Theoretical Population Biology 141 (2021) 34–43

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

The distribution of waiting distances in ancestral recombination graphs

Yun Deng [a,*], Yun S. Song [b,c,d], Rasmus Nielsen [b,e]

[a] Center for Computational Biology, University of California, Berkeley, CA 94720, United States of America
[b] Department of Statistics, University of California, Berkeley, CA 94720, United States of America
[c] Computer Science Division, University of California, Berkeley, CA 94720, United States of America
[d] Chan Zuckerberg Biohub, San Francisco, CA 94158, United States of America
[e] Department of Integrative biology, University of California, Berkeley, CA 94720, United States of America

ARTICLE INFO

Article history:
Received 30 December 2020
Available online 26 June 2021

Keywords:
Ancestral recombination graph
Sequentially Markovian coalescent
Tree inference methods
Recombination evolution

ABSTRACT

The *ancestral recombination graph* (ARG) contains the full genealogical information of the sample, and many population genetic inference problems can be solved using inferred or sampled ARGs. In particular, the waiting distance between tree changes along the genome can be used to make inference about the distribution and evolution of recombination rates. To this end, we here derive an analytic expression for the distribution of waiting distances between tree changes under the sequentially Markovian coalescent model and obtain an accurate approximation to the distribution of waiting distances for topology changes. We use these results to show that some of the recently proposed methods for inferring sequences of trees along the genome provide strongly biased distributions of waiting distances. In addition, we provide a correction to an undercounting problem facing all available ARG inference methods, thereby facilitating the use of ARG inference methods to estimate temporal changes in the recombination rate.

1. Introduction

At each position of the genome, the relationship among individuals in a sample can be characterized by a tree, known as a genealogical or coalescent tree, and it can be regarded as the result of a generative process called the *coalescent* (Kingman, 1982a,b). In the presence of recombination, the genealogies can vary at different positions of the genome, and the *ancestral recombination graph* (ARG) — the structure which fully describes the joint distribution of coalescent trees along the genome — provides all the information about the genealogical history of a sample, including the locations of recombination events. The full ARG can be also seen as the result of generative process, *the coalescent with recombination* (Griffiths, 1981; Hudson, 1983). Although it is straightforward to simulate under this process (Hudson, 2002; Kelleher et al., 2016), inferring ARGs from population genetic variation data is a very challenging problem. Indeed, much algorithmic work has been done on reconstructing parsimonious histories with recombination (Hein, 1993; Gusfield et al., 2003; Song and Hein, 2003; Wang et al., 2001; Lyngsø et al., 2005; Gusfield, 2014; Ignatieva et al., 2020) and sampling ARGs under the coalescent with recombination (Griffiths and Marjoram, 1996; Fearnhead and Donnelly, 2001; Jenkins and Griffiths, 2011; Rasmussen et al., 2014).

The coalescent with recombination (Griffiths, 1981; Hudson, 1983) was originally formulated as a stochastic process over time, but Wiuf and Hein (1999) later showed that it can also be reformulated as a spatial process along the genome. This spatial process is not Markovian, because of the long-range dependency caused by coalescence events between distant genomic positions. However, constraining the spatial process to be Markovian (McVean and Cardin, 2005; Marjoram and Wall, 2006; Hobolth and Jensen, 2014) has led to useful, practical approximations of the full coalescent with recombination while retaining accuracy in many aspects. The first Markovian approximation is called the *sequentially Markovian coalescent* (SMC) (McVean and Cardin, 2005), and a subsequent improvement (Marjoram and Wall, 2006), known as SMC', incorporates an additional class of genealogical events.

The Markovian approximations have successfully been applied in the estimation of changes in population size (e.g., Li and Durbin (2011), Schiffels and Durbin (2014) and Terhorst et al. (2017)), by representing the genealogy as states of a Hidden Markov Model (HMM), whose transition probabilities then can be calculated under the SMC or SMC' assumptions. However, these methods

* Corresponding author.
E-mail addresses: yun_deng@berkeley.edu (Y. Deng), rasmus_nielsen@berkeley.edu (R. Nielsen).

https://doi.org/10.1016/j.tpb.2021.06.003
0040-5809/

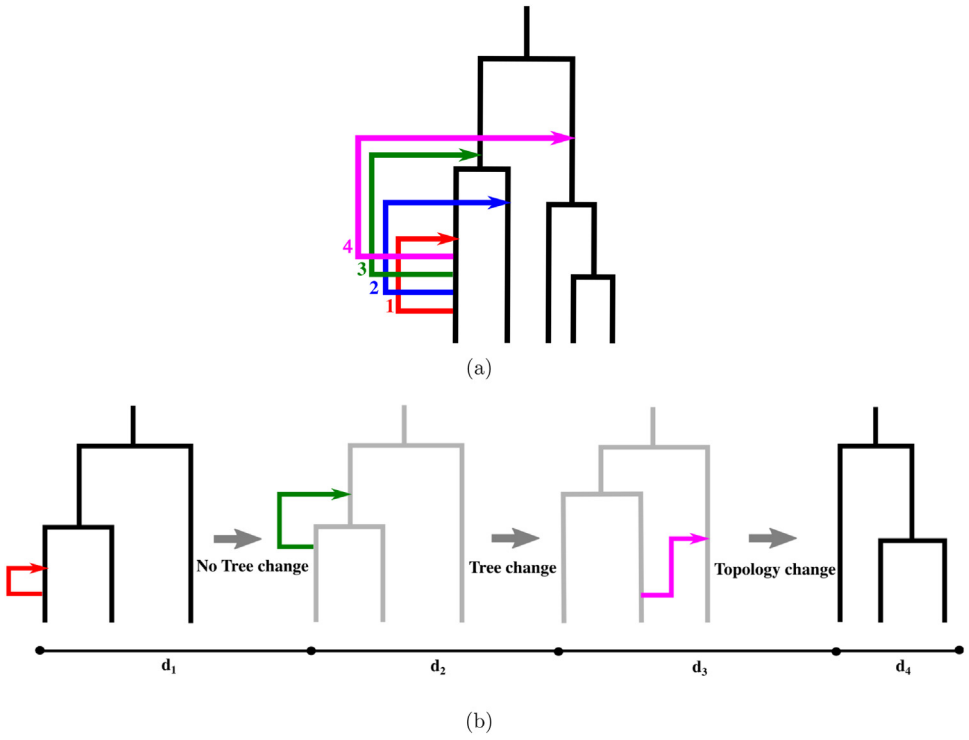

Fig. 1. Difficulties in genealogy inference. (a) Different types of recombination events, in which type 1 does not change the tree, type 2 and 3 change the tree but not the topology, and type 4 changes the topology; (b) Illustration of tree transition omission in tree inference methods, in which the shaded trees are harder to detect as they are not produced by topology changes.

are restrained to at most analyzing a few individuals due to the exploding size of the state space with increasing sample size. However, recently several different methods for inferring genealogies in models with recombination have been proposed (e.g., Rasmussen et al. (2014), Kelleher et al. (2019) and Speidel et al. (2019)). *ARGweaver* by Rasmussen et al. (2014) is capable of full posterior sampling of ARGs under the SMC or SMC' approximations using Markov Chain Monte Carlo (MCMC), but becomes prohibitively slow for large sample sizes (typically $>$ 50). *Relate* by Speidel et al. (2019) and *tsinfer* by Kelleher et al. (2019) are capable of inferences for larger sample sizes, but do not perform full posterior sampling. Both methods only provide a single estimate of the tree topology, although *Relate* also samples coalescent times using MCMC, conditionally on the estimated local coalescent tree topology.

1.1. Motivation

In the SMC, the SMC', and the full coalescent process with recombination, the waiting distance d until the next recombination event along the chromosome is exponentially distributed, conditionally on the total tree length $L(\mathcal{T})$ of the current tree $\mathcal{T}$:

$$p_r(d \mid \mathcal{T}) = \frac{\rho}{2} L(\mathcal{T}) \exp\left[-\frac{\rho}{2} L(\mathcal{T}) d\right], \tag{1}$$

where $L(\mathcal{T})$ is in coalescent unit of $2N_e$ generations and $\frac{\rho}{2} = 2N_e r$ denotes the population-scaled recombination rate per bp. SMC' provides a closer approximation to the full coalescent with recombination than does SMC, as the former allows for events in which a lineage splits off and coalesces back to the same branch (type 1 in Fig. 1a). This type of event occurs with particularly high probability when the sample size is small (Wilton et al., 2015). However, it does not change the genealogy (second tree in Fig. 1b) and is therefore not sampled in *ARGweaver* or reported in the output of *msprime* simulations (Kelleher et al., 2016). Therefore, the waiting distance between adjacent trees in *ARGweaver* or *msprime* will not follow the exponential distribution shown in (1). Furthermore, there are two additional types of events (type 2 and 3 in Fig. 1a) that may occur in both SMC and SMC' approximations where some coalescence times change, but not the tree topology Hudson and Kaplan (1985) considered different recombination types along this line and studied their statistical properties, and Hein et al. (2004) refined this classification). Hence, the waiting distance distribution until next topology change is even further biased away from (1), and similar problem was explored in the context of incompatibility by Hudson and Kaplan (1985). To quantify and better visualize these phenomena, we perform simulations using the SMC' mode in *msprime*, with $n = 8, N_e = 1 \times 10^4, \mu = r = 1 \times 10^{-8}$. We then divide trees into bins based on their total branch lengths, and collect waiting distances until tree and topology changes. We compare the empirical distribution to the exponential distribution using the midpoint of the tree length to represent each bin (Figs. 2a and 2b).

To use the waiting distance distribution between trees inferred by SMC or SMC' models, or as reported by common simulation programs such as *msprime*, for understanding patterns of recombination, or for validating the accuracy of inference methods or simulation methods, it is necessary to understand the distribution of waiting distances between tree topology changes induced by events of type 1, 2, or 3 in Fig. 1a.

In this paper we derive the waiting distance distributions for the SMC' model and we use these results to benchmark three common methods for inferring tree topologies along the length of a chromosome: *ARGweaver* (Rasmussen et al., 2014), *Relate* (Speidel et al., 2019), and *tsinfer* (Kelleher et al., 2019). We also illustrate the utility of the results, when used in combination with methods for inferring trees, for inferences regarding temporal changes in recombination rate.

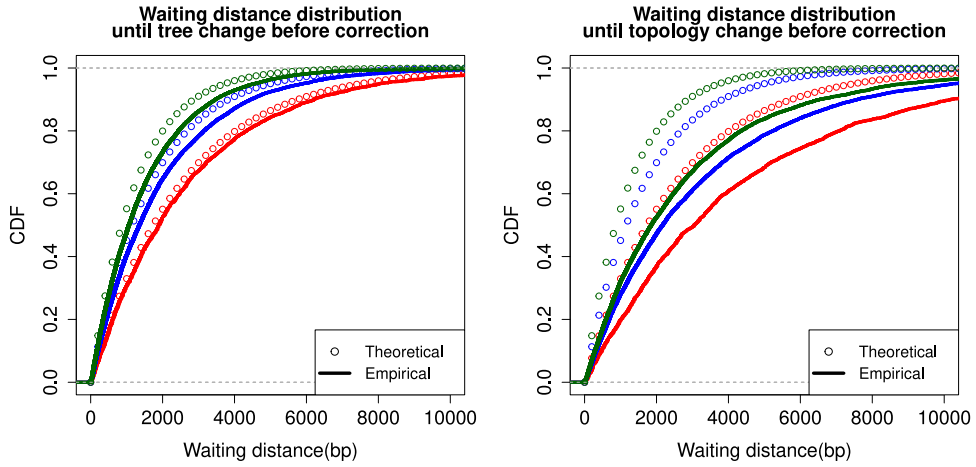

Fig. 2. Uncorrected distributions of waiting distance until tree or topology changes. The distribution of simulated waiting distances until tree or topology changes (solid lines) is compared to the exponential distribution in (1) (circles). Green, blue, and red colors correspond the total tree length, $L(\mathcal{T})$, falling within $[3\times10^4, 5\times10^4)$, $[5\times10^4, 7\times10^4)$, and $[7\times10^4, 9\times10^4)$, respectively. (a) The waiting distance until tree change differs from the simple exponential distribution in (1). (b) The waiting distance until topology change differs even more from the simple exponential distribution in (1).

2. Theoretical results

2.1. Notation

Suppose we have a sample of size n from a diploid population with constant effective population size N_e (all times mentioned henceforth will be in coalescent units of $2N_e$ generations). We use T_i to denote the length of the epoch when there are i ancestral lineages in the coalescent tree and introduce the notation

$$\sigma_i = T_n + T_{n-1} + \cdots + T_i, \quad \text{for } i = 2, \ldots, n,$$
$$\sigma_{n+1} = 0,$$

to denote the times when coalescence occurs. We let $A_n(t)$ denote the number of lineages at time t ancestral to the sample and define it to be right-continuous. There are $2n-2$ branches in the coalescent tree before the most recent common ancestor (TMRCA) since there are n leaf nodes and $n-2$ internal nodes created by coalescent events prior to the TMRCA, and we index them arbitrarily by $b = 1, \ldots, 2n-2$. For each branch b, we use t_b^l and t_b^u to denote its lower and upper times, respectively (see Fig. 3).

2.2. Waiting distance distribution until next tree change

The waiting distance until next tree change can be modeled as a waiting distance in a thinned Poisson process, where we color the events differently depending on whether they produce identical trees. Since the intensity of the un-thinned process is just the product of the tree length and the recombination rate, the only thing which needs to be identified is the thinning parameter, the probability of a recombination leading to a tree change. A simpler version of the problem was previously solved using the same idea for the waiting distance distribution until a TMRCA change for $n = 2$ (Carmi et al., 2014) and $n = 3$ (https://github.com/scarmi/arg_tree_change_distance_n3/blob/main/TreeChangeDistances.pdf).

Now suppose the current tree is $\mathcal{T}$, and the recombination happened on branch b at time t, we have the following result regarding the probability of this recombination not changing the tree under the SMC' model (a type 1 event):

Proposition 1. *Under the SMC', the probability of a recombination not to change the tree, given its breakpoint on branch b at time t on tree $\mathcal{T}$ is:*

$$\mathbb{P}(\text{tree unchanged} \mid b, t, \mathcal{T}) = \frac{1}{i} + e^{it}\left[\sum_{j=A_n(t_b^u)+1}^{i} P_{ij}\right], \tag{2}$$

where

$$i = A_n(t),$$
$$P_{ii} = -\frac{1}{i}e^{-i\sigma_i},$$
$$P_{ij} = \exp\left(-i\sigma_i - \sum_{k=j+1}^{i-1} kT_k\right)\frac{1}{j}[1 - e^{-jT_j}].$$

Note that when the lower index is larger than the upper index in a summation, that sum is here defined to be 0.

By a change in tree, we here mean a change in at least one coalescence time, but the topology may stay the same. This result, which is proved in the Appendix, shows that the probability of a recombination event being detected depends on where it occurs in the tree. We will also later show that this result facilitates inferences of temporal changes in recombination rates using inferred ARGs. We also note that under the SMC model $\mathbb{P}(\text{tree unchanged} \mid b, t, \mathcal{T}) = 0$.

The unconditional probability of type 1 event on a branch b under the SMC' can then be found by integrating over the breakpoint time, t, with respect to its conditionally uniform distribution on the branch:

Proposition 2. *Under the SMC', the probability of a recombination happening on branch b not to change the tree is:*

$$\mathbb{P}(\text{tree unchanged} \mid b, \mathcal{T}) = \frac{1}{t_b^u - t_b^l}\sum_{i=A_n(t_b^u)+1}^{A_n(t_b^l)} p_b^{(i)}, \tag{3}$$

where

$$p_b^{(i)} = \frac{1}{i}\left[T_i + (e^{i\sigma_i} - e^{i\sigma_{i+1}})\sum_{j=A_n(t_b^u)+1}^{i} P_{ij}\right].$$

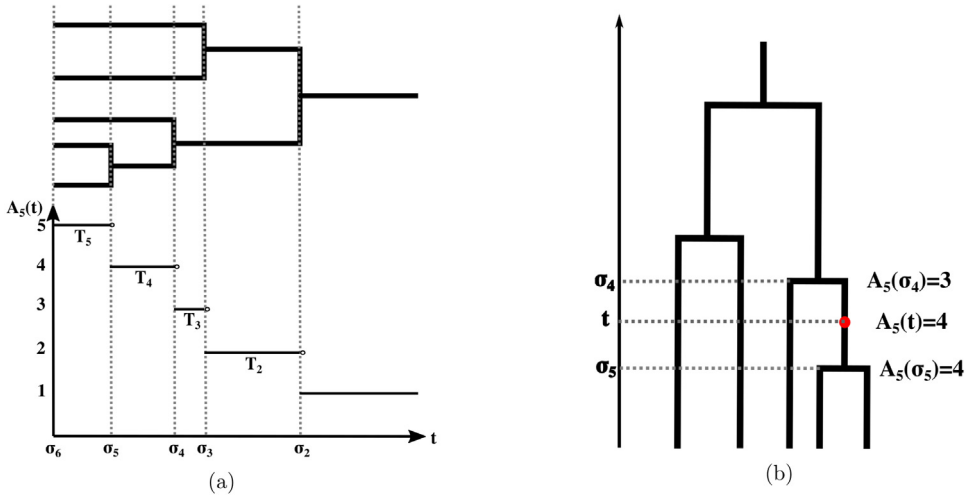

Fig. 3. Illustration of the notation used. (a) $A_n(t)$, T_i and σ_j are illustrated for an example genealogy with sample size $n = 5$. (b) A recombination event occurs on a branch at time t, and the corresponding upper and lower times of the branch are $t_b^u = \sigma_4$ and $t_b^l = \sigma_5$.

Now, the probability of a recombination event not changing the tree can simply be given by a weighted sum of the probabilities in (3) over all branches in the tree, weighting by the branch lengths:

Theorem 1 (*Tree-unchanging Probability*)**.** *Under the SMC', the probability of a recombination event not changing the tree is:*

$$\mathbb{P}(\text{tree unchanged} \mid \mathcal{T}) = \sum_{b=1}^{2n-2}\left[\frac{t_b^u - t_b^l}{L(\mathcal{T})}\right]\mathbb{P}(\text{tree unchanged} \mid b, \mathcal{T}) = \frac{1}{L(\mathcal{T})}\sum_{b=1}^{2n-2}\sum_{i=A_n(t_b^u)+1}^{A_n(t_b^l)} p_b^{(i)}.$$

The waiting distance between tree changes under the SMC' is then determined as follows.

Theorem 2 (*Waiting Distance Distribution Until Next Tree Change*)**.** *Under the SMC', the distribution of waiting distance until next tree change given the current tree $\mathcal{T}$ is given by:*

$$p_r(d \mid \mathcal{T}) = \frac{\rho}{2}\alpha(\mathcal{T})L(\mathcal{T})\exp\left[-\frac{\rho}{2}\alpha(\mathcal{T})L(\mathcal{T})d\right], \tag{4}$$

where

$$\alpha(\mathcal{T}) = 1 - \mathbb{P}(\text{tree unchanged} \mid \mathcal{T}).$$

Note that the waiting distance until next tree change is still exponential, but the intensity is reduced by a factor of $\alpha(\mathcal{T})$. Fig. 4a is generated similarly to Fig. 2, except that we are collecting waiting distances of $L(\mathcal{T})\alpha(\mathcal{T})$ in bins, to show that this expression indeed accurately describes the waiting distance distribution observed in *msprime* simulations using the SMC' mode.

2.3. Waiting distance distribution until next topology change

Although inferring every tree change would be desirable for ARG inference, it is notoriously hard, especially when scaling to hundreds, or thousands of genomes. Two of the most recent genealogy inference programs, *Relate* and *tsinfer*, use efficient approximations to infer genomic series of trees. However, inferences of tree changes are mostly guided by information regarding topologies, and neither method is designed to detect changes in the tree that does not involve a change in topology. For benchmarking and comparing these, and other programs, is therefore also important to understand the distribution of waiting distances between changes in topology.

To derive the waiting distance distribution between topology changes, we will again use the idea of a thinned process. The quantity of interest is the probability that a recombination event will change the tree topology. This probability can be calculated in very similar manner to that in Theorem 2, except that there are two more types of events of recombination and coalescent to consider (type 2 and 3) so some extra bookkeeping is needed. In Appendix A.3, we prove the following result:

Theorem 3 (*Topology-unchanging Probability*)**.** *For a given branch b in tree $\mathcal{T}$, let b' denote the branch that b coalesces with and let c denote their parental branch. Then, under the SMC', the conditional probability that a recombination event will not change the tree topology given the current tree $\mathcal{T}$ is*

$$\mathbb{P}(\text{topology unchanged} \mid \mathcal{T}) = \frac{1}{L(\mathcal{T})}\sum_{b=1}^{2n-2}\left[\sum_{i=A_n(t_{b'}^l)+1}^{A_n(t_b^l)} p_{b,1}^{(i)} + \sum_{i=A_n(t_b^u)+1}^{A_n(t_{b'}^l)} p_{b,2}^{(i)}\right],$$

where

$$p_{b,1}^{(i)} = \frac{1}{i}\left[T_i + (e^{i\sigma_i} - e^{i\sigma_{i+1}})\left(\sum_{j=A_n(t_{b'}^u)+1}^{A_n(t_{b'}^l)} P_{ij} + \sum_{j=A_n(t_c^u)+1}^{i} P_{ij}\right)\right],$$

$$p_{b,2}^{(i)} = \frac{1}{i}\left[2T_i + (e^{i\sigma_i} - e^{i\sigma_{i+1}})\left(2\sum_{j=A_n(t_c^u)+1}^{i} P_{ij} - \sum_{j=A_n(t_c^u)+1}^{A_n(t_c^l)} P_{ij}\right)\right].$$

Unlike the result for the distribution of waiting distance until next tree change, the waiting distance to the next topology changing event is not exponentially distributed because of the possibility of recombination events that change the tree but do not change the topology. Such events will change the intensity of the process so that it is no longer time-homogeneous. Arriving at an exact formula is, therefore, difficult. However, we notice that there is a rather accurate approximation method based on the following empirical observation:

The product of total tree length and the topology-changing probability remains approximately the same after a topology-unchanging recombination. This makes intuitive sense because a tree with larger total branch length is likely to have

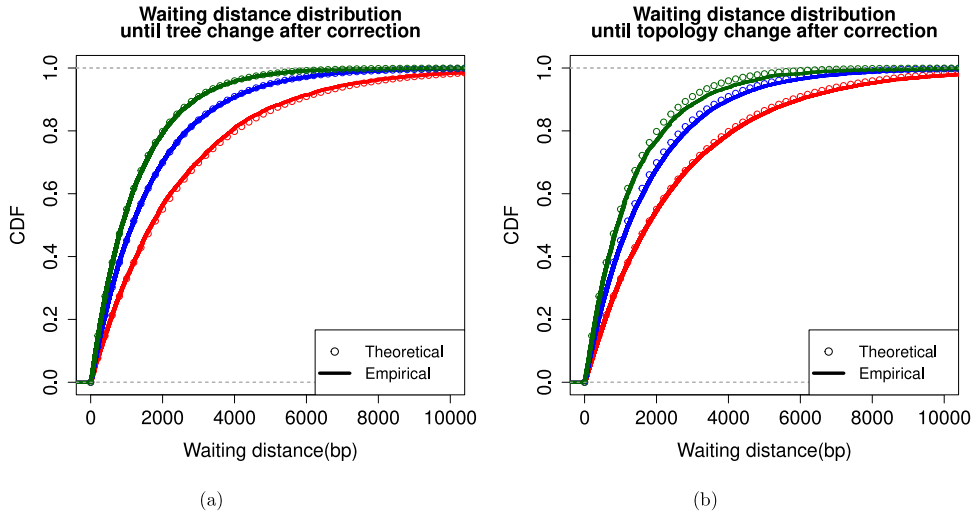

Fig. 4. Corrected distributions of waiting distance until tree or topology changes. The distribution of simulated waiting distances until tree or topology changes (solid lines) is compared to the corrected distribution in (4) or (5) (circles). Green, blue, and red colors indicate thinned intensities, $L(\mathcal{T})\alpha(\mathcal{T})$ in (a) and $L(\mathcal{T})\beta(\mathcal{T})$ in (b), that are in the intervals $[3\times10^4, 5\times10^4)$, $[5\times10^4, 7\times10^4)$, and $[7\times10^4, 9\times10^4)$, respectively. In (a), notice that the waiting distance until tree change matches the corrected exponential distribution of (4). Similarly, in (b), the waiting distance distribution until topology change matches the corrected exponential distribution of (5).

long deep branches, which favor type 1, 2 and 3 recombinations, leading to a lower topology-changing probability.

Note in Fig. 5 that while both the tree length, $L(\mathcal{T})$, and the probability of a topology changing recombination event, $\beta(\mathcal{T}) = 1-\mathbb{P}(\text{topologyunchanged} \mid \mathcal{T})$, after a topology-unchanging event, have relatively high variance, the variance of their product is much smaller. In other words, a recombination+coalescent event that increases the total tree length will decrease the probability that the next event is topology changing and *vice versa*. The net effect is that the product of the two, $L(\mathcal{T})\beta(\mathcal{T})$, is much more stable than either are individually. Using this concentration relation we can build an approximation for the distribution of waiting distances until next topology change:

The distribution of waiting distance until next topology change can be approximated by:

$$p_r(d \mid \mathcal{T}) = \frac{\rho}{2}\beta(\mathcal{T})L(\mathcal{T})\exp\left[-\frac{\rho}{2}\beta(\mathcal{T})L(\mathcal{T})d\right], \tag{5}$$

where

$$\beta(\mathcal{T}) = 1 - \mathbb{P}(\text{topology unchanged} \mid \mathcal{T}).$$

This approximation also agrees well with the simulations in Fig. 4b, showing waiting distances until topology changes conditioning on $L(\mathcal{T})\beta(\mathcal{T})$ falling in certain bins.

Our code for calculating $\alpha(\mathcal{T})$ and $\beta(\mathcal{T})$ using *tskit* format is available here: https://github.com/YunDeng98/Correction_probability_calculation.

3. Applications

3.1. Benchmark of tree inference methods

One application of the theory introduced in this article is the use for benchmarking of some ARG/genealogy inference methods. If the method detects every tree/topology change, or if it samples from a correct posterior distribution of waiting distances, then the waiting distance between adjacent trees should be described by (4) or (5).

To investigate this, we did simulations with realistic choices of parameters ($\mu = r = 1\times10^{-8}$, $N_e = 1\times10^4$ with 8 haplotypes) and used *ARGweaver*, *Relate*, and *tsinfer* to infer the genealogy. Then we compared the distribution of waiting distance between adjacent trees in the output with our theoretical prediction (4) or (5).

Although *ARGweaver* does not report some tree transitions caused by type 1 recombination (Fig. 6a), causing the waiting distance distribution to be different from the exponential distribution in (1), it is indeed capable of doing approximately posterior sampling of tree changes following (4) (Fig. 6d). However, we observe a very small bias which might be due to the discretization in *ARGweaver*.

However, both *Relate* and *tsinfer* are undersampling trees. Under realistic choices of mutation and recombination rate of $\mu = r = 1\times10^{-8}$ with 8 haplotypes, the waiting distance distribution in *tsinfer* is quite different from (5), suggesting that *tsinfer* undersamples topology changes and overestimates waiting distances (Fig. 6e). *Relate* has an even stronger tendency to undersample topology changes and to overestimate waiting distances (Fig. 6f).

3.2. Inference of temporal variation of recombination rate

It is generally hard to study the evolution of recombination rate through time without specific reference to ARGs. After sampling the ARG of a region using a program that accurately samples tree changes, such as *ARGweaver*, a naive way of estimating time-specific recombination rates is to count the number of recombination happening within each time interval, divided by the total branch length appearing in the interval. In order to test this procedure, we simulated data under a mutation rate of $\mu = 1\times10^{-8}$ and a constant recombination rate $r = 1\times10^{-8}$ with 8 haplotypes of 1Mb. However, the naive estimation procedure provides biased estimates in at least two aspects (Fig. 7): first, the inferred rates are significantly lower than the true values; second, the inferred trajectory is not constant, leading to a false conclusion of temporal changes in the recombination rate.

The bias mainly comes from the fact that type 1 recombinations do not result in tree change at all and are unreported by *ARGweaver*, which instead reports recombination events according to (2). As the probability of a type 1 event depends on the number of lineages, this induces the appearance of temporal recombination rate changes. The intuition here is that because there are many more coalescence events happening recently, where branches tend to be shorter and where there are more of

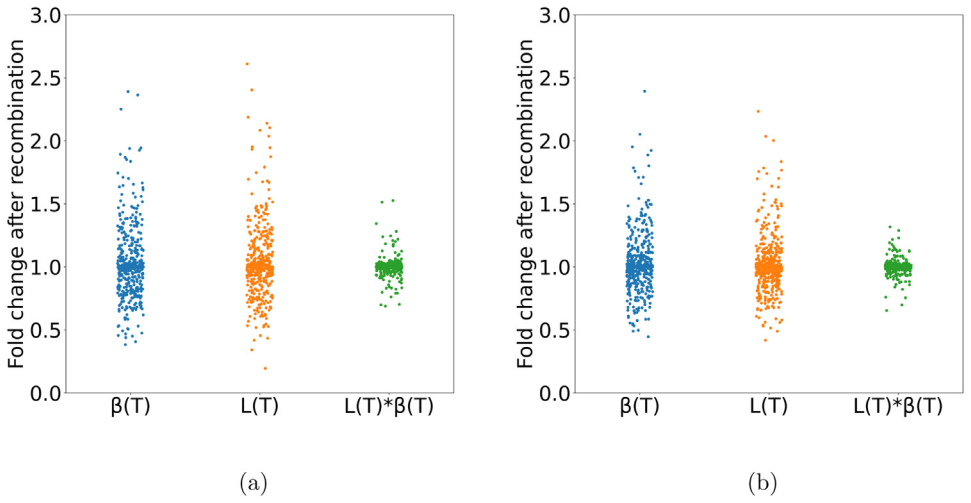

Fig. 5. Concentration of $L(\mathcal{T})\beta(\mathcal{T})$ after a topology-unchanging recombination. Simulations were performed using *msprime* for $n = 20$ and $n = 50$, and fold-changes of $L(\mathcal{T})$, $\beta(\mathcal{T})$ and $L(\mathcal{T})\beta(\mathcal{T})$ after topology-unchanging recombinations were calculated. Although neither $L(\mathcal{T})$ nor $\beta(\mathcal{T})$ remains stable after a topology-unchanging recombination, their product remains approximately constant with high probability. This relation is shown here empirically with simulation results using sample size $n = 20$ (a) and $n = 50$ (b).

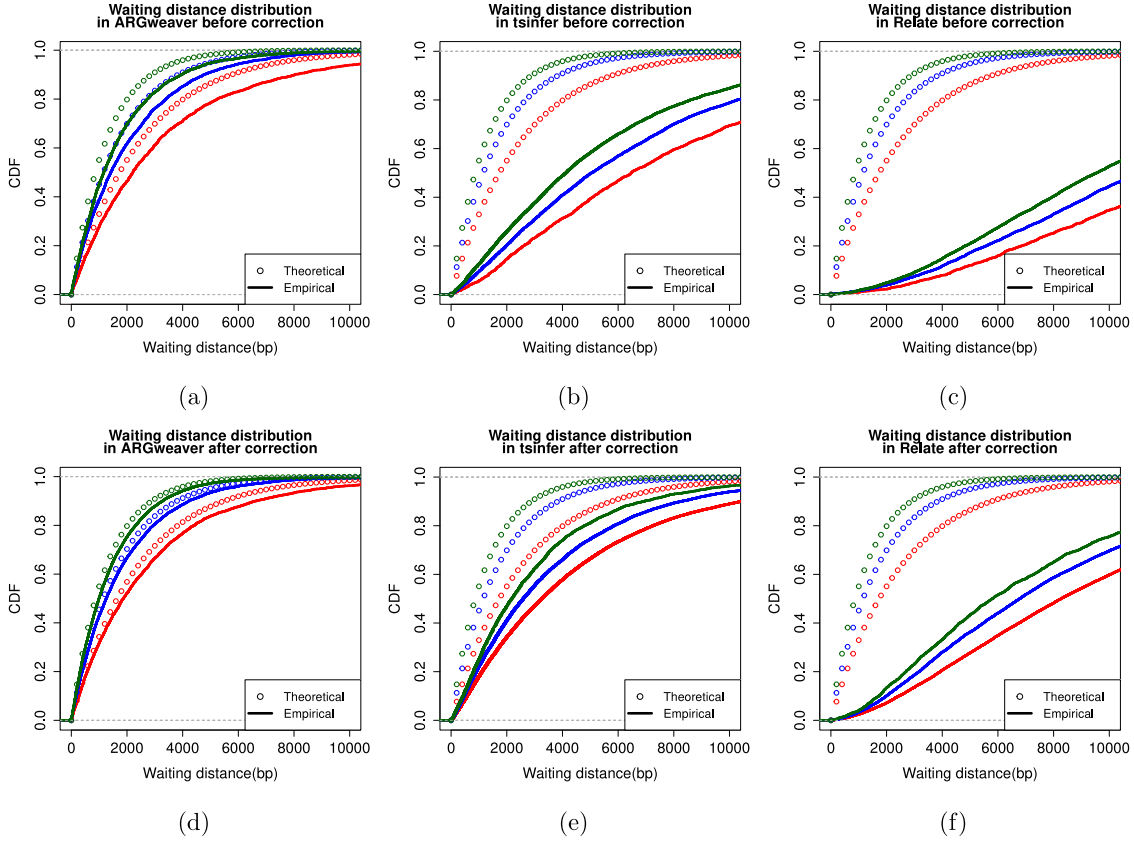

Fig. 6. The waiting distance distribution in *ARGweaver*, *tsinfer* and *Relate*. Simulations are done in *msprime* using $n = 8, N_e = 1 \times 10^4, \mu = r = 1 \times 10^{-8}$ and the distribution of waiting distances between adjacent trees in the outputs (solid lines) is compared to (1) and (4)/(5) (circles). Green, blue, and red mean the conditioned quantity to be within $(3 \times 10^4, 5 \times 10^4), (5 \times 10^4, 7 \times 10^4), (7 \times 10^4, 9 \times 10^4)$. The conditioned quantities are $L(\mathcal{T})$ in (a), (b) and (c), $L(\mathcal{T})\alpha(\mathcal{T})$ in (d), and $L(\mathcal{T})\beta(\mathcal{T})$ in (e) and (f). (a) The waiting distance distribution in *ARGweaver* is not well-characterized by the simple exponential distribution (1); (b) The waiting distance distribution in *tsinfer* is biasing away from the exponential distribution even more; (c) The waiting distance distribution in *Relate* is significantly biasing away from the exponential distribution; (d) The correction (4) provides a much better fit to the waiting distance distribution in *ARGweaver*; (e) The correction (5) is closer to the empirical distribution in *tsinfer*, but the correction does not completely solve the problem; (f) The correction (5) on *Relate* helps little.

them, there are fewer opportunities for "silent" recombinations, while in the older deep lineages many more recombinations can be missed.

The way to correct this is to introduce "effective counts" instead of naive counts of the recombination events in each time interval. For a recombination happening on branch b at time t on

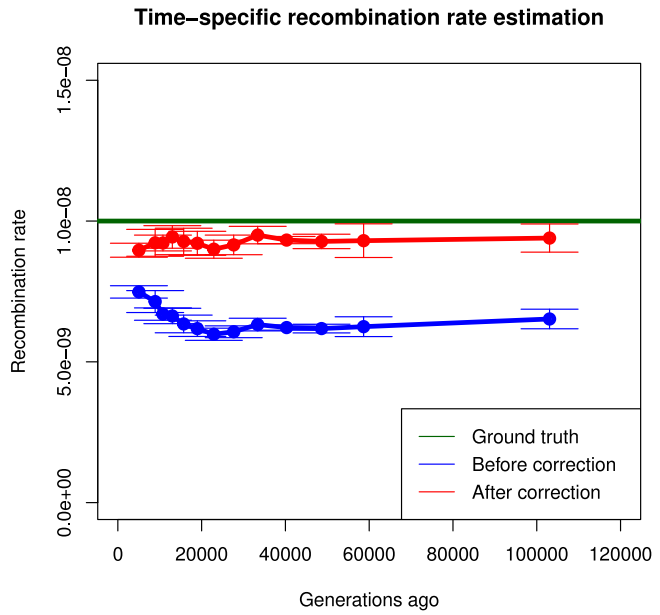

Fig. 7. The time specific recombination estimation before and after correction using (2). The naive estimation without correction (blue line) is biased significantly from the true values (green line) and falsely indicates a decrease in recombination rates the past, whereas the correction leads to a much more reasonable estimate (red line) which is only slightly underestimating the recombination rate. 500 ARGs of 8 haplotypes are sampled from *ARGweaver* and divided into 5 groups, and each estimate is obtained as the average over 100 ARGs from each group. The error bars represent 1 SE among the five estimates.

tree $\mathcal{T}$, instead of counting it as one, we assign it the following weights:

$$\frac{1}{\mathbb{P}(\text{tree changed} \mid b, t, \mathcal{T})} = \frac{1}{1 - \mathbb{P}(\text{tree unchanged} \mid b, t, \mathcal{T})}.$$

After applying this correction, we observe that the estimate is significantly closer to the true values (Fig. 7) and no false evidence of temporal changes in the recombination rate is observed. In real data analyses it is, therefore, possible to infer changes in recombination rates using a combination of *ARGweaver* inferred trees and the correction derived here. We note that we did not test this procedure on simulated data with time-varying recombination rates because there is no available program to efficiently do such simulations.

4. Discussion

In this paper, we derived analytical formulae for the distribution of waiting distance to the next tree change and close approximations for the waiting distance to the next topology change under the SMC' model. We use these results to show that tree transition omission in ARG/genealogy inference methods is a common problem, causing biases away from the SMC/SMC' assumption that the waiting distance is exponentially distributed with intensity equal to the product of recombination rate and tree length. This challenges the use of such methods for making inferences about recombination rates and recombination rate evolution. The waiting distance between adjacent trees reported by *ARGweaver* is close to what is predicted by theory, and matches that expected for a valid Bayesian ARG sample, suggesting the potential of using *ARGweaver* to estimate recombination rates in a more principled way with this correction. In fact, recombination map estimation was mentioned in Rasmussen et al. (2014), and using the SMC' mode and applying the correction should help. A recent method, *iSMC* (Barroso et al., 2019), uses the pairwise ARG to estimate recombination maps, and we anticipate that it should also benefit from using the correction. However, neither *Relate* nor *tsinfer* provides easily interpretable waiting distances between trees, and under realistic setting in humans, undercounts topology changes and overestimates waiting distances. We emphasize here that this tree missing phenomenon can impact other aspects of inference as well; e.g., because *Relate* misses many trees, the inference result of marginal trees is actually an average of a series of trees, likely leading to biased age estimates of the most recent and ancient coalescence events.

We also highlight the use of ARG inference in understanding the temporal variation of recombination rates, by estimating the time-specific recombination rates using the correction proposed here, which can serve as a tool for understanding the evolution of recombination rates. However, we note that the constant population size demography model may not apply to most scenarios for real data, which will affect the calculation of the probability in (2). We argue that this problem can still be solved because the probabilities can still be calculated with by incorporating the underlying demography model, and the extension is straightforward when the demographic model only involves a single population with changing size. Also, ARGs could be sampled using *ARGweaver-D* (Hubisz et al., 2020) instead of *ARGweaver* under a non-standard demography model. Finally, we note that importance sampling approaches could be used to adjust for model misspecifications.

Acknowledgments

We thank Aaron Stern and Debora Brandt for helpful discussions, Shai Carmi for providing independent calculations for the specific case when $n = 3$ used to verify our results, and Shang Liu for programming suggestions. This research is supported in part by an NIH grant R35-GM134922 to YSS and NIH grant R01GM138634 to RN. YSS is a Chan Zuckerberg Biohub Investigator.

Appendix. Proofs

A.1. Proof of Proposition 1

In order for the tree to not change after a recombination, the recombination needs to start from breakpoint t and end at time τ on the same branch b. At time τ, there will be $A_n(\tau)$ lineages, and only one of them will be the original branch b. The probability density of starting from t and ending at τ while not changing the tree is thus $\frac{1}{A_n(\tau)}p(\tau \mid t)$, where $p(\tau \mid t)$ corresponds to the probability density of the rejoining time τ given the recombination time t. Then we need to integrate over τ since essentially any time older than t and younger than the upper end of b could be a possible choice. The probability for a recombination to start at t on branch b can be thus calculated as:

$$\begin{aligned}
&\mathbb{P}(\text{tree unchanged} \mid b, t, \mathcal{T}) \\
&= \int_t^{t_b^u} \frac{1}{A_n(\tau)} p(\tau \mid t) d\tau \\
&= \int_t^{t_b^u} \exp\left[-\int_t^{\tau} A_n(s) ds\right] d\tau.
\end{aligned}$$

To simplify the notation we define $i = A_n(t)$ so that $t \in [\sigma_{i+1}, \sigma_i]$. Hence, we are basically considering the case when there are i lineages left when the recombination happened. Because the recombination end time τ ranges from t to t_b^u, we can simply break it into intervals, so that in each one of them the number

of remaining lineages, $A_n(\tau)$, is a constant, which leads to:

$$\begin{aligned}&\mathbb{P}(\text{tree unchanged} \mid b, t, \mathcal{T})\\ &= \int_t^{\sigma_i} \exp\left[-\int_t^{\tau} A_n(s)ds\right]d\tau\\ &\quad + \sum_{j=A_n(t_b^u)+1}^{i-1} \int_{\sigma_{j+1}}^{\sigma_j} \exp\left[-\int_t^{\tau} A_n(s)ds\right]d\tau.\end{aligned}$$

Note that for the first term, $A_n(s) = i$ always holds because τ is constrained to $[t, \sigma_i]$ and s is constrained to $[t, \tau]$, which leads to the following simplification:

$$\begin{aligned}\int_t^{\sigma_i} \exp\left(-\int_t^{\tau} A_n(s)ds\right)d\tau &= \int_t^{\sigma_i} \exp\Big(-i(\tau - t)\Big)d\tau\\ &= \frac{1}{i} - \frac{1}{i}e^{-i\sigma_i}e^{it}\\ &= \frac{1}{i} + P_{ii}e^{it}.\end{aligned}$$

For the other terms, note that we can break the inner integral into 2 parts, $[t, \sigma_{j+1}]$ and $[\sigma_{j+1}, \tau]$, such that the first part is independent from τ and the second part has $A_n(s)$ as a constant j. Further leveraging that $A_n(s)$ is a piece-wise constant function, we have

$$\begin{aligned}&\int_{\sigma_{j+1}}^{\sigma_j} \exp\left(-\int_t^{\tau} A_n(s)ds\right)d\tau\\ &= \exp\left(-\int_t^{\sigma_{j+1}} A_n(s)ds\right)\left[\int_{\sigma_{j+1}}^{\sigma_j} \exp\left(-\int_{\sigma_{j+1}}^{\tau} A_n(s)ds\right)d\tau\right]\\ &= \exp\left(-i(\sigma_i - t) - \sum_{k=j+1}^{i-1} kT_k\right)\int_{\sigma_{j+1}}^{\sigma_j} \exp\Big(-j(\tau - \sigma_{j+1})\Big)d\tau\\ &= e^{it}\exp\left(-i\sigma_i - \sum_{k=j+1}^{i-1} kT_k\right)\frac{1}{j}(1 - e^{-jT_j})\\ &= P_{ij}e^{it}.\end{aligned}$$

Combining these terms, we find

$$\mathbb{P}(\text{tree unchanged} \mid b, t, \mathcal{T}) = \frac{1}{i} + e^{it}\left[\sum_{j=A_n(t_b^u)+1}^{i} P_{ij}\right],$$

which completes the proof of Proposition 1.

A.2. Proof of Proposition 2

The probability of type 1 event on a branch b, which is the probability that a recombination does not change the tree conditioned on it happening on branch b, can be found by integrating over the breakpoint time, t, with respect to its conditionally uniform distribution on the branch. As in the proof of Proposition 1, we can break $[t_b^l, t_b^u]$ into multiple intervals as

$$\begin{aligned}\mathbb{P}(\text{tree unchanged} \mid b, \mathcal{T}) &= \frac{1}{t_b^u - t_b^l}\\ &\quad \times \int_{t_b^l}^{t_b^u} \mathbb{P}(\text{tree unchanged} \mid b, t, \mathcal{T})dt\\ &= \frac{1}{t_b^u - t_b^l}\sum_{i=A_n(t_b^u)+1}^{A_n(t_b^l)} \int_{\sigma_{i+1}}^{\sigma_i} \mathbb{P}(\text{tree unchanged} \mid\\ &\quad \times b, t, \mathcal{T})dt,\end{aligned}$$

and using Proposition 1 for the integrand gives

$$\begin{aligned}\int_{\sigma_{i+1}}^{\sigma_i} \mathbb{P}(\text{tree unchanged} \mid b, t, \mathcal{T})dt &= \frac{1}{i}\left[T_i + (e^{i\sigma_{i+1}} - e^{i\sigma_i})\right.\\ &\quad \left.\times \sum_{j=A_n(t_b^u)}^{i} P_{ij}\right]\\ &= p_b^{(i)},\end{aligned}$$

thus yielding

$$\mathbb{P}(\text{tree unchanged} \mid b, \mathcal{T}) = \frac{1}{t_b^u - t_b^l}\sum_{i=A_n(t_b^u)+1}^{A_n(t_b^l)} p_b^{(i)}.$$

A.3. The proof of Theorem 3

For each branch b in the tree $\mathcal{T}$, we denote by b' the lineage that coalesces with it, and denote the parental lineage of b and b' by c (Fig. 8).

We note that the steps of this proof are essentially the same as for the proof of Theorem 1, which is to find $\mathbb{P}(\text{topology unchanged} \mid b, \mathcal{T})$ by integrating $\mathbb{P}(\text{topology unchanged} \mid b, t, \mathcal{T})$, and then use the law of total probability to get $\mathbb{P}(\text{topology unchanged} \mid \mathcal{T})$. The probability calculation will now take into account recombinations of type 1, 2 and 3, instead of just type 1, yielding

$$\begin{aligned}&\mathbb{P}(\text{topology unchanged} \mid b, \mathcal{T})\\ &= \frac{1}{t_b^u - t_b^l}\int_{t_b^l}^{t_b^u} \mathbb{P}(\text{topology unchanged} \mid b, t, \mathcal{T})dt\\ &= \frac{1}{t_b^u - t_b^l}\left[\left(\int_{t_b^l}^{t_{b'}^l} + \int_{t_{b'}^l}^{t_{b'}^u}\right)\mathbb{P}(\text{topology unchanged} \mid b, t, \mathcal{T})dt\right]\\ &= \frac{1}{t_b^u - t_b^l}\left[\left(\sum_{i=A_n(t_{b'}^l)+1}^{A_n(t_b^l)} + \sum_{i=A_n(t_{b'}^u)+1}^{A_n(t_{b'}^l)}\right)\right.\\ &\quad \left.\times \int_{\sigma_{i+1}}^{\sigma_i} \mathbb{P}(\text{topology unchanged} \mid b, t, \mathcal{T})dt\right].\end{aligned} \tag{6}$$

Here, without loss of generality, we assume that the lower end of b is younger than that of b' (i.e., $t_b^l \le t_{b'}^l$), and note that $t_b^u = t_{b'}^u$ since b and b' coalesce (Fig. 8). The reason why we break the range into $[t_b^l, t_{b'}^l]$ and $[t_{b'}^l, t_{b'}^u]$ is explained in the caption of Fig. 8.

First case: $t \in [\sigma_{i+1}, \sigma_i] \subset [t_b^l, t_{b'}^l]$ In this scenario, there are 3 possible ways in which a recombination lineage can coalesce to retain the original topology: the first is to join branch b in $[t, t_{b'}^l]$, the second is to join branch b or b' in $[t_{b'}^l, t_{b'}^u]$, and the third is to join branch c in $[t_{b'}^u, t_c^l]$. They lead to

$$\begin{aligned}&\mathbb{P}(\text{topology unchanged} \mid b, t, \mathcal{T})\\ &= \int_t^{t_{b'}^l} \frac{1}{A_n(\tau)}p(\tau \mid t)d\tau + \int_{t_{b'}^l}^{t_{b'}^u} \frac{2}{A_n(\tau)}p(\tau \mid t)d\tau\\ &\quad + \int_{t_c^l}^{t_c^u} \frac{1}{A_n(\tau)}p(\tau \mid t)d\tau\end{aligned}$$

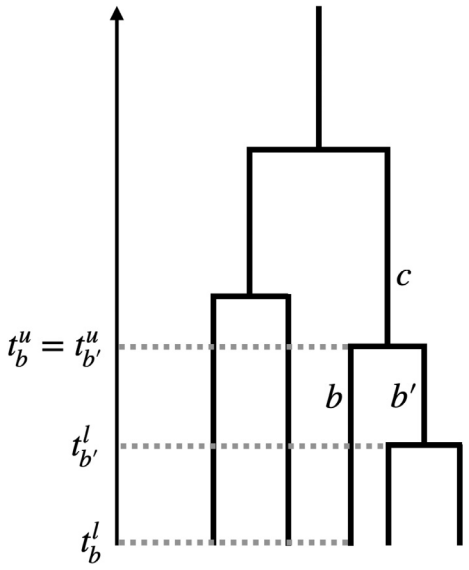

Fig. 8. Illustration of the additional notational conventions. Here we use b and b' to denote a pair of coalescing lineages, with b being the one with lower end, and c to denote the parental lineage of them. The reason why we need to break the proof of Theorem 3 into two cases is that in $[t_b^l, t_{b'}^l]$, the recombining lineage must rejoin b for the topology not to change, while in $[t_{b'}^l, t_{b'}^u]$ it can rejoin either b or b' without changing the topology.

$$= \underbrace{\int_t^{t_{b'}^l} \exp\left(-\int_t^\tau A_n(s)ds\right)d\tau}_{P_1} + 2\underbrace{\int_{t_{b'}^l}^{t_{b'}^u} \exp\left(-\int_t^\tau A_n(s)ds\right)d\tau}_{P_2} + \underbrace{\int_{t_c^l}^{t_c^u} \exp\left(-\int_t^\tau A_n(s)ds\right)d\tau}_{P_3}, \tag{7}$$

and, using similar ideas to those in the proof of Proposition 1, we can simplify the summands as follows:

$$\begin{aligned}
P_1 &= \int_t^{\sigma_i} \exp\left(-\int_t^\tau A_n(s)ds\right)d\tau \\
&\quad + \sum_{j=A_n(t_{b'}^l)+1}^{i-1} \int_{\sigma_{j+1}}^{\sigma_j} \exp\left(-\int_t^\tau A_n(s)ds\right)d\tau \\
&= \frac{1}{i} + P_{ii}e^{it} + \sum_{j=A_n(t_{b'}^l)+1}^{i-1} P_{ij}e^{it} \\
P_2 &= 2\sum_{j=A_n(t_{b'}^u)+1}^{A_n(t_{b'}^l)} \int_{\sigma_{j+1}}^{\sigma_j} \exp\left(-\int_t^\tau A_n(s)ds\right)d\tau \\
&= 2\sum_{j=A_n(t_{b'}^u)+1}^{A_n(t_{b'}^l)} P_{ij}e^{it} \\
P_3 &= \sum_{j=A_n(t_c^u)+1}^{A_n(t_c^l)} \int_{\sigma_{j+1}}^{\sigma_j} \exp\left(-\int_t^\tau A_n(s)ds\right)d\tau \\
&= \sum_{j=A_n(t_c^u)+1}^{A_n(t_c^l)} P_{ij}e^{it}.
\end{aligned}$$

Plugging these expressions back into (7), we obtain

$$\begin{aligned}
&\mathbb{P}(\text{topology unchanged} \mid b, t, \mathcal{T}) \\
&= \frac{1}{i} + A_{ii}e^{it} + \sum_{j=A_n(t_{b'}^l)+1}^{i-1} P_{ij}e^{it} + 2\sum_{j=A_n(t_{b'}^u)+1}^{A_n(t_{b'}^l)} P_{ij}e^{it} \\
&\quad + \sum_{j=A_n(t_c^u)+1}^{A_n(t_c^l)} P_{ij}e^{it} \\
&= \frac{1}{i} + \sum_{j=A_n(t_c^u)+1}^{i} P_{ij}e^{it} + \sum_{j=A_n(t_{b'}^u)+1}^{A_n(t_{b'}^l)} P_{ij}e^{it},
\end{aligned}$$

and we can easily do the integration with respect to t to obtain

$$\begin{aligned}
&\int_{\sigma_{i+1}}^{\sigma_i} \mathbb{P}(\text{topology unchanged} \mid b, t, \mathcal{T})dt \\
&= \frac{1}{i}\left[t_i + \left(A_{ii} + \sum_{j=A_n(t_{b'}^l)+1}^{i-1} P_{ij} + 2\sum_{j=A_n(t_{b'}^u)+1}^{A_n(t_{b'}^l)} P_{ij} + \sum_{j=A_n(t_c^u)+1}^{A_n(t_c^l)} P_{ij} \right) \right. \\
&\quad \left. \times (e^{i\sigma_i} - e^{i\sigma_{i+1}}) \right] \\
&= \frac{1}{i}\left[t_i + \left(\sum_{j=A_n(t_{b'}^u)+1}^{A_n(t_{b'}^l)} P_{ij} + \sum_{j=A_n(t_c^u)+1}^{i} P_{ij} \right)(e^{i\sigma_i} - e^{i\sigma_{i+1}}) \right] \\
&= p_{b,1}^{(i)}.
\end{aligned}$$

Second case: $t \in [\sigma_{i+1}, \sigma_i] \subset [t_{b'}^l, t_{b'}^u]$ In this case there are 2 possible ways in which a recombining lineage can coalesce to retain the original topology: the first is to join branch b or b' in $[t_{b'}^l, t_{b'}^u]$, and the other is to join branch c in $[t_{b'}^u, t_c^l]$. Hence,

$$\begin{aligned}
&\mathbb{P}(\text{topology unchanged} \mid b, t, \mathcal{T}) \\
&= \int_t^{t_{b'}^u} \frac{2}{A_n(\tau)} p(\tau \mid t)d\tau + \int_{t_c^l}^{t_c^u} \frac{1}{A_n(\tau)} p(\tau \mid t)d\tau \\
&= 2\int_t^{t_{b'}^u} \exp\left(-\int_t^\tau A_n(s)ds\right)d\tau + \int_{t_c^l}^{t_c^u} \exp\left(-\int_t^\tau A_n(s)ds\right)d\tau \\
&= 2\left[\int_t^{\sigma_i} \exp\left(-\int_t^\tau A_n(s)ds\right)d\tau \right. \\
&\quad \left. + \sum_{j=A_n(t_{b'}^u)+1}^{i-1} \int_{\sigma_{j+1}}^{\sigma_j} \exp\left(-\int_t^\tau A_n(s)ds\right)d\tau \right] \\
&\quad + \sum_{j=A_n(t_c^u)+1}^{A_n(t_c^l)} \int_{\sigma_{j+1}}^{\sigma_j} \exp\left(-\int_t^\tau A_n(s)ds\right)d\tau \\
&= 2\left(\frac{1}{i} + P_{ii}e^{it} + \sum_{j=A_n(t_{b'}^u)+1}^{i-1} P_{ij}e^{it} \right) + \sum_{j=A_n(t_c^u)+1}^{A_n(t_c^l)} P_{ij}e^{it} \\
&= 2\left(\frac{1}{i} + \sum_{j=A_n(t_{b'}^u)+1}^{i} P_{ij}e^{it} \right) + \sum_{j=A_n(t_c^u)+1}^{A_n(t_c^l)} P_{ij}e^{it},
\end{aligned}$$

and we can integrate with respect to t again to obtain

$$\begin{aligned}
&\int_{\sigma_{i+1}}^{\sigma_i} \mathbb{P}(\text{topology unchanged} \mid b, t, \mathcal{T}) dt \\
&= \frac{1}{i}\left[2t_i + \left(2A_{ii} + 2\sum_{j=A_n(t_{b'}^u)+1}^{i-1} P_{ij} + \sum_{j=A_n(t_c^u)+1}^{A_n(t_c^l)} P_{ij}\right)(e^{i\sigma_i} - e^{i\sigma_{i+1}})\right] \\
&= \frac{1}{i}\left[2t_i + \left(2\sum_{j=A_n(t_c^u)+1}^{i} P_{ij} - \sum_{j=A_n(t_c^u)+1}^{A_n(t_c^l)} P_{ij}\right)(e^{i\sigma_i} - e^{i\sigma_{i+1}})\right] \\
&=: p_{b,2}^{(i)}.
\end{aligned}$$

Now, we can plug these results into (6) to get

$$\begin{aligned}
&\mathbb{P}(\text{topology unchanged} \mid b, \mathcal{T}) \\
&= \frac{1}{t_b^u - t_b^l}\left[\sum_{i=A_n(t_{b'}^l)+1}^{A_n(t_b^l)} p_{b,1}^{(i)} + \sum_{i=A_n(t_{b'}^u)+1}^{A_n(t_{b'}^l)} p_{b,2}^{(i)}\right],
\end{aligned}$$

which leads to the following conclusion using the law of total probability:

$$\begin{aligned}
\mathbb{P}(\text{topology unchanged} \mid \mathcal{T}) &= \sum_{b=1}^{2n-2} \frac{t_b^u - t_b^l}{L(\mathcal{T})} \\
&\quad \times \mathbb{P}(\text{topology unchanged} \mid b, \mathcal{T}) \\
&= \frac{1}{L(\mathcal{T})}\sum_{b=1}^{2n-2}\left[\sum_{i=A_n(t_{b'}^l)+1}^{A_n(t_b^l)} p_{b,1}^{(i)} \right. \\
&\quad \left. + \sum_{i=A_n(t_{b'}^u)+1}^{A_n(t_{b'}^l)} p_{b,2}^{(i)}\right].
\end{aligned}$$

References

Barroso, G.V., Puzović, N., Dutheil, J.Y., 2019. Inference of recombination maps from a single pair of genomes and its application to ancient samples. PLoS Genet. 15 (11), e1008449.

Carmi, S., Wilton, P.R., Wakeley, J., Pe'er, I., 2014. A renewal theory approach to IBD sharing. Theor. Popul. Biol. 97, 35–48.

Fearnhead, P., Donnelly, P., 2001. Estimating recombination rates from population genetic data. Genetics 159 (3), 1299–1318.

Griffiths, R., 1981. Neutral two-locus multiple allele models with recombination. Theor. Popul. Biol. 19, 169–186.

Griffiths, R.C., Marjoram, P., 1996. Ancestral inference from samples of DNA sequences with recombination. J. Comput. Biol. 3 (4), 479–502.

Gusfield, D., 2014. ReCombinatorics: The Algorithmics of Ancestral Recombination Graphs and Explicit Phylogenetic Networks. MIT press.

Gusfield, D., Eddhu, S., Langley, C., 2003. Efficient reconstruction of phylogenetic networks with constrained recombination. In: Computational Systems Bioinformatics. CSB2003. Proceedings of the 2003 IEEE Bioinformatics Conference. CSB2003. IEEE, pp. 363–374.

Hein, J., 1993. A heuristic method to reconstruct the history of sequences subject to recombination. J. Mol. Evol. 36 (4), 396–405.

Hein, J., Schierup, M., Wiuf, C., 2004. Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory. Oxford University Press, USA.

Hobolth, A., Jensen, J.L., 2014. MarkovIan approximation to the finite loci coalescent with recombination along multiple sequences. Theor. Popul. Biol. 98, 48–58.

Hubisz, M.J., Williams, A.L., Siepel, A., 2020. Mapping gene flow between ancient hominins through demography-aware inference of the ancestral recombination graph. PLoS Genet. 16 (8), e1008895.

Hudson, R.R., 1983. Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. 23 (2), 183–201.

Hudson, R., 2002. Generating samples under the wright-Fisher neutral model of genetic variation. Bioinformatics 18 (2), 337–338.

Hudson, R.R., Kaplan, N.L., 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 111 (1), 147–164.

Ignatieva, A., Lyngsø, R.B., Jenkins, P., Jotun, H., 2020. Kwarg: Parsimonious reconstruction of ancestral recombination graphs with recurrent mutation.. bioRxiv preprint, https://doi.org/10.1101/2020.12.17.423233.

Jenkins, P.A., Griffiths, R.C., 2011. Inference from samples of DNA sequences using a two-locus model. J. Comput. Biol. 18 (1), 109–127.

Kelleher, J., Etheridge, A.M., McVean, G., 2016. Efficient coalescent simulation and genealogical analysis for large sample sizes. PLoS Comput. Biol. 12 (5), e1004842.

Kelleher, J., Wong, Y., Wohns, A.W., Fadil, C., Albers, P.K., McVean, G., 2019. Inferring whole-genome histories in large population datasets. Nature Genet. 51 (9), 1330–1338.

Kingman, J.F.C., 1982a. The coalescent. Stochastic Process. Appl. 13 (3), 235–248.

Kingman, J.C., 1982b. On the genealogy of large populations. J. Appl. Probab. 19A, 27–43.

Li, H., Durbin, R., 2011. Inference of human population history from individual whole-genome sequences. Nature 475 (7357), 493–496.

Lyngsø, R.B., Song, Y.S., Hein, J., 2005. Minimum recombination histories by branch and bound. In: International Workshop on Algorithms in Bioinformatics. Springer, pp. 239–250.

Marjoram, P., Wall, J.D., 2006. Fast "coalescent" simulation. BMC Genet. 7 (1), 16.

McVean, G.A., Cardin, N.J., 2005. Approximating the coalescent with recombination. Philos. Trans. R. Soc. B 360 (1459), 1387–1393.

Rasmussen, M.D., Hubisz, M.J., Gronau, I., Siepel, A., 2014. Genome-wide inference of ancestral recombination graphs. PLoS Genet. 10 (5), e1004342.

Schiffels, S., Durbin, R., 2014. Inferring human population size and separation history from multiple genome sequences. Nature Genet. 46 (8), 919–925.

Song, Y.S., Hein, J., 2003. Parsimonious reconstruction of sequence evolution and haplotype blocks. In: International Workshop on Algorithms in Bioinformatics. Springer, pp. 287–302.

Speidel, L., Forest, M., Shi, S., Myers, S.R., 2019. A method for genome-wide genealogy estimation for thousands of samples. Nature Genet. 51 (9), 1321–1329.

Terhorst, J., Kamm, J.A., Song, Y.S., 2017. Robust and scalable inference of population history from hundreds of unphased whole genomes. Nature Genet. 49 (2), 303–309.

Wang, L., Zhang, K., Zhang, L., 2001. Perfect phylogenetic networks with recombination. J. Comput. Biol. 8 (1), 69–78.

Wilton, P.R., Carmi, S., Hobolth, A., 2015. The smc' is a highly accurate approximation to the ancestral recombination graph. Genetics 200 (1), 343–355.

Wiuf, C., Hein, J., 1999. Recombination as a point process along sequences. Theor. Popul. Biol. 55 (3), 248–259.