The distribution of branch duration and detection of inversions in ancestral recombination graphs

Anastasia Ignatieva[1,2†], Martina Favero[3*], Jere Koskela[3*], Jaromir Sant[1*], and Simon R. Myers[1,4]

[1]Department of Statistics, University of Oxford
[2]School of Mathematics and Statistics, University of Glasgow
[3]Department of Statistics, University of Warwick
[4]Wellcome Centre for Human Genetics, University of Oxford
[*]Contributed equally, ordered alphabetically
[†]Correspondence: anastasia.ignatieva@glasgow.ac.uk

July 11, 2023

Abstract

Recent breakthroughs in the reconstruction of ancestral recombination graphs (ARGs) have meant that inference of genome-wide genealogies from large sequencing datasets is now possible. This development has also opened several new directions where research is urgently needed: improving our understanding of which aspects of genealogies are (and which are not) faithfully captured by reconstruction tools, as well as the development of inference methods that directly use ARGs as inputs. By deriving the distribution of branch duration under the SMC' model, defined as the genomic interval spanned by a branch of the ARG until it is disrupted by recombination, we benchmark the quality of ARGs produced by several reconstruction tools: ARGweaver, Relate, tsinfer and tsdate, and ARG-Needle. Further, we develop an accurate and powerful ARG-based test for the presence of chromosomal inversions, based on the idea that suppression of recombination in individuals heterozygous for an inversion has a detectable impact on branch duration.

1 Introduction

In the presence of recombination, the genealogical history of a sample can be fully captured in the form of an ancestral recombination graph (ARG). This was first described by Griffiths and Marjoram (1997) as a realisation of the coalescent with recombination (CwR), a stochastic process operating backwards in time, generating a genealogy through a sequence of coalescence and recombination events (Hudson, 1983). Wiuf and Hein (1999) reframed the CwR as a stochastic process operating spatially along the genome: starting with the leftmost endpoint, local trees are generated sequentially moving to the right, reshaped by recombination events. While calculating the properties of the ARG under both frameworks is generally intractable, this seminal work spurred on a suite of simplifying approximations, enabling applications to large-scale genomic data.

The sequentially Markovian coalescent (SMC) model, proposed by McVean and Cardin (2005), imposed the assumption that the process along the genome is Markovian, which, in essence, prohibits recombination events in genetic material not ancestral to the sample. This was followed by the SMC' extension (Marjoram and Wall, 2006), which was shown to be an excellent approximation to the CwR, based on the joint distribution of pairwise coalescent times and a quantification of bias in population size estimates (Wilton et al., 2015), and the distribution of the next local tree conditional on the current one in a two-locus model (Hobolth and Jensen, 2014). Thus, for a small trade-off in accuracy, the SMC' model offered a substantially more tractable way of calculating analytic approximations to various quantities of interest, such as the correlation between coalescence time and linkage probability for a randomly sampled pair of sequences (Eriksson et al., 2009), and identity-by-descent tract length distributions and related quantities (Harris and Nielsen, 2013; Carmi

et al., 2014). It also enabled the development of powerful new inference methods: for instance, by considering the genealogy of a single pair of sequences, Li and Durbin (2011) developed a HMM-based approach (the pairwise SMC, or PSMC) for inferring the history of human population sizes, which was subsequently extended by Schiffels and Durbin (2014) to multiple samples.

Meanwhile, the definition of the ARG has become decoupled from its initial description as the realisation of a stochastic process, to more broadly denote a genealogical network that captures genetic inheritance. Using this looser definition, the problem of explicitly reconstructing plausible ARGs from sequencing data has seen significant recent progress driven by the use of heuristic methods and principled approximations to the CwR. ARGweaver (Rasmussen et al., 2014) implements an MCMC scheme based on a time-discretised version of the SMC (or SMC') to obtain posterior samples of ARGs compatible with a given dataset. Relate (Speidel et al., 2019) and tsinfer/tsdate (Kelleher et al., 2019; Wohns et al., 2022) reconstruct a single ARG from data, by using methods based on the Li and Stephens (2003) framework to first reconstruct the topologies and then estimating the branch lengths using Bayesian approaches with coalescent-based priors. ARG-Needle (Zhang et al., 2023) reconstructs a single ARG by sequentially threading in each sample, by first identifying the most closely related samples already in the ARG via genotype hashing, and subsequently estimating coalescence times under the Ascertained Sequentially Markovian Coalescent (ASMC) model (a coalescent-based HMM). These methods scale to thousands of human genome-length samples and have already been applied to many large-scale datasets, resulting in powerful inference of evolutionary events and parameters, such as the history of human demography (Wohns et al., 2022), past population sizes (Speidel et al., 2019), signals of selection (Hejase et al., 2022), and genetic associations for complex traits (Zhang et al., 2023).

Methods for quantifying the quality of ARG reconstruction have so far been limited to simulation studies calculating the accuracy of inferred local tree topologies (Speidel et al., 2019; Kelleher et al., 2019) and pairwise coalescence times (Brandt et al., 2022). While calculating quantities of interest is generally very difficult under the CwR model, the tractability of the SMC' model makes it possible to derive metrics under which reconstructed ARGs can be compared against the ground truth in coalescent-based simulations, and hence identify which features are and are not captured faithfully. Only a small number of studies have so far explored this direction. Deng et al. (2021) calculated the probability that a single recombination event is tree-changing or topology-changing under the SMC' and hence derived an approximation to the distribution of the genomic distance between local trees, finding that Relate and tsinfer generally overestimate this quantity; McKenzie and Eaton (2022) extended these results to the multispecies coalescent model.

Moreover, ARGs can, in principle, capture the effects of all the evolutionary forces that have shaped the observed genetic diversity of a sample of sequences, while providing a much more efficient representation of genomes than multiple sequence alignments (Kelleher et al., 2019). Thus, ARG-based statistical inference methods have the potential to provide very powerful insights into various evolutionary events and parameters. This is a rapidly developing field, however progress has been hampered by the fact that methods that work very well on simulated ARGs often lose power and accuracy when applied to reconstructed genealogies, for reasons that are, in general, poorly understood. We thus have two main goals: to use the tractable SMC' model to better understand the ways in which reconstructed ARGs differ from simulated ones, and to use this model to develop genealogy-based inference methods.

We present a number of results in this direction, focusing on the idea of branch duration. For a given edge of the ARG, this can be defined as the span of the genome in which the branch is present in the local tree (as illustrated in Figure 1), which is determined by recombination events that change local tree topologies and branch lengths. Intuitively, the longer (resp. shorter) the *time-length* of a branch, the more (resp. less) likely it is to be broken up by recombination as we move along the genome, so its *duration* along the genome should be shorter (resp. longer). This is a fundamental property of the ARG which has several implications. For instance, in the absence of recombination,

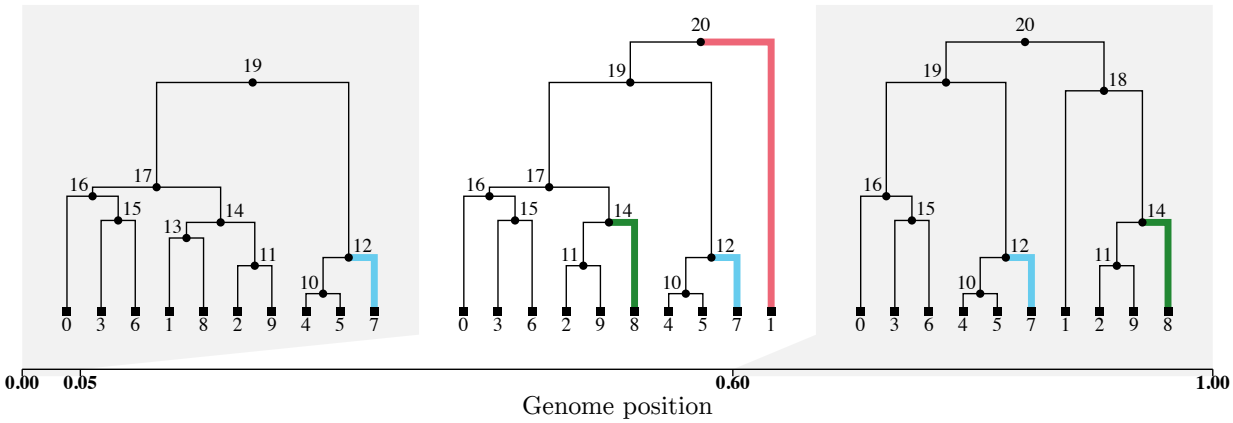

Figure 1: Local trees of an ARG with $n = 10$ samples. Branch highlighted in red ($1 \to 20$) spans one local tree and has duration 0.55; branch highlighted in green ($8 \to 14$) spans two local trees and has duration 0.95; branch highlighted in blue ($7 \to 12$) spans all three local trees and has duration 1. ARG simulated using msprime and visualised using tskit (Baumdicker et al., 2022; Kelleher et al., 2016).

the genealogy is a single tree with all branch durations equal to the length of the genome, and the probability that a given branch carries at least one mutation is inversely proportional to its time-length (assuming mutations occur as a Poisson process along the branches). On the other hand, in the presence of recombination, the mutation rate per branch is less variable, because of the interplay between branch time-length and duration. Thus, the distribution of the duration of a given branch conditional on the local tree in which it first arises is a fundamental property of the ARG arising from the correlations between local trees and, to the best of our knowledge, has not been previously derived analytically.

We calculate the probability that a particular branch of the ARG is disrupted by the next recombination event under the SMC' model and use this to construct an accurate approximation of the distribution of branch duration along the genome. Through simulation studies, we use this to compare how closely the SMC and SMC' models approximate the CwR, and how well the aforementioned reconstruction methods estimate branch durations for ARGs simulated under these models. We also obtain several results demonstrating the effect of a recombination event on the height and total branch length of a given local tree.

Further, we propose a very close approximation to the distribution of *clade* duration along the genome and use this to construct an ARG-based test for detecting chromosomal inversions. This is based on the idea that recombination is suppressed in individuals heterozygous for an inversion, which manifests in the ARG as a clade of samples that persists for a longer stretch of the genome than would otherwise be expected. We demonstrate the utility of this test for both simulated and reconstructed ARGs.

Code implementing the results and used to produce the figures is publicly available at `github.com/a-ignatieva/argbd`.

2 Methods

2.1 Notation and background

The notation is illustrated in Figure 2. Let $\mathcal{T}$ be a fixed local tree with n leaves. Denote by T_i (for $i \in \{2, \cdots, n\}$) the population-scaled time at which the number of lineages in $\mathcal{T}$ jumps from i to $i-1$, with T_2 being the time of MRCA and setting $T_{n+1} := 0$. Let $n(t)$ be the number of lineages at time t, so $n(0) = n$, with $n(T_j) = j - 1$ and $n(t) = 1$ for $t \geq T_2$.

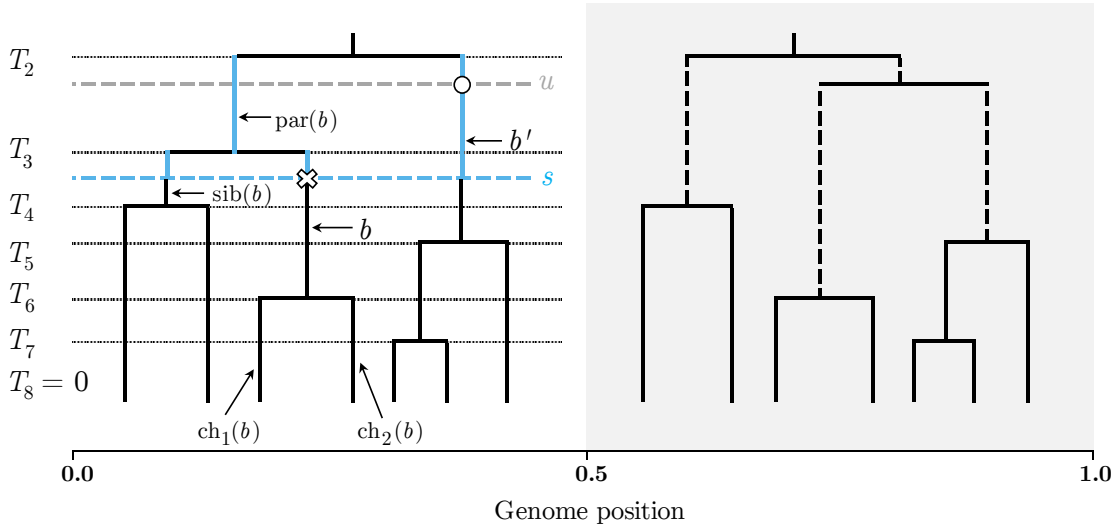

Figure 2: Illustration of the notation used throughout. The ARG has two marginal trees, where the tree on the left is $\mathcal{T}$, with $n = 7$. Coalescent event times are shown as black dotted lines. For the branch labelled b, $t^{\uparrow}(b) = T_3$, $t^{\downarrow}(b) = T_6$, $d^{\leftarrow}(b) = 0$, $d^{\rightarrow}(b) = 0.5$ (so the duration of the branch is 0.5); $\mathcal{A}(b) = \{b, \text{sib}(b), \text{par}(b)\}$ and $\mathcal{B}(b) = \{b, \text{sib}(b), \text{ch}_1(b), \text{ch}_2(b)\}$. The recombination event occurs at genomic position 0.5; the recombination point $\mathcal{R} = (b, s)$ is shown as a cross; the coalescence point $\mathcal{C} = (b', u)$ is shown as a circle; $n(s) = 3$ and $L_{\mathcal{T}}(s)$ gives the total length of the branches shown in blue. The tree on the right $\mathcal{T}'$ is obtained by pruning the subtree below $\mathcal{R}$ and reattaching at $\mathcal{C}$; solid vertical lines show branches that have not been affected by the recombination event.

For a branch $b \in \mathcal{T}$, denote the lower end time by $t^{\downarrow}(b) \geq 0$ and the upper end time by $t^{\uparrow}(b) \leq T_2$, with the *time-length* of the branch given by $\bar{t}(b) = t^{\uparrow}(b) - t^{\downarrow}(b)$. Let $d^{\leftarrow}(b)$ and $d^{\rightarrow}(b)$ be the leftmost and rightmost endpoints of the genomic span of branch b, respectively, with its *duration* given by $d^{\rightarrow}(b) - d^{\leftarrow}(b)$.

Let $\text{par}(b), \text{sib}(b), \text{ch}_1(b)$ and $\text{ch}_2(b)$ denote the parent, sibling, left child and right child branch of b respectively, such that we have the following relations:

$$
\begin{aligned}
&t^{\downarrow}(\text{par}(b)) = t^{\uparrow}(b) \\
&t^{\uparrow}(\text{sib}(b)) = t^{\uparrow}(b), \\
&t^{\uparrow}(\text{ch}_1(b)) = t^{\uparrow}(\text{ch}_2(b)) = t^{\downarrow}(b), \\
&\text{ch}_1(b) = \text{ch}_2(b) = \emptyset \text{ if } t^{\downarrow}(b) = 0 \text{ (} b \text{ extends from a leaf node).}
\end{aligned}
$$

Define the sets of branches $\mathcal{A}(b) := \{b, \text{sib}(b), \text{par}(b)\}$ and $\mathcal{B}(b) := \{b, \text{sib}(b), \text{ch}_1(b), \text{ch}_2(b)\}$, and denote by b_r the root lineage extending past the MRCA node.

Let $L_{\mathcal{T}}(t)$ be the sum of branch lengths in $\mathcal{T}$ above time t and up to T_2:

$$
L_{\mathcal{T}}(t) = \sum_{b \in \mathcal{T}: t^{\uparrow}(b) > t} \left[t^{\uparrow}(b) - \max(t, t^{\downarrow}(b)) \right],
$$

so that $L_{\mathcal{T}}(0)$ is the total branch length of $\mathcal{T}$. Denote by $\mathcal{T}_x$ the local tree at position x along the genome.

Under the SMC', moving along the genome, a recombination event happens after an exponentially distributed waiting time with rate $\mathcal{L}_{\mathcal{T}}(0) \cdot \rho/2$; when this event happens, a location $\mathcal{R}$ is selected uniformly at random along the branches of $\mathcal{T}$, say on branch b at time s, which we denote as $\mathcal{R} \in b$ or $\mathcal{R} = (b, s)$. A new coalescence point $\mathcal{C}$ is selected by allowing the recombining lineage to

coalesce at rate 1 with all the lineages present above time s (including b_r). We denote a coalescence point on branch b' at time u as $\mathcal{C} \in b'$ or $\mathcal{C} = (b', u)$. The next tree along the genome is then formed by pruning the subtree below the recombination point $\mathcal{R}$ and reattaching it at the chosen coalescence point $\mathcal{C}$. We write $\mathcal{R} \in \mathcal{B}(b)$ to mean that the recombination point is on one of the branches in $\mathcal{B}(b)$.

The difference with the spatial formulation of the CwR is that the coalescence point is restricted to be on the local tree $\mathcal{T}$, whereas under the CwR it could be placed on any branches of the ARG corresponding to the full sequence of trees to the left of the recombination position. In essence, the SMC' approximation disallows any recombination events that occur in non-ancestral material, making the process Markovian along the genome.

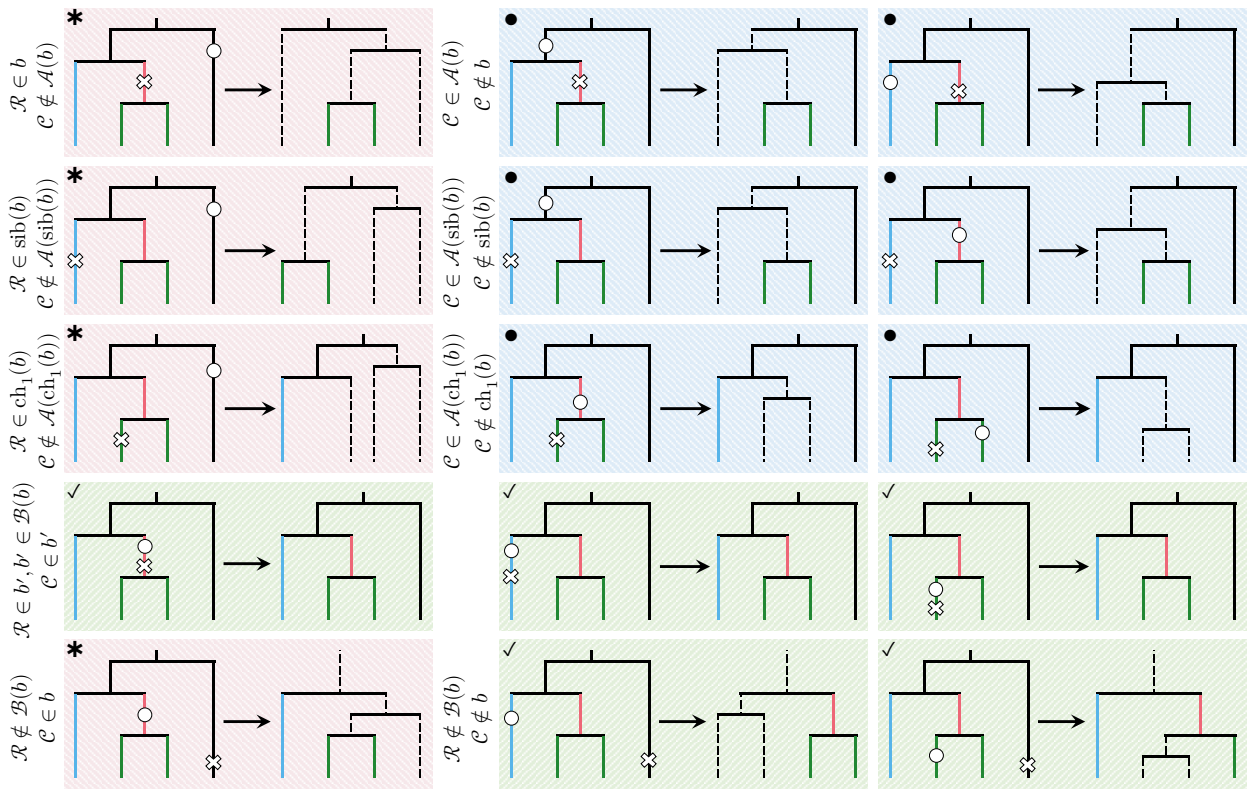

Figure 3: Possible events that do and do not disrupt branch b (shown in red); $\mathrm{sib}(b)$ is shown in blue, $\mathrm{ch}_1(b)$ and $\mathrm{ch}_2(b)$ in green. Recombination points are shown as crosses; coalescence points as circles. Branches that are disrupted (or newly added) are shown as dashed lines. Events highlighted in green (marked with ticks) do not disrupt b. Events highlighted in red (marked with stars) disrupt the branch in terms of both branch length and topology (b is topologically disrupted); those highlighted in blue (marked with dots) disrupt b via changing only its branch length.

The difference between the SMC and SMC' models is that under the SMC', the coalescence point can be chosen on the same branch b containing the recombination point, above time s (so that recombinations can occur that do not change the tree topology or branch lengths), whereas events of this type are disallowed under the SMC.

2.2 Probability that a branch is disrupted

Considering a fixed branch $b \in \mathcal{T}$, when a recombination event occurs, we would like to know the probability that b is affected by this recombination event. This includes both changes in the branch length of b and events that change the topology of the clade around b (we will say in these

cases that b is *topologically* disrupted by the recombination). This can happen because either (1) the recombination point is on $b' \in \mathcal{B}(b)$ and the coalescence point is not on b', or (2) the recombination point is not on $\mathcal{B}(b)$ and the coalescence point is on b. The possible scenarios that do and do not disrupt b are illustrated in Figure 3.

We have, for a given tree $\mathcal{T}$,

$$\begin{aligned} 1 &= \mathbb{P}_{\mathcal{T}}(\mathcal{R} \in \mathcal{B}(b)) + \mathbb{P}_{\mathcal{T}}(\mathcal{R} \notin \mathcal{B}(b)) \\ &= \sum_{b' \in \mathcal{B}(b)} \left[\mathbb{P}_{\mathcal{T}}(\mathcal{C} \notin b' | \mathcal{R} \in b') + \mathbb{P}_{\mathcal{T}}(\mathcal{C} \in b' | \mathcal{R} \in b')\right] \cdot \mathbb{P}_{\mathcal{T}}(\mathcal{R} \in b') \\ &\quad + \left[\mathbb{P}_{\mathcal{T}}(\mathcal{C} \in b | \mathcal{R} \notin \mathcal{B}(b)) + \mathbb{P}_{\mathcal{T}}(\mathcal{C} \notin b | \mathcal{R} \notin \mathcal{B}(b))\right] \cdot \mathbb{P}_{\mathcal{T}}(\mathcal{R} \notin \mathcal{B}(b)). \end{aligned}$$

The probability that b is disrupted by the recombination event is thus

$$\begin{aligned} \mathbb{P}_{\mathcal{T}}(b \text{ disrupted}) &= \sum_{b' \in \mathcal{B}(b)} \mathbb{P}_{\mathcal{T}}(\mathcal{R} \in b' \text{ and } \mathcal{C} \notin b') + \mathbb{P}_{\mathcal{T}}(\mathcal{R} \notin \mathcal{B}(b) \text{ and } \mathcal{C} \in b) \\ &= \sum_{b' \in \mathcal{B}(b)} \mathbb{P}_{\mathcal{T}}(\mathcal{C} \notin b' | \mathcal{R} \in b') \cdot \mathbb{P}_{\mathcal{T}}(\mathcal{R} \in b') \\ &\quad + \mathbb{P}_{\mathcal{T}}(\mathcal{C} \in b | \mathcal{R} \notin \mathcal{B}(b)) \cdot \mathbb{P}_{\mathcal{T}}(\mathcal{R} \notin \mathcal{B}(b)). \end{aligned} \tag{2.1}$$

We now calculate each of these probabilities in turn for an arbitrary branch $\beta \in \mathcal{T}$ under the SMC'.

2.2.1 Probability recombination point is on branch β

Under the SMC', the recombination point is chosen uniformly at random along the branches of the tree. Thus, the probability that the recombination event happens on branch β is the ratio of the branch length to the total branch length of $\mathcal{T}$, so

$$\mathbb{P}_{\mathcal{T}}(\mathcal{R} \in \beta) = \frac{\bar{t}(\beta)}{L_{\mathcal{T}}(0)}, \tag{2.2}$$

and

$$\mathbb{P}_{\mathcal{T}}(\mathcal{R} \notin \mathcal{B}(\beta)) = 1 - \sum_{b' \in \mathcal{B}(\beta)} \frac{\bar{t}(b')}{L_{\mathcal{T}}(0)}. \tag{2.3}$$

2.2.2 Probability coalescence point is not on β given recombination point is on β

The probability that, conditional on the recombination event happening on branch β at time s, the coalescence point is on β has been derived by Deng et al. (2021), which in our notation is as follows.

Proposition 2.1 (Deng et al. (2021), Proposition 1)**.** *Letting $k = n(s)$, so that T_k is the first coalescence time just above time s,*

$$\mathbb{P}_{\mathcal{T}}(\mathcal{C} \in \beta | \mathcal{R} = (\beta, s)) = \frac{1}{k} + e^{ks} \cdot \sum_{j=n(t^{\uparrow}(\beta))+1}^{k} Q_{kj}, \tag{2.4}$$

where

$$Q_{kk} := -\frac{1}{k} e^{-kT_k}, \tag{2.5}$$

and

$$Q_{kj} = e^{-kT_k} e^{-L_{\mathcal{T}}(T_k)} \frac{1}{j} \left(e^{L_{\mathcal{T}}(T_{j+1})} - e^{L_{\mathcal{T}}(T_j)}\right). \tag{2.6}$$

Marginalising out the recombination time s, hence summing over $k = n(s)$ in (2.4), gives the following.

Proposition 2.2 (Deng et al. (2021), Proposition 2)**.**

$$\mathbb{P}_{\mathcal{T}}(\mathcal{C} \notin \beta | \mathcal{R} \in \beta) = 1 - \frac{1}{\bar{t}(\beta)} \sum_{k=n(t^{\uparrow}(\beta))+1}^{n(t^{\downarrow}(\beta))} \left(\widetilde{Q}^1(k) + \widetilde{Q}^2(k, k, n(t^{\uparrow}(\beta)) + 1, 0, 1) \right), \tag{2.7}$$

where

$$\widetilde{Q}^1(k) := \frac{1}{k} \left(T_k - T_{k+1} \right), \tag{2.8}$$

and for $x, y, A, B \in \mathbb{Z}$, $x \geq k$, $2 \leq y \leq x$,

$$\widetilde{Q}^2(k, x, y, A, B) := \frac{1}{k} \left(e^{kT_k} - e^{kT_{k+1}} \right) \sum_{j=y}^{x} (Aj + B) \cdot Q_{kj}. \tag{2.9}$$

The proofs, translated into our notation, are given in the Appendix, Sections A1 and A2.

2.2.3 Probability coalescence point is on β given recombination point is not on $\mathcal{B}(\beta)$

We start by conditioning on the recombination time s to obtain the following.

Proposition 2.3. *Conditional on the recombination point* $\mathcal{R}$ *being at time* s *and on a branch outside the set* $\mathcal{B}(\beta)$*, with* $k = n(s)$*, the probability that* β *is disrupted is*

$$\mathbb{P}_{\mathcal{T}}(\mathcal{C} \in \beta | \mathcal{R} \notin \mathcal{B}(\beta), \mathcal{R} = (\cdot, s)) = \begin{cases} e^{ks} \sum_{j=n(t^{\uparrow}(\beta))+1}^{n(t^{\downarrow}(\beta))} Q_{kj} & s < t^{\downarrow}(\beta) \\ \frac{1}{k} + e^{ks} \sum_{j=n(t^{\uparrow}(\beta))+1}^{k} Q_{kj} & t^{\downarrow}(\beta) \leq s < t^{\uparrow}(\beta) \\ 0 & \textit{otherwise.} \end{cases} \tag{2.10}$$

with Q_{kk} *and* Q_{kj} *as defined in* (2.5) *and* (2.6)*, respectively.*

The proof is given in the Appendix, Section A3.

Let t_1, t_2, t_3, t_4 denote the event times $t^{\downarrow}(\mathrm{ch}_1(\beta)), t^{\downarrow}(\mathrm{ch}_2(\beta)), t^{\downarrow}(\mathrm{sib}(\beta)), t^{\downarrow}(\beta)$ sorted in increasing order, and define $t_0 := 0$ and $t_5 := t^{\uparrow}(\beta)$. Integrating out the recombination time in (2.10), we have the following.

Proposition 2.4. *Conditional on the recombination point being on a branch outside the set* $\mathcal{B}(\beta)$*, the probability that* β *is disrupted is*

$$\begin{aligned} \mathbb{P}_{\mathcal{T}}(\mathcal{C} \in \beta | \mathcal{R} \notin \mathcal{B}(\beta)) = \frac{1}{L_{\mathcal{T}}(0) - \sum_{b' \in \mathcal{B}(\beta)} \bar{t}(b')} \Bigg\{ & \sum_{k=n(t_1)+1}^{n} k \widetilde{Q}^2(k, n(t^{\downarrow}(\beta)), n(t^{\uparrow}(\beta)) + 1, 0, 1) \\ & + \sum_{k=n(t_2)+1}^{n(t_1)} (k-1) \widetilde{Q}^2(k, n(t^{\downarrow}(\beta)), n(t^{\uparrow}(\beta)) + 1, 0, 1) \\ & + \sum_{k=n(t_3)+1}^{n(t_2)} (k-2) \widetilde{Q}^2(k, n(t^{\downarrow}(\beta)), n(t^{\uparrow}(\beta)) + 1, 0, 1) \end{aligned}$$

$$
\begin{aligned}
&+ \sum_{k=n(t_4)+1}^{n(t_3)} \Big[\mathbb{1}(t^{\downarrow}(\mathrm{sib}(\beta)) < t^{\downarrow}(\beta))(k-3)\widetilde{Q}^2(k, n(t^{\downarrow}(\beta)), n(t^{\uparrow}(\beta)) + 1, 0, 1) \\
&\quad + \mathbb{1}(t^{\downarrow}(\mathrm{sib}(\beta)) \geq t^{\downarrow}(\beta))(k-1) \Big(\widetilde{Q}^1(k) + \widetilde{Q}^2(k, k, n(t^{\uparrow}(\beta)) + 1, 0, 1) \Big) \Big] \\
&+ \sum_{k=n(t_5)+1}^{n(t_4)} (k-2) \Big(\widetilde{Q}^1(k) + \widetilde{Q}^2(k, k, n(t^{\uparrow}(\beta)) + 1, 0, 1) \Big) \Bigg\}, \qquad (2.11)
\end{aligned}
$$

with $\widetilde{Q}^1$ and $\widetilde{Q}^2$ as defined in (2.8) *and* (2.9)*, respectively.*

The proof is given in the Appendix, Section A4.

Substituting the expressions (2.2), (2.3), (2.7) and (2.11) into (2.1) gives the desired probability that branch b is disrupted by the next recombination event.

2.3 Probability that a branch is topologically disrupted

Most ARG reconstruction algorithms focus on identifying the presence of recombination events through finding patterns of mutations not consistent with tree-like evolution. This, in general, does not allow for the detection of recombination events that only change branch lengths (panels highlighted in blue in Figure 3). We therefore also calculate the probability that a branch b is topologically disrupted (corresponding to panels highlighted in red in Figure 3).

Theorem 2.1. *The probability that a branch b is topologically disrupted by a recombination event is given by*

$$
\begin{aligned}
\mathbb{P}_{\mathcal{T}}(b \textit{ topologically disrupted}) = &\sum_{b' \in \mathcal{B}(b)} \mathbb{P}_{\mathcal{T}}(\mathcal{C} \notin \mathcal{A}(b') | \mathcal{R} \in b') \cdot \mathbb{P}_{\mathcal{T}}(\mathcal{R} \in b') \\
&+ \mathbb{P}_{\mathcal{T}}(\mathcal{C} \in b | \mathcal{R} \notin \mathcal{B}(b)) \cdot \mathbb{P}_{\mathcal{T}}(\mathcal{R} \notin \mathcal{B}(b)), \qquad (2.12)
\end{aligned}
$$

with

$$
\mathbb{P}_{\mathcal{T}}(\mathcal{C} \notin \mathcal{A}(\beta) | \mathcal{R} \in \beta) = 1 - \frac{1}{\bar{t}(\beta)} \sum_{k=n(t^{\uparrow}(\beta))+1}^{n(t^{\downarrow}(\beta))} G_\beta(k),
$$

where, for $k \leq n(t^{\downarrow}(\mathrm{sib}(\beta)))$,

$$
G_\beta(k) \;:=\; 2 \cdot \widetilde{Q}^1(k) + 2 \cdot \widetilde{Q}^2(k, k, n(t^{\uparrow}(\beta)) + 1, 0, 1) + \widetilde{Q}^2(k, n(t^{\uparrow}(\beta)), n(t^{\uparrow}(\mathrm{par}(\beta))) + 1, 0, 1),
$$

and for $k > n(t^{\downarrow}(\mathrm{sib}(\beta)))$

$$
\begin{aligned}
G_\beta(k) := \widetilde{Q}^1(k) + \widetilde{Q}^2(k, k, n(t^{\downarrow}(\mathrm{sib}(\beta))) + 1, 0, 1) + 2 \cdot \widetilde{Q}^2(k, n(t^{\downarrow}(\mathrm{sib}(\beta))), n(t^{\uparrow}(\beta)) + 1, 0, 1) \\
+ \widetilde{Q}^2(k, n(t^{\uparrow}(\beta)), n(t^{\uparrow}(\mathrm{par}(\beta))) + 1, 0, 1),
\end{aligned}
$$

and $\widetilde{Q}^1$ and $\widetilde{Q}^2$ are as defined in (2.8) *and* (2.9)*, respectively.*

The proof is given in the Appendix, Section A5.

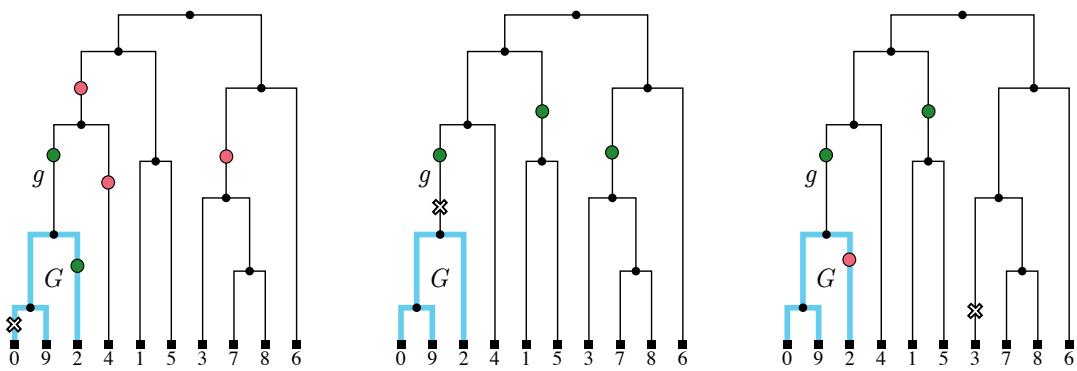

Figure 4: Recombination events that do and do not disrupt a clade. Clade G, subtended by branch g, contains samples $\{0, 2, 9\}$; branches belonging to G are shown in blue. In each tree, for the given recombination point (marked by a cross), red (resp. green) circles show examples of coalescence points that would (resp. would not) result in G being disrupted.

2.4 Probability that a clade is disrupted

We now calculate the probability that a particular *clade* of branches is disrupted by the next recombination event, i.e. that the membership of sample nodes in the clade changes from one local tree to the next (but allowing for events that disrupt branches within the clade without changing the group of subtended samples). This can happen when a lineage within the clade recombines and coalesces outside the clade or its root branch, or if a lineage from outside the clade recombines and coalesces into the clade, as illustrated in Figure 4.

Let G be the set of branches subtended by a branch g, with clade MRCA time $t^{\downarrow}(g) = T_m$, $n_G(t)$ the number of lineages in clade G at time t, abusing notation to write $G \cup g := G \cup \{g\}$. Let $n_{G\cup g}(t)$ be the number of lineages in $G \cup g$ at time t and $L_G(t)$ the total branch length within the clade above time t. Then

$$\begin{aligned} 1 = \mathbb{P}(\mathcal{R} \in G) \cdot (\mathbb{P}(\mathcal{C} \notin G \cup g | \mathcal{R} \in G) + \mathbb{P}(\mathcal{C} \in G \cup g | \mathcal{R} \in G)) + \\ + \mathbb{P}(\mathcal{R} \in g) \\ + \mathbb{P}(\mathcal{R} \notin G \cup g) \cdot (\mathbb{P}(C \notin G | \mathcal{R} \notin G \cup g) + \mathbb{P}(C \in G | \mathcal{R} \notin G \cup g)) \,. \end{aligned}$$

Considering only the events that disrupt the clade, we obtain

$$\mathbb{P}(G \text{ disrupted}) = \mathbb{P}(\mathcal{R} \in G) \cdot \mathbb{P}(\mathcal{C} \notin G \cup g | \mathcal{R} \in G) + \mathbb{P}(\mathcal{R} \notin G \cup g) \cdot \mathbb{P}(C \in G | \mathcal{R} \notin G \cup g).$$

Theorem 2.2. *The probability that a clade G is disrupted by a recombination event is*

$$\begin{aligned} \mathbb{P}_{\mathcal{T}}(G \textit{ disrupted}) = \frac{L_G(0)}{L_{\mathcal{T}}(0)} - \frac{1}{L_{\mathcal{T}}(0)} \sum_{k=n(t^{\downarrow}(g)+1)}^{n} \left[n_{G\cup g}(T_{k+1}) \widetilde{Q}^1(k) + \widetilde{Q}^4(k, G \cup g) \right] n_G(T_{k+1}) \\ + \frac{1}{L_{\mathcal{T}}(0)} \sum_{k=n(t^{\downarrow}(g))+1}^{n} (k - n_{G\cup g}(T_{k+1})) \left[n_G(T_{k+1}) \widetilde{Q}^1(k) + \widetilde{Q}^4(k, G) \right], \end{aligned} \tag{2.13}$$

where

$$\widetilde{Q}^4(k, A) = \frac{1}{k} \left(e^{kT_k} - e^{kT_{k+1}} \right) \sum_{j=n(t^{\uparrow}(A))+1}^{k} n_A(T_{j+1}) Q_{kj},$$

taking $t^{\uparrow}(G \cup g) = t^{\uparrow}(g)$ and $t^{\uparrow}(G) = t^{\downarrow}(g)$.

The proof is given in the Appendix, Section A9.

2.5 Change in total branch length of tree

Conditioning on the tree $\mathcal{T}$, we now consider the distribution of

$$C = L_{\mathcal{T}'}(0) - L_{\mathcal{T}}(0),$$

the change in total branch length following a single recombination event. First, considering the sign of the change, we have the following.

Proposition 2.5. *The probability of C being negative, zero, or positive is given by*

$$\mathbb{P}_{\mathcal{T}}(C<0) = \frac{1}{L_{\mathcal{T}}(0)} \sum_{b\in\mathcal{T}} \sum_{k=n(t^{\uparrow}(b))+1}^{n(t^{\downarrow}(b))} \left((k-1)\widetilde{Q}^1(k) + \widetilde{Q}^2(k,k,n(t^{\uparrow}(b)+1,1,-1)\right) \tag{2.14}$$

$$\mathbb{P}_{\mathcal{T}}(C=0) = \frac{1}{L_{\mathcal{T}}(0)} \sum_{b\in\mathcal{T}} \sum_{k=n(t^{\uparrow}(b))+1}^{n(t^{\downarrow}(b))} \left(\widetilde{Q}^1(k) + \widetilde{Q}^2(k,k,n(t^{\uparrow}(b))+1,0,1)\right) \tag{2.15}$$

$$\mathbb{P}_{\mathcal{T}}(C>0) = \frac{1}{L_{\mathcal{T}}(0)} \sum_{b\in\mathcal{T}} \sum_{k=n(t^{\uparrow}(b))+1}^{n(t^{\downarrow}(b))} \left(\widetilde{Q}^2(k,n(t^{\uparrow}(b)),2,1,0) + \widetilde{Q}^3(k)\right), \tag{2.16}$$

respectively, where

$$\widetilde{Q}^3(k) = \frac{1}{k}\left(e^{-L_{\mathcal{T}}(T_k)} - e^{-L_{\mathcal{T}}(T_{k+1})}\right). \tag{2.17}$$

The proof is given in the Appendix, Section A6.

To explore the distribution of the magnitude of the change in branch length (when this is non-zero), we derive its density.

Proposition 2.6. *Conditional on the change in total branch length being non-zero, the density of C is given by*

$$p^C_{\mathcal{T}}(c|C\neq 0) = \frac{1}{L_{\mathcal{T}}(0)} \sum_{\substack{b\in\mathcal{T}:\\ c\geq -\bar{t}(b)}} \xi_b(c), \tag{2.18}$$

where

$$\xi_b(c) = \begin{cases} (n(c+t^{\uparrow}(b))-1)\left(e^{-(l-1)(c+t^{\uparrow}(b))} \cdot \sum_{k=l}^{n(t^{\downarrow}(b))} \widetilde{P}^1_{kl} - \frac{e^{(l-1)(T_l-t^{\uparrow}(b)-c)}-1}{l-1}\right) & -\bar{t}(b)\leq c<0 \\ n(c+t^{\uparrow}(b))e^{-\left[L_{\mathcal{T}}(t^{\uparrow}(b))-L_{\mathcal{T}}(c+t^{\uparrow}(b))\right]} \cdot \sum_{k=n(t^{\uparrow}(b))+1}^{n(t^{\downarrow}(b))} \widetilde{P}^2_k & 0<c\leq T_2-t^{\uparrow}(b) \\ \frac{1}{2}e^{-(c+t^{\uparrow}(b)-T_2)/2} \cdot \sum_{k=n(t^{\uparrow}(b))+1}^{n(t^{\downarrow}(b))} \widetilde{P}^3_k & c>T_2-t^{\uparrow}(b) \\ 0 & \textit{otherwise}, \end{cases}$$

and

$$\widetilde{P}^1_{kl} := \frac{1}{k-1}\left(e^{(k-1)T_k} - e^{(k-1)T_{k+1}}\right)\exp\left(l\cdot T_l - k\cdot T_k + L_{\mathcal{T}}(T_l) - L_{\mathcal{T}}(T_k)\right)$$

$$\widetilde{P}^2_k := \frac{1}{k-1}\left(e^{(k-1)T_k} - e^{(k-1)T_{k+1}}\right)\exp\left(t^{\uparrow}(b) - k\cdot T_k - L_{\mathcal{T}}(T_k) + L_{\mathcal{T}}(t^{\uparrow}(b))\right)$$

$$\widetilde{P}^3_k := \frac{1}{k-1}\left(e^{(k-1)T_k} - e^{(k-1)T_{k+1}}\right)\exp\left(t^{\uparrow}(b) - k\cdot T_k - L_{\mathcal{T}}(T_k)\right).$$

The proof is given in the Appendix, Section A7.

2.6 Change in tree height

The height of the tree, $H(\mathcal{T}) = T_2$, can change following a recombination event if (1) the coalescence point is above T_2, in which case a new root is formed and the tree height increases, or (2) the recombination happens on one of the two lineages descending from the MRCA, then the tree height can either increase or decrease. Let the set $\mathcal{M} := \{\mathrm{ch}_1(b_r), \mathrm{ch}_2(b_r)\}$ contain the two branches descending from the MRCA, and let $H = H(\mathcal{T}') - H(\mathcal{T})$ be the magnitude of the change in height. Then we have the following.

Proposition 2.7. *Conditional on $\mathcal{T}$, the probability of the change in tree height being negative, zero, or positive is given by*

$$\mathbb{P}_{\mathcal{T}}(H < 0) = \frac{1}{L_{\mathcal{T}}(0)} \sum_{b \in \mathcal{M}} \sum_{k=2}^{n(t^{\downarrow}(b))} \left\{ (k-1)\widetilde{Q}^1(k) - \widetilde{Q}^2(k,k,2,0,1) - \widetilde{Q}^3(k) \right\} \tag{2.19}$$

$$\mathbb{P}_{\mathcal{T}}(H = 0) = \frac{1}{L_{\mathcal{T}}(0)} \left\{ \sum_{b \in \mathcal{M}} \sum_{k=2}^{n(t^{\downarrow}(b))} \left(\widetilde{Q}^1(k) + \widetilde{Q}^2(k,k,2,0,1) \right) + \sum_{b \notin \mathcal{M}} \left(\bar{t}(b) - \sum_{k=n(t^{\uparrow}(b))+1}^{n(t^{\downarrow}(b))} \widetilde{Q}^3(k) \right) \right\} \tag{2.20}$$

$$\mathbb{P}_{\mathcal{T}}(H > 0) = \frac{1}{L_{\mathcal{T}}(0)} \sum_{b \in \mathcal{T}} \sum_{k=n(t^{\uparrow}(b))+1}^{n(t^{\downarrow}(b))} \widetilde{Q}^3(k), \tag{2.21}$$

where $\widetilde{Q}^1$, $\widetilde{Q}^2$ and $\widetilde{Q}^3$ are as defined in (2.8), (2.9) *and* (2.17), *respectively.*

The proof is given in the Appendix, Section A8.

2.7 Distribution of branch duration along the genome

For a given branch b, the distribution of its duration can be characterised by considering the rate at which branch-disrupting recombination events arrive as we move left-to-right along the genome. However, the instantaneous rate at which b is disrupted at position τ may not be the same as that at position $\tau' > \tau$, due to the effect of other recombination events that might occur between τ and τ'. Thus, the duration of a branch is the waiting time to the next branch-disrupting recombination event, but the rate at which this happens is inhomogeneous along the genome (and is, in fact, itself random).

Similarly to Deng et al. (2021), however, we find that if a recombination event between adjacent trees $\mathcal{T}$ and $\mathcal{T}'$ does not disrupt b, then

$$P_{\mathcal{T}}(b \text{ disrupted}) \cdot \mathcal{L}_{\mathcal{T}}(0) \approx P_{\mathcal{T}'}(b \text{ disrupted}) \cdot \mathcal{L}_{\mathcal{T}'}(0),$$

where $P_{\mathcal{T}}(b \text{ disrupted})$ is given by (2.1), based on simulation results as shown in the Appendix, Figure A2. Thus, an approximation to the distribution of branch duration can be constructed by assuming that the rate at which branch-disrupting recombination events arrive *is* homogeneous along the genome, which is equivalent to assuming that recombination events that do not disrupt the branch b also do not change the local tree: so if $\mathcal{T}$ is the local tree at position $d^{\leftarrow}(b)$, after each recombination event the newly formed local tree is $\mathcal{T}' = \mathcal{T}$. Recombination events occur as a Poisson process along the genome with rate $\mathcal{L}_{\mathcal{T}}(0) \cdot \rho/2$, allowing us to thin the process by multiplying this rate by the probability that the event is branch-disrupting, thereby offering a tractable approximation for the arrival of branch-disrupting recombination events. Then conditional on $d^{\leftarrow}(b)$, the branch duration $d^{\rightarrow}(b) - d^{\leftarrow}(b)$ is distributed as the waiting time to the first event in a Poisson process with rate

$$\mathbb{P}_{\mathcal{T}}(b \text{ disrupted}) \cdot L_{\mathcal{T}}(0) \cdot \frac{\rho}{2}.$$

That is,

$$d^{\rightarrow}(b) - d^{\leftarrow}(b) \mid d^{\leftarrow}(b) \sim \text{Exp}\left(\mathbb{P}_{\mathcal{T}}(b \text{ disrupted}) \cdot L_{\mathcal{T}}(0) \cdot \frac{\rho}{2}\right),$$

or, by rescaling,

$$\mathbb{P}_{\mathcal{T}}(b \text{ disrupted}) \cdot L_{\mathcal{T}}(0) \cdot \frac{\rho}{2}(d^{\rightarrow}(b) - d^{\leftarrow}(b)) \mid d^{\leftarrow}(b) \sim \text{Exp}(1).$$

Analogously, if the recombination rate is not constant along the genome, with the population-scaled recombination rate at position w given by $\rho(w)/2$, then the intensity of the process at position w is instead given by

$$\mathbb{P}_{\mathcal{T}}(b \text{ disrupted}) \cdot L_{\mathcal{T}}(0) \cdot \frac{\rho(w)}{2},$$

and we have

$$\mathbb{P}_{\mathcal{T}}(b \text{ disrupted}) \cdot L_{\mathcal{T}}(0) \cdot \int_{d^{\leftarrow}(b)}^{d^{\rightarrow}(b)} \frac{\rho(w)}{2} dw \mid d^{\leftarrow}(b) \sim \text{Exp}\,(1)\,. \tag{2.22}$$

The quality of this approximation can be verified by simulation, using the probability integral transform as follows. For the i-th branch $b_i \in \{b_1, \ldots, b_m\}$ of a simulated ARG, take $\mathcal{T}$ to be the local tree at position $d^{\leftarrow}(b_i)$, compute q_i as

$$q_i := \mathbb{P}_{\mathcal{T}}(b_i \text{ disrupted}) \cdot L_{\mathcal{T}}(0) \cdot \int_{d^{\leftarrow}(b_i)}^{d^{\rightarrow}(b_i)} \frac{\rho(w)}{2} dw,$$

and let $p_i = 1 - e^{-q_i}$. Then a Q-Q plot can be constructed by plotting the ordered quantities $p_{(1)} \leq \ldots \leq p_{(m)}$ against the corresponding quantiles of the uniform distribution $\frac{1}{1+m}, \ldots, \frac{m}{1+m}$. If the approximation fits well, the points should lie on the diagonal. A Kolmogorov–Smirnov (K–S) goodness of fit test can be used to test the null hypothesis that the computed p_i values are uniformly distributed.

If we were to consider only events that topologically disrupt the branch, we instead have the approximation

$$\mathbb{P}_{\mathcal{T}}(b \text{ topologically disrupted}) \cdot L_{\mathcal{T}}(0) \cdot \int_{d^{\leftarrow}(b)}^{d^{\rightarrow}(b)} \frac{\rho(w)}{2} dw \mid d^{\leftarrow}(b) \sim \text{Exp}\,(1)\,. \tag{2.23}$$

A similar procedure to that described above can be used to check goodness of fit.

2.8 Distribution of clade duration along the genome

Similar approximations to those employed when investigating the distribution of branch duration along the genome can be used for the distribution of the waiting time until a clade G is broken up by a recombination event (that is, when the clade either gains or loses one or more samples as a consequence of recombination). Consider an inhomogeneous Poisson process with intensity at position w given by

$$\mathbb{P}_{\mathcal{T}_{d^{\leftarrow}(G)}}(G \text{ disrupted}) \cdot L_{\mathcal{T}}(0) \cdot \frac{\rho(w)}{2},$$

conditional on the local tree at $d^{\leftarrow}(G)$ (defined as the leftmost position along the genome where the clade arises). Again through rescaling time, we have

$$\mathbb{P}_{\mathcal{T}_{d^{\leftarrow}(G)}}(G \text{ disrupted}) \cdot L_{\mathcal{T}_{d^{\leftarrow}(G)}}(0) \cdot \int_{d^{\leftarrow}(G)}^{d^{\rightarrow}(G)} \frac{\rho(w)}{2} dw \mid d^{\leftarrow}(G) \sim \text{Exp}\,(1)\,. \tag{2.24}$$

Thus, for each clade G_i in a simulated ARG, we can calculate its left and right endpoints $d^{\leftarrow}(G_i)$ and $d^{\rightarrow}(G_i)$, respectively. Letting

$$q_i := \mathbb{P}_{\mathcal{T}_{d^{\leftarrow}(G_i)}}(G_i \text{ disrupted}) \cdot L_{\mathcal{T}_{d^{\leftarrow}(G_i)}}(0) \cdot \int_{d^{\leftarrow}(G_i)}^{d^{\rightarrow}(G_i)} \frac{\rho(w)}{2} dw, \tag{2.25}$$

the quality of the approximation can again be checked using a Q-Q plot as described above.

3 Results

3.1 Probability of a branch being disrupted by a recombination event

The left panel of Figure 5 shows (2.1) calculated for each branch of an ARG simulated under the SMC' with stdpopsim (Adrion et al., 2020; Lauterbur et al., 2022), against the normalised age of the branch $t^{\uparrow}(b)/L_{\mathcal{T}}(0)$. Middle and right panels show the same quantities, but calculated for each branch of an ARG reconstructed from the simulated data: using Relate (v1.1.9) (middle), and tsinfer (v0.3.0) and tsdate (v0.1.5) (right).

For the simulated ARG, the calculated probability that the branch is disrupted by a recombination event is higher for older branches: this is as expected, as branches at the top of a local tree tend to have greater time-length and co-exist with fewer other lineages. However, old branches are generally difficult to accurately reconstruct from the sequences at the leaves due to lack of signal, unless they are strongly supported by mutations (which implies longer duration along the genome). As a result, reconstructed ARGs have fewer old branches (with fewer points lying towards the right of the plots), and the old branches that *are* present tend to have lower probability of disruption than those of a similar age in the simulated ARG. This bias is seen for both Relate and tsinfer/tsdate.

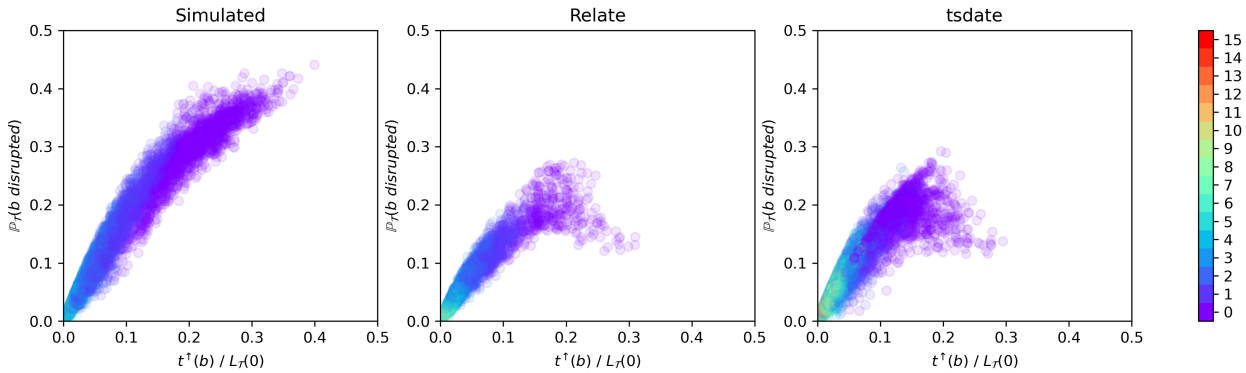

Figure 5: Value of (2.1) calculated for each branch b, against $t^{\uparrow}(b)/L_{\mathcal{T}}(0)$. Left panel: for a uniform random sample of 10 000 branches in one simulated ARG ($n = 100$, stdpopsim human Chr21, HapMapII GRCh38 recombination map, constant population size model with $N_e = 10\,000$ diploids). Middle and right: same quantity calculated for each branch of the ARG reconstructed from the simulated data using Relate (middle) and tsinfer/tsdate (right). Colour shows number of edges between the top end of the branch and the root of $\mathcal{T}$ (purple dots correspond to branches extending from the MRCA node).

3.2 Quality of approximation to the distribution of branch duration

We first assess the quality of the approximation derived in Section 2.7 by simulating an ARG under the SMC' and checking if the simulated branch durations follow (2.22), by using the procedure described in Section 2.7. Using stdpopsim, we simulated two ARGs under the SMC' with $n = 100$ (haploid samples) and the following sets of parameters:

- dataset 1: Chr21 with HapMapII_GRCh38 recombination map, mutation rate $1.29 \cdot 10^{-8}$ per site per generation, $N_e = 10\,000$ diploids, constant population size model;
- dataset 2: 5Mb of Chr21 with flat recombination map, recombination rate $1.2 \cdot 10^{-8}$ per site per generation, mutation rate $1.29 \cdot 10^{-8}$ per site per generation, $N_e = 10\,000$ diploids, constant population size model.

We sampled 10 000 branches from each ARG (uniformly at random) for testing, to speed up computation. The corresponding Q-Q plots are shown in Figure 6 (blue points). The points adhere very

closely to the diagonal, demonstrating that the approximation provides an excellent fit. The K–S p-values of 0.31 (left panel) and 0.75 (right panel) also suggest good agreement. Grouping branches by their depth (the number of branches on the way to the MRCA) or clade size (number of samples subtended by the branch) in the tree and constructing Q-Q plots for each group also did not reveal significant deviation from the diagonal (as seen in the Appendix, Figure A4), suggesting that the approximation holds for all branches.

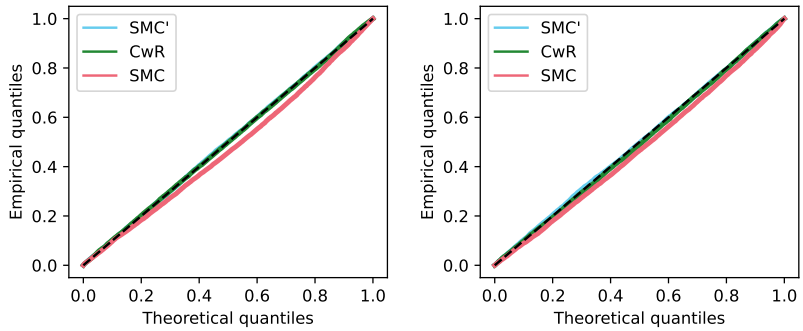

Figure 6: Q-Q plots using (2.22) computed from ARGs simulated using the SMC' (blue), CwR (green), and SMC (red) models with $n = 100$. Note the blue and green points closely overlap, and overlay the diagonal. Left panel: dataset 1 parameters; right panel: dataset 2 parameters. Dashed line: diagonal from $(0, 0)$ to $(1, 1)$.

3.3 Effects of recombination on a local tree

To calculate a Monte Carlo estimate of the marginal probability that the change in total branch length is negative, zero, or positive, we average over local trees simulated using msprime under the SMC' model. The results are shown in Figure 7. Recombination has a stabilising effect on the total branch length of local trees: when the total branch length is small (resp. large), the probability that the recombination event will increase the total branch length increases (resp. decreases).

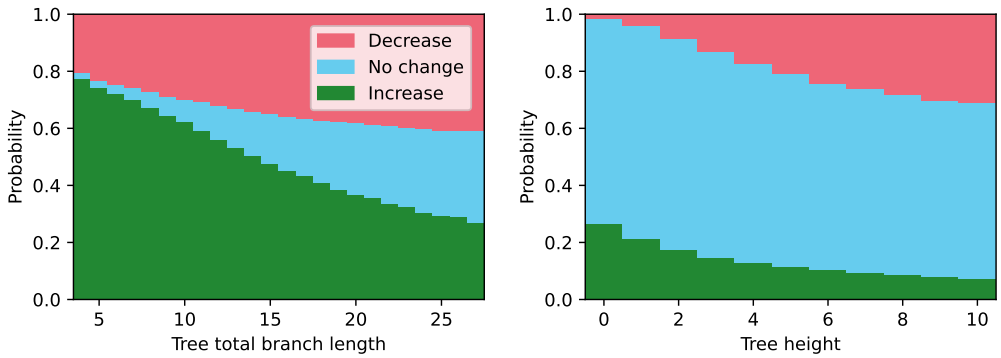

Figure 7: Mean probability of change in total branch length (left) and tree height (right) being negative (red, top stack), zero (blue, middle stack) or positive (green, bottom stack). Trees were simulated and binned according to total branch length (left) or height (right), with 100 trees simulated per bin, and sample size $n = 100$. For each tree, probabilities were calculated using (2.14)-(2.16), the stacked bar plot shows the mean probabilities for each bin.

Similarly, we can average over trees simulated under the SMC' model to estimate the marginal probability that the change in tree height following a recombination event is negative, zero, or positive. The results are shown in Figure 7 (right panel). As with total branch length, recombination tends to increase (resp. decrease) the tree height with higher probability when the tree height is small (resp. large).

Figure 8 shows the density (2.18) for three simulated trees with varying total branch lengths $L_{\mathcal{T}}(0)$, for $n = 10$ and $n = 100$. In both cases, the density is concentrated around zero and skewed to the left (resp. right) when the total branch length is small (resp. large); it is roughly symmetric about zero for middling values of $L_{\mathcal{T}}(0)$.

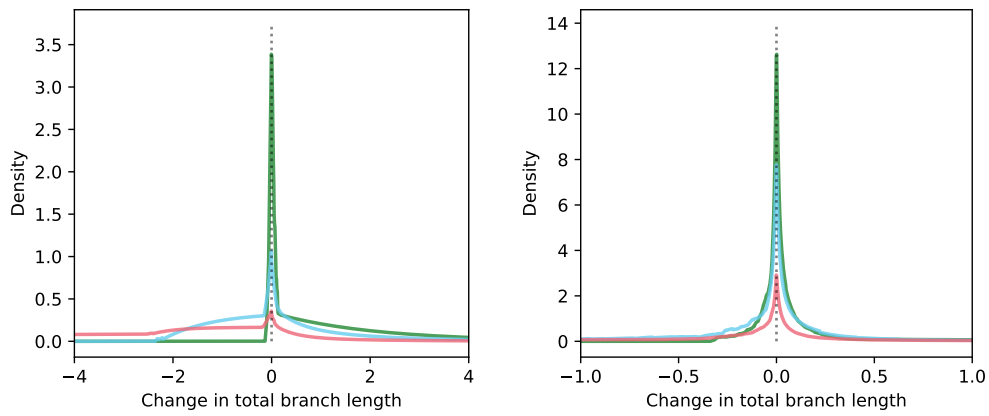

Figure 8: Density of change in total branch length (2.18) for three simulated trees, with $n = 10$ (left panel) $n = 100$ (right panel). Left: trees have total branch length 1 (green), 6 (blue) and 24 (red). Right: trees have total branch length 5 (green), 10 (blue) and 28 (red).

These results shed light on why, despite the strong assumption that recombination events that do not disrupt the given branch also do not change the rest of the local tree, our approximation to the distribution of branch duration gives an extraordinarily close fit for data simulated under the SMC' model. For a branch that is close to the leaves, the probability that the branch is disrupted by the next recombination event is very small (Figure 5). Thus, many recombination events will occur before this branch is disrupted. As can be seen in Figure 7, recombination has the effect of stabilising the total branch length, with events causing an increase (decrease) in total branch length being more likely if the current total branch length is relatively small (large). Thus, it seems the fluctuations in total branch length average out and do not significantly affect the overall rate of branch-disrupting recombination events. On the other hand, for a branch that is close to the root of the tree, per Figure 5 the probability of the branch being disrupted by the next recombination is relatively high. Thus, a relatively small number of recombination events are likely to occur before they affect the given branch. As illustrated in Figure 8, when a recombination changes the total branch length of the tree, the magnitude of this change is concentrated around 0. Thus, the effect of recombination on the rest of the tree does not appear to significantly affect the probability that the branch is disrupted.

3.4 Comparison of simulation models

We next simulate ARGs under the CwR and under the SMC, with the same two parameter sets as given in Section 3.2 and again compare the resulting branch durations to (2.22). For the CwR, the duration of a branch is taken to be the sum of all the genomic intervals where that branch appears in the local tree (to account for the presence of recombination events that occur in non-ancestral material). Figure 6 shows that the approximation is an excellent fit to the CwR (green points), with K–S p-values of 0.06 (left panel) and 0.94 (right panel). This suggests that the distribution of branch durations under the SMC' and that under the CwR are remarkably close.

Under the SMC (red points), branch durations are shorter in general, with points falling below the diagonal (with both K–S p-values < 0.001). This is due to the model disallowing recombination events that do not change the local tree, so branches are more frequently disrupted by recombination.

3.5 Reconstructed ARGs

We next compare the distribution of branch duration under the SMC' to that of ARGs reconstructed using Relate, tsinfer/tsdate, ARG-Needle, and ARGweaver. For each dataset simulated under the SMC' as described in Section 3.2, we used each tool to reconstruct an ARG, using the true simulation parameters as inputs for each tool (and for ARGweaver, a time discretisation grid with 100 points and selecting the MAP ARG from 1000 posterior samples, for ARG-Needle using 50 time discretisation points). We sense-checked the output of the ARG reconstruction methods using a number of metrics (comparing the MRCA times, local tree topologies, and LD decay against the simulated ARGs).

The corresponding Q-Q plots are shown in Figure 9 (calculated for a uniform random sample of 10 000 branches for each ARG; the corresponding histograms are also shown in the Appendix, Figure A3). For the simulated ARGs and those reconstructed by ARGweaver and ARG-Needle, we check branch duration against the approximation (2.22); for ARGs reconstructed using Relate and tsinfer/tsdate, we check against (2.23) which only takes topology-changing events into account (to adjust for the fact that these tools do not detect recombination events that affect only branch lengths). Moreover, Relate does not attempt to infer branch duration when the branch is not supported by at least one mutation, in which case the local topology is resampled from one local tree to the next (while tsinfer more directly captures the duration of branches in the reconstructed ARG based on shared ancestry). Therefore, we only test branches in Relate ARGs which carry at least one mutation. This requires conditioning the approximation (2.23) on the number of mutations being greater than zero; the corresponding calculations are provided in the Appendix, Section A10.

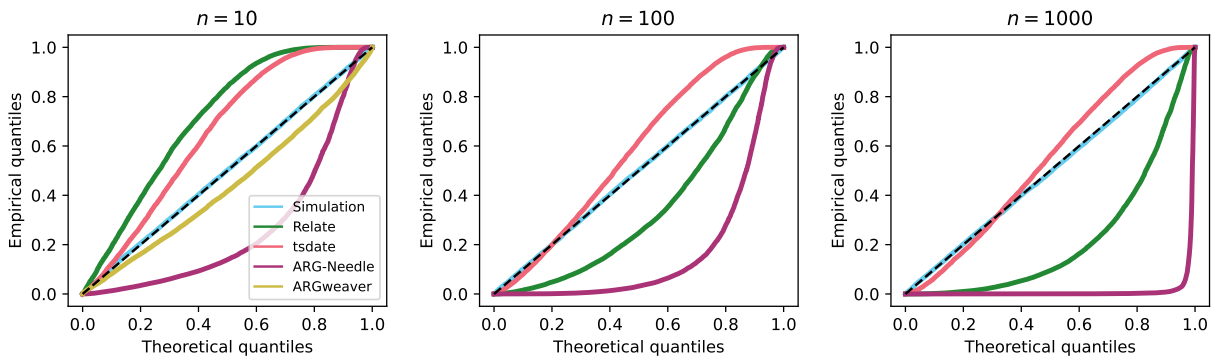

Figure 9: Q-Q plots for ARGs simulated with same parameters as dataset 1 (dataset 2 for ARGweaver) in Section 3.2, with $n = 10$ (left panel), $n = 100$ (middle panel), $n = 1\,000$ (right panel). Dashed line: diagonal from $(0, 0)$ to $(1, 1)$. ARGweaver results only shown for $n = 10$ due to excessive runtimes for larger sample sizes.

ARGweaver accurately captures the branch duration distribution under the SMC model, showing a similar deviation from the diagonal as for data simulated under the SMC in Figure 6, although results were only calculated for $n = 10$ due to excessive runtimes for larger sample sizes. Note that while it is possible to use the SMC' model with the latest version of ARGweaver, we found that the resulting ARGs contained cycles, which prevented us from calculating the required probabilities.

ARGs reconstructed by tsinfer/tsdate consistently have an excess of long-lasting branches, while for Relate this depends on the sample size; both tools however produce skewed distributions of branch duration. This is likely to be, in part, due to the waiting distances between trees being generally skewed, as shown by Deng et al. (2021). The ARGs produced by tsinfer contain polytomies (nodes with more than two descendants), which is also not accounted for under the SMC' model. This results in the total branch length of a reconstructed local tree to often be greater than it would be if the polytomies were resolved to make the tree binary. Moreover, some recombination events that would disrupt the branch under the SMC' (such as recombination points located on

child branches), may not do so when the tree contains polytomies.

Relate first reconstructs the sequence of (correlated) local trees along the entire genome and then identifies which branches persist between trees, using an argument based on the similarity of clades subtended by the branch in successive trees. Thus, branch durations are calculated approximately, rather than being explicitly inferred, which we do not account for.

The threading procedure used by ARG-Needle to reconstruct the ARG uses the ASMC model to estimate coalescence times between the added sequence and the closely-related samples already in the ARG, with the lowest possible resulting coalescence time dictating which branch the sequence is threaded to and over what genomic length. This procedure (which relies on a combination of maximum a posteriori and posterior mean estimates for the coalescence times) is optimised for metrics having a direct effect on downstream analyses, and appears to lead to ARG-Needle consistently underestimating branch durations. We note that this is somewhat affected by the choice of time discretisation used within the ASMC (particularly for small sample sizes), but our overall findings do not change significantly when toggling this parameter.

Considering the un-corrected distributions of the branch durations in the simulated and reconstructed ARGs (i.e. directly calculating the observed duration of each branch, without adjustment, and plotting the histogram) results in similar conclusions (shown in the Appendix, Figure A5). Further, Figure A6 (Appendix) shows histograms of the expected number of mutations per branch, being the product of the observed branch duration (in bp), observed time-length (in generations), and the mutation rate (per bp per generation), demonstrating large deviations between the distributions for simulated and reconstructed ARGs.

For Relate and tsinfer/tsdate, we find that using approximation (2.23) results in a marginally better fit than (2.22). This is as expected, since these algorithms are not designed to detect recombination events that only affect branch times. Additionally, we find that goodness of fit depends on the mutation rate: as the mutation rate increases, more recombination events are detected by the algorithms and thus branch time estimates become more accurate.

3.6 ARG-based test for inversions

Chromosomal inversions are a type of structural variant, whereby as a result of recombination, the genome breaks at two points and the segment between the breakpoints is reinserted in the opposite orientation. While recombination can proceed normally in individuals homozygous for the inversion, in heterozygotes recombination is substantially suppressed in the region containing the inversion, since a crossover recombination within the region is likely to result in the production of unbalanced gametes (Kirkpatrick, 2010). Inversions are a relatively common phenomenon in many organisms, including humans (Puig et al., 2015) and mice (Hammer et al., 1989), and can range in size from hundreds of kilobases to several megabases (Hanlon et al., 2022). They have been posited to play an important role in speciation and local adaptation (Kirkpatrick, 2010), and a large body of literature exists examining the interplay of inversions and selection from the population genetics perspective (e.g. Kirkpatrick and Barton, 2006).

Detecting inversions computationally from sequencing data typically relies on paired-end mapping (detecting reads mapping in the opposite orientation to the reference), split-read methods (detecting reads that map onto the reference with gaps), and *de novo* assembly (directly reconstructing the sequenced genome to look for differences with the reference) (Tattini et al., 2015). For instance, DELLY (Rausch et al., 2012) implements a combination of these approaches and was used to identify hundreds of inversions in data generated by the 1000 Genomes Project (Sudmant et al., 2015). In general, however, such methods suffer from high false positive rates and poor sensitivity, with their performance depending on the size of the inverted region, particularly for short-read sequencing data (Lucas Lledó and Cáceres, 2013); the detection of structural variants from long-read data is challenging due to high sequencing error rates (Sedlazeck et al., 2018).

Suppose that in a given ARG, an inversion happens on a branch g which subtends a clade

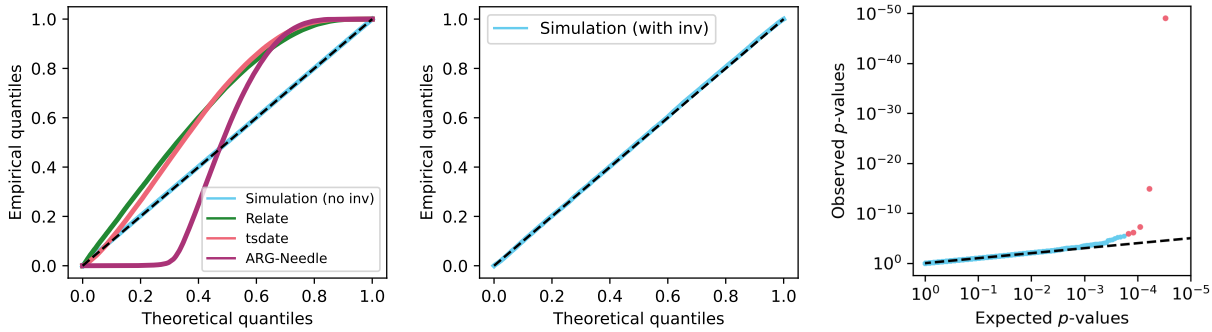

Figure 10: Left: Q-Q plot for clade duration in simulated and reconstructed ARGs with $n = 100$ and parameters as for dataset 1 given in Section 3.2. Middle and right: Q-Q plot and p-value plot for ARG simulated using SLiM with one inversion under balancing selection (red points correspond to clades with p-value below the Bonferroni-corrected significance threshold).

G. Suppression of recombination in heterozygotes implies that if a lineage within G undergoes a recombination, it will coalesce with lineages in G with high probability; likewise, if a recombination event happens on a branch not carrying the inversion, with high probability it will coalesce with branches outside G. This implies that inversions can be detected by looking for clades that last for "too long" along the genome. Phrasing this as a hypothesis test, for each clade G_i, we calculate its left and right endpoints $d^{\leftarrow}(G_i)$ and $d^{\rightarrow}(G_i)$, and estimate the probability of observing a clade duration greater than $d^{\rightarrow}(G_i) - d^{\leftarrow}(G_i)$ via $p_i = e^{-q_i}$, with q_i as defined in (2.25). Simulation studies confirm that for ARGs simulated under the SMC' model without inversions, these p-values are approximately uniformly distributed (Figure 10, left panel, blue points), as expected.

Note that we are imposing the simplifying assumption that recombination in heterozygotes is suppressed completely in the inverted region (so the clade G remains completely intact). In reality, recombination can occur in heterozygotes: multiple breakpoints occurring in the inverted region would allow for this, however such events have relatively low probability unless the inversion region is very large or the recombination rate is very high.

3.6.1 Simulated ARGs

To demonstrate the power of this test in detecting inversions, we used SLiM (Haller and Messer, 2019; Haller et al., 2019) to simulate an ARG with one inversion under balancing selection ($n = 100$, 5Mb with recombination rate $1{\cdot}10^{-8}$ per bp per generation, constant population size 10 000, inverted segment length 200kb, neutral mutations at rate $1{\cdot}10^{-8}$ per bp per generation added using msprime).

Figure 10 shows the corresponding Q-Q plot (middle panel) and p-values for each clade (right panel). The durations of most clades are unaffected by the inversion, so most of the points on the Q-Q plot adhere tightly to the diagonal. However, there are outliers in the tail (right panel, shown in red) with p-value $< 0.05/m$, where $m = 32\,564$ is the number of clades. Figure 11 shows in red the genomic positions of the clades with significant p-values; all of these overlap the location of the inverted region, and the clade subtended by the edge carrying the inversion is a significant outlier with the lowest p-value. The equivalent Q-Q and p-value plots for a neutral simulation, using SLiM with no inversion and otherwise the same parameters, are shown in the Appendix, Figure A7, demonstrating that in this case the p-values are approximately uniformly distributed and there are no clades with significant p-values.

3.6.2 Reconstructed ARGs

As can be seen from Figure 10 (left panel), similarly to branch duration, the distribution of clade duration in ARGs reconstructed using Relate, tsinfer/tsdate and ARG-Needle is skewed, with clade

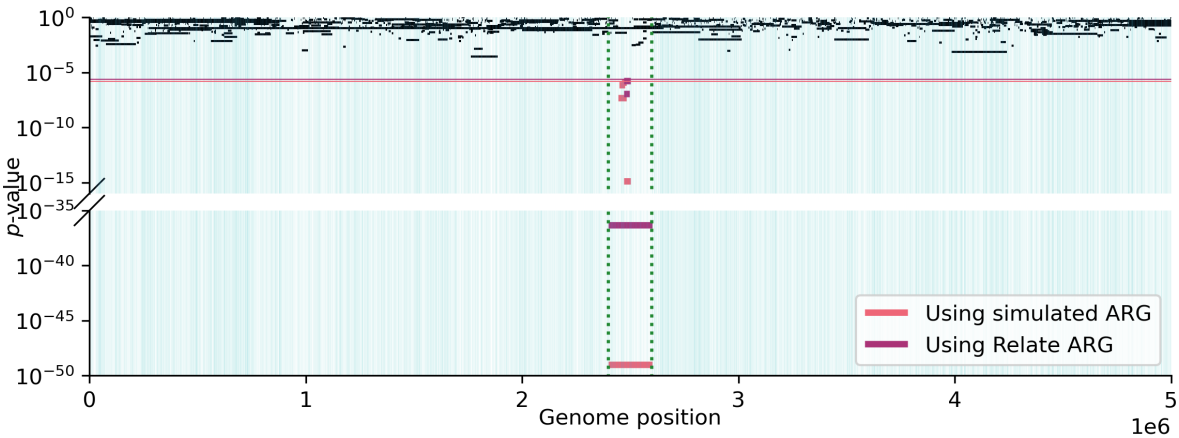

Figure 11: The spans of all clades classed as outliers based on (2.24) for the simulated ARG (resp. ARG reconstructed using Relate) are plotted as red (resp. purple) thick horizontal lines; a random sample of 2 000 other clades also shown in black, for reference. Corresponding p-values are shown on the y-axis. Blue vertical lines show breakpoints between local trees; green dotted lines delineate the region of the inversion; thin red (resp. purple) horizontal line shows Bonferroni-corrected significance threshold for calculations using the simulated (resp. Relate) ARG.

duration generally overestimated by these methods. Thus, directly applying the test as described above will lead to a high false positive rate. We now describe a correction which can be applied to counteract two main problematic features of reconstructed ARGs, focusing particularly on Relate (due to the presence of polytomies for tsinfer/tsdate being difficult to correct for, and the large bias seen with ARG-Needle which both under- and overestimates clade duration).

Let $\mathcal{G} = \{G_1, G_2, \ldots, G_N\}$ be a list of all clades in the reconstructed ARG. The first issue is that due to a lack of mutations around the leftmost and rightmost endpoints of a clade, Relate may overestimate its duration, causing false positives. To correct for this, we proceed as follows. Suppose that the root branches of a clade G_i in trees $\mathcal{T}_{d^{\leftarrow}(G_i)}, \ldots, \mathcal{T}_{d^{\rightarrow}(G_i)}$ have mutations at positions $m_1^i < m_2^i < \ldots < m_K^i$. We (1) remove from $\mathcal{G}$ all clades with fewer than two mutations in total and fewer than M mutations per kb on average, and (2) measure an adjusted clade duration $\widetilde{d}$ using the positions of the leftmost and rightmost mutations that support the given clade. That is, we define

$$\widetilde{d}^{\leftarrow}(G_i) = m_1^i,\ \widetilde{d}^{\rightarrow}(G_i) = m_K^i.$$

The second issue is that the clade carrying the inversion may not be supported by mutations uniformly along the inverted region, causing it to appear and disappear multiple times in quick succession in the reconstructed ARG, which can cause false negatives. We correct for this by "merging" pairs of clades that are nearby on the genome (in terms of genetic distance, to allow for varying recombination rates). For two clades $G_i, G_j \in \mathcal{G}$ that have identical sets of sample descendants and are less than L centiMorgans apart, that is

$$100 \cdot \int_{d^{\rightarrow}(G_i)}^{d^{\leftarrow}(G_j)} \rho(w)dw < L,$$

form $G_{i,j} := G_i$ but setting $d^{\leftarrow}(G_{i,j}) := d^{\leftarrow}(G_i)$, $d^{\rightarrow}(G_{i,j}) := d^{\rightarrow}(G_j)$, $\widetilde{d}^{\leftarrow}(G_{i,j}) := \widetilde{d}^{\leftarrow}(G_i)$, $\widetilde{d}^{\rightarrow}(G_{i,j}) := \widetilde{d}^{\rightarrow}(G_j)$, and update $\mathcal{G}$ as

$$\mathcal{G} = \{G_{i,j}\} \cup \mathcal{G} \backslash \{G_i, G_j\}.$$

We apply this to all pairs of clades in $\mathcal{G}$ iteratively, until no more clades can be merged together.

For each clade in the reduced list $G_i \in \mathcal{G}$, we thus calculate an adjusted version of (2.25):

$$\widetilde{q}_i := \mathbb{P}_{\mathcal{T}_{\widetilde{d}^{\leftarrow}(G_i)}}(G_i \text{ disrupted}) \cdot L_{\mathcal{T}_{\widetilde{d}^{\leftarrow}(G_i)}}(0) \cdot \int_{\widetilde{d}^{\leftarrow}(G_i)}^{\widetilde{d}^{\rightarrow}(G_i)} \frac{\rho(w)}{2} dw, \tag{3.1}$$

again taking $\widetilde{p}_i = e^{-\widetilde{q}_i}$. Note that with the above definitions, this can be computed even though the clades in the reduced list will now not necessarily exist (with the re-defined durations) in the ARG itself. We thus obtain adjusted p-values for each clade, applying a significance threshold of $0.05/N$, where N is the original number of clades in the reconstructed ARG.

We apply this to an ARG reconstructed using Relate for the data simulated using SLiM (as described in Section 3.6.1), setting $L = 0.01$ and $M = 0.05$. We detect three significant clades, shown in purple in Figure 11; all of these overlap the inverted region. The clade spanning the entire inverted region remains a significant outlier with the lowest p-value. The corresponding Q-Q and p-value plots are shown in the top row of Figure A8 (Appendix), showing that the correction brings the points on the Q-Q plot very close to the diagonal. The equivalent plots for a simulation with no inversion are shown in the bottom row, demonstrating that the p-values are approximately uniformly distributed and there are no false positives.

3.6.3 Test error rates

We examined the performance of the test by applying it to ARGs simulated using SLiM with the parameters given in Section 3.6.1, varying the length of the inverted region from 0 to 200kb (100 ARGs in each case). Defining positive detection as there being at least one clade within the inverted region with a significant p-value, the resulting ROC curve is shown in Figure 12 (left panel). Fixing the false positive rate at 5% (corresponding roughly to one false positive per 100Mb) gives the confusion matrix shown in Table 1a, demonstrating very high sensitivity. For each inversion length, out of all the clades with significant p-values across the simulations, a high percentage lie within the inverted region.

Reconstructing an ARG using Relate for each simulated dataset and applying the adjustments described in Section 3.6.2 gives the ROC curve shown in Figure 12 (right panel); the results in Table 1b show high sensitivity is maintained for inversions longer than around 100kb. These results demonstrate very good performance in detecting the presence of inversions, as well as pinpointing the candidate clade and its position along the genome.

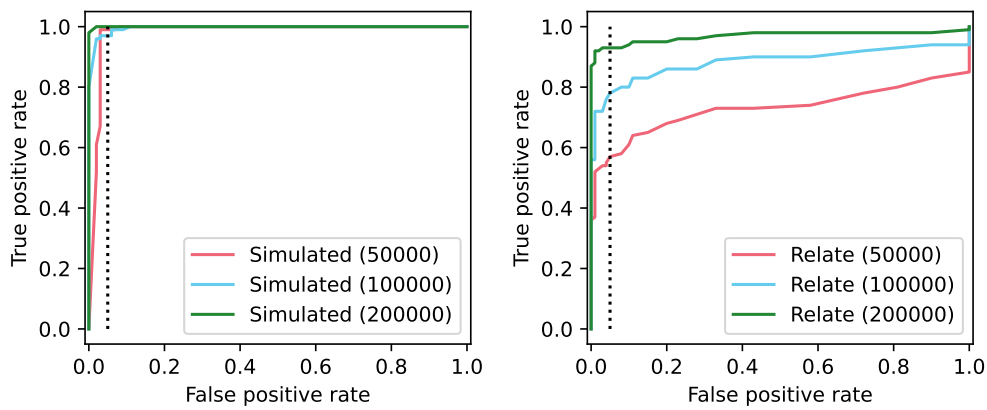

Figure 12: ROC curve for inversion detection test, based on 100 simulations for each given length of the inverted region. Left: using the simulated ARGs; right: ARGs reconstructed using Relate. Dotted line corresponds to a false positive rate of 5% (false positive being defined as an ARG simulated with no inversion but having at least one significant clade). Colours correspond to the different inversion lengths.

(a)		Inversion length (kb)			
		200	100	50	0
Inv. detected	+	100	97	99	5
	−	0	3	1	95
Clades tested		3.3m	3.4m	3.4m	3.4m
Significant clades		201	144	127	5
Within inv. region		96%	92%	97%	0%

(b)		Inversion length (kb)			
		200	100	50	0
Inv. detected	+	93	78	57	5
	−	7	22	43	95
Clades tested		116k	116k	117k	122k
Significant clades		227	123	79	5
Within inv. region		96%	97%	96%	0%

Table 1: Confusion matrices and results summaries for inversion detection test, based on 100 simulations for each given length of the inverted region: using (a) the simulated ARGs, (b) ARGs reconstructed using Relate.

4 Discussion

We have found that the distribution of branch duration is very accurately captured by the SMC' approximation to the CwR. Relate and tsinfer/tsdate seemingly do not capture the distribution of branch duration well, echoing the results of Deng et al. (2021) regarding the length of intervals between local trees.

The differences between the approximate distribution of branch durations and the distribution of branch durations in ARGs produced using reconstruction tools are due to both model misspecification and the particularities of each algorithm. We suggest that the bias seen in ARGs reconstructed using tsinfer stems from the presence of polytomies, which lead to an excess of deep branches with long durations. It is nontrivial to adjust for their presence within our model, and it is not currently possible to break polytomies at random using tskit without re-sampling branches in each tree independently (which would prevent a proper calculation of their duration). In general, deep branches can arise either through true demographic events or due to inaccuracies in ARG reconstruction; our method can be used to detect deep branches that are likely to be artefactual, which may, for instance, help with avoiding spurious inference of introgression.

We note that apart from ARGweaver, none of the ARG reconstruction methods we consider are explicitly optimised to recover the distribution of branch duration, focusing instead on other aspects such as node times, local tree topologies, and patterns of linkage disequilibrium (which *are* recovered well in our simulation studies). Our results can potentially be used to improve the estimates of branch duration produced by tsinfer and Relate during the topology reconstruction step and hence also improve the downstream inference of node times.

The concept of branch duration is closely related to the ARG-based definition of haplotype blocks proposed by Shipilina et al. (2023): the "set of genomic regions that descend from a particular edge in the ARG, which is defined by a unique coalescence event, and by the set of descendant samples". Our results thus provide a theoretical derivation for the distribution of the length of haplotype blocks under the SMC' approximation. Moreover, since we demonstrate that branch durations are not, in general, well reconstructed by Relate, tsinfer/tsdate and ARG-Needle, this definition may be difficult to apply to data in practice.

The SMC'-based approximation we construct for the distribution of clade duration also appears to provide an excellent fit based on simulations. The corresponding test we develop for detecting chromosomal inversions has excellent performance on simulated ARGs, as well as (with appropriate adjustment) ARGs reconstructed using Relate. Since the test detects long clade duration after adjusting for its age and the local tree topology, and hence specifically detects localised suppression of recombination, it has the power to discriminate inversions from other genealogy-distorting events such as point mutations under balancing selection. This opens up the exciting possibility of detecting inversions using ARGs reconstructed from large-scale genetic datasets.

Following the recent breakthroughs in scalable ARG reconstruction, there has been a flurry of interest in developing ARG-based methods for statistical inference of evolutionary events and

parameters. Such methods have the potential to leverage the efficiency with which ARGs store genetic information and avoid the loss of power faced by methods relying on summary statistics. However, structural variants are largely ignored by the currently available methods for ARG storage and reconstruction, while being an important part of the evolution of many species. Developments in this direction would bring us closer to accurately recording and reconstructing genealogies, paving the way to more powerful genealogy-based inference.

5 Data availability statement

Data sharing is not applicable to this article as no new data were generated or analysed.

6 Acknowledgements

This work was initiated at the *Stochastic modelling in the life sciences* Junior Trimester Programme held at the Hausdorff Research Institute for Mathematics, University of Bonn (funded by the Deutsche Forschungsgemeinschaft under Germany's Excellence Strategy EXC-2047/1-390685813). SRM and AI are supported by the Wellcome Trust (Investigator Award 212284/Z/18/Z), MF by the Knut and Alice Wallenberg Foundation (Program for Mathematics 2021, grant 2020.072), JK by the EPSRC (research grant EP/V049208/1), and JS by the ERC (Starting Grant ARGPHENO 850869). We thank Yan Wong and Pier Palamara for helpful discussions.

References

Adrion, J. R., Cole, C. B., Dukler, N., Galloway, J. G., Gladstein, A. L., Gower, G., Kyriazis, C. C., Ragsdale, A. P., Tsambos, G., Baumdicker, F., Carlson, J., Cartwright, R. A., Durvasula, A., Gronau, I., Kim, B. Y., McKenzie, P., Messer, P. W., Noskova, E., Ortega-Del Vecchyo, D., Racimo, F., Struck, T. J., Gravel, S., Gutenkunst, R. N., Lohmueller, K. E., Ralph, P. L., Schrider, D. R., Siepel, A., Kelleher, J. and Kern, A. D. (2020). A community-maintained standard library of population genetic models. *eLife*, **9**, e54967.

Baumdicker, F., Bisschop, G., Goldstein, D., Gower, G., Ragsdale, A. P., Tsambos, G., Zhu, S., Eldon, B., Ellerman, E. C., Galloway, J. G., Gladstein, A. L., Gorjanc, G., Guo, B., Jeffery, B., Kretzschumar, W. W., Lohse, K., Matschiner, M., Nelson, D., Pope, N. S., Quinto-Cortés, C. D., Rodrigues, M. F., Saunack, K., Sellinger, T., Thornton, K., van Kemenade, H., Wohns, A. W., Wong, Y., Gravel, S., Kern, A. D., Koskela, J., Ralph, P. L. and Kelleher, J. (2022). Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*, **220**(3), iyab229.

Brandt, D. Y. C., Wei, X., Deng, Y., Vaughn, A. H. and Nielsen, R. (2022). Evaluation of methods for estimating coalescence times using ancestral recombination graphs. *Genetics*, **221**(1), iyac044.

Carmi, S., Wilton, P. R., Wakeley, J. and Pe'er, I. (2014). A renewal theory approach to IBD sharing. *Theoretical Population Biology*, **97**, 35–48.

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

Eriksson, A., Mahjani, B. and Mehlig, B. (2009). Sequential Markov coalescent algorithms for population models with demographic structure. *Theoretical Population Biology*, **76**(2), 84–91.

Griffiths, R. C. and Marjoram, P. (1997). An ancestral recombination graph. In P. Donnelly and S. Tavaré, editors, *Progress in population genetics and human evolution*, pp. 257–270. Springer, New York.

Haller, B. C., Galloway, J., Kelleher, J., Messer, P. W. and Ralph, P. L. (2019). Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes. *Molecular Ecology Resources*, **19**(2), 552–566.

Haller, B. C. and Messer, P. W. (2019). SLiM 3: Forward genetic simulations beyond the Wright–Fisher model. *Molecular Biology and Evolution*, **36**(3), 632–637.

Hammer, M. F., Schimenti, J. and Silver, L. M. (1989). Evolution of mouse chromosome 17 and the origin of inversions associated with T haplotypes. *Proceedings of the National Academy of Sciences*, **86**(9), 3261–3265.

Hanlon, V. C., Lansdorp, P. M. and Guryev, V. (2022). A survey of current methods to detect and genotype inversions. *Human Mutation*, **43**(11), 1576–1589.

Harris, K. and Nielsen, R. (2013). Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS Genetics*, **9**(6), e1003521.

Hejase, H. A., Mo, Z., Campagna, L. and Siepel, A. (2022). A deep-learning approach for inference of selective sweeps from the ancestral recombination graph. *Molecular Biology and Evolution*, **39**(1), msab332.

Hobolth, A. and Jensen, J. L. (2014). Markovian approximation to the finite loci coalescent with recombination along multiple sequences. *Theoretical Population Biology*, **98**, 48–58.

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

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

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

Kirkpatrick, M. (2010). How and why chromosome inversions evolve. *PLoS Biology*, **8**(9), e1000501.

Kirkpatrick, M. and Barton, N. (2006). Chromosome inversions, local adaptation and speciation. *Genetics*, **173**(1), 419–434.

Lauterbur, M. E., Cavassim, M. I. A., Gladstein, A. L., Gower, G., Pope, N. S., Tsambos, G., Adrion, J., Belsare, S., Biddanda, A., Caudill, V., Cury, J., Echevarria, I., Haller, B. C., Hasan, A. R., Huang, X., Iasi, L. N. M., Noskova, E., Obšteter, J., Pavinato, V. A. C., Pearson, A., Peede, D., Perez, M. F., Rodrigues, M. F., Smith, C. C. R., Spence, J. P., Teterina, A., Tittes, S., Unneberg, P., Vazquez, J. M., Waples, R. K., Wohns, A. W., Wong, Y., Baumdicker, F., Cartwright, R. A., Gorjanc, G., Gutenkunst, R. N., Kelleher, J., Kern, A. D., Ragsdale, A. P., Ralph, P. L., Schrider, D. R. and Gronau, I. (2022). Expanding the stdpopsim species catalog, and lessons learned for realistic genome simulations. *bioRxiv*. doi:10.1101/2022.10.29.514266.

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

Li, N. and Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, **165**(4), 2213–2233.

Lucas Lledó, J. I. and Cáceres, M. (2013). On the power and the systematic biases of the detection of chromosomal inversions by paired-end genome sequencing. *PLoS One*, **8**(4), e61292.

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

McKenzie, P. F. and Eaton, D. A. R. (2022). Estimating waiting distances between genealogy changes under a multi-species extension of the sequentially Markov coalescent. *bioRxiv*. doi: 10.1101/2022.08.19.504573.

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

Puig, M., Casillas, S., Villatoro, S. and Cáceres, M. (2015). Human inversions and their functional consequences. *Briefings in Functional Genomics*, **14**(5), 369–379.

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

Rausch, T., Zichner, T., Schlattl, A., Stütz, A. M., Benes, V. and Korbel, J. O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, **28**(18), i333–i339.

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

Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., Von Haeseler, A. and Schatz, M. C. (2018). Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods*, **15**(6), 461–468.

Shipilina, D., Pal, A., Stankowski, S., Chan, Y. F. and Barton, N. H. (2023). On the origin and structure of haplotype blocks. *Molecular Ecology*, **32**(6), 1441–1457.

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

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M. et al. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, **526**(7571), 75–81.

Tattini, L., D'Aurizio, R. and Magi, A. (2015). Detection of genomic structural variants from next-generation sequencing data. *Frontiers in Bioengineering and Biotechnology*, **3**, 92.

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

Wiuf, C. and Hein, J. (1999). Recombination as a point process along sequences. *Theoretical Population Biology*, **55**(3), 248–259.

Wohns, A. W., Wong, Y., Jeffery, B., Akbari, A., Mallick, S., Pinhasi, R., Patterson, N., Reich, D., Kelleher, J. and McVean, G. (2022). A unified genealogy of modern and ancient genomes. *Science*, **375**(6583), eabi8264.

Zhang, B. C., Biddanda, A., Gunnarsson, Á. F., Cooper, F. and Palamara, P. F. (2023). Biobank-scale inference of ancestral recombination graphs enables genealogical analysis of complex traits. *Nature Genetics*, **55**, 768–776.

Appendix

A1 Proof of Proposition 2.1

Conditioning on the recombination happening on branch β, the density of the recombination event time is

$$p^S_{\mathcal{T}}(s|\mathcal{R}\in\beta) = \begin{cases} \frac{1}{\bar{t}(\beta)} & t^{\downarrow}(\beta)\le s\le t^{\uparrow}(\beta) \\ 0 & \text{otherwise,} \end{cases}$$

as the recombination time is chosen uniformly at random along the length of the branch. The conditional density of the coalescence time is

$$p^U_{\mathcal{T}}(u|\mathcal{R}=(\beta,s)) = p^U_{\mathcal{T}}(u|\mathcal{R}=(\cdot,s)) = n(u)\exp\left(-\int_s^u n(t)dt\right), \tag{A1}$$

for $u>s$ and 0 otherwise. Conditional on the coalescence time u, the probability that the coalescence point is on β is

$$\mathbb{P}_{\mathcal{T}}(\mathcal{C}\in\beta|\mathcal{C}=(\cdot,u),\mathcal{R}=(\beta,s)) = \begin{cases} \frac{1}{n(u)} & s<u<t^{\uparrow}(\beta) \\ 0 & \text{otherwise.} \end{cases}$$

Letting $k=n(s)$, so that T_k is the first coalescence time just above time s,

$$\begin{aligned}
\mathbb{P}_{\mathcal{T}}(\mathcal{C}\in\beta|\mathcal{R}=(\beta,s)) &= \int_0^\infty \mathbb{P}(\mathcal{C}\in\beta|\mathcal{C}=(\cdot,u),\mathcal{R}=(\beta,s))\cdot p^U_{\mathcal{T}}(u|\mathcal{R}=(\cdot,s))du \\
&= \int_s^{t^{\uparrow}(\beta)} \exp\left(-\int_s^u n(t)dt\right)du \\
&= \int_s^{T_k} \exp\left(-\int_s^u k\,dt\right)du + \sum_{j=n(t^{\uparrow}(\beta))+1}^{k-1}\int_{T_{j+1}}^{T_j}\exp\left(-\int_s^u n(t)\,dt\right)du,
\end{aligned}$$

note that $T_{n(t^{\uparrow}(b))+1} = t^{\uparrow}(b)$. The first term is

$$\begin{aligned}
\int_s^{T_k}\exp\left(-\int_s^u k\,dt\right)du &= \int_s^{T_k}\exp\left(-k(u-s)\right)du \\
&= \left[-\frac{1}{k}e^{-k(u-s)}\right]_s^{T_k} \\
&= \frac{1}{k}-\frac{1}{k}e^{-kT_k}e^{ks},
\end{aligned} \tag{A2}$$

and the summands of the second term are

$$\begin{aligned}
\int_{T_{j+1}}^{T_j}\exp\left(-\int_s^u n(t)\,dt\right)du &= \int_{T_{j+1}}^{T_j}\exp\left(-\int_s^{T_{j+1}} n(t)\,dt-\int_{T_{j+1}}^u n(t)\,dt\right)du \\
&= \exp\left(-\int_s^{T_{j+1}} n(t)\,dt\right)\int_{T_{j+1}}^{T_j}\exp\left(-\int_{T_{j+1}}^u j\,dt\right)du \\
&= \exp\left(-k(T_k-s)-\sum_{i=j+1}^{k-1} i(T_i-T_{i+1})\right)\left[-\frac{1}{j}e^{-j(u-T_{j+1})}\right]_{T_{j+1}}^{T_j} \\
&= e^{ks}\exp\left(-kT_k+L_{\mathcal{T}}(T_{j+1})-L_{\mathcal{T}}(T_k)\right)\frac{1}{j}\left(1-e^{-j(T_j-T_{j+1})}\right).
\end{aligned}$$

Thus,

$$\mathbb{P}_{\mathcal{T}}(\mathcal{C} \in \beta | \mathcal{R} = (\beta, s)) = \frac{1}{k} + e^{ks} \cdot \sum_{j=n(t^{\uparrow}(\beta))+1}^{k} Q_{kj},$$

where

$$Q_{kk} := -\frac{1}{k} e^{-kT_k},$$

and

$$\begin{aligned} Q_{kj} &:= \frac{1}{j}\left(1 - e^{-j(T_j - T_{j+1})}\right) \exp\left(-kT_k + L_{\mathcal{T}}(T_{j+1}) - L_{\mathcal{T}}(T_k)\right) \\ &= \frac{1}{j} e^{-kT_k} e^{-L_{\mathcal{T}}(T_k)} \left(e^{L_{\mathcal{T}}(T_{j+1})} - e^{L_{\mathcal{T}}(T_{j+1}) - j(T_j - T_{j+1})}\right) \\ &= e^{-kT_k} e^{-L_{\mathcal{T}}(T_k)} \frac{1}{j} \left(e^{L_{\mathcal{T}}(T_{j+1})} - e^{L_{\mathcal{T}}(T_j)}\right). \end{aligned}$$

A2 Proof of Proposition 2.2

Marginalising out the recombination time in (2.4),

$$\begin{aligned} \mathbb{P}_{\mathcal{T}}(\mathcal{C} \notin \beta | \mathcal{R} \in \beta) &= 1 - \int_{t^{\downarrow}(\beta)}^{t^{\uparrow}(\beta)} \mathbb{P}_{\mathcal{T}}(\mathcal{C} \in \beta | \mathcal{R} = (\beta, s)) p_{\mathcal{T}}^{S}(s | \mathcal{R} \in \beta)\, ds \\ &= 1 - \frac{1}{\bar{t}(\beta)} \sum_{k=n(t^{\uparrow}(\beta))+1}^{n(t^{\downarrow}(\beta))} \int_{T_{k+1}}^{T_k} \left(\frac{1}{k} + e^{ks} \cdot \sum_{j=n(t^{\uparrow}(\beta))+1}^{k} Q_{kj}\right) ds \\ &= 1 - \frac{1}{\bar{t}(\beta)} \sum_{k=n(t^{\uparrow}(\beta))+1}^{n(t^{\downarrow}(\beta))} \left(\widetilde{Q}^1(k) + \widetilde{Q}^2(k, k, n(t^{\uparrow}(\beta)) + 1, 0, 1)\right), \end{aligned}$$

where

$$\widetilde{Q}^1(k) := \frac{1}{k}\left(T_k - T_{k+1}\right),$$

and for $x, y, A, B \in \mathbb{Z}$, $x \geq k$, $2 \leq y \leq x$,

$$\widetilde{Q}^2(k, x, y, A, B) := \frac{1}{k}\left(e^{kT_k} - e^{kT_{k+1}}\right) \sum_{j=y}^{x} (Aj + B) \cdot Q_{kj}.$$

A3 Proof of Proposition 2.3

We have

$$\mathbb{P}_{\mathcal{T}}(\mathcal{C} \in \beta | \mathcal{C} = (\cdot, u), \mathcal{R} \notin \mathcal{B}(\beta), \mathcal{R} = (\cdot, s)) = \begin{cases} \frac{1}{n(u)} & \max(s, t^{\downarrow}(\beta)) < u < t^{\uparrow}(\beta) \\ 0 & \text{otherwise.} \end{cases}$$

For $s < t^{\downarrow}(\beta)$,

$$\begin{aligned} \int_{t^{\downarrow}(\beta)}^{t^{\uparrow}(\beta)} \exp\left(-\int_s^u n(t) dt\right) du &= \sum_{j=n(t^{\uparrow}(\beta))+1}^{n(t^{\downarrow}(\beta))} \int_{T_{j+1}}^{T_j} \exp\left(-\int_s^u n(t) dt\right) du \\ &= \sum_{j=n(t^{\uparrow}(\beta))+1}^{n(t^{\downarrow}(\beta))} \exp\left(-\int_s^{T_{j+1}} n(t)\, dt\right) \int_{T_{j+1}}^{T_j} \exp\left(-\int_{T_{j+1}}^u j\, dt\right) du \end{aligned}$$

$$= e^{ks} \cdot \sum_{j=n(t^{\uparrow}(\beta))+1}^{n(t^{\downarrow}(\beta))} Q_{kj},$$

and the case $s \geq t^{\downarrow}(\beta)$ is given by (A2). Thus,

$$\mathbb{P}_{\mathcal{T}}(\mathcal{C} \in \beta | \mathcal{R} \notin \mathcal{B}(\beta), \mathcal{R} = (\cdot, s)) = \int_{\max(s, t^{\downarrow}(\beta))}^{t^{\uparrow}(\beta)} \exp\left(-\int_s^u n(t)dt\right) du$$

$$= \begin{cases} e^{ks} \sum_{j=n(t^{\uparrow}(\beta))+1}^{n(t^{\downarrow}(\beta))} Q_{kj} & s < t^{\downarrow}(\beta) \\ \frac{1}{k} + e^{ks} \sum_{j=n(t^{\uparrow}(\beta))+1}^{k} Q_{kj} & t^{\downarrow}(\beta) \leq s < t^{\uparrow}(\beta) \\ 0 & \text{otherwise.} \end{cases}$$

A4 Proof of Proposition 2.4

Consider all of the possible orderings of the event times $t_1, \ldots, t_4$, as illustrated in Figure A1.

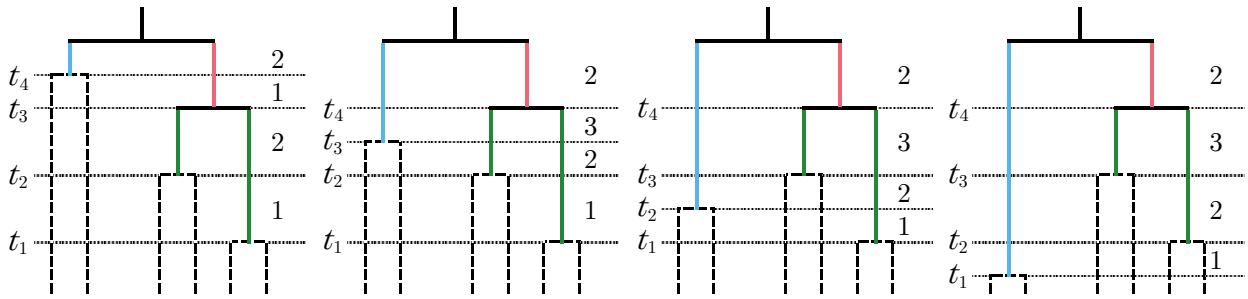

Figure A1: All possible orderings of $t^{\downarrow}(\text{ch}_1(\beta)), t^{\downarrow}(\text{ch}_2(\beta)), t^{\downarrow}(\text{sib}(\beta)), t^{\downarrow}(\beta)$. The branch β is shown in red, $\text{sib}(\beta)$ in blue, $\text{ch}_1(\beta)$ and $\text{ch}_2(\beta)$ in green. Numbers to the right of each tree show the number of lineages in $\mathcal{B}(\beta)$ in each time interval. For instance, in the leftmost tree, $t_1 = t^{\downarrow}(\text{ch}_2(\beta))$, $t_2 = t^{\downarrow}(\text{ch}_1(\beta))$, $t_3 = t^{\downarrow}(\beta)$ and $t_4 = t^{\downarrow}(\text{sib}(\beta))$.

The number of lineages in set $\mathcal{B}(\beta)$ at time s can be written as

$$n_{\mathcal{B}(\beta)}(s) := \sum_{b' \in \mathcal{B}(\beta)} \mathbb{1}(s \in [t^{\downarrow}(b'), t^{\uparrow}(b')])$$

$$= \begin{cases} 1 & t_1 \leq s < t_2 \\ 2 & t_2 \leq s < t_3 \\ 1 + 2 \cdot \mathbb{1}(t^{\downarrow}(\text{sib}(\beta)) < t^{\downarrow}(\beta)) & t_3 \leq s < t_4 \\ 2 & t_4 \leq s < t_5 = t^{\uparrow}(\beta) \\ 0 & \text{otherwise,} \end{cases}$$

where $\mathbb{1}(\cdot)$ is the indicator function. Then conditional on the recombination point not being on a branch in the set $\mathcal{B}(\beta)$, the density of the recombination event time is

$$p^S_{\mathcal{T}}(s | \mathcal{R} \notin \mathcal{B}(\beta)) = \begin{cases} \frac{n(s) - n_{\mathcal{B}(\beta)}(s)}{L_{\mathcal{T}}(0) - \sum_{b' \in \mathcal{B}(\beta)} \bar{t}(b')} & 0 \leq s \leq T_2 \\ 0 & \text{otherwise.} \end{cases}$$

Marginalising out the recombination time in (2.10),

$$\begin{aligned}
\mathbb{P}_{\mathcal{T}}(\mathcal{C} \in \beta | \mathcal{R} \notin \mathcal{B}(\beta)) &= \int_0^{t^{\uparrow}(\beta)} \mathbb{P}_{\mathcal{T}}(\mathcal{C} \in \beta | \mathcal{R} \notin \mathcal{B}(\beta), \mathcal{R} = (\cdot, s)) \cdot p^S_{\mathcal{T}}(s | \mathcal{R} \notin \mathcal{B}(\beta))\, ds. \\
&= \frac{1}{L_{\mathcal{T}}(0) - \sum_{b' \in \mathcal{B}(\beta)} \bar{t}(b')} \Bigg\{ \sum_{k=n(t_1)+1}^{n} k \widetilde{Q}^2(k, n(t^{\downarrow}(\beta)), n(t^{\uparrow}(\beta)) + 1, 0, 1) \\
&\quad + \sum_{k=n(t_2)+1}^{n(t_1)} (k-1) \widetilde{Q}^2(k, n(t^{\downarrow}(\beta)), n(t^{\uparrow}(\beta)) + 1, 0, 1) \\
&\quad + \sum_{k=n(t_3)+1}^{n(t_2)} (k-2) \widetilde{Q}^2(k, n(t^{\downarrow}(\beta)), n(t^{\uparrow}(\beta)) + 1, 0, 1) \\
&\quad + \sum_{k=n(t_4)+1}^{n(t_3)} \Big[\mathbb{1}(t^{\downarrow}(\mathrm{sib}(\beta)) < t^{\downarrow}(\beta))(k-3) \widetilde{Q}^2(k, n(t^{\downarrow}(\beta)), n(t^{\uparrow}(\beta)) + 1, 0, 1) \\
&\qquad + \mathbb{1}(t^{\downarrow}(\mathrm{sib}(\beta)) \geq t^{\downarrow}(\beta))(k-1) \left(\widetilde{Q}^1(k) + \widetilde{Q}^2(k, k, n(t^{\uparrow}(\beta)) + 1, 0, 1) \right) \Big] \\
&\quad + \sum_{k=n(t_5)+1}^{n(t_4)} (k-2) \left(\widetilde{Q}^1(k) + \widetilde{Q}^2(k, k, n(t^{\uparrow}(\beta)) + 1, 0, 1) \right) \Bigg\}.
\end{aligned}$$

A5 Proof of Theorem 2.1

If we now say that a branch is disrupted only if there is a change in topology (but not branch length), we can allow the events shown in Figure 2 in blue (marked with dots), i.e. those where the recombination point is on branch $\beta \in \mathcal{B}(\beta)$, and the coalescence point is on one of the branches in $\mathcal{A}(\beta)$. Thus,

$$\begin{aligned}
\mathbb{P}_{\mathcal{T}}(b \text{ topologically disrupted}) = &\sum_{b' \in \mathcal{B}(b)} \mathbb{P}_{\mathcal{T}}(\mathcal{C} \notin \mathcal{A}(b') | \mathcal{R} \in b') \cdot \mathbb{P}_{\mathcal{T}}(\mathcal{R} \in b') \\
&+ \mathbb{P}_{\mathcal{T}}(\mathcal{C} \in b | \mathcal{R} \notin \mathcal{B}(b)) \cdot \mathbb{P}_{\mathcal{T}}(\mathcal{R} \notin \mathcal{B}(b)).
\end{aligned}$$

To calculate the probability $\mathbb{P}_{\mathcal{T}}(\mathcal{C} \notin \mathcal{A}(\beta) | \mathcal{R} \in \beta)$, we follow the same approach as the proofs of Propositions 2.1 and 2.2, first conditioning on the recombination point $\mathcal{R} = (\beta, s)$. Let $k = n(s)$, so that T_k is the first coalescence time just above time s. Let

$$r(u) := \exp\left(- \int_s^u n(t) dt \right).$$

Then if $t^{\downarrow}(\mathrm{sib}(b)) < s$,

$$\begin{aligned}
\mathbb{P}_{\mathcal{T}}(\mathcal{C} \in \mathcal{A}(\beta) | \mathcal{R} = (\beta, s)) &= \int_0^{\infty} \mathbb{P}(\mathcal{C} \in \mathcal{A}(\beta) | \mathcal{C} = (\cdot, u), \mathcal{R} = (\beta, s)) \cdot p^U_{\mathcal{T}}(u | \mathcal{R} = (\cdot, s)) du \\
&= \int_s^{t^{\uparrow}(\beta)} 2 \cdot r(u)\, du + \int_{t^{\uparrow}(\beta)}^{t^{\uparrow}(\mathrm{par}(\beta))} r(u)\, du \\
&= \frac{2}{k} + e^{ks} \cdot \left(\sum_{j=n(t^{\uparrow}(\beta))+1}^{k} 2 \cdot Q_{kj} + \sum_{j=n(t^{\uparrow}(\mathrm{par}(\beta)))+1}^{n(t^{\uparrow}(\beta))} Q_{kj} \right),
\end{aligned}$$

and if $t^{\downarrow}(\text{sib}(b)) \geq s$,

$$\mathbb{P}_{\mathcal{T}}(\mathcal{C} \in \mathcal{A}(\beta)|\mathcal{R} = (\beta, s)) = \int_s^{t^{\downarrow}(\text{sib}(\beta))} r(u)\, du + \int_{t^{\downarrow}(\text{sib}(\beta))}^{t^{\uparrow}(\beta)} 2 \cdot r(u)\, du + \int_{t^{\uparrow}(\beta)}^{t^{\uparrow}(\text{par}(\beta))} r(u)\, du$$
$$= \frac{1}{k} + e^{ks} \cdot \left(\sum_{j=n(t^{\downarrow}(\text{sib}(\beta)))+1}^{k} Q_{kj} + \sum_{j=n(t^{\uparrow}(\beta))+1}^{n(t^{\downarrow}(\text{sib}(\beta)))} 2 \cdot Q_{kj} + \sum_{j=n(t^{\uparrow}(\text{par}(\beta)))+1}^{n(t^{\uparrow}(\beta))} Q_{kj} \right).$$

Marginalising out the recombination time,

$$\mathbb{P}_{\mathcal{T}}(\mathcal{C} \notin \mathcal{A}(\beta)|\mathcal{R} \in \beta) = 1 - \int_{t^{\downarrow}(\beta)}^{t^{\uparrow}(\beta)} \mathbb{P}_{\mathcal{T}}(\mathcal{C} \in \mathcal{A}(\beta)|\mathcal{R} = (\beta, s)) p_{\mathcal{T}}^{S}(s|\mathcal{R} \in \beta)\, ds$$
$$= 1 - \frac{1}{\bar{t}(\beta)} \sum_{k=n(t^{\uparrow}(\beta))+1}^{n(t^{\downarrow}(\beta))} G_{\beta}(k),$$

where for $k \leq n(t^{\downarrow}(\text{sib}(\beta)))$,

$$G_{\beta}(k) \;=\; 2 \cdot \widetilde{Q}^1(k) + 2 \cdot \widetilde{Q}^2(k, k, n(t^{\uparrow}(\beta)) + 1, 0, 1) + \widetilde{Q}^2(k, n(t^{\uparrow}(\beta)), n(t^{\uparrow}(\text{par}(\beta))) + 1, 0, 1),$$

and for $k > n(t^{\downarrow}(\text{sib}(\beta)))$

$$G_{\beta}(k) = \widetilde{Q}^1(k) + \widetilde{Q}^2(k, k, n(t^{\downarrow}(\text{sib}(\beta))) + 1, 0, 1) + 2 \cdot \widetilde{Q}^2(k, n(t^{\downarrow}(\text{sib}(\beta))), n(t^{\uparrow}(\beta)) + 1, 0, 1)$$
$$+ \widetilde{Q}^2(k, n(t^{\uparrow}(\beta)), n(t^{\uparrow}(\text{par}(\beta))) + 1, 0, 1).$$

A6 Proof of Proposition 2.5

A6.1 Probability of no change in total branch length

The probability that the recombination event does not result in a change in total branch length is

$$\mathbb{P}_{\mathcal{T}}(C = 0) = \sum_{b \in \mathcal{T}} \frac{\bar{t}(b)}{L_{\mathcal{T}}(0)} \left(1 - P_{\mathcal{T}}(\mathcal{C} \notin b|\mathcal{R} \in b)\right)$$
$$= \frac{1}{L_{\mathcal{T}}(0)} \sum_{b \in \mathcal{T}} \sum_{k=n(t^{\uparrow}(\beta))+1}^{n(t^{\downarrow}(\beta))} \left(\widetilde{Q}^1(k) + \widetilde{Q}^2(k, k, n(t^{\uparrow}(\beta)) + 1, 0, 1) \right),$$

using (2.7). This is equal to the probability derived by Deng et al. (2021, Theorem 1).

A6.2 Probability that change in total branch length is negative

We have

$$\mathbb{P}_{\mathcal{T}}(C < 0|\mathcal{R} = (b, s)) = \int_s^{t^{\uparrow}(b)} \frac{n(u) - 1}{n(u)} \cdot n(u) \exp\left(- \int_s^u n(t)\, dt \right) du$$
$$= \frac{k-1}{k} + e^{ks} \sum_{j=n(t^{\uparrow}(b))+1}^{k} (j-1) Q_{kj}.$$

Integrating over the recombination time,

$$\mathbb{P}_{\mathcal{T}}(C<0|\mathcal{R}\in b) = \frac{1}{\bar{t}(b)}\int_{t^{\downarrow}(b)}^{t^{\uparrow}(b)}\left(\frac{k-1}{k}+e^{ks}\sum_{j=n(t^{\uparrow}(b))+1}^{k}(j-1)Q_{kj}\right)ds$$
$$= \frac{1}{\bar{t}(b)}\sum_{k=n(t^{\uparrow}(b))+1}^{n(t^{\downarrow}(b))}\left((k-1)\widetilde{Q}^1(k)+\widetilde{Q}^2(k,k,n(t^{\uparrow}(\beta))+1,1,-1)\right).$$

Then

$$\mathbb{P}_{\mathcal{T}}(C<0) = \frac{1}{L\mathcal{T}(0)}\sum_{b\in\mathcal{T}}\sum_{k=n(t^{\uparrow}(b))+1}^{n(t^{\downarrow}(b))}\left((k-1)\widetilde{Q}^1(k)+\widetilde{Q}^2(k,k,n(t^{\uparrow}(\beta))+1,1,-1)\right).$$

A6.3 Probability that change in total branch length is positive

Similarly, the probability that the change in branch length is positive is given by

$$\mathbb{P}_{\mathcal{T}}(C>0|\mathcal{R}=(b,s)) = \int_{t^{\uparrow}(b)}^{\infty} n(u)\exp\left(-\int_s^u n(t)\,dt\right)du$$
$$= \int_{t^{\uparrow}(b)}^{T_2} n(u)\exp\left(-\int_s^u n(t)\,dt\right)du + \int_{T_2}^{\infty}\exp\left(-\int_s^u n(t)\,dt\right)du$$
$$= e^{ks}\sum_{j=2}^{n(t^{\uparrow}(b))} jQ_{kj} + e^{-L\mathcal{T}(s)},$$

giving

$$\mathbb{P}_{\mathcal{T}}(C>0|\mathcal{R}\in b) = \frac{1}{\bar{t}(b)}\sum_{k=n(t^{\uparrow}(b))+1}^{n(t^{\downarrow}(b))}\left(\widetilde{Q}^2(k,n(t^{\uparrow}(b)),2,1,0)+\widetilde{Q}^3(k)\right),$$

where

$$\widetilde{Q}^3(k) = \int_{T_{k+1}}^{T_k} e^{-L\mathcal{T}(s)}\,ds$$
$$= \int_{T_{k+1}}^{T_k} e^{ks}\exp\left(-kT_k - L\mathcal{T}(T_k)\right)ds$$
$$= \frac{1}{k}\left(e^{kT_k}-e^{kT_{k+1}}\right)\exp\left(-kT_k - L\mathcal{T}(T_k)\right)$$
$$= \frac{1}{k}\left(e^{-L\mathcal{T}(T_k)}-e^{-L\mathcal{T}(T_{k+1})}\right).$$

Thus,

$$\mathbb{P}_{\mathcal{T}}(C>0) = \frac{1}{L\mathcal{T}(0)}\sum_{b\in\mathcal{T}}\sum_{k=n(t^{\uparrow}(b))+1}^{n(t^{\downarrow}(b))}\left(\widetilde{Q}^2(k,n(t^{\uparrow}(b)),2,1,0)+\widetilde{Q}^3(k)\right).$$

A7 Proof of Proposition 2.6

We first calculate the density of change in total branch length conditional on the recombination point $\mathcal{R}=(b,s)$, then marginalise out the recombination time and branch.

A7.1 Density of change in total branch length conditional on recombination point

Suppose that the recombination point is on branch b at time s. Then given the coalescence time u, the change in total branch length $C > t^{\downarrow}(b) - t^{\uparrow}(b) = -\bar{t}(b)$, given $C \neq 0$ is

$$C = \begin{cases} u - t^{\uparrow}(b) + u - T_2 = 2u - t^{\uparrow}(b) - T_2 & u > T_2 \\ u - t^{\uparrow}(b) & \text{otherwise.} \end{cases}$$

Let $l = n(u)$, so T_l is the time of the first coalescence event just above time u. Through conditioning on the coalescence point not being on branch b, and a change of variable in (A1), for $s - t^{\uparrow}(b) < c < 0$

$$\begin{aligned} p_{\mathcal{T}}^{C}(c|C \neq 0, \mathcal{R} = (b,s)) &= (n(u)-1)\exp\left(-\int_s^u (n(t)-1)\,dt\right) \\ &= (n(u)-1)\exp\left(\left[-\int_s^{T_k} - \int_{T_k}^{T_l} + \int_u^{T_l}\right](n(t)-1)\,dt\right) \\ &= e^{(k-1)s} P_{kl}^1 (n(u)-1) e^{-(l-1)u} \\ &= e^{(k-1)s} P_{kl}^1 (n(c+t^{\uparrow}(b))-1) e^{-(l-1)(c+t^{\uparrow}(b))}, \end{aligned}$$

where

$$\begin{aligned} P_{kl}^1 &:= \exp\left(-(k-1)T_k - \sum_{i=l}^{k-1}(i-1)(T_i - T_{i+1}) + (l-1)T_l\right) \\ &= \exp\left(l \cdot T_l - k \cdot T_k + L_{\mathcal{T}}(T_l) - L_{\mathcal{T}}(T_k)\right). \end{aligned}$$

For $0 < c \leq T_2 - t^{\uparrow}(b)$,

$$\begin{aligned} p_{\mathcal{T}}^{C}(c|C \neq 0, \mathcal{R} = (b,s)) &= n(u)\exp\left(-\int_s^{t^{\uparrow}(b)} (n(t)-1)\,dt - \int_{t^{\uparrow}(b)}^u n(t)\,dt\right) \\ &= \exp\left(-\int_s^{t^{\uparrow}(b)} (n(t)-1)\,dt\right) n(u)\exp\left(-\int_{t^{\uparrow}(b)}^u n(t)\,dt\right) \\ &= e^{(k-1)s} P_k^2 n(c+t^{\uparrow}(b))\exp\left(-\int_{t^{\uparrow}(b)}^{c+t^{\uparrow}(b)} n(t)\,dt\right) \\ &= e^{(k-1)s} P_k^2 n(c+t^{\uparrow}(b))\exp\left(-\left[L_{\mathcal{T}}(t^{\uparrow}(b)) - L_{\mathcal{T}}(c+t^{\uparrow}(b))\right]\right), \end{aligned}$$

where

$$\begin{aligned} P_k^2 &:= \exp\left(-(k-1)T_k - \sum_{i=n(t^{\uparrow}(b))+1}^{k-1}(i-1)(T_i - T_{i+1})\right) \\ &= \exp\left(t^{\uparrow}(b) - k \cdot T_k - L_{\mathcal{T}}(T_k) + L_{\mathcal{T}}(t^{\uparrow}(b))\right). \end{aligned}$$

Finally, for $c > T_2 - t^{\uparrow}(b)$,

$$\begin{aligned} p_{\mathcal{T}}^{C}(c|C \neq 0, \mathcal{R} = (b,s)) &= \exp\left(-\int_s^{t^{\uparrow}(b)} (n(t)-1)\,dt - \int_{t^{\uparrow}(b)}^{T_2} n(t)\,dt - \int_{T_2}^u 1\,dt\right) \\ &= \exp\left(-\int_s^{T_2} n(t)\,dt + \int_s^{t^{\uparrow}(b)} 1\,dt - \int_{T_2}^u 1\,dt\right) \end{aligned}$$

$$
\begin{aligned}
&= e^{(k-1)s} P_k^3 e^{-(u-T_2)} \\
&= \frac{1}{2} e^{(k-1)s} P_k^3 e^{-(c+t^{\uparrow}(b)-T_2)/2},
\end{aligned}
$$

where

$$
\begin{aligned}
P_k^3 &:= \exp\left(-(k-1)T_k - \sum_{i=2}^{k-1} i(T_i - T_{i+1}) + \sum_{i=n(t^{\uparrow}(b))+1}^{k-1} (T_i - T_{i+1}) \right) \\
&= \exp\left(t^{\uparrow}(b) - k \cdot T_k - L_{\mathcal{T}}(T_k) \right).
\end{aligned}
$$

Note that

$$
e^{(k-1)s} P_k^3 = \exp\left(-\left[L_{\mathcal{T}}(s) - (t^{\uparrow}(b) - s) \right] \right)
$$

gives the probability that the coalescence event happens above T_2. The conditional density of C is thus

$$
p_{\mathcal{T}}^C(c|C \neq 0, \mathcal{R} = (b,s)) = \begin{cases} e^{(k-1)s} P_{kl}^1 (n(c + t^{\uparrow}(b)) - 1) e^{-(l-1)(c+t^{\uparrow}(b))} & s - t^{\uparrow}(b) \leq c < 0 \\ e^{(k-1)s} P_k^2 n(c + t^{\uparrow}(b)) e^{-\left[L_{\mathcal{T}}(t^{\uparrow}(b)) - L_{\mathcal{T}}(c+t^{\uparrow}(b))\right]} & 0 < c \leq T_2 - t^{\uparrow}(b) \\ \frac{1}{2} e^{(k-1)s} P_k^3 e^{-(c+t^{\uparrow}(b)-T_2)/2} & c > T_2 - t^{\uparrow}(b) \\ 0 & \text{otherwise.} \end{cases}
$$

A7.2 Density of change in total branch length

Marginalising out the position and time of the recombination point, we have

$$
\begin{aligned}
p_{\mathcal{T}}^C(c|C \neq 0) &= \sum_{b \in \mathcal{T}} \mathbb{P}(\mathcal{R} \in b) \int_0^{\infty} p_{\mathcal{T}}^S(s|\mathcal{R} \in b)\, p_{\mathcal{T}}^C(c|C \neq 0, \mathcal{R} = (b,s))\, ds \\
&= \sum_{\substack{b \in \mathcal{T}: \\ c \geq -\bar{t}(b)}} \frac{\bar{t}(b)}{L_{\mathcal{T}}(0)} \int_{t^{\downarrow}(b)}^{t^{\uparrow}(b)+\min(0,c)} \frac{1}{\bar{t}(b)}\, p_{\mathcal{T}}^C(c|C \neq 0, \mathcal{R} = (b,s))\, ds \\
&= \frac{1}{L_{\mathcal{T}}(0)} \sum_{\substack{b \in \mathcal{T}: \\ c \geq -\bar{t}(b)}} \int_{t^{\downarrow}(b)}^{t^{\uparrow}(b)+\min(0,c)} p_{\mathcal{T}}^C(c|C \neq 0, \mathcal{R} = (b,s))\, ds.
\end{aligned}
$$

Let

$$
\begin{aligned}
\widetilde{P}_{kl}^1 &:= \int_{T_{k+1}}^{T_k} e^{(k-1)s} P_{kl}^1\, ds = \frac{1}{k-1} \left(e^{(k-1)T_k} - e^{(k-1)T_{k+1}} \right) P_{kl}^1, \\
\widetilde{P}_k^2 &:= \int_{T_{k+1}}^{T_k} e^{(k-1)s} P_k^2\, ds = \frac{1}{k-1} \left(e^{(k-1)T_k} - e^{(k-1)T_{k+1}} \right) P_k^2, \\
\widetilde{P}_k^3 &:= \int_{T_{k+1}}^{T_k} e^{(k-1)s} P_k^3\, ds = \frac{1}{k-1} \left(e^{(k-1)T_k} - e^{(k-1)T_{k+1}} \right) P_k^3.
\end{aligned}
$$

Then for $t^{\downarrow}(b) - t^{\uparrow}(b) \leq c < 0$,

$$
\int_{t^{\downarrow}(b)}^{t^{\uparrow}(b)+c} e^{(k-1)s} P_{kl}^1 (n(c + t^{\uparrow}(b)) - 1) e^{-(l-1)(c+t^{\uparrow}(b))}\, ds
$$

$$= (n(c+t^{\uparrow}(b))-1)e^{-(l-1)(c+t^{\uparrow}(b))}\left(\sum_{k=l}^{n(t^{\downarrow}(b))}\int_{T_{k+1}}^{T_k} e^{(k-1)s}P_{kl}^1\,ds - \int_{c+t^{\uparrow}(b)}^{T_l} e^{(l-1)s}\,ds\right)$$

$$= (n(c+t^{\uparrow}(b))-1)e^{-(l-1)(c+t^{\uparrow}(b))}\left(\sum_{k=l}^{n(t^{\downarrow}(b))}\widetilde{P}_{kl}^1 - \frac{1}{l-1}\left(e^{(l-1)T_l} - e^{(l-1)(c+t^{\uparrow}(b))}\right)\right)$$

$$= (n(c+t^{\uparrow}(b))-1)\left(e^{-(l-1)(c+t^{\uparrow}(b))}\cdot\sum_{k=l}^{n(t^{\downarrow}(b))}\widetilde{P}_{kl}^1 - \frac{1}{l-1}\left(e^{(l-1)(T_l-t^{\uparrow}(b)-c)}-1\right)\right).$$

For $0 < c \le T_2 - t^{\uparrow}(b)$,

$$\int_{t^{\downarrow}(b)}^{t^{\uparrow}(b)} e^{(k-1)s}P_k^2 n(c+t^{\uparrow}(b))e^{-\left[L_{\mathcal{T}}(t^{\uparrow}(b))-L_{\mathcal{T}}(c+t^{\uparrow}(b))\right]}\,ds$$

$$= n(c+t^{\uparrow}(b))e^{-\left[L_{\mathcal{T}}(t^{\uparrow}(b))-L_{\mathcal{T}}(c+t^{\uparrow}(b))\right]}\sum_{k=n(t^{\uparrow}(b))+1}^{n(t^{\downarrow}(b))}\int_{T_{k+1}}^{T_k} e^{(k-1)s}P_k^2\,ds$$

$$= n(c+t^{\uparrow}(b))e^{-\left[L_{\mathcal{T}}(t^{\uparrow}(b))-L_{\mathcal{T}}(c+t^{\uparrow}(b))\right]}\sum_{k=n(t^{\uparrow}(b))+1}^{n(t^{\downarrow}(b))}\widetilde{P}_k^2.$$

For $c > T_2 - t^{\uparrow}(b)$,

$$\int_{t^{\downarrow}(b)}^{t^{\uparrow}(b)} \frac{1}{2}e^{(k-1)s}P_k^3 e^{-(c+t^{\uparrow}(b)-T_2)/2}\,ds$$

$$= \frac{1}{2}e^{-(c+t^{\uparrow}(b)-T_2)/2}\sum_{k=n(t^{\uparrow}(b))+1}^{n(t^{\downarrow}(b))}\int_{T_{k+1}}^{T_k} e^{(k-1)s}P_k^3\,ds$$

$$= \frac{1}{2}e^{-(c+t^{\uparrow}(b)-T_2)/2}\sum_{k=n(t^{\uparrow}(b))+1}^{n(t^{\downarrow}(b))}\widetilde{P}_k^3.$$

Thus,

$$p_{\mathcal{T}}^{C}(c|C\neq 0) = \frac{1}{L_{\mathcal{T}}(0)}\sum_{\substack{b\in\mathcal{T}:\\ c\ge -\bar{t}(b)}}\xi_b(c),$$

where

$$\xi_b(c) = \begin{cases} (n(c+t^{\uparrow}(b))-1)\left(e^{-(l-1)(c+t^{\uparrow}(b))}\cdot\sum_{k=l}^{n(t^{\downarrow}(b))}\widetilde{P}_{kl}^1 - \frac{e^{(l-1)(T_l-t^{\uparrow}(b)-c)}-1}{l-1}\right) & -\bar{t}(b)\le c<0 \\ n(c+t^{\uparrow}(b))e^{-\left[L_{\mathcal{T}}(t^{\uparrow}(b))-L_{\mathcal{T}}(c+t^{\uparrow}(b))\right]}\cdot\sum_{k=n(t^{\uparrow}(b))+1}^{n(t^{\downarrow}(b))}\widetilde{P}_k^2 & 0<c\le T_2-t^{\uparrow}(b) \\ \frac{1}{2}e^{-(c+t^{\uparrow}(b)-T_2)/2}\cdot\sum_{k=n(t^{\uparrow}(b))+1}^{n(t^{\downarrow}(b))}\widetilde{P}_k^3 & c>T_2-t^{\uparrow}(b) \\ 0 & \text{otherwise.} \end{cases}$$

A8 Proof of Proposition 2.7

The probability of a negative change in tree height, conditional on the location and time of the recombination point, is

$$\mathbb{P}_{\mathcal{T}}(H<0|\mathcal{R}=(b,s)) = \begin{cases} \frac{k-1}{k} - e^{ks}\left(\exp\left(-kT_k - L_{\mathcal{T}}(T_k)\right) + \sum_{j=2}^{k} Q_{kj}\right) & b \in \mathcal{M} \\ 0 & b \notin \mathcal{M}. \end{cases}$$

Integrating over the recombination time gives

$$\mathbb{P}_{\mathcal{T}}(H<0|\mathcal{R}\in b) = \begin{cases} \frac{1}{\bar{t}(b)} \sum\limits_{k=2}^{n(t^{\downarrow}(b))} \left\{(k-1)\widetilde{Q}^1(k) - \widetilde{Q}^2(k,k,2,0,1) - \widetilde{Q}^3(k)\right\} & b \in \mathcal{M} \\ 0 & b \notin \mathcal{M}. \end{cases}$$

Summing over the branches and multiplying by the corresponding probability, the unconditional probability that the change in tree height is negative is thus

$$\mathbb{P}_{\mathcal{T}}(H<0) = \frac{1}{L_{\mathcal{T}}(0)} \sum_{b\in\mathcal{M}} \sum_{k=2}^{n(t^{\downarrow}(b))} \left\{(k-1)\widetilde{Q}^1(k) - \widetilde{Q}^2(k,k,2,0,1) - \widetilde{Q}^3(k)\right\}.$$

The probability of no change in tree height, conditional on the recombination point, is

$$\mathbb{P}_{\mathcal{T}}(H=0|\mathcal{R}=(b,s)) = \begin{cases} \frac{1}{k} + e^{ks}\sum_{j=2}^{k} Q_{kj} & b \in \mathcal{M} \\ 1 - e^{ks}\exp\left(-kT_k - L_{\mathcal{T}}(T_k)\right) & b \notin \mathcal{M}, \end{cases}$$

and

$$\mathbb{P}_{\mathcal{T}}(H=0|\mathcal{R}\in b) = \begin{cases} \frac{1}{\bar{t}(b)} \sum\limits_{k=2}^{n(t^{\downarrow}(b))} \left(\widetilde{Q}^1(k) + \widetilde{Q}^2(k,k,2,0,1)\right) & b \in \mathcal{M} \\ 1 - \frac{1}{\bar{t}(b)} \sum\limits_{k=n(t^{\uparrow}(b))+1}^{n(t^{\downarrow}(b))} \widetilde{Q}^3(k) & b \notin \mathcal{M}. \end{cases}$$

The probability of no change in tree height is thus

$$\mathbb{P}_{\mathcal{T}}(H=0) = \frac{1}{L_{\mathcal{T}}(0)} \left\{ \sum_{b\in\mathcal{M}} \sum_{k=2}^{n(t^{\downarrow}(b))} \left(\widetilde{Q}^1(k) + \widetilde{Q}^2(k,k,2,0,1)\right) + \sum_{b\notin\mathcal{M}} \left(\bar{t}(b) - \sum_{k=n(t^{\uparrow}(b))+1}^{n(t^{\downarrow}(b))} \widetilde{Q}^3(k)\right) \right\}.$$

Finally, the probability of a positive change in height, conditional on the recombination point, is

$$\mathbb{P}_{\mathcal{T}}(H>0|\mathcal{R}=(b,s)) = e^{ks}\exp\left(-kT_k - L_{\mathcal{T}}(T_k)\right),$$

and

$$\mathbb{P}_{\mathcal{T}}(H>0|\mathcal{R}\in b) = \frac{1}{\bar{t}(b)} \sum_{k=n(t^{\uparrow}(b))+1}^{n(t^{\downarrow}(b))} \widetilde{Q}^3(k).$$

The probability that the change in tree height is positive is

$$\mathbb{P}_{\mathcal{T}}(H>0) = \frac{1}{L_{\mathcal{T}}(0)} \sum_{b\in\mathcal{T}} \sum_{k=n(t^{\uparrow}(b))+1}^{n(t^{\downarrow}(b))} \widetilde{Q}^3(k).$$

A9 Proof of Theorem 2.2

Conditional on $\mathcal{T}$, the recombination point is chosen uniformly along the branches, so

$$\mathbb{P}(\mathcal{R} \in G) = \frac{L_G(0)}{L_{\mathcal{T}}(0)}, \quad \mathbb{P}(\mathcal{R} \notin G \cup g) = \frac{L_{\mathcal{T}} - L_G(0) - \bar{t}(g)}{L_{\mathcal{T}}(0)}.$$

Conditional on the recombination point being in G and letting the clade MRCA time be $t^{\downarrow}(g) = T_m$, the density of the recombination time is

$$p_S(s|\mathcal{R} \in G) = \begin{cases} \frac{n_G(s)}{L_G(0)} & \text{for } s \leq T_m \\ 0 & \text{otherwise,} \end{cases}$$

and similarly

$$p_S(s|\mathcal{R} \notin G \cup g) = \begin{cases} \frac{n(s) - n_{G\cup g}(s)}{L_{\mathcal{T}}(0) - L_G(0) - \bar{t}(g)} & \text{for } s \leq T_2 \\ 0 & \text{otherwise.} \end{cases}$$

First conditioning on the recombination time,

$$\begin{aligned}
\mathbb{P}(\mathcal{C} \in G \cup g|\mathcal{R} = (G, s)) &= \int_s^{t^{\uparrow}(g)} \frac{n_{G\cup g}(u)}{n(u)} n(u) \exp\left(-\int_s^u n(t)dt\right) du \\
&= \int_s^{T_k} n_{G\cup g}(T_{k+1}) \exp\left(-\int_s^u k dt\right) du \\
&+ \sum_{j=n(t^{\uparrow}(g))+1}^{k-1} \int_{T_{j+1}}^{T_j} n_{G\cup g}(T_{j+1}) \exp\left(-\int_s^u n(t)dt\right) du \\
&= \frac{1}{k} n_{G\cup g}(T_{k+1}) + e^{ks} \sum_{j=n(t^{\uparrow}(g))+1}^{k} n_{G\cup g}(T_{j+1}) Q_{kj},
\end{aligned}$$

and so

$$\begin{aligned}
\mathbb{P}(\mathcal{C} \in G \cup g|\mathcal{R} \in G) &= \frac{1}{L_G(0)} \sum_{k=n(t^{\downarrow}(g))+1}^{n} \int_{T_{k+1}}^{T_k} n_G(T_{k+1}) \mathcal{P}(\mathcal{C} \in G \cup g|\mathcal{R} \in G, \mathcal{R} = (\cdot, s)) ds \\
&= \frac{1}{L_G(0)} \sum_{k=n(t^{\downarrow}(g))+1}^{n} \left[n_{G\cup g}(T_{k+1}) \widetilde{Q}^1(k) + \widetilde{Q}^4(k, G \cup g) \right] n_G(T_{k+1}),
\end{aligned}$$

with

$$\widetilde{Q}^4(k, A) = \frac{1}{k} \left(e^{kT_k} - e^{kT_{k+1}} \right) \sum_{j=n(t^{\uparrow}(A))+1}^{k} n_A(T_{j+1}) Q_{kj}.$$

Similarly,

$$\begin{aligned}
\mathbb{P}(\mathcal{C} \in G|\mathcal{R} \notin G \cup g, \mathcal{R} = (\cdot, s)) &= \int_s^{t^{\downarrow}(g)} \frac{n_G(u)}{n(u)} n(u) r(u) du \\
&= \frac{1}{k} n_G(T_{k+1}) + \sum_{j=n(t^{\downarrow}(g))+1}^{k} e^{ks} n_G(T_{j+1}) Q_{kj},
\end{aligned}$$

and

$$\mathbb{P}(\mathcal{C} \in G|\mathcal{R} \notin G \cup g) = \frac{1}{L_{\mathcal{T}}(0) - L_G(0) - \bar{t}(g)} \sum_{k=n(t^{\downarrow}(g))+1}^{n} \int_{T_{k+1}}^{T_k} (k - n_{G\cup g}(T_{k+1}))$$

$$\cdot\, \mathbb{P}(\mathcal{C} \in G | \mathcal{R} \notin G \cup g, \mathcal{R} = (\cdot, s))ds$$

$$= \frac{1}{L_{\mathcal{T}}(0) - L_G(0) - \bar{t}(g)} \sum_{k=n(t^{\downarrow}(g))+1}^{n} (k - n_{G\cup g}(T_{k+1})) \left[n_G(T_{k+1})\widetilde{Q}^1(k) + \widetilde{Q}^4(k, G) \right].$$

A10 Conditioning on branch having at least one mutation

Suppose that mutations occur as a Poisson process along the branches with constant rate θ, and the recombination rate at position x is $\rho(x)/2$. Conditional on the left endpoint of the given branch $d^{\leftarrow}$, let D be its right endpoint, which using (2.23) has the density

$$p^D_{\mathcal{T}}(\delta | d^{\leftarrow}) = \lambda(d^{\leftarrow}) \exp\left(- \int_{d^{\leftarrow}}^{\delta} \lambda(x)dx \right),$$

where

$$\lambda(x) = \mathbb{P}(b \text{ topologically disrupted}) \cdot L_{\mathcal{T}}(0) \cdot \frac{\rho(x)}{2}.$$

Let $\bar{t}$ be the time-length and M the number of mutations on the branch. Then the conditional distribution of D is given by

$$P^D_{\mathcal{T}}(d^{\rightarrow} | M > 0, d^{\leftarrow}) = \frac{\int_{d^{\leftarrow}}^{d^{\rightarrow}} \mathbb{P}(M > 0 | \delta, d^{\leftarrow}) p^D_{\mathcal{T}}(\delta | d^{\leftarrow}) d\delta}{\int_{d^{\leftarrow}}^{\infty} \mathbb{P}(M > 0 | \delta, d^{\leftarrow}) p^D_{\mathcal{T}}(\delta | d^{\leftarrow}) d\delta},$$

with

$$\mathbb{P}(M > 0 | \delta, d^{\leftarrow}) = 1 - \exp(-\theta \bar{t}(\delta - d^{\leftarrow})).$$

We have

$$\mathbb{P}(M > 0 | \delta, d^{\leftarrow}) p^D_{\mathcal{T}}(\delta | d^{\leftarrow}) = \lambda(d^{\leftarrow}) \exp\left(\int_{d^{\leftarrow}}^{\delta} \lambda(x)dx \right) - \lambda(d^{\leftarrow}) \exp\left(-\theta \bar{t}(\delta - d^{\leftarrow}) - \int_{d^{\leftarrow}}^{\delta} \lambda(x)dx \right).$$

Assuming that the recombination map is piecewise constant, we split the part of the genome to the right of $d^{\leftarrow}$ into portions where the recombination rate is constant between the (ordered) breakpoints $d^{\leftarrow} =: w_0 < w_1 < w_2 < \ldots < w_k < \ldots$, adding an extra breakpoint $w_k := d^{\rightarrow}$. Then we can write

$$\begin{aligned}
\int_{d^{\leftarrow}}^{\infty} \mathbb{P}(M > 0 | \delta, d^{\leftarrow}) p^D_{\mathcal{T}}(\delta | d^{\leftarrow}) d\delta &= 1 - \lambda(d^{\leftarrow}) \int_{d^{\leftarrow}}^{\infty} \exp\left(-\theta \bar{t}(\delta - d^{\leftarrow}) - \int_{d^{\leftarrow}}^{\delta} \lambda(x)dx \right) d\delta \\
&= 1 - \lambda(d^{\leftarrow}) \sum_{i=0}^{\infty} \exp\left(- \int_{d^{\leftarrow}}^{w_i} \lambda(x)dx \right) \int_{w_i}^{w_{i+1}} \exp\left(-\theta \bar{t}(\delta - d^{\leftarrow}) - \int_{w_i}^{\delta} \lambda(x)dx \right) d\delta \\
&= 1 - \sum_{i=0}^{\infty} S_i,
\end{aligned}$$

where, by integrating,

$$S_i = \frac{\lambda(d^{\leftarrow}) \exp\left(- \int_{d^{\leftarrow}}^{w_i} \lambda(x)dx \right) \left[\exp(-\theta \bar{t}(w_i - d^{\leftarrow})) - \exp(-\theta \bar{t}(w_{i+1} - d^{\leftarrow}) - \lambda(w_i)(w_{i+1} - w_i)) \right]}{\theta \bar{t} + \lambda(w_i)}.$$

Similarly,

$$\int_{d^{\leftarrow}}^{d^{\rightarrow}} \mathbb{P}(M > 0 | \delta, d^{\leftarrow}) p^D_{\mathcal{T}}(\delta | d^{\leftarrow}) d\delta = 1 - \exp\left(-\Lambda(b) \right) - \sum_{i=0}^{k-1} S_i,$$

where

$$\Lambda(b) = \int_{d^{\leftarrow}}^{d^{\rightarrow}} \lambda(x)dx.$$

Combining, we have

$$P_{\mathcal{T}}^{D}(d^{\rightarrow}|M>0,d^{\leftarrow}) = \frac{1-\exp(-\Lambda(b)) - \sum_{j=0}^{k-1} S_j}{1-\sum_{j=0}^{\infty} S_j}. \tag{A3}$$

Note that in the limit $\theta \to \infty$, this reduces to $1-\exp(-\Lambda(b))$, as expected (since this effectively removes the conditioning).

For the case where the recombination rate is constant along the genome, with $\rho(x)=\rho$ and $\lambda(x)=\lambda$, we obtain

$$P_{\mathcal{T}}^{D}(d^{\rightarrow}|M>0,d^{\leftarrow}) = 1 - \frac{\theta\bar{t}+\lambda}{\theta\bar{t}}\exp(-\lambda(d^{\rightarrow}-d^{\leftarrow})) + \frac{\lambda}{\theta\bar{t}}\exp(-(\lambda+\theta\bar{t})(d^{\rightarrow}-d^{\leftarrow})).$$

The p-values for each branch of a simulated ARG can now be computed by evaluating the cdf (A3) for the given values of $d^{\leftarrow}$ and $d^{\rightarrow}$, and the Q-Q plot can again be constructed by plotting these against the corresponding quantiles of the uniform distribution. However, this potentially requires summing over a large number of increments of the recombination map. Instead, we propose to approximate (A3) by

$$\widetilde{P}_{\mathcal{T}}^{D}(d^{\rightarrow}|M>0,d^{\leftarrow}) := 1 - \frac{\theta\bar{t}+\Lambda(b)}{\theta\bar{t}}\exp(-\Lambda(b)) + \frac{\Lambda(b)}{\theta\bar{t}}\exp(-(\Lambda(b)+\theta\bar{t}d)).$$

We find this to be a very close match to the exact distribution based on simulations with human-like parameters, while being very fast to compute.

A11 Change in total branch length and the probability that a branch is disrupted, following a recombination event

For a given branch b_i that exists in local trees $\mathcal{T}_{(1)}, \ldots, \mathcal{T}_{(k)}$, let

$$f(b_i, l) := P_{\mathcal{T}_{(l)}}(b_i \text{ disrupted}) \cdot \mathcal{L}_{\mathcal{T}_{(l)}}(0),$$

for $1 \leq l \leq k$. Let

$$H_1^l(b_i) := f(b_i, l+1)/f(b_i, l),$$

for $1 \leq l \leq k-1$, and

$$H_2(b_i) = \max_l f(b_i, l)/\min_l f(b_i, l).$$

These quantities are calculated for a uniform random sample of 1 000 branches from an ARG simulated with stdpopsim ($n = 100$, Chr21 with HapMapII_GRCh38 recombination map, $N_e = 10\,000$, constant population size model); the corresponding histograms are shown in Figure A2. These suggest that the quantity $P_{\mathcal{T}}(b_i \text{ disrupted}) \cdot \mathcal{L}_{\mathcal{T}}(0)$ stays relatively conserved following each recombination event, even if individually $P_{\mathcal{T}}(b_i \text{ disrupted})$ and $\mathcal{L}_{\mathcal{T}}(0)$ may vary more significantly. This is similar to the findings of Deng et al. (2021, Figure 5).

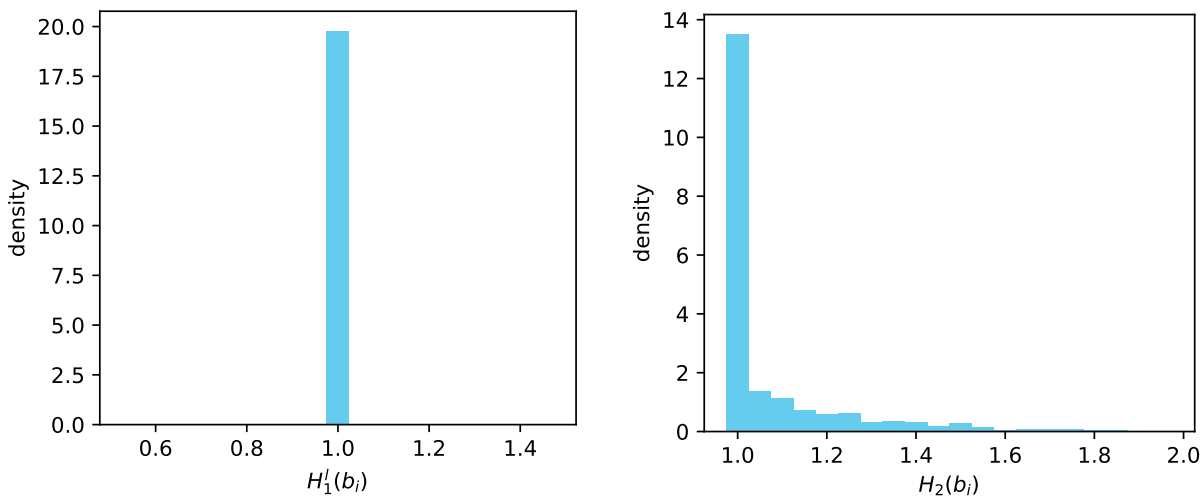

Figure A2: Histograms of $H_1^l(b_i)$ and $H_2(b_i)$ for a simulated ARG.

A12 Histograms corresponding to Q-Q plots

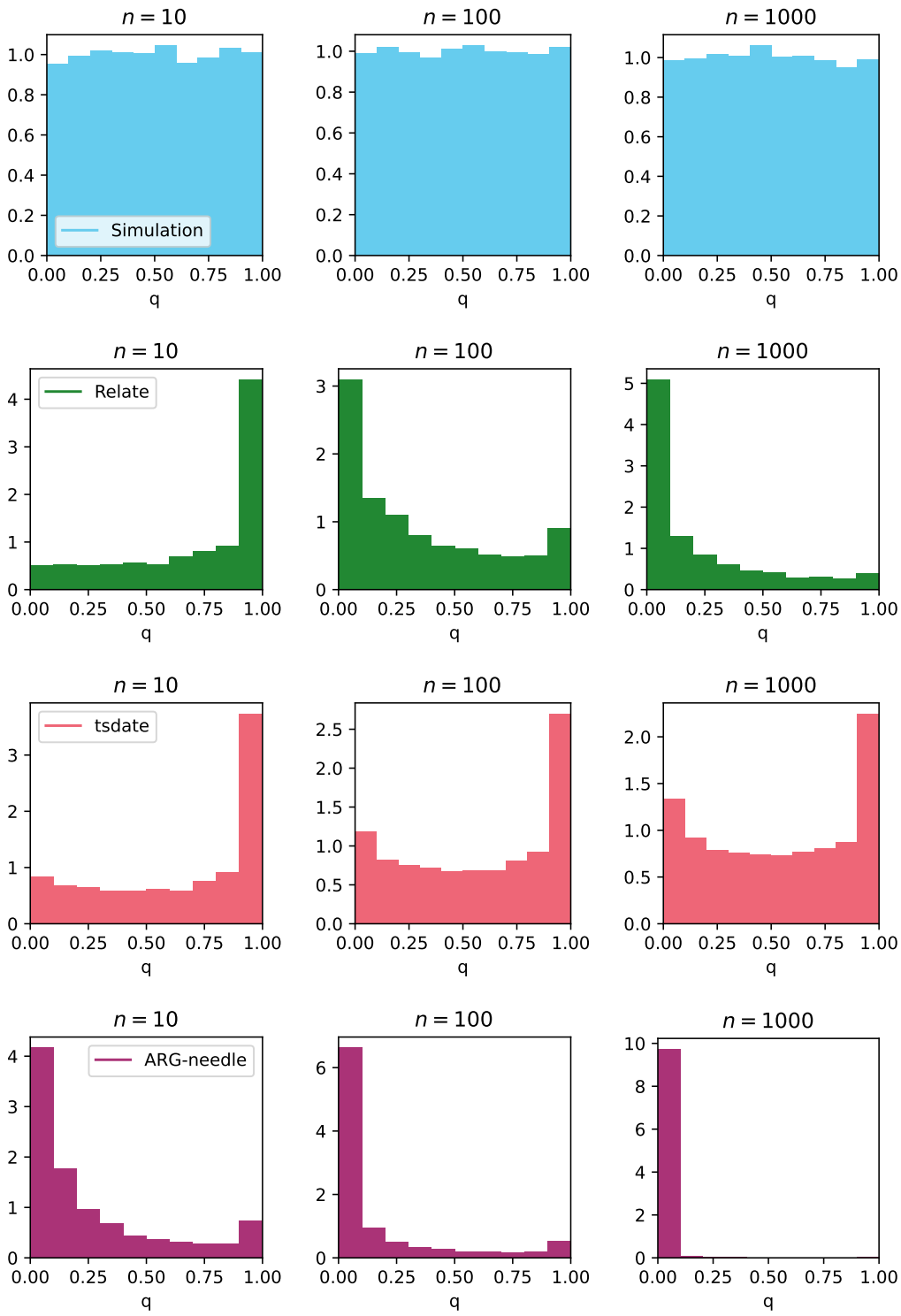

Figure A3: Histograms corresponding to Q-Q plots in Figure 9.

A13 Q-Q plots split by clade size and depth

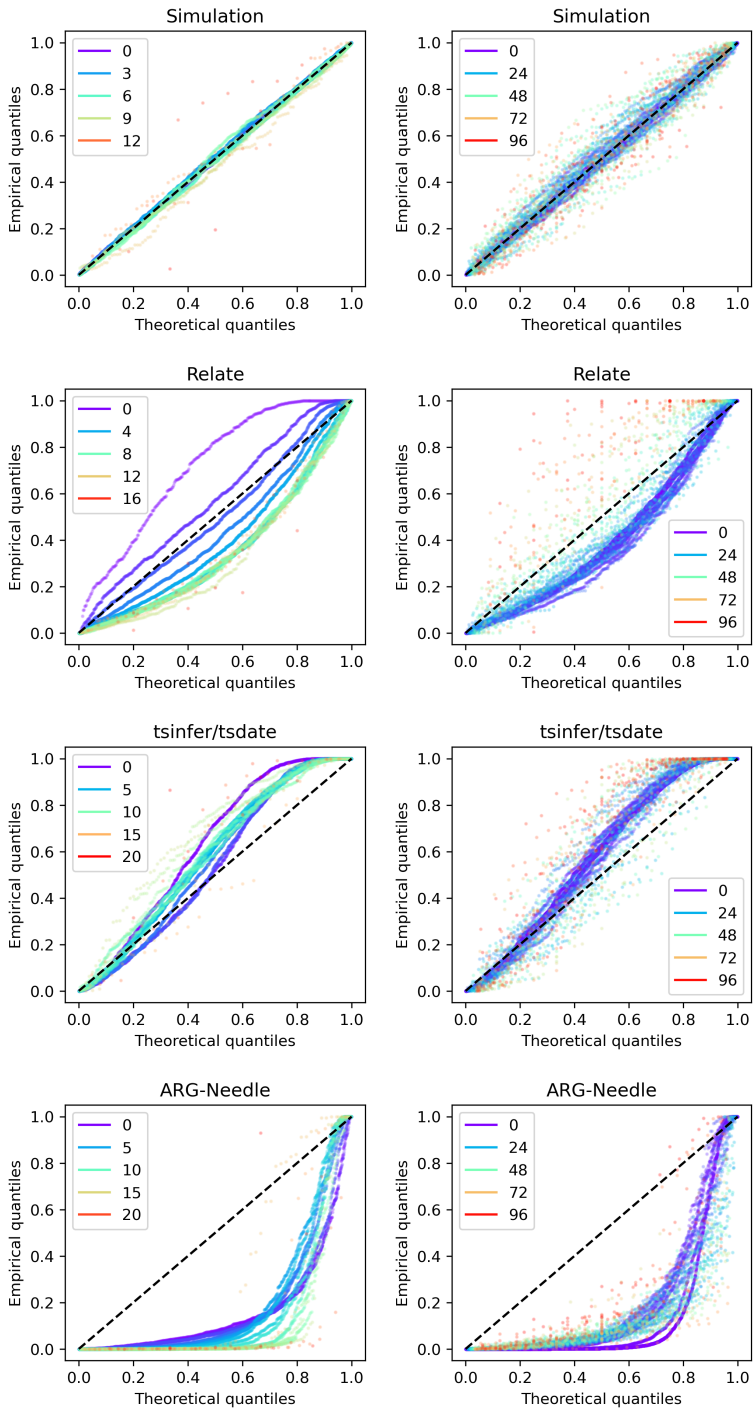

Figure A4: Q-Q plots using (2.22) computed from an ARG simulated using dataset 1 parameters under the SMC' with $n = 100$. Left panels: split by depth of the branch (defined as the number of branches to the root of the tree); right panels: split by clade size (defined as the number of samples subtended by the branch). Dashed line: diagonal from $(0, 0)$ to $(1, 1)$. For the simulated ARGs, none of the corresponding K–S p-values for each group are significant using a 0.05 significance threshold.

A14 Histograms of observed branch duration for simulated and reconstructed ARGs

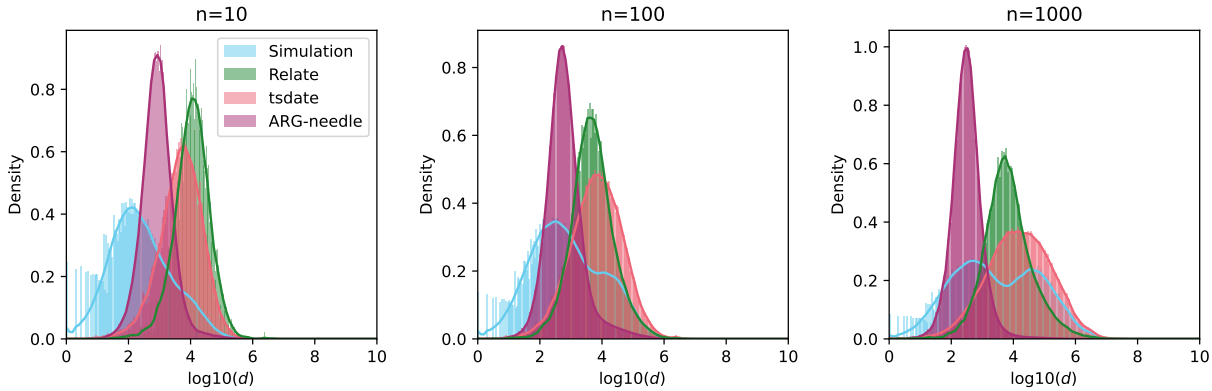

Figure A5: Histograms of (observed) branch duration, calculated as $d(b) = d^{\rightarrow}(b) - d^{\leftarrow}(b)$ for each branch b in simulated and reconstructed ARGs (same as those in Figure 9). Note log scale on the x-axis.

A15 Histograms of expected number of mutations per branch for simulated and reconstructed ARGs

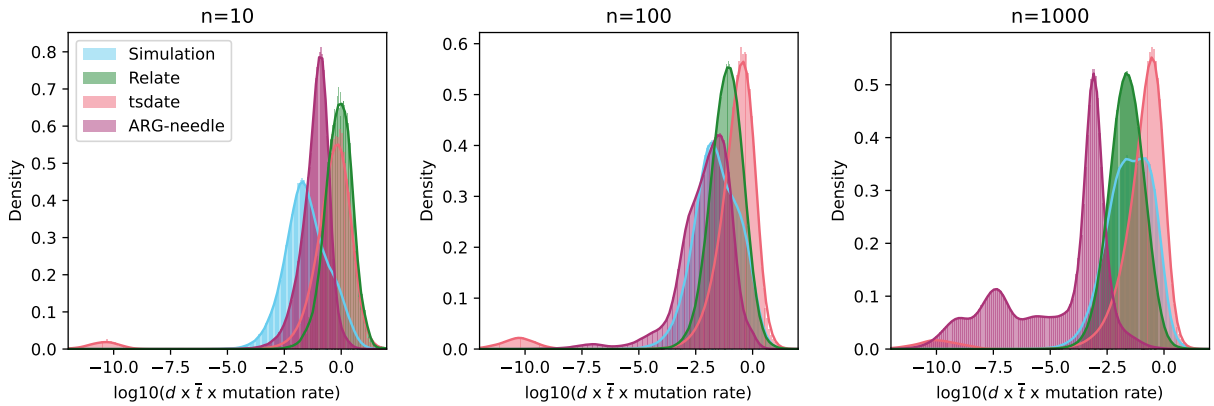

Figure A6: Histograms of (observed) expected number of mutations per branch, calculated as $(d^{\rightarrow}(b) - d^{\leftarrow}(b)) \cdot \bar{t}(b) \cdot \mu$ for each branch b in simulated and reconstructed ARGs (same as those in Figure 9). Note log scale on the x-axis.

A16 Q-Q and p-value plots (simulated ARG without inversion)

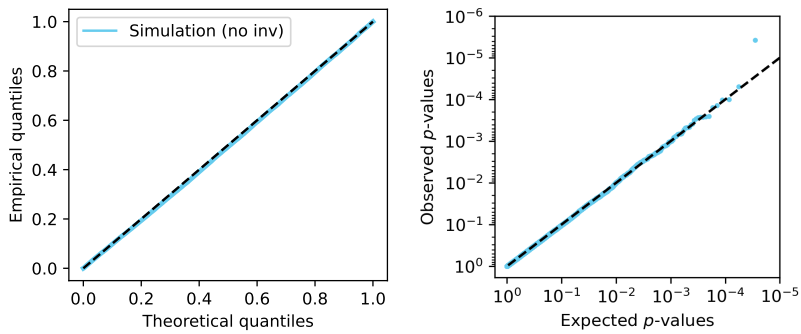

Figure A7: Q-Q plot (left panel) and p-value plot (right panel) for ARG simulated using SLiM (without inversions and otherwise same parameters as in Section 3.6.1). No clades have p-values below the Bonferroni-corrected significance threshold.

A17 Q-Q and p-value plots (ARGs reconstructed using Relate)

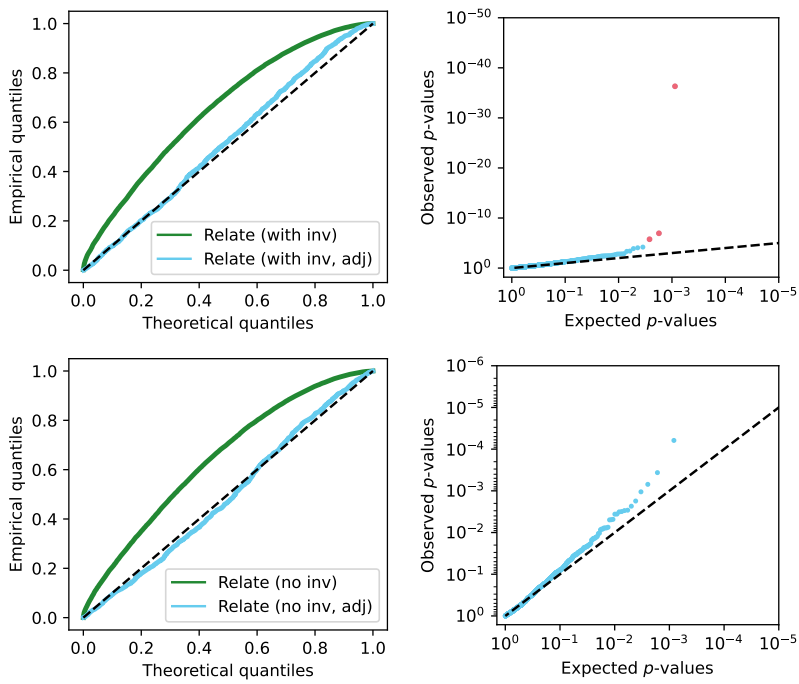

Figure A8: Top row: Q-Q plot (left panel) and p-value plot (right panel) for ARG reconstructed using Relate from data simulated using SLiM with one inversion under balancing selection (as described in Section 3.6.1). Blue (resp. green) points show values calculated after (resp. before) applying the adjustments described in Section 3.6.2; red points correspond to clades with p-value below the Bonferroni-corrected significance threshold. Bottom row: same using SLiM simulation without inversions (and otherwise same parameters).