On the Effect of Intralocus Recombination on Triplet-Based Species Tree Estimation*

Max Hill[†] Sebastien Roch[‡]

November 5, 2021

Abstract

We consider species tree estimation from multiple loci subject to intralocus recombination. We focus on R^* , a summary coalescent-based methods using rooted triplets. We demonstrate analytically that intralocus recombination gives rise to an inconsistency zone, in which correct inference is not assured even in the limit of infinite amount of data. In addition, we validate and characterize this inconsistency zone through a simulation study that suggests that differential rates of recombination between closely related taxa can amplify the effect of incomplete lineage sorting and contribute to inconsistency.

*Acknowledgments: MH was supported by supported by NSF grants DMS-1902892 (to SR) and DMS-2023239 (TRIPODS Phase II). SR was supported by NSF grants DMS-1902892 and DMS-2023239 (TRIPODS Phase II). MH and SR are grateful for the feedback from Cecile Ane and her lab members as well as Claudia Solis-Lemus.

[†]Department of Mathematics, University of Wisconsin–Madison.

[‡]Department of Mathematics, University of Wisconsin–Madison.

1 Introduction

Species tree estimation from genomic data is complicated by various biological phenomena, among them hybridization, horizontal gene transfer, gene duplication and loss, and incomplete lineage sorting (ILS) [SS20]. ILS in particular is a source of phylogenetic conflict, in which gene trees and species tree do not share the same topology, especially for species trees with short internal branches [SS20]. Of some interest is the existence of an anomaly zone for species trees, in which the most probable topology in the gene tree distribution differs from the topology of the species tree [DR06, DR09, Deg13] (see also [RELY20, BH20] for a more recent discussion of these and other relevant issues).

The existence of an anomaly zone has served as an impetus for the development of summary coalescent-based methods, such as R^* , MP-EST, BUCKy, ASTRAL, and others [DDBR09, LYE10, LKDA10, MRB⁺14]. Some of these methods are based on the facts that rooted triples and unrooted quartets are special cases in which no anomaly zone exists [Deg13, KYT20] and also provide sufficient information to reconstruct the full phylogeny [SS⁺03, Ste16]. Provided that the gene trees are estimated without error, such methods can provide statistically consistent methods of estimating species tree topology [War17].

A common assumption of coalescent-based models based on the multispecies coalescent (MSC) [RY03, RELY20] is that recombination occurs between genes (or loci)—so that gene trees may be assumed unlinked or statistically independent—but that *intralocus recombination* (i.e., recombination occurring *within* gene sequences), does not occur [EXJ⁺16, BH20]. The significance of the latter assumption—that is, the impact of intralocus recombination on phylogenetic inference—is a matter of present interest [ZY21, BH20] and much debate about its significance when unaccounted for [LK12, SG18, EXJ⁺16]. One justification for assuming no intralocus recombination is that within-gene recombination may break gene function [SS20].

An influential simulation study argued that even high levels of intralocus recombination do not present a significant challenge for species tree estimation relative to other biological phenomena [LK12]. On the other hand, the authors of [SG16] suggest the absence of intralocus recombination may be an unreasonable assumption in real data, such as protein-coding genes in eukaryotes [MLH19, BH20], and particularly in the case of species phylogenies with many taxa [SG18]. In particular, the potential for intralocus recombination to distort gene tree frequencies has been recognized as a challenge to summary coalescent-based methods, and [LK12] has been critiqued for its focus on shallow divergences and limitation to a low number of loci and taxa [SG18].

In this paper we take an analytical approach to investigate the effect of intralocus recombination. We prove that intralocus recombination has the potential to confound R^* , a summary coalescent-based methods based on inferring rooted triples. That is, we show that correct inference of rooted triplets cannot be guaranteed in the presence of intralocus recombination, assuming a distance-based approach is used for gene tree reconstruction. We then present a simulation study which characterizes the “inconsistency zone”, i.e. the regime of parameters for S in which rooted triple inference does not converge to S as $m \rightarrow \infty$. We find that the effect arises when differential rates of recombination are exhibited between closely-related taxa. We also discuss related implications for quartet-based methods.

1.1 Key Definitions

A *species phylogeny* $S = (V, E; r, \bar{\rho}, \bar{\tau}, \bar{\theta})$ is a directed binary tree with vertex set V , edge set E , root $r \in V$, and n labeled leaves $L = [n]$, such that each edge $e \in E$ is associated with a length $\tau_e \in (0, \infty)$, expressed in coalescent units, a recombination rate $\rho_e \in [0, \infty)$, and a mutation rate $\theta_e \in [0, \infty)$. It is assumed that there exists an ancestral population common to all leaves of S , i.e., a population above the root, with respective mutation and recombination parameters. In this paper, mutation rates are assumed to be per site per coalescent unit (a coalescent unit being $2N_e$ generations for diploid organisms, where N_e is the effective population size); the recombination rates are per locus per coalescent unit.

The general question considered here is how to reconstruct the topology of the species phylogeny from gene sequence data sampled from its leaves. This sequence data takes the form of multiple sequence alignments; a *multiple sequence alignment* (MSA) is an $n \times k$ matrix M whose entries are letters in the nucleotide alphabet $\{A, T, C, G\}$ such that entries in the same column are assumed to share a common ancestor. The phylogenetic reconstruction problem in this paper is to recover the topology of S from m independent samples of M .

A *rooted triple* is a rooted binary phylogenetic tree with label set of size three; we use the notation $XY|Z$

(or equivalently $YX|Z$) to denote a rooted triple with leaves X, Y, Z having the property that the path from X to Y does not intersect the path from Z to the root [SS⁺03]. The term *species triplet* refers to a restriction of S to three of its leaves. A rooted triple $XY|Z$ is said to be *uniquely favored* if it appears in more gene samples than either of the other two rooted triples $XZ|Y$ or $YZ|X$.

1.2 Inference Methods

This paper considers *Majority-Rule Rooted Triple*, or R^* , a consensus-based pipeline for species tree estimation. R^* utilizes the fact that the full topology of S is uniquely determined by, and hence can be recovered from, its rooted triples [Ste16]. The R^* pipeline has three steps: first, for each gene, infer a rooted triple for each triplet of leaves $X, Y, Z \in L$. Second, make a list of uniquely favored triples from the m sampled genes. Finally, construct the most-resolved topology containing only uniquely favored triples. When gene trees are drawn independently according to the MSC, it holds that for every set of three taxa, the most probable rooted triple in the gene tree distribution matches the rooted triple obtained by restricting the species tree S to that set of three taxa; for this reason, the topology of the R^* consensus tree converges to that of S [DDBR09].

Since we are interested in the inference of the species-tree topology from *sequence data*, we consider a distance-based approach in which a species triplet with leaves X, Y, Z is inferred to have topology $XY|Z$ if

$$\delta_{XY} < \delta_{XZ} \wedge \delta_{YZ}. \quad (1)$$

where $\delta_{XY} = \delta_{XY}(M_k)$ is the number of mismatching nucleotides between sequences $\mathbf{s}_X$ and $\mathbf{s}_Y$ ($X, Y \in L$). We refer to this inference procedure as **R^* with sequence distances**.

1.3 Multispecies Coalescent with Recombination

The model considered here, which we term the *Multispecies Coalescent with Intralocus Recombination*, or MSCR, uses the ancestral recombination graph (ARG) model from [GM97] (see also [Are13]) within the framework of the multi-species coalescent (MSC) [RY03, RELY20, DR09]. In the single-population ARG [GM97], ancestors are represented by edges in the graph (see Figure 1a), and the number N of ancestors, or *gene lineages*, at time t is a bottom-up birth-death process in which births (recombination events) occur at rate ρN and deaths (coalescent events) occur at rate $N(N-1)/2$. When a coalescent event happens, two edges are chosen at random and merged into one. When recombination occurs, a randomly chosen lineage splits into two parent lineages. Each recombination vertex is labeled by a number b , chosen uniformly on $[0, 1]$; this number is the *breakpoint* of the recombination.

The single-population ARG can be extended to multiple species in a manner similar to the MSC: at time $t = 0$, each leaf of S begins with a single lineage, and these lineages evolve in a bottom-up manner according to the ARG process along each edge of a fixed species tree (see Figure 1b). If $\mathcal{G}$ is a rooted directed graph with edge lengths and leaf and breakpoint labels obtained in this manner, then we say that $\mathcal{G}$ is **generated according to the MSCR process on S** . In this scheme, the locus is modeled by the unit interval, and for each site $x \in [0, 1]$, a *marginal gene tree* $\mathcal{T}(x)$ can be obtained by tracing upward along the edges of $\mathcal{G}$ starting from the leaves; if a recombination vertex is reached with breakpoint b , take the left path if $x \leq b$ and the right path if $x > b$. This yields a rooted edge-weighted binary trees; a simple example is shown in Figure 2. The set of marginal gene trees $\mathcal{M} := \{\mathcal{T}(x) : 0 \leq x \leq 1\}$ is almost surely finite [GM97]. For each $T_g \in \mathcal{M}$, define $I(T_g) = \{x \in [0, 1] : \mathcal{T}(x) = T_g\}$, and define $w_g = |I(T_g)|$, where $|\cdot|$ denotes Lebesgue measure. In words, $I(T_g)$ is the identical-by-descent segment of the locus having genealogy T_g , and w_g is the proportion of sites with genealogy T_g .

Measuring time in coalescent units, this paper assumes that the per-site mutation rate is given by a fixed number $\theta > 0$ which does not vary on S . For each $x \in [0, 1]$, site x evolves independently according to the Jukes-Cantor process [JC⁺69, Ste16] on the tree $\mathcal{T}(x)$. A somewhat more general description of this algorithm can be found in [DMNR17].

Thus, to model the evolution of a genetic locus consisting of k sites in which recombination breakpoints are distributed uniformly between them, a two-step process is followed. First, a multispecies ARG $\mathcal{G}$ is generated according to the MSCR process on S , from which a marginal gene tree $\mathcal{T}(x)$ is obtained for each $x \in [0, 1]$. Second, for each $x \in [0, 1]$ the Jukes-Cantor process is run with input tree $\mathcal{T}(x)$ in order to

Figure 1: Two depictions of an ARG, in a single population (left) and in the multispecies case (right). In Figure 1a, two lineages enter the population at time 0 and three exit at time t_{end} . Coalescent events occurred at times t_2 and t_5 . Recombinations with breakpoints b_1, b_2, b_3 occurred at times t_1, t_3 , and t_4 . Figure 1b depicts a multispecies ARG (in blue) within the four-taxon tree S considered in Corollary 1.

generate a nucleotide $\mathcal{N}(i, x) \in \{A, T, C, G\}$ for each $i \in L$. The MSA M_k is then defined as the $n \times k$ random matrix with rows $\mathbf{s}_1, \dots, \mathbf{s}_n$ where for each $X \in [n]$, $\mathbf{s}_X = (s_X(1), \dots, s_X(k))$ where $s_X(j) = \mathcal{N}(X, \frac{j}{k-1})$, $j = 0, 1, \dots, k-1$. In this case, we say that M_k is **generated according to the MSCR-JC(k) process on S** . The phylogenetic reconstruction problem considered here can then be stated precisely as follows:

Problem: Let S be a species phylogeny with leaf labels $L = [n]$. Fix $k \geq 2$. Given m independent samples $M_k^{(1)}, \dots, M_k^{(m)}$, each generated according to the MSCR-JC(k) process on S , recover the topology of S .

1.4 Estimating Sequence Distances

Let $\mathcal{G}$ be generated according to the MSCR process on S , and $\mathcal{M}$ the corresponding set of marginal gene trees. Given a marginal gene tree $T_g \in \mathcal{M}$, let $d_{XY}^{T_g}$ be the *evolutionary distance* between leaves X and Y on T_g , defined as the expected number of mutations per site along the unique path between X and Y . It follows from the assumptions about the mutation process that $d_{XY}^{T_g} = 2\theta t$, where t is the time of the most recent common ancestor of X and Y on T_g . For example in Figure 2, $d_{AB}^{T_1} = 2\theta t_4$ and $d_{AB}^{T_2} = 2\theta t_2$. Define the *breakpoint-weighted uncorrected distance* by

$$\Delta_{XY} := \frac{3}{4} \sum_{T_g \in \mathcal{M}} w_g \left(1 - e^{-\frac{4}{3} d_{XY}^{T_g}} \right). \quad (2)$$

This formula, due to [Wan17], generalizes the uncorrected Jukes-Cantor distance to the setting of intralocus recombination; if no intralocus recombination occurs, then the right-hand side has only a single summand and reduces to the inverse of the Jukes-Cantor distance correction formula for a single non-recombining locus.

Our first lemma shows that δ_{XY} can be approximated by $k\Delta_{XY}$ when k is large.

Lemma 1. *If M_k is generated according to the MSCR-JC(k) process on S then for all $X, Y \in L(S)$, conditioned on $\mathcal{G}$, $\delta_{XY}(M_k) = k\Delta_{XY} + o(k)$ almost surely as $k \rightarrow \infty$.*

Figure 2: On the left, an ancestral recombination graph (in blue) is shown within a 3-taxon tree S (in black). The times of coalescence and recombination events are labeled $t_1, \dots, t_4$ on the time axis, and the breakpoint associated with the recombination event is labeled $b \in [0, 1]$. On the right, the corresponding marginal gene trees T_1 and T_2 are shown.

See Appendix A for a proof.

2 Inconsistency of R^*

2.1 Statement and Overview

The main result is the following:

Theorem 1. *For k sufficiently large, R^* using sequence distances is not statistically consistent under the MSCR-JC(k) model. That is, there exists a species phylogeny S such that the topology of the output of R^* using sequence distances does not converge in probability to the topology of the species tree.*

We discuss implications for quartet-based methods in Appendix C.

To prove Theorem 1, it suffices to consider a species tree S with $L = \{A, B, C\}$ and topology $AB|C$. Denote edges of S , or *populations*, by the letters A, B, C, AB , and ABC as depicted in Figure 3 where A, B, C correspond to the leaf populations, AB is the parent edge of A and B , and ABC is edge extending above the root. The key idea is to allow recombination only in population A . In order to keep the analysis tractable, the recombination rate and length of edge A are chosen so that with high probability the number of recombinations is 0 or 1, so that the number of lineages on the ARG exiting population A (backwards-in-time) is either one or two. By choosing the internal branch length τ_{AB} sufficiently small, ILS occurs along that edge with high probability, so that all coalescent events on the ancestral recombination graph occur in the root population ABC . In that case, as long as the mutation rate is not too large, we show that, on the event R_1C_0 (see Figure 3), taxa B and C are more likely to be inferred as more closely related than taxa A and B , so that R^* converges to the wrong topology $BC|A$ as the number m of samples grows.

The mutation rate θ is assumed to be the same in all populations. The vector of recombination rates $\bar{\rho}$ is defined by setting $\rho_A = \rho > 0$ and $\rho_X = 0$ for all $X \neq A$. Assume S to be ultrametric. The populations A and B have length $\tau_A = \tau_B > 0$, the internal population AB has length τ_{AB} , the age of the root t_{root} is given by $t_{\text{root}} = \tau_A + \tau_{AB} = \tau_C$. For now assume that $\tau_{AB} > 0$ and $\tau_A > 0$; their precise values will be determined later in the proof.

Let M_k be generated according to the MSCR-JC(k) process on S , and let $E_{XY|Z}$ be the event that the rooted triple inferred from M_k using (1) is $XY|Z$. The following lemma implies that to prove Theorem 1, it will suffice to prove

$$\mathbb{P}[E_{YZ|X}] > \mathbb{P}[E_{XY|Z}]. \quad (3)$$

The *consistency zone* for R^* with sequence distances under the MSCR-JC(k) model is the set of species phylogenies S such that the topology of the R^* consensus tree converges in probability to the topology of S as $m \rightarrow \infty$.

Lemma 2. *A necessary and sufficient condition for S to lie in the consistency zone for R^* with sequence distances under the MSCR-JC(k) model is that for all $XY|Z \in \mathcal{R}(S)$,*

$$\mathbb{P}[E_{XY|Z}] > \mathbb{P}[E_{XZ|Y}] \vee \mathbb{P}[E_{YZ|X}] \quad (4)$$

Here $\mathcal{R}(S) = \{S|J : J \subseteq L, |J| = 3, \text{ and } S|J \text{ is binary}\}$ is the set of restricted rooted triples of S (see [SS⁺03]).

See Appendix A for a proof.

By Lemma 1, with probability one, an ancestral recombination graph $\mathcal{G}$ generated according to the MSCR process has the property that sequences of increasing length k generated on it by the Jukes-Cantor process satisfy the almost sure limit $\frac{1}{k}\delta_{XY}(M_k) \rightarrow \Delta_{XY}$ as $k \rightarrow \infty$. Since almost sure convergence implies convergence in distribution, it holds that under the joint process which combines both genealogical and mutational processes, $\frac{1}{k}\delta_{XY}(M_k) \Rightarrow \Delta_{XY}$ as $k \rightarrow \infty$ for all $X, Y \in L(S)$. Therefore, since the distribution function of Δ_{XY} is continuous, $\mathbb{P}[E_{XY|Z}] \rightarrow \mathbb{P}[E]$ and $\mathbb{P}[E_{YZ|X}] \rightarrow \mathbb{P}[F]$ as $k \rightarrow \infty$, where $E := [\Delta_{AB} < \Delta_{AC} \wedge \Delta_{BC}]$ and $F := [\Delta_{BC} < \Delta_{AB} \wedge \Delta_{AC}]$. Therefore inequality (3) will hold for sufficiently large k provided that

$$\mathbb{P}[F] > \mathbb{P}[E]. \quad (5)$$

We detail the proof next.

2.2 Key Lemmas

Since recombination occurs only in population A , the number of recombination events is governed by the recombination rate ρ and the duration τ_A of population A . The following lemma, proved in the appendix, shows that τ_A can be chosen sufficiently small that with high probability, zero or one recombination occurs. In the next lemma and what follows, set intersection is denoted with product notation, i.e. $XY = X \cap Y$ for events X, Y , and the important events are

$$\begin{aligned} R_i &= [\text{exactly } i \text{ recombinations occur in the time interval } (0, \tau_A)] \\ C_i &= [\text{exactly } i \text{ coalescences occur during the time interval } (0, t_{\text{root}})] \\ C_{0,X} &= [\text{no coalescence occurs in population } X]. \end{aligned}$$

Lemma 3 (Recombination Probabilities). $\mathbb{P}[R_0] = e^{-\rho\tau_A}$, $\mathbb{P}[R_1] \geq \mathbb{P}[R_1 C_{0,A}] \geq \rho\tau_A e^{-(1+2\rho)\tau_A}$, and $\mathbb{P}[\cup_{k \geq 2} R_k] = O(\rho^2 \tau_A^2)$ as $\tau_A \rightarrow 0^+$.

For the case where no recombination occurs, the probabilities of E and F are estimated in the following lemma using elementary MSC calculations which can be found in the appendix.

Lemma 4 (No Recombination Case). $\mathbb{P}[E|R_0] - \mathbb{P}[F|R_0] \leq \tau_{AB}$.

For the case where *exactly one* recombination occurs, the following lemma characterizes the behavior of coalescent events occurring below the root of S . Intuitively, it says that coalescence in population AB is rare when τ_{AB} is small; for a detailed proof see the appendix.

Lemma 5 (Effect of Small Internal Edge). As $\tau_{AB} \rightarrow 0^+$, $\mathbb{P}[C_0|R_1] = K + O(\tau_{AB})$, $\mathbb{P}[C_{0,A}|R_1 C_1] = O(\tau_{AB})$, and $\mathbb{P}[C_2|R_1] = O(\tau_{AB})$, where $K = \mathbb{P}[C_{0,A}|R_1] \in (0, 1)$ depends only on τ_A and ρ , and satisfies $\lim_{\tau_A \rightarrow 0} K = 1$ for any fixed $\rho > 0$.

Next we apply Lemma 5 to show that $\mathbb{P}[E|R_1 C_1] - \mathbb{P}[F|R_1 C_1]$ is small, tending to zero as $\tau_{AB} \rightarrow 0^+$.

Lemma 6. $\mathbb{P}[E|R_1 C_1] - \mathbb{P}[F|R_1 C_1] = O(\tau_{AB})$ as $\tau_{AB} \rightarrow 0^+$,

Proof Sketch. By a symmetry argument similar to that given in the proof of Lemma 4, one can show that conditional on $R_1 C_1 C_{0,A}^c$, the events E and F are equally likely. Therefore by the Law of Total Probability,

$$\mathbb{P}[E|R_1 C_1] - \mathbb{P}[F|R_1 C_1] = (\mathbb{P}[E|R_1 C_1 C_{0,A}] - \mathbb{P}[F|R_1 C_1 C_{0,A}]) \mathbb{P}[C_{0,A}|R_1 C_1] \leq \mathbb{P}[C_{0,A}|R_1 C_1],$$

and the right-hand side is $O(\tau_{AB})$ by Lemma 5. $\square$

Figure 3: A depiction of the event R_1C_0 . The portion of the ancestral recombination graph more ancient than t_{root} is not shown.

We now come to a key part of the calculation: the event R_1C_0 , depicted in Figure 3. The next lemma demonstrates that as long as θ is not too large, conditional on R_1C_0 , the event F is more likely than E .

Lemma 7. *The quantity $\bar{\alpha} := \mathbb{P}[F|R_1C_0] - \mathbb{P}[E|R_1C_0]$ depends only on θ and is positive if $\theta \in (0, 3/4)$.*

Proof Sketch. We sketch the proof idea here; the proof can be found in the Appendix. Conditional on R_1C_0 , four distinct lineages enter population ABC at time t_{root} . Denote these lineages by A_1, A_2, B , and C , as shown in Figure 3. Since no recombination occurs in population ABC , the order in which the lineages coalesce determines a *labeled history* (an ultrametric rooted binary tree with labeled tips and internal nodes rank-ordered according to age [RELY20]), whose tips are taken to be the lineages A_1, A_2, B and C at time t_{root} . There are 18 such labeled histories $\gamma_1, \dots, \gamma_{18}$. Since pairs of lineages coalesce uniformly at random under the coalescent, $\mathbb{P}[\gamma_j|R_1C_0] = \frac{1}{18}$ for all j , and hence

$$\mathbb{P}[F|R_1C_0] - \mathbb{P}[E|R_1C_0] = \frac{1}{18} \sum_{j=1}^{18} (\mathbb{P}[F|R_1C_0\gamma_j] - \mathbb{P}[E|R_1C_0\gamma_j]). \quad (6)$$

Having conditioned a particular labeled history, the probabilities $\mathbb{P}[E|R_1C_0\gamma_j] - \mathbb{P}[F|R_1C_0\gamma_j]$ for $j = 1, \dots, 18$ are computed in a straightforward manner, so that the right hand side of (6) is positive provided that not too much signal is lost by a high mutation rate. In particular, since there are *two* lineages from A and only one from each of B and C , at least one of the A lineages is more likely to be included in the final coalescing pair, favoring greater pairwise distances between A and the other two taxa than those between B and C . $\square$

The next lemma applies Lemmas 5, 6, and 7 to show that $\mathbb{P}[F|R_1] > \mathbb{P}[E|R_1]$ when the internal branch length τ_{AB} is small and the mutation rate θ is not too large.

Lemma 8. *If $\theta \in (0, 3/4)$, then $\mathbb{P}[E|R_1] - \mathbb{P}[F|R_1] = -\bar{\alpha}K + O(\tau_{AB})$ as $\tau_{AB} \rightarrow 0^+$ (where the term $-\bar{\alpha}K$ does not depend on τ_{AB}).*

Proof. By Law of Total Probability, since R_1C_0, R_1C_1 , and R_1C_2 partition R_1 ,

$$\mathbb{P}[E|R_1] - \mathbb{P}[F|R_1] = \sum_{i=0}^2 (\mathbb{P}[E|R_1C_i] - \mathbb{P}[F|R_1C_i]) \mathbb{P}[C_i|R_1]. \quad (7)$$

By Lemma 7, $\bar{\alpha} = \mathbb{P}[F|R_1C_0] - \mathbb{P}[E|R_1C_0]$ is positive and does not depend on τ_{AB} . Moreover, by Lemma 5, $K = \mathbb{P}[C_{0,A}|R_1]$ does not depend on τ_{AB} and $\mathbb{P}[C_0|R_1] = K + O(\tau_{AB})$, hence

$$(\mathbb{P}[E|R_1C_0] - \mathbb{P}[F|R_1C_0]) \mathbb{P}[C_0|R_1] = -\bar{\alpha}K + O(\tau_{AB}).$$

It remains to show that the other two summands on the right hand side of (7) are of magnitude at most $O(\tau_{AB})$. Indeed, the $i = 1$ term is at most $O(\tau_{AB})$ by Lemma 6. Moreover, by Lemma 5, $\mathbb{P}[C_2|R_1] = O(\tau_{AB})$, so the $i = 2$ term is at most $O(\tau_{AB})$ as well. $\square$

2.3 Proof of Theorem 1

Proof of Theorem 1. It suffices to prove (5) for some choice of parameters ρ, θ, τ_A , and τ_{AB} . Let $\rho > 0$ and $\theta \in (0, 3/4)$ be arbitrary; we will show that τ_A , and τ_{AB} can be chosen sufficiently small that (5) holds. Conditioning on the number of recombination events in population A ,

$$\mathbb{P}[F] - \mathbb{P}[E] > (\mathbb{P}[F|R_0] - \mathbb{P}[E|R_0])\mathbb{P}[R_0] + (\mathbb{P}[F|R_1] - \mathbb{P}[E|R_1])\mathbb{P}[R_1] - \mathbb{P}[\cup_{k \geq 2} R_k].$$

Therefore by Lemma 4 and the trivial inequality $\mathbb{P}[R_0] \leq 1$,

$$\mathbb{P}[F] - \mathbb{P}[E] > -\tau_{AB} + (\mathbb{P}[F|R_1] - \mathbb{P}[E|R_1])\mathbb{P}[R_1] - \mathbb{P}[\cup_{k \geq 2} R_k].$$

By Lemma 8, there exists $\delta > 0$ such that $\mathbb{P}[F|R_1] - \mathbb{P}[E|R_1] > \bar{\alpha}K/2$ whenever $0 < \tau_{AB} < \delta$. Assume further that $\tau_{AB} \in (0, \delta)$. Then

$$\mathbb{P}[F] - \mathbb{P}[E] > -\tau_{AB} + \frac{\bar{\alpha}K}{2}\mathbb{P}[R_1] - \mathbb{P}[\cup_{k \geq 2} R_k].$$

By Lemma 3, there exists constants $C, D > 0$ not depending on τ_{AB} such that $\mathbb{P}[R_1] \geq C\rho\tau_A$ and $\mathbb{P}[\cup_{i \geq 2} R_i] \leq D\rho^2\tau_A^2$, so that

$$\mathbb{P}[F] - \mathbb{P}[E] > -\tau_{AB} + \left(\frac{1}{2}\bar{\alpha}KC - D\rho\tau_A\right)\rho\tau_A.$$

Since K does not depend on τ_{AB} and $K \rightarrow 1$ as $\tau_A \rightarrow 0$ by Lemma 5, there exists $\tau_A > 0$ sufficiently small that both $K > 1/2$ and $\epsilon := \bar{\alpha}C/4 - D\rho\tau_A > 0$. It follows that $\mathbb{P}[F] - \mathbb{P}[E] > -\tau_{AB} + \epsilon\rho\tau_A$. Since ϵ does not depend on τ_{AB} , it follows that $\mathbb{P}[F] - \mathbb{P}[E] > 0$ for τ_{AB} sufficiently small. $\square$

3 Simulation Study

We performed a simulation study to characterize the inconsistency zone established in Theorem 1. Code and documentation can be found at <https://github.com/max-hill/MSCR-simulator.git>. In all simulations, sequence data is generated according to the MSCR process on an ultrametric species phylogeny S with three species A, B, C , and rooted topology $AB|C$. In all cases, $k = 500$, $\tau_A = 1$ and θ does not vary among populations. We use the notation $\hat{p}_{XY|Z}$ to denote the proportion of the m samples from which the rooted triple $XY|Z$ was inferred, and $\hat{t}$ to denote the R^* uniquely favored rooted triple of the m samples. By the strong law of large numbers, $\hat{p}_{XY|Z}$ serves as an estimate of $\mathbb{P}[E_{XY|Z}]$ for large m , where $E_{XY|Z}$ is defined as in Lemma 2.

The range of recombination rates considered in these simulations are comparable to those in [LK12], who suggest they encompass biologically plausible values. As for mutation rates, typical rates in eukaryotes are on the order of $\mu = 10^{-9}$ to 10^{-8} per site per generation [Hah18, Lyn10] and effective eukaryotic population sizes N_e range from 10^4 to 10^8 [LM15], making the values considered here of $\theta = 2N_e\mu \in \{0.01, 0.1\}$ plausible as well. Computational constraints limited the ability to consider mutation rates lower than these, as doing so would have necessitated an increase in k or m to compensate; however the analytic results here predict that the inconsistency zone will persist, and may grow, for smaller values of θ : the computed difference $\bar{\alpha} = \mathbb{P}[F|R_1C_0] - \mathbb{P}[E|R_1C_0]$ actually increases as $\theta \rightarrow 0$, as shown in Figure 10 in the appendix, suggesting that phylogenetic conflict may be greater under regimes with smaller mutation rates than those simulated here.

In the first experiment, we simulated the MSCR-JC(k) process under a variety of parameter regimes in order to characterize the anomaly zone and evaluate the robustness of triplet-based inference in the presence of intralocus recombination. In particular $m = 10^4$ replicates were generated independently under each parameter regime, with the aim of estimating how frequently the correct topology was inferred. The parameters used were $\theta = 0.1$, $\tau_{AB} \in \{0.01, 0.02, \dots, 0.20\}$, $\rho_A \in \{0, \dots, 20\}$, and $\rho_X = 1$ for all $X \neq A$, so that recombination occurred only in population A . Figure 4 shows the value of $\hat{t}$ for each simulated parameter regime, and Figure 5 plots the surface $z = \hat{p}_{AB|C} - \hat{p}_{BC|A}$ as a function of ρ_A and τ_{AB} , so that parameter regimes with negative z values indicates inconsistent inference.

Figure 4: R^* inconsistency zone. The color of each dot represents a simulation of $m = 10000$ replicates.

Figure 5: The surface $z = \hat{p}_{AB|C} - \hat{p}_{BC|A}$ as a function of τ_{AB} and ρ_A .

We also evaluated R^* inference with rooted triples inferred not by equation (1), but rather by maximum-likelihood under the (false) assumption of no intralocus recombination; in this mode, which we call **R^* with maximum likelihood**, binary sequences were simulated and the maximum likelihood rooted triple was computed analytically using the method in [Yan00]. A plot almost identical to Figure 4 was obtained. For the very short internal branch length $\tau_{AB} = 0.01$, simulations were run with similar parameters and higher number of replicates ($m = 15,000$), with inference performed using both R^* with sequence distances and R^* with maximum likelihood. Figure 6 plots the difference $y = \hat{p}_{BC|A} - \hat{p}_{AB|C}$ as a function of ρ_A obtained from these simulations.

These results show that the combination of intralocus recombination in population A along with a very short internal branch length τ_{AB} resulted in the rooted triple $BC|A$ being more slightly likely to be inferred than the correct topology $AB|C$. Figure 6 shows clearly that this effect increases for larger values of ρ_A . Nonetheless, as both Figures 5 and 6 show, the magnitude of this effect is relatively small: even when $\hat{p}_{BC|A} - \hat{p}_{AB|C}$ is positive, it is never greater than 0.1. Moreover, as Figures 4 and 5 show, this effect disappears when τ_{AB} is increased (ILS being less likely to occur on longer edges of S). Notably, even for high rates of recombination, R^* under both sequence distance mode and maximum likelihood mode always correctly inferred the topology of S when $\tau_{AB} > 0.1$ coalescent units.

In our second experiment, we relaxed the assumption that recombination occurs only in population A by allowing for recombination in population B as well. For this simulation, $\tau_{AB} = 0.01$ and $\theta = 0.01$, with inference performed using R^* with sequence distances. Figure 7 shows the uniquely favored rooted triple for each choice of ρ_A and ρ_B , with each estimate obtained from $m = 10^5$ samples. When this experiment was repeated with $\tau_{AB} = 0.1$, all but one parameter regimes resulted in correct inference; the exception was when $\rho_A = 0$ and $\rho_B = 20$, in which case $\hat{t} = AC|B$. These results support the hypothesis that taxa exhibiting higher rates of recombination relative to other taxa are more likely to be inferred as more distantly related, but that the effect is small and manifests only in species triplets with very short internal branches.

The third experiment tested the effect when all populations in S (excluding the root population ABC) experience recombination at comparable rates. The simulation parameters were $\rho := \rho_A = \rho_B = \rho_C = \rho_{AB} \in \{0, 1, \dots, 20\}$ and $\rho_{ABC} = 0$, along with $\theta = 0.1$, $\tau_{AB} = 0.01$, and $m = 10^6$, with inference performed using R^* with sequence distances. The results, shown in Figure 8, suggest that when recombination rates are similar on the edges of S , greater recombination rates does *not* lead to incorrect inference of rooted triples: in all cases, $\hat{p}_{AB|C} > \hat{p}_{AC|B} \vee \hat{p}_{BC|A}$, suggesting consistent inference despite the very short internal branch length, a result which agrees with the conclusions of [LK12] that even high recombination rates are not a significant source of error, at least when rates are comparable across species. Thus, the existence of

Figure 6: The effect of increasing ρ_A on inference using R^* with sequence distances and maximum likelihood.

differential rates of recombination between closely related taxa appears to be a necessary condition for a species tree S to lie in the inconsistency zone.

4 Discussion

The primary focus of this study is the effect of intralocus recombination on the inference of rooted triples. In contrast to previous simulation studies [LK12, WL16, Con20], the current work considers the effect of intralocus recombination on inference of species phylogenies *with recombination rate heterogeneity across taxa*. Our main result is a proof that within the parameter space of species phylogenies there exists a subset—the inconsistency zone—in which phylogenetic conflict between the topology of the species phylogeny and the topology of inferred gene trees is of a sufficient level to render certain majority vote methods statistically inconsistent. We further quantify and characterize this inconsistency zone through simulations, showing that it includes biologically plausible recombination and mutation rates for eukaryotes, and suggesting that it arises on species phylogenies exhibiting both (1) very short internal branch lengths (less than 0.1 coalescent units) and (2) differential rates of recombination between closely related taxa. These results highlight a way in which intralocus recombination can exacerbate ILS and lead to overestimation of the divergence times of those taxa exhibiting disproportionately high intralocus recombination rates relative to other taxa.

These findings do not necessarily contradict the conclusions of [LK12] that the effect of unrecognized intralocus recombination can be minor. Indeed, our simulation experiments provide further evidence that inference of rooted triples is hampered by unrecognized intralocus recombination only in cases where the internal branch length of the species tree is short, that is in cases where ILS is already high. The size of the observed effect is also relatively small; even when the uniquely favored rooted triple do not agree with the species tree, it is usually only slightly more common than the true rooted triple. Furthermore, if differential rates of recombination between closely-related taxa are rare, then summary coalescent-based methods which take no account of intralocus recombination may nonetheless indeed be robust even when recombination rates are high.

Our results raise a number of questions for future study. Our analysis focused on a simple idealized case consisting of a rooted ultrametric three-taxa species phylogeny with mutations modeled by the Jukes-Cantor process. The nature and significance of the inconsistency zone may be affected by factors such as variable population sizes as well as elements of mutation and recombination rate heterogeneity not considered here. In addition, our theoretical results only consider distance-based gene tree estimation. Extending these results to likelihood-based inference would be of interest.

Figure 7: R^* inference with recombination in both populations A and B .

Figure 8: Equal recombination rates in A, B, C and AB .

References

- [Are13] Miguel Arenas. The importance and application of the ancestral recombination graph. *Frontiers in genetics*, 4:206, 2013.
- [BH20] David Bryant and Matthew W Hahn. The concatenation question. In C. Scornavacca, F. Delsuc, and N. Galtier, editors, *Phylogenetics in the Genomic Era*, chapter 3.4, pages 3.4:1–3.4:23. No commercial publisher— Authors open access book. This book is freely available online at <https://hal.inria.fr/PGE>, 2020.
- [Con20] Michael Conry. *Determining the Impact of Recombination on Phylogenetic Inference*. PhD thesis, The Florida State University, 2020.
- [DDBR09] James H Degnan, Michael DeGiorgio, David Bryant, and Noah A Rosenberg. Properties of consensus methods for inferring species trees from gene trees. *Systematic Biology*, 58(1):35–54, 2009.
- [Deg13] James H Degnan. Anomalous unrooted gene trees. *Systematic biology*, 62(4):574–590, 2013.
- [DMNR17] Gautam Dasarathy, Elchanan Mossel, Robert Nowak, and Sebastien Roch. Coalescent-based species tree estimation: a stochastic farris transform. *arXiv preprint arXiv:1707.04300*, 2017.
- [DR06] James H Degnan and Noah A Rosenberg. Discordance of species trees with their most likely gene trees. *PLoS genet*, 2(5):e68, 2006.
- [DR09] James H. Degnan and Noah A. Rosenberg. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution*, 24(6):332–340, 2009.
- [EXJ⁺16] Scott V. Edwards, Zhenxiang Xi, Axel Janke, Brant C. Faircloth, John E. McCormack, Travis C. Glenn, Bojian Zhong, Shaoyuan Wu, Emily Moriarty Lemmon, Alan R. Lemmon, Adam D. Leaché, Liang Liu, and Charles C. Davis. Implementing and testing the multispecies coalescent model: A valuable paradigm for phylogenomics. *Molecular Phylogenetics and Evolution*, 94:447–462, 2016.

- [GM97] Robert C Griffiths and Paul Marjoram. An ancestral recombination graph. In P. Donnelly and S Tavaré, editors, *Progress in Population Genetics and Human Evolution*, volume 87, page 257, 1997.
- [Hah18] Matthew William Hahn. *Molecular population genetics*. Oxford University Press, 2018.
- [JC⁺69] Thomas H Jukes, Charles R Cantor, et al. Evolution of protein molecules. *Mammalian protein metabolism*, 3:21–132, 1969.
- [KYT20] Paschalia Kapli, Ziheng Yang, and Maximilian J Telford. Phylogenetic tree building in the genomic age. *Nature Reviews Genetics*, 21(7):428–444, 2020.
- [LK12] Hayley C Lanier and L Lacey Knowles. Is recombination a problem for species-tree analyses? *Systematic Biology*, 61(4):691–701, 2012.
- [LKDA10] Bret R. Larget, Satish K. Kotha, Colin N. Dewey, and Cécile Ané. BUCKy: Gene tree/species tree reconciliation with Bayesian concordance analysis. *Bioinformatics*, 26(22):2910–2911, 09 2010.
- [LM15] Michael Lynch and Georgi K Marinov. The bioenergetic costs of a gene. *Proceedings of the National Academy of Sciences*, 112(51):15690–15695, 2015.
- [LYE10] Liang Liu, Lili Yu, and Scott V Edwards. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC evolutionary biology*, 10(1):1–18, 2010.
- [Lyn10] Michael Lynch. Evolution of the mutation rate. *TRENDS in Genetics*, 26(8):345–352, 2010.
- [MLH19] Fábio K Mendes, Andrew P Livera, and Matthew W Hahn. The perils of intralocus recombination for inferences of molecular convergence. *Philosophical Transactions of the Royal Society B*, 374(1777):20180244, 2019.
- [MRB⁺14] Siavash Mirarab, Rezwana Reaz, Md S Bayzid, Théo Zimmermann, M Shel Swenson, and Tandy Warnow. Astral: genome-scale coalescent-based species tree estimation. *Bioinformatics*, 30(17):i541–i548, 2014.
- [PG16] Fabio Pardi and Olivier Gascuel. Distance-based methods in phylogenetics, 2016.
- [RELY20] Bruce Rannala, Scott V. Edwards, Adam Leaché, and Ziheng Yang. The multi-species coalescent model and species tree inference. In C. Scornavacca, F. Delsuc, and N. Galtier, editors, *Phylogenetics in the Genomic Era*, chapter 3.3, pages 3.3:1–3.3:21. No commercial publisher — Authors open access book. This book is freely available online at <https://hal.inria.fr/PGE>, 2020.
- [RY03] Bruce Rannala and Ziheng Yang. Bayes estimation of species divergence times and ancestral population sizes using dna sequences from multiple loci. *Genetics*, 164(4):1645–1656, 2003.
- [SG16] Mark S. Springer and John Gatesy. The gene tree delusion. *Molecular Phylogenetics and Evolution*, 94:1–33, 2016.
- [SG18] Mark S Springer and John Gatesy. Delimiting coalescence genes (c-genes) in phylogenomic data sets. *Genes*, 9(3):123, 2018.
- [SS⁺03] Charles Semple, Mike Steel, et al. *Phylogenetics*, volume 24. Oxford University Press on Demand, 2003.
- [SS20] Dominik Schrempf and Gergely Szöllösi. The sources of phylogenetic conflicts. In C. Scornavacca, F. Delsuc, and N. Galtier, editors, *Phylogenetics in the Genomic Era*, chapter 3.1, pages 3.1:1–3.1:23. No commercial publisher — Authors open access book. This book is freely available online at <https://hal.inria.fr/PGE>, 2020.
- [Ste16] Mike Steel. *Phylogeny: discrete and random processes in evolution*. SIAM, 2016.

- [Wan17] Kun-Chieh Wang. *Phylogenetic reconstruction accuracy in the face of heterogeneity, recombination, and reticulate evolution*. The University of Wisconsin-Madison, 2017.
- [War17] Tandy Warnow. *Computational phylogenetics: an introduction to designing methods for phylogeny estimation*. Cambridge University Press, 2017.
- [WL16] Zhiwei Wang and Kevin J Liu. A performance study of the impact of recombination on species tree analysis. *BMC genomics*, 17(10):165–174, 2016.
- [Yan00] Ziheng Yang. Complexity of the simplest phylogenetic estimation problem. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 267(1439):109–116, 2000.
- [ZY21] Tianqi Zhu and Ziheng Yang. Complexity of the simplest species tree problem. *Molecular Biology and Evolution*, 38(9):3993–4009, 01 2021.

A Preliminary lemmas

This section contains the proof of some preliminary lemmas.

A.1 Proof of Lemma 1

We begin with some auxiliary lemmas.

Lemma 9. *Suppose $\mathcal{G}$ is generated according to the MSCR process on S and $\mathcal{M}$ is its associated collection of marginal gene trees. With probability one, for all $T \in \mathcal{M}$, $I(T)$ is a finite union of nondegenerate subintervals of $[0, 1]$ such that the endpoints of each of these intervals are elements of $B(\mathcal{G}) \cup \{0, 1\}$ where $B(\mathcal{G})$ is the set of recombination breakpoints of $\mathcal{G}$.*

Proof. Let $\mathcal{H}$ denote the collection consisting of the empty set together with all nondegenerate subintervals of $[0, 1]$ whose endpoints are elements of $B(\mathcal{G})$, and define

$$\Pi = \{S \subset [0, 1] : S = I_1 \cup \dots \cup I_j \text{ with } I_1, \dots, I_j \in \mathcal{H} \text{ for some } j = 0, 1, \dots\}.$$

For each $e \in E(\mathcal{G})$ define $\mathcal{P}_e = \{x \in [0, 1] : e \in E(\mathcal{T}(x))\}$. As noted in [GM97], with probability one $\mathcal{P}_e$ is either empty or is a finite union of intervals such that the endpoints of these intervals are a subset of $B(\mathcal{G})$. Therefore $\mathcal{P}_e \in \Pi$ for all e . Let $T \in \mathcal{M}$ and denote the edge set of T by $E(T)$. Let $x \in [0, 1]$. Then $\mathcal{T}(x) = T$ if and only if $e \in E(\mathcal{T}(x))$ for all $e \in E(T)$. Therefore

$$I(T) = \{x \in [0, 1] : \mathcal{T}(x) = T\} = \{x \in [0, 1] : E(T) \subset E(\mathcal{T}(x))\} = \bigcap_{e \in E(T)} \mathcal{P}_e$$

Since $E(T)$ is finite, it will suffice to show that Π is closed under finite intersections, as the statement of the lemma will then follow from the inclusion $I(T) \in \Pi$. Suppose $I = I_1 \cup \dots \cup I_p$ and $J = J_1 \cup \dots \cup J_q$ with $I_i, J_j \in \mathcal{H}$ for all i, j . Then by the distributive property of unions and intersections,

$$I \cap J = \bigcup_{i=1}^p \bigcup_{j=1}^q I_i \cap J_j$$

and we note that each set $I_i \cap J_j$ is either empty or is an interval with endpoints in $B(\mathcal{G})$. Therefore $I_i \cap J_j \in \mathcal{H}$ and hence $I \cap J \in \Pi$. The conclusion that Π is closed under finite intersections follows by induction. $\square$

Lemma 10. *With probability one, if $\mathcal{G}$ is generated according the MSCR process on S , then*

$$\frac{1}{k} \# \{i : i \in S_k \cap I(T_g)\} \rightarrow |I(T_g)|$$

as $k \rightarrow \infty$ for all $T_g \in \mathcal{M}$.

Proof. We first prove the following claim: If I is a subinterval of $[0, 1]$ with endpoints $a, b \in B(\mathcal{G})$, where $a \leq b$, then

$$\lim_{k \rightarrow \infty} \frac{1}{k} \# \{i : i \in S_k \cap I\} = b - a. \quad (8)$$

To prove this claim, first observe that

$$\frac{1}{k} \# \{i : i \in S_k \cap I\} = \frac{p_k - q_k}{k}$$

where $p_k = \sup \{\ell \geq 0 : \frac{\ell}{k} < b\}$ and $q_k = \sup \{\ell \geq 0 : \frac{\ell}{k} < a\}$. Moreover, by definition of supremum

$$\frac{p_k}{k} < b \leq \frac{p_k + 1}{k} \quad \text{and} \quad \frac{q_k}{k} < a \leq \frac{q_k + 1}{k}.$$

These inequalities imply that $\lim_{k \rightarrow \infty} \frac{p_k - q_k}{k} = b - a$. This proves the claim.

We now prove the statement of the lemma. Using the result of Lemma 9, with probability one there exists a pairwise disjoint sequence of intervals $I_1, \dots, I_h \in \mathcal{H}$ such that $I(T_g) = \cup_{i=1}^h I_i$. Therefore

$$\{i : i \in S_k \cap I(T_g)\} = \bigcup_{j=1}^h \{i : i \in S_k \cap I_j\}$$

and hence

$$\frac{1}{k} \# \{i : i \in S_k \cap I(T_g)\} = \sum_{j=1}^h \frac{1}{k} \# \{i : i \in S_k \cap I_j\}.$$

By Equation (8), $\frac{1}{k} \# \{i : i \in S_k \cap I_j\} \rightarrow |I_j|$ for all j , and hence

$$\lim_{k \rightarrow \infty} \frac{1}{k} \# \{i : i \in S_k \cap I(T_g)\} = \sum_{i=1}^h |I_i| = |I(T_g)|.$$

□

Finally, we are in a position to prove Lemma 1.

Proof of Lemma 1. Let $k \geq 2$, $S_k = \{\frac{j}{k-1} : j = 0, 1, \dots, k-1\}$, and $\xi_i = \mathbf{1}_{[\mathcal{N}(X,i) \neq \mathcal{N}(Y,i)]}$, $i \in S_k$. Then $\xi_i, i \in S_k$ are independent and $\delta_{XY}(M_k) = \sum_{i \in S_k} \xi_i$. For all $T_g \in \mathcal{M}$,

$$\mathbb{E}[\xi_i] = \mathbb{P}[\mathcal{N}(X,i) \neq \mathcal{N}(Y,i) | \mathcal{T}(i) = T_g] = \frac{3}{4} \left(1 - e^{-\frac{4}{3}d_{XY}^{T_g}}\right) \quad (9)$$

a.s. for all $i \in I(T_g)$ by definition of the Jukes-Cantor process (see e.g., [PG16]). Since $\mathcal{M}$ is a.s. finite,

$$\frac{1}{k} \delta_{XY}(M_k) = \frac{1}{k} \sum_{i \in S_k} \xi_i = \frac{1}{k} \sum_{T_g \in \mathcal{M}} \sum_{i \in S_k \cap I(T_g)} \xi_i = \sum_{T_g \in \mathcal{M}: N_k(T_g) > 0} \frac{N_k(T_g)}{k} \left(\frac{1}{N_k(T_g)} \sum_{i \in S_k \cap I(T_g)} \xi_i \right) \quad (10)$$

where $N_k(T_g) = \#\{i : i \in S_k \cap I(T_g)\}$. Since $I(T_g)$ is a.s. finite union of nondegenerate intervals with endpoints in S_k for all $T_g \in \mathcal{M}$, it follows that $\lim_{k \rightarrow \infty} N_k(T_g) = \infty$ and $\lim_{k \rightarrow \infty} N_k(T_g)/k = w_g$. (By Lemmas 9 and 10.) From these limits, together with (9) and the strong law of large numbers, the right-hand side of (10) converges to Δ_{XY} almost surely. □

A.2 Proof of Lemma 2

Proof of Lemma 2. Suppose $M_k^{(1)}, \dots, M_k^{(m)}$ are generated independently according to the MSCR-JC(k) process on S . For each $\ell = 1, \dots, m$ and each triplet $X, Y, Z \in L$, let $E_{XY|Z}^{(\ell)}$ to be the event that the rooted triple inferred from $M^{(\ell)}$ is $XY|Z$. Let $U_S^{(m)}$ be the event that the topology of S is successfully reconstructed from these m independent samples using R^* with sequence distances. It suffices to show $\lim_{m \rightarrow \infty} \mathbb{P}[U_S^{(m)}] = 1$ if and only if (4) holds. By definition of the R^* pipeline, $U_S^{(m)} = \bigcap_{XY|Z \in \mathcal{R}(S)} U_{XY|Z}^{(m)}$ where

$$U_{XY|Z}^{(m)} = \left[\sum_{\ell=1}^m \mathbf{1}_{E_{XY|Z}^{(\ell)}} > \max \left\{ \sum_{\ell=1}^m \mathbf{1}_{E_{XZ|Y}^{(\ell)}}, \sum_{\ell=1}^m \mathbf{1}_{E_{YZ|X}^{(\ell)}} \right\} \right]. \quad (11)$$

Since the samples $M_k^{(1)}, \dots, M_k^{(m)}$ are i.i.d., the law of large numbers implies

$$\lim_{m \rightarrow \infty} \frac{1}{m} \sum_{\ell=1}^m \mathbf{1}_{E_{XY|Z}^{(\ell)}} = \mathbb{P}[E_{XY|Z}^{(1)}] \quad (12)$$

for all triplets $X, Y, Z \in L$. It follows from (11) and (12) that (4) holds if and only if $\lim_{m \rightarrow \infty} \mathbb{P}[U_{XY|Z}^{(m)}] = 1$ for all $XY|Z \in \mathcal{R}(S)$. Since $|\mathcal{R}(S)|$ is finite, the equivalence of (4) and $\lim_{m \rightarrow \infty} \mathbb{P}[U_S^{(m)}] = 1$ is established. □

B Key lemmas

We now prove the key lemmas leading to our main result.

B.1 Proof of Lemma 3

Proof of Lemma 3. Since recombination occurs only in population A of S , the event R_i occurs if and only if exactly i recombinations occur in population A . Since the process starts with a single lineage in A at time t , the first event (if an event occurs in A) must be a recombination, which occurs with rate ρ . Therefore there is an exponentially distributed waiting time e_ρ with rate ρ such that a the lineage in A recombines at time e_ρ if $e_\rho < \tau_A$ and not otherwise. Therefore $\mathbb{P}[R_0] = \mathbb{P}[e_\rho > \tau_A] = e^{-\rho\tau_A}$.

Given $e_\rho < \tau_A$, population A has two lineages at time e_ρ , so the next event (coalescence or recombination) occurs at rate $1 + 2\rho$. Therefore there is an exponentially distributed waiting time $e_{1+2\rho}$ with rate $1 + 2\rho$ and independent of e_ρ such that if $e_{1+2\rho} < \tau_A - e_\rho$ then an event occurs at time $e_\rho + e_{1+2\rho}$ and not otherwise. Therefore $R_1 C_{0,A} = [e_\rho < \tau_A, e_{1+2\rho} > \tau_A - e_\rho]$, and hence by the numerical inequality $e^x - 1 \geq x$,

$$\begin{aligned}\mathbb{P}[R_1 C_{0,A}] &= \int_0^{\tau_A} \mathbb{P}[e_{1+2\rho} > \tau_A - s | e_\rho = s] \rho e^{-\rho s} ds \\ &= \int_0^{\tau_A} e^{-(1+2\rho)(\tau_A - s)} \rho e^{-\rho s} ds \\ &= \frac{\rho}{1 + \rho} e^{-(1+2\rho)\tau_A} (e^{\tau_A(1+\rho)} - 1) \\ &\geq \rho \tau_A e^{-(1+2\rho)\tau_A}.\end{aligned}$$

To estimate $\mathbb{P}[\cup_{i=2}^\infty R_i]$, we may restrict our analysis to the single-population case since the recombination rate is positive only in population A . Define

$$\begin{aligned}\tau_1 &= \inf \{t \geq 0 : \text{recombination occurs at time } t\} \quad \text{and} \\ \tau_2 &= \inf \{t > \tau_1 : \text{recombination occurs at time } t\},\end{aligned}$$

so that $\cup_{i=2}^\infty R_i \subset [\tau_2 < \tau_A]$. When $t = 0$, there is only one lineage in population A , and hence if an event occurs it must be a recombination event, which occurs at rate ρ . Therefore

$$\mathbb{P}[\tau_2 < \tau_A] = \int_0^{\tau_A} \mathbb{P}[\tau_2 < \tau_A | \tau_1 = z] \rho e^{-\rho z} dz$$

Conditional on $\tau_1 < \tau_A$, let s_1 denote the time until the next event. There are two lineages at time τ_1 , so that events happen at rate $1 + 2\rho$, and hence $s_1 \sim \exp(1 + 2\rho)$, with the event at $\tau_1 + s_1$ being either a recombination with probability $\frac{2\rho}{1+2\rho}$ or a coalescence with probability $\frac{1}{1+2\rho}$. Therefore

$$\mathbb{P}[\tau_2 < \tau_A] = \int_0^{\tau_A} \int_0^{\tau_A - z} \mathbb{P}[\tau_2 < \tau_A | \tau_1 = z, s_1 = y] (1 + 2\rho) e^{-(1+2\rho)y} dy \rho e^{-\rho z} dz \quad (13)$$

Conditional on $s_1 + \tau_1 < \tau_A$, let H_{coal} be the event that a coalescence occurs at time $\tau_1 + s_1$ rather than a recombination. Assume that $z \in (0, \tau_A)$ and $y \in (0, \tau_A - z)$. Then by the Law of Total Probability,

$$\mathbb{P}[\tau_2 < \tau_A | \tau_1 = z, s_1 = y] = \frac{1}{1 + 2\rho} \mathbb{P}[\tau_2 < \tau_A | \tau_1 = z, s_1 = y, H_{\text{coal}}] + \frac{2\rho}{1 + 2\rho} \cdot 1. \quad (14)$$

Moreover, conditional on $[\tau_1 = z, s_1 = y] \cap H_{\text{coal}}$, there is exactly one lineage at time $s_1 + \tau_1$ in population A , so the time s_2 until the next event is an independent exponential with rate ρ and the next event must be a recombination. Therefore

$$\mathbb{P}[\tau_2 < \tau_A | \tau_1 = z, s_1 = y, H_{\text{coal}}] = \mathbb{P}[s_2 < \tau_A - y - z] \leq \mathbb{P}[s_2 < \tau_A] = 1 - e^{-\rho\tau_A} \leq \rho\tau_A$$

By this inequality and (14), $\mathbb{P}[\tau_2 < \tau_A | \tau_1 = z, s_1 = y] \leq \frac{\rho}{1+2\rho} (\tau_A + 2)$. Therefore by (13),

$$\mathbb{P}[\tau_2 < \tau_A] \leq \int_0^{\tau_A} \int_0^{\tau_A - z} \rho^2 (\tau_A + 2) dy dz = \rho^2 \tau_A^2 (\tau_A + 2) = O(\rho^2 \tau_A^2),$$

□

B.2 Proof of Lemma 4

Proof of Lemma 4. Since the recombination rate is zero except in population A , it follows that conditional on R_0 , the ARG is a tree, so that the analysis is exactly the same as in the case of the multi-species coalescent. The lineages from A and B coalesce in population AB with probability $1 - e^{-\tau_{AB}}$. If that happens, the lineages from A and B coalesce at some time $t_c < t_{\text{root}}$ so that

$$\Delta_{AB} = \frac{3}{4} \left(1 - e^{-\frac{8}{3}\theta t_c}\right) < \frac{3}{4} \left(1 - e^{-\frac{8}{3}\theta t_{\text{root}}}\right) < \Delta_{AC} \wedge \Delta_{BC}$$

hence E occurs with probability one on that event. On the other hand, if the lineages A and B do *not* coalesce in population AB , then by a symmetry argument (each pair of lineages being equally likely to coalesce in population ABC), the random variables Δ_{AB} , Δ_{AC} , and Δ_{BC} are identically distributed, so that $\mathbb{P}[E|R_0C_0] = \mathbb{P}[F|R_0C_0] = 1/3$. It follows by the law of total probability that $\mathbb{P}[E|R_0] - \mathbb{P}[F|R_0] = \mathbb{P}[C_1|R_0] = 1 - e^{-\tau_{AB}} \leq \tau_{AB}$. $\square$

B.3 Proof of Lemma 5

Proof of Lemma 5. Since $K \in (0, 1)$ depends only on the portion of the ARG in population A , it is clear that K does not depend on τ_{AB} . By Lemma 3, $\mathbb{P}[R_1C_{0,A}] \geq \rho\tau_A e^{-\tau_A(1+2\rho)}$ and $\mathbb{P}[R_1] \leq 1 - \mathbb{P}[R_0] = 1 - e^{-\rho\tau_A} \leq \rho\tau_A$. Therefore,

$$K = \mathbb{P}[C_{0,A}|R_1] = \frac{\mathbb{P}[R_1C_{0,A}]}{\mathbb{P}[R_1]} \geq e^{-\tau_A(1+2\rho)}.$$

Therefore $\lim_{\tau_A \rightarrow 0} K = 1$ for all $\rho > 0$ uniformly in τ_{AB} .

Let $e_2 \sim \exp(1)$ and $e_3 \sim \exp(3)$ be independent exponential clocks for the arrival of the next coalescence event given 2 and 3 lineages respectively. Then

$$\mathbb{P}[C_0|R_1] = \mathbb{P}[C_{0,A}C_{0,AB}|R_1] = K\mathbb{P}[C_{0,AB}|R_1C_{0,A}] = K\mathbb{P}[e_3 > \tau_{AB}] = Ke^{-3\tau_{AB}},$$

which implies

$$\mathbb{P}[C_0|R_1] = K + O(\tau_{AB}). \quad (15)$$

Next we claim that

$$\mathbb{P}[C_1|R_1C_{0,A}] = O(\tau_{AB}). \quad (16)$$

To prove this claim, observe that conditional on $R_1C_{0,A}$ three lineages enter population AB (i.e., two lineages from A and one from B) and one leaves (backwards in time). Thus, the event C_1 occurs if and only if exactly one of the three pairs of lineages entering population AB coalesces and no further coalescence events occur in population AB (i.e., during the time interval $(\tau_A, t_{\text{root}})$). Letting $\tilde{e}_3$ be the time until the first coalescence given three lineages, and $\tilde{e}_2$ be the (independent) clock for two lineages, it follows that C_1 occurs if and only if $\tilde{e}_3 < \tau_{AB}$ and $\tilde{e}_3 + \tilde{e}_2 > \tau_{AB}$. Therefore since $\tilde{e}_3 \sim \exp(3)$ and $\tilde{e}_2 \sim \exp(1)$,

$$\begin{aligned} \mathbb{P}[C_1|R_1C_{0,A}] &= \mathbb{P}[\tilde{e}_3 < \tau_{AB}, \tilde{e}_2 + \tilde{e}_3 > \tau_{AB}] \\ &= 3 \int_0^{\tau_{AB}} e^{-3x} \mathbb{P}[\tilde{e}_2 > \tau_{AB} - x] dx \\ &= 3 \int_0^{\tau_{AB}} e^{-2x - \tau_{AB}} dx \\ &= \frac{3}{2} e^{-\tau_{AB}} (1 - e^{-2\tau_{AB}}). \end{aligned}$$

This proves (16). On the other hand, conditional on $R_1C_{0,A}^c$, the event C_1 occurs if and only if the two lineages (A and B) entering population AB fail to coalesce during the time interval $(\tau_A, t_{\text{root}})$. Therefore

$$\mathbb{P}[C_1|R_1C_{0,A}^c] = \mathbb{P}[e_2 > \tau_{AB}] = e^{-\tau_{AB}}. \quad (17)$$

By the Law of Total Probability, $\mathbb{P}[C_1|R_1] = \mathbb{P}[C_1|R_1C_{0,A}]K + \mathbb{P}[C_1|R_1C_{0,A}^c](1 - K)$, and hence by (16) and (17),

$$\mathbb{P}[C_1|R_1] = 1 - K + O(\tau_{AB}). \quad (18)$$

Conditional on R_1 , exactly one of the events C_0, C_1 , and C_2 occurs. Therefore $\mathbb{P}[C_2|R_1] + \mathbb{P}[C_0|R_1] + \mathbb{P}[C_1|R_1] = 1$. Therefore by (15) and (18), $\mathbb{P}[C_2|R_1] = O(\tau_{AB})$. Finally, applying Bayes' theorem along with the estimates from (16), (18), one obtains

$$\mathbb{P}[C_{0,A}|R_1C_1] = \frac{\mathbb{P}[C_{0,A}|R_1] \mathbb{P}[C_1|R_1C_{0,A}]}{\mathbb{P}[C_1|R_1]} = O(\tau_{AB}),$$

uniformly in ρ, τ_A as $\tau_{AB} \rightarrow 0^+$. $\square$

B.4 Proof of Lemma 7

Proof of Lemma 7. Conditional on R_1C_0 , four distinct lineages enter population ABC at time t_{root} . We denote these lineages by A_1, A_2, B , and C , with B and C with the letter corresponding the originating population, as shown in Figure 3. Since no recombination occurs in population ABC , the order in which the lineages coalesce determines a *labeled history* (defined as an ultrametric rooted binary tree with labeled tips and internal nodes rank-ordered according to age [RELY20]), whose tips are taken to be the lineages A_1, A_2, B and C at time t_{root} . Denote by $((WX)Y)Z$ the event that the labeled history exhibited in population ABC has a caterpillar topology, W and X are the first pair of lineages to coalesce, and the most recent common ancestor of W and Z coalesces with the lineage ancestral to Y before coalescing with that of Z . Denote by $(WX)YZ$ the event that the labeled history has balanced topology with lineages W, X coalescing first.

Let t_k be the duration of the epoch in which population ABC has k distinct lineages. Then t_k is an independent exponential random variable with parameter $k(k-1)/2$. Conditional on R_1C_0 , the set $\mathcal{M}$ has two distinct elements, T_1 and T_2 , where T_i is the triplet with leaves A_i, B , and C for $i = 1, 2$. Hence $w_i := |I(T_i)|$ is uniformly distributed on $[0, 1]$ for $i = 1, 2$ and $w_1 + w_2 = 1$. Since recombination is assumed only to occur in population A , it follows that $d_{BC}^{T_1} = d_{BC}^{T_2}$, and we shall denote this quantity by d_{BC} .

There are 18 labeled histories $\gamma_1, \dots, \gamma_{18}$ in which the four labeled lineages A_1, A_2, B , and C may coalesce. Since pairs of lineages are chosen to coalesce uniformly at random, $\mathbb{P}[\gamma_j|R_1C_0] = \frac{1}{18}$ for all j , and hence

$$\mathbb{P}[F|R_1C_0] - \mathbb{P}[E|R_1C_0] = \frac{1}{18} \sum_{j=1}^{18} (\mathbb{P}[F|R_1C_0\gamma_j] - \mathbb{P}[E|R_1C_0\gamma_j]). \quad (19)$$

The task will therefore be to calculate $\mathbb{P}[F|R_1C_0\gamma_j] - \mathbb{P}[E|R_1C_0\gamma_j]$ for $j = 1, \dots, 18$. To do this we group together labeled histories that involve similar calculations into five claims.

Claim 1. Let $\gamma_1 = ((A_1A_2)B)C$, $\gamma_2 = ((A_1B)A_2)C$, and $\gamma_3 = ((A_2B)A_1)C$. Then

$$\mathbb{P}[F|R_1C_0\gamma_j] - \mathbb{P}[E|R_1C_0\gamma_j] = -1 \quad \text{for } j = 1, 2, 3.$$

Proof. We prove the case for $j = 1$ as the other two cases are similar. The labeled history corresponding to γ_1 is depicted in Figure 9. Since C is the last lineage to coalesce, $d_{AB}^{T_1} < 2\theta(t_{\text{root}} + t_4 + t_3 + t_2) = d_{AC}^{T_1} = d_{BC}^{T_1}$ for $i = 1, 2$. Letting $\phi(x) := \frac{3}{4} \left(1 - e^{-\frac{4}{3}x}\right)$, it follows that

$$\begin{aligned} \Delta_{AB} &= w\phi\left(d_{AB}^{T_1}\right) + (1-w)\phi\left(d_{AB}^{T_2}\right) \\ \Delta_{AC} &= w\phi\left(d_{AC}^{T_1}\right) + (1-w)\phi\left(d_{AC}^{T_2}\right) \\ \Delta_{BC} &= \phi(d_{BC}). \end{aligned} \quad (20)$$

Therefore, since ϕ is increasing, $\Delta_{AB} < \Delta_{AC} \wedge \Delta_{BC}$ a.s., hence $\mathbb{P}[E|R_1C_0\gamma_1] = 1$. This proves the claim. $\square$

Claim 2. Let $\gamma_4 = ((A_1A_2)C)B$, $\gamma_5 = ((A_1C)A_2)B$, and $\gamma_6 = ((A_2C)A_1)B$. Then

$$\mathbb{P}[F|R_1C_0\gamma_j] - \mathbb{P}[E|R_1C_0\gamma_j] = 0 \quad \text{for } j = 4, 5, 6.$$

Proof. The proof is similar to that of Claim 1. Since B is the last lineage to coalesce, $d_{AC}^{T_i} < 2\theta(t_{\text{root}} + t_4 + t_3 + t_2) = d_{AB}^{T_i} = d_{BC}^{T_i}$ for $i = 1, 2$. Therefore $\Delta_{AC} < \Delta_{AB} \wedge \Delta_{BC}$ almost surely, so that $\mathbb{P}[E|R_1C_0\gamma_j] = \mathbb{P}[F|R_1C_0\gamma_j] = 0$. This proves the claim. $\square$

Figure 9: The labeled histories corresponding to Claims 1 (left) and 4 (right).

Claim 3. Let $\gamma_7 = (A_1 A_2)BC$, $\gamma_8 = (BC)A_1 A_2$, $\gamma_9 = ((BC)A_1)A_2$, and $\gamma_{10} = ((BC)A_2)A_1$. Then

$$\mathbb{P}[F|R_1 C_0 \gamma_j] - \mathbb{P}[E|R_1 C_0 \gamma_j] = 1 \quad \text{for } j = 7, 8, 9, 10.$$

Proof. It suffices to show $\Delta_{BC} < \Delta_{AB} \wedge \Delta_{AC}$ a.s., conditional on $R_1 C_0 \gamma_j$. Therefore by (20) it suffices to show that $d_{BC} < d_{AB}^{T_i} \wedge d_{AC}^{T_i}$ almost surely for $i = 1, 2$. Indeed, if $j = 7, 8$ then $d_{BC} \leq 2\theta(t_{\text{root}} + t_4 + t_3) < 2\theta(t_{\text{root}} + t_4 + t_3 + t_2) = d_{AB}^{T_i} \wedge d_{AC}^{T_i}$, and if $j = 9, 10$ then $d_{BC} = 2\theta(t_{\text{root}} + t_4) < 2\theta(t_{\text{root}} + t_4 + t_2) = d_{AB}^{T_i} \wedge d_{AC}^{T_i}$. This proves the claim. $\square$

Claim 4. Let $\gamma_{11}, \gamma_{12}, \gamma_{13}, \gamma_{14}$ be $((A_1 C)B)A_2, ((A_2 C)B)A_1, ((A_1 B)C)A_2$ and $((A_2 B)C)A_1$ respectively. Then

$$\sum_{j=11}^{14} \mathbb{P}[F|R_1 C_0 \gamma_j] - \mathbb{P}[E|R_1 C_0 \gamma_j] = 6\alpha - 2$$

where

$$\alpha = 3 \int_0^\infty \int_0^\infty \frac{1 - e^{-\frac{8}{3}\theta x}}{e^{\frac{8}{3}\theta y} - e^{-\frac{8}{3}\theta x}} e^{-3y-x} dx dy. \quad (21)$$

Proof. Consider the general case $((UX)Y)V$ where $\{U, V\} = \{A_1, A_2\}$ and $\{X, Y\} = \{B, C\}$. This corresponds to the labeled history depicted in Figure 9

Let $u = \begin{cases} w_1 & : U = A_1 \\ 1 - w_1 & : U = A_2 \end{cases}$, it follows that

$$\begin{aligned} \Delta_{AX} &= u\phi(2\theta(t_{\text{root}} + t_4)) + (1 - u)\phi(2\theta(t_{\text{root}} + t_4 + t_3 + t_2)) \\ \Delta_{AY} &= u\phi(2\theta(t_{\text{root}} + t_4 + t_3)) + (1 - u)\phi(2\theta(t_{\text{root}} + t_4 + t_3 + t_2)) \\ \Delta_{BC} &= \Delta_{XY} = \phi(2\theta(t_{\text{root}} + t_4 + t_3)). \end{aligned} \quad (22)$$

Since ϕ is increasing, $\Delta_{BC} < \Delta_{AY}$ almost surely, and hence

$$\mathbb{P}[E_{AY|X}|R_1 C_0 \gamma_j] = 0. \quad (23)$$

Therefore there exists some real number α with $0 \leq \alpha \leq 1$ such that

$$\mathbb{P}[E_{BC|A}|R_1 C_0 \gamma_j] = \alpha \quad \text{and} \quad \mathbb{P}[E_{AX|Y}|R_1 C_0 \gamma_j] = 1 - \alpha. \quad (24)$$

Using (22) and (24),

$$\begin{aligned} \alpha &= \mathbb{P}[\Delta_{BC} < \Delta_{AX}|R_1 C_0 \gamma_j] \\ &= \mathbb{P}\left[1 - e^{-\frac{8}{3}\theta(t_{\text{root}} + t_4 + t_3)} < u(1 - e^{-\frac{8}{3}\theta(t_{\text{root}} + t_4)}) + (1 - u)(1 - e^{-\frac{8}{3}\theta(t_{\text{root}} + t_4 + t_3 + t_2)})|R_1 C_0 \gamma_j\right] \\ &= \mathbb{P}\left[1 > ue^{\frac{8}{3}\theta t_3} + (1 - u)e^{-\frac{8}{3}\theta t_2}|R_1 C_0 \gamma_j\right]. \end{aligned}$$

Conditional on $R_1 C_0$, the random variables t_2, t_3 are exponentially distributed with rates 1 and 3 respectively and are independent of γ_j and each other. Conditional on $R_1 C_0 \gamma_j$, u is uniformly distributed on $(0, 1)$ for any j . Therefore

$$\begin{aligned}\alpha &= \int_0^1 \int_0^\infty \int_0^\infty \mathbf{1}\left[1 > w_1 e^{\frac{8}{3}\theta y} + (1 - w_1) e^{-\frac{8}{3}\theta x}\right] e^{-x} dx \ 3e^{-3y} dy \ dw_1 \\ &= 3 \int_0^\infty \int_0^\infty \int_0^1 \mathbf{1}\left[w_1 < \frac{1 - e^{-\frac{8}{3}\theta x}}{e^{\frac{8}{3}\theta y} - e^{-\frac{8}{3}\theta x}}\right] dw_1 \ e^{-3y-x} dx \ dy \\ &= 3 \int_0^\infty \int_0^\infty \frac{1 - e^{-\frac{8}{3}\theta x}}{e^{\frac{8}{3}\theta y} - e^{-\frac{8}{3}\theta x}} e^{-3y-x} dx \ dy.\end{aligned}$$

This shows that the value of α does not depend on j . Therefore by (23) and (24),

$$\mathbb{P}[E|R_1 C_0 \gamma_j] = \begin{cases} 1 - \alpha & \text{if } j = 11, 12 \\ 0 & \text{if } j = 13, 14 \end{cases} \quad (25)$$

and by (24),

$$\mathbb{P}[F|R_1 C_0 \gamma_j] = \alpha \quad \text{for } j = 11, 12, 13, 14. \quad (26)$$

The statement of the claim then follows from (25) and (26). This proves the claim. $\square$

Claim 5. Let $\gamma_{15}, \gamma_{16}, \gamma_{17}$ and γ_{18} be $(A_1 B)A_2 C$, $(A_2 B)A_1 C$, $(A_1 C)A_2 B$, and $(A_2 C)A_1 B$ respectively. Then

$$\sum_{j=15}^{18} \mathbb{P}[F|R_1 C_0 \gamma_j] - \mathbb{P}[E|R_1 C_0 \gamma_j] = -2. \quad (27)$$

Proof. To prove this, we consider the general case $(UX)VY$, where $\{U, V\} = \{A_1, A_2\}$ and $\{X, Y\} = \{B, C\}$. This corresponds to the following labeled history:

Letting $u = \begin{cases} w_1 & : U = A_1 \\ 1 - w_1 & : U = A_2 \end{cases}$, it follows that

$$\begin{aligned}\Delta_{AX} &= u\phi(2\theta(t_{\text{root}} + t_4)) + (1 - u)\phi(2\theta(t_{\text{root}} + t_4 + t_3 + t_2)) \\ \Delta_{AY} &= u\phi(2\theta(t_{\text{root}} + t_4 + t_3 + t_2)) + (1 - u)\phi(2\theta(t_{\text{root}} + t_4 + t_3)) \\ \Delta_{BC} &= \Delta_{XY} = \phi(2\theta(t_{\text{root}} + t_4 + t_3 + t_2)).\end{aligned} \quad (28)$$

Since $\Delta_{BC} > \Delta_{AX}$, it follows that

$$\mathbb{P}[F|R_1 C_0 \gamma_j] = \mathbb{P}[E_{BC|A}|R_1 C_0 \gamma_j] = 0 \quad (29)$$

whenever $15 \leq j \leq 18$. Therefore

$$\mathbb{P}[E_{AX|Y}|R_1 C_0 \gamma_j] = \beta \quad \text{and} \quad \mathbb{P}[E_{AY|X}|R_1 C_0 \gamma_j] = 1 - \beta. \quad (30)$$

where $\beta = \mathbb{P} \left[d_{AX}^{(\ell)} < d_{AY}^{(\ell)} | R_1 C_0 \gamma_j \right]$. By (28),

$$\beta = \mathbb{P} \left[u + (1 - 2u)e^{-\frac{8}{3}\theta(t_3+t_2)} - (1 - u)e^{-\frac{8}{3}\theta t_3} > 0 | R_1 C_0 \gamma_j \right].$$

Moreover, similarly as in Claim 4, conditional on $R_1 C_0$, the random variables t_2, t_3 are exponentially distributed with rates 1 and 3 respectively and are independent of γ_j and each other, and conditional on $R_1 C_0 \gamma_j$, u is uniformly distributed on $(0, 1)$ for any j . Therefore

$$\beta = \mathbb{P} \left[\hat{u} + (1 - 2\hat{u})e^{-\frac{8}{3}\theta(\hat{t}_3+\hat{t}_2)} - (1 - \hat{u})e^{-\frac{8}{3}\theta\hat{t}_3} > 0 \right]$$

where $\hat{u} \sim \text{unif}(0, 1)$, $\hat{t}_2 \sim \text{exp}(1)$ and $\hat{t}_3 \sim \text{exp}(3)$, with $\hat{u}, \hat{t}_2, \hat{t}_3$ independent. In particular, this implies β does not depend on j . Moreover, equation (30) implies $\mathbb{P}[E | R_1 C_0 \gamma_{15}] = \mathbb{P}[E | R_1 C_0 \gamma_{16}] = \beta$ and $\mathbb{P}[E | R_1 C_0 \gamma_{17}] = \mathbb{P}[E | R_1 C_0 \gamma_{18}] = 1 - \beta$. Therefore $\sum_{j=15}^{18} \mathbb{P}[E | R_1 C_0 \gamma_j] = 2$. Combining this with (29) implies equation (27). This proves the claim. $\square$

We now return to the proof of the lemma. Combining the results from Claims 1-5 with (19) yields

$$\mathbb{P}[F | R_1 C_0] - \mathbb{P}[E | R_1 C_0] = \frac{1}{18} (6\alpha - 3) = \frac{\alpha}{3} - \frac{1}{6}$$

and substituting the formula for α from (21) gives

$$\mathbb{P}[F | R_1 C_0] - \mathbb{P}[E | R_1 C_0] = \int_0^\infty \int_0^\infty \frac{1 - e^{-8\theta x/3}}{e^{8\theta y/3} - e^{-8\theta x/3}} e^{-3y-x} dx dy - \frac{1}{6}. \quad (31)$$

It will suffice to show that the right hand side of (31) is a decreasing function of θ and equals zero when $\theta = 3/4$, as this will imply that it is positive for all $0 < \theta < 3/4$. By the substitution $c = 8\theta/3$, it suffices to show that the function

$$\Psi(c) := \int_0^\infty \int_0^\infty \frac{1 - e^{-cx}}{e^{cy} - e^{-cx}} e^{-3y-x} dx dy$$

satisfies $\Psi(2) = 1/6$ and is strictly decreasing. Numerical plots of Ψ are given in Figure 10. Making the substitution $u = e^{-(x+y)}$, $du = -e^{-(x+y)} dx$,

$$\Psi(2) = \int_0^\infty \int_0^\infty \frac{1 - e^{-2x}}{e^{2y} - e^{-2x}} e^{-3y-x} dx dy = \int_0^\infty \int_0^{e^{-y}} \frac{e^{-4y}}{1 - u^2} - \frac{u^2 e^{-2y}}{1 - u^2} du dy,$$

and by the Fubini-Tonelli theorem,

$$\begin{aligned} \Psi(2) &= \int_0^1 \frac{1}{1 - u^2} \int_0^{-\log u} e^{-4y} dy - \frac{u^2}{1 - u^2} \int_0^{-\log u} e^{-2y} dy du \\ &= \int_0^1 \frac{1}{1 - u^2} \left[\frac{1}{4} (1 - u^4) \right] - \frac{u^2}{1 - u^2} \left[\frac{1}{2} (1 - u^2) \right] du dy \\ &= \int_0^1 \frac{1}{4} (1 + u^2) - \frac{1}{2} u^2 du = \frac{1}{6}. \end{aligned}$$

It remains to show that $\Psi(c) > \Psi(2)$ for all $0 < c < 2$. Fix $x, y > 0$ and let $\psi(c) = \frac{1 - e^{-cx}}{e^{cy} - e^{-cx}} e^{-3y-x}$. Differentiating ψ gives

$$\psi'(c) = \frac{(x + y)e^{c(y-x)} - ye^{cy} - xe^{-cx}}{(e^{cy} - e^{-cx})^2} = \frac{e^{c(y-x)}}{(e^{cy} - e^{-cx})^2} [y(1 - e^{cx}) + x(1 - e^{-cy})].$$

Since $1 - e^a < -a$ for all $a \neq 0$, $y(1 - e^{cx}) + x(1 - e^{-cy}) < y(-cx) + x(cy) = 0$, it follows that $\psi'(c) < 0$ for all $c > 0$ and hence Ψ is strictly decreasing on $(0, \infty)$, as required. $\square$

Figure 10: Numerical plots of the function $\Psi(c)$, shown in blue. Note that Ψ is decreasing and $\Psi(2) = 1/6$.

C Inconsistency of Unrooted Quartet Majority

Our main result extends to the unrooted case. For $W, X, Y, Z \in L$, the unrooted quartet topology inferred by the *four point method* is $WZ|YZ$ if $\delta_{WX} + \delta_{YZ} < (\delta_{WY} + \delta_{XZ}) \wedge (\delta_{WZ} + \delta_{XY})$. Let S be an ultrametric species phylogeny with $L = \{A, B, C, D\}$ and topology $((AB)C)D$. Then the unrooted quartet topology is $AB|CD$, but we will show that it may not be the most likely unrooted quartet topology to be inferred by the four point method when the data M_k is generated according to the MSCR-JC(k) process on S for k sufficiently large.

Specifically, take S be the tree from the proof of Theorem 1. We may append a distantly related fourth leaf D to an edge above the root as depicted in Figure 1b. Then the resulting tree has unrooted quartet topology $AB|CD$. However the residual bias in favor of the $BC|A$ rooted triple rather than $AB|C$ in subtree S causes the four point method to infer the unrooted quartet topology $BC|AD$ more often than $AB|CD$. This idea motivates the following corollary.

Corollary 1. *For k sufficiently large, there exists a species phylogeny S with four leaves such that the most likely unrooted quartet topology to be inferred using the four-point method from MSAs generated under the MSCR-JC(k) model is different from the unrooted quartet topology of S .*

Proof. Let q_k be the unrooted quartet topology inferred from M_k by the four-point method. We wish to show that $\mathbb{P}[q_k = AB|CD] < \mathbb{P}[q_k = AD|BC]$ for k sufficiently large. For each quartet $W, X, Y, Z \in L$, let

$$\mathcal{E}_{WX|YZ} = [\Delta_{WX} + \Delta_{YZ} < (\Delta_{WY} + \Delta_{XZ}) \wedge (\Delta_{WZ} + \Delta_{XY})],$$

so that $\lim_{k \rightarrow \infty} \mathbb{P}[q_k = WX|YZ] = \mathbb{P}[\mathcal{E}_{WX|YZ}]$ by Lemma 1. Therefore it suffices to show that $\mathbb{P}[\mathcal{E}_{AB|CD}] < \mathbb{P}[\mathcal{E}_{AD|BC}]$.

Let $\epsilon = \mathbb{P}[F] - \mathbb{P}[E]$, where $E = [\Delta_{AB} < \Delta_{AC} \wedge \Delta_{BC}]$ and $F = [\Delta_{BC} < \Delta_{AB} \wedge \Delta_{AC}]$. Since the restricted subtree $S[\{A, B, C\}]$ satisfies the assumptions of Theorem 1, it follows that for k sufficiently large there exists a choice of parameters ρ_X and $\theta = \theta_X$ for all $X \in \{A, B, C, AB, ABC\}$, along with τ_A, τ_{AB} such that $\epsilon > 0$ for all τ_{ABC} sufficiently large. See Figure 1b.

Further assign $\rho_{ABC} = \rho_D = 0$ and $\theta_{ABCD} = \theta$. Since S is ultrametric, $\tau_D = \tau_A + \tau_{AB} + \tau_{ABC}$, which is the age of the root of S . The only parameter of S which remains to be chosen is τ_{ABC} . Let z be the age of the MRCA on the ancestral recombination graph of the lineages originating in populations A, B , and C . Since $\tau_D \rightarrow \infty$ as $\tau_{ABC} \rightarrow \infty$,

$$\lim_{\tau_{ABC} \rightarrow \infty} \mathbb{P}[X|z < \tau_D] = \lim_{\tau_{ABC} \rightarrow \infty} \frac{\mathbb{P}[X, z < \tau_D]}{\mathbb{P}[z < \tau_D]} = \mathbb{P}[X]$$

for $X \in \{E, F\}$. Therefore we may choose $\tau_{ABC} > 0$ sufficiently large that

$$\mathbb{P}[F|z < \tau_D] - \mathbb{P}[E|z < \tau_D] > \frac{\epsilon}{2} \quad \text{and} \quad \mathbb{P}[z < \tau_D] \geq \frac{2}{2 + \epsilon}. \quad (32)$$

If $z < \tau_D$, then since S is ultrametric and $\theta_X = \theta$ for all X , it holds that for all $T_g \in \mathcal{M}$, $d_{AD}^{T_g} = d_{BD}^{T_g} = d_{CD}^{T_g} = 2h$ where h is the height of the ancestral recombination graph. Therefore $\Delta_{AD} = \Delta_{BD} = \Delta_{CD}$. Using this identity, it is easy to check that, conditional on $z < \tau_D$, the events E and $\mathcal{E}_{AB|CD}$ are equivalent, and the events F and $\mathcal{E}_{AC|BC}$ are equivalent. Therefore by the Law of Total Probability and (32),

$$\begin{aligned} \mathbb{P}[\mathcal{E}_{AD|BC}] - \mathbb{P}[\mathcal{E}_{AB|CD}] &= (\mathbb{P}[\mathcal{E}_{AD|BC}|z < \tau_D] - \mathbb{P}[\mathcal{E}_{AB|CD}|z < \tau_D]) \mathbb{P}[z < \tau_D] \\ &\quad + (\mathbb{P}[\mathcal{E}_{AD|BC}|z \geq \tau_D] - \mathbb{P}[\mathcal{E}_{AB|CD}|z \geq \tau_D]) \mathbb{P}[z \geq \tau_D] \\ &\geq (\mathbb{P}[F|z < \tau_D] - \mathbb{P}[E|z < \tau_D]) \mathbb{P}[z < \tau_D] - \mathbb{P}[z \geq \tau_D] \\ &> \frac{\epsilon}{2} \left(\frac{2}{2 + \epsilon} \right) - \left(1 - \frac{2}{2 + \epsilon} \right) = 0. \end{aligned}$$

□