On the use of quartets for reconstructing cell lineage trees under an unbiased error and missingness model

Yunheng Han[1,2][0000−0003−0200−5924] and Erin K. Molloy[1,2,*][0000−0001−5553−3312]

[1] *Department of Computer Science, University of Maryland, College Park, USA*

[2] *University of Maryland Institute for Advanced Computer Studies, College Park, USA*

**Email: ekmolloy@umd.edu*

April 4, 2023

Abstract

A cell lineage tree, or phylogeny, describes the evolutionary history of cells in a biological sample. Reconstructing cell lineage trees for tumor biopsies may shed light on cancer progression and response to treatment; thus the development of fast and accurate methods is of critical importance. Many popular methods assume that mutations occur according to the infinite sites model and that deviations from a perfect phylogeny are due to sequencing error, as data generated by single-cell platforms are notoriously error-prone. To address this issue, sequencing error is modeled, with false positives, false negatives, and missing values being independent and identically distributed across mutations and cells. Here, we consider how to estimate phylogenies under this model using quartets (i.e., unrooted, binary, four-leaf trees), which are implied by mutations being present in two cells and absent from two cells. Our main result is that the most probable quartet identifies the unrooted cell lineage tree; it follows that the most frequent quartet from the input data is a consistent estimator. This leads us to consider how heuristics developed for maximum quartet support supertrees may further enable cell lineage tree reconstruction. Overall, our work connects statistical theory from species phylogenetics to tumor phylogenetics.

1 Introduction

Systematics has transformed over the last few decades from the observation that gene trees evolve within a species tree (Maddison 1997; Edwards 2009). A gene tree reflects the genealogical history of a gene, which is passed down from ancestor to descendant. A species tree, in contrast, governs the pool of potential ancestors. The Multi-Species Coalescent (MSC) (Pamilo and Nei 1988; Rannala and Yang 2003) generates gene trees given a model species tree with branch lengths in coalescent units. Gene trees generated under the MSC can disagree (i.e., be topologically discordant) with the model species tree, in which case we say that incomplete lineage sorting (ILS) has occurred. An important result of the last few decades is that the most probable gene tree under the MSC can disagree with the model species tree (Degnan and Rosenberg 2006). In such situations, the model species tree is said to be in the anomaly zone or the offending gene tree is said to be anomalous. It is now widely recognized that anomalous gene trees can challenge traditional species tree estimation methods (Kubatko and Degnan 2007; Roch and Steel 2015).

Importantly, there are no anomalous quartets (unrooted, binary, four-leaf trees) under the MSC (Allman et al. 2011; Degnan 2013). It follows that the most frequent quartet in the input gene trees is a consistent estimator of the unrooted species tree. This result has driven methodological development over the last decade, with some of the most popular approaches to date being based on quartets (Larget et al. 2010; Mirarab et al. 2014). Moreover, new and improved quartet-based methods are continually being developed (Mirarab and Warnow 2015; Zhang et al. 2018; Dibaeinia et al. 2021; Mahbub et al. 2021; Han and Molloy 2023). Similar results exist for triplets (rooted, binary, three-leaf trees); see Degnan and Rosenberg (2006); Liu et al. (2010); Islam et al. (2020).

Here, we connect statistical theory and computational techniques from species phylogenetics to tumor phylogenetics by examining the use of quartets and triplets for reconstructing cell lineage trees from single-cell sequencing data. This task is important for cancer genomics, as the evolutionary histories of tumors can shed light on cancer progression and response to treatment (Schwartz and Schäffer 2017). Many popular methods for tumor phylogenetics assume that mutations occur on the cell lineage tree according to the infinite sites (IS) model, meaning that there are no parallel mutations or reversals. Thus, a mutation being present in two different cells implies they have descended from a common ancestor cell. Deviations from a perfect phylogeny are attributed to sequencing error, modeled as the introduction of false positives, false negatives, and missing values with probability α, β, and γ, respectively (Ross and Markowetz 2016; Jahn et al. 2016; Wu 2019; Kizilkale et al. 2022). We refer to this as the unbiased error (UE) and missingness (M) model because errors and missing values are independent and identically distributed across mutations and cells.

The IS+UEM model generates 2-state characters: state 0 indicates the mutation is absent from a cell and state 1 indicates the mutation is present (note that these mutations are called in reference to a healthy cell, which is the root of the cell lineage tree). Although 2-state characters are generated under the IS+UEM model rather than (gene) trees, quartets are implied by characters with two cells in state 0 and two cells in

state 1. Our main result is that there are no anomalous quartets under the IS+UEM model; it follows that the most frequent quartet is a consistent estimator of the unrooted cell lineage tree. As we will show, this result generalizes to other models of 2-state character evolution, denoted $\mathcal{M}$, provided there are no anomalous quartets under $\mathcal{M}$. An example of such a model is the neutral Wright-Fisher (WF) plus IS model (Fisher 1923; Wright 1931) and its approximations (Hudson 2002). Under the WF+IS model, mutations follow the IS assumption but evolve within a model species tree. Thus, deviations from a perfect phylogeny are due to ILS. The WF+IS model has been used in phylogenetics to reconstruct species trees from low-homoplasy retroelement insertions (Mendes and Hahn 2017; Springer et al. 2019; Molloy et al. 2021; Korstian et al. 2022); it is also commonly used in population genetics, although in this context mutations typically evolve within a network rather than a tree (Patterson et al. 2012; Pickrell and Pritchard 2012; Peter 2016). Because there are no anomalous quartets under $\mathcal{M}$+UEM models, we consider the use of quartets for reconstructing phylogenetic trees from 2-state characters, focusing on issues that may arise in tumor phylogenetics.

Lastly, we extend our results to triplets (rooted, binary, three-leaf trees), finding that some additional requirements are necessary (e.g., $\alpha = 0$) for there to be no anomalous triplets under $\mathcal{M}$+UEM models. This has suggests that triplets are less robust to error than quartets, perhaps limiting their utility for cell lineage tree reconstruction.

2 Preliminaries

We begin with some preliminaries on phylogenetic trees and models of character evolution.

2.1 Phylogenetic Trees

A *phylogenetic tree* or X-tree is defined by the tripple (g, X, ϕ), where g is a connected acyclic graph, X is a set of labels (often representing species or cells), and ϕ is a bijection between the labels in X and leaves (i.e., vertices with degree 1) of g. Phylogenetic trees can be either *unrooted* or *rooted*, and we denote the unrooted version of a rooted tree T as $u(T)$. Edges in an unrooted tree are undirected, whereas edges in a rooted tree are directed away from the root, a special vertex with in-degree 0 (all other vertices have in-degree 1). Vertices that are neither leaves nor the root are called internal vertices, and an interval vertex with degree greater than 3 is called a *polytomy*. A polytomy can be introduced to tree T by contracting one of its edges (i.e., deleting the edge and identifying its endpoints). A *refinement* of a polytomy is the opposite of a *contraction*. If there are no polytomies in T, we say that T is *binary* or *fully resolved*; otherwise, we say that T is *non-binary* or *unresolved*.

As previously mentioned, many methods for species tree estimation are based on *quartets*. A quartet is an unrooted, binary tree with four leaves. We denote the three possible quartets on $X = \{A, B, C, D\}$ as $q_1 = A, B|C, D$, $q_2 = A, C|B, D$, and $q_3 = A, D|B, C$. A set of quartets can be created from an unrooted tree T by restricting T to every possible subset of four labels (i.e., deleting the other leaves and suppressing

vertices of degree 2). The resulting set $Q(T)$ is referred to as the quartet encoding of T, and we say that T displays quartet q if $q \in Q(T)$. Importantly, if T contains polytomies, not every selection of four labels will produce a quartet. Some selections will produce star trees, which do not provide any topological information. The concepts for quartets can be extended to triplets. A *triplet* is a rooted, binary tree with three leaves. We denote the three possible triplets on $X = \{A, B, C\}$ as $t_A = A|B, C$, $t_B = B|A, C$, and $t_C = C|A, B$. Lastly, a *bipartition* or *split* of leaf label set X partitions it into disjoint subsets. It is easy to see that 2-state characters are bipartitions and that each edge in an unrooted tree induces a bipartition. The set of bipartitions induced by the edges of T is denoted $Bip(T)$.

2.2 Characters and Models of Evolution

A character matrix is an $n \times m$ matrix, where n is the number of leaves (representing cells or species) and m is the number of characters (representing mutations). The columns of the matrix are commonly referred to as mutations, characters, or site patterns. Our focus here is on 2-state characters, with state 0 indicating the mutation is absent (i.e., the ancestral state) and state 1 indicating mutation is present (i.e., the derived state). We assume that characters are generated under some model of evolution parameterized by a rooted phylogenetic tree topology and some associated numeric parameters (Figure 1). However, the resulting matrix G is not directly observed due to error. Error is commonly modeled as

$$P(D_{i,j} = x | G_{i,j} = y) = \begin{cases} (1-\alpha) & \text{if } D_{i,j} = 0 \text{ and } G_{i,j} = 0 \\ \alpha & \text{if } D_{i,j} = 1 \text{ and } G_{i,j} = 0 \\ \beta & \text{if } D_{i,j} = 0 \text{ and } G_{i,j} = 1 \\ (1-\beta) & \text{if } D_{i,j} = 1 \text{ and } G_{i,j} = 1 \end{cases} \tag{1}$$

where $G_{i,j}$ is the state generated from the model cell lineage tree for leaf i and mutation j and $D_{i,j}$ is the observed state. It is also common to assume that entries of D can be missing with probability γ; this is easily accommodated by multiplying each of the cases in Equation 1 by $(1 - \gamma)$.

For much of this paper, we will assume that G is generated under the IS model given some cell lineage tree. The model cell lineage tree T includes a rooted topology as well as a set of numeric parameters, containing the probability of a mutation occurring on each of the edges of T. Edges on which mutations cannot occur are contracted so that the probability of a mutation on any edge in T is strictly greater than zero. An exception occurs because the model cell lineage tree must be extended so that each internal vertex is also represented as a "fake" leaf. Mutations are not allowed on "fake" edges, which are incident to "fake" leaves. This extension process is necessary because single-cell sequencing can produce data for cells that are ancestral to other cells in the same data set.

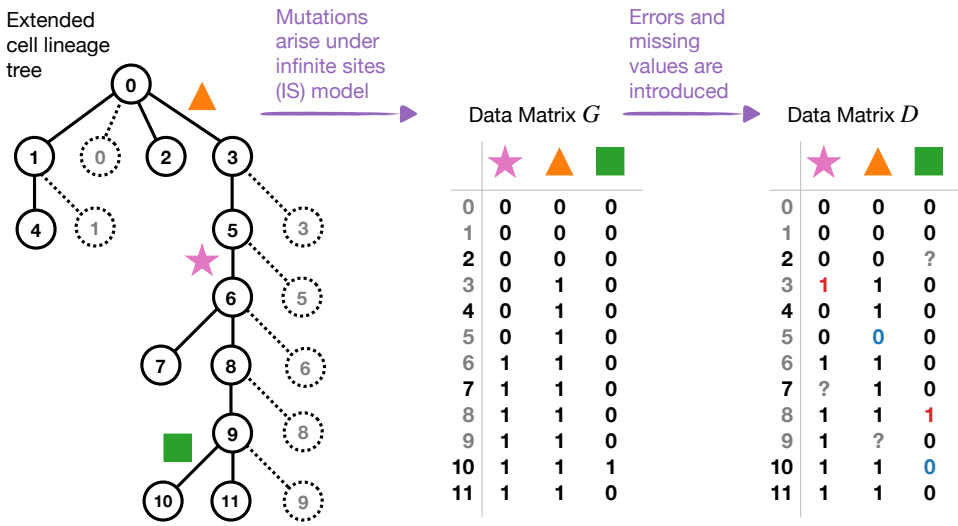

Figure 1: Above we show a model cell lineage tree, where the dashed lines and circles are "fake" edges and vertices, respectively. If we assume a mutation occurs on any non-fake edges with equal probability (e.g., Kizilkale et al. 2022), then probability of a mutation on any solid edge will be 1/11. Mutations cannot occur on any of the dashed edges. Data are generated from this model cell lineage tree in two steps. First, mutations arise on the tree according to the infinite sites model, producing data matrix G. Second, false positives (0 flips to 1; shown in red), false negatives (1 flips to 0; shown in blue), and missing values (0/1 flips to ?; shown in grey) are introduced to G under the UEM model, producing data matrix D.

3 No anomalous quartets under an unbiased error and missingness model

We now derive our main result: there are no anomalous quartets under the IS+UEM model, which is commonly used in tumor phylogenetics.

First, we note that an important feature of the IS model tree is that it can be decomposed into sub-models for any subset of leaves. A submodel of T is created by selecting a subset of leaves from T, restricting T to this leaf set, and updating the numeric parameters accordingly. For the IS model, the edge probabilities are added together (as this represents an 'or' condition, i.e., a mutation can occur on this branch or on that branch). We can also add branch lengths (in coalescent units) for the neutral WF+IS model, although we focus on the IS model for now.

For a four-leaf sub-model σ, there are five possible topologies (Figure 2). Two of them display a star when unrooted, and the other three correspond to a quartet when unrooted. Our goal is to estimate σ (unrooted topology only) from the character data, restricted to the same four leaves (cells) as σ. There are 16 possible patterns when considering a single mutation (column) on cells $\{A, B, C, D\}$. Two patterns (0011 and 1100) support quartet $q_1 = A, B|C, D$, two patterns (0101 and 1010) support quartet $q_2 = A, C|B, D$, two patterns (0110 and 1001) support quartet $q_3 = A, D|B, C$, and the other 10 patterns support a star (i.e., they do not contain topological information). Thus, we denote the probability of quartets under IS model given σ as

$\mathbb{P}(q_1|\sigma) = \mathbb{P}(1100|\sigma) + \mathbb{P}(0011|\sigma)$, $\mathbb{P}(q_2|\sigma) = \mathbb{P}(1010|\sigma) + \mathbb{P}(0101|\sigma)$, and $\mathbb{P}(q_3|\sigma) = \mathbb{P}(1001|\sigma) + \mathbb{P}(0110|\sigma)$. The same notation is used for the IS+UEM model but including the additional parameters: α, β, and γ.

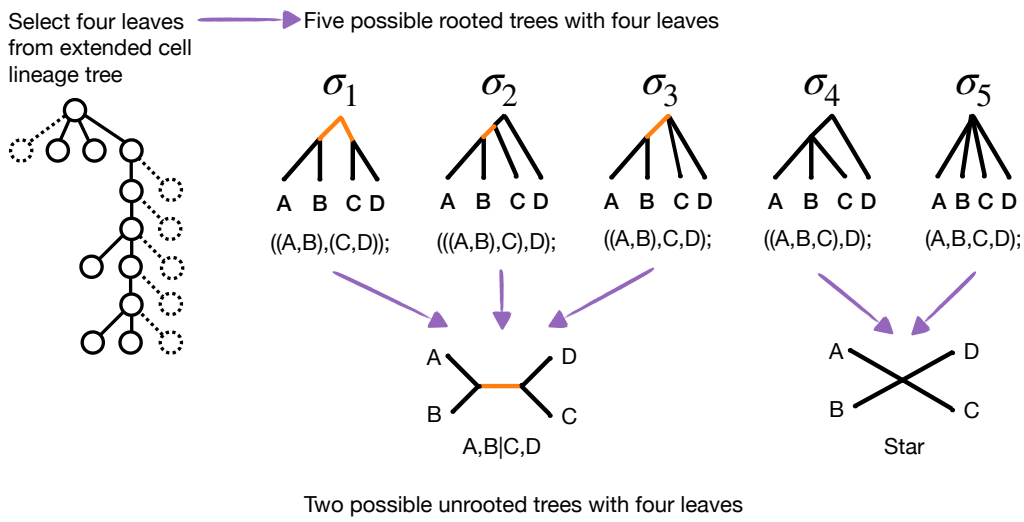

Figure 2: The five possible cell lineage trees with four leaves. Three of these trees (σ_1, σ_2, and σ_3) have a non-trivial unrooted topology (quartet: $A, B|C, D$). The other two trees (σ_4 and σ_5) have a trivial unrooted topology (star). Looking at the cell lineage tree in Figure 1, σ_1–σ_5 can be recovered from sampling cells $X_1 = \{1, 4, 10, 11\}$, $X_2 = \{1, 7, 10, 11\}$, $X_3 = \{1, 2, 10, 11\}$, $X_4 = \{1, 9, 10, 11\}$, and $X_5 = \{0, 1, 2, 3\}$, respectively.

Before coming to our main result, we provide a useful lemma.

Lemma 1. *Let $0 \leq \alpha < 1$ and $0 \leq \beta < 1$. Then,*

$$\left((1-\beta)^2(1-\alpha)^2 + \beta^2\alpha^2\right) - 2\beta(1-\beta)\alpha(1-\alpha) = \left(1 - (\alpha+\beta)\right)^2 > 0 \tag{2}$$

for $\alpha + \beta \neq 1$. If $\alpha + \beta = 1$, the inequality in Equation 2 becomes an equality.

See Supplementary Materials for proof.

Theorem 1. *Let σ be an IS model tree with four leaves. Then, under the IS+UE model, $\mathbb{P}(q_1|\alpha, \beta, \sigma) > \mathbb{P}(q_2|\alpha, \beta, \sigma) = \mathbb{P}(q_3|\alpha, \beta, \sigma)$ if σ has unrooted topology q_1 and $\mathbb{P}(q_1|\alpha, \beta, \sigma) = \mathbb{P}(q_2|\alpha, \beta, \sigma) = \mathbb{P}(q_3|\alpha, \beta, \sigma)$ if the unrooted version of σ is a star, assuming $\alpha + \beta \neq 1$. Thus, there are no anomalous quartet under the IS+UE model.*

Proof. Across all four-leaf model trees (Figure 2), there are nine characters that can be generated under the IS model: two invariant characters (0000 and 1111), five singleton characters (1000, 0100, 0010, 0001, and 1110), and two phylogenetically informative characters (1100 and 0011); see Supplementary Table S1. For these 9 characters, we enumerate all possible ways of introducing false positives and false negatives.

This gives us the probability of observing the 16 possible characters under the UE model; see Table 1 and Supplementary Tables S2–S7. By combining the probabilities of the informative characters, we find that the probability of quartet q under the IS+UE model is

$$\begin{aligned}\mathbb{P}(q|\alpha,\beta,\sigma) = &\left(2\alpha^2(1-\alpha)^2\right)\cdot\mathbb{P}(0000|\sigma) + \left(2\beta^2(1-\beta)^2\right)\cdot\mathbb{P}(1111|\sigma)\\ &+\Big(\big(\alpha(1-\alpha)^2(1-\beta)\big)+\big(\alpha^2(1-\alpha)\beta\big)\Big)\cdot\big(\mathbb{P}(0001|\sigma)+\mathbb{P}(0010|\sigma)+\mathbb{P}(0100|\sigma)+\mathbb{P}(1000|\sigma)\big)\\ &+\Big(\big(\beta(1-\beta)^2(1-\alpha)\big)+\big(\beta^2(1-\beta)\alpha\big)\Big)\cdot\mathbb{P}(1110|\sigma)\\ &+f(q,\alpha,\beta)\cdot\big(\mathbb{P}(1100|\sigma)+\mathbb{P}(0011|\sigma)\big)\end{aligned}$$

where

$$f(q,\alpha,\beta) = \begin{cases}\left((1-\beta)^2(1-\alpha)^2+\beta^2\alpha^2\right) & \text{if } q=q_1\\ 2\beta(1-\beta)\alpha(1-\alpha) & \text{if } q=q_2 \text{ or } q=q_3\end{cases}$$

Now we can compute the difference in probabilities between q_1 and q_j for $j\in\{2,3\}$.

$$\begin{aligned}\mathbb{P}(q_1|\alpha,\beta,\sigma)-\mathbb{P}(q_j|\alpha,\beta,\sigma) &= \Big(\left((1-\alpha)^2(1-\beta)^2+\alpha^2\beta^2\right)-2\alpha\beta(1-\alpha)(1-\beta)\Big)\cdot\mathbb{P}(q_1|\sigma)\\ &= \big(1-(\alpha+\beta)\big)^2\cdot\mathbb{P}(q_1|\sigma)\quad\text{(Lemma 1)}\end{aligned}\tag{3}$$

Assuming $\alpha+\beta\neq 1$, this quantity is zero if $\mathbb{P}(q_1|\sigma)=\mathbb{P}(1100|\sigma)+\mathbb{P}(0011|\sigma)=0$ and greater than zero otherwise. The former will be the case if the unrooted version of σ is a star (i.e, $\sigma\in\{\sigma_4,\sigma_5\}$); the latter will be the case otherwise (i.e, $\sigma\in\{\sigma_1,\sigma_2,\sigma_3\}$). □

It is worth noting that the quantity $\alpha+\beta$ is unlikely to equal 1 in practice, as both probabilities should be less than 0.5.

Corollary 1. *There are no anomalous quartets under the IS+UEM model, where $0\leq\gamma<1$ is the probability that an entry of the character matrix is missing.*

Proof. Let x be a character of four states, each corresponding to a leaf in the model tree. If one or more of the values in character x is missing, then no quartet is displayed by x. Therefore, we replace $\mathbb{P}(x|\sigma)$ with $\mathbb{P}(x|\gamma,\sigma)=(1-\gamma)^4\cdot\mathbb{P}(x|\sigma)$ in the proof of Theorem 1. This update does not impact the result. □

We now provide a similar result for a more general models of 2-state character evolution, beyond the IS model.

Theorem 2. *Let $\mathcal{M}$ be some model of evolution that generates 2-state characters, and let σ be a model tree with four leaves. Suppose that $\mathbb{P}(q_1|\sigma)>\mathbb{P}(q_2|\sigma)$ and $\mathbb{P}(q_1|\sigma)>\mathbb{P}(q_3|\sigma)$ if σ has unrooted topology q_1 and that $\mathbb{P}(q_1|\sigma)=\mathbb{P}(q_2|\sigma)=\mathbb{P}(q_3|\sigma)$ if the unrooted version of σ is a star. Then, these equations hold for the*

$\mathcal{M}$+UEM model, assuming $\alpha + \beta \neq 1$. Put simply, if there are no anomalous quartets under $\mathcal{M}$, then there are no anomalous quartets under $\mathcal{M}$+UEM.

See Supplementary Materials for proof, in which we compute the difference in probability between quartet q_1 and an alternative quartet q_j:

$$\mathbb{P}(q_1|\alpha,\beta,\gamma,\sigma) - \mathbb{P}(q_j|\alpha,\beta,\gamma,\sigma) = \big(1-(\alpha+\beta)\big)^2 \cdot (1-\gamma)^4 \cdot \big(\mathbb{P}(q_1|\sigma) - \mathbb{P}(q_j|\sigma)\big). \tag{4}$$

for $j \in \{2,3\}$. If $\mathcal{M}$ is the IS model, then $P(q_1|\sigma) > 0$ and $P(q_j|\sigma) = 0$ for $j \in \{2,3\}$; this gives us our previous result (Theorem 1 and Corollary 1). In this case, the difference in quartet probabilities (Equation 4) depends on (1) the error probabilities (specifically how close their sum is to one), (2) the probability of missing values, and (3) the probability of observing an informative pattern (0011 and 1100) under the IS model. In the simulations performed by Kizilkale et al. (2022), the largest values of α, β, and γ were 0.001, 0.2, and 0.05, respectively. This puts the difference in probabilities at $0.52 \cdot \mathbb{P}(q_1|\sigma)$. Thus, the magnitude of $\mathbb{P}(q_1|\sigma)$, which depends on the larger cell lineage tree as well as the four cells sampled from it, is likely to have a larger impact on the difference in quartet probabilities. Another example of $\mathcal{M}$ for which there are no anomalous quartets is the neutral WF+IS model (Theorem 1 in Molloy et al. 2021).

We now consider statistical estimation under the $\mathcal{M}$+UEM model, in the special case where the model tree has four leaves.

# of FP	# of FN	$\mathbb{P}(D_{:,j}\|G_{:,j})$	$D_{:,j}$ given $G_{:,j}$ of **1100**	**0011**	Unrooted Tree
0	0	$(1-\beta)^2(1-\alpha)^2$	1100	0011	$A,B\|C,D$
0	1	$\beta(1-\beta)(1-\alpha)^2$	0100	0001	star
			1000	0010	star
0	2	$\beta^2(1-\beta)^2$	0000	0000	star
1	0	$(1-\beta)^2\alpha(1-\alpha)$	1110	1011	star
			1101	0111	star
1	1	$\beta(1-\beta)\alpha(1-\alpha)$	0110	1001	$A,D\|B,C$
			0101	1010	$A,C\|B,D$
			1010	0101	$A,C\|B,D$
			1001	0110	$A,D\|B,C$
1	2	$\beta^2\alpha(1-\alpha)$	0010	1000	star
			0001	0100	star
2	0	$(1-\beta)^2\alpha^2$	1111	1111	star
2	1	$\beta(1-\beta)\alpha^2$	0111	1101	star
			1011	1110	star
2	2	$\beta^2\alpha^2$	0011	1100	$A,B\|C,D$

Table 1: Above we show the characters that can be generated from **character #12** ($ABCD = 1100$) and **character #3** (0011) under the UE model as well as their probabilities.

Theorem 3. *Let $\mathcal{C}$ be a set of 2-state characters generated from σ under the $\mathcal{M}$+UEM model. Suppose that σ is binary with four leaves and that there are no anomalous quartets under $\mathcal{M}$. Then, the most frequent quartet in $\mathcal{C}$ is a consistent estimator of the unrooted version of σ under the $\mathcal{M}$+UEM model.*

Proof. By Theorem 2, the quartet corresponding to $u(\sigma)$, denoted q_1, has greater probability under the $\mathcal{M}$+UEM model than the two alternatives: q_2 and q_3. Therefore, as the number of characters in $\mathcal{C}$ goes to infinity, the frequency of q_1 will be greater q_2 and q_3 with high probability. It follows that the most frequent quartet will correspond to $u(\sigma)$ with high probability. □

4 Anomalous triplets under an unbiased error and missingness model

We now provide some related results for triplets. For a three-leaf sub-model σ, there are two possible topologies (Supplementary Figure S1). One of them is non-binary or the other is binary. Once again, our goal is to estimate σ (rooted topology only) from character data, restricted to the same three leaves (cells) as σ. There are 8 possible patterns when considering a single mutation (column) on cells $\{A, B, C\}$. One supports (110) supports triplet $t_C = C|A, B$, one (101) supports triplet $t_B = B|A, C$, and one (011) supports triplet $t_A = A|B, C$, provided we make the IS assumption because then two cells being in state 1 and one cell being in state 0 as this implies that the cells in state 1 have a lower common ancestor than either of these cells paired with the cell in state 0. Thus, we denote the probability of triplets given σ as $\mathbb{P}(t_A|\sigma) = \mathbb{P}(011|\sigma)$, $\mathbb{P}(t_B|\sigma) = \mathbb{P}(101|\sigma)$, and $\mathbb{P}(t_C|\sigma) = \mathbb{P}(110|\sigma)$.

Theorem 4. *Let $\mathcal{M}$ be some model of evolution that generates 2-state characters with the IS assumption, and let σ be a model tree with three leaves $\{A, B, C\}$. Suppose that $\mathbb{P}(t_C|\sigma) > \mathbb{P}(t_A|\sigma)$ and $\mathbb{P}(t_C|\sigma) > \mathbb{P}(t_B|\sigma)$ if σ corresponds to triplet t_C and that $\mathbb{P}(t_C|\sigma) = \mathbb{P}(t_A|\sigma) = \mathbb{P}(t_B|\sigma)$ if σ is non-binary. Then, these equations hold and thus there are no anomalous quartets under the $\mathcal{M}$+UEM model, provided that $\alpha = 0$ or that $\alpha + \beta \neq 1$ and singletons (*100, 010, *and* 001*) occur with equal probability. Otherwise, there can be anomalous triplets under $\mathcal{M}$+UEM.*

Proof. Across all three-leaf model trees, there are eight characters that can be generated: two invariant characters (000 and 111), three singleton characters (100, 010, 001), and three phylogenetically informative characters (110, 101, and 011). For these eight characters, we enumerate all possible ways of introducing false positives and false negatives. This gives us the probability of observing the 8 possible characters under the UE model; see Supplementary Tables S9–S16. By combining the probabilities of the informative characters,

we find that the probability of triplet t_C under the $\mathcal{M}$+UE model is

$$\begin{aligned}\mathbb{P}(t_C|\alpha,\beta,\sigma) = \alpha^2(1-\alpha)\cdot\mathbb{P}(000|\sigma) + \beta(1-\beta)^2\cdot\mathbb{P}(111|\sigma)\\ + \alpha^2\beta\cdot\mathbb{P}(001|\sigma) + \alpha(1-\alpha)(1-\beta)\cdot\big(\mathbb{P}(100|\sigma)+\mathbb{P}(010|\sigma)\big)\\ + (1-\alpha)(1-\beta)^2\cdot\mathbb{P}(110|\sigma) + \alpha\beta(1-\beta)\cdot\big(\mathbb{P}(101|\sigma)+\mathbb{P}(011|\sigma)\big)\end{aligned}$$

Similar probabilities can be computed for t_B and t_A. To provide a general formula, we define

$$g(\alpha,\beta,\sigma) = \alpha^2(1-\alpha)\cdot\mathbb{P}(000|\sigma) + \beta(1-\beta)^2\cdot\mathbb{P}(111|\sigma)$$

and set $x_A = 100$, $x_B = 010$, and $x_C = 001$. This allows us to write the probability of any triplet as

$$\begin{aligned}\mathbb{P}(t_i|\alpha,\beta,\sigma) = g(\alpha,\beta,\sigma)\\ + \alpha^2\beta\cdot\mathbb{P}(x_i|\sigma) + \alpha(1-\alpha)(1-\beta)\cdot\big(\mathbb{P}(x_j|\sigma)+\mathbb{P}(x_k|\sigma)\big)\\ + (1-\alpha)(1-\beta)^2\cdot\mathbb{P}(t_i|\sigma) + \alpha\beta(1-\beta)\cdot\big(\mathbb{P}(t_j|\sigma)+\mathbb{P}(t_k|\sigma)\big)\end{aligned} \tag{5}$$

where $i,j,k \in \{A,B,C\}$ with $i \neq j \neq k$. Now we can compute the differences in probabilities between any pair of triplets t_j and t_k.

$$\begin{aligned}&\mathbb{P}(t_j|\alpha,\beta,\sigma) - \mathbb{P}(t_k|\alpha,\beta,\sigma)\\ &= \alpha^2\beta\cdot\big(\mathbb{P}(x_j|\sigma)-\mathbb{P}(x_k|\sigma)\big) + \alpha(1-\alpha)(1-\beta)\cdot\big(\mathbb{P}(x_k|\sigma)-\mathbb{P}(x_j|\sigma)\big)\\ &\quad + (1-\alpha)(1-\beta)^2\cdot\big(\mathbb{P}(t_j|\sigma)-\mathbb{P}(t_k|\sigma)\big) + \alpha\beta(1-\beta)\cdot\big(\mathbb{P}(t_k|\sigma)-\mathbb{P}(t_j|\sigma)\big)\\ &= \big(\alpha^2\beta - \alpha(1-\alpha)(1-\beta)\big)\cdot\big(\mathbb{P}(x_j|\sigma)-\mathbb{P}(x_k|\sigma)\big) + \big((1-\alpha)(1-\beta)^2 - \alpha\beta(1-\beta)\big)\cdot\big(\mathbb{P}(t_j|\sigma)-\mathbb{P}(t_k|\sigma)\big)\\ &= \alpha\big(1-(\alpha+\beta)\big)\cdot\big(\mathbb{P}(x_k|\sigma)-\mathbb{P}(x_j|\sigma)\big) + (1-\beta)\big(1-(\alpha+\beta)\big)\cdot\big(\mathbb{P}(t_j|\sigma)-\mathbb{P}(t_k|\sigma)\big)\end{aligned} \tag{6}$$

Assuming that $\alpha = 0$ or that $\alpha+\beta = 0$ and $\mathbb{P}(x_k|\sigma) = \mathbb{P}(x_j|\sigma)$, this quantity is zero if $\mathbb{P}(t_j|\sigma) = \mathbb{P}(t_k|\sigma)$ and greater than zero if $\mathbb{P}(t_j|\sigma) > \mathbb{P}(t_k|\sigma)$. By our assumptions on $\mathcal{M}$, the former will be the case if σ is non-binary; the latter will be the case if σ is binary. Thus, there are no anomalous triplets under $\mathcal{M}+UE$, provided that false positives cannot occur or that $\alpha+\beta = 1$ and singletons occur with equal probability. If these conditions do not hold, there exists an anomalous triplet when

$$\frac{\alpha}{(1-\beta)}\cdot\big(\mathbb{P}(x_C|\sigma)-\mathbb{P}(x_k|\sigma)\big) - \big(\mathbb{P}(t_C|\sigma)-\mathbb{P}(t_k|\sigma)\big) > 0 \tag{7}$$

for $k \in \{A,B\}$. □

Examples of models for which there are no anomalous triplets are the IS model and the neutral WF+IS model (see Supplementary Materials of Kuritzin et al. 2016). As previously mentioned, the IS model is

commonly used in tumor phylogenetics where α and β are used to model single-cell sequencing error. In this setting, the largest values of α and β were 0.001 and 0.2, respectively, in the simulations performed by Kizilkale et al. (2022). This would put $\alpha/(1-\beta) = 0.00125$ so $\mathbb{P}(x_C|\sigma) - \mathbb{P}(x_k|\sigma)$ would need to be 800 times greater than $\mathbb{P}(t_C|\sigma) - \mathbb{P}(t_k|\sigma)$ for t_k to be anomalous.

Although this seems drastic, it could occur in the context of sampling three cells from a larger cell lineage tree. Consider sampling cells 1, 4, and 10 from the model cell lineage tree T^* in Figure 1, assuming mutations occur on all (non-fake) edges with equal probability. The resulting sub-model σ has topology $t_{10} = 10|1,4$. Then, under the IS model, the probability of singletons is $\mathbb{P}(x_1|\sigma) = 0$, $\mathbb{P}(x_4|\sigma) = 1/11$, and $\mathbb{P}(x_{10}|\sigma) = 6/11$. Similarly, the probability of triplets is $\mathbb{P}(t_1|\sigma) = 0$, $\mathbb{P}(t_4|\sigma) = 0$, and $\mathbb{P}(t_{10}|\sigma) = 1/11$. Looking at Equation 7 from Theorem 4, we see that triplet $t_1 = 1|4,10$ will be an anomalous triplet whenever $\alpha/(1-\beta) > 1/6$, which could occur for $\beta = 0$ and $\alpha = 0.2$. Of course, these parameter settings are not realistic; under more realistic error rates, we might expect something like $0.001/0.8 = 0.00125$. However, from this example, it is easy to see that if we add 1000 descendants between cell 9 and cell 10, then t_1 will be anomalous for these reasonable error rates. This leads us to conjecture that triplets may be less robust than quartets for reconstructing large cell lineage trees.

Before moving on to phylogeny estimation, we consider the the presence of missing data.

Corollary 2. *Theorem 4 holds for the $\mathcal{M}$+UEM model, where $0 \leq \gamma < 1$ is the probability that an entry of the character matrix is missing.*

See Supplementary Materials for proof.

5 Phylogeny Estimation from Quartets

We now consider how to estimate phylogenies from quartets displayed by the input character data. To begin, we assume that the model tree T is binary, which is the convention for the MSC model (which generates gene trees) and the neutral WF+IS model (which generates 2-state characters). For both of these models, there are no anomalous quartets, motivating the development of quartet-based methods for species tree estimation. By quartet-based methods, we mean heuristics for the Maximum Quartet Support Supertree (MQSS) problem (Wilkinson et al. 2005) (also see Section 7.7 and in Warnow 2017).

Definition 1. *The MQSS problem is defined by*

Input: *A set of unrooted trees $\mathcal{P} = \{T_1, T_2, \ldots, T_k\}$, with tree T_i on leaf label set X_i*

Output: *A unrooted, binary tree B on leaf label set $\cup_{i=1}^{k} X_i$ that maximizes $\sum_{q \in Q(B)} w(q, \mathcal{P})$, where $w(q, \mathcal{P})$ is the number of input trees in $\mathcal{P}$ that display q*

The MQSS problem is NP-hard (Jiang et al. 2001). When all input trees are on the same leaf label set, MQSS becomes weighted quartet consensus (WQC). The optimal solution to WQC is a consistent estimator of the unrooted model tree topology under the MSC model (Theorem 2 in Mirarab et al. 2014). Although

WQC is NP-hard (Lafond and Scornavacca 2019), fast and accurate heuristics have been developed, with the most well-known heuristic being ASTRAL (Mirarab et al. 2014; Mirarab and Warnow 2015; Zhang et al. 2018). ASTRAL allows the input trees to be incomplete—and is consistent under the MSC model in this case, provided that the missingness is unbiased (Nute et al. 2018).

We consider two main issues when using quartet-based methods to reconstruct cell lineage trees. The first issue is that the input to MQSS/WQC are unrooted trees rather than 2-state characters. This issue was addressed by (Springer et al. 2019; Molloy et al. 2021), who observe that 2-state characters can be viewed as unrooted trees with at most one internal branch (Figure 3). Given this transformation, it is possible to run ASTRAL and other quartet-based methods on character data sets.

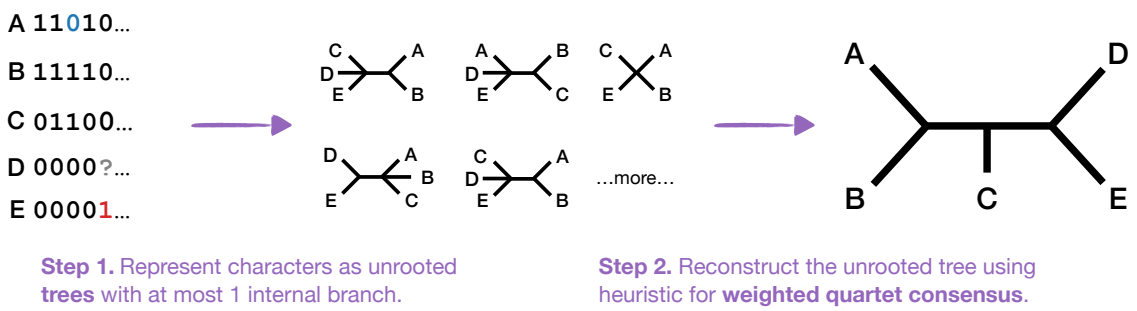

Figure 3: This figure shows a 2-state character matrix with one false negative (in blue), one false positive (in red), and one missing entry. The first step indicates how each character (column in the matrix) corresponds to an unrooted tree with at most one internal branch. The second step is to estimate the phylogeny by applying a quartet-based method. This is motivated by our result that there are no anomalous quartets under the IS+UEM model (Theorem 1).

The second issue is that MQSS/WQC seek a binary tree; however, this restriction is unreasonable for cell lineage trees given the clonal model of tumor evolution (see review by Schwartz and Schäffer 2017). In need, methods for tumor phylogenetics are often evaluated on data simulated from highly unresolved trees (Kizilkale et al. 2022). Our next result suggests that MQSS/WQC are still relevant when reconstructing cell lineage trees.

Theorem 5. *Let $\mathcal{C}$ be a set of 2-state characters generated from N under the $\mathcal{M}$+UEM model. Suppose that N is non-binary and that there are no anomalous quartets under $\mathcal{M}$. Then, as the number of characters in $\mathcal{C}$ goes to infinity, the optimal solution to MQSS given $\mathcal{C}$ will be a refinement of $u(N)$ with high probability. In other words, the optimal solution to MQSS given $\mathcal{C}$ will contain zero false negative branches with high probability (on the other hand, it is guaranteed to contain false positive branches).*

Proof. Let $L(N)$ be the leaf label set associated with N. Let B and T be unrooted, binary trees on $L(N)$ such that B is a refinement of $u(N)$ and T is NOT (note that B is a refinement of $u(N)$ if B can be obtained from $u(N)$ in a sequence of refinement operations). For the optimal solution to MQSS given $\mathcal{C}$ to be a

refinement of $u(N)$, we need to show that for any pair B and T

$$\sum_{q \in Q} w_{\mathcal{C}}(q, B) > \sum_{q \in Q} w_{\mathcal{C}}(q, T) \tag{8}$$

where Q is the set of all possible quartets on $L(N)$ and $w_{\mathcal{C}}(q, M)$ is a function that returns the number of characters in $\mathcal{C}$ that display quartet q if $q \in Q(M)$ and returns 0 if $q \notin Q(M)$.

For every subset X of four leaves in $L(N)$, there are three possible quartets: q_1, q_2, and q_3. Exactly one of these quartets is displayed by B and by T. When B and T display the same quartet, say q_1, then $w_{\mathcal{C}}(q_1, B) = w_{\mathcal{C}}(q_1, T)$ and $w_{\mathcal{C}}(q_2, B) = w_{\mathcal{C}}(q_2, T) = w_{\mathcal{C}}(q_3, B) = w_{\mathcal{C}}(q_3, T) = 0$. Thus, this case does not impact which tree has a higher score. The case where B and T display different quartets could impact which tree has a higher score. Without loss of generality, suppose B displays q_1 and T displays q_2. If $q_2 \in Q(u(N))$, it must also be in $Q(B)$ as $Q(u(N)) \subset Q(B)$; thus $q_2 \notin Q(u(N))$. Therefore, if $q_1 \in Q(u(N)) \subset Q(B)$, $\mathbb{P}(q_1|\alpha, \beta, \gamma, \sigma) > \mathbb{P}(q_2|\alpha, \beta, \gamma, \sigma)$ under the $\mathcal{M}$+UEM model by Theorem 2. Alternatively, if $q_1 \notin Q(u(N)) \subset Q(B)$, then N restricted to X is a star; it follows from Theorem 2 that $\mathbb{P}(q_1|\alpha, \beta, \gamma, \sigma) = \mathbb{P}(q_2|\alpha, \beta, \gamma, \sigma)$ under the $\mathcal{M}$+UEM model. Putting this together, as $|\mathcal{C}| \to \infty$, $w_{\mathcal{C}}(q_1, B) \geq w_{\mathcal{C}}(q_2, T)$ and $w_{\mathcal{C}}(q_2, B) = w_{\mathcal{C}}(q_3, B) = w_{\mathcal{C}}(q_1, T) = w_{\mathcal{C}}(q_3, T) = 0$, with high probability. Thus, with high probability,

$$\sum_{q \in \{q_1, q_2, q_3\}} w_{\mathcal{C}}(q, B) \geq \sum_{q \in \{q_1, q_2, q_3\}} w_{\mathcal{C}}(q, T) \tag{9}$$

Equation 9 holds for all subsets of four leaves and is an inequality for at least one of these subsets, as $Q(T) \neq Q(u(N)) \subset Q(B)$. This proves Equation 8. □

We now consider the split-constrained Maximum Quartet Support Supertree (sc-MQSS) problem (Bryant and Steel 2001; Mirarab et al. 2014).

Definition 2. *The sc-MQSS problem is defined by*

Input: *A set of unrooted trees $\mathcal{P} = \{T_1, T_2, \ldots, T_k\}$, with tree T_i on leaf label set X_i and a set Σ of bipartitions of $X = \cup_{i=1}^{k} X_i$*

Output: *A unrooted, binary tree B on leaf label set X such that $Bip(B) \subseteq \Sigma$ and $\sum_{q \in Q(B)} w(q, \mathcal{P})$ is maximized.*

Note that when all input trees are on the same leaf label set, sc-MQSS becomes sc-WQC.

Theorem 6. *Suppose the conditions of Theorem 5 hold. Then, as the number of characters in $\mathcal{C}$ goes to infinity, ASTRAL given $\mathcal{C}$ returns a refinement of $u(N)$ with high probability, provided the every character (i.e., bipartition of X) occurs with non-zero probability under $\mathcal{M}$+UEM model.*

Proof. ASTRAL operates by (1) setting Σ so that it contains all bipartitions induced by the input "trees" and then (2) solving sc-MQSS for $(\mathcal{C}, \Sigma)$ exactly. If every character (i.e., bipartition of X) has non-zero probability under the $\mathcal{M}$+UEM model, then as the number of characters in $\mathcal{C}$ goes to infinity, Σ will contain

all bipartitions induced by one or more refinements of $u(N)$ with high probability. Then, by Theorem 5, ASTRAL will return a refinement of N with high probability, giving us our result. □

Note that every character is possible when $0 < \alpha, \beta < 1$ or when $\mathcal{M}$ is the WF+IS model.

Corollary 3. *Suppose the conditions of Theorem 5 hold except that N is binary. Then, the optimal solution to MQSS is a consistent estimator of $u(N)$. Moreover, the ASTRAL method is statistically consistent under $\mathcal{M}$+UEM model, provided the every character occurs with non-zero probability under $\mathcal{M}$+UEM model.*

Proof. The result easily follows from the proof of Theorem 5 because a binary tree can only refine itself so $B = N$. □

Corollary 3 addresses an open question from Molloy et al. (2021) about the impact of error and missingness on species tree estimation from low-homoplasy retroelement insertions under the WF+IS model. It is also worth noting that the results above for non-binary model trees extend to data generated under the MSC model. Of course, we are primarily interested in cell lineage tree reconstruction from 2-state characters where error and missingness are unbiased. Theorems 5 and 6 suggest the potential of quartet-based methods in this context, assuming false positive branches in the output tree can be tolerated or dealt with effectively; we come back to this issue in the discussion.

Lastly, we derive related results for triplets by viewing 2-state characters as a rooted trees with at most one internal branch (Supplementary Figure S2) and replacing MQSS with Maximum Triplet Support Supertrees (MTSS). Importantly, the consistency results for MTSS require there to be no anomalous triplets under the $\mathcal{M}$+UEM model (Supplemental Theorem S1 and Supplemental Corollary S1). This can be achieved by setting $\alpha = 0$, suggesting triplet-based methods are less robust to error than quartet-based methods.

6 Discussion

Quartet-based approaches have garnered much success for estimating species phylogenies under coalescent models. Here, we consider their application for estimating cell lineage trees. There are two important differences when estimating cell lineage trees compared to species trees. First, errors and missing values can arise from single-cell sequencing and thus are typically modeled. Second, the model cell lineage tree may be highly unresolved. To address these issues, we first show that there are no anomalous quartets under the infinite sites (IS) plus unbiased error and missingness (UEM) model. We then show that under the IS+UEM model, the optimal solution to the Maximum Quartet Support Supertree (MQSS) problem is a refinement of the model cell lineage tree. Moreover, our results generalize from the IS to any model of 2-state character evolution with no anomalous quartets; this addresses an open question posed by Molloy et al. (2021) who leverage quartet-based methods to estimate species phylogenies from low-homoplasy retroelement insertions under the neutral Wright-Fisher plus IS model. Lastly, we consider the case of triplets, showing that there

can be anomalous triplets under reasonable conditions. This result suggests that quartets are more likely to be useful for reconstructing cell lineage trees than triplets.

In the context of tumor phylogenetics, we are interested in the rooted cell lineage tree (which is likely to be highly unresolved) or a mutation tree. Quartet-based methods return unrooted, binary trees; however, the resulting tree can be rooted on the edge incident to the healthy cell with no mutations. Moreover, the resulting tree can be transformed into a highly unresolved tree by contracting branches. Identifying which branches to contract is related to mapping mutations onto branches in the cell lineage tree and identifying which cells are members of the same clone or subclone. These tasks may need to be performed when using likelihood-based methods for cell lineage tree reconstruction, as binary trees are often returned such methods, including ScisTree (Wu 2019), SiFit (Zafar et al. 2017), and CellPhy (Kozlov et al. 2022) (it is worth noting that only ScisTree makes the IS assumption; the other two use finite sites).

Likelihood-based methods require explicitly estimating numeric parameters, like α and β, as well as exploring the space of cell lineage trees, which grows exponentially in the numbers of cells. In contrast, quartet-based approaches allow for error implicitly and are often based on algorithmic techniques, like divide-and-conquer, that are quite fast in practice. That being said, quartet-based methods have been designed for species phylogenetics, where the number of leaves (species) is typically much less than the number of gene trees or characters. In tumor phylogenetics, the number of leaves (cells) can be much greater than than characters. This seems likely to impact runtime, in addition to accuracy (just consider that our consistency guarantees assume an infinite number of 2-state characters). Even though the number of characters is small, the underlying model tree may be highly unresolved. This suggests that sampling different cells around the same branch is essentially providing more data (this observation is leveraged by Kizilkale et al. 2022). It would be interesting to explore the impact of data set size and unresolved model trees on quartet-based methods via simulation and analyses of biological data.

Lastly, we note that quartet-based methods, as presented here, fail to address doublets and copy number aberrations (CNAs), which also challenge cell lineage tree reconstruction. A doublet is a sequencing artifact where data provided for a single cell is really a mixture of two cells. This "hybrid" cell challenges the notion of tree-like evolution, motivating the development of methods for correcting doublets (Weber et al. 2021). CNAs include large sections of chromosomes being duplicated or lost (see review on methods for detecting CNAs by Mallory et al. 2020). CNA losses have motivated the development of many methods for reconstructing tumor phylogenies (El-Kebir 2018; Malikic et al. 2019; Satas et al. 2020). Some of these methods view CNA losses as false negatives (note that these false negatives will be biased towards particular cells and mutations). SCARLET (Satas et al. 2020), in particular, reconstructs a CNA tree and then uses it to constrain phylogeny reconstruction. Constraints are also leveraged in species phylogenetics, including with quartet-based methods (Rabiee and Mirarab 2020). As such, it seems possible that quartet-based approaches could be tailored to address practical challenges in tumor phylogenetics, and we are optimistic about their potential for cell lineage tree reconstruction.

7 Acknowledgements

This work was conducted in preparation for the talk "Models and methods for reconstructing population-level evolutionary histories and relationships to tumor phylogenetics", given at the U.S. National Cancer Institute's Spring School in Algorithmic Cancer Biology on March 15, 2023. EKM thanks the organizers (M. El Kebir, S. Malikic, T.M. Przytycka, B. Raphael, S.C. Sahinalp, and M. Singh) for the invitation and the audience for their questions and engagement.

8 Funding

This work was funded by the State of Maryland.

9 Conflicts of Interest

The authors have no competing interests to declare.

References

Allman ES, Degnan JH, Rhodes JA. 2011. Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. *Journal of Mathematical Biology* **62**: 833–862.

Bryant D, Steel M. 2001. Constructing optimal trees from quartets. *Journal of Algorithms* **38**: 237–259.

Degnan JH. 2013. Anomalous unrooted gene trees. *Systematic Biology* **62**: 574–590.

Degnan JH, Rosenberg NA. 2006. Discordance of species trees with their most likely gene trees. *PLOS Genetics* **2**: 1–7.

Dibaeinia P, Tabe-Bordbar S, Warnow T. 2021. FASTRAL: improving scalability of phylogenomic analysis. *Bioinformatics* **37**: 2317–2324.

Edwards SV. 2009. Is a new and general theory of molecular systematics emerging? *Evolution* **63**: 1–19.

El-Kebir M. 2018. SPhyR: tumor phylogeny estimation from single-cell sequencing data under loss and error. *Bioinformatics* **34**: i671–i679.

Fisher RA. 1923. Xxi.—on the dominance ratio. *Proceedings of the Royal Society of Edinburgh* **42**: 321–341.

Han Y, Molloy EK. 2023. TREE-QMC: Improving quartet graph construction for scalable and accurate species tree estimation from gene trees. *bioRxiv* DOI: 10.1101/2022.06.25.497608.

Hudson RR. 2002. Generating samples under a Wright–Fisher neutral model of genetic variation . *Bioinformatics* **18**: 337–338.

Islam M, Sarker K, Das T, Reaz R, Bayzid MS. 2020. STELAR: a statistically consistent coalescent-based species tree estimation method by maximizing triplet consistency. *BMC Genomics* **21**: 136.

Jahn K, Kuipers J, Beerenwinkel N. 2016. Tree inference for single-cell data. *Genome Biology* **17**: 86.

Jiang T, Kearney P, Li M. 2001. A polynomial time approximation scheme for inferring evolutionary trees from quartet topologies and its application. *SIAM Journal on Computing* **30**: 1942–1961.

Kizilkale C, Mehrabadi FR, Azer ES, Pérez-Guijarro E, Marie KL, Lee MP, Day CP, Merlino G, Ergün F, Buluç A, Sahinalp SC, Malikić S. 2022. Fast intratumor heterogeneity inference from single-cell sequencing data. *Nature Computational Science* **2**: 577–583.

Korstian J, Paulat N, Platt II R, Stevens R, Ray D. 2022. Sine-based phylogenomics reveal extensive introgression and incomplete lineage sorting in myotis. *Genes* **13**: 399.

Kozlov A, Alves JM, Stamatakis A, Posada D. 2022. CellPhy: accurate and fast probabilistic inference of single-cell phylogenies from scdna-seq data. *Genome Biology* **23**: 37.

Kubatko LS, Degnan JH. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic Biology* **56**: 17–24.

Kuritzin A, Kischka T, Schmitz J, Churakov G. 2016. Incomplete lineage sorting and hybridization statistics for large-scale retroposon insertion data. *PLOS Computational Biology* **12**: 1–20.

Lafond M, Scornavacca C. 2019. On the weighted quartet consensus problem. *Theoretical Computer Science* **769**: 1–17.

Larget BR, Kotha SK, Dewey CN, Ané C. 2010. BUCKy: Gene tree/species tree reconciliation with bayesian concordance analysis. *Bioinformatics* **26**: 2910–2911.

Liu L, Yu L, Edwards SV. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evolutionary Biology* **10**: 302.

Maddison W. 1997. Gene trees in species trees. *Systematic Biology* **46**: 523–536.

Mahbub M, Wahab Z, Reaz R, Rahman MS, Bayzid MS. 2021. wQFM: highly accurate genome-scale species tree estimation from weighted quartets. *Bioinformatics* **37**: 3734–3743.

Malikic S, Mehrabadi FR, Ciccolella S, Rahman MK, Ricketts C, Haghshenas E, Seidman D, Hach F, Hajirasouliha I, Sahinalp SC. 2019. PhISCS: a combinatorial approach for subperfect tumor phylogeny reconstruction via integrative use of single-cell and bulk sequencing data. *Genome Research* **29**: 1860–1877.

Mallory X, Edrisi M, Navin N, Nakhleh L. 2020. Methods for copy number aberration detection from single-cell dna-sequencing data. *Genome Biology* **21**: 208.

Mendes FK, Hahn MW. 2017. Why Concatenation Fails Near the Anomaly Zone. *Systematic Biology* **67**: 158–169.

Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson MS, Warnow T. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* **30**: i541–i548.

Mirarab S, Warnow T. 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* **31**: i44–i52.

Molloy EK, Gatesy J, Springer MS. 2021. Theoretical and practical considerations when using retroelement insertions to estimate species trees in the anomaly zone. *Systematic Biology* **71**: 721–740.

Nute M, Chou J, Molloy EK, Warnow T. 2018. The performance of coalescent-based species tree estimation methods under models of missing data. *BMC Genomics* **19(Suppl 5)**: 286.

Pamilo P, Nei M. 1988. Relationships between gene trees and species trees. *Molecular Biology and Evolution* **5**: 568–583.

Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D. 2012. Ancient Admixture in Human History. *Genetics* **192**: 1065–1093.

Peter BM. 2016. Admixture, Population Structure, and F-Statistics. *Genetics* **202**: 1485–1501.

Pickrell JK, Pritchard JK. 2012. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLOS Genetics* **8**: 1–17.

Rabiee M, Mirarab S. 2020. Forcing external constraints on tree inference using astral. *BMC Genomics* **21**: 218.

Rannala B, Yang Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* **164**: 1645–1656.

Roch S, Steel M. 2015. Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. *Theoretical Population Biology* **100**: 56–62.

Ross EM, Markowetz F. 2016. OncoNEM: inferring tumor evolution from single-cell sequencing data. *Genome Biology* **17**: 69.

Satas G, Zaccaria S, Mon G, Raphael BJ. 2020. SCARLET: Single-cell tumor phylogeny inference with copy-number constrained mutation losses. *Cell Systems* **10**: 323–332.e8.

Schwartz R, Schäffer AA. 2017. The evolution of tumour phylogenetics: principles and practice. *Nature Reviews Genetics* **18**: 213–229.

Springer MS, Molloy EK, Sloan DB, Simmons MP, Gatesy J. 2019. ILS-aware analysis of low-homoplasy retroelement insertions: Inference of species trees and introgression using quartets. *Journal of Heredity* **111**: 147–168.

Warnow T. 2017. *Computational Phylogenetics: An Introduction to Designing Methods for Phylogeny Estimation*. Cambridge University Press, Cambridge, United Kingdom.

Weber LL, Sashittal P, El-Kebir M. 2021. doubletD: detecting doublets in single-cell DNA sequencing data. *Bioinformatics* **37**: i214–i221.

Wilkinson M, Cotton JA, Creevey C, Eulenstein O, Harris SR, Lapointe FJ, Levasseur C, Mcinerney JO, Pisani D, Thorley JL. 2005. The Shape of Supertrees to Come: Tree Shape Related Properties of Fourteen Supertree Methods. *Systematic Biology* **54**: 419–431.

Wright S. 1931. Evolution in mendelian populations. *Genetics* **16**: 97–159.

Wu Y. 2019. Accurate and efficient cell lineage tree inference from noisy single cell data: the maximum likelihood perfect phylogeny approach. *Bioinformatics* **36**: 742–750.

Zafar H, Tzen A, Navin N, Chen K, Nakhleh L. 2017. SiFit: inferring tumor trees from single-cell sequencing data under finite-sites models. *Genome Biology* **18**: 178.

Zhang C, Rabiee M, Sayyari E, Mirarab S. 2018. ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformics* **19**: 153.