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 fitting species tree models.

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 species 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. 1), 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 (Ne) is the key parameter determining the rate of coalescence, and can vary among different lineages.

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

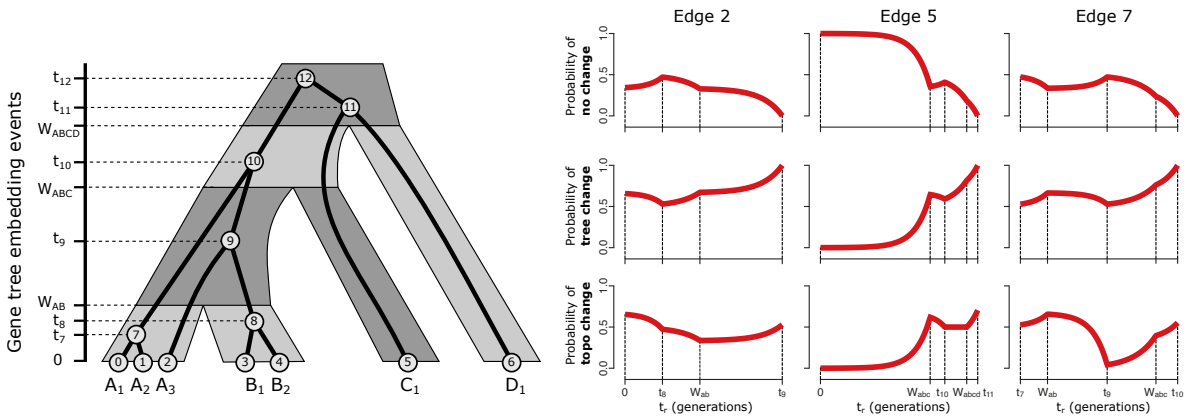

Figure 1. A gene tree embedded in a species tree. (a) Species tree divergence times (W) define ancestral populations. The population structure constrains which lineages of the genealogy can coalesce. Further, each population which can have a different effective population size (N_b), which determine rates of coalescence within that population. Viewed backwards in time, coalescent events (e.g., t_7) in the genealogy reduce the number of samples, whereas species tree divergence events (e.g., W_{AB}) increase the number of samples present. (b-d) Recombination is modeled by detaching a branch at a specific time and sampling a waiting time until it re-coalesces, leading to either no change, a tree change (edge lengths), or topology change. We arbitrarily selected edges 2, 5, and 7 to illustrate how piecewise variation in coalescent rates back in time through the species tree leads to changing probabilities of each outcome (discussed in Demonstration). Note that the values in (b) and (c) must sum to 1, since one of those two events must happen, while (d) is showing the probabilities of a subset of the events in (c). See Table 1 for the full gene tree embedding table for this example.

is a consequence of shared ancestry among samples at nearby regions, which decays over time and distance as recombination events reduce their shared ancestry. This process can be approximated by randomly breaking an edge on a genealogy and sampling a waiting time (based on the lineage effective population size) until it reconnects to the genealogy at a different shared ancestor (Fig. 2). In this way, the coalescent with recombination can be viewed as a process where a set of samples at the present substitutes one ancestor for another on either side of a recombination breakpoint.

An algorithm to simulate the coalescent with recombination was developed early on (Hudson, 1983), as a method for yielding a distribution of linked genealogies as well as the breakpoints across the genome where recombination occurred (i.e., "waiting distances" between genealogies). As a data structure, this can be represented either as an ancestral recombination graph (ARG) (Griffiths & Marjoram, 1996) or tree-sequence (Kelleher *et al.*, 2016) (hereafter we will refer to it as an ARG). Unfortunately, although generating ARGs is relatively simple, inferring an ARG from sequence data remains highly challenging, as it is computationally intensive, highly limited by sequence information, and can be nearly infinite in possibilities for large genome lengths and numbers of samples (McVean & Cardin, 2005). A major advance was achieved through development of the Sequentially Markov Coalescent (SMC), a simplification of the space of possible ARGs that restricts the types of coalescent events that can occur between sequential genealogies, in a way that enables modeling changes between them as a Markov process (McVean & Cardin, 2005). Implicit to SMC-based methods is the expectation that recombination occurs at some rate from which the waiting distance between recombination events can be modeled as an exponentially distributed random variable. Under these assumptions a tractable likelihood framework can be developed. Because neither genealogies or segment lengths can be observed directly, most SMC-based methods use hidden Markov model (HMM) methods to treat these as hidden states to be inferred from sequence data. Examples of inference tools built on the SMC framework include PSMC (Li & Durbin, 2011) and MSMC (Schiffels & Durbin, 2014), for inferring changes in effective population sizes of single or multiple populations through time, respectively, based on pairwise coalescent times between sequential genealogies, (ARGweaver; Hubisz & Siepel, 2020; Rasmussen *et al.*, 2014). as well as 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 that result in no change between neighboring genealogies, which has been shown to significantly improve SMC-based inferences (Wilton *et al.*, 2015). Under the SMC', there are thus four categories of possible outcomes of recombination (Fig. 2): (1) no change between sequential genealogies; (2) shortening of a coalescent time; (3) lengthening of a coalescent time; and (4) a coalescent event that changes the topology (relationships). These can be grouped more generally into three categories: "no change" (category 1), "tree change" (categories 2 or 3), and "topology change" (category 4). Recently, Deng *et al.* (2021) derived a set of solutions for the expected waiting distances between these three categories of outcomes for a single population with constant effective population size. This provides an important advance in establishing a neutral expectation not only for generic recombination events to occur, but for different categories of events that leave different detectable signatures in the genome.

Here, we extend the methods of Deng *et al.* (2021) to an MSC framework to approximate the expected

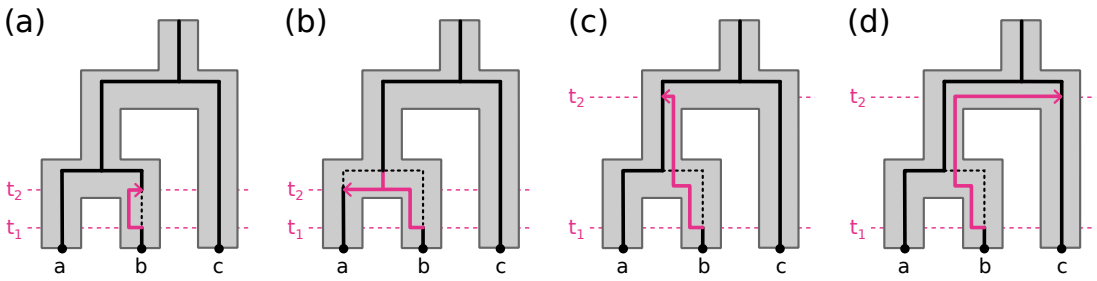

Figure 2. Examples of the four different categories of outcomes from a recombination event (adapted from (Deng *et al.*, 2021)). (a) A category 1 event occurs when the dislocated lineage reattaches to the original lineage. (b) A category 2 event occurs when a dislocated lineage attaches to its sibling lineage (the lineage that it originally coalesced with), shortening both of their branch lengths and extending that of its parental lineage. (c) A category 3 event occurs when the dislocated lineage attaches to its parental lineage, lengthening the original lineage and its sibling lineage, and shortening the parental lineage. (d) A category 4 recombination event occurs when the dislocated lineage attaches with any lineage other than itself, its sibling lineage, and its parental lineage. This is the only category of recombination event that changes the topology of the genealogy.

waiting distances between different categories of genealogy changes under a parameterized species tree model. In this respect, the waiting distances between recombination events of category 4 may be of greatest interest, as topology changes 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. However, events of categories 1-3 occur disproportionately often in small sample sizes (Wilton *et al.*, 2015), which are especially 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 categories of recombination events that will be observed. Consequently, the distributions of waiting distances between different categories 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 effective population sizes assigned to each branch (N_b), a genealogy ($\mathcal{G}$) for any number of sampled gene copies 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. The rate of coalescence is $\frac{1}{2N_b}$, and the expected waiting time between coalescence events that will reduce the number of samples from k to k-1 in a population follows the Kingman coalescent:

$$\mathbb{E}(t_k|N) = \frac{k(k-1)}{2N} \tag{1}$$

In a single population model with constant N this rate decreases monotonically as the number of remaining samples after each coalescence event decreases. In an MSC model, however, the rate of coalescence typically waxes and wanes through time, as the transition from one branch interval to the next can be associated with different N_b and an increase in k as samples from descendant branch intervals are merged.

Based on this generative framework for sampling genealogies, a set of likelihood solutions have been developed to fit coalescent model parameters, such as N in single population models ((Kingman, 1982)), or N_b and W values 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 equation 1 as a rate parameter (λ) to calculate the likelihood of an observed coalescent waiting time (t_k) between k and k-1 lineages from an exponential probability density:

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

2.3 MS-SMC' model description and notation

Under the SMC' model, sampling of a linked genealogy requires considering not only coalescent 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 an edge, separating a subtree below the cut from the rest of the genealogy (Fig. 2). The subtree must then be reconnected by sampling a new common ancestor between it and the other samples still connected to the genealogy, which must occur above the time of the cut. This re-connection point is sampled probabilistically with an expected waiting time until re-coalescence the same as described above (equations 1 and 2).

In a single population model with constant N the re-coalescent probability after a recombination event decreases monotonically, as each subsequent coalescence 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 effective population sizes can also vary among species tree intervals. Thus, the probability that a subtree re-connects to the genealogy can vary along its path of possible recon-

nection points through different species tree intervals. A species tree can thus be decomposed into a series of relevant intervals between events that change the rate of coalescence which we refer to as the gene tree embedding table (Table 1).

Each branch (b) of $\mathcal{G}$ spans one or more intervals from which a lower and upper (min, max) bound for that edge can also be extracted (t_b^l and t_b^u, respectively). The time of the start of each interval (i) is σ_i. An integer index ($\mathcal{I}$) stores the number of intervals that each branch occurs in, such that enumerating over the range of ($\mathcal{I}$) can visit each interval of a genealogy branch. These variables, and the gene tree embedding table, include all relevant parameters for the derivations below.

To better understand the relationship between the MS-SMC' and the gene tree embedding table, let us consider edge 7 as an example. For this edge, we are interested in defining intervals where coalescent rate parameters change. To do so, we must identify times at which there are relevant coalescent events between other lineages in the same population, or at which there are relevant species merging events. This involves examining the genealogical tree as it is embedded within the species tree. With the gene tree embedding table, we can easily isolate the intervals specific to edge 7 by querying the "branches" column of the table and extracting the rows where this column contains our focal edge (for edge 7, these are rows 1, 6, 7, and 8). The number of rows we extract (in this case, 4) is equal to the number of intervals in the edge (I_b). If we retain the relative order of the rows and re-index them from 0 to $I_b - 1$, the information in the columns (e.g. n_i, a_i) is easily adapted for use in the equations presented below.

To summarize, the gene tree embedding table is simply a set containing all unique intervals (and their properties) over all edges in the genealogy, and it maps those intervals to the genealogy edges to which they belong. This allows us to easily query relevant intervals for any edge of the genealogical tree. For extracting relationships among branches (e.g. parent, sibling, as required for the topology cahnge equations), we can reference a separate table specific to the structure of the genealogical tree only (see Appendix, Table 2).

Table 1. A gene tree embedding table for MS-SMC' calculations for the gene tree and species tree in Figure 1.

	start (t_i^l)	end (t_i^u)	pop	N (n_i)	N edges (a_i)	coal	length (d_i)	branches
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

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

A recombination can result in one of four categories of outcomes (Fig. 2), resulting in either no change to the genealogy, a change only to its branch lengths, or a change to its topology. Our first step in deriving probability statements for these different outcomes is to calculate the probability of no change. Then, from the law of total probability we can calculate probabilities of changes that fall into the other categories.

2.4.1 Probability that recombination at t_r on branch b does not change the tree:

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. The index i refers to the gene tree embedding interval in which recombination occurs. For no genealogy change to occur the detached branch must re-attach to its original lineage during the same interval in which it detached. If it connects to any other existing lineages during this same interval, or coalesces in a later interval, this would lead to a change in either the tree or topology. A derivation of this equation can be found in the Appendix. The first two terms correspond to the probability of detaching from the interval containing t_r and re-connecting to the same branch within it (i.e., ii), whereas the later terms describe the probability of re-connecting in a different interval on the same branch (i.e., ij).

$$\mathbb{P}(\text{tree unchanged}|\mathcal{S},\mathcal{G},b,t_r) = \left(\frac{1}{a_i} + P_{ii}\exp\left\{\frac{a_i}{n_i}t_r\right\}\right) + \sum_{j=i+1}^{\mathcal{I}_b-1} P_{ij}\exp\left\{\frac{a_i}{n_i}t_r\right\} \tag{3}$$

This includes the following two constant values.

$$P_{ii} = -\frac{1}{a_i}\exp\left\{-\frac{a_i}{n_i}\sigma_{i+1}\right\} \tag{4}$$

$$P_{ij} = \left(\frac{1}{a_j}(1-\exp\left\{-\frac{a_j}{n_j}d_j\right\}\right) \times \exp\left\{-\frac{a_i}{n_i}\sigma_{i+1} - \sum_{q=i+1}^{j-1}\frac{a_q}{n_q}d_q\right\} \tag{5}$$

2.4.2 Probability recombination on branch b will not change the tree:

By integrating the previous equation 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 on this branch does not change the tree:

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

$$= \frac{1}{t_b^u - t_b^l}\sum_{i=0}^{\mathcal{I}_b-1}\frac{1}{a_i}d_i + \frac{n_i}{a_i}\left(\exp\left\{\frac{a_i}{n_i}\sigma_{i+1}\right\} - \exp\left\{\frac{a_i}{n_i}\sigma_i\right\}\right)\left(\sum_{j=i}^{\mathcal{I}_b-1}P_{ij}\right) \tag{6}$$

2.4.3 Probability that a recombination event will not change the tree:

Finally, by summing across all branches on the tree while assuming a uniform recombination rate through time and across branches, we get the probability that a recombination event will result in a category 1 outcome.

$$\mathbb{P}(\text{tree unchanged}|\mathcal{S},\mathcal{G}) = \sum_{b\in\mathcal{G}} \left[\frac{t_b^u - t_b^l}{L(\mathcal{G})}\right] \mathbb{P}(\text{tree unchanged}|\mathcal{S},\mathcal{G},b) \tag{7}$$

2.4.4 Waiting distance to the next tree change:

Under the SMC' the rate at which recombination events occur (λ_r) is a product of the per-site per-generation recombination rate (r) and total branch lengths of the current genealogy.

$$\lambda_r = L(\mathcal{G}) \times r \tag{8}$$

And thus a waiting distance (x) to the next recombination event is distributed as an exponential probability density.

$$f(x) = \lambda e^{-\lambda x} \tag{9}$$

We can then modify this equation to instead yield a probability density for waiting distances to a recombination event that causes a tree change. Here, distances continue to be exponentially distributed, however, the new rate parameter (λ_g) is reduced proportionally by the probability that a recombination does not change the tree.

$$\lambda_g = L(\mathcal{G}) \times r \times (1 - \mathbb{P}(\text{tree unchanged}|\mathcal{S},\mathcal{G})) \tag{10}$$

2.5 Deriving probabilities of changes in the genealogical topology

We can next derive an analogous probability density for the waiting distance to a change in the *topology* of a genealogy. This involves excluding events of categories 2 and 3 to isolate the waiting distance to a category 4 event. Once again, we start with a branch-specific probability and then sum across all branches on the tree. In order to isolate tree changes that only affect edge lengths from those that affect the topology we must take into account which specific branches the detached subtree from branch b re-coalesces with. The two relevant branches are its sibling (b') and parent (c). If it re-coalesces with a sibling the topology remains the same but with a shortened coalescent time, and if it re-coalesces with its parent then its original coalescent time is simply lengthened.

Retaining the correct order of intervals is important in these calculations, particularly for P_{ij}, which involves summing over not only the information in the i and j intervals, but also all of intervals that lie between them. If we define bc as the ordered union of the sets of intervals on branches b and c, then $\mathcal{I}_{bc}$ is the length of intervals that can be used to index ordered intervals in this set. Similarly, we define the intersection of the sets of intervals in b and b' as bb', which includes only intervals that both of these branches occur in (i.e., it excludes intervals during which the branches are embedded in separate species tree branches).

We then define m as the index of the lowest interval in bb' (occurring at time t_b^m) such that by indexing from m to I_b we can visit all intervals that are shared by the sibling branches.

2.5.1 Probability recombination at t_r on branch b does not change the topology:

Once again we begin by assuming knowledge of the branch on which a recombination event occurs and its timing. Unlike the previous approach, where we simply calculated the probability of a category 1 event, we now calculate the probability of an event falling into categories 1, 2, or 3. To do this, we must immediately break the problem into two different cases: when t_r occurs below t_b^m, and when it occurs above t_b^m.

First case – Given $t \in [\sigma_i, \sigma_{i+1}] \subset [t_b^l, t_b^m]$: In this case, we have integrated coalescence probabilities through three sections in which a re-coalescence can produce a category 1, 2, or 3 event: the lower part of branch b when the disconnected branch can reconnect with itself, the upper part of branch b when it can reconnect with either itself or with its sibling, and all of branch c, with which a coalescent would lengthen branches without changing the topology (Fig. S1a).

$$\mathbb{P}(\text{topology unchanged}|\mathcal{S}, \mathcal{G}, b, t) = \frac{1}{a_i} + \sum_{k=i}^{\mathcal{I}_{b+c}-1} P_{ik} \exp\left\{\frac{a_i}{n_i} t\right\} + \sum_{k=m}^{\mathcal{I}_b-1} P_{ik} \exp\left\{\frac{a_i}{n_i} t\right\} \tag{11}$$

Second case – Given $t \in [\sigma_i, \sigma_{i+1}] \subset [t_b^m, t_b^u]$: If t_r occurs after t_b^m then the detached branch can immediately coalesce with either the original branch or its sibling, and so there are only two distinct sections of intervals we must integrate across: the first is when the detached branch can re-coalesce with itself or its sibling, and the second when it can coalesce with its parent (Fig. S1b).

$$\mathbb{P}(\text{topology unchanged}|\mathcal{S}, \mathcal{G}, b, t) = 2\left(\frac{1}{a_i} + \sum_{k=i}^{\mathcal{I}_b-1} P_{ik} \exp\left\{\frac{a_i}{n_i} t\right\}\right) + \sum_{\mathcal{I}_b}^{\mathcal{I}_{b+c}-1} P_{ik} \exp\left\{\frac{a_i}{n_i} t\right\} \tag{12}$$

2.5.2 Probability that a recombination event on branch b will not change the topology:

We next present the probability that a recombination event occurring on a focal branch b will change the tree topology. As in Section 2.4, this results from integrating through the solution that specifies both a specific branch and time.

$$\mathbb{P}(\text{topology unchanged}|\mathcal{S},\mathcal{G},b) = \frac{1}{t_b^u - t_b^l}\left[\sum_{i=0}^{m-1} p_{b,1}^{(i)} + \sum_{i=m}^{\mathcal{I}_b-1} p_{b,2}^{(i)}\right]$$

$$where:$$

$$p_{b,1}^{(i)} = \frac{1}{a_i}\left[d_i + n_i\left(\exp\left\{\frac{a_i}{n_i}\sigma_{i+1}\right\} - \exp\left\{\frac{a_i}{n_i}\sigma_i\right\}\right)\left(\sum_{k=i}^{\mathcal{I}_{b+c}-1} P_{ik} + \sum_{k=m}^{\mathcal{I}_b-1} P_{ik}\right)\right] \tag{13}$$

$$and:$$

$$p_{b,2}^{(i)} = \frac{1}{a_i}\left[2d_i + n_i\left(\exp\left\{\frac{a_i}{n_i}\sigma_{i+1}\right\} - \exp\left\{\frac{a_i}{n_i}\sigma_i\right\}\right)\left(2\sum_{k=i}^{\mathcal{I}_b-1} P_{ik} + \sum_{k=\mathcal{I}_b}^{\mathcal{I}_{b+c}-1} P_{ik}\right)\right]$$

2.5.3 Probability that a recombination event will not change the topology:

Finally, we can simply sum the previous equation across all branches in the genealogy to determine the probability that a recombination event falling uniformly on the tree will change the topology.

$$\begin{aligned}\mathbb{P}(\text{topology unchanged}|\mathcal{S},\mathcal{G}) &= \sum_{b\in\mathcal{G}} \frac{t_b^u - t_b^l}{L(\mathcal{G})} \times \mathbb{P}(\text{topology unchanged}|\mathcal{S},\mathcal{G},b) \\ &= \frac{1}{L(\mathcal{G})}\sum_{b\in\mathcal{G}}\left[\sum_{i=0}^{m-1} p_{b,1}^{(i)} + \sum_{i=m}^{\mathcal{I}_b-1} p_{b,2}^{(i)}\right]\end{aligned} \tag{14}$$

2.5.4 Waiting distance to the next topology change:

Unlike with the expected waiting distance to a *tree* change, the exact waiting distance distribution to a *topology* change is more complicated to derive, because of possible intermediate recombination events that change the branch lengths but not the topology. As the branch lengths of intermediate genealogies change this affects the rate of subsequent recombination events, and thus the probability that such events could change the topology of the next genealogy. This problem similarly arises in a single population model, where Deng *et al.* (2021) demonstrated the its effect can be ignored. This is because of the inverse relationship between genealogy length and the probability of a topology-changing recombination event. Even as branch lengths increase, reducing the waiting time to the next recombination event, the probability that this event will change the topology decreases. Therefore, we follow their approach by approximating the waiting distance to a change in topology using an exponential probability density with the following rate:

$$\lambda_t = L(\mathcal{G}) \times r \times (1 - \mathbb{P}(\text{topology unchanged}|\mathcal{S},\mathcal{G})) \tag{15}$$

3 Demonstration

We have implemented our solutions in the Python package *ipcoal* (McKenzie & Eaton, 2020), which is an MSC-focused tool that includes a wrapper around the coalescent simulation software *msprime*

(Baumdicker *et al.*, 2022) paired with tree visualizations from *toytree* (Eaton, 2020). Here we demonstrate the accuracy of our solutions relative to simulated tree sequences. Source code is available at `https://github.com/eaton-lab/ipcoal` with reproducible examples implemented in jupyter notebooks.

First, given a species tree model and initial genealogy, we can visualize the probabilities of different types of outcomes of recombination as a function of the position (time) on a genealogy branch where a recombination event occurs (Fig. 1). As an example, we show probabilities along three selected edges of a genealogy embedded in a species tree model. Our species tree was proposed arbitrarily with a root height of 1e6 generations and includes branch-specific N values. Our sampling in the genealogy includes three individuals for population A, two individuals for population B, and one individual for each population C and D. The exact parameterization for the model, as well as the code to replicate the analysis, is documented in our supplementary notebooks. Fig. 1 panels b, c, and d show the probability that a recombination event falling at each time t_r on the 3 selected genealogy branches results in no change (b), results in a tree change (c), and results in a topology change (d). The tendency of the y-values in (b) to approach 0 and (c) to approach 1 as x-values increase make clear that recombination events falling at the tops of branches (i.e., far right on the x-axis) are increasingly certain to change the branch lengths of the genealogy. This is because any re-coalescence of the detached branch can only occur above the recombination breakpoint. The values in part (d), however, never approach 1 (i.e. certain to produce a topology change). This is because re-coalescence with the parent branch, which would not result in a topology change, is still possible even if a recombination event occurs at the very top of a focal branch.

Next, we tested the effect of variation in basic species tree parameters on expected waiting distances to tree changes and topology changes (Fig. 3). We started with four different species trees: two imbalanced and two balanced. For each pair, we used a large tree (10 tips), and a small tree (5 tips) that was pruned from the large tree, with internode distances kept the same. On each species tree, we generated 1000 random *unlinked* genealogies using MSC probabilities and using each of 9 evenly spaced Ne values, ranging from 50000 to 250000. We calculated the expected waiting distance to a tree change and to a topology change for each genealogy/species tree pair, assuming a recombination rate of 1e-9 recombs/bp/generation. In Fig. 3, we show the mean expected waiting distances to tree and topology changes for each species tree and for each Ne value. The decline in expected waiting distances across Ne values and from small to large trees demonstrates that, on average, waiting distances to tree and topology changes decrease with increasing numbers of tips sampled and with increasing Ne values.

Finally, we demonstrated that a known sequence of genealogies paired with the accompanying observed waiting distances to next topological change for each can be used to estimate the Ne value for species trees with different topologies (Fig. 4). We used two different species trees: one imbalanced, 5-tip tree with internode branches of equal length, and one 10-tip tree with an irregular topology and branches of variable lengths. Both species trees were assigned equal root heights of 1e6 generations and equal Ne values of 150000 on all branches, and we simulated a 5e6-bp chromosome for each using *ipcoal*. We broke each chromosome down into its component "initial genealogies" – those that start each segment bounded by changes in topology – and the accompanying segment lengths (i.e. waiting distances). Using these values and a known recombination rate of 1e-9 recombs/bp/generation, we calculated the likelihood that each of 41 proposed Ne values produced the set of genealogies and waiting distances. Using this approach, we

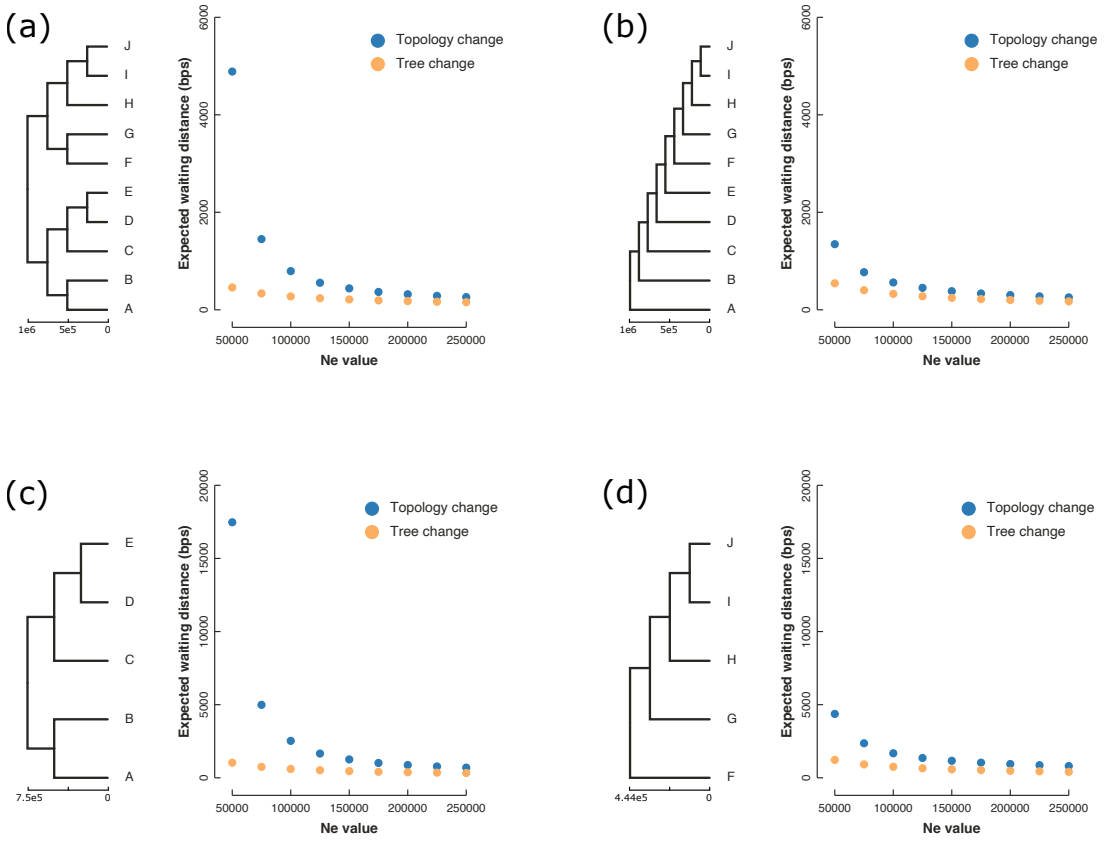

Figure 3. Mean expected waiting distances for differently parameterized species tree models. We used four different species trees: two 10-tip trees (a, b), each with 1e6-generation root heights, and two 5-tip trees (c, d) that are subselected from each of the larger trees. For each species tree, we iterated over 9 Ne values ranging from 50000 to 250000, generating 1000 unlinked MSC genealogies at each Ne value and calculating the expected waiting distance to a tree change and to a topology change for each. The mean of the expected waiting distances at each Ne value is presented in the scatterplot accompanying each species tree. Note that the y-axis scale changes from the top row (a, b) to the bottom row (c, d).

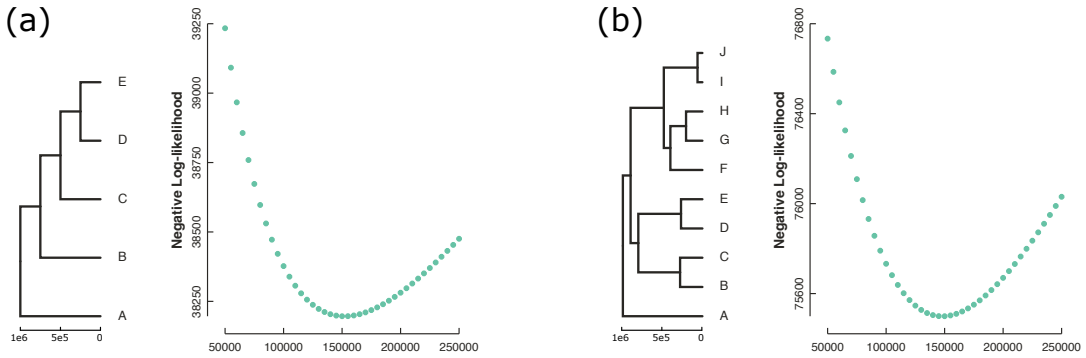

Figure 4. Inference of species tree Ne values from genealogies, observed waiting distances to topology changes, and recombination rate. We generated two different species trees: Both have root heights of 1e6 generations, but one has five tips, an unbalanced topology, and identical internode lengths, while the other has ten tips, an irregular topology, and irregular branch lengths. Using a recombination rate of 1e-9 recombs/bp/generation and a constant Ne value of 150000 over all branches, we generated a 5MB tree sequence from each species tree using *ipcoal*. Then, we decomposed the alignment into segments bounded by topology changes in the genealogies, and we recorded the initial genealogy for each segment. Finally, we calculated the likelihood of the set of genealogies and waiting distances, along with the recombination rate, at 41 different proposed Ne values. For both sequence alignments, 150000 was correctly inferred as the Ne value.

recovered a correctly inferred Ne value of 150000 for both trees.

4 Conclusions

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.

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.

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.

4.0.1 Acknowledgements

This work was supported by the National Science Foundation (NSF DEB-2046813 awarded to D.A.R.E. and NSF GRFP awarded to P.F.M.). Thanks to ... for feedback on ...

References

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

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, 10, 16, 17, 19, 20

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. 3

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. 3

Kelleher, J., Etheridge, A.M. & McVean, G. (2016). Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS computational biology*, 12, e1004842. 3

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

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. 14

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. (2020). ipcoal: an interactive Python package for simulating and analyzing genealogies and sequences on a species tree or network. *Bioinformatics*. 10

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

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, 4

5 Appendix 1: Genealogy Table

6 Appendix 2: Step-by-Step Solution

6.1 Notation Summary

6.2 Notation

Assume we have sampled from a species tree $\mathcal{S}$ with $\mathcal{N}$ tips, where each tip represents a sampled individual belonging to one of the species. The species tree consists of a topology describing the history of population divergences, where each branch of the topology has a constant effective diploid population size N_b associated with it. Within each branch, coalescence occurs at a constant per-generation rate of $\frac{1}{2N_b}$.

We begin with a genealogical tree $\mathcal{G}$ sampled from the species tree according to coalescent probabilities. This tree is embedded within the species tree so that the time of coalescence for any two individuals

Table 2. A table summarizing the relationships among branches in the genealogical tree embedded in the species tree in Figure 1.

Branch	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}

from different species in $\mathcal{G}$ is constrained to occur farther back in time than their species' coalescence times in $\mathcal{S}$.

In Deng *et al.* (2021), the genealogical tree was broken into a series of intervals so that the number of remaining (not-coalesced) lineages was constant within each interval. In their method, the number of remaining lineages decreases monotonically through time from tipward to rootward intervals as lineages coalesce. Our approach is similar, except that the intervals are branch-specific and correspond to intervals of constant numbers of lineages to coalesce with (A) and constant effective population size (N). Breakpoints may therefore exist where divergences occur in the species tree (potentially changing both A and N) and where coalescent events occur in the genealogical tree (reducing A). Unlike in Deng *et al.* (2021), A is no longer monotonic, since *species* coalescent events can increase the number of lineages available for coalescence, while *genealogical* coalescent events always reduce that number.

The intervals within each branch of $\mathcal{T}$ are each assigned an index increasing from 0 to $\mathcal{I}_b - 1$, where $\mathcal{I}_b$ is the total number of intervals in the branch. The times in generations marking the lower and upper bounds of each branch b are notated t_l^b and t_u^b, respectively. Each interval of index x is bounded by times σ_x and σ_{x+1}.

We begin with a genealogical tree $\mathcal{G}$ sampled from the species tree according to coalescent probabilities. This tree is embedded within the species tree so that the time of coalescence for any two individuals from different species in $\mathcal{G}$ is constrained to occur farther back in time than their species' coalescence times in $\mathcal{S}$.

6.2.1 Given a branch and a time

We begin by assuming a specific branch and time of a recombination event. We then integrate across possible times on that branch, and we sum across all branches on the tree to solve for the probability of the genealogy being unchanged given any recombination event. Finally, we incorporate this probability into the exponential distribution presented in Equation 4.

Table 3. Summary of variables used in waiting distance equations.

Variable	Description
i	Integer-valued interval index to which t_r belongs
a_x	The value of $A(t)$ for all t_r in interval x
n_x	The value of $N(t_r)$ for all t_r in interval x
σ_x	The lower boundary, in generations, of interval x
d_x	The length, in generations, of the interval x – i.e., $\sigma_{x+1} - \sigma_x$
t_l^b	The lower boundary, in generations, of branch b
t_u^b	The upper boundary, in generations, of branch b
t_m^b	The time at which a focal branch b is able to coalesce with its sibling branch
m	Integer-valued interval index whose lower boundary is t_m^b
$\mathcal{I}_b$	The number of intervals on branch b
$\mathcal{G}$	The genealogical tree
$\mathcal{S}$	The species tree with divergence times and branch-specific effective population sizes
$L(\mathcal{G})$	The total length of branches in genealogy $\mathcal{G}$

The intervals within each branch of $\mathcal{G}$ are each assigned an index increasing from 0 to $\mathcal{I}_b - 1$, where $\mathcal{I}_b$ is the total number of intervals in the branch. The times in generations marking the lower and upper bounds of each branch b are notated t_l^b and t_u^b, respectively. Each interval of index x is bounded by times σ_x and σ_{x+1}.

Where $A(\tau)$ is the number of lineages able to be coalesced with at any time τ, and $p(\tau|t_r)$ is the exponential probability density of coalescing any time after the recombination event. Thus, we are integrating over the probability of coalescence along the path of species tree intervals traversed by branch b, starting at the time of recombination and ending at the end of the branch. If the branch spans only a single gene tree embedding interval then $A(\tau)$ and $N(\tau)$

Through this series of one or more gene tree embedding intervals the

up to that time through the species tree intervals that are traversed by branch b, which can involve changing numbers of samples over time (A(t)), and changing effective population sizes (N(t)). Here we define the coalescence rate at time τ as $\frac{A(\tau)}{N(\tau)}$. Note that this differs from Deng *et al.* (2021), since our branch lengths are in units of generations rather than in coalescent units.

$$\mathbb{P}(\text{tree unchanged}|b, t, \mathcal{G}, \mathcal{S}) = \int_t^{t_b^u} \frac{1}{A(\tau)} p(\tau|t) d\tau \tag{16}$$

$$= \int_t^{t_b^u} \frac{1}{A(\tau)} \frac{A(\tau)}{N(\tau)} e^{-\int_t^\tau \frac{A(s)}{N(s)} ds} d\tau \tag{17}$$

$$= \int_t^{t_b^u} \frac{1}{N(\tau)} e^{-\int_t^\tau \frac{A(s)}{N(s)} ds} d\tau \tag{18}$$

We can now take advantage of the fact that the rate of coalescence is piecewise-constant. We split the equation into one interval i whose length depends on t (it ranges from t until the upper limit of its interval, σ_{i+1}) and we add to it integrals across intervals that are above it on the same branch, which are predefined by their upper and lower σ values:

$$= \int_t^{\sigma_{i+1}} \frac{1}{N(\tau)} \exp\left(-\int_t^{\tau} \frac{A(s)}{N(s)} ds\right) d\tau + \sum_{k=i+1}^{\mathcal{I}_b - 1} \int_{\sigma_k}^{\sigma k+1} \frac{1}{N(\tau)} \exp\left(-\int_t^{\tau} \frac{A(s)}{N(s)} ds\right) d\tau \quad (19)$$

We solve first term and later terms separately:

First term –

$$= \frac{1}{a_i} - \frac{1}{a_i} e^{-\frac{a_i}{n_i}\sigma_{i+1}} e^{\frac{a_i}{n_i}t} \quad (20)$$

$$= \frac{1}{a_i} + P_{ii} e^{\frac{a_i}{n_i}t} \quad (21)$$

Later terms –

$$= \sum_{k=i+1}^{\mathcal{I}_b - 1} e^{\frac{a_i}{n_i}t} \exp\left(-\frac{a_i}{n_i}\sigma_{i+1} - \sum_{q=i+1}^{k-1} \frac{a_q}{n_q} T_q\right) \left(\frac{1}{a_k}(1 - e^{-\frac{a_k}{n_k}T_k})\right) \quad (22)$$

$$= \sum_{k=i+1}^{\mathcal{I}_b - 1} e^{\frac{a_i}{n_i}t} P_{ik} \quad (23)$$

Adding these together, we are left with the solution:

$$\mathbb{P}(\text{tree unchanged}|b, t, \mathcal{G}, \mathcal{S}) = \frac{1}{a_i} + \sum_{k=i}^{\mathcal{I}_b - 1} P_{ik} e^{\frac{a_i}{n_i}t} \quad (24)$$

6.2.2 Across a full branch

Having solved for the probability of the genealogical tree being unchanged given the time t of the recombination event, our next step is to integrate this equation across the branch with respect to t:

$$\mathbb{P}(\text{tree unchanged}|b, \mathcal{G}, \mathcal{S}) = \frac{1}{t_b^u - t_b^l} \int_{t_b^l}^{t_b^u} \mathbb{P}(\text{tree unchanged}|b, t, \mathcal{G}, \mathcal{S}) dt \quad (25)$$

Plugging in Equation 17 above, we find:

$$= \frac{1}{t_b^u - t_b^l} \sum_{i=0}^{\mathcal{I}_b - 1} \frac{1}{a_i} d_i + \left(e^{\frac{a_i}{n_i}\sigma_{i+1}} - e^{\frac{a_i}{n_i}\sigma_i}\right) \left[-\frac{n_i}{a_i^2} e^{-\frac{a_i}{n_i}\sigma_{i+1}} + \frac{n_i}{a_i} \left(\sum_{k=i+1}^{\mathcal{I}_b - 1} \exp\left(-\frac{a_i}{n_i}\sigma_{i+1} - \sum_{q=i+1}^{k-1} \frac{a_q}{n_q} T_q\right) \frac{1 - \exp(-\frac{a_k}{n_k}T_k)}{a_k}\right.\right. \quad (26)$$

6.2.3 Across the whole tree

At last, we can calculate the probability that, given a recombination event, the genealogical tree 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:

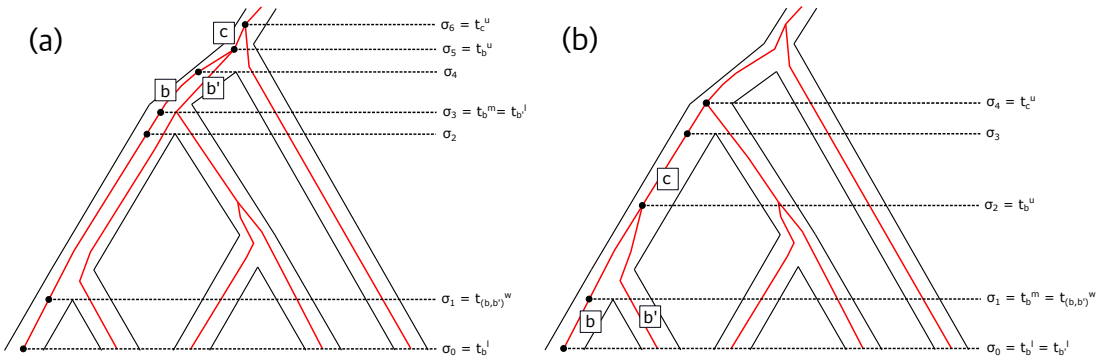

Figure 5. Illustrating the parameters for calculating the distribution of distances to a change in the topology of a genealogy. Panels (a) and (b) show slightly different genealogies embedded in the same species tree. In both (a) and (b), the focal branch b is the one leading from the left-most species. Branches c and b' – the parent and sibling branches, respectively – are also both labeled. Note that in (a), t_b^m corresponds to σ_3, while in (b) it corresponds to σ_1.

$$\mathbb{P}(\text{tree unchanged}|\mathcal{G},\mathcal{S}) = \sum_{b=1}^{2N-2}\left[\frac{t_b^u - t_b^l}{L(\mathcal{G})}\right]\mathbb{P}(\text{tree unchanged}|b,\mathcal{G},\mathcal{S}) \tag{27}$$

To derive the waiting distance to the next tree change, we incorporate the value of $1-\mathbb{P}(\text{tree unchanged}|\mathcal{G},\mathcal{S})$ into the exponential distribution describing the waiting distance to the next recombination event (introduced in Equation 4):

$$\begin{aligned} &p_r(d|\mathcal{G},\mathcal{S}) = r\alpha_{\mathcal{S}}(\mathcal{G})L(\mathcal{G})\exp\left[-r\alpha_{\mathcal{S}}(\mathcal{G})L(\mathcal{G})d\right], \\ &where \\ &\alpha_{\mathcal{S}}(\mathcal{G}) = 1 - \mathbb{P}(\text{tree unchanged}|\mathcal{G},\mathcal{S}). \end{aligned} \tag{28}$$

6.3 The distribution of distances to a change in genealogical topology

Next, we attempt to further exclude recombination events. We filter out recombination events of categories 2 and 3, which just change branch lengths, to isolate the probability that a recombination event changes the *topology* of the tree.

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 6).

$$\mathbb{P}(\text{topology unchanged}|b,\mathcal{G},\mathcal{S}) = \frac{1}{t_b^u - t_b^l}\int_{t_b^l}^{t_b^u}\mathbb{P}(\text{topology unchanged}|b,t,\mathcal{G},\mathcal{S})dt \tag{29}$$

$$= \frac{1}{t_b^u - t_b^l} \left[\left(\int_{t_b^l}^{t_b^m} + \int_{t_b^m}^{t_b^u} \right) \mathbb{P}(\text{topology unchanged}|b, t, \mathcal{G}, \mathcal{S}) dt \right] \tag{30}$$

Where $t_b^m = \max\{t^w_{(b,b')}, t^l_{b'}, t^l_b\}$.

6.3.1 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.

First case – Given $t \in [\sigma_i, \sigma_{i+1}] \subset [t_b^l, t_b^m]$:

$$\mathbb{P}(\text{topology unchanged}|b, t, \mathcal{G}, \mathcal{S}) = \int_t^{t_b^m} \frac{1}{A(\tau)} \rho(\tau|t) d\tau + \int_{t_b^m}^{t_b^u} \frac{2}{A(\tau)} \rho(\tau|t) d\tau + \int_{t_b^u}^{t_c^u} \frac{1}{A(\tau)} \rho(\tau|t) d\tau$$

$$= \frac{1}{a_i} + \sum_{k=i}^{\mathcal{I}_{b+c}-1} P_{ik} e^{\frac{a_i}{n_i} t} + \sum_{k=m}^{\mathcal{I}_b - 1} P_{ik} e^{\frac{a_i}{n_i} t} \tag{31}$$

The second case is solved in a similar manner, although we eliminate the first integral:

Second case – Given $t \in [\sigma_i, \sigma_{i+1}] \subset [t_b^m, t_b^u]$:

$$\mathbb{P}(\text{topology unchanged}|b, t, \mathcal{G}, \mathcal{S}) = \int_t^{t_b^u} \frac{2}{A(\tau)} \rho(\tau|t) d\tau + \int_{t_b^u}^{t_c^u} \frac{1}{A(\tau)} \rho(\tau|t) d\tau$$

$$= 2 \left(\frac{1}{a_i} + \sum_{k=i}^{\mathcal{I}_b - 1} P_{ik} e^{\frac{a_i}{n_i} t} \right) + \sum_{\mathcal{I}_b}^{\mathcal{I}_{b+c}-1} P_{ik} e^{\frac{a_i}{n_i} t} \tag{32}$$

6.3.2 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 –

$$\int_{t_b^l}^{t_b^m} \mathbb{P}(\text{topology unchanged}|b, t, \mathcal{G}, \mathcal{S}) = \sum_{i=0}^{m-1} \frac{1}{a_i} \left[d_i + n_i \left(\exp\left(\frac{a_i}{n_i} \sigma_{i+1} \right) - \exp\left(\frac{a_i}{n_i} \sigma_i \right) \right) \left(\sum_{k=i}^{\mathcal{I}_{b+c}-1} P_{ik} + \sum_{k=m}^{\mathcal{I}_b - 1} P_{ik} \right) \right] \tag{33}$$

Second case –

$$\int_{t_b^m}^{t_b^u} \mathbb{P}(\text{topology unchanged}|b,t,\mathcal{G},\mathcal{S}) = \sum_{i=m}^{\mathcal{I}_b-1} \frac{1}{a_i}\left[2d_i + n_i\left(\exp\left(\frac{a_i}{n_i}\sigma_{i+1}\right) - \exp\left(\frac{a_i}{n_i}\sigma_i\right)\right)\left(2\sum_{k=i}^{\mathcal{I}_b-1} P_{ik} + \sum_{k=\mathcal{I}_b}^{\mathcal{I}_{b+c}-1} P_{ik}\right)\right. \tag{34}$$

Result –

$$\mathbb{P}(\text{topology unchanged}|b,\mathcal{G},\mathcal{S}) = \frac{1}{t_b^u - t_b^l}\left[\sum_{i=0}^{m-1} p_{b,1}^{(i)} + \sum_{i=m}^{\mathcal{I}_b-1} p_{b,2}^{(i)}\right] \tag{35}$$

6.3.3 Across the whole tree

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

$$\begin{aligned}\mathbb{P}(\text{topology unchanged}|\mathcal{G},\mathcal{S}) &= \sum_{b\in\mathcal{G}} \frac{t_b^u - t_b^l}{L(\mathcal{G})} \times \mathbb{P}(\text{topology unchanged}|b,\mathcal{G},\mathcal{S}) \\ &= \frac{1}{L(\mathcal{G})}\sum_{b\in\mathcal{G}}\left[\sum_{i=0}^{m-1} p_{b,1}^{(i)} + \sum_{i=m}^{\mathcal{I}_b-1} p_{b,2}^{(i)}\right]\end{aligned} \tag{36}$$