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

Abstract–

Under neutrality, genomes are expected to be mosaics of many genealogical trees, each separated by an ancestral recombination event. In phylogenetics, the expected discordance between *unlinked* parts of the genome due to this phenomenon is well-known, and inference under the multispecies coalescent model (MSC) has become popular to account for this expected discordance. However, the amount of discordance expected between *linked* parts of the genome, explicitly considering the effect of recombination, is not well characterized. Recently, Deng et al. (2021) presented equations describing the distribution of waiting distances between changes in genealogical trees along the genome for single-population models with constant effective population sizes (Ne). Here, we generalize their results to yield the distribution of waiting distances between changes in genealogical trees and topologies given a species tree model with an arbitrary topology and arbitrary, branch-specific Ne values. These solutions help establish an expectation for the expected amount of linkage between nearby regions in a genome while providing a new source of comparison to evaluate the fit of a species tree to sequence data.

1 Introduction

The multispecies coalescent (MSC) is an extension of the coalescent (Kingman, 1982), a model describing the distribution of genealogical relationships among a set of samples at some region of the genome. Whereas the coalescent models a single panmictic population, the MSC includes constraints that prevent samples from different species lineages from sharing a most recent genealogical ancestor until prior to a divergence event that separates the species (Maddison, 1997; Maddison and Knowles, 2006). Analytically, the MSC can be viewed as a piecewise likelihood function composed of the standard coalescent applied to each edge of a "species tree", representing a hierarchical model of the relationships and divergence times among constraining lineages – i.e., it can be used to calculate the likelihood of a container tree within which a distribution of genealogies must be embedded (Rannala and Yang, 2003; Degnan and Rosenberg, 2009). In both the coalescent and MSC models, effective population size (Ne) is the key parameter determining the rate of coalescence, and it can vary over time and/or among different lineages.

Importantly, both the coalescent and MSC are models of the expected distribution of *unlinked* (non-autocorrelated) genealogies sampled from throughout the genome. By contrast, two linked genealogies that are drawn from nearby regions of the 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. In this way, the effect of recombination on the coalescent can be viewed as a process where the set of samples at the present might substitute one ancestor for another on either side of the recombination breakpoint. A simulation algorithm to approximate the coalescent with recombination was developed early on (Hudson, 1983), however only with recent advances has it become possible to scale coalescent simulations with recombination to larger, genome-wide scales (Kelleher et al., 2016).

While our ability to simulate the coalescent with recombination has made substantial progress, likelihood-based inference of recombination events and coalescent times (that is, the ancestral recombination graph – "ARG") is notoriously difficult (Griffiths and Marjoram, 1996). Full inference of the ARG is computationally intensive, highly limited by sequence information, and nearly infinite in possibilities for large genome lengths and numbers of samples (McVean and Cardin, 2005). However, by modeling the coalescent with recombination using a Markovian approximation called the Sequentially Markov Coalescent (SMC), ARG inference becomes more tractable. Under the SMC, the likelihood of a sequential set of genealogies and recombination breakpoints (i.e., an ARG) is computed as the product of the set of independent likelihoods of observing each change from one genealogy to the next, given an effective population size and recombination rate. While explicit likelihood-based inference is still very difficult, the SMC has enabled several new statistical methods that utilize information about recombination events, and the similarities among sequential genealogies, to make historical inferences about changes in effective population sizes (Li and Durbin, 2011), recombination rates, and selection (Rasmussen et al., 2014; Hubisz and Siepel, 2020).

Despite the many advances that have been developed for incorporating recombination into coalescent modeling, the MSC framework (and related fields in phylogenetics, such as the network multispecies coalescent (Degnan, 2018)) continues to largely treat recombination as a source of error as opposed to information. This stems from the widely known expectation that, in parts of parameter space, concatenation of sequences from multiple distinct genealogical histories can mislead phylogenetic inference. If a concatenated sequence is used to infer a single tree representing the species relationships, it might converge on a topology that is different from the species tree topology (Degnan and Rosenberg, 2006; Kubatko and Degnan, 2007). The same concatenation biases can affect the distribution of inferred gene trees at individual loci as well, with potential effects on species tree or network inference methods that take multiple gene trees as inputs – a process termed "concatalescence" (Gatesy and Springer, 2013). The extent to which genetic loci (e.g., exons, UCEs, genome windows) represent a single versus multiple genealogical histories can be estimated with coalescent simulations under an MSC model with recombination (McKenzie and Eaton, 2020b). However, analytical solutions have not been developed to describe these as distributions within a statistical phylogenetic framework.

Recently, Deng et al. (2021) derived a solution describing the distribution of waiting distances to genealogical changes in an ARG under the SMC' (Marjoram and Wall, 2006), a popular extension of the SMC that includes "invisible" recombination events that result in no change between neighboring genealogical trees. The solutions in Deng et al. (2021) assume a single population with a constant effective population size. Their solutions are an important advance for establishing a neutral expectation for turnover in genealogical topologies (and for connecting the coalescent parameters to these outcomes) and could be useful for ARG inference. While the expected waiting distance until any recombination event was

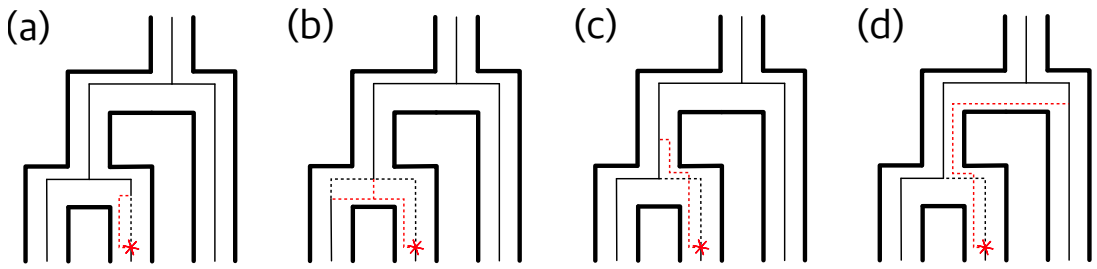

Figure 1. Examples of the four different categories of outcomes from a recombination event (adapted from (Deng et al., 2021)). (a) A category 1 event occurs when the dislocated lineage reattaches to the original lineage. (b) A category 2 event occurs when a dislocated lineage attaches to its sibling 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.

known previously, Deng et al. (2021) partitioned these recombination events into those that have no affect on the genealogical tree (category 1), into those that cause "tree changes" (i.e. that they affect either the branch lengths or the topology of the genealogical tree, as in any of categories 2, 3, and 4), and into those that cause "topology changes" (only category 4) (**Figure 1**).

Recombination events of type 4 are the most inherently interesting for most applications, because they will produce the most obvious signal in the sequence data. However, events of types 1-3 occur disproportionately often in small sample sizes (Wilton et al., 2015), which is especially common in MSC-based studies, where genealogies are embedded within a species tree, such that very few samples typically represent each lineage. But the relative frequency of the different events is contingent upon the Ne values of the species tree branches and on the species divergence times. In other words, the partitioning of coalescence events among lineages of the species tree topology is expected to constrain the types of recombination events that are more likely to be observed. This results in distributions of waiting distances between genealogical changes being highly dependent upon parameters of the species tree model.

With this in mind, we extend the method of Deng et al. (2021) to the MSC to predict the expected waiting distances between genealogy changes under a parameterized species tree model. This new solution is important for establishing an expectation for the neutral rate of genealogy turnover under different species tree parameterizations. It may also have applications for ARG inference under complex demographic models (e.g., Hubisz et al., 2020), potentially serving as a component of a likelihood function for inferring species trees from linked genealogies – a future theoretical goal.

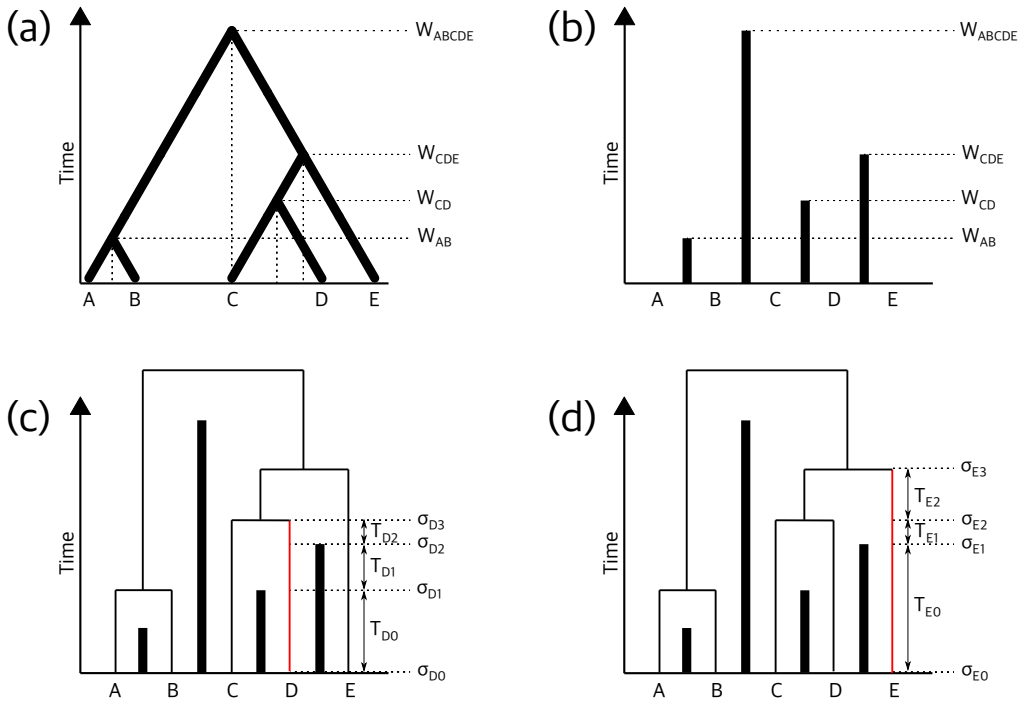

Figure 2. Illustrative depiction of the species tree model to introduce key parameters. a) A species tree topology with divergence times (W). Each branch has a characteristic effective population size (not shown). b) Alternative representation of the species tree that preserves topology and divergence times. This representation highlights how population divergences act as "walls" between populations, preventing coalescences between individuals from separate populations. c) A sampled genealogy embedded within the species tree, with focal branch D. The focal branch is broken into intervals with constant coalescence rates. The number of intervals in this branch, $\mathcal{I}_D$, is 3. d) The same genealogy, but with focal branch E.

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 genealogical changes for a single population of constant size. We modified the single-population model to include 1) the barriers to coalescence imposed by the species tree topology, and 2) the changing coalescence rates due to variable effective population sizes. We have intentionally reproduced our equations with the same structure, and many of the same variables, as Deng et al. (2021), which should help highlight the changes we made to generalize their solution.

2.2 Model description and notation

We assume a species tree model with divergence times and branch-specific effective population sizes as described by the MSC (Maddison 1997). Generating unlinked genealogical trees under the MSC is straightforward: Coalescent events between samples can only happen after (moving backward in time) co-

alescence of the species that contain them, and the coalescent rate in each part of the tree is dependent on both the number of genealogical lineages that can coalesce in that section and on the effective population size in that section.

Under the single-population SMC' considered by Deng et al. (2021), a recombination event occurs at any location, uniformly, on a focal genealogical tree. The recombination event detaches a lineage from the tree, and the effective population size of the population specifies the rate at which the detached lineage reattaches either to the same lineage or to another lineage moving backwards in time on the tree. In this single-population model, available lineages for coalescence decrease monotonically backwards in time. Because it is a panmictic population, all samples (or ancestors of samples) are available to coalesce with any others at all times. Finally, the effective population size remains constant. Therefore, the model is specified to produce piecewise-constant coalescence rates, with interval breakpoints designated where any coalescence event occurs (reducing the coalescence rate).

In our MSC adaptation of this framework, we also specify a model with piecewise-constant coalescence rates. However, intervals under the MSC adaptation are bounded not just by coalescences in the *genealogical* tree (decreasing the number of available lineages for coalescence), but also by relevant coalescences in the *species* tree. Species coalesence events increase the number of genealogical lineages available for coalescence, and they also can correspond to changes in effective population size (**Figure 2**).

Unlike in the single-population SMC', the intervals in the multispecies model are branch-specific. For example, the breakpoint σ_{D1} in Figure 2c corresponds to the species tree coalescence of species C and species D. This increases the number of lineages available for the D lineage to coalesce with (from 1 – itself – to 2) and might also change the effective population size. However, this breakpoint does not appear in Figure 2d since the lineage in species E is not able to coalesce with lineages from species C or D until σ_{E1}, when E coalesces with the ancestor of C and D in the species tree.

Having described our model, we also adopt the following notation:

- Intervals on each branch are assigned integer-valued indices, increasing from 0 at the base of the branch to $\mathcal{I}_b - 1$ at the top of the branch.
- n_i: the constant-valued effective population size at any time in the interval with index i.
- a_i: the constant-valued number of available lineages to coalesce with at any time in the interval with index i.
- T_i: the length, in generations, of interval with index i.
- σ_i: the time, in generations, of the lower bound of interval with index i.
- t_b^u: the time, in generations, of the upper bound of branch b.
- t_b^l: the time, in generations, of the lower bound of branch b.
- $\mathcal{T}$: the genealogical tree
- $\mathcal{S}$: the species tree
- $\mathcal{I}_b$: the number of intervals in branch b

- $\mathcal{N}$: the number of tips in the species tree
- $L(\mathcal{T})$: the total tree length, in generations
- All summations where the stopping value is less than the starting value are equal to zero.

2.3 Deriving probabilities of changes in the genealogical tree

Any recombination event will result in one of four categories of outcomes (Figure 1). In category 1, there is no change to the genealogical tree. In categories 2 and 3, the branch lengths of the genealogical tree will change, while the topology remains the same. In category 4, the topology of the resulting tree is different from the original tree.

Our first steps are to calculate the probability of a recombination event falling into category 1, where the resulting genealogical tree is identical to the prior tree. The law of total probability then allows us to calculate the waiting distance to a change that falls into any of categories 2, 3, or 4.

2.3.1 Probability that a recombination event at time t on branch b will not change the tree:

We begin by calculating the probability that a recombination event that occurs on branch b and at time t will result in a category 1 outcome:

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

$$where$$

$$P_{ii} = -\frac{1}{a_i}e^{-\frac{a_i}{n_i}\sigma_{i+1}}, \tag{1}$$

$$and$$

$$P_{ik} = \exp\left(-\frac{a_i}{n_i}\sigma_{i+1} - \sum_{q=i+1}^{k-1}\frac{a_q}{n_q}T_q\right)\left(\frac{1}{a_k}(1-e^{-\frac{a_k}{n_k}T_k})\right),$$

when i is the interval index to which t belongs.

2.3.2 Probability that a recombination event on branch b will not change the tree:

We now integrate the previous equation through values of t across the branch and assume a uniform distribution of recombination events across the branch.

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

$$= \frac{1}{t_b^u - t_b^l}\sum_{i=0}^{\mathcal{I}_b-1}\frac{1}{a_i}T_i + \frac{n_i}{a_i}\left(e^{\frac{a_i}{n_i}\sigma_{i+1}} - e^{\frac{a_i}{n_i}\sigma_i}\right)\left(\sum_{k=i}^{\mathcal{I}_b-1}P_{ik}\right) \tag{2}$$

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

Finally, we can sum across all branches on the tree, while assuming a uniform distribution of recombination events with respect to total tree length, to determine the probability that a recombination event will result in a category 1 outcome.

$$\mathbb{P}(\text{tree unchanged}|\mathcal{T},\mathcal{S}) = \sum_{b=1}^{2N-2} \left[\frac{t_b^u - t_b^l}{L(\mathcal{T})}\right] \mathbb{P}(\text{tree unchanged}|b,\mathcal{T},\mathcal{S}) \tag{3}$$

2.3.4 Waiting distance to the next tree change:

Under the SMC', the distribution of waiting distances to the next recombination event is well-known:

$$p_r(d|\mathcal{T}) = rL(\mathcal{T}) \exp\left[-rL(\mathcal{T})d\right]. \tag{4}$$

where r is the recombination rate in recombs/bp/generation, and where d is the waiting distance.

We modify this equation to calculate the waiting distance to a recombination event that changes the genealogical tree.

$$\begin{aligned} &p_r(d|\mathcal{T},\mathcal{S}) = r\alpha_{\mathcal{S}}(\mathcal{T})L(\mathcal{T}) \exp\left[-r\alpha_{\mathcal{S}}(\mathcal{T})L(\mathcal{T})d\right], \\ &where \\ &\alpha_{\mathcal{S}}(\mathcal{T}) = 1 - \mathbb{P}(\text{tree unchanged}|\mathcal{T},\mathcal{S}) \end{aligned} \tag{5}$$

In this modified equation, distances continue to be exponentially distributed. However, the lambda parameter is reduced proportionally based on the probability that a recombination event will not change the tree.

2.4 Deriving probabilities of changes in the genealogical topology

Our next step is to derive analogous equations to describe the waiting distance distribution to a change in the *topology* of a genealogical tree. This involves excluding events of categories 2 and 3 to isolate the waiting distance to a category 4 event.

As previously, we find the branch-specific probability and then sum across all branches on the tree. The primary difference in our approach from the previous section is that for any branch b we have to also consider coalescence possibilities with the branch above it (branch c) and the branch with which it coalesces (branch b'). When labeling interval indices, we continue up branch c so that the final index is equal to $\mathcal{I}_{b+c} - 1$.

We also introduce a new variable, t_b^m, which corresponds to the lowest point at which branch b can potentially coalesce with its sibling, branch b'. t_b^m will always be the maximum of three values: the time at which b and b' are separated by a species divergence (if at all), the lowest point on branch b' (i.e., $t_{b'}^l$), and t_b^l. We occasionally refer to interval m, which is simply the interval whose lower bound is t_b^m.

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

We first present the probability that a recombination event occurring on a focal branch b will change the tree topology. As in Section 2.3, this results from integrating through the solution when specifying a particular branch and time (see Appendix).

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

$$where:$$

$$p_{b,1}^{(i)} = \frac{1}{a_i}\left[T_i + n_i\left(\exp\left(\frac{a_i}{n_i}\sigma_{i+1}\right) - \exp\left(\frac{a_i}{n_i}\sigma_i\right)\right)\left(\sum_{k=i}^{\mathcal{I}_{b+c}-1} P_{ik} + \sum_{k=m}^{\mathcal{I}_b-1} P_{ik}\right)\right] \tag{6}$$

$$and:$$

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

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

We 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.

$$\begin{aligned}\mathbb{P}(\text{topology unchanged}|\mathcal{T},\mathcal{S}) &= \sum_{b=1}^{2N-2} \frac{t_b^u - t_b^l}{L(\mathcal{T})} \times \mathbb{P}(\text{topology unchanged}|b,\mathcal{T},\mathcal{S}) \\ &= \frac{1}{L(\mathcal{T})}\sum_{b=1}^{2N-2}\left[\sum_{i=0}^{m-1} p_{b,1}^{(i)} + \sum_{i=m}^{\mathcal{I}_b-1} p_{b,2}^{(i)}\right]\end{aligned} \tag{7}$$

2.4.3 Waiting distance to the next topology change:

Unlike with the waiting distance to a *tree* change, the exact waiting distance distribution to a change in *topology* is complicated to derive because of possible intermediate recombination events that change the branch lengths but not the topology. These events impact the rate of recombination events and the probability that each further recombination event changes the topology of the newly generated genealogy. However, Deng et al. (2021) demonstrate the approximate negation of this effect due to 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 the recombination event will change the topology decreases. Therefore, we follow their approach by approximating the waiting distance to a change in topology using an exponential distribution:

$$\begin{aligned}&p_r(d|\mathcal{T},\mathcal{S}) = r\beta_{\mathcal{S}}(\mathcal{T})L(\mathcal{T})\exp\left[-r\beta_{\mathcal{S}}(\mathcal{T})L(\mathcal{T})d\right], \\ &where \\ &\beta_{\mathcal{S}}(\mathcal{T}) = 1 - \mathbb{P}(\text{topology unchanged}|\mathcal{T},\mathcal{S}).\end{aligned} \tag{8}$$

3 Examples

We adapted our solutions to python code, and we implemented them alongside the phylogenomic simulation package *ipcoal* (an MSC-focused wrapper for the popular population genomic simulator *msprime*) (McKenzie and Eaton, 2020a; Baumdicker et al., 2021). Doing so facilitated three simple, illustrative analyses.

First, when given a species tree and initial genealogy, we could examine the probability of changes in the genealogy given a recombination event on a specific branch and at a specific time. **Figure 3** illustrates the selection of three branches from a genealogy embedded in a species tree model. The species tree was unbalanced with five tips, had equal internode distances of 2.5e5 generations, and had Ne of 2e5. Parts b, c, and d show the probability that a recombination event falling at each time on the 3 selected genealogy branches does not change the topology (i.e., the branch lengths might change, but the topology remains the same) and does not change the tree (i.e., both the branch lengths and topology remain unchanged). These results 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, since any re-coalescence of the detached branch can only occur above the recombination breakpoint. This also makes clear that recombination events occurring when there are only two lineages remaining are certain to not change the topology, since the two lineages are only able to re-coalesce with each other (as shown in the upper panels of c and d).

Next, we tested the effect of variation in basic species tree parameters on expected waiting distances to tree changes and topology changes (**Figure 4**). We started with four different species trees: two imbalanced and two balanced. For each pair, there is a large tree (10 tips), and a small tree (5 tips) pruned from the large tree, with internode distances kept the same. On each species tree, we generated 1000 random MSC genealogies using each of 9 different 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 Figure 4, 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 (**Figure 5**). 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 recovered a correctly inferred Ne value of 150000 for both trees.

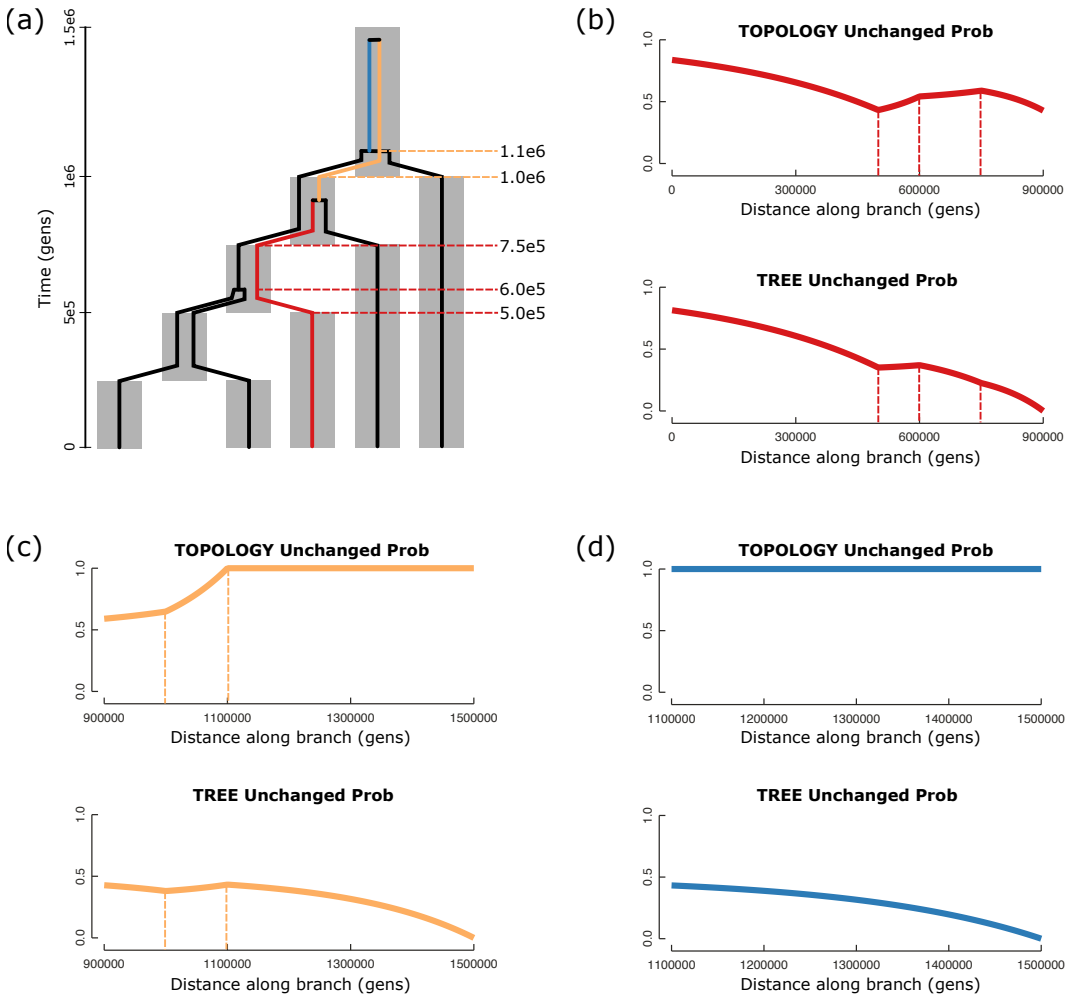

Figure 3. Probability that the topology and the tree remain unchanged, given a recombination event at a specific time on a specific branch. The species tree (a) was arbitrarily parameterized as being unbalanced with 5 tips, having internode distances of 2.5e5, and having constant Ne of 2e5. The embedded genealogical tree was randomly generated by *ipcoal* using MSC probabilities. We selected three branches, indicated by three different colors, across which we calculated the probability that the topology would remain unchanged and the tree would remain unchanged if a recombination event were to occur at each time point on the branch (b-d). Dashed lines indicate the beginning and ending times of different intervals on each branch – that is, either a relevant species tree divergence event or genealogy coalescence event.

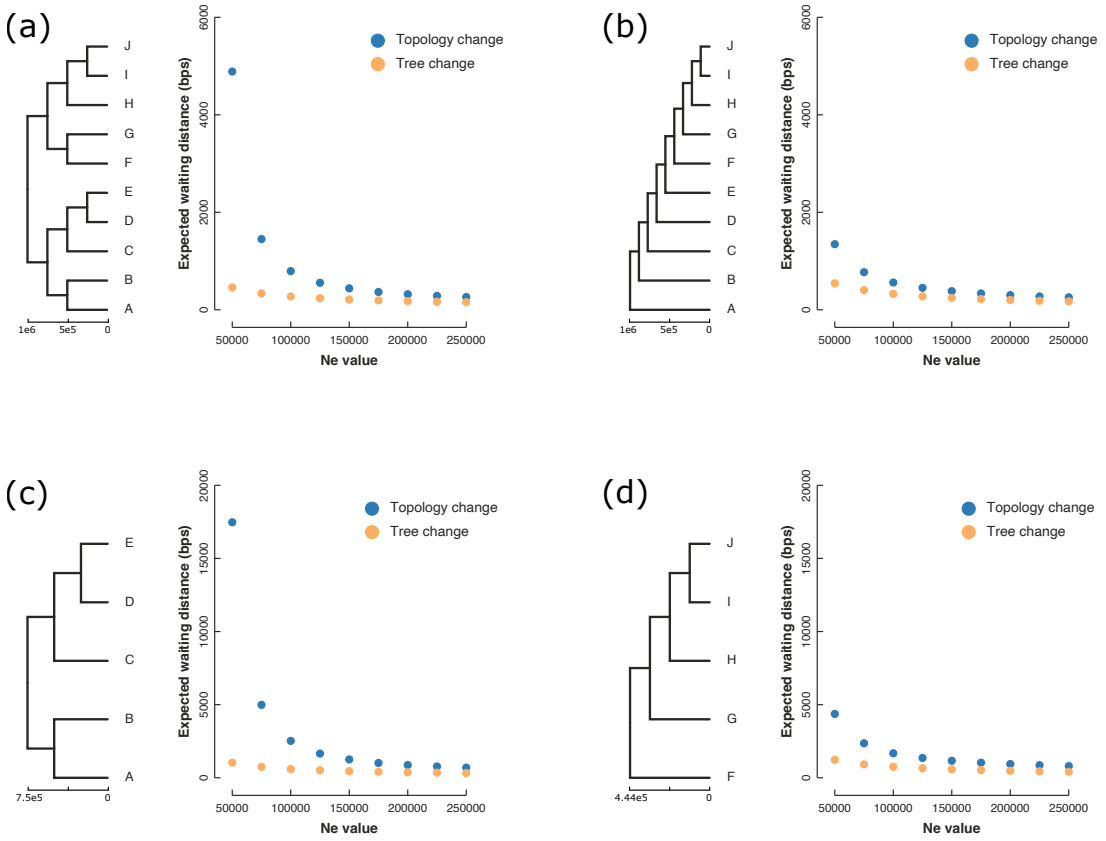

Figure 4. Mean expected waiting distances for differently parameterized species tree models. We used four different species trees: two 10-tip trees (a, b), each with 1e6-generation root heights, and two 5-tip trees (c, d) that are subselected from each of the larger trees. For each species tree, we iterated 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).

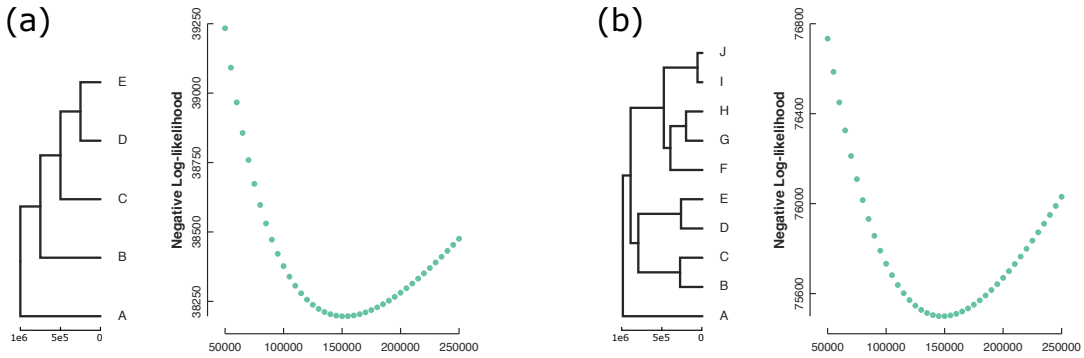

Figure 5. Inference of species tree Ne values from genealogies, observed waiting distances to topology changes, and recombination rate. We generated two different species trees: Both have root heights of 1e6 generations, but one has five tips, an unbalanced topology, and identical internode lengths, 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.

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.

References

Franz Baumdicker, Gertjan Bisschop, Daniel Goldstein, Graham Gower, Aaron P Ragsdale, Georgia Tsambos, Sha Zhu, Bjarki Eldon, E Castedo Ellerman, Jared G Galloway, Ariella L Gladstein, Gregor Gorjanc, Bing Guo, Ben Jeffery, Warren W Kretzschumar, Konrad Lohse, Michael Matschiner, Dominic Nelson, Nathaniel S Pope, Consuelo D Quinto-Cortés, Murillo F Rodrigues, Kumar Saunack, Thibaut Sellinger, Kevin Thornton, Hugo van Kemenade, Anthony W Wohns, Yan Wong, Simon Gravel, Andrew D Kern, Jere Koskela, Peter L Ralph, and Jerome Kelleher. Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*, 220(3), 12 2021. ISSN 1943-2631. doi:10.1093/genetics/iyab229. URL https://doi.org/10.1093/genetics/iyab229. iyab229. 9

James H Degnan. Modeling hybridization under the network multispecies coalescent. *Systematic biology*, 67 (5):786–799, 2018. 2

James H Degnan and Noah A Rosenberg. Discordance of species trees with their most likely gene trees. *PLoS genetics*, 2(5):e68, 2006. 2

James H Degnan and Noah A Rosenberg. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in ecology & evolution*, 24(6):332–340, 2009. 1

Yun Deng, Yun S. Song, and Rasmus Nielsen. The distribution of waiting distances in ancestral recombination graphs. *Theoretical Population Biology*, 141:34–43, 2021. ISSN 0040-5809. doi:https://doi.org/10.1016/j.tpb.2021.06.003. URL https://www.sciencedirect.com/science/article/pii/S0040580921000484. 1, 2, 3, 4, 5, 8, 15, 16, 18

John Gatesy and Mark S. Springer. Concatenation versus coalescence versus “concatalescence”. *Proceedings of the National Academy of Sciences*, 110 (13):E1179–E1179, 2013. doi:10.1073/pnas.1221121110. URL https://www.pnas.org/doi/abs/10.1073/pnas.1221121110. 2

RC Griffiths and P Marjoram. An ancestral recombination graph, pp. 257–270 in ima volume on mathematical population genetics, edited by donnelly p., tavaré s, 1996. 2

Melissa Hubisz and Adam Siepel. Inference of ancestral recombination graphs using argweaver. In *Statistical Population Genomics*, pages 231–266. Humana, New York, NY, 2020. 2

Melissa J Hubisz, Amy L Williams, and Adam Siepel. Mapping gene flow between ancient hominins through demography-aware inference of the ancestral recombination graph. *PLoS genetics*, 16(8): e1008895, 2020. 3

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

Jerome Kelleher, Alison M Etheridge, and Gilean McVean. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS computational biology*, 12(5):e1004842, 2016. 2

John Frank Charles Kingman. The coalescent. *Stochastic processes and their applications*, 13(3):235–248, 1982. 1

Laura Salter Kubatko and James H Degnan. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic biology*, 56(1):17–24, 2007. 2

Gang Li, Henrique V Figueiró, Eduardo Eizirik, and William J Murphy. Recombination-aware phylogenomics reveals the structured genomic landscape of hybridizing cat species. *Molecular biology and evolution*, 36(10):2111–2126, 2019. 12

Heng Li and Richard Durbin. Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357):493–496, 2011. 2

Wayne P Maddison. Gene trees in species trees. *Systematic biology*, 46(3):523–536, 1997. 1

Wayne P Maddison and L Lacey Knowles. Inferring phylogeny despite incomplete lineage sorting. *Systematic biology*, 55(1):21–30, 2006. 1

Paul Marjoram and Jeff D Wall. Fast" coalescent" simulation. *BMC genetics*, 7(1):1–9, 2006. 2

Patrick F McKenzie and Deren A R Eaton. ipcoal: an interactive Python package for simulating and analyzing genealogies and sequences on a species tree or network. *Bioinformatics*, 36(14):4193–4196, 05 2020a. ISSN 1367-4803. doi:10.1093/bioinformatics/btaa486. URL https://doi.org/10.1093/bioinformatics/btaa486. 9

Patrick F McKenzie and Deren AR Eaton. The multispecies coalescent in space and time. *bioRxiv*, 2020b. 2

Gilean AT McVean and Niall J Cardin. Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1459):1387–1393, 2005. 2

Bruce Rannala and Ziheng Yang. Bayes estimation of species divergence times and ancestral population sizes using dna sequences from multiple loci. *Genetics*, 164(4):1645–1656, 2003. 1

Matthew D Rasmussen, Melissa J Hubisz, Ilan Gronau, and Adam Siepel. Genome-wide inference of ancestral recombination graphs. *PLoS genetics*, 10(5):e1004342, 2014. 2

Peter R Wilton, Shai Carmi, and Asger Hobolth. The smc' is a highly accurate approximation to the ancestral recombination graph. *Genetics*, 200(1):343–355, 2015. 3

5 Appendix: Step-by-Step Solution

5.1 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 $\mathcal{T}$ 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 $\mathcal{T}$ 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}.

5.2 The distribution of distances to any change in a genealogical tree

Our goal is to derive a distribution of waiting distances to the next tree change from an input species tree and current genealogical tree.

5.2.1 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.

$$\mathbb{P}(\text{tree unchanged}|b, t, \mathcal{T}, \mathcal{S}) = \int_t^{t_b^u} \frac{1}{A(\tau)} p(\tau|t) d\tau \tag{9}$$

Where $A(\tau)$ is the number of lineages able to be coalesced with at time τ, and $p(\tau|t)$ is the coales-

cence rate.

$$= \int_t^{t_b^u} \frac{1}{A(\tau)} \frac{A(\tau)}{N(\tau)} e^{-\int_t^\tau \frac{A(s)}{N(s)} ds} d\tau \tag{10}$$

$$= \int_t^{t_b^u} \frac{1}{N(\tau)} e^{-\int_t^\tau \frac{A(s)}{N(s)} ds} d\tau \tag{11}$$

Here we have explicitly defined our 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.

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:

$$= \int_t^{\sigma_{i+1}} \frac{1}{N(\tau)} \exp\left(-\int_t^\tau \frac{A(s)}{N(s)} ds\right) d\tau + \sum_{k=i+1}^{\mathcal{I}_b - 1} \int_{\sigma_k}^{\sigma k+1} \frac{1}{N(\tau)} \exp\left(-\int_t^\tau \frac{A(s)}{N(s)} ds\right) d\tau \tag{12}$$

We solve first term and later terms separately:

First term –

$$= \frac{1}{a_i} - \frac{1}{a_i} e^{-\frac{a_i}{n_i}\sigma_{i+1}} e^{\frac{a_i}{n_i}t} \tag{13}$$

$$= \frac{1}{a_i} + P_{ii} e^{\frac{a_i}{n_i}t} \tag{14}$$

Later terms –

$$= \sum_{k=i+1}^{\mathcal{I}_b - 1} e^{\frac{a_i}{n_i}t} \exp\left(-\frac{a_i}{n_i}\sigma_{i+1} - \sum_{q=i+1}^{k-1} \frac{a_q}{n_q} T_q\right) \left(\frac{1}{a_k}(1 - e^{-\frac{a_k}{n_k}T_k})\right) \tag{15}$$

$$= \sum_{k=i+1}^{\mathcal{I}_b - 1} e^{\frac{a_i}{n_i}t} P_{ik} \tag{16}$$

Where:

- i is the interval index to which t belongs
- a_x is the value of A(t) for all t in interval x
- n_x is the value of N(t) for all t in interval x
- $T_x = \sigma_{x+1} - \sigma_x$
- $\mathcal{I}_b$ is the number of intervals on branch b
- $\mathcal{T}$ is the current genealogy, and

- $\mathcal{S}$ is the species tree with divergence times and branch-specific effective population sizes.

Adding these together, we are left with the solution:

$$\mathbb{P}(\text{tree unchanged}|b,t,\mathcal{T},\mathcal{S}) = \frac{1}{a_i} + \sum_{k=i}^{\mathcal{I}_b-1} P_{ik} e^{\frac{a_i}{n_i}t} \tag{17}$$

5.2.2 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:

$$\mathbb{P}(\text{tree unchanged}|b,\mathcal{T},\mathcal{S}) = \frac{1}{t_b^u - t_b^l}\int_{t_b^l}^{t_b^u} \mathbb{P}(\text{tree unchanged}|b,t,\mathcal{T},\mathcal{S})dt \tag{18}$$

Plugging in Equation 17 above, we find:

$$= \frac{1}{t_b^u - t_b^l}\sum_{i=0}^{\mathcal{I}_b-1}\frac{1}{a_i}T_i + \left(e^{\frac{a_i}{n_i}\sigma_{i+1}} - e^{\frac{a_i}{n_i}\sigma_i}\right)\left[-\frac{n_i}{a_i^2}e^{-\frac{a_i}{n_i}\sigma_{i+1}} + \frac{n_i}{a_i}\left(\sum_{k=i+1}^{\mathcal{I}_b-1}\exp\left(-\frac{a_i}{n_i}\sigma_{i+1} - \sum_{q=i+1}^{k-1}\frac{a_q}{n_q}T_q\right)\frac{1-\exp(-\frac{a_k}{n_k}T_k)}{a_k}\right.\right. \tag{19}$$

5.2.3 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:

$$\mathbb{P}(\text{tree unchanged}|\mathcal{T},\mathcal{S}) = \sum_{b=1}^{2N-2}\left[\frac{t_b^u - t_b^l}{L(\mathcal{T})}\right]\mathbb{P}(\text{tree unchanged}|b,\mathcal{T},\mathcal{S}) \tag{20}$$

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

$$\begin{aligned} &p_r(d|\mathcal{T},\mathcal{S}) = r\alpha_{\mathcal{S}}(\mathcal{T})L(\mathcal{T})\exp\left[-r\alpha_{\mathcal{S}}(\mathcal{T})L(\mathcal{T})d\right], \\ &where \\ &\alpha_{\mathcal{S}}(\mathcal{T}) = 1 - \mathbb{P}(\text{tree unchanged}|\mathcal{T},\mathcal{S}). \end{aligned} \tag{21}$$

5.3 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

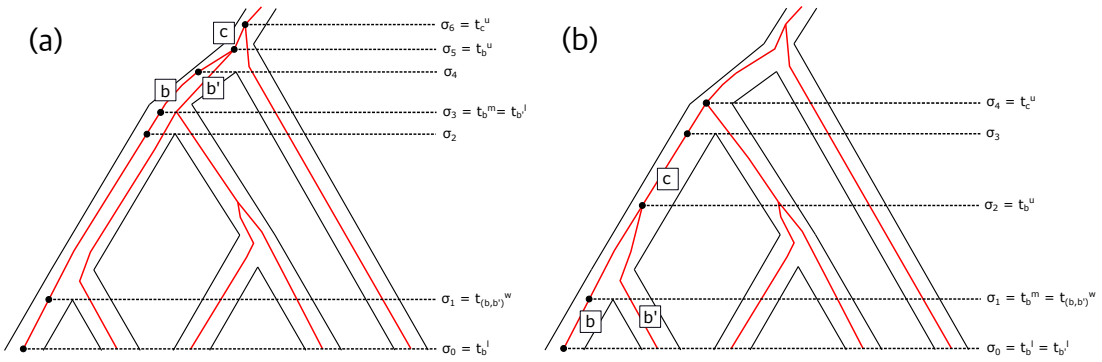

Figure 6. Illustrating the parameters for calculating the distribution of distances to a change in the topology of a genealogy. Panels (a) and (b) show slightly different genealogies embedded in the same species tree. In both (a) and (b), the focal branch b is the one leading from the left-most species. 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.

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).

$$\mathbb{P}(\text{topology unchanged}|b,\mathcal{T},\mathcal{S}) = \frac{1}{t_b^u - t_b^l}\int_{t_b^l}^{t_b^u} \mathbb{P}(\text{topology unchanged}|b,t,\mathcal{T},\mathcal{S})dt \tag{22}$$

$$= \frac{1}{t_b^u - t_b^l}\left[\left(\int_{t_b^l}^{t_b^m} + \int_{t_b^m}^{t_b^u}\right)\mathbb{P}(\text{topology unchanged}|b,t,\mathcal{T},\mathcal{S})dt\right] \tag{23}$$

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.

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

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

$$= \frac{1}{a_i} + \sum_{k=i}^{\mathcal{I}_{b+c}-1} P_{ik}e^{\frac{a_i}{n_i}t} + \sum_{k=m}^{\mathcal{I}_b-1} P_{ik}e^{\frac{a_i}{n_i}t} \tag{24}$$

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

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

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

$$= 2\left(\frac{1}{a_i} + \sum_{k=i}^{\mathcal{I}_b-1} P_{ik}e^{\frac{a_i}{n_i}t}\right) + \sum_{\mathcal{I}=n_b}^{\mathcal{I}_{b+c}-1} P_{ik}e^{\frac{a_i}{n_i}t} \tag{25}$$

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 –

$$\int_{t_b^l}^{t_b^m} \mathbb{P}(\text{topology unchanged}|b, t, \mathcal{T}, \mathcal{S}) = \sum_{i=0}^{m-1} \frac{1}{a_i}\left[T_i + n_i\left(\exp\left(\frac{a_i}{n_i}\sigma_{i+1}\right) - \exp\left(\frac{a_i}{n_i}\sigma_i\right)\right)\left(\sum_{k=i}^{\mathcal{I}_{b+c}-1} P_{ik} + \sum_{k=m}^{\mathcal{I}_b-1} P_{ik}\right)\right] \tag{26}$$

Second case –

$$\int_{t_b^m}^{t_b^u} \mathbb{P}(\text{topology unchanged}|b, t, \mathcal{T}, \mathcal{S}) = \sum_{i=m}^{\mathcal{I}_b-1} \frac{1}{a_i}\left[2T_i + n_i\left(\exp\left(\frac{a_i}{n_i}\sigma_{i+1}\right) - \exp\left(\frac{a_i}{n_i}\sigma_i\right)\right)\left(2\sum_{k=i}^{\mathcal{I}_b-1} P_{ik} + \sum_{k=\mathcal{I}_b}^{\mathcal{I}_{b+c}-1} P_{ik}\right)\right] \tag{27}$$

Result –

$$\mathbb{P}(\text{topology unchanged}|b, \mathcal{T}, \mathcal{S}) = \frac{1}{t_b^u - t_b^l}\left[\sum_{i=0}^{m-1} p_{b,1}^{(i)} + \sum_{i=m}^{\mathcal{I}_b-1} p_{b,2}^{(i)}\right] \tag{28}$$

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:

$$\mathbb{P}(\text{topology unchanged}|\mathcal{T}, \mathcal{S}) = \sum_{b=1}^{2N-2} \frac{t_b^u - t_b^l}{L(\mathcal{T})} \times \mathbb{P}(\text{topology unchanged}|b, \mathcal{T}, \mathcal{S})$$

$$= \frac{1}{L(\mathcal{T})} \sum_{b=1}^{2N-2}\left[\sum_{i=0}^{m-1} p_{b,1}^{(i)} + \sum_{i=m}^{\mathcal{I}_b-1} p_{b,2}^{(i)}\right] \tag{29}$$