# Quartets enable statistically consistent estimation of 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 11, 2023

**Abstract**

Cancer progression and treatment can be informed by reconstructing its evolutionary history from tumor cells. However, traditional methods assume the input data are error-free and the output tree is fully resolved. These assumptions are challenged in tumor phylogenetics because single-cell sequencing produces sparse, error-ridden data and because tumors evolve clonally. Here, we find that methods based on quartets (four-leaf, unrooted trees) withstand these barriers. We consider a popular tumor phylogenetics model, in which mutations arise on a (highly unresolved) tree and then (unbiased) errors and missing values are introduced. Quartets are implied by mutations present in two cells and absent from two cells. Our main result is that the most probable quartet identifies the unrooted model tree on four cells. This motivates seeking a tree such that the number of quartets shared between it and the input mutations is maximized. We prove an optimal solution is a consistent estimator of the unrooted cell lineage tree; this guarantee includes the case where the model tree is highly unresolved, with error defined as the number of false negative branches. Lastly, we outline how quartet-based methods might be employed when there are copy number aberrations and other challenges specific to tumor phylogenetics.

## 1 Introduction

Cancer progression and treatment can be informed by reconstructing the evolutionary history of tumor cells (Lim et al. 2020). Although many methods exist to estimate evolutionary trees (called phylogenies) from molecular sequences, traditional approaches assume the input data are error-free and the output tree is fully resolved. These assumptions are challenged in tumor phylogenetics because single-cell sequencing produces sparse, error-ridden data and because tumors evolve clonally so the underlying tree is highly unresolved (Jahn

et al. 2016; Schwartz and Schäffer 2017). Here, we demonstrate that quartet-based methods withstand these barriers.

A *quartet* is an unrooted, phylogenetic tree with four leaves. Quartets have long been used as the building blocks for reconstructing the evolutionary history of species (Wilkinson et al. 2005). The reason quartet-based methods have garnered such success in species phylogenetics is their good statistical properties under the Multi-Species Coalescent (MSC) model (Pamilo and Nei 1988; Rannala and Yang 2003). An MSC model species tree generates gene trees (note that a gene tree reflects the genealogical history of a gene, which is passed down from ancestor to descendant, whereas the species tree governs the pool of potential ancestors). One of the most important results from the last decade of systematics states: for four species, the most probable unrooted gene tree under the MSC is topologically equivalent to the unrooted model species tree (Allman et al. 2011). Notably, for trees with more than four leaves, the most probable unrooted gene tree can be topologically discordant with the unrooted model species tree (Degnan 2013). 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).

The statistical theory above has motivated the development of quartet-based methods (e.g., Larget et al. 2010; Mirarab et al. 2014) and is central to their proofs of statistical consistency under the MSC. ASTRAL (Mirarab et al. 2014), in particular, has become a gold standard approach to multi-locus species tree estimation. 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; Zhang and Mirarab 2022). Similar theory and methodology has been given for *triplets*: three-leaf, rooted, phylogenetic trees (Degnan and Rosenberg 2006; Liu et al. 2010; Islam et al. 2020).

Inspired by these efforts, we pursue related problems for a popular tumor phylogenetics model (Jahn et al. 2016; Ross and Markowetz 2016; Wu 2019; Kizilkale et al. 2022), in which mutations arise on a (highly unresolved) cell lineage tree according to the infinite sites model, and then errors and missing values are introduced to the resulting mutation data in an unbiased fashion. The idea is that deviations from a perfect phylogeny can be attributed to sequencing errors, as data produced by single-cell protocols are notoriously error-prone and sparse. Although the infinite sites plus unbiased error and missingness (IS+UEM) model generates mutations rather than gene trees, quartets are implied by mutations that are present in two cells and absent from two cells.

Our main result is that there are no anomalous quartets under the IS+UEM model; this motivates seeking a cell lineage tree such that the number of quartets shared between it and the input mutations is maximized. We prove an optimal solution to this problem is a consistent estimator of the unrooted cell lineage tree; this guarantee extends to the case of highly unresolved model trees, with error defined as the number of false negative branches. Somewhat surprisingly, our positive results for quartets do not extend to triplets, as there can be anomalous triplets under the IS+UEM model under reasonable conditions. The theory

presented here generalizes beyond the IS model, making it relevant to several recent systematic studies of placental mammals, birds, and bats. Of course, our main focus is tumor phylogenetics, which is additionally challenged by copy number aberrations and doublets. We conclude by outlining how quartet-based methods might be employed given these unique challenges.

# 2 Preliminaries

We now provide some background on phylogenetic trees, models of evolution, and statistical consistency.

## 2.1 Phylogenetic Trees

A *phylogenetic tree* is defined by the tripple $(g, X, \phi)$, where $g$ is a connected acyclic graph, $X$ is a set of labels (often representing species or cells), and $\phi$ 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 use $u(T)$ to denote the unrooted version of a rooted tree $T$. 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 a tree 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*. We use the phrase *highly unresolved* to indicate that $T$ contains many polytomies and/or that the polytomies in $T$ have high degrees.

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 leaves (i.e., deleting the other leaves from $T$ and then 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, restricting $T$ to some subsets of four labels will not produce a quartet. Some selections will produce star trees, which do not provide any topological information. We use $T|_S$ to denote $T$ restricted to label set $S$ (note that if branch parameters are associated with $T$, they are added together when suppressing vertices of degree 2).

The concepts above for quartets extend to triplets. A *triplet* is a rooted, binary tree with three leaves, and 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 label set $X$ partitions it into two disjoint subsets. It is easy to see that each edge in an unrooted tree induces a bipartition, and we use $Bip(T)$ to denote the set of bipartitions induced by all edges in $T$.

## 2.2 Mutations and Models of Evolution

A mutation matrix $M$ is an $n \times k$ matrix, where $n$ is the number of rows (representing cells or species) and $k$ is the number of columns (representing mutations). Columns are also referred to as characters or site patterns. Our focus here is on 2-state characters, with $M_{i,j} = 0$ indicating that mutation $j$ is absent from cell $i$ and $M_{i,j} = 1$ indicating that mutation $j$ is present in cell $i$. In tumor phylogenetics, 0 represents the ancestral state and 1 represents mutant/derived state because mutations are called in reference to a healthy cell, which is the root of the cell lineage. However, this interpretation of states 0 and 1 is only important for triplets and not quartets.

Throughout this paper, we assume the mutation matrix $D$ is generated under a *hierarchical model* with two steps (Figure 1).

1. A mutation matrix $G$ is generated under some model $\mathcal{M}$, parameterized by a rooted phylogenetic tree topology $\sigma$ and a set $\Theta$ of associated numeric parameters. Importantly, model $\mathcal{M}$ given $(T, \Theta)$ defines a probability distribution on mutation patterns, and we assume mutations in $G$ are *independent and identically distributed (i.i.d.)* according to this model. For simplicity of notation, we typically omit the dependence on the numeric parameters in $\Theta$.

2. Errors and missing values are introduced to the ground truth matrix $G$ according to the UEM model (described below). This result of this process is the observed matrix $D$.

Hierarchical models of this form, denoted $\mathcal{M}$+UEM, define a probability distribution on mutation patterns given their parameters. Thus, if we say that $D$ is generated under the $\mathcal{M}$+UEM model, then we assume the mutations in $D$ are *i.i.d.* according to this model. We now describe the data generation steps in greater detail for a popular tumor phylogenetics model (Jahn et al. 2016; Ross and Markowetz 2016; Wu 2019; Kizilkale et al. 2022).

**Step 1: Infinite Sites (IS) model.** For tumor phylogenetics, we take $\mathcal{M}$ to be the infinite sites (IS) model, so mutations are generated under the IS model given a rooted cell lineage tree $\sigma$ and a set $\Theta$ of edge probabilities that sum to 1. Specifically, every edge $e$ in $\sigma$ is associated with a numeric value $p(e) \in \Theta$, indicating the probability that a mutation occurs on $e$. When a mutation occurs on $e$, all cells on a directed path from $e$ to any of the leaves of $\sigma$ are set to state 1; all other cells are set to state 0. Thus, a mutation corresponds to the bipartition induced by the branch on which it occurred. Edges on which mutations cannot occur are contracted so that the probability of a mutation on any edge in $\sigma$ 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.

**Step 2: Unbiased Error and Missingness (UEM) model.** If mutation matrix $G$ is generated under the IS model given $(\sigma, \Theta)$, then reconstructing $\sigma$ is trivial. However, for tumor phylogenetics, false positives and false negatives are introduced to $G$, producing the observed matrix $D$. This is done according to

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

where $0 \leq \alpha < 1$ and $0 \leq \beta < 1$ are the probability of false positives and false negatives, respectively. Simultaneously, missing values are introduced to $G$ with probability $0 \leq \gamma < 1$; this can be incorporated into the model by multiplying each of the cases in Equation 1 by $(1 - \gamma)$.

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 the 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 under the IS 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 $D$.

Our goal is to estimate cell lineage trees under the IS+UEM model. An important property for phylogeny estimation methods is whether they are statistically consistent under the model of interest.

**Definition 1.** *[Statistical Consistency; see Section 1.1 of Warnow 2017] Let $\mathcal{M}$ be some model that generates mutations, and let $D$ be a mutation matrix, with $n$ rows (cells or species) and $k$ columns (mutations), generated under $\mathcal{M}$ given tree $\sigma$. We say that an estimation method is statistically consistent under $\mathcal{M}$ if for any $\epsilon > 0$, there exists a constant $K > 0$ such that when $D$ contains at least $K$ mutations, the method given*

*$D$ returns $\sigma$ with probability at least $1-\epsilon$. Alternatively, we might say that the error in the tree estimated from $D$ is zero with probability at least $1-\epsilon$.*

The idea is that as the number $k$ of mutations goes towards infinity, the error in the estimated tree is zero with high probability. Tree error is typically defined as the number of *false negative branches* (i.e., branches in $\sigma$ that are missing from the estimated tree) plus the number of *false positive branches* (i.e., branches in the estimated tree that are missing from $\sigma$).

# 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. To begin, we assume that the rooted cell lineage tree $\sigma$ has four leaves; therefore, it must have one of five topologies shown in Figure 2. Two of them display a star when unrooted, and the other three correspond to a quartet when unrooted. Quartets are also implied by the mutation matrix $D$, which is on the same four cells as $\sigma$. There are 16 possible patterns when considering a single mutation on cells $\{A, B, C, D\}$. A quartet is implied by two cells being in state 1 and two cells being in state 0. Two patterns ($ABCD = 0011$ 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 do provide topological information. Thus, we can denote the probability of quartets under IS model given $\sigma$ 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 $\mathbb{P}(q_3|\sigma) = \mathbb{P}(1001|\sigma) + \mathbb{P}(0110|\sigma)$. Now we turn to computing the probability of quartets under the IS+UEM model. The following lemma will later prove useful.

**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.** *Suppose that $\sigma$ is an IS model tree with four leaves and that $\alpha+\beta \neq 1$. 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 $\sigma$ unrooted is $q_1$ and $\mathbb{P}(q_1|\alpha,\beta,\sigma) = \mathbb{P}(q_2|\alpha,\beta,\sigma) = \mathbb{P}(q_3|\alpha,\beta,\sigma)$ if the $\sigma$ unrooted is a star. Thus, there are no anomalous quartets under the IS+UE model.*

*Proof.* Across all four-leaf model trees (Figure 2), there are nine mutation patterns that can be generated under the IS model and thus may be present in $G$. These are the two invariant patterns (0000 and 1111), five of the eight variant but uninformative patterns (1000, 0100, 0010, 0001, and 1110), and the two informative patterns for $u(\sigma) = q_1$ (1100 and 0011); see Supplementary Table S1. For each pattern $g$ listed above, we

enumerate all possible ways of introducing false positives and false negatives; this gives us the probability of mutation patterns under the UE model given $(\alpha, \beta, g)$; see Table 1 and Supplementary Tables S2–S7. Putting everything together, we find that the probability of quartet $q$ under the IS+UE model is

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

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 $j \in \{2,3\}$.

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

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 $\sigma$ is a star (i.e, $\sigma \in \{\sigma_4, \sigma_5\}$); the latter will be the case otherwise (i.e, $\sigma \in \{\sigma_1, \sigma_2, \sigma_3\}$). □

Note that the quantity $\alpha + \beta$ is unlikely to equal 1 in practice, as both probabilities should be less than 0.5. We now extend the result above to address missing data.

**Corollary 1.** *There are no anomalous quartets under the IS+UEM model, assuming that $\alpha + \beta \neq 1$.*

*Proof.* Let $x$ be a mutation pattern with four states, each corresponding to a leaf in the model tree. If one or more of the values in $x$ is missing, then no quartet is implied. 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. □

Our final extension generalizes the result to other mutation models for which there are no anomalous quartets.

**Theorem 2.** *Let $\mathcal{M}$ be some model that generates mutations given tree $\sigma$. Suppose that $\sigma$ has four leaves and that under model $\mathcal{M}$, (1) $\mathbb{P}(q_1|\sigma) > \mathbb{P}(q_2|\sigma)$ and $\mathbb{P}(q_1|\sigma) > \mathbb{P}(q_3|\sigma)$ if $\sigma$ unrooted is $q_1$ and (2) $\mathbb{P}(q_1|\sigma) = \mathbb{P}(q_2|\sigma) = \mathbb{P}(q_3|\sigma)$ if $\sigma$ unrooted is a star. Then, these relationships hold under the $\mathcal{M}$+UEM model, assuming that $\alpha + \beta \neq 1$. To summarize, if there are no anomalous quartets under $\mathcal{M}$, then there are no anomalous quartets under $\mathcal{M}$+UEM.*

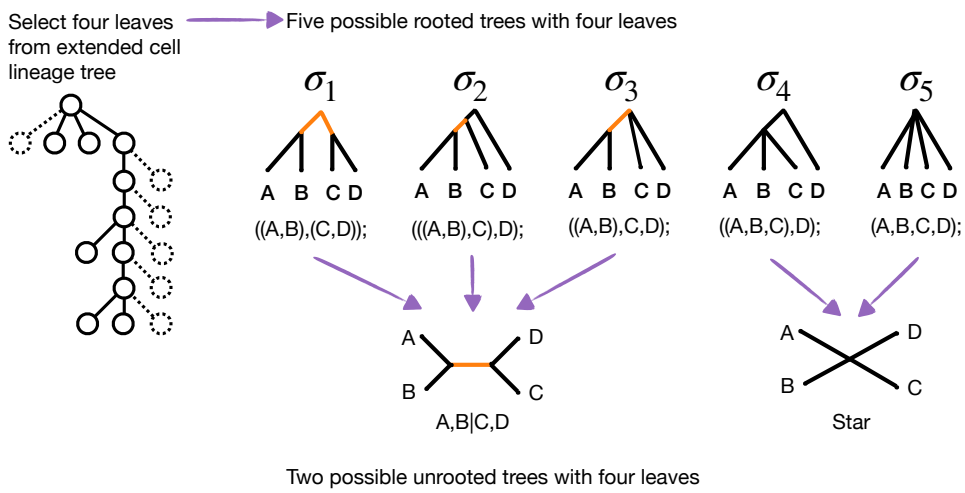


Figure 2: There are five possible trees with four leaves. Three of these trees ($\sigma_1$, $\sigma_2$, and $\sigma_3$) have a non-trivial unrooted topology (quartet: $A, B|C, D$). The other two trees ($\sigma_4$ and $\sigma_5$) have a trivial unrooted topology (star). Looking at Figure 1, $\sigma_1$–$\sigma_5$ 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.

| # of FP | # of FN | $\mathbb{P}(D_{*,j}\|\alpha, \beta, G_{*,j})$ | $D_{*,j}$ derived from $G_{*,j} = \mathbf{1100}$ | $D_{*,j}$ derived from $G_{*,j} = \mathbf{0011}$ | Quartet supported by $D_{*,j}$ |
|---|---|---|---|---|---|
| 0 | 0 | $(1-\beta)^2(1-\alpha)^2$ | 1100 | 0011 | $A, B\|C, D$ |
| 0 | 1 | $\beta(1-\beta)(1-\alpha)^2$ | 0100 | 0001 | none |
| | | | 1000 | 0010 | none |
| 0 | 2 | $\beta^2(1-\beta)^2$ | 0000 | 0000 | none |
| 1 | 0 | $(1-\beta)^2\alpha(1-\alpha)$ | 1110 | 1011 | none |
| | | | 1101 | 0111 | none |
| 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 | none |
| | | | 0001 | 0100 | none |
| 2 | 0 | $(1-\beta)^2\alpha^2$ | 1111 | 1111 | none |
| 2 | 1 | $\beta(1-\beta)\alpha^2$ | 0111 | 1101 | none |
| | | | 1011 | 1110 | none |
| 2 | 2 | $\beta^2\alpha^2$ | 0011 | 1100 | $A, B\|C, D$ |

Table 1: Above we show the mutation patterns with four cells $\{A, B, C, D\}$ that can be generated under the UE model given **pattern #12** ($G_{*,j} = 1100$) and **pattern #3** ($G_{*,j} = 0011$) as well as their probabilities. The red values indicate a false positive introduced to $G_{*,j}$ by flipping 0 to 1. The blue values indicate a false negative introduced to $G_{*,j}$ by flipping 1 to 0. Similar tables for the other 14 patterns are provided in the Supplementary Materials.

See Supplementary Materials for proof, in which we show that the difference in probability between quartet $q_1$ and an alternative quartet $q_j$ equals

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

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 $j \in \{2,3\}$, so we get our previous result (Theorem 1 and Corollary 1). Observe that the difference in quartet probabilities (Equation 4) depends on (1) the probability of false positives and negatives (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 given $\sigma$. In the simulations performed by Kizilkale et al. (2022), the largest values of $\alpha$, $\beta$, and $\gamma$ were 0.001, 0.2, and 0.05, respectively. This puts the difference in probabilities at $0.52 \cdot \mathbb{P}(q_1|\sigma)$. The magnitude of $\mathbb{P}(q_1|\sigma)$ depends on the cell lineage tree and the four cells sampled, which we will discuss later in the context of triplets. For now, we note that in practical settings, $\mathbb{P}(q_1|\sigma)$ may have a greater impact on the difference in quartet probabilities than $\alpha$, $\beta$, or $\gamma$.

# 4 Anomalous triplets under an unbiased error and missingness model

We now extend our results to triplets. Thus, the rooted cell lineage tree $\sigma$ has three leaves, and it must have one of two topologies: binary or non-binary (Supplementary Figure S1). Triplets are also implied by the mutation matrix $D$, which is on the same three cells as $\sigma$. There are 8 possible patterns when considering a single mutation on cells $\{A,B,C\}$. A triplet is implied by two cells being in state 1 (the mutant/derived state) and one cell being in state 1 (the ancestral state). One pattern ($ABC = 110$) supports triplet $t_C = C|A,B$, one pattern (101) supports triplet $t_B = B|A,C$, one pattern (011) supports triplet $t_A = A|B,C$, and the other five patterns do not provide topological information. We denote the probability of triplets under model $\mathcal{M}$ given $\sigma$ as $\mathbb{P}(t_A|\sigma) = \mathbb{P}(011|\sigma)$, $\mathbb{P}(t_B|\sigma) = \mathbb{P}(101|\sigma)$, and $\mathbb{P}(t_C|\sigma) = \mathbb{P}(110|\sigma)$. Now we turn to computing the probability of triplets under the $\mathcal{M}$+UEM model.

**Theorem 3.** *Let $\mathcal{M}$ be some model that generates mutations given $\sigma$. Suppose that $\sigma$ has three leaves $\{A,B,C\}$ and that under model $\mathcal{M}$, (1) $\mathbb{P}(t_C|\sigma) > \mathbb{P}(t_A|\sigma)$ and $\mathbb{P}(t_C|\sigma) > \mathbb{P}(t_B|\sigma)$ if $\sigma = t_C$ and (2) $\mathbb{P}(t_C|\sigma) = \mathbb{P}(t_A|\sigma) = \mathbb{P}(t_B|\sigma)$ if $\sigma$ is non-binary. Then, these relationships hold under the $\mathcal{M}$+UEM model, assuming that $\alpha = 0$ or that $\alpha + \beta \neq 1$ and all singleton patterns (*100, 010, *and* 001*) occur with equal probability. Otherwise, there can be anomalous triplets under the $\mathcal{M}$+UEM model.*

*Proof.* There are eight mutation patterns with three states. With the IS assumption, we interpret these as the two invariant patterns ($ABC = 000$ and 111), three singleton patterns (100, 010, 001), and three phylogenetically informative patterns (110, 101, and 011). For each pattern $g$ listed above, we enumerate all possible ways of introducing false positives and false negatives; this gives us the probability of mutation

patterns under the UE model given $(\alpha, \beta, g)$; see Supplementary Tables S9–S16. Putting everything together, 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 $t_C$ and $t_j$ for $j \in \{A, B\}$. We find

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

Assuming that $\alpha = 0$ or that $\alpha + \beta = 0$ and $\mathbb{P}(x_C|\sigma) = \mathbb{P}(x_j|\sigma)$, this quantity is zero if $\mathbb{P}(t_C|\sigma) = \mathbb{P}(t_j|\sigma)$ and is greater than zero if $\mathbb{P}(t_C|\sigma) > \mathbb{P}(t_j|\sigma)$. By our assumptions on $\mathcal{M}$, the former will be the case if $\sigma$ is non-binary; the latter will be the case if $\sigma = t_C$. 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

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

for $j \in \{A, B\}$. □

We now consider whether we should expect there to be anomalous triplets in the context of tumor

phylogenetics, so setting $\mathcal{M}$ to be the IS model. As previously mentioned, in the simulations performed by Kizilkale et al. (2022), the largest values of $\alpha$ and $\beta$ were 0.001 and 0.2, respectively. This would put $\alpha/(1-\beta) = 0.00125$ in Equation 7, so $\mathbb{P}(x_C|\sigma) - \mathbb{P}(x_j|\sigma)$ would need to be 800 times greater than $\mathbb{P}(t_C|\sigma) - \mathbb{P}(t_j|\sigma)$ for $t_j$ to be anomalous. Although this seems drastic, as wel will show, it could occur when $\sigma$ is created by restricting a larger model tree to a subset three leaves (adding branch parameters together when suppressing vertices of degree 2). This produces a valid result when $\mathcal{M}$ is the IS model because addition represents an 'or' condition (so a mutation on this branch or on that branch will result in the same pattern when looking at only a subset of cells).

Now suppose that $\sigma$ is created by restricting the cell lineage tree in Figure 1 to cells $\{1, 4, 10\}$ (recall that in this example, mutations occur on all non-fake edges with equal probability). Thus, $\sigma$ has rooted topology $t_{10} = 10|1, 4$ and gives the following probability distribution on mutations under the IS model. For the singletons, we have $\mathbb{P}(x_1|\sigma) = 0$, $\mathbb{P}(x_4|\sigma) = 1/11$, and $\mathbb{P}(x_{10}|\sigma) = 6/11$ because they correspond to a mutation on one of the terminal branches of $\sigma$. Similarly, for triplets, we have $\mathbb{P}(t_1|\sigma) = 0$, $\mathbb{P}(t_4|\sigma) = 0$, and $\mathbb{P}(t_{10}|\sigma) = 1/11$ because they correspond to a mutation on the internal branch of $\sigma$. Again looking at Equation 7, we see that $t_1 = 1|4, 10$ will be an anomalous triplet whenever $\alpha/(1-\beta) > 1/6$, which could occur for $\beta = 0$ and $\alpha = 1/5$. Of course, these parameter settings are not realistic. Under more realistic error rates, we might expect something like $\alpha/(1-\beta) = 0.001/0.8 = 0.00125$. However, if we add 1000 descendants between cell 9 and cell 10, then $t_1$ will be anomalous even for these reasonable error rates. This leads us to conjecture that triplets may be less robust to error than quartets, especially when sampling three cells from a larger tree.

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

**Corollary 2.** *Theorem 3 holds under the $\mathcal{M}$+UEM model.*

The argument is similar to the proof of Corollary 1.

# 5 Phylogeny Estimation from Quartets

We now consider the utility of quartet-based methods for estimating cell lineage trees from mutation data. By quartet-based methods, we mean heuristics for the Maximum Quartet Support Supertree (MQSS) problem (Wilkinson et al. 2005; also see Section 7.7 in Warnow 2017).

**Definition 2.** *The Maximum Quartet Support Supertree (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 $QS_D(B) = \sum_{q \in Q(B)} w_{\mathcal{P}}(q)$, where $w_{\mathcal{P}}(q)$ is the number of input trees in $\mathcal{P}$ that display $q$*

When all input trees are on the same leaf label set, MQSS becomes weighted quartet consensus (WQC). An optimal solution to WQC is a consistent estimator of the unrooted species tree topology under the MSC

model (Theorem 2 in Mirarab et al. 2014). Although MQSS and WQC are NP-hard (Jiang et al. 2001; Lafond and Scornavacca 2019), fast and accurate heuristics have been developed, with the most well-known being ASTRAL (Mirarab et al. 2014). Since version 2 (Mirarab and Warnow 2015; Zhang et al. 2018), ASTRAL allows the input trees to be incomplete—and is still statistically consistent under the MSC model, provided some assumptions on missing data (see Nute et al. 2018 for details).

There are two notable differences when using quartet-based methods to reconstruct cell lineage trees, rather than species trees. The first difference is that input are mutations rather than unrooted (gene) trees. This issue was addressed by Springer et al. (2019), who treat mutations as unrooted trees with at most one internal branch (Figure 3). Given this transformation of the input, it is possible to run ASTRAL and other quartet-based methods on mutation data. The second difference is that ASTRAL outputs a binary tree; however, the model cell lineage tree is unlikely to be binary given the clonal model of tumor evolution (see review by Schwartz and Schäffer 2017). Indeed, 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 useful problem formulations, even when data are generated from non-binary trees.

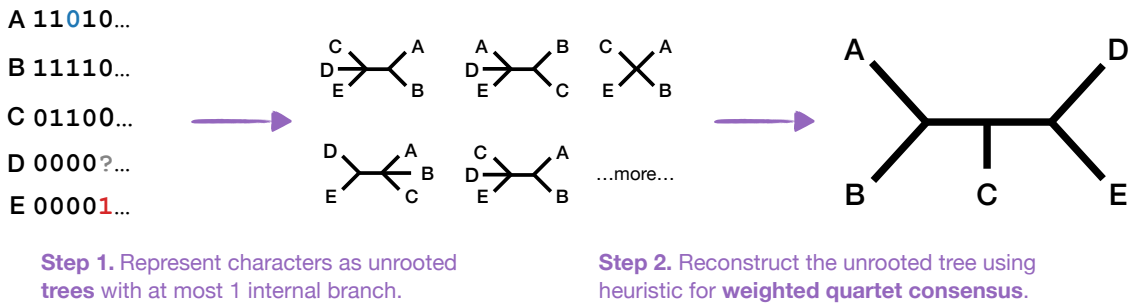


Figure 3: This figure shows a mutation matrix with one false negative (in blue), one false positive (in red), and one missing entry. The first step indicates how each mutation (column in the matrix) corresponds to an unrooted tree with at most one internal branch. The second step is to estimate the phylogeny 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).

**Theorem 4.** *Let $\mathcal{M}$ be some model that generates mutations given tree $N$ with $n$ leaves, and let $D$ be an $n \times k$ mutation matrix generated under the $\mathcal{M}$+UEM model given $(\alpha, \beta, \gamma, N)$. Suppose that $N$ has at least one internal branch when unrooted (so it can be non-binary) and that there are no anomalous quartets under $\mathcal{M}$. Then, an optimal solution to MQSS given $D$ is a consistent estimator of $u(N)$ under the $\mathcal{M}$+UEM model, assuming $\alpha + \beta \neq 1$ and with tree error defined as the number of false negative branches.*

*Proof.* Let $B$ be an unrooted, binary tree on the same label set as $u(N)$. The number of false negative branches between $B$ and $u(N)$ is zero if $B$ is a refinement of $u(N)$, meaning that $B$ can be obtained from $u(N)$ in a sequence of refinement operations (this sequence has length zero if $N$ is binary). Thus, to prove

consistency with tree error defined as the number of false negative branches, we revise Definition 1 to say that for any $\epsilon > 0$, there exists a constant $K > 0$ such that when $D$ contains at least $K$ mutations, an optimal solution to MQSS given $D$ is a refinement of $u(N)$ with probability at least $1 - \epsilon$. The remainder of the proof follows from Lemma 2. □

**Lemma 2.** *Suppose the conditions of Theorem 4 hold. Let $L(N)$ be the label set of $N$, and let $B$ and $T$ be unrooted, binary trees on $L(N)$. Suppose that $B$ is a refinement of $u(N)$ and that $T$ is NOT. Then, for any pair $B$ and $T$ and for any $\epsilon > 0$, there exists a constant $K > 0$ such that when $D$ contains at least $K$ mutations, $QS_D(B) > QS_D(T)$ with probability at least $1 - \epsilon$.*

*Proof.* To begin, we restate the inequality as

$$QS_D(B) - QS_D(T) = \sum_{S \in \mathcal{X}} w_D(B|_S) - \sum_{S \in \mathcal{X}} w_D(T|_S) > 0 \tag{8}$$

where $\mathcal{X}$ is the set of all possible subsets of four elements from $L(N)$ and $w_D(q)$ is the number of mutations in $D$ that imply quartet $q$.

**Claim 1:** First, we claim that as $k \to \infty$, $w_D/k$ converges to its expectation $F^*$ under the $\mathcal{M}$+UEM model given $(\alpha, \beta, \gamma, N)$ with probability 1. Claim 1 holds by the strong law of large numbers, as noted in the proofs of consistency for quartet and triplet-based methods under the MSC model (see Liu et al. 2010 for an example). Let $q_1^S$, $q_2^S$, $q_3^S$ denote the three possible quartets on $S$. Then, we can re-state claim 1 as follows. As $k \to \infty$, $w_D(q_i^S)/k$ converges to $F^*(q_i^S) = \mathbb{P}(q_i^S|\alpha, \beta, \gamma, N)$ with probability 1 for all $i \in \{1, 2, 3\}$ and for all $S \in \mathcal{X}$.

**Claim 2:** Second, we claim there exists a $\delta$ such that whenever $\|w_D/k - F_*\|_\infty < \delta$, Equation 8 holds. We show claim 2 is true for $\delta = \pi/2|\mathcal{X}|$, where

$$\pi = \min_{>0, S \in \mathcal{X}} \{|F^*(q_1^S) - F^*(q_2^S)|, |F^*(q_1^S) - F^*(q_3^S)|, |F^*(q_2^S) - F^*(q_3^S)|\}. \tag{9}$$

There are three cases to consider.

*Case 1:* Let $\mathcal{W} \subset \mathcal{X}$ include all $S \in \mathcal{X}$ such that $B|_S$ and $T|_S$ are the same quartet. Then, $w_D(B|_S) = w_D(T|_S)$ for all $S \in \mathcal{W}$, giving us

$$\sum_{S \in \mathcal{W}} \left( \frac{w_D(B|_S)}{k} - \frac{w_D(T|_S)}{k} \right) = 0. \tag{10}$$

Now suppose that $B$ and $T$ restricted to $S$ display different quartets, denoted $q_1^S$ and $q_2^S$, respectively. If $q_2^S \in Q(u(N))$, it must also be in $Q(B)$ as $Q(u(N)) \subset Q(B)$. Therefore, $q_2^S \notin Q(u(N))$. This gives us two additional cases to consider.

*Case 2:* Let $\mathcal{Y} \subset \mathcal{X}$ include all $S \in \mathcal{X}$ such that $B|_S$ and $T|_S$ are different quartets, denoted $q_1^S$ and $q_2^S$, respectively, with $q_1^S \in Q(u(N))$ and $q_2^S \notin Q(u(N))$. Then, for all $S \in \mathcal{Y}$, $F^*(q_1^S) > F^*(q_2^S)$

and $F^*(q_1^S) > F^*(q_3^S)$ by Theorem 2 given our assumptions. Therefore, whenever $\|w_D/k - F_*\|_\infty < \pi/2$, $w_D(B|_S) > w_D(T|_S)$ for all $S \in \mathcal{Y}$, giving us

$$\sum_{S \in \mathcal{Y}} \left( \frac{w_D(B|_S)}{k} - \frac{w_D(T|_S)}{k} \right) > 0. \tag{11}$$

*Case 3 (only needed for non-binary $N$):* Let $\mathcal{Z} \subset \mathcal{X}$ include all $S \in \mathcal{X}$ such that $B|_S$ and $T|_S$ are different quartets, denoted $q_1^S$ and $q_2^S$, respectively, with $q_1^S, q_2^S \notin Q(u(N))$. By our assumptions on $B$, $N|_S$ is a star, so $F^*(q_1^S) = F^*(q_2^S) = F^*(q_3^S)$ by Theorem 2 given all of our assumptions. However, even as $k \to \infty$, we are not guaranteed to get an exact equality $w_D(q_1^S) = w_D(q_2^S) = w_D(q_3^S)$. Thus, we need to put an upper bound $\delta$ on $\|w_D/k - F_*\|_\infty$ so that Equation 8 holds even when $w_D(B|_S) < w_D(T|_S)$ for all $S \in \mathcal{Z}$. This happens for $\delta = \pi/2|\mathcal{X}|$. When $\|w_D/k - F_*\|_\infty < \pi/2|\mathcal{X}|$, we have

$$\sum_{S \in \mathcal{Z}} \left| \frac{w_D(B|_S)}{k} - \frac{w_D(T|_S)}{k} \right| < |\mathcal{Z}| \cdot \frac{\pi}{|\mathcal{X}|} < |\mathcal{Y}| \cdot \left( \pi - \frac{\pi}{|\mathcal{X}|} \right) < \sum_{S \in \mathcal{Y}} \left( \frac{w_D(B|_S)}{k} - \frac{w_D(T|_S)}{k} \right) \tag{12}$$

because $(\mathcal{Y} + \mathcal{Z})/\mathcal{X} \leq 1$ and $1 \leq |\mathcal{Y}|$ (otherwise $T$ is a refinement of $u(N)$ or $u(N)$ has no internal branches).

*Putting the cases together:* If $u(N)$ is binary, then $\mathcal{W}|\mathcal{Y}$ is a partition $\mathcal{X}$, so we can combine equations 10 and 11 to get Equation 8. If $u(N)$ is non-binary, then $\mathcal{W}|\mathcal{Y}|\mathcal{Z}$ is a partition of $\mathcal{X}$, so we can combine Equations 10 and 12 to get Equation 8.

**Wrap up.** By claim 1, for any $\epsilon > 0$, there exists a constant $K > 0$ such that when $D$ contains at least $K$ mutations, $\|w_D/k - F^*\|_\infty < \pi/2|\mathcal{X}|$ with probability at least $1 - \epsilon$. Then, by claim 2, for any $\epsilon > 0$, when $D$ contains at least $K$ mutations, $QS_D(B) > QS_D(T)$ with probability at least $1 - \epsilon$. □

A related result can be derived for triplets by viewing mutations as a rooted trees with at most one internal branch (Supplementary Figure S2).

**Theorem 5.** *Suppose that $N$ has at least one internal branch (so it can be non-binary) and that there are no anomalous triplets under $\mathcal{M}$. Then, an optimal solution to Maximum Triplet Support Supertree (MTSS) problem given $D$ is a consistent estimator of $N$ under the $\mathcal{M}$+UEM model, assuming $\alpha = 0$ and with tree error defined as the number of false negative branches.*

The proof easily follows from Theorem 4 and Lemma 2 but replacing quartet with triplet and rooted with unrooted. An important difference is that we require $\alpha = 0$ (instead of $\alpha + \beta \neq 1$) so that there are no anomalous triplets under the $\mathcal{M}$+UEM model (Theorem 3).

**Theorem 6.** *Suppose the conditions of Theorem 4 hold and that every complete mutation pattern occurs with non-zero probability under the $\mathcal{M}$+UEM model. Then, ASTRAL given $D$ is statistically consistent under $\mathcal{M}$+UEM model, with tree error defined as the number of false negative branches.*

*Proof.* ASTRAL solves MQSS exactly within a constrained version of the solution space, denoted $\Sigma$. Its algorithm has two main steps: first, set $\Sigma$ so that it contains all bipartitions induced by the input "trees"

(i.e., mutations in $D$), and second find a solution $B$ to MQSS such that $Bip(B) \subseteq \Sigma$. Because all mutation patterns (without missing values) have non-zero probability under the $\mathcal{M}$+UEM model, for any $\epsilon_1 > 0$, there exists a constant $K_1 > 0$ such that when $k \geq K_1$, $\Sigma$ will contain all bipartitions induced by at least one refinement of $u(N)$ with probability $1 - \epsilon_1$. By Lemma 2, for any $\epsilon_2 > 0$, there exists a constant $K_2 > 0$ such that when $k \geq K_2$, $QS_D(B) > QS_D(T)$ for any pair $B$ and $T$ of unrooted binary trees on $L(N)$ such that $B$ is a refinement of $u(N)$ and $T$ is NOT. Now let $\epsilon > 0$ and select $\epsilon_1, \epsilon_2 > 0$ such that $\epsilon_1 + \epsilon_2 < \epsilon$. Then, when $D$ contains at least $\max\{K_1, K_2\}$ mutations, ASTRAL given $D$ returns a refinement of $u(N)$ with probability at least $(1 - \epsilon_1)(1 - \epsilon_2) > (1 - \epsilon)$. It follows the number of false negative branches is zero with probability at least $1 - \epsilon$. □

Note that every complete mutation pattern is possible under the IS+UEM model when $0 < \alpha, \beta < 1$ and $0 \leq \gamma < 1$.

# 6 Discussion

Quartet-based approaches have garnered much success for estimating species phylogenies under the Multi-Species Coalescent (Mirarab et al. 2014; Mirarab and Warnow 2015; Zhang et al. 2018). Here, we considered their application for estimating cell lineage trees, focusing on two important differences between 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 because tumors evolve clonally. 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, which is widely used in tumor phylogenetics (this is an *identifiability result*). We then show that under the IS+UEM model, an optimal solution to the Maximum Quartet Support Supertree (MQSS) problem is a refinement of the model cell lineage tree (this is a *consistency result* when tree error is defined as the number of false negative branches). Lastly, we consider the case of triplets, showing that there can be anomalous triplets when the probability of false positive errors is greater than zero. Our result suggests that quartets may be more robust to error than triplets when reconstructing cell lineage trees.

Moreover, our results generalize from the IS model to any model $\mathcal{M}$ of 2-state character evolution for which there are no anomalous quartets or triplets. An example of such a model is the infinite sites plus neutral Wright-Fisher (IS+nWF) model (Fisher 1923; Wright 1931) and its approximations (Hudson 2002). The IS+nWF model is a forward-time version of the MSC model. Mutations follow the IS assumption but evolve within a species tree, so deviations from a perfect phylogeny are due to genetic drift. Nevertheless, there are no anomalous triplets (see Supplementary Materials of Kuritzin et al. 2016) and or quartets (Theorem 1 in Molloy et al. 2021; also see Mendes and Hahn 2017) under the IS+nWF model. This led Springer and Gatesy (2016) use quartet-based methods to estimate species trees from retroelement insertion presence/absence patterns, and Molloy et al. (2021) showed this approach is consistent under the IS+nWF model. Since then,

quartet-based methods have been applied to retroelement data sets for placental mammals (Doronina et al. 2022), bats (Korstian et al. 2022), and birds (Gatesy and Springer 2022).

Missing values are prevalent in these data sets; however, little was known about the performance of quartet-based methods when the input data are imperfect. Here, we show that quartet-based methods are still consistent when there are unbiased errors and missingness (Therorem 4 and Corollary 6), addressing an open question from Molloy et al. (2021) and giving a positive result for these recent systematic studies. On the other hand, unbiased error is appropriate for modeling sequencing error, but error could be biased towards particular species or genes when calling retroelement insertions. Future work should investigate this issue further, looking at error and missingness biased towards particular species or (orthologous) positions of the genome. Lastly, although species trees are typically assumed to be binary, there could be hard polytomies, in which case the model tree would be binary model. Our results for consistency with error defined as the number of false negatives (Theorem 4 and Corollary 6) extend to the MSC model, suggesting the utility of quartet-based methods in this setting.

Coming back to tumor phylogenetics, our results (Theorem 4 and Corollary 6) suggest the potential of quartet-based methods for reconstructing highly unresolved cell lineage trees, provided that the tree can be rooted and that false positive branches in the output tree can be effectively handled. The former is doable because the tree can be rooted on the edge incident to the healthy cell with no mutations. The latter is related to mapping mutations onto branches in the cell lineage tree (see Kozlov et al. 2022) as well as identifying which cells are members of the same clone or subclone (Zafar et al. 2019). These tasks may need to be performed when using likelihood-based methods designed for cell lineage tree reconstruction, as such methods also return binary trees. Examples include ScisTree (Wu 2019), SiFit (Zafar et al. 2017), and CellPhy (Kozlov et al. 2022) (note that only ScisTree makes the IS assumption; the other two use finite sites).

In general, likelihood-based methods require explicitly estimating numeric parameters, like $\alpha$ and $\beta$, 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 (without explicit estimation of $\alpha$ and $\beta$) 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 have consequences on runtime and accuracy (just consider that our consistency guarantees assume an infinite number of characters). Theorem 6 sheds light on a potential issue when using ASTRAL, namely that the construction of the constrained solution space $\Sigma$ may not be very successful for mutation data, especially if the number of mutations is small. However, there are other high quality heuristics for MQSS, including wQMC (Snir and Rao 2010, 2012; Avni et al. 2014) and wQFM (Reaz et al. 2014; Mahbub et al. 2021). wQFM, in particular, was recently shown to outperform ASTRAL in simulations (Mahbub et al. 2021). Moreover, even when the number of characters

is small compared to the number of cells, the underlying model tree is likely to be highly unresolved. In this case, sampling different cells around the same branch may be a means of providing more data for estimation (this observation has already been leveraged by Kizilkale et al. 2022).

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). If doublets can be effectively corrected, then their impact on quartet-based methods would be minimal. Alternatively, quartets may be useful for detecting doublets.

CNAs include duplications and losses of large sections of chromosomes (see review on methods for detecting CNAs by Mallory et al. 2020). CNA losses, in particular, have motivated the development of many new 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 (although these false negatives will be biased towards particular cells and mutations). In contrast, SCARLET (Satas et al. 2020) reconstructs a CNA tree and then uses it to constrain phylogeny reconstruction with the mutation data. Constraints have also been leveraged in species phylogenetics, including with ASTRAL (Rabiee and Mirarab 2020). Thus, the output of quartet-based methods could similarly be forced to obey the constraints of a CNA tree. To summarize, it seems possible that quartet-based methods can 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. The authors also thank Michael Nute for discussions.

# 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.

Avni E, Cohen R, Snir S. 2014. Weighted quartets phylogenetics. *Systematic Biology* **64**: 233–242.

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.

Doronina L, Hughes GM, Moreno-Santillan D, Lawless C, Lonergan T, Ryan L, Jebb D, Kirilenko BM, Korstian JM, Dávalos LM, Vernes SC, Myers EW, Teeling EC, Hiller M, Jermiin LS, Schmitz J, Springer MS, Ray DA. 2022. Contradictory phylogenetic signals in the laurasiatheria anomaly zone. *Genes* **13**.

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

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

Gatesy J, Springer MS. 2022. Phylogenomic coalescent analyses of avian retroelements infer zero-length branches at the base of neoaves, emergent support for controversial clades, and ancient introgressive hybridization in afroaves. *Genes* **13**.

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.

Lim B, Lin Y, Navin N. 2020. Advancing cancer research and medicine with single-cell genomics. *Cancer Cell* **37**: 456–470.

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.

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.

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.

Reaz R, Bayzid MS, Rahman MS. 2014. Accurate phylogenetic tree reconstruction from quartets: A heuristic approach. *PLOS ONE* **9**: 1–13.

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.

Snir S, Rao S. 2010. Quartets MaxCut: A divide and conquer quartets algorithm. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **7**: 704–718.

Snir S, Rao S. 2012. Quartet MaxCut: A fast algorithm for amalgamating quartet trees. *Mol Phylogenetics and Evolution* **62**: 1–8.

Springer MS, Gatesy J. 2016. The gene tree delusion. *Molecular Phylogenetics and Evolution* **94**: 1–33.

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, Navin N, Chen K, Nakhleh L. 2019. SiCloneFit: Bayesian inference of population structure, genotype, and phylogeny of tumor clones from single-cell genome sequencing data. *Genome Res.* **29**: 1847–1859.

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, Mirarab S. 2022. Weighting by gene tree uncertainty improves accuracy of quartet-based species trees. *Molecular Biology and Evolution* **39**: msac215.

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.