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 developed 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 from 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 among different lineages.

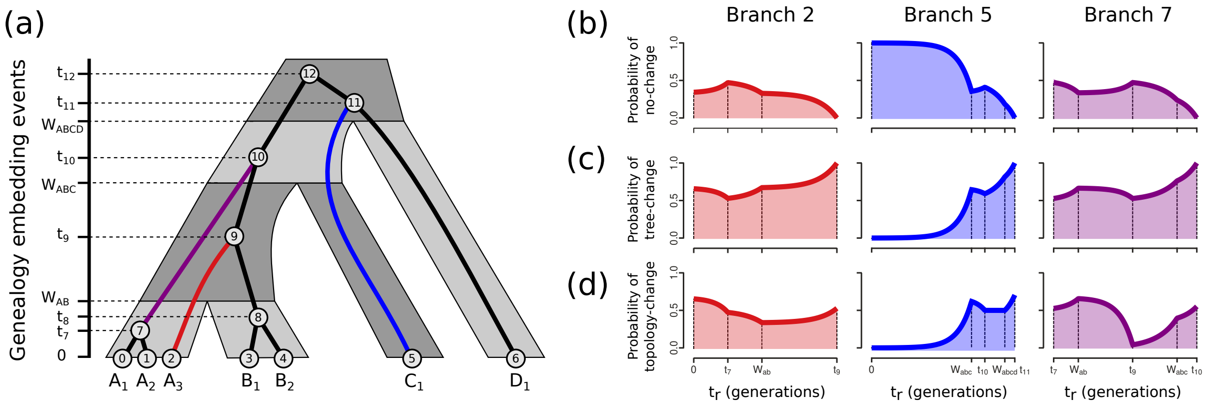


Figure 1. The probability of different outcomes of recombination occurring on a genealogy embedded in a species tree can be modeled using the MS-SMC'. (a) A parameterized MSC model is composed of discrete intervals separated by population divergence events (W), where the rate of coalescence can vary among intervals with different effective population sizes. The expected waiting time between coalescence events (e.g., t_9–t_{10}) is a product of these rates and the number of samples present, such that a table can be constructed of discrete intervals with constant rates (see Table 1). A recombination event can occur on any branch of a genealogy to cause one of three possible outcomes (b-d): "no-change", "tree-change", or "topology-change", and each branch exhibits piecewise variation in these probabilities across discrete intervals. Probabilities were calculated here for three selected branches as a function of the time at which recombination occurs (t_r), given the MSC model in (a) with parameters from Table S2.

Importantly, both the coalescent and MSC are models of the expected distribution of *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 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 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 or 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 ARGweaver (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, where a detached lineage re-attaches with its own ancestral lineage prior to the time of its next coalescent event. This leads to no change between the genealogies in two sequential intervals despite the occurrence of a recombination even 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" (type 1), "tree-change" (types 2-4), and "topology-change" (type 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 different categorical types of events that leave different detectable signatures in the genome, and extend over greater spatial distances.

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. In this respect, 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, 2006; Knowles & Kubatko, 2011). The waiting distance until a topology-change event can 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, since samples are partitioned among species tree intervals. The partitioning of coalescent events among species tree intervals is thus expected to constrain the types of recombination events that will be observed. Consequently, the distributions of

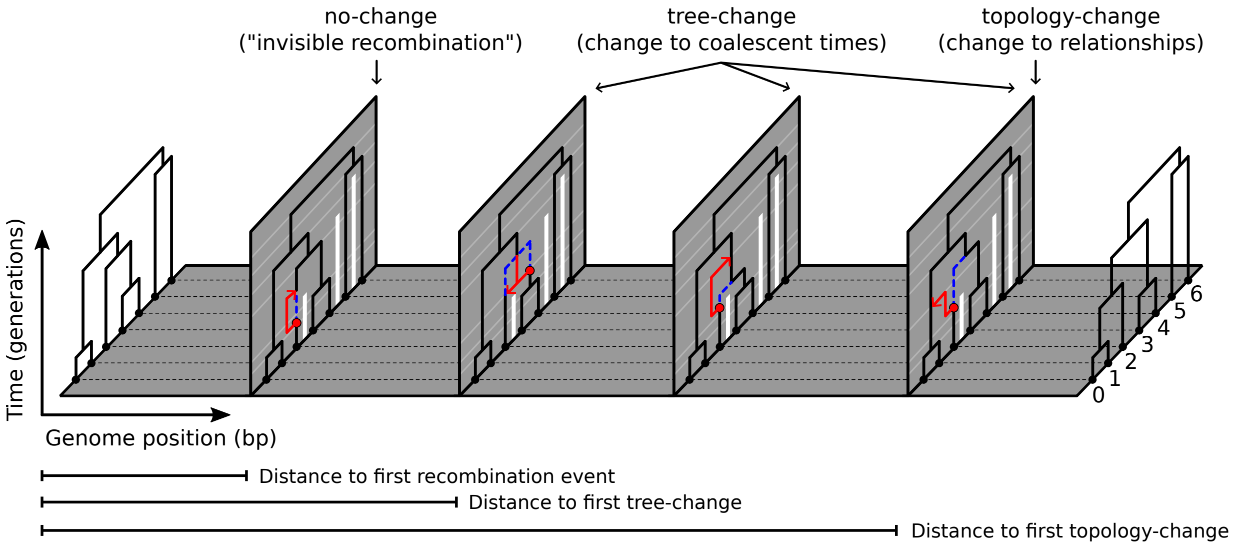

no-change
("invisible recombination")
tree-change
(change to coalescent times)
topology-change
(change to relationships)
Time (generations)
Genome position (bp)
0
1
2
3
4
5
6
Distance to first recombination event
Distance to first tree-change
Distance to first topology-change

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 start and end positions on a small chromosome in the history of a set of 7 samples constrained by a 4-tip species tree model (as in Fig. 1). Recombination events are indicated as vertical panels, showing the process by which a subtree (below the red circle) is detached and then re-coalesces (red arrow) with the remaining ancestral 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 a topology-change is 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.

waiting distances between different types of genealogy changes should be highly dependent on, and thus informative about, the species tree model. We refer to this 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. We have intentionally reproduced our equations in a similar structure and using many of the same variable names as in Deng *et al.* (2021).

2.2 MSC model description

Given an MSC model composed of a species tree topology ($\mathcal{S}$), with divergence times (W) in units of generations, and constant diploid effective population sizes assigned to each branch (N_b), a genealogy ($\mathcal{G}$) 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 that will reduce the number of samples from k to k-1 in a population is given by equation 1. From this, we can model the expected waiting time ($\mathbb{E}[t_k]$) until the next coalescence event as an exponentially distributed random variable with rate parameter λ_k:

$$\begin{aligned} &\lambda_k = \mathbb{P}(\text{coal event } | N_e, k) = \frac{k(k-1)}{2N_e} \\ &\ and: \\ &\mathbb{E}[t_k] = 1/\lambda_k \end{aligned} \tag{1}$$

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_b value and an increase in the number of samples.

Based on this generative framework for sampling genealogies, a set of likelihood solutions have been developed to fit coalescent model parameters, such as N_e in single population models (Kingman, 1982), or N_b and W in MSC models (Rannala & Yang, 2003), based on inferred coalescent 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 (λ_k) to calculate the likelihood of observed waiting times between coalescent events (t_k) in each population interval from an exponential probability density function:

$$f(t_k; \lambda) = \lambda e^{-\lambda t_k} \tag{2}$$

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 $\mathcal{G}$, and the sum of its edge lengths as $L(\mathcal{G})$, then under the assumption of a constant recombination rate through time, a recombination break point can be uniformly sampled from $L(\mathcal{G})$ to occur with equal probability anywhere on $\mathcal{G}$. A recombination event creates a bisection on a branch, separating a subtree below the cut from

Table 1. A genealogy embedding table for MS-SMC' calculations for the genealogy and species tree in Figure 1. The rate of coalescence within each interval can be calculated from a and n, and its length from d. To integrate over the probability of coalescence on a specific genealogy branch (e.g., branch 7 from Figure 1) involves integrating over all intervals that include that branch (e.g., rows 1, 6, 7, and 8).

	start (σ)	end (μ)	pop	N (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	-	W_{ABCD}	6
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

the rest of the genealogy (Fig. 2, Fig. S1). The subtree must then re-coalesce to connect with one of the remaining edges on the genealogy at a time above the recombination event. The waiting time until this re-coalescence event occurs is sampled stochastically with an expectation determined by the number of samples and the coalescent rate, similar to equation 1.

In a single population model with constant N_e the expected waiting time until re-coalescence increases monotonically with each coalescence event backwards in time, since each event decreases k. Once again, the MSC model differs from this: coalescent events similarly decrease k, but the merging of species tree branches into ancestral intervals increases k, and N_b can also vary among species tree intervals. Thus, the probability that a detached subtree re-coalesces to the genealogy can vary through time along its path of possible reconnection points through different species tree intervals (Fig. 1b-d). To calculate these probabilities, a species tree can be decomposed into a series of relevant intervals between events that change rates of coalescence, which we refer to as the genealogy embedding table (Table 1). From this table it is possible to calculate the probabilities of different recombination event type outcomes, and consequently, to model the expected waiting distances until specific recombination event types occur.

To better understand the relationship between the MS-SMC' and the genealogy embedding table, consider a specific branch (b) from Fig. 1. Branch 7 on the genealogy spans four relevant intervals, labeled as rows 1, 6, 7, and 8 in Table 1. This ordered set is defined as $\mathcal{I}_b$, such that, for example, $\mathcal{I}_7$ = {1,6,7,8}. The lower and upper bounds of each interval are defined as σ_i and μ_i, respectively, and its length as d_i. Similarly, the lower and upper bounds of each branch are defined as t_b^l and t_b^u, respectively. Each interval has a constant number of genealogy branches k_i, and effective population size n_i. Some additional indexing variables are also described below, and a summary of all variables is available in table **??**).

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

A recombination event occurring on $\mathcal{G}$ 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 a tree-change event. Finally, to calculate the probability of a topology-change event, we first derive a statement for the probability of a tree-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 in which ancestral lineages it can re-coalesce with. More detailed derivations of the solutions below can be found in the Appendix.

2.4.1 Probability of a no-change event

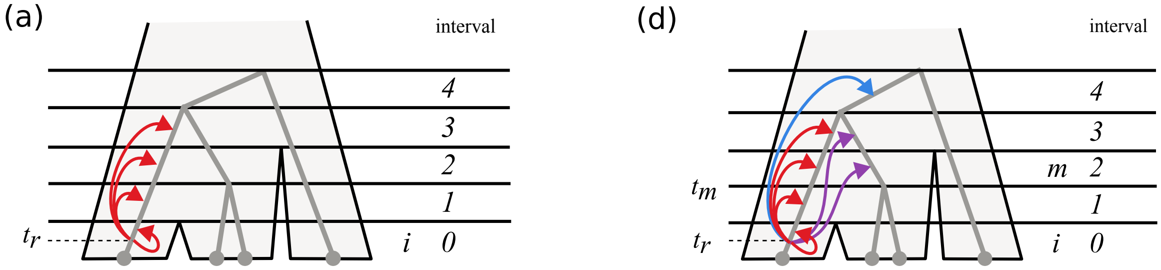
We begin by assuming knowledge of when and where recombination takes place, in terms of a recombination event bisecting branch b at time t_r. For no change to occur to the genealogy the detached subtree must re-coalesce with its original branch, either in the same interval from which it detached, or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

$$\mathbb{P}(\text{no-change}|\mathcal{S}, \mathcal{G}, b, t_r) = \frac{1}{k_i} + f(i,i)\exp\left\{\frac{k_i}{2n_i}t_r\right\} + \sum_{j \in \mathcal{J}_b(i)} f(i,j)\exp\left\{\frac{k_i}{2n_i}t_r\right\} \tag{3}$$

This includes the following function, $f(i,j)$, which returns the piece-wise constant probabilities of re-coalescence between pairs of intervals. When $j = i$, this involves the probability of coalescing over the remaining length of interval i; when $j > i$ it involves the probability of coalescing in interval j and not coalescing in interval i, or in any other intervals between i and j. For this latter term we define an indexing variable $\mathcal{Q}_b(i,j) = \{q \in \mathcal{I}_b \mid j > q > i\}$, for iterating over the ordered intervals above i and below j on branch b (e.g., Fig. 3c). The function $f(i,j)$ will reappear in later equations.

$$f(i,j) = \begin{cases} -\dfrac{1}{k_i}\exp\left\{-\dfrac{k_i}{2n_i}\mu_i\right\}, & \text{if } i = j \\ \dfrac{1}{k_j}\left(1 - \exp\left\{-\dfrac{k_j}{2n_j}d_j\right\}\right)\exp\left\{-\dfrac{k_i}{2n_i}\mu_i - \displaystyle\sum_{q \in \mathcal{Q}_b(i,j)} \dfrac{k_q}{2n_q}d_q\right\}, & \text{if } i \neq j \end{cases} \tag{4}$$

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



(b)

$$\begin{aligned}\mathcal{I}_b &= \{0,1,2,3\}\\ i &= 0\\ \mathcal{J}_b(i) &= \{1,2,3\}\end{aligned}$$

(c)

$f(i,j)$	$\mathcal{Q}_b(i,j)$
$f(0,0)$	$\{\}$
$f(0,1)$	$\{\}$
$f(0,2)$	$\{1\}$
$f(0,3)$	$\{1,2\}$

(e)

$$\begin{aligned}\mathcal{I}_{bc} &= \{0,1,2,3,4\}\\ \mathcal{I}_c &= \{4\}\\ \mathcal{M}_b &= \{2,3\}\\ \mathcal{L}_b &= \{0,1\}\end{aligned}$$

(f)

$f(i,j)$	$\mathcal{Q}_b(i,j)$
$f(0,0)$	$\{\}$
$f(0,1)$	$\{\}$
$f(0,2)$	$\{1\}$
$f(0,3)$	$\{1,2\}$
$f(0,4)$	$\{1,2,3\}$

Figure 3. Rates of coalescence are piece-wise constant within discrete intervals of the genealogy embedding table. The probabilities of different outcomes are calculated from probability statements for recombination occurring at time t_r in interval i and re-coalescing in interval j. The function $f(i,j)$ takes i and j as arguments for a given branch, and iterates over the probability of recombination in i, and not re-coalescing in the intervals between i and j, which is described as $\mathcal{Q}_b(i,j)$.

this branch does not change the tree:

$$\begin{aligned}&\mathbb{P}(\text{no-change}|\mathcal{S},\mathcal{G},b) = \frac{1}{t_b^u - t_b^l}\int_{t_b^l}^{t_b^u} \mathbb{P}(\text{no-change}|\mathcal{S},\mathcal{G},b,t)dt = \\ &\frac{1}{t_b^u - t_b^l}\sum_{i\in\mathcal{I}_b}\frac{1}{k_i}d_i + \left(\frac{2n_i}{k_i}\sum_{j\in\mathcal{J}_b(i)} f(i,j)\left(\exp\left\{\frac{k_i}{2n_i}\mu_i\right\} - \exp\left\{\frac{k_i}{2n_i}\sigma_i\right\}\right)\right)\end{aligned} \tag{5}$$

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

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

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

Under the SMC' recombination events can be modeled as a Poisson point process, where the time between events is exponentially distributed with rate parameter λ_r (equation 7), the product of the per-site per-generation recombination rate and summed branch lengths of the current genealogy (Wiuf & Hein, 1999). 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).

$$\lambda_r = L(\mathcal{G}) \times r \tag{7}$$

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

Because a no-change event is a subset of possible recombination events, and we have derived its probability, we can calculate the rate at which no-change events occur as a proportion of the rate of total recombination events. Here, waiting distances continue to be exponentially distributed, however, the new rate parameter, λ_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), which yields rate parameter λ_g for the exponential probability distribution of waiting distances between tree-change events.

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

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

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 as the opposite of a no-change (tree-unchanged) event, here we calculate topology-change as the opposite of a topology-unchanged event, where topology-unchanged represents the sum of probabilities of a no-change and a subset of possible tree-change events that only affect branch lengths, but not the topology. Our approach follows closely to that of Deng *et al.* (2021). In order to isolate re-coalesce 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 topology 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 interval in which both b and b' are present and exist within the same species tree interval is labeled m, and the lower boundary of this interval as t_b^m. For a branch b with intervals $\mathcal{I}_b$, the subset of intervals below m is as $\mathcal{L}_b$, and the subset including m and 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 its timing. For the latter, we break the problem into two different cases: when t_r occurs below t_b^m, and when it occurs above t_b^m. This equation uses the function $f(i, j)$ to return the piecewise constant probabilities where recombination occurs in interval i and re-coalescence in interval j. As before, this involves iterating over intervening intervals between i and j if present (Fig. 3d,f).

$$\mathbb{P}(\text{topology-unchanged}|\mathcal{S},\mathcal{G},b,t_r) = \\ \begin{cases} \dfrac{1}{k_i} + \displaystyle\sum_{j\in\mathcal{I}_{bc}} f(i,j)\exp\left\{\frac{k_i}{2n_i}t_r\right\} + \sum_{j\in\mathcal{M}_b} f(i,j)\exp\left\{\frac{k_i}{2n_i}t_r\right\}, & \text{if } t_r < t_b^m \\ \\ 2\left(\dfrac{1}{k_i} + \displaystyle\sum_{j\in\mathcal{I}_b} f(i,j)\exp\left\{\frac{k_i}{2n_i}t_r\right\}\right) + \sum_{j\in\mathcal{I}_c} f(i,j)\exp\left\{\frac{k_i}{2n_i}t_r\right\}, & \text{if } t_r \geq t_b^m \end{cases} \tag{11}$$

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 the positions of branch b falling into either of the two cases in equation 11.

$$\begin{aligned} &\mathbb{P}(\text{topology-unchanged}|\mathcal{S},\mathcal{G},b) = \frac{1}{t_b^u - t_b^l}\left[\sum_{i\in\mathcal{L}_b} p_{b,1}^{(i)} + \sum_{i\in\mathcal{M}_b} p_{b,2}^{(i)}\right] \\ &\quad where: \\ &p_{b,1}^{(i)} = \frac{1}{k_i}\left[d_i + 2n_i\left(\exp\left\{\frac{k_i}{2n_i}\mu_i\right\} - \exp\left\{\frac{k_i}{2n_i}\sigma_i\right\}\right)\left(\sum_{j\in\mathcal{I}_{bc}} f(i,j) + \sum_{j\in\mathcal{M}_b} f(i,j)\right)\right] \\ &\quad and: \\ &p_{b,2}^{(i)} = \frac{1}{k_i}\left[2d_i + 2n_i\left(\exp\left\{\frac{k_i}{2n_i}\mu_i\right\} - \exp\left\{\frac{k_i}{2n_i}\sigma_i\right\}\right)\left(2\sum_{j\in\mathcal{I}_b} f(i,j) + \sum_{j\in\mathcal{I}_c} f(i,j)\right)\right] \end{aligned} \tag{12}$$

Finally, by summing equation 12 across all branches on a genealogy, and 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.

$$\begin{aligned} \mathbb{P}(\text{topology-unchanged}|\mathcal{S},\mathcal{G}) = &\sum_{b\in\mathcal{G}} \frac{t_b^u - t_b^l}{L(\mathcal{G})} \times \mathbb{P}(\text{topology-unchanged}|\mathcal{S},\mathcal{G},b) = \\ \frac{1}{L(\mathcal{G})}\sum_{b\in\mathcal{G}}\left[\sum_{i\in\mathcal{L}_b} p_{b,1}^{(i)} + \sum_{i\in\mathcal{M}_b} p_{b,2}^{(i)}\right] & \end{aligned} \tag{13}$$

2.4.4 Waiting distance to topology-change events

A recombination event either does or does not change the topology of a genealogy, and therefore, we can 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, λ_t can be calculated from equation 13 for the probability of a topology-change event.

$$\lambda_t = L(\mathcal{G}) \times r \times (1 - \mathbb{P}(\text{topology-unchanged}|\mathcal{S}, \mathcal{G})) \tag{14}$$

This final expectation comes with an important caveat. Unlike the exact solution for the expected waiting distance to a tree-change, the waiting distance for a topology-change is an approximation, since we ignore the possible effects of intermediate tree-change events that could occur during the waiting distance until a topology-change (e.g., the second and third recombination events in Fig. 2). In other words, the probability of topology-change is not guaranteed to be homogeneous across some distance of the genome. This problem similarly arises in a single population model, where Deng *et al.* (2021) have previously demonstrated that its effect can likely be ignored. We re-examine this potential bias with respect to the presence of population structure in our validations below.

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 includes functions that accept a parameterized MSC model, initial genealogy, and recombination rate as input, and return the probabilities of different recombination event types. The probabilities can be calculated for specific branches and times, for entire branches, or for entire genealogies. Functions are also available to calculate the likelihood of observed waiting distances between different genealogy changes in an ARG using an exponential probability density parameterized by the MS-SMC'. The latter code is written in *numpy* (Harris *et al.*, 2020) using jit-compilation with *numba* (Lam *et al.*, 2015), and tree operations with *toytree* (Eaton, 2020). Below we explore the influence of MSC model parameters on waiting distances, and validate our solutions against expectations from coalescent simulations implemented in *ipcoal* and *msprime* (Baumdicker *et al.*, 2022), and investigate potential biases. 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 outcomes of recombination 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 lineages C and D (Fig. 1a; see Fig. S2 for detailed parameter settings). 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 as a consequence of recombination. 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 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-coalescence in one of those intervals.

For example, consider branch 2, which initially 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, as recombination occurs closer to the top, it is approaching the next species tree divergence event, where the number of samples will increase again, 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 that point, and on the same or different available branches, where a detached branch 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. 2c-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 well, 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 has the effect of either smoothing probabilities across intervals when N$_e$ is high, or accentuating differences among intervals when N$_e$ is low.

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, 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"

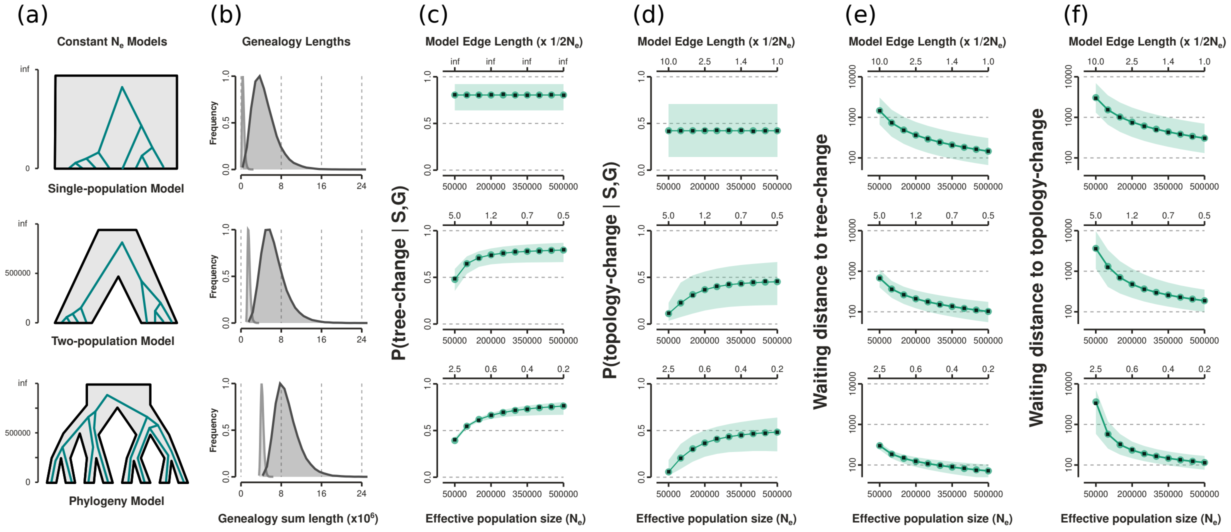


Figure 4. MS-SMC' estimates of waiting distances between recombination event types validated against coalescent simulations. Results are shown for three models (a) containing either 1, 2, or 8 populations, among which genealogies for 8 samples are embedded. For each model 10K tree sequences were simulated. (b) The sum of edge lengths across all starting genealogies for each model for the lowest and highest N$_e$ values examined (50K and 500K, respectively). (c-d) The mean and 95% CI (green circle and fill) of analytical solutions for the probabilities of recombination event types calculated from the first genealogy in each tree sequence. Results overlap with the mean frequencies of each event type in simulated tree sequences (black squares). The probabilities of tree-change or topology-change are constant with respect to N$_e$ in a 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) Waiting distances until tree or topology-change events exhibit less variance in models with population structure. The MS-SMC' calculated waiting distance expectations match very closely to simulated values.

model, which is an approximation), unless specified, and used 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 (λ) 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 ($L(\mathcal{G})$), and the probability ($\mathbb{P}$) of the specified event type (equations 9, 10, and 14). Across the three models examined, r remains constant, but both $L(\mathcal{G})$ and $\mathbb{P}$ can vary due to population structure, where the effect of structure is scaled by N$_e$. Therefore, we examined $L(\mathcal{G})$, $\mathbb{P}$, 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. Given each starting genealogy, the probability of each event type, and the expected waiting distance to that event type, were calculated under the MS-SMC'. To compare this with the result of a simulated coalescent with recombination process, we recorded the waiting distance until each recombination event type first occurred in each tree sequence, and also the frequency at which each event type occurred as the first event in a tree sequence, as a measurement of its empirical probability.

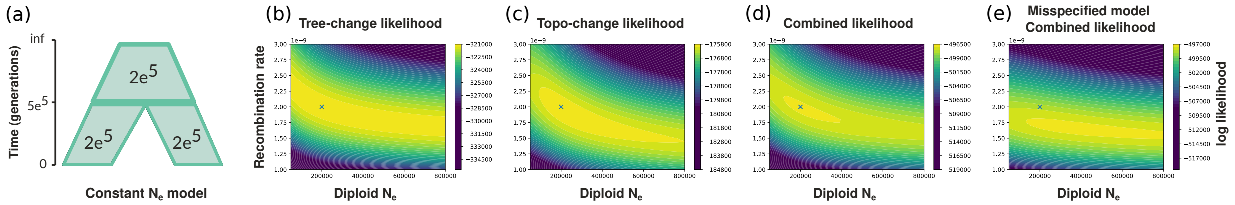


Figure 5. MS-SMC' likelihood framework. (a) ARGs were simulated under a two-population species tree model with a constant N_e=200K and r=2e^{-9}. (b) A joint log-likelihood surface for N_e and r inferred from the distances between tree-change events, (c) topology-change events; or (d) both. The true 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.

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 an effect on $L(\mathcal{G})$ at both low and high N$_e$ (Fig. 4b) of shifting the minimum and mean $L(\mathcal{G})$ higher. Because the frequency of recombination is positively correlated with $L(\mathcal{G})$ (the opportunity over which recombination can occur), larger $L(\mathcal{G})$ decreases waiting distances between recombination events, all else being equal. However, all else does not remain equal. Population structure also has an opposing effect on waiting distances by decreasing the probability of tree or topology changes (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 tree or topology change events is strongly associated with N_e in the former, while not affected at all in the latter. Consequently, in MSC models the waiting distances between each event type (Fig. 4e-f) represent a balance of the positive and negative effects of population structure on $L(\mathcal{G})$ and $\mathbb{P}$, 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 the probabilities of tree or topology changes, and consequently, also in the expected waiting distances. Although expected waiting distances do correlate with the population N_e in this model, the differences are small. By contrast, multispecies models exhibit much less variance in predicted probabilities of tree or topology changes given a set of MSC model parameters (Fig. 4c-d). This leads to a much tighter relationship (less variance) between MSC model parameters and expected waiting distances to tree or topology changes (Fig. 4e-f). Overall, this suggests that tree and topology-change distances may provide more information for inferring demographic parameters (or vice versa) when population structure is present.

3.4 MS-SMC' likelihood inference framework

Based on the expectation that waiting distances are informative about MSC model parameters, we developed a maximum likelihood framework for inferring MSC model parameters from observed waiting distances between tree and/or topology change events. Here, we apply this method using the true distances between events in simulated ARGs, however, it could similarly be applied to inferred distances between events in an ARG proposed from sequence data.

Given one or more ARGs each composing a sequence of genealogies G = $(\mathcal{G}_1, \mathcal{G}_2, ..., \mathcal{G}_n)$ and interval lengths X = $(x_1, x_2, ..., x_n)$, a subset of genealogies can be extracted representing each tree-change event, e.g., G_g = $(\mathcal{G}_1, \mathcal{G}_i, ...)$. The lengths of intervals between these events can be summed as tree-change waiting distances X$_g$ = $(\sum_{i=1}^{j} X_i, \sum_{i=j}^{k} X_i, ...,)$. The same could be done for topology-change events. Then, given a parameterized species tree $\mathcal{S}$ and recombination rate r we can obtain a sequence of exponential rate parameters $\Lambda_g = (\lambda_1, \lambda_i, ...)$ from equations 10 or 14. The likelihood of each waiting distance (X$_g$) between tree or topology change events can then be calculated from each exponential probability density function (equation 8) parameterized by the sequence of rate parameters (Λ_g). The set of species tree parameters that maximize the summed log-likelihood of all observed distances between events is the maximum likelihood solution.

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 sampled four haplotypes per population. This yielded 51,487 tree-change events with a mean length of 194bp, and 25,446 topology-change events with a mean length of 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 values between $r = 1e^{-9}$ to 3e^{-9}, and N_e=50K to 800K. The likelihood surface for tree-change distances exhibited a ridge that contained 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, representing a single population with constant N_e. In this case, the Λ_g rate parameters are calculated without enforcing a population split. This clearly introduces a significant bias in estimation, shifting the likelihood surface towards lower r and higher N_e. This demonstrates that even for a simple two population model, ignoring population structure can significantly bias spatial genealogical inference methods.

3.5 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 also act in opposing directions, such that their combined effect is further reduced.

The SMC' approximation to the coalescent with recombination is expected to deviate more significantly from the full model that it is approximating as the number of recombination events that the SMC' does not model (among ancestors that do not contribute genetic material to sampled descendants) increases. By not modeling some recombination events the SMC' will tend to over-estimate waiting distances between tree or topology change events. Because our waiting distance predictions are built upon assumptions of the SMC' model, they too will exhibit this bias. To investigate this source of error we repeated the simulations from our validation scenario using the *msprime* setting ancestry="smc_prime", to simulate tree sequences that exclude 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

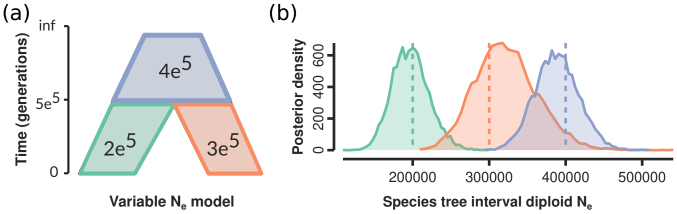


Figure 6. MS-SMC' likelihood framework. ARGs were simulated under a two-population species tree model with variable N_e. (e) A joint posterior distribution of N_e parameters inferred by Bayesian MCMC from tree and topology-change waiting distances.

full coalescent with recombination (Fig. S3). When quantified, As expected, we observe even less error between our predictions and coalescent simulations when data are simulated under the SMC' model (Fig. **??**). Whereas the error rate is generally below 5% when data are simulated under the full coalescent with recombination model, and only exceeds 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 there is an inverse relationships 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 than a single population model from this effect.

4 Discussion

We began with a simple question: what is the expected turnover rate in topologies along a genome under a species tree model? We generalized a recent solution that used a single population and constant Ne, instead structuring the equations to accept an arbitrary species tree topology and a different, arbitrary Ne value for each species tree branch. These solutions lay a groundwork for exploring how species tree structures affect neutral expectations of genealogical heterogeneity across chromosomes.

It is interesting here to also take note of the scale of waiting distances across different MSC model scenarios. - Previously the expected linkage in phylogenetic datasets could only be computed through exhaustive simulations (e.g., (McKenzie & Eaton, 2020b). - In addition to MSC parameters of divergence times and effective population sizes, other factors can also clearly affect the expected distances until ... Tree shape, tree size, the distribution of samples among species tree lineages

By accounting for the genomic heterogeneity that is expected due to incomplete lineage sorting, the multispecies coalescent model has facilitated the widespread use of multilocus data for phylogenetic inference. As phylogenetic systematics continues to turn toward whole genomes, methods should seek to take advantage of the increased resolution offered by genomic data. The traditional multispecies coalescent model overlooks the process of recombination, assuming that loci represent a single genetic history and that they are completely unlinked. In reality, under a neutral model, species tree parameters and recombination rates will influence the degree of genealogical turnover along a chromosome, potentially resulting in linked loci and/or multiple genealogical topologies per locus. Therefore, not accounting for recombination might mislead analyses. Conversely, incorporating the rates of turnover observed in the data as information could offer further clues for inference of the species tree model that generated it.

Inference of genealogies along a chromosome is a common goal in population genetics, and similar efforts have been undertaken in phylogenetics. Often, the goal of such efforts is to detect signals of introgression or selection. However, the phylogenetic approaches usually do not explicitly incorporate a model for recombination (e.g., Li *et al.*, 2019). Applications of our solutions might provide valuable null hypotheses of genealogical turnover rates by which introgression or selection might be detected. Beyond its use for detecting patterns resulting from non-neutral processes, incorporating recombination in phylogenetic-scale models could also help improve species tree inference. For example, to the extent that they are observable, the empirical distribution of waiting distances to topology changes might be inferred and compared against the expected waiting distances for a proposed species tree model (e.g., **Figure 5**). Further approaches could determine the distribution of waiting distances to specific types of topology changes, such as those that split up a focal clade.

Topological gravity wells. Even in the case of a neutral coalescent process with constant effective populations sizes and constant recombination rate across the span of a genome, highly structured demographic models can exhibit high spatial variance in the expected waiting distances until topology changes. This finding is particularly important for the practice of identifying evidence of selection or introgression based on the spatial distribution of gene tree patterns. Examples: Heliconius.

4.0.1 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 guidance on preliminary ideas and to...

References

Baum, D.A. (2007). Concordance Trees, Concordance Factors, and the Exploration of Reticulate Genealogy. *Taxon*, 56, 417–426. 3

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. 11

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. 2

Degnan, J.H. & Rosenberg, N.A. (2006). Discordance of species trees with their most likely gene trees. *PLoS genetics*, 2, e68. 3

Degnan, J.H. & Rosenberg, N.A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in ecology & evolution*, 24, 332–340. 1

Deng, Y., Song, Y.S. & Nielsen, R. (2021). The distribution of waiting distances in ancestral recombination graphs. *Theoretical Population Biology*, 141, 34–43. 3, 4, 5, 9, 11, 12, 16, 20, 24

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

Griffiths, R. & Marjoram, P. (1996). An ancestral recombination graph. In: *Progress in population genetics and human evolution*. Springer-Verlag, Berlin, pp. 257–270. 2

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. 11

Hubisz, M. & Siepel, A. (2020). Inference of ancestral recombination graphs using argweaver. In: *Statistical Population Genomics*. Humana, New York, NY, pp. 231–266. 3

Hudson, R.R. (1983). Properties of a neutral allele model with intragenic recombination. *Theoretical population biology*, 23, 183–201. 2

Kelleher, J., Etheridge, A.M. & McVean, G. (2016). Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS computational biology*, 12, e1004842. 2

Kingman, J.F.C. (1982). The coalescent. *Stochastic processes and their applications*, 13, 235–248. 1, 5

Knowles, L.L. & Kubatko, L.S. (2011). *Estimating Species Trees: Practical and Theoretical Aspects*. John Wiley and Sons. 3

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. 11

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. 17

Li, H. & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, 475, 493–496. 3

Maddison, W.P. (1997). Gene trees in species trees. *Systematic biology*, 46, 523–536. 1

Maddison, W.P. & Knowles, L.L. (2006). Inferring phylogeny despite incomplete lineage sorting. *Systematic biology*, 55, 21–30. 1

Marjoram, P. & Wall, J.D. (2006). Fast" coalescent" simulation. *BMC genetics*, 7, 1–9. 3

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*. 11

McKenzie, P.F. & Eaton, D.A.R. (2020b). The Multispecies Coalescent in Space and Time. *bioRxiv*, p. 2020.08.02.233395. 17

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

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

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

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

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, 8

5 Appendix: Derivations

5.1 Notation

Information from the genealogy embedding table can be used to calculate the probabilities of no-change, tree-change, and topology-change events under the MS-SMC'. These equations, described below, use the terms defined in Table 2. A species tree $\mathcal{S}$ is a parameterized multispecies coalescent model (demographic 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 $\mathcal{G}$ 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 intervals can be defined that are delimited by events that change the rate of coalescence. This is termed a genealogy embedding table (e.g., Table 1). In the waiting distance solutions for a single population with constant N_e by Deng *et al.* (2021),

Table 2. Summary of variables used in waiting distance equations.

Variable	Description
$\mathcal{S}$	An MSC model with topology, divergence times and effective population sizes.
$\mathcal{G}$	A genealogy that can be embedded in $\mathcal{S}$.
$L(\mathcal{G})$	Sum of edge lengths of genealogy $\mathcal{G}$.
b	A focal branch in $\mathcal{G}$.
i	Interval in the genealogy embedding table in which recombination occurs.
$\mathcal{I}_b$	Ordered set of intervals on branch b.
$\mathcal{I}_c$	Ordered set of intervals on branch c, the parent of branch b.
$\mathcal{I}_{bc}$	Ordered union of sets $\mathcal{I}_b$ and $\mathcal{I}_c$.
$\mathcal{J}_b(i)$	Ordered set of intervals above i on branch b.
$\mathcal{Q}_b(i,j)$	Ordered set of intervals above i and below j on branch b.
$\mathcal{K}(b,t)$	Number of edges of $\mathcal{G}$ in the interval containing branch b at time t.
k_x	Number of edges of $\mathcal{G}$ in interval x; piece-wise constant of $A(b,t)$.
$\mathcal{N}(b,t)$	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.
σ_x	The lower boundary of interval x, in generations.
μ_x	The upper boundary of interval x, in generations.
d_x	The length of interval x, in generations.
t_b^l	The lower boundary of branch b, in generations.
t_b^u	The upper boundary of branch b, in generations.
t_b^m	The time at which a focal branch b is able to coalesce with its sibling branch.
m	Index of an interval in the genealogy embedding table with lower boundary t_b^m.
$\mathcal{M}_b$	Ordered set of intervals from m and above on branch b.
$\mathcal{L}_b$	Ordered set of intervals below m on branch b.

this table is delimited only by coalescent events. Because N$_e$ is constant in their framework, only k differs between intervals, and this determines differences in rates of coalescence, which decrease 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, and correspond to time intervals with a constant k and N_e. 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).

Each branch on $\mathcal{G}$ will span one or more genealogy embedding intervals. The ordered set of intervals on a specific branch, b, is defined as $\mathcal{I}_b$. The lower and upper bound of each interval is σ_x and μ_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_b^l and t_b^u, respectively.

5.2 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 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_b^u - t_b^l). 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, τ, defined as $\mathcal{N}(b,\tau)$, and the number of other genealogy branches in the interval that includes branch b at time τ, defined as $\mathcal{K}(b,\tau)$. 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 $\lambda = \frac{\mathcal{K}(b,\tau)}{2\mathcal{N}(b,\tau)}$, similar to equations 1-2.

5.2.1 Probability of no-change 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 over the length from the time of recombination (t_r) to the top of the branch (t_b^u), 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 τ, where $\mathcal{K}(b,\tau)$ and $\mathcal{N}(b,\tau)$ can vary across the length of the branch if it spans different intervals.

$$\mathbb{P}(\text{tree-unchanged}|\mathcal{S},\mathcal{G},b,t_r) = \int_{t_r}^{t_b^u} \frac{1}{\mathcal{K}(b,\tau)} f(\tau - t_r;\lambda) d\tau \tag{15}$$

The exponential probability density function can be expanded, as in equation 2, where λ is $\mathcal{K}(b,\tau)$ over $2\mathcal{N}(b,\tau)$:

$$= \int_{t_r}^{t_b^u} \frac{1}{\mathcal{K}(b,\tau)} \frac{\mathcal{K}(b,\tau)}{2\mathcal{N}(b,\tau)} \exp\left\{ - \int_{t_r}^{\tau} \frac{\mathcal{K}(b,s)}{2\mathcal{N}(b,s)} ds \right\} d\tau \tag{16}$$

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 τ.

$$= \int_{t_r}^{t_b^u} \frac{1}{2\mathcal{N}(b,\tau)} \exp\bigg\{ -\int_{t_r}^{\tau} \frac{\mathcal{K}(b,s)}{2\mathcal{N}(b,s)} ds \bigg\} d\tau \tag{17}$$

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. (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 u_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 (σ_j) to the top (μ_j) of the interval. The ordered set of all intervals above i in $\mathcal{I}_b$ is defined as $\mathcal{J}_b(i) = \{j \in \mathcal{I}_b \mid j > i\}$.

$$= \int_{t_r}^{\mu_i} \frac{1}{2\mathcal{N}(b,\tau)} \exp\bigg\{ -\int_{t_r}^{\tau} \frac{\mathcal{K}(b,s)}{2\mathcal{N}(b,s)} ds \bigg\} d\tau + \sum_{j\in\mathcal{J}_b(i)} \int_{\sigma_j}^{\mu_j} \frac{1}{2\mathcal{N}(\tau)} \exp\bigg\{ -\int_{t_r}^{\tau} \frac{\mathcal{K}(s)}{2\mathcal{N}(s)} ds \bigg\} d\tau \tag{18}$$

We can now solve this equation and substitute constant values for $\mathcal{K}(b,\tau)$ and $\mathcal{N}(b,\tau)$ 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 $f(i,i)$.

$$= \frac{1}{k_i} - \frac{1}{k_i} \exp\bigg\{ -\frac{k_i}{2n_i}\mu_i \bigg\} \exp\bigg\{ \frac{k_i}{2n_i} t_r \bigg\} \tag{19}$$

$$= \frac{1}{k_i} + f(i,i) \exp\bigg\{ \frac{k_i}{2n_i} t_r \bigg\} \tag{20}$$

Later terms – 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:

$$= \sum_{j\in\mathcal{J}_b(i)} \frac{1}{k_j} \bigg(1 - \exp\bigg\{ -\frac{k_j}{2n_j} d_j \bigg\} \bigg) \exp\bigg\{ -\frac{k_i}{2n_i}\mu_i - \sum_{q\in\mathcal{Q}_b(i,j)} \frac{k_q}{2n_q} d_q \bigg\} \exp\bigg\{ \frac{k_i}{2n_i} t_r \bigg\} \tag{21}$$

$$= \sum_{j\in\mathcal{J}_b} f(i,j) \exp\bigg\{ \frac{k_i}{2n_i} t_r \bigg\} \tag{22}$$

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:

$$\mathbb{P}(\text{no-change}|\mathcal{S},\mathcal{G},b,t_r) = \frac{1}{k_i} + f(i,i)\exp\left\{\frac{k_i}{2n_i}t_r\right\} + \sum_{j\in\mathcal{J}_b(i)} f(i,j)\exp\left\{\frac{k_i}{2n_i}t_r\right\} \tag{3}$$

5.2.2 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:

$$\mathbb{P}(\text{tree-unchanged}|\mathcal{S},\mathcal{G},b) = \frac{1}{t_b^u - t_b^l}\int_{t_b^l}^{t_b^u} \mathbb{P}(\text{tree-unchanged}|\mathcal{S},\mathcal{G},b,t_r)dt_r \tag{23}$$

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

$$\begin{aligned}
&= \frac{1}{t_b^u - t_b^l}\sum_{i\in\mathcal{I}_b}\frac{1}{k_i}d_i + \left(\exp\left\{\frac{k_i}{2n_i}\mu_i\right\} - \exp\left\{\frac{k_i}{2n_i}\sigma_i\right\}\right)\\
&\left[-\frac{2n_i}{k_i^2}\exp\left\{-\frac{k_i}{2n_i}\mu_i\right\}+\right.\\
&\left.\frac{2n_i}{k_i}\left(\sum_{j\in\mathcal{J}_b(i)}\exp\left\{-\frac{k_i}{2n_i}\mu_i - \sum_{q\in\mathcal{Q}_b(i,j)}\frac{k_q}{2n_q}d_q\right\}\frac{-1}{k_j}\exp\left\{-\frac{k_j}{2n_j}d_j\right\}\right)\right]
\end{aligned} \tag{24}$$

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.

$$\begin{aligned}
\mathbb{P}(\text{tree-unchanged}|\mathcal{S},\mathcal{G},b) &= \frac{1}{t_b^u - t_b^l}\int_{t_b^l}^{t_b^u}\mathbb{P}(\text{tree-unchanged}|\mathcal{S},\mathcal{G},b,t)dt\\
&= \frac{1}{t_b^u - t_b^l}\sum_{i\in\mathcal{I}_b}\frac{1}{k_i}d_i + \left(\frac{2n_i}{k_i}\sum_{j\in\mathcal{J}_b(i)} f(i,j)\left(\exp\left\{\frac{k_i}{2n_i}\mu_i\right\} - \exp\left\{\frac{k_i}{2n_i}\sigma_i\right\}\right)\right)
\end{aligned} \tag{5}$$

5.2.3 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 unchanging probabilities for all branches:

$$\mathbb{P}(\text{tree-unchanged}|\mathcal{S},\mathcal{G}) = \sum_{b\in G}\left[\frac{t_b^u - t_b^l}{L(\mathcal{G})}\right]\mathbb{P}(\text{tree-unchanged}|\mathcal{S},\mathcal{G},b) \tag{25}$$

To derive the expected waiting distance to the next tree change, we use describe the probability of a tree-change event as $1 - \mathbb{P}(\text{tree-unchanged}|\mathcal{S},\mathcal{G})$. Following the approach described in equations 7-10, we can calculate an expected waiting distance by treating this as an exponentially distributed random variable and calculating a rate parameter.

5.3 The distribution of distances to a change in genealogical topology

As in the first section, we treat branches individually and then sum across them. However, we have to designate new variables. We now consider two lineages in addition to the focal lineage b: the lineage b' that b coalesces with, and the lineage c that is ancestral to b and b'. We also designate the timepoint t_b^m, which we use to break the problem into two cases. While the single-population example in Deng *et al.* (2021) uses $t_{b'}^l$ as this breakpoint, the species tree introduces more complexity, and we instead use the maximum of three values: $t_{b'}^l$, t_b^l, and $t_{(b,b')}^w$, which we define as the merging time for the species tree branches separating b and b' (Figure-supplement).

In the first case (Fig. 3a), where $t_r < t_b^m$ and $t_r \in [\sigma_i, \sigma_{i+1}] \subset [t_b^l, t_b^m]$ it is necessary to integrate over coalescence probabilities in three distinct sections: from t_r to t_b^m, t_b^m to t_b^u, and t_b^u to t_c^u. (Note, these sections may be composed of multiple genealogy embedding intervals if they contain additional species divergence events). The first section is notable for representing the core difference between this first case and the second case below. When a recombination event occurs in the first section on b there is a span from t_r to t_b^m where b can re-coalescence with itself but not yet with b', since at least one species divergence event separates them. Thus, although re-coalescence can occur in this span of time, it cannot lead to changes in the genealogy. In the second section, b can re-coalesce with b or b', leading to either no change or a tree change (shortened coalesence time). Finally, in the third section b can only re-coalesce with c, leading to a tree change that lengthens b's coalescence time. An integration over the probabilities of each allowable event on branch b over all three distinct sections leads to the first probability statement below. In the second case (Fig. 3b), where $t_r > t_b^m$ and $t_r \in [\sigma_i, \sigma_{i+1}] \subset [t_b^m, t_b^u]$, it is only necessary to integrate over probabilities across a subset of the second section, and across the entire third section described above, leading to the second statement below:

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. We define bc as the ordered union of the sets of intervals on branches b and c (Fig. 3c), such that $\mathcal{I}_{bc}$ is the summed number of intervals in this set. Similarly, we define the intersection of the sets of intervals in b and b' as bb', which includes only intervals in which both of these branches occur (i.e., it excludes intervals where they are embedded in separate species tree branches). Finally, we define m as the index of the lowest interval in bb' (occurring at time t_b^m) where both b and b' occur.

$$\mathbb{P}(\text{topology unchanged}|\mathcal{S}, \mathcal{G}, b) = \frac{1}{t_b^u - t_b^l} \int_{t_b^l}^{t_b^u} \mathbb{P}(\text{topology-unchanged}|\mathcal{S}, \mathcal{G}, b, t_r)dt_r \tag{26}$$

$$= \frac{1}{t_b^u - t_b^l} \left[\left(\int_{t_b^l}^{t_b^m} + \int_{t_b^m}^{t_b^u} \right) \mathbb{P}(\text{topology-unchanged}|\mathcal{S}, \mathcal{G}, b, t_r)dt_r \right] \tag{27}$$

Where $t_b^m = \max\{t_{(b,b')}^w, t_{b'}^l, t_b^l\}$.

5.3.1 Given a branch and a time

As in Deng *et al.* (2021), we break the problem into two cases: a first case in which t belongs to the interval from the base of the focal branch to t_b^m, and a second case in which t belongs to the interval from t_b^m to t_b^u.

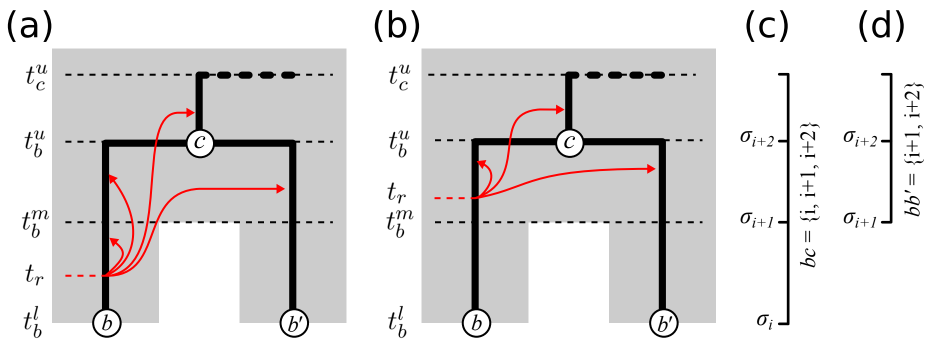


Figure 7. To calculate the probability that recombination on a genealogy branch (b) leads to a topology change involves summing over the probabilities that a detached lineage does not re-coalesce with either itself (b), its sibling (b') or its parent (c) branches. At the time recombination occurs (t_r) the branches b and b' can either exist in different species tree intervals (a) or the same interval (b). This restricts the probability of tree-change outcomes (e.g., shortened coalescent time) until the lowest shared interval between b and b' at time t_m. This leads to two ordered sets of intervals (c-d) used in equations 11 and 12.

First case – Given $t \in [\sigma_i, \sigma_{i+1}] \subset [t_b^l, t_b^m]$: The final solution here is shown as equation XX of the main text. It is an integral over part 1, two times the integral in part 2, since this length exists on both branches b and b', and then the integral over part 3.

$$\begin{gathered}\mathbb{P}(\text{topology-unchanged}|\mathcal{S},\mathcal{G},b,t_r) = \\ \int_{t_r}^{t_b^m} \frac{1}{\mathcal{K}(b,\tau)} f(\tau - t_r;\lambda)d\tau + \int_{t_b^m}^{t_b^u} \frac{2}{\mathcal{K}(b,\tau)} f(\tau - t_r;\lambda)d\tau + \int_{t_b^u}^{t_c^u} \frac{1}{\mathcal{K}(b,\tau)} f(\tau - t_r;\lambda)d\tau = \\ \frac{1}{k_i} + \sum_{j\in\mathcal{I}_{bc}} f(i,j)\exp\left\{\frac{k_i}{n_i}t_r\right\} + \sum_{j\in\mathcal{M}_b} f(i,j)\exp\left\{\frac{k_i}{n_i}t_r\right\}\end{gathered} \tag{28}$$

Second case – Given $t \in [\sigma_i, \sigma_{i+1}] \subset [t_b^m, t_b^u]$. This is similar to the first, but we can eliminate the first integral. The final result is shown as equation XX in the main text.

$$\begin{gathered}\mathbb{P}(\text{topology-unchanged}|\mathcal{S},\mathcal{G},b,t_r) = \\ \int_{t_r}^{t_b^u} \frac{2}{\mathcal{K}(b,\tau)} f(\tau - t_r;\lambda)d\tau + \int_{t_b^u}^{t_c^u} \frac{1}{\mathcal{K}(b,\tau)} f(\tau - t_r;\lambda)d\tau = \\ 2\left(\frac{1}{k_i} + \sum_{k\in\mathcal{I}_b} f(i,j)\exp\left\{\frac{k_i}{n_i}t_r\right\}\right) + \sum_{j\in\mathcal{I}_c} f(i,j)\exp\left\{\frac{k_i}{n_i}t_r\right\}\end{gathered}$$

5.3.2 Across a full branch

Now we derive the overall probability that a recombination event falling on a specific branch will change the topology. We integrate across values of t and sum the two cases defined above, while assuming a uniform probability of the recombination event occurring at any time along the branch:

First case –

$$\int_{t_b^l}^{t_b^m} \mathbb{P}(\text{topology-unchanged}|\mathcal{S},\mathcal{G},b,t_r) = \\ \sum_{i\in\mathcal{M}_b} \frac{1}{k_i}\left[d_i + n_i\left(\exp\left\{\frac{k_i}{n_i}\mu_i\right\} - \exp\left\{\frac{k_i}{n_i}\sigma_i\right\}\right)\left(\sum_{j\in\mathcal{I}_{bc}} f(i,j) + \sum_{j\in\mathcal{M}_b} f(i,j)\right)\right] \tag{29}$$

Second case –

$$\int_{t_b^m}^{t_b^u} \mathbb{P}(\text{topology-unchanged}|\mathcal{S},\mathcal{G},b,t_r) = \\ \sum_{i\in\mathcal{M}_b} \frac{1}{k_i}\left[2d_i + n_i\left(\exp\left\{\frac{k_i}{n_i}\mu_i\right\} - \exp\left\{\frac{k_i}{n_i}\sigma_i\right\}\right)\left(2\sum_{j\in\mathcal{I}_b} f(i,j) + \sum_{j\in\mathcal{I}_c} f(i,j)\right)\right] \tag{30}$$

Result –

$$\mathbb{P}(\text{topology-unchanged}|\mathcal{S},\mathcal{G},b) = \frac{1}{t_b^u - t_b^l}\left[\sum_{i\in\mathcal{L}_b} p_{b,1}^{(i)} + \sum_{i\in\mathcal{M}_b} p_{b,2}^{(i)}\right] \tag{31}$$

5.3.3 Across the whole tree

Finally, we sum across all branches to find the probability of a recombination event changing the topology of the tree:

$$\mathbb{P}(\text{topology-unchanged}|\mathcal{S},\mathcal{G}) = \sum_{b\in\mathcal{G}} \frac{t_b^u - t_b^l}{L(\mathcal{G})} \times \mathbb{P}(\text{topology-unchanged}|\mathcal{S},\mathcal{G},b) = \\ \frac{1}{L(\mathcal{G})}\sum_{b\in\mathcal{G}}\left[\sum_{i\in\mathcal{L}_b} p_{b,1}^{(i)} + \sum_{i\in\mathcal{M}_b} p_{b,2}^{(i)}\right] \tag{32}$$

6 Supplementary Information

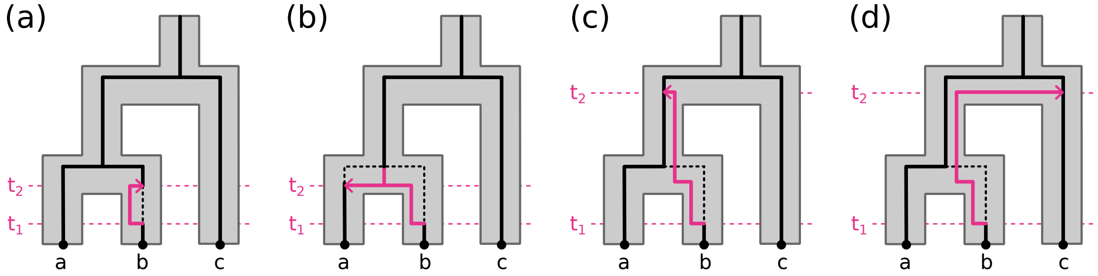
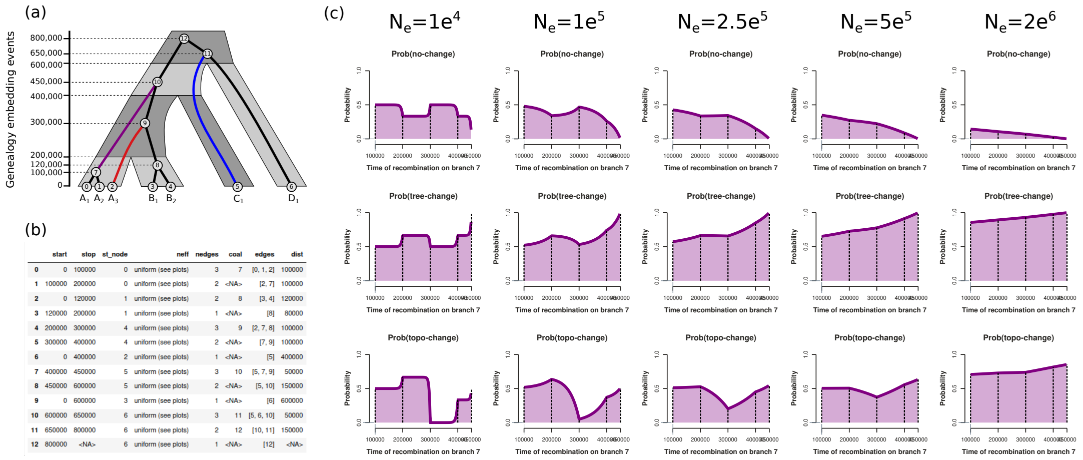


Figure S1. Four categories of outcomes of a recombination event occurring on a genealogy at time t$_1$, where the detached subtree re-coalesces with the remaining lineages under the SMC' process at time t$_2$. (a) The detached subtree re-coalesces with the original lineage from which it was detached leading to no change between the starting genealogy and subsequent genealogy. (b) The detached subtree re-coalesces with its sister 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 sister lineage. (d) The detached subtree re-coalesces with a lineage other than itself, its sister, or its parent lineage, leading to a topology-change.



	start	stop	st_node	neff	nedges	coal	edges	dist
0	0	100000	0	uniform (see plots)	3	7	[0, 1, 2]	100000
1	100000	200000	0	uniform (see plots)	2	<NA>	[2, 7]	100000
2	0	120000	1	uniform (see plots)	2	8	[3, 4]	120000
3	120000	200000	1	uniform (see plots)	1	<NA>	[8]	80000
4	200000	300000	4	uniform (see plots)	3	9	[2, 7, 8]	100000
5	300000	400000	4	uniform (see plots)	2	<NA>	[7, 9]	100000
6	0	400000	2	uniform (see plots)	1	<NA>	[5]	400000
7	400000	450000	5	uniform (see plots)	3	10	[5, 7, 9]	50000
8	450000	600000	5	uniform (see plots)	2	<NA>	[5, 10]	150000
9	0	600000	3	uniform (see plots)	1	<NA>	[6]	600000
10	600000	650000	6	uniform (see plots)	3	11	[5, 6, 10]	50000
11	650000	800000	6	uniform (see plots)	2	12	[10, 11]	150000
12	800000	<NA>	6	uniform (see plots)	1	<NA>	[12]	<NA>

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

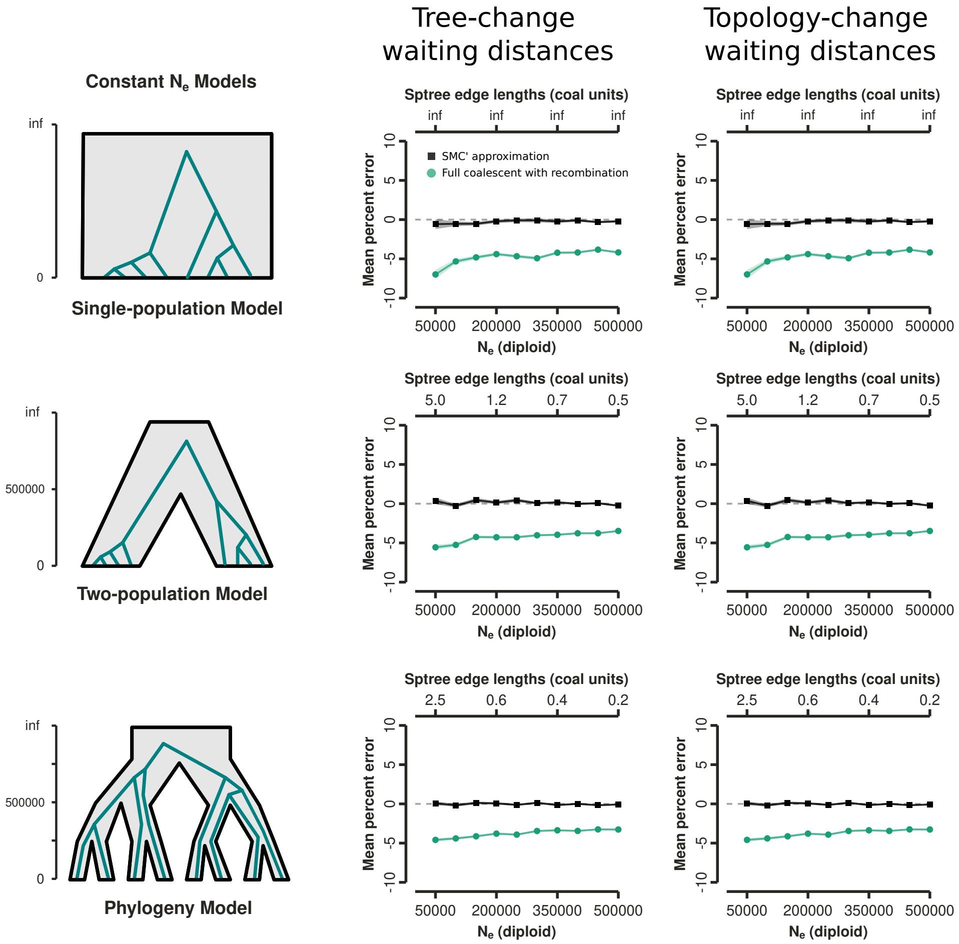


Figure S3. Error in ...

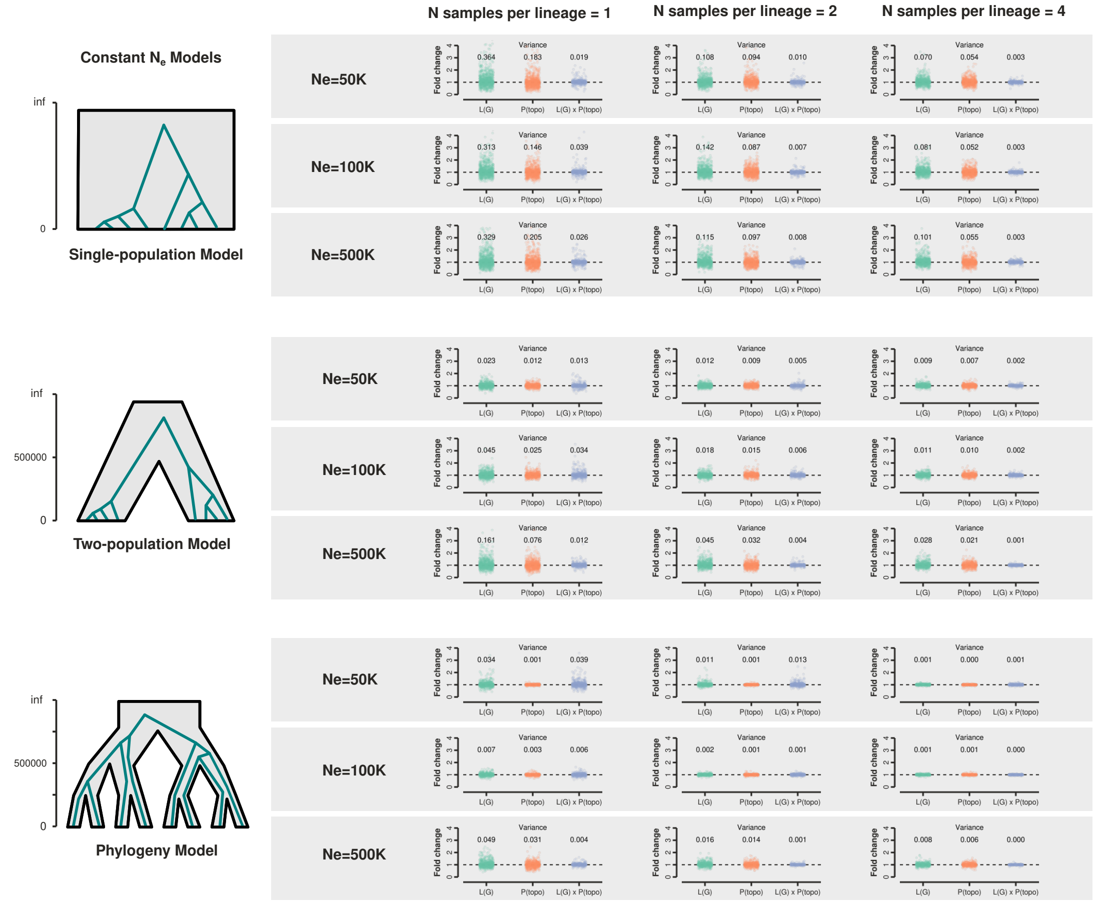
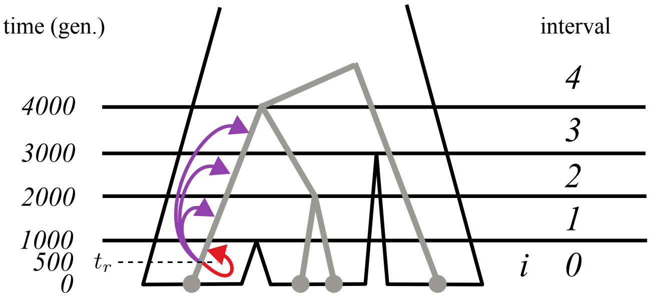


Figure S4. Variance in the fold-change in components affecting the expected waiting distance to a topology change between the starting tree and a subsequent tree that has experienced a tree-change event that changed coalescent times but not the topology. The sum of genealogical edge lengths (L(G)), the 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.



The probability of no-change given recombination on a given branch at a given time:

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

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:

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

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

$$f(i,i) = -\frac{1}{a_i}\exp\left\{-\frac{a_i}{2n_i}\mu_i\right\}$$

$$f(0,0) = -\frac{1}{1}\exp\left\{-\frac{1}{2(1000)}1000\right\}$$

$$f(i,j) = \frac{1}{a_j}\left(1-\exp\left\{-\frac{a_j}{2n_j}d_j\right\}\right)\exp\left\{-\frac{a_i}{2n_i}\mu_i - \sum_{q\in\mathcal{Q}_b}\frac{a_q}{2n_q}d_q\right\}$$

$$f(0,1) = \frac{1}{3}\left(1-\exp\left\{-\frac{3}{2(1000)}1000\right\}\right)\exp\left\{-\frac{1}{2(1000)}1000\right\}$$

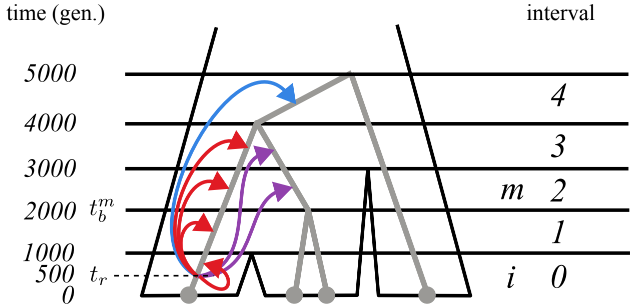
$$f(0,2) = \frac{1}{2}\left(1-\exp\left\{-\frac{2}{2(1000)}1000\right\}\right)\exp\left\{-\frac{1}{2(1000)}1000-\left(\frac{3}{2(1000)}1000\right)\right\}$$

$$f(0,3) = \frac{1}{3}\left(1-\exp\left\{-\frac{3}{2(1000)}1000\right\}\right)\exp\left\{-\frac{1}{2(1000)}1000-\left(\frac{3}{2(1000)}1000+\frac{2}{2(1000)}1000\right)\right\}$$

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

$$\begin{aligned}\mathbb{P}(\text{no-change}|\mathcal{S},\mathcal{G},b,t_r) &= 1 + f(0,0)\times 1.284 + \sum_{j\in\{1,2,3\}} f(0,j)\times 1.284\\ &= 1 + (-0.6065\times 1.284) + (0.1571\times 1.284) + (0.0428\times 1.284) + (0.0129\times 1.284)\\ &= 0.4944\end{aligned}$$

Figure S5. A step-by-step calculation of the probability of a tree-unchanged event under the MS-SMC' given a species tree and genealogy.



The probability the topology is unchanged given recombination on a given branch at a given timee:

$$\mathbb{P}(\text{topology-unchanged}|\mathcal{S},\mathcal{G},b,t_r) = \begin{cases} \dfrac{1}{a_i} + \sum\limits_{\underline{\underline{j\in\mathcal{I}_{bc}}}} f(i,j)\exp\left\{\dfrac{a_i}{2n_i}t_r\right\} + \sum\limits_{\underline{j\in\mathcal{M}_b}} f(i,j)\exp\left\{\dfrac{a_i}{2n_i}t_r\right\}, & \text{if } t_r < t_b^m \\ 2\left(\dfrac{1}{a_i} + \sum\limits_{\underline{\underline{j\in\mathcal{I}_b}}} f(i,j)\exp\left\{\dfrac{a_i}{2n_i}t_r\right\}\right) + \sum\limits_{\underline{j\in\mathcal{I}_c}} f(i,j)\exp\left\{\dfrac{a_i}{2n_i}t_r\right\}, & \text{if } t_r \geq t_b^m \end{cases}$$

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:

$$\mathbb{P}(\text{topology-unchanged}|\mathcal{S},\mathcal{G},b,t_r) = \frac{1}{a_i} + \sum_{j\in\mathcal{I}_{bc}} f(i,j)\exp\left\{\frac{a_i}{2n_i}t_r\right\} + \sum_{j\in\mathcal{M}_b} f(i,j)\exp\left\{\frac{a_i}{2n_i}t_r\right\}$$

$$= \frac{1}{1} + \sum_{j\in\{0,1,2,3,4\}} f(i,j)\exp\left\{\frac{1}{2(1000)}500\right\} + \sum_{j\in\{2,3\}} f(i,j)\exp\left\{\frac{1}{2(1000)}500\right\}$$

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

$$f(i,j) = \frac{1}{a_j}\left(1-\exp\left\{-\frac{a_j}{2n_j}d_j\right\}\right)\exp\left\{-\frac{a_i}{2n_i}\mu_i - \sum_{q\in\mathcal{Q}_b}\frac{a_q}{2n_q}d_q\right\}$$

$$f(0,0) = -\frac{1}{1}\exp\left\{-\frac{1}{2(1000)}1000\right\}$$

$$f(0,1) = \frac{1}{3}\left(1-\exp\left\{-\frac{3}{2(1000)}1000\right\}\right)\exp\left\{-\frac{1}{2(1000)}1000\right\}$$

$$f(0,2) = \frac{1}{2}\left(1-\exp\left\{-\frac{2}{2(1000)}1000\right\}\right)\exp\left\{-\frac{1}{2(1000)}1000-\left(\frac{3}{2(1000)}1000\right)\right\}$$

$$f(0,3) = \frac{1}{3}\left(1-\exp\left\{-\frac{3}{2(1000)}1000\right\}\right)\exp\left\{-\frac{1}{2(1000)}1000-\left(\frac{3}{2(1000)}1000+\frac{2}{2(1000)}1000\right)\right\}$$

$$f(0,4) = \frac{1}{2}\left(1-\exp\left\{-\frac{2}{2(1000)}1000\right\}\right)\exp\left\{-\frac{1}{2(1000)}1000-\left(\frac{3}{2(1000)}1000+\frac{2}{2(1000)}1000+\frac{3}{2(1000)}1000\right)\right\}$$

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

$$= 1+(-0.6065\times1.284)+(0.1571\times1.284)+(0.0428\times1.284)+(0.0129\times1.284)+(0.0035\times1.284)+(0.0428\times1.284)+(0.0129\times1.284)$$

$$= 0.5703$$

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

Table S1. A table summarizing the relationships among branches in the genealogical tree embedded in the species tree in Figure 1.

Branch	$\mathcal{I}_b$	t_b^l	t_b^u	Parent	Sibling	t_b^m
0	1	0	t_7	7	1	0
1	1	0	t_7	7	0	0
2	3	0	t_9	9	8	W_{AB}
3	1	0	t_8	8	4	0
4	1	0	t_8	8	3	0
5	4	0	t_{11}	11	6	W_{ABCD}
6	2	0	t_{11}	11	5	W_{ABCD}
7	4	t_7	t_{10}	10	9	t_9
8	2	t_8	t_9	9	2	W_{AB}
9	2	t_9	t_{10}	10	7	t_9
10	3	t_{10}	t_{12}	12	11	t_{11}
11	1	t_{11}	t_{12}	12	10	t_{11}