}
\]  

(11)

Next, the probability of a topology-unchanged event given recombination anywhere on a branch can be derived by integrating the previous equation over the entire length of a branch. Here, the terms $p_{b,1}$ and $p_{b,2}$ correspond to recombination occurring on branch $b$ during a time that falls into either of the two cases in equation 11.

\[
\mathbb{P}(\text{topology-unchanged}|S, G, b) = \frac{1}{t^m_b - t^l_b} \left[ \sum_{i \in \mathcal{L}_b} p_{b,1}^{(i)} + \sum_{i \in \mathcal{M}_b} p_{b,2}^{(i)} \right]
\]

where:

\[
p_{b,1}^{(i)} = \frac{1}{k_i} \left[ d_i + 2n_i \left( \exp \left( \frac{k_i}{2n_i} \mu_i \right) - \exp \left( \frac{k_i}{2n_i} \sigma_i \right) \right) \right] \left( \sum_{j \in \mathcal{L}_{bc}} f(i, j) + \sum_{j \in \mathcal{M}_b} f(i, j) \right)
\]

and:

\[
p_{b,2}^{(i)} = \frac{1}{k_i} \left[ 2d_i + 2n_i \left( \exp \left( \frac{k_i}{2n_i} \mu_i \right) - \exp \left( \frac{k_i}{2n_i} \sigma_i \right) \right) \right] \left( \sum_{j \in \mathcal{L}_b} f(i, j) + \sum_{j \in \mathcal{I}_c} f(i, j) \right)
\]

(12)

Finally, by summing equation 12 across all branches on a genealogy while weighting each by its proportion of summed branch lengths, we get the probability that a recombination event falling uniformly on the genealogy will result in a topology-unchanged event.
\[ \mathbb{P}(\text{topology-unchanged}|S, G) = \sum_{b \in G} \frac{t_b^{\text{un}} - t_b^{\text{t}}} {L(G)} \times \mathbb{P}(\text{topology-unchanged}|S, G, b) \]

\[ = \frac{1}{L(G)} \sum_{b \in G} \left[ \sum_{i \in L_b} p_{b,1}^{(i)} + \sum_{i \in M_b} p_{b,2}^{(i)} \right] \]  

(13)

### 2.4.4 Waiting distance to topology-change events

A recombination event either does or does not change the topology of a genealogy, and we can therefore get the probability of a topology-change event using our topology-unchanged probability statement. As with the previous waiting distance distributions, the distance between topology-change events given a parameterized MSC model can be modeled as an exponential probability distribution. Similar to how a rate parameter was derived for the distribution of waiting distances until a recombination event (equation 7), no-change event (equation 9), or tree-change event (equation 10), a rate parameter \( \lambda_t \) can be calculated from equation 13 for the probability of a topology-change event.

\[ \lambda_t = L(G) \times r \times (1 - \mathbb{P}(\text{topology-unchanged}|S, G)) \]  

(14)

Unlike the exact solution for the expected waiting distance to a tree-change, the waiting distance for a topology-change is an approximation. This is because topology-change probabilities are not guaranteed to be homogeneous across some distance of the genome between topology-change events, since intermediate tree-change events could occur (e.g., the second and third recombination events in Fig. 2). We examine this and other potential sources of bias in our validations below and in the Supplementary Materials.

### 3 Results

#### 3.1 Implementation

We have implemented our solutions for waiting distance calculations under the MS-SMC in the Python package ipcoal (McKenzie & Eaton, 2020a). This software includes functions that accept a parameterized MSC model and initial genealogy as input and return the probabilities of different recombination event types. These probabilities can be calculated for specific branches and times, for entire branches, or for entire genealogies. Functions are also available to calculate expected waiting distances and the likelihoods of observed waiting distances in an ARG using an exponential probability density parameterized by the MS-SMC. Our implementation is built upon toytree (Eaton, 2020), scipy (Virtanen et al., 2020), and numpy (Harris et al., 2020), and it uses jit-compilation with numba (Lam et al., 2015). Below we use ipcoal (v.0.4.0) to demonstrate the impact of MSC model parameters on waiting distances and to validate our solutions against expectations from coalescent simulations implemented in msprime (v.1.1.1) (Baumdicker et al., 2022). Source code is available at https://github.com/eaton-lab/ipcoal. Jupyter notebooks demonstrating the MS-SMC calculations and with reproducible code used for validations in this study are available at https://github.com/eaton-lab/waiting-distances.
3.2 Demonstration

Given a parameterized MSC model and initial genealogy, the probabilities of different types of recombination outcomes can be calculated and visualized as a function of when and where recombination occurs. This is demonstrated on an imbalanced 4-tip species tree with constant effective population size and with a genealogy of seven samples embedded, including three from lineage A, two from lineage B, and one from each of lineages C and D (Fig. 1a; Fig. S2). The probabilities of no-change, tree-change, or topology-change events, given a recombination event occurring on a branch at a particular time (equations 3 and 11, respectively) are shown for three selected branches on the example genealogy (Fig. 1b-d). Note that the probability of no-change and tree-change events are inversely related and sum to 1, since one or the other must occur at any recombination event. By contrast, a topology-change event is a subset of the probability of a tree-change; it is a tree-change event where the detached branch re-coalesces with a branch other than itself, its sibling, or its parent.

In general, the probability of a no-change event decreases, and the probability of a tree-change event increases, as recombination occurs closer to the top end of a branch (further back in time). This makes intuitive sense, since when recombination occurs at the top of a branch there is less time for it to re-coalesce with its same branch. Although this is a general trend, these probabilities do not behave monotonically along the length of a branch as they would in a single-population model with constant $N_e$ (Deng et al., 2021). Instead, probabilities increase or decrease through the length of each interval as a function of the rates of coalescence in subsequent intervals and the probability that a detached lineage will re-coalesce in one of those intervals.

For example, consider branch 2, which exhibits an increase in the probability of no-change through its first branch interval from time 0 to $t_7$ but then a decrease through the next interval from $t_7$ to $W_{AB}$ (Fig. 1b). The observed increase through the first interval is influenced by the fact that a re-coalescence in the subsequent interval is more likely to cause a no-change event, since that interval contains only two samples instead of three. By contrast, within the second interval, recombination events near the top are approaching the next species tree divergence event. After that event, the number of samples will increase back from 2 to 3, thus decreasing the probability of a no-change event. This visualization demonstrates how the probabilities of different recombination event types represent an integration over all the positions on a branch where recombination could occur, and all positions at or above each of these points (whether on the same or different available branches) where a detached subtree could re-coalesce.

Branch 7 provides a clear example for examining the probabilities of tree- and topology-change events. Of particular interest is the interval from $W_{AB}$ to $W_{ABC}$ where these probabilities diverge significantly (Fig. 1c-d). The probability of topology-change decreases faster than the probability of tree-change as recombination occurs closer to node $t_9$. This is because following $t_9$ there is a large stretch of time during which re-coalescence can only occur with the same branch or its sibling, neither of which can cause a topology-change event. It is only after $W_{abc}$ that it is once again possible for re-coalescence to occur with a more distant branch that would result in topology-change. If the effective population size of this species tree interval (AB) were greater, then the probability of re-coalescence in a deeper interval would be more likely, and the probability of topology-change would decrease less severely near $t_9$. This is true more generally, as can be seen by comparing edge probabilities across MSC models with different effective population sizes (Fig. S2). Effective population size affects the rate of re-coalescence and thus either
Figure 4. MS-SMC predictions validated against coalescent simulations. (a) Results are shown for three models containing 1, 2, or 8 populations. For each model, 10K tree sequences were simulated for 10 different constant \( N_e \) values between 50K and 500K. (b) The distributions of summed edge lengths of the first genealogy in each tree sequence. (c-d) The mean frequency (black square) with which the first observed recombination event was a tree-change (c) or topology-change (d) in a simulated tree sequence, and the mean (colored circle) and 95% CI (grey fill) of the predicted probability of tree or topology-change calculated from the first embedded genealogy in each tree sequence. Probabilities are constant with respect to \( N_e \) in the single population model but vary in models with population structure (also shown with respect to species tree interval lengths in coalescent units, across the top axis). (e-f) The mean waiting distance (black square) until the first observed tree-change (e) or topology-change (f) in a simulated tree sequence, and the mean (colored circle) and 95% CI (grey fill) of predicted waiting distances calculated using the first embedded genealogy in each tree sequence.

3.3 Validation

To validate our analytical solutions for the probabilities of different recombination event outcomes and their associated waiting distances, we compared predictions of the MS-SMC with results from stochastic coalescent simulations. We set up three scenarios with increasing amounts of population structure: a single-population model, a two-population model, and an 8-tip phylogeny model (Fig. 4a). All analyses used a constant per-site per-generation recombination rate of 2e-9 and simulated tree sequences using the coalescent with recombination (i.e., the "hudson" ancestry model in msprime as opposed to the "smc_prime" model, which is an approximation), and the argument "record_full_arg=True" (to retain records of invisible recombination events). For each model we simulated genealogies for the same total number of samples (8, unless specified), divided evenly among lineages when models include multiple populations (Fig. 4a).

The exponential rate parameter \( \lambda \) for a probability distribution of waiting distances is a product of the per-site per-generation recombination rate \( r \), the sum of edge lengths on the current genealogy smooths probabilities across intervals when \( N_e \) is high or accentuates differences among intervals when \( N_e \) is low.
Across the three models examined, $r$ remains constant, but both $L(G)$ and $P$ can vary due to population structure, where the effect of structure is scaled by $N_e$. Therefore, we examined $L(G)$, $P$, and the expected waiting distance calculated from their product, for each demographic model across a range of $N_e$ values (Fig. 4b-f). For each value of $N_e$, 10K tree sequences were simulated that included at least one topology-change event. The empirical probabilities of tree-change and topology-change events were then simply measured as the mean observed frequency at which the first recombination event in each tree sequence was a tree or topology change. Similarly, empirical waiting distances were measured as the mean distance until the first observation of each event type. These simulated expectations were then compared with analytical expectations computed under the MS-SMC, where probabilities and waiting distances were based only on the embedding of the first genealogy from each tree sequence into an MSC model parameterized to the specific constant $N_e$ value.

Population structure enforces a lower limit on the length of coalescent times by requiring that genealogies can be embedded in a species tree. This has the effect of shifting both the minimum and mean of $L(G)$ higher (Fig. 4b). Because the per-generation recombination rate interacts with $L(G)$ (the opportunity over which recombination can occur) to determine the frequency of recombination, larger $L(G)$ induced by population structure will tend to decrease waiting distances between recombination events, all else being equal. However, all else does not remain equal. Population structure has a simultaneous opposing effect on waiting distances by decreasing the probability of tree or topology changes ($P$; Fig. 4c-d), especially at low $N_e$ values, where species tree constraints can make tree or topology changes unlikely to occur. This is a stark difference between the MS-SMC framework and a single population model: the probability of a tree- or topology-change event is strongly associated with $N_e$ in the former but not affected at all in the latter. Consequently, the waiting distances between each event type in MSC models (Fig. 4e-f) represent a balance of the positive and negative effects of population structure on $L(G)$ and $P$, respectively.

Our analytical predictions under the MS-SMC converge accurately on the mean results from stochastic coalescent simulations (Fig. 4c-f). Moreover, by examining the variance in these predictions with respect to MSC model parameters we further gain insights into the information contained in spatial genealogical patterns. For example, in the single population model there is high variance in both the probabilities of tree and topology changes, as well as in genealogy lengths, at any given $N_e$ value. Consequently, waiting distances also exhibit high variance. Although waiting distances correlate with population $N_e$ in this model, the differences in mean waiting distances are small relative to the variance. By contrast, multi-species models exhibit much less variance in predicted probabilities of tree or topology changes given a set of MSC model parameters (Fig. 4c-d) and also exhibit less variation in genealogy lengths. This leads to a stronger relationship between MSC model parameters and expected waiting distances (Fig. 4e-f), such that models with different parameters have less overlap in waiting distance expectations. Overall, this suggests that tree- and topology-change distances may contain more information in MSC models than in single population models.

### 3.4 Bias in waiting distance estimation

Estimated waiting distances under the MS-SMC harbor two potential sources of bias, the first stemming from assumptions of the SMC’ approximation, and the second from the approximate nature of waiting
Figure 5. MS-SMC likelihood framework. (a) ARGs were simulated under a two-population species tree model with a constant $N_e=200K$ and $r=2e-9$. (b) A joint log-likelihood surface for $N_e$ and $r$ inferred from the distances between tree-change events; (c) topology-change events; or (d) both. The true parameters are marked by an X. (e) If the MSC model is misspecified as a single-population model but the data derive from a two-population model, likelihood inference is highly biased.

distance estimation for topology-change events. We examined both of these sources of error through comparison to stochastic simulations and found that their effects are generally small, and that multispecies models exhibit a similar magnitude of error as single population coalescent models (Supplementary Materials; Fig. S3, Fig. S4). The SMC’ approximation contributes the greatest error in all models, leading to a slight under-estimation of waiting distances. This is most observable when variance in waiting distances is high, as with topology-change waiting distances in models with low $N_e$. Future work could seek to find a correction for this bias.

3.5 MS-SMC likelihood inference framework

The MS-SMC provides a statistical framework for predicting waiting distances until genealogy changes in an ARG given a parameterized MSC model. Because waiting distances between recombination events are exponentially distributed, we can use the exponential probability density (equation 8) as a likelihood function to describe the joint probability of observed waiting distances as a function of MSC model parameters. Applications of this approach could potentially provide improvements to ARG or species tree inference methods. Here, to explore the potential of this approach, we developed a maximum likelihood framework for the latter: we apply the MS-SMC to fit MSC model parameters to observed distances between tree-change and/or topology-change events in simulated ARGs.

Given one or more ARGs, each composing a sequence of genealogies $G = (G_1, G_2, ..., G_n)$ and their interval lengths in the genome $X = (x_1, x_2, ..., x_n)$, a subset of genealogies can be extracted that represent the unique trees between tree-change events, $G_g = (G_1, G_i, ...)$, and waiting distances can be measured as the summed interval lengths between these trees, $X_g = (\sum_{i=1}^{j} x_i, \sum_{i=j}^{k} x_i, ...)$). The same could be done for topology-change events. Given a parameterized species tree $S$ and recombination rate $r$ we can then obtain a sequence of exponential rate parameters $\Lambda_g = (\lambda_1, \lambda_i, ...)$ corresponding to each genealogy in $G_g$ (equations 10 or 14). The likelihood of each observed waiting distance in $X_g$ can then be calculated from the exponential probability density function (equation 8) with rate parameters from $\Lambda_g$. The maximum likelihood solution can be found by searching for the set of MSC model parameters – which affect $\Lambda_g$ – that maximize the summed log-likelihood of the observed waiting distances.

We first implemented this approach for a simple two-population model with constant $N_e=200K$ and a population divergence at 500K generations. We simulated 100 independent ARGs, each 100Kb in length,
using $r=2e^{-9}$, and we sampled four haplotypes per population. This yielded 51,487 tree-change events, with mean length 194bp, and 25,446 topology-change events, with mean length 393bp. To examine the likelihood surface, we calculated the log-likelihood of joint parameters for $N_e$ and $r$, while keeping a fixed divergence time parameter, over a grid search of 41 evenly spaced values from $r=1e^{-9}$ to $3e^{-9}$ and $N_e=50K$ to 800K. The likelihood surface for tree-change distances exhibited a ridge containing the true parameter values (Fig. 5b), while the topology-change likelihood surface exhibited a distinct peak at the true values (Fig. 5c). The summed log-likelihoods from both tree and topology change distances provided the most informative likelihood surface (Fig. 5d). We also inferred a likelihood surface for a misspecified model, in which the genealogies were generated under a 2-population model, but the $\Lambda_g$ rate parameters were calculated for a single population model with constant $N_e$. This model misspecification introduced a significant bias in parameter estimation, shifting the likelihood surface towards lower $r$ and higher $N_e$ (Fig. 5e). This demonstrates that even for a simple two population model, ignoring population structure can significantly bias spatial genealogical inference methods.

To examine parameter inference under more complex MSC models we implemented a Bayesian approach using a Metropolis-Hastings Markov chain Monte Carlo (MCMC) algorithm to compute the joint posterior probability distribution of multiple $N_e$ parameters for a 2-population model with variable $N_e$ of 200K, 300K, or 400K (Fig. 6a). The input ARGs were simulated under this more complex MSC model but otherwise used the same settings as in the example above. We set a uniform prior on all three $N_e$ parameters between $1e^2$ and $1e^7$, and ran a single MCMC chain to sample 10K values, sampling every 5th iteration following a 1K iteration burnin. For simplicity, we fixed the population divergence time and recombination rate to their correct values to focus solely on $N_e$ estimation. All parameters converged with ESS values exceeding 2000, and the 95% highest posterior density intervals (HPDI) for each posterior probability distribution included the true parameter (Fig. 6b). This suggests that multiple MSC model parameters can be accurately inferred from waiting distance information alone.

### 4 Discussion

Genealogical relationships vary spatially across chromosomes, reflecting a history of recombination between genome segments inherited from different ancestors. Such variation can be modeled by the sequentially Markov coalescent, which provides a generative process upon which many statistical methods have been developed (McVean & Cardin, 2005; Spence et al., 2018). However, most applications of the SMC’
remain highly limited with regard to the scale over which they extract information from genomes—extending forward just one recombination event at a time. By contrast, the recent development of solutions for predicting tree- and topology-change waiting distances under the SMC’ (Deng et al., 2021) effectively adds two additional, longer-range sources of information for any position in a genome. In their model, these distances are a result of the probabilities of different phenomenological outcomes of the SMC’ process given a genealogy embedded in a single population coalescent model with constant $N_e$. Here we extended this framework, deriving new solutions for the probabilities of tree-change and topology-change events for a genealogy embedded in any arbitrarily parameterized structured coalescent model. These solutions lay a groundwork for exploring how variation in species tree parameters affects neutral expectations of genealogical heterogeneity across chromosomes.

4.1 Applications of the MS-SMC

Our MS-SMC approach provides a predictive model for the relationship between a parameterized species tree and the length of a genomic interval over which a genealogy is expected to be observed. We demonstrated the accuracy of our implementation relative to stochastic simulations performed under both the full coalescent with recombination and the SMC’ approximation (Hudson, 1983; McVean & Cardin, 2005; Wiuf & Hein, 1999). Within our framework, MSC model parameters determine probabilities of coalescence in species tree intervals and prevent coalescence between lineages that are separated by species divergence events. The effects of MSC model parameters on spatial genealogical variation can be statistically modeled and visualized (Fig. 1; Fig. S2), and our examples show that the neutral rate of turnover in genealogical trees and topologies across a genome is highly dependent on species tree model parameters.

A complex relationship exists between a parameterized MSC model, the distribution of genealogies that can arise under that model, and the spatial distances over which those genealogies are expected to span. Previously, such patterns could only be examined through stochastic simulations. For example, McKenzie & Eaton (2020b) used exhaustive simulations to examine the effect of species tree parameters on genetic linkage by varying species tree length, size, and shape. While this approach is practical for estimating the mean linkage across a large set of sampled genealogies under a specific demographic model, it is impractical for estimating the persistence of individual genealogies. The analytical solutions presented here not only enable fast calculations but also the opportunity to develop likelihood-based methods for fitting models that link MSC parameters to ARGs.

Many extensions of the MS-SMC are possible, including applications to more complex demographic models and the development of specific hypothesis testing frameworks. For example, genealogies could instead be embedded in a species network following the multispecies network coalescent model (Wen et al., 2016). This would involve modeling the coalescent histories as a probabilistic process that could follow one of multiple possible paths through a network. Differences between the expected waiting distances in genomic regions derived from one embedding path history versus another may be highly informative about the network model in which genealogies are embedded. Therefore, incorporating linked genealogical information may prove valuable towards resolving intractable problems in species tree or network inference by moving beyond the limited information available from the frequencies of unlinked gene trees.
4.2 Phylogenetic and local ancestry inference

By linking species tree parameters to observable patterns in ARGs, the MS-SMC can be used to improve both gene tree and species tree inference methods. One practical application of waiting distance expectations calculated under the MS-SMC is to guide the selection of appropriate locus lengths for MSC analyses so as to avoid intra-locus recombination. This process yields data within a locus that derive from multiple variable genealogies, and so inferred gene trees represent concatenation artifacts. Topology-change waiting distances are often much shorter than sampled locus lengths for multispecies datasets (McKenzie & Eaton, 2020b), and this is especially likely when gene trees are inferred from large genomic sliding windows (e.g., Li et al., 2019). When repeated across many loci, concatenation artifacts can cause the distribution of gene trees, or of their summary statistics, to deviate from expectations under the MSC – a process termed concatalescence (Gatesy & Springer, 2013).

One way to improve gene tree inference for individual loci is to re-conceive the problem as one in which loci can have multiple gene tree histories and are thus better represented as ARGs. Applications of ARG inference have historically been applied at the scale of entire chromosomes, and usually to samples that are modeled as being in a single population (Kelleher et al., 2019; Rasmussen et al., 2014; Speidel et al., 2019). However, ARGs can exist for any stretch of the genome, including short aligned contigs, and recent methods can infer ARGs conditional on a structured demographic model that includes splits between more distantly related genomes (Hubisz & Siepel, 2020). Although our example likelihood-based implementation of the MS-SMC showed that MSC model parameters can be estimated from an observed ARG (Fig. 5, Fig. 6), we expect that many useful applications of the MS-SMC will likely come from the inverse application: inferring ARGs given a parameterized species tree model. Or, from joint inference of both MSC models and ARGs. For example, although ARGWeaver-D (Hubisz & Siepel, 2020) can currently propose an ARG conditional on a structured demographic model, which can generate genealogies with tree- and topology-change waiting distances as predicted under the SMC’, these distances do not contribute to the likelihood score of an ARG. We speculate that incorporating this information could improve both accuracy and convergence.

In contrast to current phylogenomic inference approaches which tend to either ignore genetic linkage, or to discard the vast majority of sequenced data in effort to avoid it, one could envision an alternative, spatially aware phylogenetic inference framework that more effectively utilizes linked genomic data. This would mark a major transition in phylogenetics, where recombination could be viewed as a source of information rather than a source of error. We see the MS-SMC as an important step in this direction. For example, a joint ARG and species tree inference approach might build upon the species tree inference method SNAPP (Bryant et al., 2012). SNAPP subsamples a single unlinked SNP from each locus and integrates over a distribution of genealogies that could have produced it, with the goal of bypassing the problem of gene tree inference. A theoretical ARG-based extension of this approach might analyze many linked SNPs and integrate over a distribution of ARGs that could have produced each linked pattern, with the goal of bypassing ARG inference. By reducing the complexity of ARG inference to a problem of windows between topology-change events, as opposed to any two recombination events, the MS-SMC provides a more efficient framework that could potentially make methods linking ARGs and species tree inference possible.

Finally, a common goal of many local ancestry inference methods is to test evolutionary hypotheses
involving selection or introgression (Martin & Belleghem, 2017). For example, sliding window analyses have been used to identify putatively introgressed loci in *Heliconius* butterflies (Zhang et al., 2016) and felids (Li et al., 2019). Such conclusions are often based on the genomic interval lengths over which certain topological gene tree patterns are observed. We suggest that such conclusions should be examined critically when made without reference to a null-model-based expectation. Our results show that neutral expectations for waiting distances between genealogy changes can exhibit high variance, and that this variance will be spatially auto-correlated, such that a topology or clade may exist over long stretches of the genome simply by chance.

4.3 Acknowledgements

This work was supported by the National Science Foundation (NSF DEB-2046813 awarded to D.A.R.E. and NSF Graduate Research Fellowship DGE 16-44869 awarded to P.F.M.). Thanks to Yun Deng for discussion on waiting distance methods, and to members of the Eaton Lab for valuable feedback.

References

Baum, D.A. (2007). Concordance Trees, Concordance Factors, and the Exploration of Reticulate Genealogy. *Taxon*, 56, 417–426.

Baumdicker, F., Bisschop, G., Goldstein, D., Gower, G., Ragsdale, A.P., Tsambos, G., Zhu, S., Eldon, B., Ellerman, E.C., Galloway, J.G., Gladstein, A.L., Gorjanc, G., Guo, B., Jeffery, B., Kretzschumar, W.W., Lohse, K., Matschiner, M., Nelson, D., Pope, N.S., Quinto-Cortés, C.D., Rodrigues, M.F., Saunack, K., Sellinger, T., Thornton, K., van Kemenade, H., Wohns, A.W., Wong, Y., Gravel, S., Kern, A.D., Koskela, J., Ralph, P.L. & Kelleher, J. (2022). Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*, 220, iyab229.

Brandt, D., Wei, X., Deng, Y., Vaughn, A.H. & Nielsen, R. (2022). Evaluation of methods for estimating coalescence times using ancestral recombination graphs. *Genetics*, 221, iyac044.

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N.A. & RoyChoudhury, A. (2012). Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Molecular biology and evolution*, 29, 1917–1932.

Degnan, J.H. & Rosenberg, N.A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in ecology & evolution*, 24, 332–340.

Deng, Y., Song, Y.S. & Nielsen, R. (2021). The distribution of waiting distances in ancestral recombination graphs. *Theoretical Population Biology*, 141, 34–43.

Eaton, D.A.R. (2020). Toytree: A minimalist tree visualization and manipulation library for Python. *Methods in Ecology and Evolution*, 11, 187–191.

Gatesy, J. & Springer, M.S. (2013). Concatenation versus coalescence versus “concatalescence”. *Proceedings of the National Academy of Sciences*, 110, E1179–E1179. Publisher: Proceedings of the National Academy of Sciences.
Griffiths, R. & Marjoram, P. (1996). An ancestral recombination graph. In: Progress in population genetics and human evolution. Springer-Verlag, Berlin, pp. 257–270.

Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C. & Oliphant, T.E. (2020). Array programming with NumPy. Nature, 585, 357–362.

Hubisz, M. & Siepel, A. (2020). Inference of ancestral recombination graphs using argweaver. In: Statistical Population Genomics. Humana, New York, NY, pp. 231–266.

Hudson, R.R. (1983). Properties of a neutral allele model with intragenic recombination. Theoretical population biology, 23, 183–201.

Kelleher, J., Etheridge, A.M. & McVean, G. (2016). Efficient coalescent simulation and genealogical analysis for large sample sizes. PLoS computational biology, 12, e1004842.

Kelleher, J., Wong, Y., Wohns, A.W., Fadil, C., Albers, P.K. & McVean, G. (2019). Inferring whole-genome histories in large population datasets. Nature genetics, 51, 1330–1338.

Kingman, J.F.C. (1982). The coalescent. Stochastic processes and their applications, 13, 235–248.

Knowles, L.L. & Kubatko, L.S. (2011). Estimating Species Trees: Practical and Theoretical Aspects. John Wiley and Sons.

Lam, S.K., Pitrou, A. & Seibert, S. (2015). Numba: A llvm-based python jit compiler. In: Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC. pp. 1–6.

Li, G., Figueiró, H.V., Eizirik, E. & Murphy, W.J. (2019). Recombination-aware phylogenomics reveals the structured genomic landscape of hybridizing cat species. Molecular biology and evolution, 36, 2111–2126.

Li, H. & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. Nature, 475, 493–496.

Maddison, W.P. (1997). Gene trees in species trees. Systematic biology, 46, 523–536.

Maddison, W.P. & Knowles, L.L. (2006). Inferring phylogeny despite incomplete lineage sorting. Systematic biology, 55, 21–30.

Marjoram, P. & Wall, J.D. (2006). Fast" coalescent" simulation. BMC genetics, 7, 1–9.

Martin, S.H. & Belleghem, S.M.V. (2017). Exploring Evolutionary Relationships Across the Genome Using Topology Weighting. Genetics, 206, 429–438.

McKenzie, P.F. & Eaton, D.A.R. (2020a). ipcoal: an interactive Python package for simulating and analyzing genealogies and sequences on a species tree or network. Bioinformatics.
McKenzie, P.F. & Eaton, D.A.R. (2020b). The Multispecies Coalescent in Space and Time. *bioRxiv*, p. 2020.08.02.233395. 17, 18

McVean, G.A. & Cardin, N.J. (2005). Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360, 1387–1393. 3, 16, 17

Rannala, B. & Yang, Z. (2003). Bayes estimation of species divergence times and ancestral population sizes using dna sequences from multiple loci. *Genetics*, 164, 1645–1656. 1, 5

Rasmussen, M.D., Hubisz, M.J., Gronau, I. & Siepel, A. (2014). Genome-wide inference of ancestral recombination graphs. *PLoS genetics*, 10, e1004342. 3, 18

Schiffels, S. & Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, 46, 919–925. Number: 8 Publisher: Nature Publishing Group. 3

Speidel, L., Forest, M., Shi, S. & Myers, S.R. (2019). A method for genome-wide genealogy estimation for thousands of samples. *Nature Genetics*, 51, 1321–1329. 18

Spence, J.P., Steinrücken, M., Terhorst, J. & Song, Y.S. (2018). Inference of population history using coalescent HMMs: review and outlook. *Current Opinion in Genetics & Development*, 53, 70–76. 3, 16

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P. & SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17, 261–272. 11

Wen, D., Yu, Y. & Nakhleh, L. (2016). Bayesian Inference of Reticulate Phylogenies under the Multispecies Network Coalescent. *PLOS Genetics*, 12, e1006006. Publisher: Public Library of Science. 17

Wilton, P.R., Carmi, S. & Hobolth, A. (2015). The smc’ is a highly accurate approximation to the ancestral recombination graph. *Genetics*, 200, 343–355. 3

Wiuf, C. & Hein, J. (1999). Recombination as a Point Process along Sequences. *Theoretical Population Biology*, 55, 248–259. 2, 9, 17

Zhang, W., Dasmahapatra, K.K., Mallet, J., Moreira, G.R. & Kronforst, M.R. (2016). Genome-wide introgression among distantly related heliconius butterfly species. *Genome biology*, 17, 1–15. 19
5 Supplementary Information

5.1 Supplementary Figures
Figure S1. Four categories of outcomes from a recombination event occurring on a genealogy at time $t_1$, dictated by random subtree re-coalescence with a remaining lineage under the SMC' process at time $t_2$. (a) The detached subtree re-coalesces with the original lineage from which it was detached, leading to no change between the starting genealogy and subsequent genealogy. (b) The detached subtree re-coalesces with its sibling lineage prior to their previous coalescence, leading to a shortening of their coalescence time. (c) The detached subtree re-coalesces with its parent lineage, leading to a lengthening of the coalescent time between the detached subtree lineage and its sibling lineage. (d) The detached subtree re-coalesces with a lineage other than itself, its sibling, or its parent lineage, leading to a topology-change.

Figure S2. Probabilities of different recombination event outcomes for a selected genealogy edge as a function of the time at which recombination occurs and of the constant effective population size. (a) An MSC model with edge lengths in units of generations and an example genealogy embedded. (b) An genealogy embedding table for the example MSC model and genealogy. (c) Probabilities of different recombination event outcomes across genealogy edge 7. When $N_e$ is low, probabilities are nearly constant with respect to time within each interval since re-coalescence in later intervals is unlikely. When $N_e$ is high, probabilities change nearly monotonically across the length of an edge since population structure does little to constrain the time of re-coalescence.
5.2 Investigating bias in MS-SMC predictions

The MS-SMC harbors two potential sources of bias, the first stemming from assumptions of the SMC’ approximation and the second from the approximate nature of waiting distance estimation for topology-change events. We examined both of these sources of error through comparison to stochastic simulations and found that their effects are generally negligible and act in opposing directions such that their combined effect is further reduced.

The SMC’ approximation to the full coalescent with recombination is expected to deviate increasingly far from the full model as the number of recombination events that the SMC’ does not model (i.e., those that are among ancestors who do not contribute genetic material to sampled descendants) increases. By not modeling this subset of recombination events, the SMC’ will tend to exhibit greater waiting distances between tree or topology change events than the full model. To investigate this source of error in our waiting distance predictions, we repeated the simulations from our validation scenario using the msprime setting ancestry="smc_prime" to simulate tree sequences while excluding recombination events that would not occur in the SMC’ model. We have already seen from our validations that the error in our predictions is quite low when data are simulated under the full coalescent with recombination (Fig. S3). As expected, we observe even less error between our predictions and coalescent simulations when data are simulated under the SMC’ model. Whereas the error rate is generally below 5% when data are simulated under the full coalescent with recombination model (only exceeding this at the lowest $N_e$ values examined) mean error rates do not exceed 5% in any models for data simulated under the SMC’ assumptions. This shows that the SMC’ assumption does contribute a relatively small error to our waiting distance predictions, especially at low $N_e$ values, where it can lead to over-estimated waiting distances.

We also investigated whether inhomogeneity in the probability of topology changes during the waiting distance between topology-change events causes bias. For this, we employed a similar approach to Deng et al. (2021), who examined the fold-difference in parameters affecting waiting distance estimations between a starting tree and a subsequent genealogy that experienced a tree-change but not a topology-change. Because MSC model parameters affect both the length of genealogies and the probability of topology-change, we also examined variation in each of these parameters at different constant $N_e$ values of 50K, 100K, or 500K. For each setting we examined one topology-change event from 1K tree sequences.

In a single population model with constant $N_e$, Deng et al. (2021) previously showed that the bias in topology-change waiting distances is negligible because of an inverse relationship between the length of a genealogy and the probability of a topology change. Our analysis confirms this result, showing that variance in the product of these two parameters, which equates to the waiting distance rate parameter (equations 7, 8, 10), is very low (Fig. S4a) and becomes smaller as more gene copies are sampled (Fig. S4b-c).

In the 2-population and 8-population MSC-type models we find the same result. Here, constraints imposed by population structure lead to less variance in both the length of genealogies and probabilities of topology-change (Fig. S4d-i). When $N_e$ is low, there is very little variation between subsequent genealogies, and thus the waiting distance expectation exhibits little heterogeneity. When $N_e$ is high, there is little population structure, and so the waiting distance expectation remains relatively constant for the same reason as in a single population model. Overall, MSC models do not appear to exhibit a greater bias from this source than single population models.
**Figure S3.** Error in MS-SMC waiting distance expectations caused by the SMC’ approximation. Error was measured as the mean percent difference between expected waiting distances calculated under the MS-SMC and observed waiting distances in stochastic coalescent simulations. Simulations were performed under either the SMC’ approximation (black) or the full coalescent with recombination (green). The MS-SMC tends to under-estimate waiting distances compared to the full coalescent with recombination but shows only a slight bias at very low $N_e$ values compared to simulations under the SMC’.
**Figure S4.** Variance in the fold-change for components affecting the expected waiting distance to a topology-change event between the starting tree and a subsequent tree which has experienced a tree-change event, changing the coalescent times but not the topology. The sum of genealogical edge lengths ($L(G)$), the $P$(topology-change $|$ S,G), and the product of these two terms are shown for three different demographic models and with different constant $N_e$ values, and/or numbers of samples per lineage. When the fold-change in the product exhibits low variance around 1 the MS-SMC approximation for the expected waiting distance until a topology-change is expected to be more accurate. Larger effective population sizes and numbers of samples per lineage yield lower variance in the product.
6 Appendix: Derivations

6.1 Notation

Information from the genealogy embedding table (described in the following paragraph) can be used in equations that calculate the probabilities of no-change, tree-change, and topology-change events under the MS-SMC. These equations, described throughout rest of the Appendix, use the terms defined in Table S1.

A parameterized species tree, $S$, is a multispecies coalescent model in which a set of isolated populations are related by a bifurcating tree topology. Divergence times between lineages are in units of generations, and each edge (species tree interval) can be associated with a different constant diploid effective population size ($N_e$). A genealogy, $G$, represents the genealogical relationships – composing a topology and coalescent times in units of generations – for a set of sampled gene copies at some position in their genomes. A genealogy can be embedded in a species tree if the coalescent times between sampled gene copies from different populations are not younger than a population divergence event separating them.

Given a genealogy embedded in a species tree, a series of discrete time intervals can be defined that are delimited by events that change the rate of coalescence. We refer to this set of discrete time intervals and their associated properties as a genealogy embedding table (e.g., Table S1). In the waiting distance solutions for a single population with constant $N_e$ by Deng et al. (2021), this table is delimited only by coalescent

---

**Table S1.** Summary of variables used in waiting distance equations.

| Variable | Description |
|----------|-------------|
| $S$      | An MSC model with topology, divergence times and effective population sizes. |
| $G$      | A genealogy that can be embedded in $S$. |
| $L(G)$   | Sum of edge lengths of genealogy $G$. |
| $b$      | A focal branch in $G$. |
| $i$      | Interval in the genealogy embedding table in which recombination occurs. |
| $I_b$    | Ordered set of intervals on branch $b$. |
| $I_c$    | Ordered set of intervals on branch $c$, the parent of branch $b$. |
| $I_{bc}$ | Ordered union of sets $I_b$ and $I_c$. |
| $J_b(i)$ | Ordered set of intervals above $i$ on branch $b$. |
| $Q_b(i, j)$ | Ordered set of intervals above $i$ and below $j$ on branch $b$. |
| $K(b, t)$ | Number of edges of $G$ in the interval containing branch $b$ at time $t$. |
| $k_x$    | Number of edges of $G$ in interval $x$; piece-wise constant of $A(b, t)$. |
| $N(b, t)$ | Diploid effective population size in the interval containing branch $b$ at time $t$. |
| $n_x$    | Diploid effective population size in interval $x$; piece-wise constant of $N(b, t)$. |
| $t_r$    | Time of a recombination event, in generations. |
| $\sigma_x$ | The lower boundary of interval $x$, in generations. |
| $\mu_x$ | The upper boundary of interval $x$, in generations. |
| $d_x$    | The length of interval $x$, in generations. |
| $t'_b$   | The lower boundary of branch $b$, in generations. |
| $t''_b$  | The upper boundary of branch $b$, in generations. |
| $t'_m_0$ | The time at which a focal branch $b$ is able to coalesce with its sibling branch. |
| $M_b$    | Ordered set of intervals above $t''_m$ on branch $b$. |
| $L_b$    | Ordered set of intervals below $t''_m$ on branch $b$. |
events, and the intervals are non-overlapping. Because $N_e$ is constant in their framework, only $k$ differs between intervals. Therefore, changes in $k$ alone determine differences in rates of coalescence, with $k$ decreasing monotonically in subsequent intervals from the tips towards the root. Our approach is similar, but adds additional complexity (Fig. 1). In the multispecies framework, genealogy embedding intervals are specific to each species tree branch, with each one corresponding to a time interval with a constant $k$ and $N_e$ in a specific species tree branch. Breakpoints between intervals arise where divergences occur in the species tree (increasing $k$ and potentially changing $N_e$) and where coalescent events occur in the genealogy (reducing $k$). Genealogy embedding intervals corresponding to different species tree branches can overlap in time.

Each branch on $G$ will span one or more genealogy embedding intervals. The ordered set of intervals on a specific branch, $b$, is defined as $I_b$. The lower and upper time bounding each interval is $\sigma_x$ and $\mu_x$, respectively, where $x$ is the index of the interval in the genealogy embedding table. The lower and upper bounds of each branch are defined as $t^l_b$ and $t^u_b$, respectively.

6.2 Extending SMC’ waiting distance solutions:

The probabilities of different recombination event types under the MS-SMC are calculated from the probability that recombination occurs on a specific branch and the probabilities that the resulting detached subtree subsequently re-coalesces with any other available branch above that time. The opportunity for recombination to occur on a branch is scaled by its length in generations ($t^u_b - t^l_b$). Similarly, the probability of re-coalescence on a branch is scaled by its length and the coalescence rate. The latter can vary over the length of a branch as it spans different intervals, and is a function of the effective population size in the species tree interval that includes branch $b$ at a specified time, $\tau$, defined as $N(b, \tau)$, and the number of other genealogy branches in the interval that includes branch $b$ at time $\tau$, defined as $K(b, \tau)$. Finally, the probability that coalescence occurs over an interval of length ($t$) can be calculated from an exponential probability density $f(t; \lambda)$, where the rate parameter is $\lambda = \frac{K(b, \tau)}{2N(b, \tau)}$, similar to equations 1-2.

6.3 Probability of no-change

6.3.1 Given a branch and time of recombination

The probability that a tree is unchanged by a recombination event – meaning that no coalescent times are changed – is the probability that the detached subtree re-coalesces with the same branch it detached from. Thus, we can integrate from the time of recombination ($t_r$) to the top of the branch ($t^u_b$) over the probability of sampling the same branch times the exponential probability density of re-coalescing at any time on that branch above the time of recombination ($\tau - t_r$). We take this integral with respect to $\tau$, where $K(b, \tau)$ and $N(b, \tau)$ can vary across the length of the branch if it spans different intervals.

$$
\mathbb{P}({\text{tree-unchanged}|S, G, b, t_r}) = \int_{t_r}^{t^u_b} \frac{1}{K(b, \tau)} f(\tau - t_r; \lambda) d\tau
$$

The exponential probability density function can be expanded, as in equation 2, where $\lambda$ is $K(b, \tau)$ over $2N(b, \tau)$:
We similarly define the function \( f \) which recombination occurred. By contrast, the latter parts of this equation are performed over the entire set of all intervals above \( i \). Because the first term is computed over only part of the first interval we first solve this term separately, and then show the result for the later terms. The first term concerns the probability of re-coalescing in the same interval \( i \) as labeled as \( b \) (Recall, because we are currently computing the probability of a no-change event we only need to concern ourselves with re-coalescence on branch \( b \).) Here, the interval in which recombination occurred on branch \( b \) is labeled as \( i \).

In the equation below, the first integral describes the probability statement over only part of interval \( i \), from \( t_r \) to \( t_u \), rather than over its entire length, since re-coalescence can only occur above the time at which recombination occurred. By contrast, the latter parts of this equation are performed over the entire lengths of each remaining interval above \( i \), from the bottom (\( \sigma_j \)) to the top (\( \mu_j \)) of the interval. The ordered set of all intervals above \( i \) in \( \mathcal{I}_b \) is defined as \( \mathcal{J}_b(i) = \{ j \in \mathcal{I}_b \mid j > i \} \).

\[
= \int_{t_r}^{t_u} \frac{1}{2N(b, \tau)} \exp \left\{ - \int_{t_r}^{\tau} \frac{K(b, s)}{2N(b, s)} ds \right\} d\tau
\]

This simplifies to the following equation, which describes the probability that the subtree re-coalesces at any time \( (d\tau) \) on \( b \) above \( t_r \), and that it does not re-coalesce at any intervening time \( (ds) \) between \( t_r \) and \( \tau \).

\[
= \int_{t_r}^{t_u} \frac{1}{2N(b, \tau)} \exp \left\{ - \int_{t_r}^{\tau} \frac{K(b, s)}{2N(b, s)} ds \right\} d\tau
\]

Because the rate of re-coalescence is constant within each interval, we next split this equation into statements over each discrete interval that the detached subtree could possibly re-coalesce with on branch \( b \).

Because the first term is computed over only part of interval \( i \), rather than over its entire length, since re-coalescence can only occur above the time at which recombination occurred. By contrast, the latter parts of this equation are performed over the entire lengths of each remaining interval above \( i \), from the bottom (\( \sigma_j \)) to the top (\( \mu_j \)) of the interval. The ordered set of all intervals above \( i \) in \( \mathcal{I}_b \) is defined as \( \mathcal{J}_b(i) = \{ j \in \mathcal{I}_b \mid j > i \} \).

\[
= \int_{t_r}^{\mu_i} \frac{1}{2N(b, \tau)} \exp \left\{ - \int_{t_r}^{\tau} \frac{K(b, s)}{2N(b, s)} ds \right\} d\tau + \sum_{j \in \mathcal{J}_b(i)} \int_{\sigma_j}^{\mu_j} \frac{1}{2N(b, \tau)} \exp \left\{ - \int_{t_r}^{\tau} \frac{K(s)}{2N(s)} ds \right\} d\tau
\]

We can now solve this equation and substitute constant values for \( K(b, \tau) \) and \( N(b, \tau) \) in each interval. Because the first term is computed over only part of the first interval we first solve this term separately, and then show the result for the later terms. The first term concerns the probability of re-coalescing in the same interval \( i \) in which recombination occurred. The center part of this equation will appear again later, and so we define it as the function \( f(i, i) \).

\[
= \frac{1}{k_i} - \frac{1}{k_i} \exp \left\{ - \frac{k_i}{2n_i} \mu_i \right\} \exp \left\{ \frac{k_i}{2n_i} t_r \right\}
\]

\[
= \frac{1}{k_i} + f(i, i) \exp \left\{ \frac{k_i}{2n_i} t_r \right\}
\]

We similarly define the function \( f(i, j) \) for the later terms in this equation, which refer to the probability of re-coalescing in a later interval, \( j \), than the one in which recombination occurred, \( i \). This function requires also summing over any intervening intervals, \( q \). For this, we define the function \( \mathcal{Q}_b(i, j) = \{ q \in \mathcal{I}_b \mid j > q > i \} \) to return the ordered set of intervals on branch \( b \) between \( i \) and \( j \):

\[
= \sum_{j \in \mathcal{J}_b(i)} \frac{1}{k_j} \left( 1 - \exp \left\{ - \frac{k_j}{2n_j} d_j \right\} \right) \exp \left\{ - \frac{k_i}{2n_i} \mu_i - \sum_{q \in \mathcal{Q}_b(i, j)} \frac{k_q}{2n_q} d_q \right\} \exp \left\{ \frac{k_i}{2n_i} t_r \right\}
\]
\[ = \sum_{j \in \mathcal{J}_b} f(i, j) \exp \left\{ \frac{k_i}{2n_i} t_r \right\} \]  

(22)

The \( f(i, i) \) and \( f(i, j) \) function above is represented in the main text as equation 4. Adding the two terms that include these functions together we get equation 3 from the main text for the probability of a no-change event given the timing and branch on which recombination occurs:

\[
\mathbb{P}(\text{no-change}|S, G, b, t_r) = \frac{1}{k_i} + f(i, i) \exp \left\{ \frac{k_i}{2n_i} t_r \right\} + \sum_{j \in \mathcal{J}_b(i)} f(i, j) \exp \left\{ \frac{k_i}{2n_i} t_r \right\} 
\]

(3)

Note that in equation 3 above, while we have split the terms for clarity, the \( f(i, i) \) term could be lumped into the summation. The summation would then be indexed using \( j \in \mathcal{I}_b \), with the understanding that \( f(i, j) = 0 \) when \( i > j \) – that is, when summing over intervals that fall below the time of recombination.

6.3.2 Across a full branch

Having solved for the probability of the genealogy being unchanged given the time \( t_r \) of the recombination event, our next step is to integrate this equation across the entire branch with respect to \( t_r \):

\[
\mathbb{P}(\text{tree-unchanged}|S, G, b) = \frac{1}{t_u - t_l} \int_{t_l}^{t_u} \mathbb{P}(\text{tree-unchanged}|S, G, b, t_r) dt_r 
\]

(23)

Plugging in the piece-wise constant solutions from equation 3 we get the following solution.

\[
= \frac{1}{t_u - t_l} \sum_{i \in \mathcal{I}_b} \frac{1}{k_i} d_i + \left( \exp \left\{ \frac{k_i}{2n_i} \mu_i \right\} - \exp \left\{ \frac{k_i}{2n_i} \sigma_i \right\} \right) 
\]

\[
\left[ -\frac{2n_i}{k_i} \exp \left\{ -\frac{k_i}{2n_i} \mu_i \right\} + \frac{2n_i}{k_i} \left( \sum_{j \in \mathcal{J}_b(i)} \exp \left\{ -\frac{k_i}{2n_i} \mu_i - \sum_{q \in \mathcal{Q}_b(i,j)} \frac{k_q}{2n_q} d_q \right\} \right) \exp \left\{ -\frac{k_i}{2n_i} \sigma_i \right\} \right] 
\]

(24)

Finally, this can be simplified to the following solution by expressing the piecewise constant re-coalescence rates using the function \( f(i, j) \), as shown in equation 5 from the main text.

\[
\mathbb{P}(\text{tree-unchanged}|S, G, b) = \frac{1}{t_u - t_l} \int_{t_l}^{t_u} \mathbb{P}(\text{tree-unchanged}|S, G, b, t) dt 
\]

\[
= \frac{1}{t_u - t_l} \sum_{i \in \mathcal{I}_b} \left[ \frac{1}{k_i} d_i + \left( \frac{2n_i}{k_i} \sum_{j \in \mathcal{I}_b} f(i, j) \left( \exp \left\{ \frac{k_i}{2n_i} \mu_i \right\} - \exp \left\{ \frac{k_i}{2n_i} \sigma_i \right\} \right) \right) \right] 
\]

(5)

6.3.3 Across the whole tree

At last, we can calculate the probability that, given a recombination event, the genealogy is unchanged. We do this by weighting each branch by its proportion of the total tree length and summing across the
Figure S5. Calculating the probability that recombination on genealogy branch \( b \) leads to a topology change involves summing over the probabilities that the detached lineage does not re-coalesce with either itself, its sibling, or its parent \( (b, b' \text{ or } c) \), respectively). The possibility of a tree-change outcome (e.g., shortened coalescent time) is restricted until the lowest shared interval between \( b \) and \( b' \), designated at time \( t_m \). Opportunities for such events could be constrained by species divergences – as in (a) and (d) – or by the timing of prior coalescence events generating each branch – as in (b), (c), and (e). The possibility that recombination \( (t_r) \) occurs prior to \( t_m \) leads to the two ordered sets of intervals used in equations 11 and 12.

unchanging probabilities for all branches. This is equation 6 from the main text:

\[
P(\text{tree-unchanged}|S, G) = \sum_{b \in G} \left[ \frac{t_u^b - t_l^b}{L(G)} \right] P(\text{tree-unchanged}|S, G, b) \quad (6)
\]

We can then also derive the probability of a tree-change event as \( 1 - P(\text{tree-unchanged}|S, G) \). Following the approach described in equations 7-10, we can then calculate an exponential probability distribution for waiting distances between no-change or tree-change events using an exponential rate parameter that is scaled by the probabilities of either recombination event type.

### 6.4 Probability of topology change

Next, we derive the probability of a topology-unchanged event, from which we can get the associated probability of a topology-change event. As in the first section, we first derive a solution given an individual branch and time of recombination. We then extend this solution to an entire branch, and we finally sum across branches to get a probability for the entire genealogy. To isolate events that do not cause a topology change, we must find the union of events that cause a no-change event in addition to two types of possible tree-change events which affect only branch lengths but not the topology. These two types of events correspond to a re-coalescence with the sibling to branch \( b \), termed \( b' \), or with its parent, \( c \) (Fig. 3d). Because branch \( c \) is always ancestral to \( b \), a recombination event on \( b \) can potentially re-coalesce anywhere on \( c \); however, this is not the case for \( b' \), which may only exist or be available for re-coalescence over part of the length of \( b \). It is therefore important to define the lowest time point at which re-coalescence with \( b' \) is possible, termed \( t_m^b \).

In the single population model of Deng et al. (2021), \( t_m^b \) occurs at the maximum value of \( t_u^b \) and \( t_l^b \), representing the lower bounds of \( b \) and \( b' \), respectively. In our MSC-based model we must also incorporate potential constraints imposed by population barriers, and so \( t_m^b \) occurs at the maximum of \( t_u^b \), \( t_l^b \), and any population divergence events that separate \( b \) and \( b' \) (Fig. S5a-c).
6.4.1 Given a branch and time of recombination

Using the definition for $t_{mn}^b$ from above, we can describe the probability of a topology-unchanged event given the branch and time of recombination as a two-part solution. These two parts correspond to scenarios in which the time of recombination, $t_r$, occurs either above (Fig. S5a-c) or below (Fig. S5d-e) $t_{mn}^b$. When $t_r \leq t_{mn}^b$, there are only two distinct intervals over which re-coalescence can occur: from $t_r$ to $t_{mn}^b$ on branches $b$ or $b'$, and from $t_{mn}^b$ to $t_{mn}^c$ on branch $c$ (Fig. S5a-c). By contrast, when $t_r > t_{mn}^b$ there are three distinct intervals for re-coalescence: from $t_r$ to $t_{mn}^b$ on $b$, $t_{mn}^b$ to $t_{mn}^c$ on $b'$, and $t_{mn}^c$ to $t_{mn}^c$ on branch $c$ (Fig. S5d-e). Thus, in the first scenario the opportunity for re-coalescence is the same for branches $b$ and $b'$, whereas in the latter scenario it is different.

Retaining the correct order of intervals is important in these calculations, particularly for $f(i, j)$, which involves summing over not only the information in the $i$ and $j$ intervals, but also all of the intervals that lie between them. To iterate over ordered intervals on each branch we define additional indexing variables.

Just as $I_b$ defines the ordered set of intervals on branch $b$, $I_c$ is the ordered set of intervals on branch $c$, and $I_{bc}$ is the ordered union of these sets. In addition, we define $M_b$ as the ordered intervals on branch $b$ above $t_{mn}^b$, and $L_b$ as the ordered intervals on branch $b$ below $t_{mn}^b$. We can now derive a probability for the two distinct scenarios:

First case – Given $t_r < t_{mn}^b$, we integrate over the three distinct intervals where re-coalescence can occur. The first is unique to branch $b$, the second integral is multiplied by two since from $t_{mn}^b$ to $t_{mn}^c$ re-coalescence can occur with $b$ or $b'$, and the final integral is over the length of branch $c$. By substituting the piece-wise constant solutions for each interval into this equation it can simplified to the final form below, also shown in equation 11 of the main text:

$$\begin{align*}
\mathbb{P}(\text{topology-unchanged}|S, G, b, t_r) & = \int_{t_r}^{t_{mn}^b} \frac{1}{K(b, \tau)} f(\tau - t_r; \lambda) d\tau + \int_{t_{mn}^b}^{t_{mn}^c} \frac{1}{K(b, \tau)} f(\tau - t_r; \lambda) d\tau + \int_{t_{mn}^c}^{t_{mn}^c} \frac{1}{K(b, \tau)} f(\tau - t_r; \lambda) d\tau \\
& = \frac{1}{k_i} + \sum_{j \in I_{bc}} f(i, j) \exp \left\{ \frac{k_i}{2n_i} t_r \right\} + \sum_{j \in M_b} f(i, j) \exp \left\{ \frac{k_i}{2n_i} t_r \right\} \\
& \quad + \sum_{j \in I_c} f(i, j) \exp \left\{ \frac{k_i}{2n_i} t_r \right\}
\end{align*}$$

Second case – Given $t_r \geq t_{mn}^b$, we only need to integrate over two distinct intervals, thus we simply drop the first term from the equation above. The final form of this equation is also shown in equation 11 of the main text:

$$\begin{align*}
\mathbb{P}(\text{topology-unchanged}|S, G, b, t_r) & = \int_{t_r}^{t_{mn}^b} \frac{1}{K(b, \tau)} f(\tau - t_r; \lambda) d\tau + \int_{t_{mn}^b}^{t_{mn}^c} \frac{1}{K(b, \tau)} f(\tau - t_r; \lambda) d\tau \\
& = 2 \left( \frac{1}{k_i} + \sum_{j \in I_b} f(i, j) \exp \left\{ \frac{k_i}{2n_i} t_r \right\} \right) + \sum_{j \in I_c} f(i, j) \exp \left\{ \frac{k_i}{2n_i} t_r \right\}
\end{align*}$$
6.4.2 Across a full branch

Now we derive the overall probability that a recombination event falling on a specific branch will change the topology by integrating across the range of possible values for \( t_r \). Following the approach above, we split this problem into two parts, above and below \( t^m_b \), and we sum the two cases.

\[
P(\text{topology-unchanged}|S, G, b) = \frac{1}{t^u_b - t^l_b} \int_{t^l_b}^{t^u_b} P(\text{topology-unchanged}|S, G, b, t_r) dt_r
\]

(27)

\[
= \frac{1}{t^u_b - t^l_b} \left[ \left( \int_{t^l_b}^{t^m_b} + \int_{t^m_b}^{t^u_b} \right) P(\text{topology-unchanged}|S, G, b, t_r) dt_r \right]
\]

(28)

First case – We can simply sum over each entire interval below \( t^m_b \) where \( t_r \) could occur, and substitute piece-wise constant solutions for the probabilities that a detached subtree will re-coalesce over the subset of targeted intervals above this that do not cause a topology-change.

\[
\int_{t^l_b}^{t^m_b} P(\text{topology-unchanged}|S, G, b, t_r) dt_r
\]

(29)

Second case – Similarly, we can sum over each entire interval above \( t^m_b \) (up to \( t^u_b \)) and substitute piece-wise constant solutions for the same selected subset of intervals:

\[
\int_{t^m_b}^{t^u_b} P(\text{topology-unchanged}|S, G, b, t_r) dt_r
\]

(30)

Result – If we express the inner summed terms from the equations above, composed of piece-wise constant values from their intervals, as \( p^{(i)}_{b,1} \) and \( p^{(i)}_{b,2} \), respectively, then the final solution can be expressed more concisely. This is shown in the main text as equation 12.

\[
P(\text{topology-unchanged}|S, G, b) = \frac{1}{t^u_b - t^l_b} \left[ \sum_{i \in L_b} p^{(i)}_{b,1} + \sum_{i \in M_b} p^{(i)}_{b,2} \right]
\]

(12)

6.4.3 Across the whole tree

Finally, we sum across all branches, each weighted by their relative length, to find the probability of a recombination event changing the topology of the tree. This appears as equation 13 in the main text.
\[ \mathbb{P}(\text{topology-unchanged}|S, G) = \sum_{b \in G} \frac{t_{b}^{u} - t_{b}^{d}}{L(G)} \times \mathbb{P}(\text{topology-unchanged}|S, G, b) \]

\[ = \frac{1}{L(G)} \sum_{b \in G} \left[ \sum_{i \in \mathcal{L}_b} p_{b,1}^{(i)} + \sum_{i \in \mathcal{M}_b} p_{b,2}^{(i)} \right] \]

6.4.4 Examples

Examples showing how to compute the probability of a no-change (tree-unchanged) or topology-change event are shown with didactic step-by-step instructions in Fig. S6 and Fig. S7, respectively.
The probability of no-change given recombination on a given branch at a given time:

\[ P(\text{no-change}|\mathcal{S}, \mathcal{G}, b, t_r) = \frac{1}{k_i} + f(i, i) \exp \left\{ \frac{k_i}{2n_i} t_r \right\} + \sum_{j \in \mathcal{I}_i(i)} f(i, j) \exp \left\{ \frac{k_j}{2n_j} t_r \right\} \]

In this example, recombination occurs on branch $b$ in interval 0 ($i=0$) at time $t_r=500$, and we will assume all $N_e=1000$. We can plug these values into the equation:

\[ P(\text{no-change}|\mathcal{S}, \mathcal{G}, b, t_r) = \frac{1}{1} + f(0, 0) \exp \left\{ \frac{1}{2(1000)} 500 \right\} + \sum_{j \in \{1, 2, 3\}} f(0, j) \exp \left\{ \frac{1}{2(1000)} 500 \right\} \]

Then expand the piecewise constant functions $f(i,j)$ for each interval on $b$:

\[
\begin{align*}
    f(i, i) &= -\frac{1}{k_i} \exp \left\{ -\frac{k_i}{2n_i} \mu_i \right\} \\
    f(0, 0) &= -\frac{1}{1} \exp \left\{ -\frac{1}{2(1000)} 500 \right\} \\
    f(i, j) &= \frac{1}{k_j} \left( 1 - \exp \left\{ -\frac{k_j}{2n_j} d_j \right\} \right) \exp \left\{ -\frac{k_i}{2n_i} \mu_i - \sum_{q \in \mathcal{Q}(i,j)} \frac{k_q}{2n_q} d_q \right\} \\
    f(0, 1) &= \frac{1}{3} \left( 1 - \exp \left\{ -\frac{3}{2(1000)} 1000 \right\} \right) \exp \left\{ -\frac{1}{2(1000)} 1000 \right\} \\
    f(0, 2) &= \frac{1}{2} \left( 1 - \exp \left\{ -\frac{2}{2(1000)} 1000 \right\} \right) \exp \left\{ -\frac{1}{2(1000)} 1000 - \left( \frac{3}{2(1000)} 1000 \right) \right\} \\
    f(0, 3) &= \frac{1}{3} \left( 1 - \exp \left\{ -\frac{3}{2(1000)} 1000 \right\} \right) \exp \left\{ -\frac{1}{2(1000)} 1000 - \left( \frac{3}{2(1000)} 1000 + \frac{2}{2(1000)} 1000 \right) \right\}
\end{align*}\]

And sum to get final result: (colored to correspond with the figure above):

\[
P(\text{no-change}|\mathcal{S}, \mathcal{G}, b, t_r) = 1 + f(0, 0) \times 1.284 + \sum_{j \in \{1, 2, 3\}} f(0, j) \times 1.284
\]

\[
= 1 + (-0.6065 \times 1.284) + (0.1571 \times 1.284) + (0.0428 \times 1.284) + (0.0129 \times 1.284)
\]

\[
= 0.4944
\]

**Figure S6.** A step-by-step calculation of the probability of a tree-unchanged event under the MS-SMC given a species tree and genealogy.
Example: Calculating the probability of topology-unchanged

The probability of topology-unchanged given the branch and timing of recombination:

\[
P(\text{topology-unchanged}|S, G, b, t_r) = \begin{cases} 
\frac{1}{k_i} + \sum_{j\in T_e} f(i, j) \exp \left\{ \frac{k_i}{2n_i} t_r \right\} \sum_{j\in T_i} f(i, j) \exp \left\{ \frac{k_i}{2n_i} t_r \right\}, & \text{if } t_r < t_b^m \\
2 \left( \frac{1}{k_i} + \sum_{j\in T_e} f(i, j) \exp \left\{ \frac{k_i}{2n_i} t_r \right\} \right) + \sum_{j\in T_i} f(i, j) \exp \left\{ \frac{k_i}{2n_i} t_r \right\}, & \text{if } t_r \geq t_b^m
\end{cases}
\]

In this example, recombination occurs on branch \( b \) in interval 0 \((i=0)\) at time \( t_r=500 \), and we will assume all \( N_e=1000 \). Because \( t_r < t_b^m \), we apply the first case above:

\[
P(\text{topology-unchanged}|S, G, b, t_r) = \frac{1}{k_i} + \sum_{j\in T_e} f(i, j) \exp \left\{ \frac{k_i}{2n_i} t_r \right\} \sum_{j\in T_i} f(i, j) \exp \left\{ \frac{k_i}{2n_i} t_r \right\} = \frac{1}{k_i} + \sum_{j\in \{0,1,3,4\}} f(i, j) \exp \left\{ \frac{1}{2(1000)} 500 \right\} + \sum_{j\in \{2\}} f(i, j) \exp \left\{ \frac{1}{2(1000)} 500 \right\}
\]

Expand piece-wise constant functions \( f(i,j) \) for each interval over branches \( b \) and \( c \):

\[
\begin{align*}
 f(0, 0) &= \frac{1}{4} \exp \left\{ -\frac{1}{2(1000)} 1000 \right\} \\
 f(0, 1) &= \frac{1}{3} \left( 1 - \exp \left\{ -\frac{3}{2(1000)} 1000 \right\} \right) \exp \left\{ -\frac{1}{2(1000)} 1000 \right\} \\
 f(0, 2) &= \frac{1}{2} \left( 1 - \exp \left\{ -\frac{2}{2(1000)} 1000 \right\} \right) \exp \left\{ -\frac{1}{2(1000)} 1000 - \left( \frac{3}{2(1000)} 1000 \right) \right\} \\
 f(0, 3) &= \frac{1}{3} \left( 1 - \exp \left\{ -\frac{3}{2(1000)} 1000 \right\} \right) \exp \left\{ -\frac{1}{2(1000)} 1000 - \left( \frac{3}{2(1000)} 1000 + \frac{2}{2(1000)} 1000 \right) \right\} \\
 f(0, 4) &= \frac{1}{2} \left( 1 - \exp \left\{ -\frac{2}{2(1000)} 1000 \right\} \right) \exp \left\{ -\frac{1}{2(1000)} 1000 - \left( \frac{3}{2(1000)} 1000 + \frac{3}{2(1000)} 1000 \right) \right\}
\end{align*}
\]

Yields a final result (colored to correspond with the figure above):

\[
= 1 + (-0.6065 \times 1.284) + (0.1571 \times 1.284) + (0.0428 \times 1.284) + (0.0129 \times 1.284) + (0.0035 \times 1.284) + (0.0028 \times 1.284) + (0.0129 \times 1.284)
= 0.5703
\]

**Figure S7.** A step-by-step calculation of the probability of a topology-unchanged event under the MS-SMC given a species tree and genealogy.