Statistically Consistent Rooting of Species Trees under the Multi-Species Coalescent Model

Yasamin Tabatabaee[1[0000−0002−7811−5989]], Sebastien Roch[2[0000−0002−7608−8550]], and Tandy Warnow[1[0000−0001−7717−3514]]

[1] Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, USA
[2] Department of Mathematics, University of Wisconsin–Madison, Madison, USA
{warnow}@illinois.edu

Abstract. Rooted species trees are used in several downstream applications of phylogenetics. Most species tree estimation methods produce unrooted trees and additional methods are then used to root these unrooted trees. Recently, Quintet Rooting (QR) (Tabatabaee et al., ISMB and Bioinformatics 2022), a polynomial-time method for rooting an unrooted species tree given unrooted gene trees, was introduced. QR, which is based on a proof of identifiability of rooted 5-taxon trees in the presence of incomplete lineage sorting, was shown to have good accuracy, improving over other methods for rooting species trees when incomplete lineage sorting was the only cause of gene tree discordance, except when gene tree estimation error was very high. However, that study left the statistical consistency of QR as an open question. We present QR-STAR, a polynomial-time variant of QR that has an additional step for determining the rooted shape of each quintet tree. We prove that QR-STAR is statistically consistent under the multi-species coalescent (MSC) model. Our simulation study under a variety of model conditions shows that QR-STAR matches or improves on the accuracy of QR. QR-STAR is available in open source form at https://github.com/ytabatabaee/Quintet-Rooting

Keywords: Species Tree Estimation · Rooting · Statistical Consistency · Multi-Species Coalescent

1 Introduction

Inferring rooted species trees is important for many downstream applications of phylogenetics, such as comparative genomics [10,6] and dating [24]. These estimations use different loci from across the genomes of the selected species, and so are referred to as multi-locus analyses. If rooted gene trees can be accurately inferred, then the rooted species tree can be estimated from them [13]; however, this is not a reliable assumption [28]. Hence, the standard approach is to first estimate an unrooted species tree using multi-locus datasets, and then root that estimated tree.

The estimation of the unrooted species tree is challenged by biological processes, such as incomplete lineage sorting (ILS) or gene duplication and loss (GDL), that can result in different parts of the genome having different evolutionary trees. When ILS or GDL occur, statistically consistent estimation of the unrooted species tree requires techniques that take the source of heterogeneity into consideration [14,21]. The case of ILS, as modeled by the Multi-Species Coalescent (MSC) model [9], is the most well studied, and there are several methods for estimating unrooted species trees that have been proven statistically consistent under the MSC (see [21] for a survey of such methods).

The general problem of rooting a species tree (or indeed even a gene tree) is of independent interest, but presents many challenges. A common approach is the use of an outgroup taxon (i.e., the inclusion of a species that is outside the smallest clade containing the remaining species), so that the resultant tree is rooted on the edge leading to the outgroup [15]. However, outgroup selection has its own difficulties: if the outgroup is too distant, then it may be attached fairly randomly to the tree containing the remaining species, and if it is too close, it may even be an ingroup taxon [4,5,8,12]. Other approaches use branch lengths estimated on the tree to find the root based on specific optimization criteria; however, these approaches tend to degrade in accuracy unless the strict molecular clock holds (which assumes that all sites along the genome evolve under a constant rate) [7,17,30].

Quintet Rooting (QR) [29] is a recently introduced method that is designed to root a given species tree using the unrooted gene tree topologies, under the assumption that the gene trees can differ from the species tree due to ILS. QR is based on mathematical theory established by Allman, Degnan, and Rhodes [2], which showed that the rooted species tree topology is identifiable from the unrooted gene tree topologies whenever the number of species is at least five. In [29], QR was shown to provide good accuracy for rooting both estimated and true species trees in the presence of ILS, compared to alternative methods.

However, QR was not proven to be statistically consistent—which addresses the theoretical question of whether the root location selected by QR, when given the true species tree, will converge to the correct location as the number of gene trees in the input increases. Although much attention has been paid to establishing statistical consistency for unrooted species tree estimation methods and many methods, such as ASTRAL [18], SVDQuartets [31] and BUCKy [11] have been proven to be statistically consistent estimators of the unrooted species tree topology under the MSC, to our knowledge, no prior study has addressed the statistical consistency properties of methods for rooting species trees.

In this paper, we argue that QR is not guaranteed to be statistically consistent under the MSC, but we also present a modification to QR, which we call QR-STAR, that we prove statistically consistent. Moreover, QR-STAR, like QR, runs in polynomial time.

The rest of this paper is organized as follows: Section 2 provides background information on QR, including the theory established by Allman, Degnan, and Rhodes [2]. Section 3 introduces QR-STAR and Section 4 provides the proof of statistical consistency of QR-STAR under the MSC. In Section 5 we present the results of a simulation study exploring the parameter space of QR-STAR and evaluating its performance in comparison to QR. We conclude in Section 6. Due to space limitations, most of the proofs and results from the simulation study are presented in the Supplementary Materials.

2 Background

We present the theory from [2] first, which establishes identifiability of the rooted species tree from unrooted quintet trees, and then we describe Quintet Rooting (QR), our earlier method for rooting species trees. Together these form the basis for deriving our new method, QR-STAR, which we present in the next section.

2.1 Allman, Degnan, and Rhodes (ADR) Theory

Allman, Degnan, and Rhodes (ADR) [2] have shown that under the MSC, for four species, the unrooted topology of the species tree, and for five or more species, the rooted topology of the species tree is identifiable from the probability distribution of the unrooted topological gene trees. The first result has been the basis of several "quartet-based" species tree estimation methods such as BUCKy-pop [11], ASTRAL, wQFM [16] and QuCo [23], that can estimate the unrooted topology of the species tree. The second result however, has received less attention, but was recently used in the development of QR for rooting species trees.

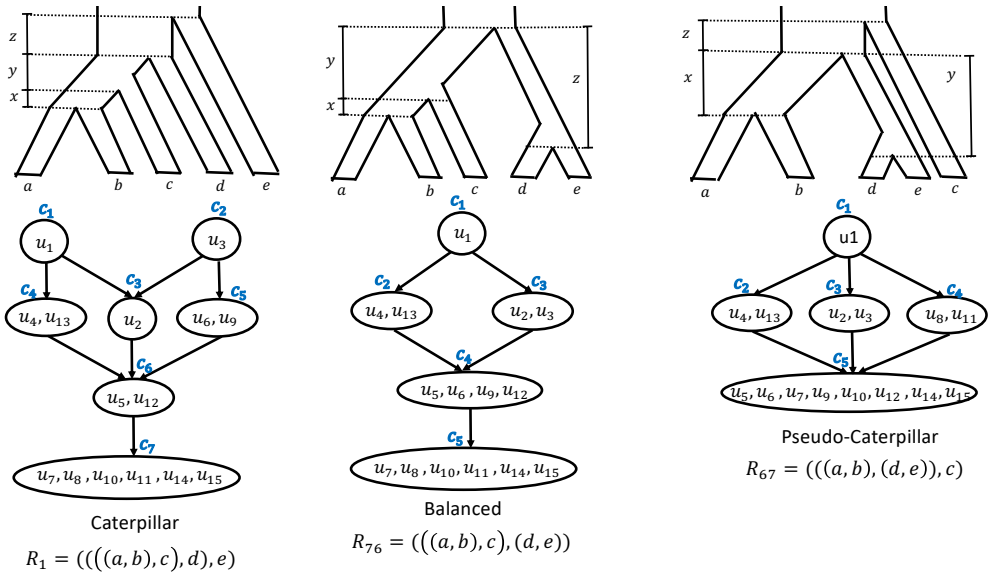

Fig. 1: **ADR invariants and inequalities for different rooted topological shapes.** The invariants and inequalities found by ADR define a partial order on the distribution of unrooted gene trees for rooted 5-taxon model species trees with different rooted shapes (caterpillar, balanced and pseudo-caterpillar). There are 15 unrooted binary trees on a given set of 5 leaves. Each of the 105 5-taxon rooted species trees define a specific distribution on the probabilities of these unrooted trees. The topology of the rooted binary species tree can be determined from this distribution (i.e., it is identifiable, as established by ADR). While the branch lengths of the rooted species tree depend on the actual probabilities, the linear invariants and inequalities that hold for these distributions is enough to determine the rooted topology of the model species tree.

ADR have described the probability distribution of unrooted topological gene trees under each 5-taxon MSC model species tree. On a given set of five taxa, there exist 105 different rooted binary trees, labeled with $R_1, \ldots, R_{105}$[3], that can be categorized into three groups based on their (unlabeled) rooted shapes: caterpillar, balanced and pseudo-caterpillar [26]. An example of a tree from each category is shown in Figure 1. Each 5-taxon model species tree defines a specific probability distribution over the 15 different unrooted topological gene trees on the same leafset, shown with $T_1, \ldots, T_{15}$. Theorem 9 in [2] states that this distribution uniquely determines the rooted tree topology and its internal branch lengths for trees with at least five taxa.

To prove this identifibaility result, the ADR theory specifies a set of linear invariants (i.e., equalities) and inequalities that must hold between the probabilities of unrooted 5-taxon gene trees, for any choice of the parameters of the model species tree. These linear invariants and inequalities define a partial order on the distribution of 5-taxon unrooted topological gene trees. In other words, two gene tree probabilities $u_i = \mathbb{P}(T_i)$ and $u_j = \mathbb{P}(T_j)$ can have one of four possible relationships: $u_i > u_j$, $u_j > u_i$, $u_i = u_j$ or u_i and u_j are not comparable. Figure 1 shows examples of these partial orders with a Hasse diagram for a particular

[3] The labeling of rooted and unrooted trees in this paper is consistent with the notations and leaf-labeling used in Tables 4-5 in [2] as well as in [29].

leaf labeling of trees from each rooted shape. In this paper, we denote the set of equivalence classes of a 5-taxon rooted tree R with C_R. As it can be seen in Figure 1, the number of equivalence classes of caterpillar, balanced and pseudo-caterpillar trees is 7, 5 and 5 respectively. Each directed edge between two equivalence classes in these Hasse diagrams defines an inequality, so that all gene tree probabilities in class c_a at the source of an edge are greater than all gene tree probabilities in class c_b at the target, and we show this by $c_a > c_b$. The exact values of the unrooted gene tree probabilities depend on the internal branch lengths of the model tree, and ADR provide a set of formulas that relate the model tree parameters to the probability distribution of the unrooted gene trees in Appendix B of [2], which will be used in our proofs.

2.2 Quintet Rooting

The input to QR is an unrooted species tree T with n leaves and a set $\mathcal{G}$ of single-copy unrooted gene trees where the gene trees draw their leaves from the leafset of T, denoted by $\mathcal{L}(T)$. Given this input, QR searches over all possible rootings of T and returns a tree most consistent with the distribution of quintets (i.e., 5-taxon trees) in the input gene trees.

QR approaches this problem by selecting a set Q of quintets of taxa from $\mathcal{L}(T)$ (called the "quintet sampling" step; refer to Appendix A for details), and scoring all rooted versions of T based on their induced trees on these quintets. The subtree $T_{|q}$, T restricted to taxa in quintet set q, can be rooted on any of its seven edges. In a preprocessing step, QR computes a score for each of these seven different rootings for all trees induced on the quintets in set Q, based on a cost function. This results in $7 \times |Q|$ computations, and therefore the preprocessing step takes $O(k(|Q| + n))$, where k is the number of input gene trees. Next, for every rooted version of T, QR sums up the costs of all its induced rooted trees on quintets in Q using the scores computed in the preprocessing step, and returns the rooting with the minimum overall cost. Since T can be rooted on any of its $2n - 3$ edges, the scoring step takes $O(n + |Q|)$ time. Therefore, the overall runtime of QR when using an $O(n)$ sampling of quintets is $O(nk)$. Figure 2 shows the pipeline of QR and its individual steps. Thus, QR provides an exact solution to the optimization problem with the following input and output:

- **Input**: An unrooted tree topology T, a set of k unrooted gene tree topologies $\mathcal{G} = \{g_1, g_2, \ldots, g_k\}$, a set Q containing quintets of taxa from leafset $\mathcal{L}(T)$ and a cost function $\mathrm{Cost}(r, \vec{u})$.
- **Output**: Rooted tree R with topology T such that $\sum_{q \in Q} \mathrm{Cost}(R|_q, \vec{\hat{u}}_q)$ is minimized, where $\vec{\hat{u}}_q$ is the distribution of unrooted gene tree quintets in $\mathcal{G}|_q = \{g_1|_q, g_2|_q, \ldots, g_k|_q\}$.

Cost Function The cost function $\mathrm{Cost}(R|_q, \vec{\hat{u}}_q)$ measures the fitness of the rooted quintet tree $R|_q$ with the distribution of the unrooted gene trees restricted to q (i.e., $\vec{\hat{u}}_q$), according to the linear invariants and inequalities derived from the ADR theory. In particular, this cost function is designed to penalize a rooted tree $R|_q$ if the estimated quintet distribution $\vec{\hat{u}}_q$ violates some of the inequalities or invariants in its partial order. To this end, a penalty term was considered for each invariant and inequality in the partial order of a 5-taxon rooted tree that is violated in a quintet distribution. The cost function was defined based on a linear combination of these penalty terms, and had the following form, where r is a 5-taxon rooted tree and $\vec{\hat{u}}$ is an estimated quintet distribution:

$$\mathrm{Cost}(r, \vec{\hat{u}}) = \underbrace{\sum_{c \in C_r} \frac{1}{|c|} \sum_{u_a, u_b \in c} |\hat{u}_a - \hat{u}_b|}_{\text{Invariants Penalty}} + \underbrace{\sum_{c > c' \in C_r} \frac{1}{|c'|} \sum_{u_a \in c, u_b \in c'} \max(0, \hat{u}_b - \hat{u}_a)}_{\text{Inequalities Penalty}}. \tag{1}$$

The normalization factors $\frac{1}{|c|}$ and $\frac{1}{|c'|}$ were used to reduce a topological bias that arose from differences in the sizes of the equivalence classes for each tree shape.

3 QR-STAR

QR-STAR is an extension to QR that has an additional step for determining the rooted shape of a quintet tree, as well as an associated penalty term in its cost function. This penalty term compares the rooted shape

of the 5-taxon tree, denoted by $S(r)$, with the rooted shape inferred by QR-STAR from the given quintet distribution, denoted by $\hat{S}(\hat{u})$. The motivation for this additional preprocessing step is that, as we argue in Appendix C, the cost function of QR does not guarantee statistical consistency. The cost function of QR-STAR takes the following general form

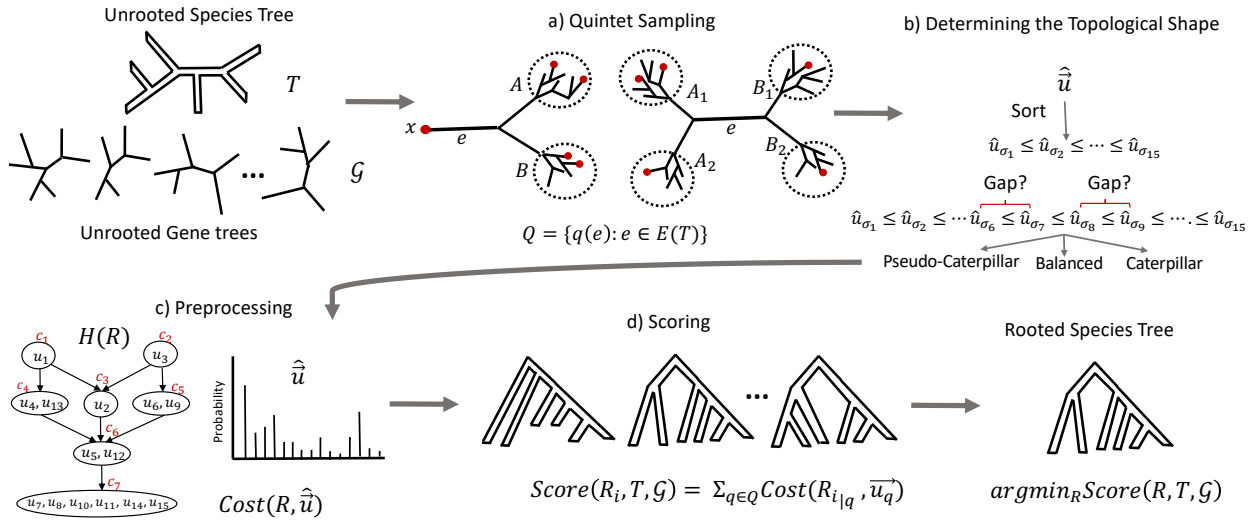

Fig. 2: **Pipeline of QR and QR-STAR.** The input is an unrooted species tree T and set of unrooted gene trees $\mathcal{G}$ on the same leafset. a) The sampling step selects a set Q of quintets from the leafset of T (shown is the linear encoding sampling). b) An additional step in QR-STAR that determines the rooted shape for each selected quintet. c) The preprocessing step computes a cost for each of the seven possible rootings of each selected quintet. d) The scoring step computes a score for each rooted tree in the search space based on the costs computed in the preprocessing step, and returns a rooting of T with minimum score.

$$\text{Cost}^*(r, \vec{\hat{u}}) = \underbrace{\sum_{c \in C_r} \sum_{u_a, u_b \in c} \alpha_{a,b} |\hat{u}_a - \hat{u}_b|}_{\text{Invariants Penalty}} + \underbrace{\sum_{c > c' \in C_r} \sum_{u_a \in c, u_b \in c'} \beta_{a,b} \max(0, \hat{u}_b - \hat{u}_a)}_{\text{Inequalities Penalty}} + \underbrace{C \mathbb{1} |S(r) \neq \hat{S}(\hat{u})|}_{\text{Shape Penalty}} \quad (2)$$

where $\alpha_{a,b} \geq 0$ and $\beta_{a,b}, C > 0$ are constant real numbers for all a, b [4]. Let $\alpha_{\max} = \max_{a,b}(\alpha_{a,b})$ and $\beta_{\min} = \min_{a,b}(\beta_{a,b})$ where a, b ranges over all pairs of indices a, b used in the penalty terms in Eq. 2.

Each of the 105 rooted binary trees on a given set of 5 leaves have a unique set of inequalities and invariants that can be derived from the ADR theory. The cost function in Eq. 2 considers a penalty term for these inequalities and invariants as well as the shape of the tree, so that $\text{Cost}^*(r, \vec{\hat{u}})$ is minimized for a rooted 5-taxon tree r that best describes the given estimated quintet distribution.

3.1 Determining the Rooted Shape

Model 5-taxon species trees with different rooted shapes (i.e., caterpillar, balanced, pseudo-caterpillar) define equivalence classes with different class sizes on the unrooted gene tree probability distribution. These class sizes can be used to determine the unlabeled shape of a rooted tree, when given the *true* gene tree probability distribution. For example, the size of the equivalence class with the smallest gene tree probabilities is 8 for the pseudo-caterpillar trees and 6 for balanced or caterpillar trees. Therefore, the size of the equivalence class corresponding to the minimal element in the partial order can differentiate a pseudo-caterpillar tree from other tree shapes. Moreover, both caterpillar and balanced trees have a unique class with the second

[4] Refer to Remark 1 for an explanation on why $\alpha_{a,b}$ does not need to be strictly positive.

smallest probability, which is of size 2 for caterpillar trees and 4 for balanced trees and this can be used to differentiate a caterpillar tree from a balanced tree. This approach is used in Theorem 9 in [2] for establishing the identifiability of rooted 5-taxon trees from unrooted gene trees.

However, given an *estimated* gene tree distribution, it is likely that none of the invariants derived from the ADR theory exactly hold, and so the class sizes cannot be directly determined and the approach above cannot be used as is to infer the shape of a rooted quintet. Here we propose a simple modification for determining the rooted shape of a tree from the estimated distribution of unrooted gene trees, by looking for significant gaps between quintet gene tree probabilities.

Let T be the unrooted species tree with $n \geq 5$ leaves given to QR-STAR and q be a quintet of taxa from $\mathcal{L}(T)$. Let $\vec{\hat{u}}$ be the quintet distribution estimated from input gene trees induced on taxa in set q. QR-STAR first sorts $\vec{\hat{u}}$ in ascending order to get $\hat{u}_{\sigma_1} \leq \hat{u}_{\sigma_2} \leq \cdots \leq \hat{u}_{\sigma_{15}}$. Let $A(k) = \sqrt{\frac{2}{k}\ln(30|Q|k)}$ (refer to Lemma 4 for the derivation of $A(k)$) where k is the number of input gene trees and $|Q|$ is the size of the set of sampled quintets in QR-STAR (it depends on the number of taxa n and is assumed fixed), and note that $\lim_{k\to\infty} A(k) = 0$. The first step of QR-STAR computes an estimate of the rooted shape of a quintet q, denoted by $\hat{S}(\hat{u})$ in Eq. 2, as follows:

- estimate the rooted shape $\hat{S}(\hat{u})$ as pseudo-caterpillar if $\hat{u}_{\sigma_7} - \hat{u}_{\sigma_6} < A(k)$;
- estimate the rooted shape $\hat{S}(\hat{u})$ as balanced if $\hat{u}_{\sigma_7} - \hat{u}_{\sigma_6} \geq A(k)$ and $\hat{u}_{\sigma_9} - \hat{u}_{\sigma_8} < A(k)$;
- estimate the rooted shape $\hat{S}(\hat{u})$ as caterpillar if $\hat{u}_{\sigma_7} - \hat{u}_{\sigma_6} \geq A(k)$ and $\hat{u}_{\sigma_9} - \hat{u}_{\sigma_8} \geq A(k)$.

The runtime of QR-STAR is the same as QR, as determining the topological shape for each quintet is done in constant time, and the overall runtime remains $O(nk)$, when a linear sampling of quintets is used.

4 Theoretical Results

In this section, we provide the main theoretical results, starting with a series of lemmas and theorems that will be used in the proof of statistical consistency of QR-STAR in Theorem 2. Throughout this paper, we assume that discordance between species trees and gene trees is solely due to ILS. In establishing statistical consistency, we assume that input gene trees are true gene trees and, thus, have no gene tree estimation error. If not otherwise specified, all trees are assumed to be fully resolved (i.e., binary). Due to space constraints, most of the proofs are provided in Appendix B. We begin with some definitions and key observations.

Definition 1 (Path length parameter). *Let R be an MSC model species tree. Let $f(R)$ be the length of the shortest internal branch of R and $g(R)$ be the length of the longest internal path (i.e., a path formed from only the internal branches) of R. We define the path length parameter of R as*

$$h(R) = \frac{1}{18}e^{-3g(R)}(1 - e^{-f(R)})^2 \tag{3}$$

Note that $h(R) \in (0, \frac{1}{18})$ since $\exp(-x) \in (0, 1)$ for all $x > 0$ and the branch lengths have positive values. The formula for Eq. 3 is derived from the proof of Lemma 2 in Appendix B.

Lemma 1. *Let R be an MSC model species tree with $n \geq 5$ leaves and q be an arbitrary set of 5 leaves from $\mathcal{L}(R)$. Then $h(R|_q) \geq h(R)$ where $R|_q$ is the rooted tree R restricted to taxa in set q.*

Lemma 2. *Let R be an MSC model species tree with 5 leaves and internal branch lengths x, y, and z. Let $\vec{u}$ be the probability distribution that it defines on the unrooted 5-taxon topological gene trees. If $\vec{\hat{u}}$ is an estimate of $\vec{u}$ such that given $\epsilon > 0$, we have $|\hat{u}_i - u_i| < \epsilon$ for all $1 \leq i \leq 15$, the following inequality holds:*

$$\forall_{c>c' \in C_R} \forall_{u_a \in c, u_b \in c'} : \hat{u}_a - \hat{u}_b > h(R) - 2\epsilon. \tag{4}$$

Definition 2. *For a 5-taxon rooted tree R, we define I_R as the set of ordered pairs (i, j), $1 \leq i \neq j \leq 15$, corresponding to inequalities in the form $u_i > u_j$ defined according to the partial order of R. The inequalities that are a result of transitivity (i.e. $u_i > u_j$ and $u_j > u_k$ implies $u_i > u_k$) are not included in I_R.*

Definition 3. *Let $V(R, R')$ be the set of violated inequalities of two rooted 5-taxon trees R and R', i.e., all pairs $\{i, j\}$ such that $(i, j) \in I_R$ and $(j, i) \in I_{R'}$.*

Figure 3 a shows an example of $V(R, R')$ computed for caterpillar trees and Figure 3 b is a heatmap showing the function $|V(R, R')|$ computed for the seven possible rootings of an unrooted quintet tree. The set $V(R, R')$ can be easily computed from I_R and $I_{R'}$ for all pairs of rooted 5-taxon trees, and I_R is derived from the ADR theory for all 105 5-taxon rooted trees in the Supplementary Materials, Sec. S2 in [29].

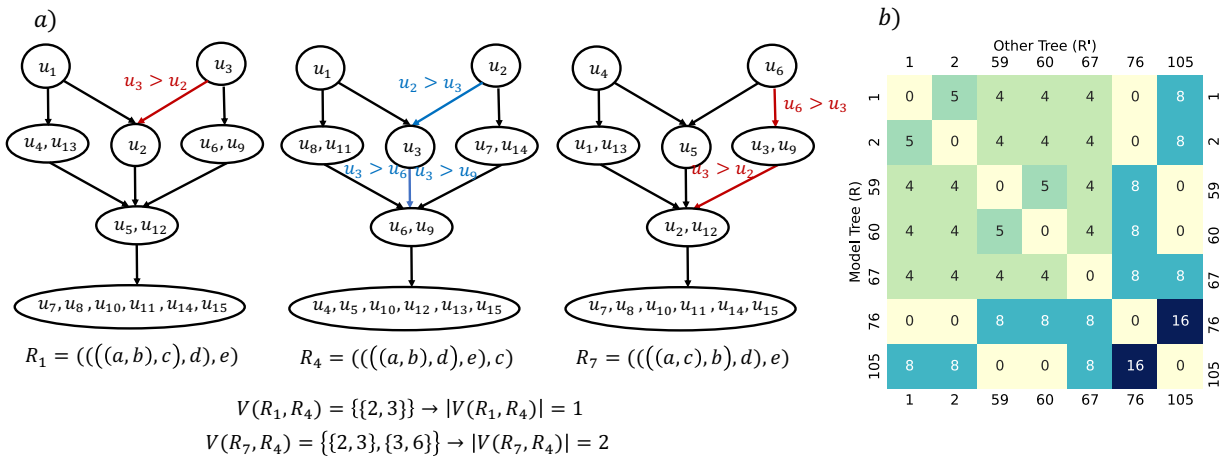

Fig. 3: **Conflicting inequality penalty terms between rooted 5-taxon species trees.** a) Set of violated inequality penalty terms in the partial orders of R_1 and R_7 with respect to R_4, which are all caterpillar trees. The red edges show violations of inequalities in tree R_4, highlighted in blue. b) Heatmap showing the number of pairwise violated penalty terms (function $|V(R, R')|$) of seven possible rooted trees having unrooted topology with bipartitions $ab|cde$ and $abc|de$. The dark colors indicate more violations, and the lightest color corresponds to no violations ($|V(R, R')| = 0$).

Lemma 3. *(a) For 5-taxon binary rooted trees R and R' with the same rooted shape, the set $V(R, R')$ is always non-empty. (b) For each balanced tree B, there exist two caterpillar trees C_1 and C_2 such that $V(B, C_i) = \emptyset$.*

4.1 Statistical Consistency

In this section, we establish statistical consistency for QR-STAR under the MSC and provide the sufficient condition for a set of sampled quintets to lead to consistency. That is, we prove that as the number of input true gene trees increases, the probability that QR-STAR and its variants correctly root the given unrooted species tree converges to 1. We first prove statistical consistency for QR-STAR when the model tree has only five taxa in Theorem 1 and then extend the proofs to trees with arbitrary numbers of taxa in Theorem 2. The main idea of the proof of consistency for 5-taxon trees is that we show as the number of input gene trees increases, the cost of the true rooted tree becomes arbitrarily close to zero, but the cost of any other rooted tree is bounded away from zero, where the bound depends on the path length parameter of the model tree, $h(R)$. Our proofs imply sample complexity bounds. For lack of space, we leave them for the journal version.

Lemma 4. *Let R be an MSC model species tree with $n \geq 5$ leaves and Q be a set of quintets of taxa from $\mathcal{L}(R)$. Given $\delta > 0$ and $k > 0$ unrooted topological gene trees, the following inequality holds, where $A_\delta(k) = \sqrt{\frac{2}{k}\ln(\frac{30|Q|}{\delta})}$*

$$\mathbb{P}\left(\forall_{q \in Q} \forall_{1 \leq i \leq 15} |(\hat{u}_q)_i - (u_q)_i| < \frac{A_\delta(k)}{2}\right) \geq 1 - \delta. \tag{5}$$

Setting $\delta = \frac{1}{k}$ in Eq. 5, we get $A(k) = \sqrt{\frac{2}{k}\ln(30|Q|k)}$, which is the bound that is used for determining the rooted shape of each quintet in the first step of QR-STAR as well as the proofs of statistical consistency. When R has only five taxa, $A(k)$ becomes $\sqrt{\frac{2}{k}\ln(30k)}$, as Q can only contain one quintet.

Lemma 5 (Correct determination of rooted shape). *Let R be a 5-taxon model species tree and $\vec{u}$ be the probability distribution that it defines on the unrooted 5-taxon topological gene trees. There is an integer $k > 0$ such that if we are given at least k unrooted gene trees drawn i.i.d. from the distribution $\vec{u}$, the first step of QR-STAR will correctly determine the rooted shape of R with probability at least $1 - \frac{1}{k}$.*

Lemma 6 (Upper bound on the cost of the model tree). *Let R be a 5-taxon model species tree and $\vec{u}$ be the probability distribution that it defines on the unrooted 5-taxon topological gene trees. There is an integer $k > 0$ such that if we are given at least k unrooted gene trees drawn i.i.d. from distribution $\vec{u}$, then $\mathrm{Cost}^*(R, \hat{\vec{u}})$ is less than $31\alpha_{\max}A(k)$ with probability at least $1 - \frac{1}{k}$.*

Theorem 1 (Statistical Consistency of QR-STAR for 5-taxon trees). *Let R be a 5-taxon model species tree and $\vec{u}$ be the distribution that it defines on the unrooted 5-taxon topological gene trees. Given a set $\mathcal{G}$ of unrooted true quintet gene trees drawn i.i.d. from $\vec{u}$, QR-STAR is a statistically consistent estimator of R under the MSC.*

Proof. According to Lemma 6, when k is large enough so that $A(k) < \frac{1}{2}h(R)$, if $\hat{\vec{u}}$ is the distribution estimated from $\mathcal{G}$, then $\mathrm{Cost}^*(R, \hat{\vec{u}})$ is at most $31\alpha_{\max}A(k)$ with probability at least $1 - \frac{1}{k}$. We now prove that for every other rooted 5-taxon tree R', $\mathrm{Cost}^*(R', \hat{\vec{u}})$ is bounded away from zero. Note that according to Lemma 3, for every rooted 5-taxon tree $R' \neq R$ with the same rooted shape as R, we have $V(R, R') \neq \emptyset$ and therefore, there exists $1 \leq x \neq y \leq 15$ such that $(x, y) \in I_R, (y, x) \in I_{R'}$.

Let $\mathcal{E}$ be the event that $|\hat{u}_i - u_i| < \frac{A(k)}{2}$ for all $1 \leq i \leq 15$. According to Lemma 2, when $\mathcal{E}$ holds, then $\hat{u}_x - \hat{u}_y > h(R) - A(k)$ as $(x, y) \in I_R$. However, since $(y, x) \in I_{R'}$, an inequality penalty term in the form of $\max(0, \hat{u}_x - \hat{u}_y)$ is added to $\mathrm{Cost}^*(R', \hat{\vec{u}})$ when R' has the same shape as R. Moreover, according to Lemma 5, when $\mathcal{E}$ holds and $A(k) < \frac{1}{2}h(R)$, the first step of QR-STAR correctly determines the rooted shape of R. Therefore, if R' has a different rooted shape than R, then the penalty $\mathbb{1}|S(R') \neq \hat{S}(\hat{u})|$ becomes 1 and a positive cost C is added to the cost of R'. Therefore, both cases (i.e. whether R' has a different topology from R or not) lead to a positive penalty in the cost function for R', that is bounded away from zero. Hence,

$$\begin{aligned}\mathrm{Cost}^*(R', \hat{\vec{u}}) &= \underbrace{\sum_{c \in C_{R'}} \sum_{u_a, u_b \in c} \alpha_{a,b}|\hat{u}_a - \hat{u}_b|}_{\text{Invariants Penalty}} + \underbrace{\sum_{c > c' \in C_{R'}} \sum_{u_a \in c, u_b \in c'} \beta_{a,b} \max(0, \hat{u}_b - \hat{u}_a)}_{\text{Inequalities Penalty}} + \underbrace{C\mathbb{1}|S(R') \neq \hat{S}(\hat{u})|}_{\text{Shape Penalty}} \\ &\Rightarrow \forall_{R' \neq R}\, \mathrm{Cost}^*(R', \hat{\vec{u}}) \geq \min(\beta_{\min}(\hat{u}_x - \hat{u}_y), C) > \min\left(\beta_{\min}\left(h(R) - A(k)\right), C\right)\end{aligned} \tag{6}$$

Eq. 6 defines a lower bound for the cost of any tree other than the true rooted tree and Lemma 6 gives an upper bound for the cost of the true tree, both with respect to the estimated quintet distribution $\hat{\vec{u}}$. Therefore, when k is large enough so that

$$A(k) < \min\left(\frac{C}{31\alpha_{\max}}, \frac{h(R)}{31\frac{\alpha_{\max}}{\beta_{\min}} + 1}, \frac{1}{2}h(R)\right) \tag{7}$$

we will have

$$\mathrm{Cost}^*(R, \hat{\vec{u}}) < 31\alpha_{\max}A(k) < \min\left(\beta_{\min}\left(h(R) - A(k)\right), C\right) < \mathrm{Cost}^*(R', \hat{\vec{u}})$$

which means that the cost of the true rooted tree will be less than the cost of any other rooted tree on the same leafset with probability at least $1 - \frac{1}{k}$, when k is sufficiently large. Precisely, $\forall_{R' \neq R}\, \mathrm{Cost}^*(R, \hat{\vec{u}}) < \mathrm{Cost}^*(R', \hat{\vec{u}})$ when Eq. 7 holds, where $\beta_{\min}, C, h(R) > 0$ and $\alpha_{\max} \geq 0$ are constants. As a result, QR-STAR will return the true rooted species tree topology with probability converging to 1 as the number of gene trees grows large, proving the statistical consistency for 5-leaf trees. □

Remark 1. Note that when $\alpha_{\max} = 0$, meaning that the invariant penalty terms are removed from the cost function, the cost of the true tree would become exactly zero according to the proof of Lemma 6, and the cost of any other tree would be positive when k is large enough so that the conditions of Theorem 1 holds. Hence in this case, the condition in Eq. 7 will reduce to $A(k) < \frac{1}{2}h(R)$.

Remark 2. Note that Lemma 3(a) holds for *all* pairs of 5-taxon rooted trees with the same rooted shape and with different permutations of the leaf-labeling, regardless of whether they have the same unrooted topology or not. Due to this property, it is possible to differentiate all pairs of 5-taxon rooted trees in a statistically consistent manner with the cost function of QR-STAR without prior knowledge about the unrooted tree topology, and hence Theorem 1 does not assume that the unrooted topology is given as input.

The next lemma and theorem extend the proof of statistical consistency to trees with $n > 5$ taxa. Linear encodings of T are defined in Appendix A.

Lemma 7 (Identifiability of the root from the linear encoding). *Let R and R' be rooted trees with unrooted topology T and distinct roots. Let $Q_{LE}(T)$ be the set of quintets of leaves in a linear encoding of T. There is at least one quintet of taxa $q \in Q_{LE}(T)$ so that $R_{|q}$ and $R'_{|q}$ have different rooted topologies.*

Lemma 7 states that no two distinct rooted trees with topology T induce the same set of rooted quintet trees on quintets of taxa in set $Q_{LE}(T)$. Clearly, the same is true for any superset $Q_{LE}(T) \subseteq Q$ including the set Q_5 of all quintets of taxa on the leafset of T. There might also be other quintet sets that are not a superset of $Q_{LE}(T)$, but have the property that no two rooted versions of T define the same set of rooted quintets on their elements. We generalize the proof of consistency to all set of sampled quintets with this property.

Definition 4. *Let T be an unrooted tree and Q be a set of quintets of taxa from $\mathcal{L}(T)$. We say Q is "root-identifying" if every rooted tree R with topology T is identifiable from T and the set of rooted quintet trees in $\{R|_q : q \in Q\}$, i.e., no two rooted trees with topology T induce the same set of rooted quintet trees on Q.*

Theorem 2 (Statistical Consistency of QR-STAR). *Let R be an MSC model species tree with $n \geq 5$ leaves and let T denote its unrooted topology. Given T and a set $\mathcal{G}$ of unrooted true gene trees on the leafset $\mathcal{L}(T)$, QR-STAR is a statistically consistent estimator of the rooted version of T under the MSC, if the set of sampled quintets Q is root-identifying.*

Proof. QR-STAR computes the score of each rooted tree R with topology T as $\sum_{q \in Q} \text{Cost}^*(R|_q, \hat{\vec{u}}_q)$. Let $\mathcal{E}_q$ be the event that $|(\hat{u}_q)_i - (u_q)_i| < \frac{A(k)}{2}$ for all $1 \leq i \leq 15$ for a quintet $q \in Q$. According to Theorem 1, when $\mathcal{E}_q$ holds and k is large enough so that $A(k) < \min\left(\frac{C}{31\alpha_{\max}}, \frac{h(R)}{31\frac{\alpha_{\max}}{\beta_{\min}}+1}, \frac{1}{2}h(R)\right)$ (that guarantees $A(k) < \min\left(\frac{C}{31\alpha_{\max}}, \frac{h(R|_q)}{31\frac{\alpha_{\max}}{\beta_{\min}}+1}, \frac{1}{2}h(R|_q)\right)$ according to Lemma 1), $\text{Cost}(R|_q, \hat{\vec{u}}_q)$ will be less than the cost of each of the six alternative rooted trees on the five taxa in q with high probability, as $R|_q$ is the *true* rooting of the unrooted quintet tree $T|_q$.

According to Lemma 4, $\mathcal{E}_q$ simultaneously holds for all $q \in Q$ with probability at least $1 - \frac{1}{k}$. In this event, for every other rooted tree $R' \neq R$ and each $q \in Q$, we have $\text{Cost}^*(R'|_q, \hat{\vec{u}}_q) \geq \text{Cost}^*(R|_q, \hat{\vec{u}}_q)$, according to Theorem 1. This means that for every rooted tree R' with topology T, $\sum_{q \in Q} \text{Cost}^*(R'|_q, \hat{\vec{u}}_q) \geq \sum_{q \in Q} \text{Cost}^*(R|_q, \hat{\vec{u}}_q)$. Also, according to Definition 4, Q has the property that no other rooted tree R' induces the exact same set of rooted quintet trees as R on Q. Hence, there exists a $q^* \in Q$ such that $\text{Cost}^*(R'|_{q^*}, \hat{\vec{u}}_{q^*}) > \text{Cost}^*(R|_{q^*}, \hat{\vec{u}}_{q^*})$, as the cost of $R|_{q^*}$ is strictly less than the cost of each of the six alternative rooted trees on q^*, when the conditions in Theorem 1 hold. Therefore, the function $\sum_{q \in Q} \text{Cost}^*(r|_q, \hat{\vec{u}}_q)$ obtains its *unique* minimum for the rooted tree R. As a result, QR-STAR returns the true rooted topology of T with probability converging to 1 as the number of input gene trees increases and this proves the statistical consistency for trees with an arbitrary number of taxa. □

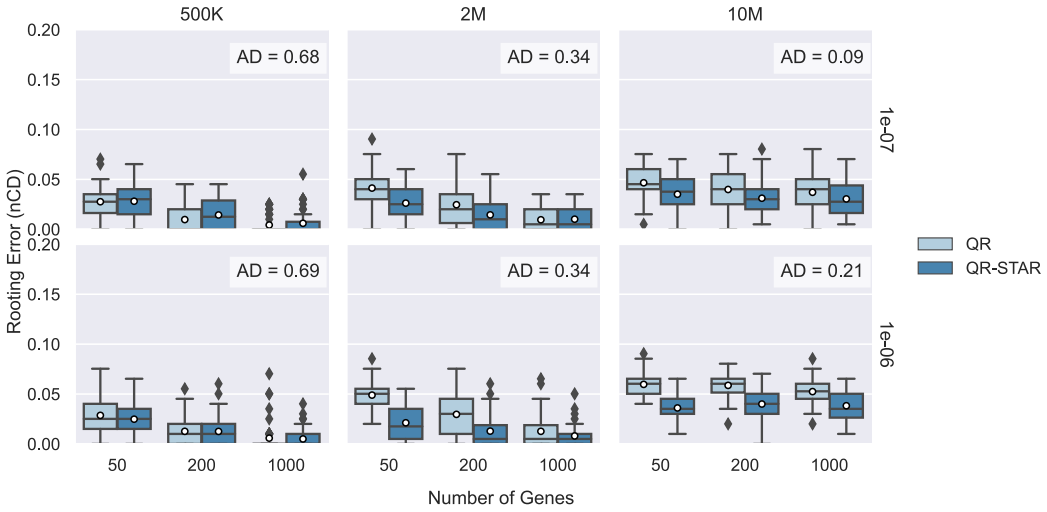

Fig. 4: **Rooting the model species tree with estimated gene trees on S200 datasets.** Comparison between QR and QR-STAR in terms of rooting error (nCD) for rooting the true unrooted species tree topology using estimated gene trees (0.35 mean GTEE averaged over six conditions) on the 201-taxon datasets from [19] with 50 replicates in each model condition. The columns show tree height (500K or high ILS, 2M or moderate ILS, and 10M or low ILS), and the rows show speciation rate (1e-06 or 1e-07).

5 Experimental Study

We performed an experimental study on simulated datasets to explore the parameter space of QR-STAR on a training data, and then compared its accuracy to QR on a test dataset. We used the 101-taxon simulated datasets from [32] as our training data, that had model conditions characterized by four levels of gene tree estimation error (GTEE) ranging from 0.23 to 0.55 (measured in terms of normalized Robinson-Foulds (RF) [25] distance between true and estimated gene trees) for 1000 genes. The normalized RF error between the model species tree and true gene trees (denoted average distance, or AD) in this dataset was 0.46, that indicates moderate ILS. We used a set of 201-taxon simulated datasets from [19] as our test data; these are characterized by two different speciation rates and three tree heights (thus six tree shapes), and three number of genes for each tree shape. The AD levels for this dataset for 1000 genes (given in Table D1) ranged from 0.09 (for the 10M, 1e-07 condition) to 0.69 (for the 500K, 1e-06 condition). The number of replicates for each model condition for both the training and test datasets was 50.

We measured the error in the rooted species tree in terms of average normalized clade distance (nCD) [29], which is an extension of RF error for rooted trees. For our training experiment, we only rooted the true species tree topology to directly observe the rooting error. In our test experiments, we rooted both the model species tree and estimated species tree, as produced by ASTRAL, using both true and estimated gene trees (which were estimated using FastTree [22]). Additional information about the simulation study, datasets, and software commands are provided in Supplementary Materials Sec. D.

In our training experiments, we explored the impact of the shape coefficient C and the ratio $\frac{\alpha_{max}}{\beta_{min}}$ (that describes the relative impact of invariants and inequalities) on the accuracy of QR-STAR. Results for the training experiments are provided in Supplementary Materials Sec. E.1 and show that there are wide ranges of settings for the algorithmic parameters that provide the best accuracy. We used these training results and theoretical considerations related to sample complexity of QR-STAR to set the algorithmic parameters to C =1e-02 and $\frac{\alpha_{max}}{\beta_{min}} = 0$.

Figure 4 shows a comparison between QR and QR-STAR in terms of rooting error for rooting the model species tree topology using estimated gene trees (GTEE values for each condition are provided in Table D1) on the test datasets. Note that QR-STAR matches or improves upon QR across these model conditions and for different numbers of genes, with improvements that are largest under the two lower ILS levels (middle and right columns). Another trend we observe is that rooting accuracy depends on ILS, with the highest

accuracy obtained under high ILS conditions; however, even under lower ILS conditions (right), the accuracy is very good. Figures E3, E4 and E5 show additional results on the test dataset when rooting species trees estimated using ASTRAL, with true or estimated gene trees. The relative accuracy of QR and QR-STAR in these figures are the same as seen in Figure 4, with QR-STAR performing better in almost all model conditions, but the error in the final rooted tree is impacted by both species tree estimation error and rooting error, and it is generally higher under high ILS settings (see Sec. 6 for a discussion on this trend).

6 Discussion and Concluding Remarks

In this work we presented QR-STAR, a polynomial time statistically consistent method for rooting species trees under the multi-species coalescent model. QR-STAR is an extension to QR, a method for rooting species trees introduced in [29]. QR-STAR differs from QR in that it has an additional step for determining the topological shape of each unrooted quintet selected in the QR algorithm, and incorporates the knowledge of this shape in its cost function, alongside the invariants and inequalities previously used in QR. We also showed that the statistical consistency for QR-STAR holds for a larger family of optimization problems based on cost functions and sampling methods.

This study suggests several directions for future research. For example, we proved statistical consistency for one class of cost functions, which was a linear combination of the invariant, inequality and shape penalty terms; however, cost functions in other forms could also be explored and proven statistically consistent. Equation 7 suggests that the sample complexity of QR-STAR depends on the function $h(R)$, which is based on both the length of the shortest branch and the longest path in the model tree. This suggests that having very short or very long branches can both confound rooting under ILS, which is also suggested in previous studies [21]. This is unlike what is known for species tree estimation methods such as ASTRAL, where the sample complexity is only affected by the shortest branch of the model tree [27,3], and trees with long branches are easier to estimate.

Another theoretical direction is the construction of the rooted species tree directly from the unrooted gene trees. As explained in Remark 2, the proof of consistency of QR-STAR for 5-taxon trees does not depend upon the knowledge of the unrooted tree topology; this suggests that it is possible to estimate the rooted topology of the species tree in a statistically consistency manner *directly* from unrooted topological gene trees. Future work could focus on developing statistically consistent methods for this problem, which is significantly harder than the problem of rooting a given tree. To our knowledge, this is the first work that address the statistical consistency of methods for rooting species trees. It remains to be investigated whether other rooting methods can also be proven statistically consistent under models of gene evolution inside species trees, such as the MSC or models of GDL.

There are also directions for improving empirical results. An important consideration in designing a good cost function is its empirical performance, as many cost functions can lead to statistical consistency but may not provide accurate estimations of the rooted tree in practice (see Figures E1 and E2). One potential direction is to incorporate estimated branch lengths, whether of the gene trees or the unrooted species tree, into the rooting procedure.

Finally, one of the perhaps surprising results of the experimental study in this paper shows that rooting with QR and QR-STAR is easier under higher levels of discordance due to ILS, and becomes more difficult as the ILS level decreases. An explanation for this could be that with less discordance, the probability that gene trees that are less likely appear in the input would decrease. In this case, some quintet probabilities would become zero, and it may not be possible to differentiate some of the rooted quintets using the inequality and invariants derived from the ADR theory. In the extreme case, when there is no discordance, there will be only one quintet gene tree with non-zero probability, and the identifiability theorem in [2] would not hold and it becomes impossible to find the root. This as well is unlike what is known for species tree estimation methods, as they show the reverse trend: the accuracy of the species tree methods decrease under higher levels of ILS [18,19]. This suggests that an interesting trade-off could exist between the level of discordance and its impact on species tree estimation and rooting. However, the results in Figures E4 and E5 suggest that the error in the final rooted species tree might be dominated by the species tree estimation error, especially when the number of genes is small or the ILS level is high, and so the overall error in the rooted species tree would increase as the ILS levels increase.

References

1. Alanzi, A.R., Degnan, J.H.: Inferring rooted species trees from unrooted gene trees using approximate bayesian computation. Molecular phylogenetics and evolution **116**, 13–24 (2017)
2. Allman, E.S., Degnan, J.H., Rhodes, J.A.: Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. Journal of Mathematical Biology **62**(6), 833–862 (2011)
3. Chan, Y.b., Li, Q., Scornavacca, C.: The large-sample asymptotic behaviour of quartet-based summary methods for species tree inference. Journal of Mathematical Biology **85**(3), 1–22 (2022)
4. Felsenstein, J.: Cases in which parsimony or compatibility methods will be positively misleading. Systematic zoology **27**(4), 401–410 (1978)
5. Graham, S.W., Olmstead, R.G., Barrett, S.C.: Rooting phylogenetic trees with distant outgroups: a case study from the commelinoid monocots. Molecular biology and evolution **19**(10), 1769–1781 (2002)
6. Skarp-de Haan, C., Culebro, A., Schott, T., Revez, J., Schweda, E.K., Hänninen, M.L., Rossi, M.: Comparative genomics of unintrogressed campylobacter coli clades 2 and 3. BMC genomics **15**(1), 1–14 (2014)
7. Hess, P.N., De Moraes Russo, C.A.: An empirical test of the midpoint rooting method. Biological Journal of the Linnean society **92**(4), 669–674 (2007)
8. Holland, B., Penny, D., Hendy, M.: Outgroup misplacement and phylogenetic inaccuracy under a molecular clock—a simulation study. Systematic biology **52**(2), 229–238 (2003)
9. Hudson, R.R.: Testing the constant-rate neutral allele model with protein sequence data. Evolution pp. 203–217 (1983)
10. Jun, S.R., Leuze, M.R., Nookaew, I., Uberbacher, E.C., Land, M., Zhang, Q., Wanchai, V., Chai, J., Nielsen, M., Trolle, T., et al.: Ebolavirus comparative genomics. FEMS microbiology reviews **39**(5), 764–778 (2015)
11. 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)
12. Li, C., Matthes-Rosana, K.A., Garcia, M., Naylor, G.J.: Phylogenetics of Chondrichthyes and the problem of rooting phylogenies with distant outgroups. Molecular phylogenetics and evolution **63**(2), 365–373 (2012)
13. Liu, L., Yu, L., Edwards, S.V.: A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC evolutionary biology **10**(1), 1–18 (2010)
14. Maddison, W.P.: Gene trees in species trees. Systematic biology **46**(3), 523–536 (1997)
15. Maddison, W.P., Donoghue, M.J., Maddison, D.R.: Outgroup analysis and parsimony. Systematic biology **33**(1), 83–103 (1984)
16. Mahbub, M., Wahab, Z., Reaz, R., Rahman, M.S., Bayzid, M.S.: wQFM: highly accurate genome-scale species tree estimation from weighted quartets. Bioinformatics **37**(21), 3734–3743 (2021)
17. Mai, U., Sayyari, E., Mirarab, S.: Minimum variance rooting of phylogenetic trees and implications for species tree reconstruction. PloS one **12**(8), e0182238 (2017)
18. 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)
19. Mirarab, S., Warnow, T.: ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics **31**(12), i44–i52 (2015)
20. Mitzenmacher, M., Upfal, E.: Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis. Cambridge university press (2017)
21. Posada, D.: Phylogenomics for systematic biology. Systematic biology **65**(3), 353–356 (2016)
22. Price, M.N., Dehal, P.S., Arkin, A.P.: Fasttree 2–approximately maximum-likelihood trees for large alignments. PloS one **5**(3), e9490 (2010)
23. Rabiee, M., Mirarab, S.: QuCo: quartet-based co-estimation of species trees and gene trees. Bioinformatics **38**(Supplement_1), i413–i421 (2022)
24. Renner, S.S., Grimm, G.W., Schneeweiss, G.M., Stuessy, T.F., Ricklefs, R.E.: Rooting and dating maples (acer) with an uncorrelated-rates molecular clock: implications for north american/asian disjunctions. Systematic biology **57**(5), 795–808 (2008)
25. Robinson, D.F., Foulds, L.R.: Comparison of phylogenetic trees. Mathematical biosciences **53**(1-2), 131–147 (1981)
26. Rosenberg, N.A.: Counting coalescent histories. Journal of Computational Biology **14**(3), 360–377 (2007)
27. Shekhar, S., Roch, S., Mirarab, S.: Species tree estimation using ASTRAL: how many genes are enough? IEEE/ACM transactions on computational biology and bioinformatics **15**(5), 1738–1747 (2017)
28. Simmons, M.P., Springer, M.S., Gatesy, J.: Gene-tree misrooting drives conflicts in phylogenomic coalescent analyses of palaeognath birds. Molecular Phylogenetics and Evolution **167**, 107344 (2022)
29. Tabatabaee, Y., Sarker, K., Warnow, T.: Quintet Rooting: rooting species trees under the multi-species coalescent model. Bioinformatics **38**(Supplement_1), i109–i117 (2022)
30. Tria, F.D.K., Landan, G., Dagan, T.: Phylogenetic rooting using minimal ancestor deviation. Nature ecology & evolution **1**(1), 1–7 (2017)

31. Wascher, M., Kubatko, L.: Consistency of SVDQuartets and maximum likelihood for coalescent-based species tree estimation. Systematic Biology **70**(1), 33–48 (2021)
32. Zhang, C., Rabiee, M., Sayyari, E., Mirarab, S.: ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics **19**(6), 15–30 (2018)