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

Max Hill(✉) and Sebastien Roch

Department of Mathematics, University of Wisconsin–Madison, Madison, USA
bacharach@wisc.edu

Abstract. We consider species tree estimation from multiple loci subject to intralocus recombination. We focus on R^*, a summary coalescent-based method 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.

Keywords: Phylogenetics · Species tree estimation · Intralocus recombination

1 Introduction

Species tree estimation from genomic data is complicated by various biological phenomena which generate phylogenetic conflict, among them hybridization, horizontal gene transfer, gene duplication and loss, and incomplete lineage sorting (ILS) [23]. In particular, ILS may cause phylogenetic conflict in which a gene tree exhibits a different topology from that of the species tree, and is of greatest concern for species trees with short internal branches [23]. 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 [5,7,8] (see also [2,21] 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, quartets, such as R^*, MP-EST, BUCKy, ASTRAL, and others [6,15,16,20]. Some of these methods are based on the fact that rooted triples and unrooted quartets are special cases in which no anomaly zone exists [5,13] and also provide sufficient information to reconstruct

MH was 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).

I. Pe'er (Ed.): RECOMB 2022, LNBI 13278, pp. 143–158, 2022.
https://doi.org/10.1007/978-3-031-04749-7_9

the full phylogeny [24,27]. Provided that the gene trees are estimated without error, such methods can provide statistically consistent methods of estimating species tree topology [30].

A common assumption of coalescent-based models based on the multispecies coalescent (MSC) [21,22] 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 [2,9]. The significance of the latter assumption—that is, the impact of intralocus recombination on phylogenetic inference—is a matter of present interest [2,32] and much debate about its significance when unaccounted for [9,14,26]. One justification for assuming no intralocus recombination is that within-gene recombination may break gene function [23].

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 [14]. On the other hand, the authors of [25] suggest the absence of intralocus recombination may be an unreasonable assumption in real data, such as protein-coding genes in eukaryotes [2,19], and particularly in the case of species phylogenies with many taxa [26]. In particular, the potential for intralocus recombination to distort gene tree frequencies has been recognized as a challenge to summary coalescent-based methods, and [14] has been critiqued for its focus on shallow divergences and limitation to a low number of loci and taxa [26].

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 $m \to \infty$. We find that the effect arises when differential rates of recombination are exhibited between closely-related taxa.

1.1 Key Definitions

A *species phylogeny* $S = (V_S, E_S; r, \bar{\rho}, \bar{\tau}, \bar{\theta})$ is a directed binary tree with vertex set V_S, edge set E_S, root $r \in V_S$, and n labeled leaves $L_S = [n]$, such that each edge $e \in E_S$ 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. 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); 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.

We define a *rooted triple* to be 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 [24]. 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 [27]. The R^* pipeline has three steps: first, for each gene, infer a rooted triple for each triplet of leaves $X, Y, Z \in L_S$. 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 [6].

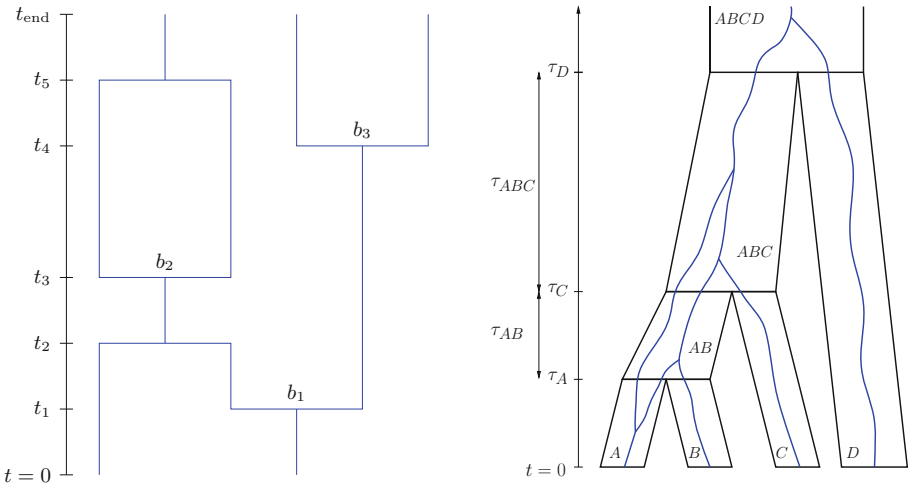
Since we are interested in the inference of the species-tree topology from *sequence data*, we consider a distance-based approach in which a species triplet with leaves X, Y, Z is inferred to have topology $XY|Z$ if

$$\delta_{XY} < \delta_{XZ} \wedge \delta_{YZ}. \tag{1}$$

where $\delta_{XY} = \delta_{XY}(M_k)$ is the number of mismatching nucleotides between sequences $\mathbf{s}_X$ and $\mathbf{s}_Y$ $(X, Y \in L_S)$. We refer to this inference procedure as $\boldsymbol{R^*}$ **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 [10] (see also [1]) within the framework of the multi-species coalescent (MSC) [8,21,22]. In the single-population ARG [10], ancestors are represented by edges in the graph (see Fig. 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



(a) A single-population ARG

(b) A multi-species ARG

Fig. 1. Two depictions of an ARG, in a single population (left) and in the multispecies case (right). In Fig. 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_2 occurred at times t_1, t_3, and t_4. In Fig. 1b, the lineages of a multispecies ARG are shown in blue within a 4-taxa species tree S (the thick tree) with fixed edge lengths $\tau_A, \tau_B, \ldots, \tau_{ABC}$. (Color figure online)

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 Fig. 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 $\boldsymbol{S}$**. 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 collection of rooted edge-weighted binary trees; a simple example is shown in Fig. 2. The set of marginal gene trees $\mathcal{M} := \{\mathcal{T}(x) : 0 \leq x \leq 1\}$ is almost surely finite [10]. 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.

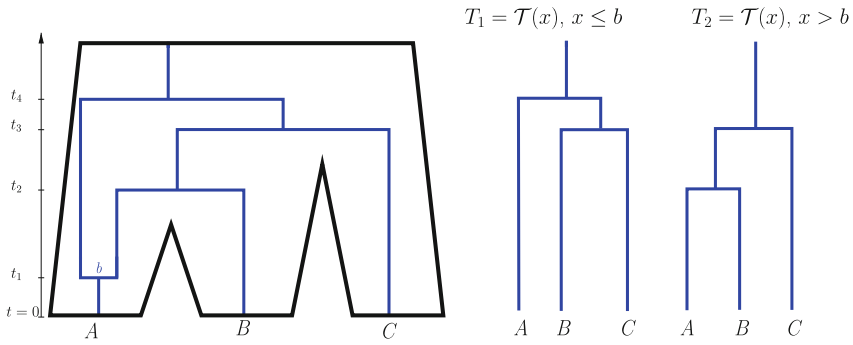


Fig. 2. On the left, an ancestral recombination graph (in blue) is shown within a 3-taxa tree S (in black). The times of coalescence and recombination events are labeled $t_1, \ldots t_4$ on the time axis, and the breakpoint associated with the recombination event is labeled $b \in [0, 1]$. On the right, the corresponding marginal gene trees T_1 and T_2 are shown. This particular example also illustrates how intralocus recombination may contribute to phylogenetic conflict by allowing for 'partial' ILS, whereby one or more of the marginal gene trees (in this case T_1) exhibits a topology different from that of S.

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 [12,27] on the tree $\mathcal{T}(x)$. A somewhat more general description of this algorithm can be found in [4].

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

In words, the MSCR-JC(k) process models the evolution of n homologous genes situated at a common genetic locus consisting of k sites, and which may have experienced intralocus recombination; these homologous genes are assumed to have been drawn from n distinct species whose true species phylogeny is represented by S. The resulting homologous aligned DNA sequences are the rows of the $n \times k$ matrix M_k. The phylogenetic reconstruction problem considered here pertains to whether the topology of S can be recovered from sequence data generated in this manner, or more precisely:

Problem: Let S be a species phylogeny with leaf labels $L_S = [n]$. Fix $k \geq 2$. Given m independent samples $M_k^{(1)}, \ldots, M_k^{(m)}$, 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 Fig. 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

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

This formula, due to [28], 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 $k \to \infty$.*

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.*

To prove Theorem 1, it suffices to consider a species tree S with $L_S = \{A, B, C\}$ and topology $AB|C$. Denote edges of S, or *populations*, by the letters A, B, C, AB, and ABC as depicted in Fig. 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 Fig. 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

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

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 $m \to \infty$.

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 $XY|Z \in \mathcal{R}(S)$,*

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

Here $\mathcal{R}(S) = \{S|J : J \subseteq L_S, |J| = 3,$ and $S|J$ is binary$\}$ is the set of restricted rooted triples of S (see [24]).

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 $\frac{1}{k}\delta_{XY}(M_k) \to \Delta_{XY}$ as $k \to \infty$. 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 $k \to \infty$ for all $X, Y \in L_S$. Therefore, since the distribution function of Δ_{XY} is continuous, $\mathbb{P}[E_{XY|Z}] \to \mathbb{P}[E]$ and $\mathbb{P}[E_{YZ|X}] \to \mathbb{P}[F]$ as $k \to \infty$, 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

$$\mathbb{P}[F] > \mathbb{P}[E]. \tag{5}$$

We detail the proof next.

2.2 Key Lemmas

In what follows, set intersection is denoted with product notation (i.e. so that $XY = X \cap Y$ for events X, Y) and the important events to be considered are

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

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 shows that τ_A can be chosen sufficiently small that with high probability, zero or one recombination occurs.

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

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

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.

Lemma 5 (Effect of Small Internal Edge). *As* $\tau_{AB} \to 0^+$, $\mathbb{P}[C_0|R_1] = K + O(\tau_{AB})$, $\mathbb{P}[C_{0,A}|R_1C_1] = O(\tau_{AB})$, *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* $\lim_{\tau_A \to 0} K = 1$ *for any fixed* $\rho > 0$.

Next we apply Lemma 5 to show that $\mathbb{P}[E|R_1C_1] - \mathbb{P}[F|R_1C_1]$ is small, tending to zero as $\tau_{AB} \to 0^+$.

Lemma 6. $\mathbb{P}[E|R_1C_1] - \mathbb{P}[F|R_1C_1] = O(\tau_{AB})$ *as* $\tau_{AB} \to 0^+$,

We now come to a key part of the calculation: the event R_1C_0, depicted in Fig. 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. 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 Fig. 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

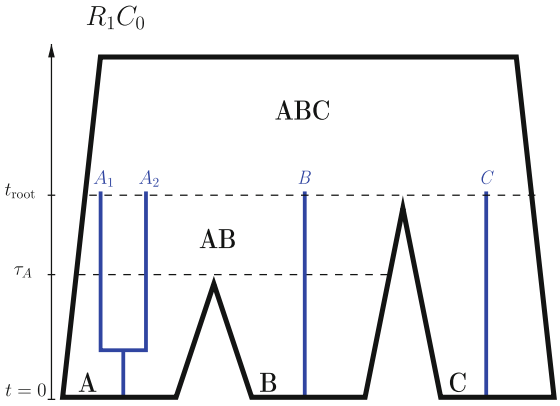


Fig. 3. A depiction of the event R_1C_0. The portion of the ancestral recombination graph more ancient than t_{root} is not shown.

to age [21]), 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 $\gamma_1, \ldots, \gamma_{18}$. 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

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

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 $j = 1, \ldots, 18$ 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. □

The next lemma applies Lemmas 5, 6, and 7 to show that $P[F|R_1] > \mathbb{P}[E|R_1]$ 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 $\tau_{AB} \to 0^+$ (where the term $-\bar{\alpha}K$ does not depend on τ_{AB}).*

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,

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

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 $K \to 1$ as $\tau_A \to 0$ 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. □

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 [14], 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 [11,17] and effective eukaryotic population sizes N_e range from 10^4 to 10^8 [18], 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 $\theta \to 0$, 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 $m = 10^5$ replicates were generated independently under each parameter regime, with the aim of estimating how frequently the correct topology was inferred. The parameters used were $\theta = 0.1$, $\tau_{AB} \in \{0.01, 0.02, \ldots, 0.15\}$, $\rho_A \in \{0, \ldots, 20\}$, and $\rho_X =$ 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 Fig. 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.

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 [31]. A plot almost identical to Fig. 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 Figs. 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 Figs. 4 and 5 show, this effect disappears when τ_{AB} is increased

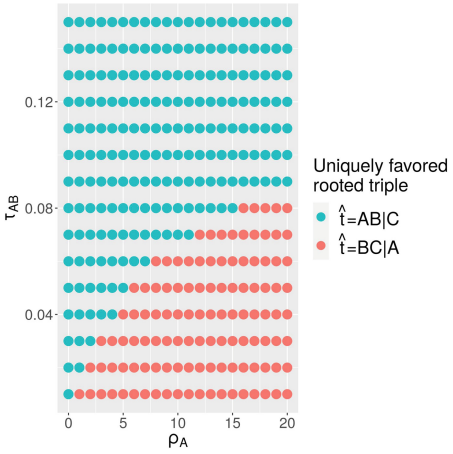


Fig. 4. R^* inconsistency zone. The color of each dot represents a simulation of $m = 10^5$ replicates.

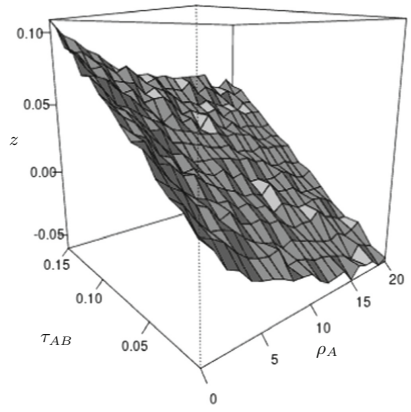


Fig. 5. The surface $z = \hat{p}_{AB|C} - \hat{p}_{BC|A}$ as a function of τ_{AB} and ρ_A.

(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.

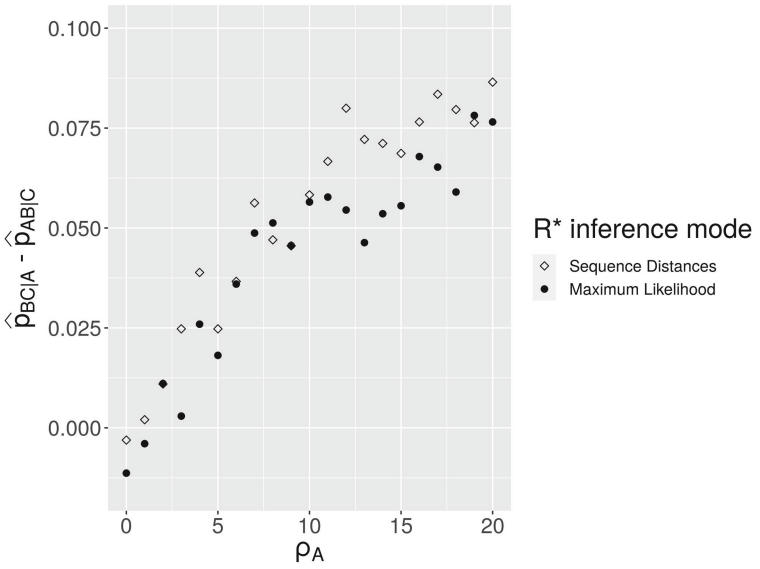


Fig. 6. The effect of increasing ρ_A on inference using R^* with sequence distances and maximum likelihood.

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 $\rho := \rho_A = \rho_B = \rho_C = \rho_{AB} \in \{0, 1, \ldots, 20\}$ 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 Fig. 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,

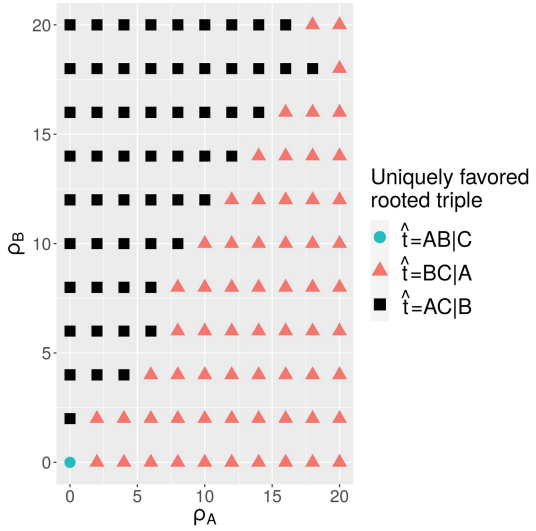


Fig. 7. R^* inference with recombination in both populations A and B.

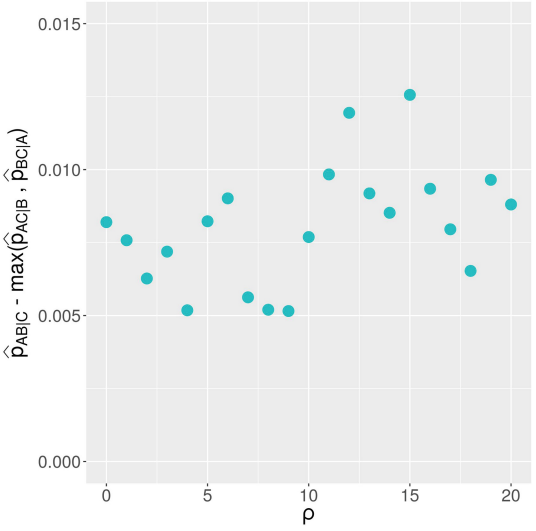


Fig. 8. Equal recombination rates in A, B, C and AB.

$\hat{p}_{AB|C} > \hat{p}_{AC|B} \vee \hat{p}_{BC|A}$, suggesting consistent inference despite the very short internal branch length, a result which agrees with the conclusions of [14] that even high recombination rates are not a significant source of error, at least when rates are comparable across species. Thus, the existence of 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 [3,14,29], 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 [14] 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 does 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.

Acknowledgements. 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.

References

1. Arenas, M.: The importance and application of the ancestral recombination graph. Front. Genet. **4**, 206 (2013). https://doi.org/10.3389/fgene.2013.00206
2. Bryant, D., Hahn, M.W.: The concatenation question. In: Scornavacca, C., Delsuc, F., Galtier, N. (eds.) Phylogenetics in the Genomic Era, Chap. 3.4, pp. 3.4:1–3.4:23 (2020). No commercial publisher—Authors open access book. https://hal.inria.fr/PGE
3. Conry, M.: Determining the impact of recombination on phylogenetic inference. Ph.D. thesis, The Florida State University (2020)
4. Dasarathy, G., Mossel, E., Nowak, R., Roch, S.: Coalescent-based species tree estimation: a stochastic Farris transform. arXiv preprint arXiv:1707.04300 (2017)
5. Degnan, J.H.: Anomalous unrooted gene trees. Syst. Biol. **62**(4), 574–590 (2013). https://doi.org/10.1093/sysbio/syt023
6. Degnan, J.H., DeGiorgio, M., Bryant, D., Rosenberg, N.A.: Properties of consensus methods for inferring species trees from gene trees. Syst. Biol. **58**(1), 35–54 (2009). https://doi.org/10.1093/sysbio/syp008
7. Degnan, J.H., Rosenberg, N.A.: Discordance of species trees with their most likely gene trees. PLoS Genet. **2**(5), e68 (2006). https://doi.org/10.1371/journal.pgen.0020068
8. Degnan, J.H., Rosenberg, N.A.: Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. **24**(6), 332–340 (2009). https://doi.org/10.1016/j.tree.2009.01.009
9. Edwards, S.V., et al.: Implementing and testing the multispecies coalescent model: a valuable paradigm for phylogenomics. Mol. Phylogene. Evol. **94**, 447–462 (2016). https://doi.org/10.1016/j.ympev.2015.10.027
10. Griffiths, R.C., Marjoram, P.: An ancestral recombination graph. In: Donnelly, P., Tavaré, S. (eds.) Progress in Population Genetics and Human Evolution. vol. 87, p. 257. Springer New York (1997)
11. Hahn, M.W.: Molecular Population Genetics. Oxford University Press, Oxford (2018)
12. Jukes, T.H., Cantor, C.R.: Evolution of protein molecules. Mammalian Protein Metab. **3**, 21–132 (1969). https://doi.org/10.1016/B978-1-4832-3211-9.50009-7
13. Kapli, P., Yang, Z., Telford, M.J.: Phylogenetic tree building in the genomic age. Nat. Rev. Gene. **21**(7), 428–444 (2020). https://doi.org/10.1038/s41576-020-0233-0
14. Lanier, H.C., Knowles, L.L.: Is recombination a problem for species-tree analyses? Syst. Biol. **61**(4), 691–701 (2012). https://doi.org/10.1093/sysbio/syr128
15. Larget, B.R., Kotha, S.K., Dewey, C.N., Ané, C.: BUCKy: Gene tree/species tree reconciliation with Bayesian concordance analysis. Bioinformatics **26**(22), 2910–2911 (2010). https://doi.org/10.1093/bioinformatics/btq539
16. Liu, L., Yu, L., Edwards, S.V.: A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evol. Biol. **10**(1), 1–18 (2010). https://doi.org/10.1186/1471-2148-10-302
17. Lynch, M.: Evolution of the mutation rate. TRENDS Genet. **26**(8), 345–352 (2010). https://doi.org/10.1016/j.tig.2010.05.003
18. Lynch, M., Marinov, G.K.: The bioenergetic costs of a gene. Proc. Nat. Acad. Sci. **112**(51), 15690–15695 (2015). https://doi.org/10.1073/pnas.1514974112
19. Mendes, F.K., Livera, A.P., Hahn, M.W.: The perils of intralocus recombination for inferences of molecular convergence. Philos. Trans. R. Soc. B **374**(1777), 20180244 (2019). https://doi.org/10.1098/rstb.2018.0244

20. Mirarab, S., Reaz, R., Bayzid, M.S., Zimmermann, T., Swenson, M.S., Warnow, T.: Astral: genome-scale coalescent-based species tree estimation. Bioinformatics **30**(17), i541–i548 (2014). https://doi.org/10.1093/bioinformatics/btu462
21. Rannala, B., Edwards, S.V., Leaché, A., Yang, Z.: The multi-species coalescent model and species tree inference. In: Scornavacca, C., Delsuc, F., Galtier, N. (eds.) Phylogenetics in the Genomic Era, Chap. 3.3, pp. 3.3:1–3.3:21 (2020). No commercial publisher—Authors open access book. https://hal.inria.fr/PGE
22. Rannala, B., Yang, Z.: Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics **164**(4), 1645–1656 (2003). https://doi.org/10.1093/genetics/164.4.1645
23. Schrempf, D., Szöllősi, G.: The sources of phylogenetic conflicts. In: Scornavacca, C., Delsuc, F., Galtier, N. (eds.) Phylogenetics in the Genomic Era, Chap. 3.1, pp. 3.1:1–3.1:23 (2020). No commercial publisher—Authors open access book. https://hal.inria.fr/PGE
24. Semple, C., Steel, M., et al.: Phylogenetics, vol. 24. Oxford University Press on Demand, London(2003)
25. Springer, M.S., Gatesy, J.: The gene tree delusion. Mol. Phylogene. Evol. **94**, 1–33 (2016). https://doi.org/10.1016/j.ympev.2015.07.018
26. Springer, M.S., Gatesy, J.: Delimiting coalescence genes (c-genes) in phylogenomic data sets. Genes **9**(3), 123 (2018). https://doi.org/10.3390/genes9030123
27. Steel, M.: Phylogeny: Discrete and Random Processes in Evolution. SIAM (2016)
28. Wang, K.C.: Phylogenetic reconstruction accuracy in the face of heterogeneity, recombination, and reticulate evolution. The University of Wisconsin-Madison (2017)
29. Wang, Z., Liu, K.J.: A performance study of the impact of recombination on species tree analysis. BMC genomics **17**(10), 165–174 (2016). https://doi.org/10.1186/s12864-016-3104-5
30. Warnow, T.: Computational Phylogenetics: An Introduction to Designing Methods for Phylogeny Estimation. Cambridge University Press, New York (2017)
31. Yang, Z.: Complexity of the simplest phylogenetic estimation problem. Proc. R. Soc. Lond. Ser. B Biol. Sci. **267**(1439), 109–116 (2000). https://doi.org/10.1098/rspb.2000.0974
32. Zhu, T., Yang, Z.: Complexity of the simplest species tree problem. Mol. Biol. Evol. **38**(9), 3993–4009 (2021). https://doi.org/10.1093/molbev/msab009