Journal of Mathematical Biology (2022) 84:36
https://doi.org/10.1007/s00285-022-01731-5

Mathematical Biology

A stochastic Farris transform for genetic data under the multispecies coalescent with applications to data requirements

Gautam Dasarathy[1] · **Elchanan Mossel**[2] · **Robert Nowak**[3] · **Sebastien Roch**[4]

Received: 4 June 2020 / Revised: 15 February 2022 / Accepted: 17 February 2022 /
Published online: 8 April 2022

Abstract

Species tree estimation faces many significant hurdles. Chief among them is that the trees describing the ancestral lineages of each individual gene—the gene trees—often differ from the species tree. The multispecies coalescent is commonly used to model this gene tree discordance, at least when it is believed to arise from incomplete lineage sorting, a population-genetic effect. Another significant challenge in this area is that molecular sequences associated to each gene typically provide limited information about the gene trees themselves. While the modeling of sequence evolution by single-site substitutions is well-studied, few species tree reconstruction methods with theoretical guarantees actually address this latter issue. Instead, a standard—but unsatisfactory—assumption is that gene trees are perfectly reconstructed before being

G. Dasarathy: Supported by the NSF grant CCF-2048223 (CAREER) and the NIH grant 1R01GM140468-01. E. Mossel: Supported by Simons-NSF Grant DMS-2031883 and by a Vannevar Bush Faculty Fellowship ONR-N00014-20-1-2826. S. Roch: Supported by NSF Grants DMS-1149312 (CAREER), DMS-1614242, CCF-1740707 (TRIPODS), DMS-1902892 and DMS-2023239 (TRIPODS Phase II), as well as a Simons Fellowship and a Vilas Associates Award.

✉ Sebastien Roch
roch@math.wisc.edu

Gautam Dasarathy
gautamd@asu.edu

Elchanan Mossel
elmos@mit.edu

Robert Nowak
rdnowak@wisc.edu

[1] School of Electrical, Computer, and Energy Engineering, Arizona State University, Tempe, USA

[2] Department of Mathematics and IDSS, Massachusetts Institute of Technology, Cambridge, USA

[3] Department of Electrical and Computer Engineering, University of Wisconsin, Madison, USA

[4] Department of Mathematics, University of Wisconsin, Madison, USA

fed into a so-called summary method. Hence much remains to be done in the development of inference methodologies that rigorously account for gene tree estimation error—or completely avoid gene tree estimation in the first place. In previous work, a data requirement trade-off was derived between the number of loci m needed for an accurate reconstruction and the length of the locus sequences k. It was shown that to reconstruct an internal branch of length f, one needs m to be of the order of $1/[f^2\sqrt{k}]$. That previous result was obtained under the restrictive assumption that mutation rates as well as population sizes are constant across the species phylogeny. Here we further generalize this result beyond this assumption. Our main contribution is a novel reduction to the molecular clock case under the multispecies coalescent, which we refer to as a stochastic Farris transform. As a corollary, we also obtain a new identifiability result of independent interest: for any species tree with $n \geq 3$ species, the rooted topology of the species tree can be identified from the distribution of its unrooted *weighted* gene trees even in the absence of a molecular clock.

Keywords Phylogenetic reconstruction · Coalescent · Gene tree/species tree · Distance methods · Data requirement

Mathematics Subject Classification 60K35 · 92D15

1 Introduction

Modern sequencing technologies have provided a wealth of data to assist biologists in the inference of evolutionary relationships between species. It is now common to sequence thousands of genes, or entire genomes, simultaneously across a range of species. With this abundance of data comes new algorithmic and statistical challenges. One such challenge arises because phylogenomic inference entails dealing with the interplay of *two* processes. While a species phylogeny depicts the history of speciation of extant organisms, each gene within the genomes of these organisms *has its own history*. That history is captured by a gene tree. In practice, by contrasting the molecular sequences of a gene (or other genomic region) across species, one can reconstruct the corresponding gene tree. Indeed the accumulation of mutations along the gene tree reflects, if imperfectly, the underlying history. Much is known about the reconstruction of single-gene trees, a subject with a long history. See Steel (2016), Warnow (2017) for an overview.

But a gene tree does not necessarily coincide with the species phylogeny. In particular, various mechanisms lead to *discordance* between gene trees. These include the transfer of genetic material between species, hybrid speciation events and a population-genetic effect known as incomplete lineage sorting (Maddison 1997). The wide availability of genomic datasets has brought to the fore the major impact these discordances have on phylogenomic inference (Degnan and Rosenberg 2009). As a result, in addition to the stochastic process governing the evolution of molecular sequences on a fixed gene tree, one is led to model the structure of the gene tree *itself*, in relation to the species phylogeny, through a separate stochastic process. The inference of these complex, two-level evolutionary models is an active area of research.

We focus on incomplete lineage sorting (ILS) and consider the problem of reconstructing a species phylogeny from multiple genes under the multispecies coalescent (Rannala and Yang 2003), a standard population-genetic model. The problem is of great practical interest in computational evolutionary biology and is currently the subject of intense study; see e.g. Nakhleh (2013), Kapli et al. (2020), Scornavacca et al. (2020). There is in particular a growing body of theoretical results in this area (Degnan and Rosenberg 2006; Degnan et al. 2009; DeGiorgio and Degnan 2010; Mossel and Roch 2010; Liu et al. 2010; Allman et al. 2011; Roch 2013; Roch and Steel 2015; DeGiorgio and Degnan 2014; Dasarathy et al. 2015; Chifman and Kubatko 2015; Roch and Warnow 2015; Mossel and Roch 2017; Shekhar et al. 2017; Long and Kubatko 2017; Roch 2018; Roch et al. 2019; Allman et al. 2018, 2019; Long and Kubatko 2019; Rhodes 2019). A significant portion of prior rigorous work on species phylogeny estimation in the presence of ILS has been aimed at the case where true gene trees are assumed to be available. In reality, one needs to estimate gene trees from molecular sequences, leading to reconstruction errors, and indeed there has been a recent thrust towards understanding the effect of this important source of error in phylogenomic inference, both from empirical (Kubatko and Degnan 2007; Bayzid and Warnow 2013; Mirarab et al. 2014, 2016) and theoretical (DeGiorgio and Degnan 2014; Roch and Steel 2015; Bayzid et al. 2015; Roch and Warnow 2015) standpoints. Another option—which we further explore here—is to bypass the reconstruction of gene trees altogether and infer the species history directly from sequence data (Dasarathy et al. 2015; Mossel and Roch 2017; Chifman and Kubatko 2015; Rusinko and McPartlon February 2017; Allman et al. 2019; Long and Kubatko 2019).

In previous work (Mossel and Roch 2017), a surprising trade-off was derived between the number of genes m needed to accurately reconstruct a species phylogeny and the length of the genes k. Specifically, it was shown that m needs to scale like $1/[f^2\sqrt{k}]$, where f is the length of the shortest branch in the tree (measured in expected number of substitutions per unit of time per site; see Sect. 2 for formal definitions). This result was obtained under a restrictive molecular clock assumption where the leaves are "equidistant" from the root; i.e., it was assumed that the mutation rates and population sizes do not vary across species.

In this work, we make progress towards designing species tree estimation methods that provably achieve the theoretical limit by relaxing the above assumption. Our key contribution is of independent interest. We show how to transform sequence data to appear as though it was generated under the multispecies coalescent with a molecular clock. We achieve this through a novel reduction which we call a *stochastic Farris transform*. Our construction relies on an identifiability result which is partly new: for any species phylogeny with $n \geq 3$ species, the rooted topology of the species tree can be identified from the distribution of the unrooted *weighted* gene trees even in the absence of a molecular clock. We state our results in Sect. 2 and describe our reduction in Sect. 3. The main proofs are in Sects. 4 and 5.

Related work A common approach to species tree estimation that bypasses gene trees is to (1) concatenate the aligned gene sequences and (2) apply a standard phylogenetic reconstruction method (under the incorrect assumption that all sites have evolved on a fixed tree), such as maximum likelihood or a distance-based method, to the concatenated data. That approach has been shown to have serious theoretical drawbacks. Indeed, in Roch and Steel (2015), it was proved that, under the multispecies coalescent with a standard site substitution model (see Sect. 2), maximum likelihood on a concatenated alignment is statistically inconsistent, that is, it can converge on the wrong phylogeny even as the amount of data grows to infinity. The result in Roch and Steel (2015) allows for gene sequence lengths to vary with the number of genes. In follow-up work (Roch et al. 2019), it was shown that fully partitioned concatenated maximum likelihood (that is, maximizing the likelihood of the sequence data under the assumption that the tree topology is fixed across genes but allowing for branch lengths to vary) is also statistically inconsistent. That result was established under bounded gene sequence lengths.

On the other hand, positive results have also been obtained for some concatenation-based approaches under the multispecies coalescent. In Dasarathy et al. (2015), a notion of distance between taxa defined over a concatenated alignment was shown to satisfy the four-point condition (see, e.g., Semple and Steel (2003)) in the limit of infinitely many genes, for any fixed gene sequence length. The latter ensures that the unrooted topology of the species phylogeny is identifiable from the sequence data, that is, that species phylogenies with distinct unrooted topologies necessarily produce distinct data distributions. The results in Dasarathy et al. (2015) allow for varying population sizes and mutation rates across branches and were proved under the Jukes-Cantor substitution model. They also come with a data requirement guarantee: using an appropriate distance-based method, for any gene sequence length $k \geq 1$, the correct unrooted species tree topology is guaranteed to be recovered with high probability provided the number of genes m scales roughly like $\propto e^{C\Delta} \log n$ where Δ is the depth of the tree (see, e.g., Erdos et al. 1999), n is the number of leaves and C is a universal constant.

The distance-based pipeline introduced in Dasarathy et al. (2015), specifically in combination with the reconstruction method Neighbor Joining (Saitou and Nei 1987) (but under a more general model of site substitution), was tested on simulated datasets in Rusinko and McPartlon (February 2017). Moreover, the theoretical identifiability results of Dasarathy et al. (2015) were extended to a significantly broader class of site substitution models in Allman et al. (2019) using the concept of log-det distance (see, e.g., Semple and Steel 2003). The model considered in Allman et al. (2019) allows for an arbitrary mixture of general time-reversible rate matrices across the genome and population sizes that vary on each branch of the species tree as a function of time.

There has also been closely related work on single nucleotide polymorphism (SNP)-based approaches (that is, the case $k = 1$). In particular, it was shown in Chifman and Kubatko (2015) that the unrooted topology of the species phylogeny is identifiable given observed data at the leaves of the tree that are assumed to have arisen from the multispecies coalescent under a time-reversible substitution process with site-specific rate variation modeled by the discrete gamma distribution and a proportion of invariable sites. The results of Chifman and Kubatko (2015), which led to a practical SNP-based approach (Chifman and Kubatko 2014), were extended in Long and Kubatko (2019) to a modified model that allows for varying population sizes and mutation rates.

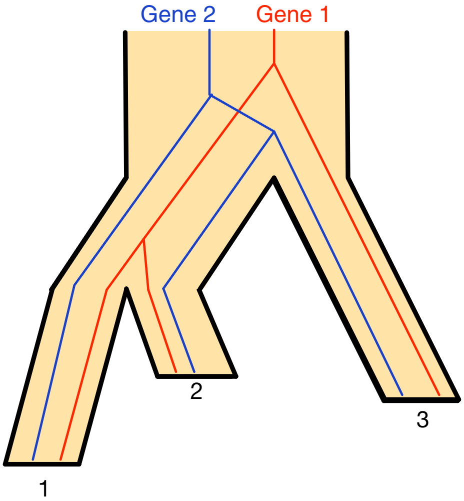

Fig. 1 Two samples from the MSC on a species phylogeny with 3 leaves. The topology of Gene 1 agrees with the topology of the underlying species phylogeny (i.e., species 1 and 2 are closest), while the topology of Gene 2 does not (here species 2 and 3 are closest instead)

Finally, the data requirement bounds in Dasarathy et al. (2015) were substantially improved in Mossel and Roch (2017) using a different reconstruction approach, but under more restrictive assumptions (see Sect. 2). One of our main contributions here is to relax these assumptions while preserving roughly the same data requirements.

2 Background and results

We begin with a description of our modeling assumptions. More details on these standard models can be found for example in Steel (2016).

Species phylogeny v. gene trees A species phylogeny is a graphical depiction of the evolutionary history of a set of species. The leaves of the tree correspond to extant species while internal vertices indicate a speciation event. Each edge (or branch) corresponds to an ancestral population and will be described here by two numbers: one that indicates the amount of time that the corresponding population lived, and a second one that specifies the rate of genetic mutation in that population. Formally, we define the species phylogeny as follows. Throughout, we use the notation $[n] = \{1, \ldots, n\}$.

Definition 1 *(Species phylogeny).* A species phylogeny $S = (V_s, E_s; r, \vec{\tau}, \vec{\mu})$ is a directed tree rooted at $r \in V_s$ with vertex set V_s, edge set E_s, and n labelled leaves $L = [n]$ such that (a) the degree of all internal vertices is 3 except for the root r which has degree 2, and (b) each edge $e \in E_s$ is associated with a length $\tau_e \in (0, \infty)$ and a mutation rate $\mu_e \in (0, \infty)$.

Our model involves several natural time scales, each of which can be used as a time unit. See, e.g., Allman et al. (2019) for a discussion. Here we choose to measure the length τ_e of a branch $e \in E_s$ in coalescent time units, which is the duration of the branch t_e in number of generations divided by twice its population size N_e. That is, $\tau_e = t_e/2N_e$. The mutation rates μ_e are expressed in expected number of substitutions per site per unit of time, where time is measured in coalescent units again. We pictorially represent species phylogenies as thick shaded trees; see Fig. 1. The goal in our applications will be to reconstruct the *rooted topology* of the species phylogeny $S = (V_s, E_s; r, \vec{\tau}, \vec{\mu})$, that is, the rooted tree (V_s, E_s, r).

While a species phylogeny describes the history of speciation, each gene has its own history captured by a gene tree.

Definition 2 *(Gene trees).* Gene tree $G^{(i)} = (V^{(i)}, E^{(i)}; R, \vec{\delta}^{(i)})$ of gene i is a directed tree rooted at R with vertex set $V^{(i)}$ and edge set $E^{(i)}$, and the same labeled leaf set $L = [n]$ as S such that (a) the degree of each internal vertex is 3, except the root R whose degree is 2, and (b) each branch $e \in E^{(i)}$ is associated with a branch length $\delta_e^{(i)} \in (0, \infty)$.

The branch lengths $\delta_e^{(i)}$ are expressed in expected number of substitutions per site. The precise relationship between the branch lengths of the species phylogeny and those of the gene trees under our modeling assumptions will be given below in (2).

Multispecies coalescent and Jukes-Cantor models While the species phylogeny is assumed to be fixed (but unknown), the gene trees are random. Their distribution depends on the species phylogeny. Specifically, we assume that a multispecies coalescent (MSC) process produces m independent gene trees $G^{(1)}, G^{(2)}, \ldots, G^{(m)}$. This process is parametrized by the species phylogeny S. In words, proceeding backwards in time, in each population, every pair of lineages entering from descendant populations merges at exponential rate one in coalescent time units. One key feature of the gene trees is the following: their topology may be *distinct from that of the species phylogeny*. This discordance, which in this context is referred to as incomplete lineage sorting, is a major challenge for species tree estimation from multiple genes. See again Figure 1 for an illustration.

Gene trees are not observed directly. They are inferred from sequence data evolving on the gene trees. We model this sequence generation process according to the standard Jukes-Cantor (JC) model. Given a gene tree $G^{(i)} = (V^{(i)}, E^{(i)}; r, \vec{\delta}^{(i)})$, we associate to each $e \in E^{(i)}$, a probability $p_e^{(i)} = \frac{3}{4}\left(1 - e^{-\frac{4}{3}\delta_e^{(i)}}\right)$. In words, the corresponding gene i is a sequence of length k in $\{\mathtt{A}, \mathtt{T}, \mathtt{G}, \mathtt{C}\}^k$. Each position in the sequence evolves independently, starting from a uniform state in $\{\mathtt{A}, \mathtt{T}, \mathtt{G}, \mathtt{C}\}$ at the root. Moving away from the root, a substitution occurs on edge e with probability $p_e^{(i)}$, in which case the state changes uniformly at random. Repeating this process for each position produces a sequence of length k for all leaves of $G^{(i)}$, for each $i \in [m]$—our input.

Tree metrics It remains to describe the relationship between the branch lengths of the species phylogeny and those of the gene trees. For this purpose, we recall some notions on tree metrics.

Definition 3 (Species metric). A species phylogeny $S = (V_s, E_s; r, \vec{\tau}, \vec{\mu})$ induces the following metric on the leaf set L. For any pair of leaves $a, b \in L$, we let

$$\mu_{ab} = \sum_{e \in \pi(a,b;S)} \tau_e \, \mu_e, \tag{1}$$

where $\pi(a, b; S)$ is the unique path connecting a and b in S interpreted as a set of edges. We will refer to $\{\mu_{ab}\}_{a,b \in L}$ as the **species metric** induced by S.

The species metric μ can be naturally extended to the entire set V_s by using (1) for an arbitrary pair of vertices. The metric $\{\mu_{ab}\}_{a,b \in L}$ is said to be ultrametric when $\mu_{ra} = \mu_{rb}$ for all $a, b \in L$, that is, when the distance from the root to every leaf is the same. We do *not* make this assumption here. In particular, we allow mutation rates and population sizes to vary across branches of the species phylogeny. Instead, the key to our contribution is a transformation of the sequence data that mimics an ultrametric case. Details on this transformation are given below in Algorithm 1 and Defintion 5.

Recall from Definition 2 that each gene tree has an associated set of branch lengths. Under the MSC, a single branch of a gene tree may span across multiple branches of the species phylogeny; this can also be seen in Fig. 1. Let $t_{\tilde{e}}$ denote the length of the branch $\tilde{e} \in E^{(i)}$ in coalescent time units. For any species phylogeny branch $e \in E_s$, let $t_{\tilde{e} \cap e}$ denote the length of the branch $\tilde{e}$ that overlaps with e. Then, $\delta_{\tilde{e}}$ and $t_{\tilde{e}}$ satisfy the following relationship

$$\delta_{\tilde{e}} = \sum_{e \in E_s} \mu_e t_{\tilde{e} \cap e}. \tag{2}$$

This set of weights defines a different metric on the leaves. Note that, under the MSC, these weights are random. They also lead to a metric which will play a central role in our approach.

Definition 4 *(Gene metric).* A gene tree $G^{(i)} = (V^{(i)}, E^{(i)}; R^{(i)}, \vec{\delta}^{(i)})$ induces the following metric on the leaf set L. For any pair of leaves $a, b \in L$, we let $\delta^{(i)}_{ab} = \sum_{e \in \pi(a,b;G^{(i)})} \delta^{(i)}_e$ where, again, $\pi(a, b; G^{(i)})$ is the unique path connecting a and b in $G^{(i)}$ interpreted as a set of edges. We will refer to $\{\delta^{(i)}_{ab}\}_{a,b \in L}$ as the **gene metric** induced by $G^{(i)}$.

The gene metric $\delta^{(i)}$ can be extended to the entire set V_s and we say that it is ultrametric if $\delta^{(i)}_{ra} = \delta^{(i)}_{rb}$ for all $a, b \in L$. Throughout, the μ_{ab}'s are deterministic (but unknown) while the δ_{ab}'s are random.

Inference problem For gene i, we let $\left\{\xi^{ij}_x \,:\, j \in [k], x \in L\right\}$ denote the data generated at the leaves L of the tree $G^{(i)}$ by the Jukes-Cantor process, the superscript j runs across the positions of the gene sequence. To simplify the notation, we denote $\xi^{ij} = (\xi^{ij}_x)_{x \in L}$. The species phylogeny estimation problem can then be stated as:

> *The $n \times m \times k$ data array $\left\{\xi^{ij}\right\}_{i \in [m], j \in [k]}$ is generated according to the Jukes-Cantor process on the m gene trees, each of which is generated by the multispecies coalescent on S. The goal is to recover the topology of the species phylogeny S from $\left\{\xi^{ij}\right\}_{i \in [m], j \in [k]}$.*

We refer to this two-step process as the MSC-JC(m, k) process on S.

A "stochastic" Farris transform Previous work in Mossel and Roch (2017) on tight data requirement trade-offs under the MSC-JC model was restricted to the case where mutation rates and population sizes do *not* vary across the species phylogeny. This results in the species metric μ_{ab} being in fact ultrametric, as defined above. That, in turn, implies that the gene metrics $\delta^{(i)}$ are also ultrametric. That property produces symmetries that are useful in the design and analysis of reconstruction algorithms. Our main contribution here is a reduction to this ultrametric case.

```
Require: Sequences $\{\xi_x^{ij} : x \in \mathcal{X} = \{1,2,3\}, i \in [m], j \in [k]\}$. Partition of the set of genes $[m] = \mathcal{M}_{\mathrm{R}} \sqcup \mathcal{M}_{\mathrm{Q}}$, with $\mathcal{M}_{\mathrm{R}} = \mathcal{M}_{\mathrm{R1}} \sqcup \mathcal{M}_{\mathrm{R2}}$,
    where $|\mathcal{M}_{\mathrm{R1}}|$ and $|\mathcal{M}_{\mathrm{R2}}|$ satisfy (14) in the proof of Theorem 1.
For each $x, y \in \mathcal{X}$ and $i \in \mathcal{M}_{\mathrm{R}}$, define $\widehat{p}^{i}_{xy} = \frac{1}{k}\sum_{j=1}^{k} \mathbb{1}\{\xi_x^{ij} \neq \xi_y^{ij}\}$, $\widehat{p}^{i\downarrow}_{xy} = \frac{2}{k}\sum_{j=1}^{k/2} \mathbb{1}\{\xi_x^{ij} \neq \xi_y^{ij}\}$, and $\widehat{p}^{i\uparrow}_{xy} = \frac{2}{k}\sum_{j=k/2+1}^{k} \mathbb{1}\{\xi_x^{ij} \neq \xi_y^{ij}\}$.
Let $\{x_i, y_i\}$, $i = 1, 2, 3$, be the three distinct (unordered) pairs of distinct leaves in $\mathcal{X}$.
for i = 1,2 do
    Fixing gene tree topologies
    Let $x = x_i$, $y = y_i$, and $z$ be the unique element in $\mathcal{X} - \{x, y\}$.
    Compute empirical quantiles $\widehat{p}^{(1/3)}_{xy}$, $\widehat{p}^{(2/3)}_{xz}$, $\widehat{p}^{(5/6)}_{xz}$, $\widehat{p}^{(2/3)}_{yz}$, $\widehat{p}^{(5/6)}_{yz}$ from the loci in $\mathcal{M}_{\mathrm{R1}}$. E.g., to compute $\widehat{p}^{(1/3)}_{xy}$, sort the set $\left\{\widehat{p}^{i}_{xy} : i \in \mathcal{M}_{\mathrm{R1}}\right\}$ in ascending order and pick the $\left\lfloor \frac{|\mathcal{M}_{\mathrm{R1}}|}{3} \right\rfloor$th element, breaking ties arbitrarily.
    Set
        $$I := \left\{i \in \mathcal{M}_{\mathrm{R2}} : \widehat{p}^{i\downarrow}_{xy} \le \widehat{p}^{(1/3)}_{xy}, \widehat{p}^{(2/3)}_{xz} \le \widehat{p}^{i\downarrow}_{xz}, \widehat{p}^{(2/3)}_{yz} \le \widehat{p}^{i\downarrow}_{yz}\right\}$$
        $$\cap \left\{i \in \mathcal{M}_{\mathrm{R2}} : \widehat{p}^{i\downarrow}_{xz} \le \widehat{p}^{(5/6)}_{xz} \text{ OR } \widehat{p}^{i\downarrow}_{yz} \le \widehat{p}^{(5/6)}_{yz}\right\}.$$
    Estimation of differences $\Delta_{xy}$
    Set $\widehat{p}^{I}_{xz} := \frac{1}{|I|}\sum_{i \in I} \widehat{p}^{i\uparrow}_{xz}$, and similarly for $\widehat{p}^{I}_{yz}$. Now calculate $\widehat{\Delta}_{xy} := -\widehat{\Delta}_{yx} := -\frac{3}{4}\log\left(\frac{1-\frac{4}{3}\widehat{p}^{I}_{yz}}{1-\frac{4}{3}\widehat{p}^{I}_{xz}}\right)$ (abort if not well-defined)
end for
Let $z_3$ be the unique element in $\mathcal{X} - \{x_3, y_3\}$ and set $\widehat{\Delta}_{x_3y_3} := \widehat{\Delta}_{x_3z_3} - \widehat{\Delta}_{y_3z_3}$ and $\widehat{\Delta}_{x_3x_3}, \widehat{\Delta}_{y_3y_3}, \widehat{\Delta}_{z_3z_3} := 0$.
Stochastic Farris transform
Find a permutation $\{x, y, z\}$ of $\mathcal{X}$ such that $\min\{\widehat{\Delta}_{zx}, \widehat{\Delta}_{zy}\} \ge 0$.
For each $i \in \mathcal{M}_{\mathrm{Q}}$ and $j \in [k]$, set $\xi^{ij}_{z,N} = \xi^{ij}_{z}$. Set $\xi^{ij}_{x,N} = \xi^{ij}_{x}$ with probability $1 - p(\widehat{\Delta}_{zx})$ and otherwise choose $\xi^{ij}_{x,N}$ uniformly from $\{\mathtt{A}, \mathtt{T}, \mathtt{G}, \mathtt{C}\} \setminus \xi^{ij}_{x}$. Same for $\xi^{ij}_{y}$ (with $\widehat{\Delta}_{yz}$ instead of $\widehat{\Delta}_{xz}$) to obtain $\xi^{ij}_{y,N}$.
Return "noisy" sequence data $\left\{\xi^{ij}_{x,N} : i \in \mathcal{M}_{\mathrm{Q}}, j \in [k], x \in \mathcal{X}\right\}$
```

Algorithm 1: Reduction step

That is, we transform the input sequences to appear as though they were generated (approximately) by a species phylogeny with an ultrametric species metric. This ultrametric reduction, inspired by a classical technique known as the Farris transform, may be of independent interest as it could be used to generalize other reconstruction algorithms. Further details and intuition about the Farris transform are given in Sect. 3.1. At a high level, this transform relates a general tree metric to an ultrametric over the same topology. Here we split the data and use one piece to estimate quantities necessary to perform a randomized version of the transform.

Although our reduction could be applied to a dataset of arbitrary size, for ease of presentation we fix a triple of leaves $\mathcal{X} = [3]$. Specifically, Algorithm 1 takes as input a set of genes $[m]$ divided into two disjoint subsets, $\mathcal{M}_{\mathrm{R}}$ and $\mathcal{M}_{\mathrm{Q}}$. The set $\mathcal{M}_{\mathrm{R}}$ is used to estimate parameters needed for the reduction. The reduction is subsequently performed on $\mathcal{M}_{\mathrm{Q}}$. For $\phi > 0$, we say that two metrics μ' and μ'' over $\mathcal{X}$ are ϕ-close if $\left|\mu'_{xy} - \mu''_{xy}\right| \leq \phi$, for all $x, y \in \mathcal{X}$. For convenience, we also say that two species phylogenies are ϕ-close if their species metrics are. In essence, we show that $m \geq 1/(k\phi^2)$ genes suffice to output a dataset that is ϕ-close to ultrametric. Given that $1/\phi^2$ independent sites are required to estimate distances within ϕ Steel and Székely (2002), our bound on the total number of sites mk likely cannot be improved. Throughout, for $a, b \in \mathbb{R}$, we use the notation $a \vee b = \max\{a, b\}$.

Theorem 1 (Ultrametric reduction). *Suppose we have sequence data* $\{\xi^{ij}\}_{i\in[m], j\in[k]}$ *generated under the MSC-JC*(m, k) *process on a three-species phylogeny* $S = (V_s, E_s; r, \vec{\tau}, \vec{\mu})$. *The mutation rates, leaf-edge lengths and internal-edge lengths are respectively in the intervals* (μ_L, μ_U), (f', g') *and* (f, g). *Then, there are constants* $c', c'' > 0$ *such that, for any* $\varepsilon, \phi \in (0, 1)$ *satisfying*

$$k \geq c' \left(\log \phi^{-2} + \log \varepsilon^{-1}\right), \tag{3}$$

with probability at least $1-\varepsilon$, *the output of Algorithm 1 is distributed according to the MSC-JC process on a species phylogeny* S' *that is* ϕ*-close to a species phylogeny with an ultrametric species metric and a rooted topology identical to that of* S, *provided*

$$m \geq c'' \left(1 \vee \frac{1}{k\phi^2}\right) \log \varepsilon^{-1}. \tag{4}$$

Application: Data requirement trade-off Our main application of the ultrametric reduction is an extension of the data requirement trade-off in Mossel and Roch (2017) without the assumption that mutation rates and population sizes do not vary across the species phylogeny. After applying our reduction, we use the quantile triplet test developed in Mossel and Roch (2017). Roughly speaking this test, which is detailed in Algorithm 2 in the appendix, compares a well-chosen quantile of the sequence-based estimates of gene metrics $\left\{\delta^{(i)}_{ab}\right\}_{a,b\in\mathcal{X}}$ in order to determine which pair of leaves is closest.

For any leaves $x, y, z \in L$, the species phylogeny S restricted to these three leaves has one of three possible rooted topologies: $xy|z$, $xz|y$, or $yz|x$. It is a classical phylogenetic result that if one is able to correctly reconstruct the topology of all triples of leaves in L, then the topology of the full species phylogeny can be correctly reconstructed as well (see e.g., Steel 2016). Hence, we restrict to the case $\mathcal{X} = [3]$. Our data requirement applies to an unknown species phylogeny in the following class.

We assume that: mutation rates are in the interval (μ_L, μ_U); leaf-edge lengths are in (f', g'); and internal-edge lengths are in (f, g). We suppress the dependence on μ_L, μ_U, f', g', g, which we think of as constants, and focus here on the role of f, which is known to play a critical role. Specifically, we answer the following question: as $f \to 0$, how many genes m of length k are needed for a correct reconstruction with high probability? We obtain the same trade-off between m and k as in Mossel and Roch (2017), whose proof only applies to the ultrametric case.

Theorem 2 (Data requirement). *Suppose that we have sequence data* $\left\{\xi^{ij}\right\}_{i\in[m],j\in[k]}$ *generated according to the MSC-JC*(m, k) *process on a species phylogeny* $S = (V_s, E_s; r, \vec{\tau}, \vec{\mu})$. *The mutation rates, leaf-edge lengths and internal-edge lengths are respectively in* (μ_L, μ_U), (f', g') *and* (f, g). *We assume further that there are* $C, C' > 0$ *such that* $k = f^{-C}$ *and* $\varepsilon, \phi \in (0, 1)$ *satisfy* (3) *with* $\phi = C' f / \log f^{-1}$. *Then, there exists a universal constant* $c''' > 0$ *such that Algorithm* 2 *correctly identifies the rooted topology of* S *restricted to* $\mathcal{X} = [3]$ *with probability at least* $1 - \varepsilon$ *provided that*

$$m \geq c''' \left(\frac{1}{f} \vee \frac{1}{\sqrt{k} f^2} \right) \log \varepsilon^{-1}. \tag{5}$$

3 Main steps of the proof of Theorem 1

The goal of the ultrametric reduction step, Algorithm 1, is to transform the sequence data to appear statistically as though it is the output of an MSC-JC process on an ultrametric species phylogeny with the same topology as S restricted to $\mathcal{X}$. Here we provide an overview of the key ideas behind this step.

3.1 Preliminary step: an identifiability result

Before diving into the description of Algorithm 1, we provide some insights into the algebra of our reduction by first proving an identifiability result, which is partly new.

It was shown in (Allman et al. 2011, Theorem 9) that the distribution of the unrooted topologies of the gene trees suffices to identify the rooted topology of the species phylogeny when the number of leaves exceeds 4. In fact, even more was shown in that case: the species metric (in coalescent time units, that is, taking $\mu_e = 1$ for all e in our notation) can be recovered from the same information. On the other hand, it was also proved in (Allman et al. 2011, Proposition 3) that, when $n = 4$, the gene tree topologies are not enough to locate the root of the species phylogeny.

We complement these previous results by revisiting the cases $n = 3, 4$ when information about gene tree branch lengths is available. Here we show that, already with three species (and therefore when $n \geq 3$), this extra information allows to recover the rooted topology of the species phylogeny. Branch length information plays a critical role in achieving the data requirement in Theorem 2. We give a constructive proof of Theorem 3 below, leading to Algorithm 1.

Theorem 3 (Identifiability of rooted topology of species phylogeny from gene metrics). *Let $S = (V_s, E_s; r, \vec{\tau}, \vec{\mu})$ be a species phylogeny with $n \geq 3$ leaves and root r and let $G = (V, E; R, \vec{\delta})$ be a gene tree sampled from the MSC with corresponding (random) gene metric $\{\delta_{ab}\}_{a,b \in L}$. Then the rooted topology of the species phylogeny is identifiable from the distribution of the gene metric.*

Our proof is inspired by the Farris transform (also related to the Gromov product; see e.g. Semple and Steel 2003), a classical technique to transform a general tree metric into an ultrametric. In a typical application of the Farris transform, one "roots" the species phylogeny S at an "outgroup" o (i.e., a species that is "far away" from the leaves of S) and then uses the quantities μ_{ox}, $x \in L$ to implicitly stretch the leaf edges appropriately, so as to make all inter-species distances to o equal, without changing the underlying topology. More specifically, let S be a species phylogeny. Suppose $\mathcal{X} = [3]$ and let $o \in L - \mathcal{X}$ be any leaf of S outside $\mathcal{X}$. Assume that $\mu_{o1} \geq \max\{\mu_{o2}, \mu_{o3}\}$ (the other cases being similar) and define the Farris transform

$$\dot{\mu}_{xy} := \mu_{xy} + 2\mu_{o1} - \mu_{ox} - \mu_{oy}, \qquad \forall x, y \in \mathcal{X}. \tag{6}$$

A standard phylogenetic result (proved for instance in Semple and Steel 2003, Lemma 7.2.2) states that $\{\dot{\mu}_{xy}\}_{x,y \in \mathcal{X}}$ is an ultrametric on $\mathcal{X}$ consistent with the topology of S re-rooted at o and, then, restricted to $\mathcal{X}$.

In the multi-gene context, however, we cannot apply a Farris transform in this manner. In particular, we do not have direct access to $\{\mu_{xy}\}$; rather, we only estimate the gene metrics $\{\delta_{xy}^{(i)}\}$. Moreover the latter vary across genes under the MSC.

> Key idea 1: We artificially fix rooted gene tree topologies through conditioning. Doing so allows us to relate species and gene metrics.

We prove Theorem 3 for $n = 3$, which suffices.[1] Let S be a species phylogeny with three leaves and recall that r is the root of S. Let $G = (V, E; R, \vec{\delta})$ be a gene tree sampled from the MSC with corresponding (random) gene metric $\{\delta_{ab}\}_{a,b \in L}$. Unlike the classical Farris transform above, we do not use an outgroup. Instead, we show how to achieve the same outcome by using only the distribution of G and, in particular, of $\{\delta_{ab}\}_{a,b \in L}$. Notice from (6) that we only need the *differences* $\Delta_{xy} := \mu_{rx} - \mu_{ry}$. It is these quantities that we derive from the distribution of the gene metric. The high-level idea is to:

1. Condition on an event such that *the rooted topology of the gene tree is guaranteed to be equal to* $xy|z$. Intuitively, we achieve this by considering an event where one pair of leaves is "somewhat close" while the other two pairs are "somewhat far." We make the latter condition precise in Proposition 1 below.
2. Recover the species-based difference $\Delta_{xy} = \mu_{rx} - \mu_{ry}$ from the gene-based difference $\delta_{xz} - \delta_{yz}$. Indeed, when the rooted topology of G is $xy|z$, then the difference $\delta_{xz} - \delta_{yz}$ is *equal to* Δ_{xy} This is established in Proposition 2 below. See Fig. 2 for an illustration.

[1] Technically, we must allow each edge of S to be a *sequence* of populations with varying population sizes and mutation rates (corresponding to a path within a larger phylogeny). The proofs of Propositions 1 and 2 are unaffected by this extension.

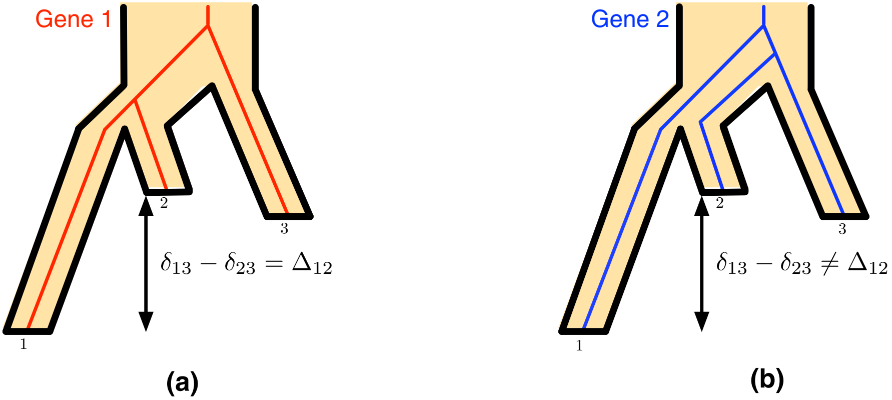
Gene 1
2
3
1
$\delta_{13} - \delta_{23} = \Delta_{12}$
(a)
Gene 2
2
3
1
$\delta_{13} - \delta_{23} \neq \Delta_{12}$
(b)

Fig. 2 **a** Gene 1 (red gene) has the topology 12|3. Therefore, the gene distance on this gene satisfies the condition that $\delta_{13} - \delta_{23} = \Delta_{12}$. **b** In this case, Gene 2 (blue gene) has the topology 1|23. Observe that therefore, $\delta_{13} - \delta_{23} \neq \Delta_{12}$

More formally, we establish the following two propositions, whose proofs are in Sect. 4. For $x, y \in L$ and $\beta \in [0, 1]$, let $\delta_{xy}^{(\beta)}$ be the βth quantile of δ_{xy}. That is, $\delta_{xy}^{(\beta)}$ is the smallest number $a \in [0, 1]$ such that $\mathbb{P}\left[\delta_{xy} \leq a\right] \geq \beta$. Note that this quantile is a function of the distribution of G.

Proposition 1 (Fixing the rooted topology of the gene tree). *Let* (x, y, z) *be an arbitrary permutation of* $(1, 2, 3)$. *The following event has positive probability and implies that the rooted topology of G is* $xy|z$*:*

$$\mathscr{E}_I = \left\{\delta_{xy} \leq \delta_{xy}^{(1/2)}, \delta_{xz} > \delta_{xz}^{(1/2)}, \delta_{yz} > \delta_{yz}^{(1/2)}\right\}. \tag{7}$$

Conditioning on $\mathscr{E}_I$, we then show how to recover the difference Δ_{xy} from $\delta_{xz} - \delta_{yz}$.

Proposition 2 (A formula for the height difference). *We have that, conditioned on* $\mathscr{E}_I$, *almost surely*

$$\delta_{xz} - \delta_{yz} = \Delta_{xy}. \tag{8}$$

The quantity on the l.h.s. of (8) is a function of the distribution of G. From the Δ_{xy} s, we can solve for μ_{rx}'s. Combining the properties of the Farris transform with Propositions 1 and 2 leads to Theorem 3. The Proof of Theorem 3 can also be found in Sect. 4.

3.2 Algorithm 1: the reduction step

We are now ready to describe the reduction algorithm (Algorithm 1). Recall that we are restricting our attention to the case of three leaves $\mathcal{X} = \{1, 2, 3\}$ with rooted species tree topology 12|3. The main idea underlying the reduction algorithm is based on the proof of the identifiability result (Theorem 3). That is, we find a set of genes whose topology is highly likely to be a fixed triplet, we estimate the height differences on this set using (8), and we perform a type of Farris transform. However, given that we do not have access to the actual gene tree distribution but only to sequence data, there are *several differences* with the identifiability proof that make the analysis and the algorithm more involved. We explain these next in details.

A first challenge is that, in the regime where sequence length is "short," i.e., when $k \ll f^{-2}$, the sequence-based estimate of the gene tree metric is much less accurate than what is needed for our reduction step.

> Key idea 2: We show how to combine genes satisfying a condition related to (7) to produce a much more accurate estimate of distance differences.

Fixing gene tree topologies Here we only have access to sequence data. In particular the δs are unknown. So, we work instead with the p-distances $\widehat{p}^i_{xy} = \frac{1}{k}\sum_{j\in[k]} \mathbb{1}\left\{\xi^{ij}_x \neq \xi^{ij}_y\right\}$, for gene i and $x, y \in \mathcal{X}$, and their empirical quantiles $\widehat{p}^{(\beta)}_{xy}$.[2] Similarly to Proposition 1, we then consider those genes for which the event

$$\left\{\widehat{p}^i_{xy} \leq \widehat{p}^{(1/3)}_{xy}, \widehat{p}^{(2/3)}_{xz} \leq \widehat{p}^i_{xz}, \widehat{p}^{(2/3)}_{yz} \leq \widehat{p}^i_{yz}\right\} \cap \left\{\widehat{p}^i_{xz} \leq \widehat{p}^{(5/6)}_{xz} \text{ OR } \widehat{p}^i_{yz} \leq \widehat{p}^{(5/6)}_{yz}\right\}. \tag{9}$$

holds for some chosen permutation (x, y, z) of $(1, 2, 3)$. We will call this set of genes I. We show that this set has a "non-trivial" size and that, with high probability, the genes satisfying (9) have topology $xy|z$ (see Proposition 3).[3] In particular, letting $p(x) = \frac{3}{4}\left(1 - e^{-4x/3}\right)$, the analysis of this construction accounts for the "sequence noise" around the expected values

$$p^i_{xy} := \mathbb{E}\left[\widehat{p}^i_{xy}|G^{(i)}\right] = \frac{3}{4}\left(1 - e^{-4\delta^i_{xy}/3}\right) =: p(\delta^i_{xy}). \tag{10}$$

Estimating distance differences Because we work with p-distances, we adapt formula (8) for the difference Δ_{xy} as follows. Using $\widehat{p}^I_{xz} = \frac{1}{|I|}\sum_{i\in I}\widehat{p}^i_{xz}$ and $\widehat{p}^I_{yz} = \frac{1}{|I|}\sum_{i\in I}\widehat{p}^i_{yz}$, our estimate of the distance differences is given by

$$\widehat{\Delta}_{xy} = \left\{-\frac{3}{4}\log\left(1 - \frac{4}{3}\widehat{p}^I_{xz}\right)\right\} - \left\{-\frac{3}{4}\log\left(1 - \frac{4}{3}\widehat{p}^I_{yz}\right)\right\}.$$

Recall from Proposition 2 and Fig. 2 that, for this formula to be valid, we need to ensure that the topology of the gene trees used is $xy|z$. The logarithmic transforms in the curly brackets are the usual distance corrections in the Jukes-Cantor sequence model (see e.g. Semple and Steel (2003)). Note, however, that we perform an average over I before the correction; this is important to obtain the needed statistical power of

[2] Actually, the quantiles are estimated from part of the gene set ($\mathcal{M}_{\mathrm{R1}}$) to avoid unwanted correlations. The rest of the analysis is done on the other part. See Algorithm 1.

[3] The p-distances in (9) are actually estimated over half the gene length to once again avoid unwanted correlations. That is, we use $\widehat{p}^{i\downarrow}_{xy}$ defined in Algorithm 1 to compute I. We use the other half to estimate the differences below.

our estimator. The non-trivial part of the analysis of this step is to bound the estimation error. Indeed, unlike the identifiability result, we have a finite amount of gene data and, moreover, we must account for the sequence noise. This is done using concentration inequalities in Proposition 4.

In Algorithm 1, we compute $\widehat{\Delta}_{xy}$ with the formula above for two of the pairs in $\mathcal{X}$, say $(1, 2)$ and $(1, 3)$, and then derive the third quantity consistently, i.e., $\widehat{\Delta}_{23} = \widehat{\Delta}_{13} - \widehat{\Delta}_{12}$. We also set $\widehat{\Delta}_{xy} = -\widehat{\Delta}_{yx}$ and $\widehat{\Delta}_{xx} = 0$ for all x, y.

Stochastic Farris transform The quantile test (see Section 1) is not a distance-based method in the traditional sense of the term. Indeed we do not define a pairwise distance matrix on the leaves. Instead, we use the *empirical distribution of the p-distances across genes*. It is for this reason that we do not simply apply the classical Farris transform of (6) to the estimated distances. Rather, we perform what we call a "stochastic" Farris transform to ensure that we properly mimic the contributions from both the multispecies coalescent and the Jukes-Cantor model to the distribution of p-distances.

> Key idea 3: We transform the sequence data itself to mimic the distribution under an ultrametric species phylogeny. This is done by adding the right amount of noise to the sequence data at each gene, as detailed next.

We will let $\oplus$ denote addition mod-4 and identify $\texttt{A}, \texttt{T}, \texttt{G}, \texttt{C}$ with $0, 1, 2, 3$ respectively in that order when doing this addition. For instance, this means that $\texttt{A} \oplus 1 = \texttt{T}$ and $\texttt{G} \oplus 2 = \texttt{A}$.

Definition 5 (Stochastic Farris transform). For a gene i, let $\{\xi_x^i\}_{x\in\mathcal{X}}$ be a sequence dataset over the species $\mathcal{X} = [3]$ and let $\Delta_{xy} = \mu_{rx} - \mu_{ry}$, $x, y \in \mathcal{X}$. Assume without loss of generality that $\min\{\Delta_{12}, \Delta_{13}\} \geq 0$. The **stochastic Farris transform** defines a new set of sequences $\{\xi_{x,N}^i\}_{x\in\mathcal{X}}$ such that $\xi_{x,N}^i = \xi_x^i \oplus \epsilon_x^i$, where $\epsilon_x^i \in \{0, 1, 2, 3\}^k$ is an independent random sequence whose jth coordinate is drawn according to: $\epsilon_x^{ij} = 0$, w.p. $1 - p(\Delta_{1x})$; otherwise it is chosen uniformly among [3].

We write this as $\{\xi_{x,N}^i\}_{x\in\mathcal{X}} = \mathcal{F}(\{\xi_x^i\}_{x\in\mathcal{X}}; \{\Delta_{xy}\}_{x,y\in\mathcal{X}})$.

By the Markov property, for $x, y \in \mathcal{X}$, the "noisy" sequence data above satisfy

$$\mathbb{P}\left[\xi_{x,N}^i \neq \xi_{y,N}^i\right] = p\left(\delta_{xy}^i + \Delta_{1x} + \Delta_{1y}\right) =: r_{xy}^i.$$

Notice that δ_{xy}^i, the random gene tree distance between x and y under gene i, can be decomposed as $\mu_{xy} + \Gamma_{xy}^i$, where Γ_{xy}^i is the random component contributed by the multispecies coalescent. On the other hand, the set of distances $\mu_{xy} + \Delta_{1x} + \Delta_{1y}$ is ultrametric by the properties of the classical Farris transform (see the Proof of Theorem 3). As a result, the stochastic Farris transform modifies the sequence data so that it appears as though it was generated from an ultrametric MSC-JC process.

In reality, we do not have access to the true differences Δ_{xy}, $x, y \in \mathcal{X}$. Instead, we employ our estimates $\widehat{\Delta}_{xy}$ for all $x, y \in \mathcal{X}$ in the previous step to obtain the following approximate stochastic Farris transform:

$$\{\xi_{x,N}^i\}_{x\in\mathcal{X}} = \mathcal{F}(\{\xi_x^i\}_{x\in\mathcal{X}}; \{\widehat{\Delta}_{xy}\}_{x,y\in\mathcal{X}}). \tag{11}$$

This is the output of the reduction. See Algorithm 1 for details. We prove Theorem 1 in Sect. 5.

4 Identifiability: detailed proofs

Proof (Proposition 1) Recall the definition of the event $\mathscr{E}_I$ from (7). Our goal is to show that it has positive probability and that it implies that the rooted topology of G is $xy|z$. This makes sense on an intuitive level because the event $\mathscr{E}_I$ requires that δ_{xy} is "somewhat small" and that δ_{xz}, δ_{yz} are "somewhat large." To prove this, we make crucial use of the following symmetry. Let w be the most recent common ancestor of x, y and z in S. By "above (respectively, below) w," we refer to the times prior to (respectively, following) the species divergence at w (forward in time). Let

$$\Gamma_{xy} = \delta_{xy} - \mu_{xy}, \tag{12}$$

and let $\mathscr{B}_{xy}$ denote the event that the coalescence between the lineages of x and y occurs below w (which is only possible if x and y happen to be sister populations in the species phylogeny). For $\beta \in [0, 1]$, let $\Gamma_{xy}^{(\beta)}$ be the βth quantile of Γ_{xy}. We define the quantities and event above similarly for the other pairs. Under $\mathscr{B}_{xy}^c$, note that Γ_{xy} is the contribution to δ_{xy} coming from the path above w on G, and the same holds for the other pairs. Hence, by the exchangeability of the coalescent process above w, we have

$$\Gamma_{xy} \mid \mathscr{B}_{xy}^c \overset{\mathrm{d}}{=} \Gamma_{xz} \mid \mathscr{B}_{xz}^c \overset{\mathrm{d}}{=} \Gamma_{yz} \mid \mathscr{B}_{yz}^c. \tag{13}$$

That observation will facilitate the comparison of quantiles.

We break up the proof into 3 cases depending on the rooted topology of S:

- **Rooted topology of** S **is** $yz|x$. In that case, the lineages from the pairs (x, y) and (x, z) coalesce *only* above w. That is, the events $\mathscr{B}_{xy}$ and $\mathscr{B}_{xz}$ occur almost surely. By (13), we then get $\Gamma_{xy} \overset{\mathrm{d}}{=} \Gamma_{xz}$ which implies that $\Gamma_{xy}^{(1/2)} = \Gamma_{xz}^{(1/2)} =: \gamma$. Because μ_{xy} and μ_{xz} are deterministic, (12) guarantees further that

  $$\delta_{xy}^{(1/2)} = \mu_{xy} + \gamma, \qquad \delta_{xz}^{(1/2)} = \mu_{xz} + \gamma.$$

  As a result, the event $\mathscr{E}_I$ means that

  $$\mu_{xy} + \Gamma_{xy} = \delta_{xy} \leq \delta_{xy}^{(1/2)} = \mu_{xy} + \gamma,$$

  and

  $$\mu_{xz} + \Gamma_{xz} = \delta_{xz} > \delta_{xz}^{(1/2)} = \mu_{xz} + \gamma.$$

  The last two inequalities are equivalent to $\Gamma_{xy} \leq \gamma < \Gamma_{xz}$ which is only possible if the rooted topology of G is $xy|z$. It remains to show that the event $\mathscr{E}_I$ occurs with positive probability. What we have shown implies also that $\Gamma_{yz} = \Gamma_{xz}$ almost surely. Moreover, conditioned on $\mathscr{B}_{yz}$, the random variable Γ_{yz} is non-positive; hence, $\Gamma_{yz}^{(1/2)} \leq \gamma < \Gamma_{xz} = \Gamma_{yz}$. So $\mathscr{E}_I$ is indeed possible and occurs with positive probability under the MSC.

- **Rooted topology of** S **is** $xz|y$**.** This is case is similar to the previous one, and is therefore omitted.
- **Rooted topology of** S **is** $xy|z$**.** In that case, the lineages from the pairs (x, z) and (y, z) coalesce only above w. On the other hand, conditioned on $\mathscr{B}_{xy}$, the random variable Γ_{xy} is non-positive. Hence, combining these observations with (13), we get $\Gamma_{xy}^{(1/2)} \leq \Gamma_{xz}^{(1/2)} = \Gamma_{yz}^{(1/2)} =: \gamma$. The event $\mathscr{E}_I$ then means that

$$\mu_{xy} + \Gamma_{xy} = \delta_{xy} \leq \delta_{xy}^{(1/2)} = \mu_{xy} + \Gamma_{xy}^{(1/2)} \leq \mu_{xy} + \gamma,$$

and

$$\mu_{xz} + \Gamma_{xz} = \delta_{xz} > \delta_{xz}^{(1/2)} = \mu_{xz} + \gamma,$$

and again $\Gamma_{xy} \leq \gamma < \Gamma_{xz}$, implying that the rooted topology of G is $xy|z$. The equality $\Gamma_{xz} = \Gamma_{yz}$ holds as well almost surely. So $\mathscr{E}_I$ has positive probability.

That proves the claim. □

‘

Proof (Proposition 2) We know from Proposition 1 that, conditioned on $\mathscr{E}_I$, coalescence between the lineages of x and z necessarily occurs in the common ancestral population of x, y and z, irrespective of the species tree topology. The same holds for y and z. Further $\Gamma_{xz} = \Gamma_{yz}$ almost surely, where the Γs were defined in the proof of Proposition 1. Using (12) it follows that, conditioned on $\mathscr{E}_I$, we have $\delta_{xz} - \delta_{yz} = \mu_{xz} - \mu_{yz} = \Delta_{xy}$ with probability 1 as claimed. □

Proof (Theorem 3) By Proposition 2, conditioning on the event $\mathscr{E}_I$—which depends only on the gene metric—we have $\delta_{xz} - \delta_{yz} = \mu_{xz} - \mu_{yz} = \Delta_{xy}$. Hence, from this information, we can construct

$$\dot{\mu}_{xy} = \mu_{xy} + \Delta_{1x} + \Delta_{1y}, \qquad x, y \in \mathcal{X}.$$

assuming that $\min\{\Delta_{12}, \Delta_{13}\} \geq 0$ (the other cases follow similarly). Recalling that $\Delta_{xy} := \mu_{rx} - \mu_{ry}$, where r is the root of S, it follows that $\dot{\mu}_{r1} = \dot{\mu}_{r2} = \dot{\mu}_{r3}$. That is, $\dot{\mu}$ is ultrametric. From this, the rooted topology of the species tree can be reconstructed. See Fig. 2 for an illustration and, for instance (Semple and Steel 2003, Lemma 7.2.2) for more details on this last step. □

5 Ultrametric reduction: detailed proofs

Before proving Theorem 1, we define S' formally. Given estimates $\widehat{\Delta}_{xy}$ for all $x, y \in \mathcal{X}$, and supposing that $\min\{\widehat{\Delta}_{12}, \widehat{\Delta}_{13}\} \geq 0$ or equivalently that the quantity $-\frac{3}{4}\log\left(1 - \frac{4}{3}\widehat{p}_{xy}^{I}\right)$ is minimized for the pair $(x, y) = (2, 3)$ (the other cases follow similarly), we let

$$\widehat{\dot{\mu}}_{xy} = \mu_{xy} + \widehat{\Delta}_{1x} + \widehat{\Delta}_{1y}, \qquad x, y \in \mathcal{X}.$$

Note that this formula is consistent with the definition of $\dot{\mu}_{xy}$ in (6). Observe further that we do *not* have access to $\widehat{\dot{\mu}}_{xy}$, as μ_{xy} is unknown—it is only used to define S' in the statement of the theorem.

Let $S' = (V_s, E_s, r, \vec{\tau}, \hat{\dot{\mu}})$ be a species phylogeny with the same topology and branch lengths as S restricted to $\mathcal{X}$, and with mutation rates $\{\widehat{\dot{\mu}}\}_{e \in E_s}$ that are chosen such that:

(a) If $e \in E_s$ is an internal branch, we let $\widehat{\dot{\mu}}_e = \mu_e$; and
(b) Otherwise, that is, if $e \in E_s$ is incident to a leaf x of S, we let $\widehat{\dot{\mu}}_e$ be chosen to satisfy

$$\widehat{\dot{\mu}}_e \tau_e = \mu_e \tau_e + \widehat{\Delta}_{1x}.$$

The goal here is to "stretch" the leaf edges so that the modified species metric between any pair of leaves $x, y \in \mathcal{X}$ is given by $\widehat{\dot{\mu}}_{xy}$. Alternatively, we could have modified the branch lengths.

5.1 Proof of Theorem 1

Proof (Theorem 1) Define two disjoint subsets $\mathcal{M}_{\mathrm{R}1}, \mathcal{M}_{\mathrm{R}2}$ of $[m]$ satisfying:

$$|\mathcal{M}_{\mathrm{R}1}| = c_2 \log(4\varepsilon^{-1}), \qquad |\mathcal{M}_{\mathrm{R}2}| = \left(1 \vee \frac{1}{k\phi^2}\right) c_2 \log(4\varepsilon^{-1}), \tag{14}$$

for a constant c_2 to be determined below, and let $\mathcal{M}_{\mathrm{R}}$ and $\mathcal{M}_{\mathrm{Q}}$ be such that $\mathcal{M}_{\mathrm{R}} = \mathcal{M}_{\mathrm{R}1} \sqcup \mathcal{M}_{\mathrm{R}2}$ and $[m] = \mathcal{M}_{\mathrm{R}} \sqcup \mathcal{M}_{\mathrm{Q}}$. For a gene i and leaves $x, y \in \mathcal{X}$, recall that $\widehat{p}^{\,i}_{xy} = \frac{1}{k} \sum_{j \in [k]} \mathbb{1}\left\{\xi^{ij}_x \neq \xi^{ij}_y\right\}$. In fact, we split this last average into two to avoid unwanted correlations as we explain below. Assuming k is even for simplicity, we denote these as

$$\widehat{p}^{\,i\downarrow}_{xy} = \frac{2}{k} \sum_{j=1}^{k/2} \mathbb{1}\{\xi^{ij}_x \neq \xi^{ij}_y\} \qquad \text{and} \qquad \widehat{p}^{\,i\uparrow}_{xy} = \frac{2}{k} \sum_{j=k/2+1}^{k} \mathbb{1}\{\xi^{ij}_x \neq \xi^{ij}_y\}.$$

For $\beta \in [0, 1]$, let $\widehat{p}^{(\beta)}_{xy}$ be the corresponding empirical quantiles computed based on the set $\{\widehat{p}^{\,i}_{xy} : i \in \mathcal{M}_{\mathrm{R}1}\}$. Fix a permutation (x, y, z) of $(1, 2, 3)$. Consider the following subset of genes in $\mathcal{M}_{\mathrm{R}2}$:

$$\begin{aligned} I = &\left\{ i \in \mathcal{M}_{\mathrm{R}2} \,:\, \widehat{p}^{\,i\downarrow}_{xy} \leq \widehat{p}^{(1/3)}_{xy},\, \widehat{p}^{(2/3)}_{xz} \leq \widehat{p}^{\,i\downarrow}_{xz},\, \widehat{p}^{(2/3)}_{yz} \leq \widehat{p}^{\,i\downarrow}_{yz} \right\} \\ &\cap \left\{ i \in \mathcal{M}_{\mathrm{R}2} \,:\, \widehat{p}^{\,i\downarrow}_{xz} \leq \widehat{p}^{(5/6)}_{xz} \text{ OR } \widehat{p}^{\,i\downarrow}_{yz} \leq \widehat{p}^{(5/6)}_{yz} \right\}. \end{aligned}$$

The proof of the theorem closely tracks that of the identifiability result (Theorem 3). We first show that the rooted topologies in I are highly likely to be $xy|z$ and prove some technical claims that will be useful in the proof of Proposition 4 (proof in Sect. 5.2).□

Proposition 3 (Fixing gene tree topologies) *There are* $c_9, c_{10}, c'_{10}, \varepsilon_0 > 0$ *such that with probability at least*

$$1 - 10\exp(-2c_9^2|\mathcal{M}_{\mathrm{R1}}|) - 6|\mathcal{M}_{\mathrm{R2}}|\exp\left(-k\varepsilon_0^2\right) - 2\exp\left(-2c_{10}^2|\mathcal{M}_{\mathrm{R2}}|\right),$$

the following hold:

(a) *The rooted topology of all gene trees in* I *is* $xy|z$,
(b) *For all* $i \in I$, $p_{xy}^i \leq p_{xy}^{(7/24)}$, $p_{xz}^{(17/24)} \leq p_{xz}^i \leq p_{xz}^{(19/24)}$, $p_{yz}^{(17/24)} \leq p_{yz}^i \leq p_{yz}^{(19/24)}$,
(c) *The size of* I *is greater than* $c'_{10}|\mathcal{M}_{\mathrm{R2}}|$.

Let $\mathscr{I}$ be the event that the conclusion of Proposition 3 holds. To simplify the notation, we use $\widetilde{\mathbb{P}}$ and $\widetilde{\mathbb{E}}$ to denote the probability and expectation operators *conditioned on* $\mathscr{I}$. Using

$$\widehat{p}_{xz}^I = \frac{1}{|I|}\sum_{i\in I}\widehat{p}_{xz}^{i\uparrow} \qquad \text{and} \qquad \widehat{p}_{yz}^I = \frac{1}{|I|}\sum_{i\in I}\widehat{p}_{yz}^{i\uparrow},$$

let

$$\widehat{\Delta}_{xy} = \left\{-\frac{3}{4}\log\left(1-\frac{4}{3}\widehat{p}_{yz}^I\right)\right\} - \left\{-\frac{3}{4}\log\left(1-\frac{4}{3}\widehat{p}_{xz}^I\right)\right\}.$$

Recall that $\Delta_{xy} = \mu_{rx} - \mu_{ry}$. We next show that $\widehat{\Delta}_{xy}$ is a good approximation to Δ_{xy} (proof in Sect. 5.2).

Proposition 4 (Estimating differences). *There is a* $c_{11} \in (0, 1)$ *such that with* $\widetilde{\mathbb{P}}$-*probability at least* $1 - 4\exp\left(-c_{11}k|\mathcal{M}_{\mathrm{R2}}|\phi^2\right)$, *it holds that* $\left|\widehat{\Delta}_{xy} - \Delta_{xy}\right| \leq \phi/2$.

We repeat the height difference estimation above for all pairs in $\mathcal{X}$. Therefore, by a union bound, we get the above guarantee for all pairs with probability at least $1 - 12\exp\left(-c_{11}k|\mathcal{M}_{\mathrm{R2}}|\phi^2\right)$.

Without loss of generality, assume that $\mu_{r1} \geq \max\{\mu_{r2}, \mu_{r3}\}$, and recall the Farris transform

$$\dot{\mu}_{xy} = \mu_{xy} + 2\mu_{r1} - \mu_{rx} - \mu_{ry} = \mu_{xy} + \Delta_{1x} + \Delta_{1y}, \qquad x, y \in \mathcal{X},$$

which defines an ultrametric, and consider the approximation

$$\widehat{\dot{\mu}}_{xy} = \mu_{xy} + \widehat{\Delta}_{1x} + \widehat{\Delta}_{1y}, \qquad x, y \in \mathcal{X}.$$

Assuming that the conclusion of Proposition 4 holds for all $x, y \in \mathcal{X}$, we have shown that $(\widehat{\mu}_{xy})$ is ϕ-close to the ultrametric $(\dot{\mu}_{xy})$. As we explained in Sect. 3.2, we produce a new sequence dataset using an approximate stochastic Farris transform $\{\xi^i_{x,N}\}_{x\in\mathcal{X}} = \mathcal{F}(\{\xi^i_x\}_{x\in\mathcal{X}}; \{\widehat{\Delta}_{xy}\}_{x,y\in\mathcal{X}})$.

Hence, again by a union bound, we get the claim of Theorem 1 except with probability

$$10\exp(-2c_9^2|\mathcal{M}_{\mathrm{R1}}|) \;\; +6|\mathcal{M}_{\mathrm{R2}}|\exp\left(-2k\varepsilon_0^2\right) + 2\exp\left(-2c_{10}^2|\mathcal{M}_{\mathrm{R2}}|\right) \\ +12\exp\left(-c_{11}k|\mathcal{M}_{\mathrm{R2}}|\phi^2\right).$$

We get the data requirement by asking for the conditions under which the above quantity is less than ε, which involves choosing c_2, c', c'' large enough in (14) and in the statement of the theorem. □

5.2 Proofs of Propositions 3 and 4

Proof (Proposition 3) First fix $0 < \varepsilon_1 < 1/24$ and let $\varepsilon_0 > 0$ be the largest value such that

$$\begin{aligned}
p_{xy}^{(7/24)} &\le p_{xy}^{(1/3-\varepsilon_1)} - 2\varepsilon_0 \le p_{xy}^{(1/3+\varepsilon_1)} + 2\varepsilon_0 \le p_{xy}^{(9/24)}\\
p_{xz}^{(15/24)} &\le p_{xz}^{(2/3-\varepsilon_1)} - 2\varepsilon_0 \le p_{xz}^{(2/3+\varepsilon_1)} + 2\varepsilon_0 \le p_{xz}^{(17/24)}\\
p_{xz}^{(19/24)} &\le p_{xz}^{(5/6-\varepsilon_1)} - 2\varepsilon_0 \le p_{xz}^{(5/6+\varepsilon_1)} + 2\varepsilon_0 \le p_{xz}^{(21/24)}\\
p_{yz}^{(15/24)} &\le p_{yz}^{(2/3-\varepsilon_1)} - 2\varepsilon_0 \le p_{yz}^{(2/3+\varepsilon_1)} + 2\varepsilon_0 \le p_{yz}^{(17/24)}\\
p_{yz}^{(19/24)} &\le p_{yz}^{(5/6-\varepsilon_1)} - 2\varepsilon_0 \le p_{yz}^{(5/6+\varepsilon_1)} + 2\varepsilon_0 \le p_{yz}^{(21/24)}.
\end{aligned} \tag{15}$$

The fact that these inequalities hold is guaranteed by (27) in the Appendix which characterizes the behavior of the quantile functions of the random variables associated with the MSC. Note that ε_0 depends on the parameters μ_U, g' and g.

Let (x, y, z) be an arbitrary permutation of the leaves $(1, 2, 3)$. The idea of the proof is to rely on Proposition 1, which we rephrase in terms of p-distances. For a gene G_i, let

$$p_{xy}^i = \frac{3}{4}\left(1 - e^{-4\delta_{xy}^i/3}\right).$$

And, for $\beta \in [0, 1]$, the corresponding βth quantile is given by

$$p_{xy}^{(\beta)} = \frac{3}{4}\left(1 - e^{-4\delta_{xy}^{(\beta)}/3}\right);$$

similarly for the other pairs. Then, by Proposition 1, the event

$$\mathscr{E}_I^i = \left\{p_{xy}^i \le p_{xy}^{(1/2)}, p_{xz}^{(1/2)} < p_{xz}^i, p_{yz}^{(1/2)} < p_{yz}^i\right\},$$

implies that the rooted topology of G_i is $xy|z$.

Hence, our *main goal* is to show that

$$\mathscr{Q}_i = \left\{ \widehat{p}_{xy}^{i\downarrow} \le \widehat{p}_{xy}^{(1/3)}, \widehat{p}_{xz}^{(2/3)} \le \widehat{p}_{xz}^{i\downarrow}, \widehat{p}_{yz}^{(2/3)} \le \widehat{p}_{yz}^{i\downarrow} \right\} \cap \left\{ \widehat{p}_{xz}^{i\downarrow} \le \widehat{p}_{xz}^{(5/6)} \text{ OR } \widehat{p}_{yz}^{i\downarrow} \le \widehat{p}_{yz}^{(5/6)} \right\}, \tag{16}$$

implies $\mathscr{E}_I^i$ with high probability. We do this by controlling the deviations of $\widehat{p}_{uw}^{(\beta)}$ (Lemma 1 below) and $\widehat{p}_{uw}^{i\downarrow}$ (Lemma 2 below). (The upper bounds on $\widehat{p}_{xz}^{i\downarrow}$ and $\widehat{p}_{yz}^{i\downarrow}$ in (16) are included for technical reasons that will be useful in proving Proposition 4.)

Recall that we use only the genes in $\mathcal{M}_\text{R}$ for the reduction step and this set in turn is divided into disjoint subsets $\mathcal{M}_{\text{R}1}$ and $\mathcal{M}_{\text{R}2}$. The quantiles are estimated using $\mathcal{M}_{\text{R}1}$, while $\mathcal{M}_{\text{R}2}$ is used to compute I. We do *not* argue about the deviation of $\widehat{p}_{uw}^{(\beta)}$ from the *true* βth quantile of the distribution of $\widehat{p}_{uw}^{i}$. Instead we show that $\widehat{p}_{uw}^{(\beta)}$ is close to the βth quantile $p_{uw}^{(\beta)}$ of the disagreement probability *under the MSC*, that is, the quantile *without the sequence noise*. We argue this way because the events that we are ultimately interested in (whether a certain coalescence event has occured in a particular population) are expressed in terms of the MSC. Note that, in order to obtain a useful bound of this type, we must assume that the sequence length is sufficiently long, that is, that the sequence noise is reasonably small. Hence this is one of the steps of our argument where we require a lower bound on k.

Lemma 1 (Deviation of $\widehat{p}_{uw}^{(\beta)}$). *Fix a pair $u, w \in \mathcal{X}$ and a constant $\beta \in (0, 1)$. For all $\varepsilon_0 > 0$ and $0 < \varepsilon_1 < \min\{\beta, 1-\beta\}$, there is a constant $c_9 > 0$ depending on β, ε_0 and ε_1 such that*

$$\mathbb{P}\left[p_{uw}^{(\beta-\varepsilon_1)} - \varepsilon_0 \le \widehat{p}_{uw}^{(\beta)} \le p_{uw}^{(\beta+\varepsilon_1)} + \varepsilon_0 \right] \ge 1 - 2\exp\left(-2c_9^2|\mathcal{M}_{\text{R}1}|\right), \tag{17}$$

provided that k is greater than a constant depending on ε_0 and ε_1.

Proof We prove one side of the first equation (the other inequalities being similar). Define the random variable

$$M = \left|\left\{ i \in \mathcal{M}_{\text{R}1} : \widehat{p}_{uw}^{i} \le p_{uw}^{(\beta+\varepsilon_1)} + \varepsilon_0 \right\}\right|,$$

and observe that

$$\mathbb{P}\left[\widehat{p}_{uw}^{(\beta)} > p_{uw}^{(\beta+\varepsilon_1)} + \varepsilon_0 \right] \le \mathbb{P}\left[M < \beta|\mathcal{M}_{\text{R}1}| \right]. \tag{18}$$

Our goal is hence to bound the probability on the r.h.s.

For this, we first bound the expectation of M by noting that

$$\begin{aligned} &\mathbb{P}\left[\widehat{p}_{uw}^{i} \le p_{uw}^{(\beta+\varepsilon_1)} + \varepsilon_0 \right] \\ &\ge \mathbb{P}\left[\widehat{p}_{uw}^{i} \le p_{uw}^{(\beta+\varepsilon_1)} + \varepsilon_0 \mid p_{uw}^{i} \le p_{uw}^{(\beta+\varepsilon_1)} \right] \mathbb{P}\left[p_{uw}^{i} \le p_{uw}^{(\beta+\varepsilon_1)} \right] \end{aligned}$$

$$\geq \left[1 - \exp\left(-k\varepsilon_0^2\right)\right](\beta + \varepsilon_1), \tag{19}$$

by Hoeffding's inequality Boucheron et al. (2013) applied to $\widehat{p}^i_{uw}$ and the definition of $p^{(\beta+\varepsilon_1)}_{uw}$; we also used that $\mathbb{E}[\widehat{p}^i_{uw} \mid p_{uw}] = p_{uw}$.

We then apply Hoeffding's inequality to M itself. From (19), it follows that

$$\mathbb{E}[M] \geq (\beta + \varepsilon_1)\left[1 - \exp\left(-k\varepsilon_0^2\right)\right]. \tag{20}$$

Therefore, letting

$$c_9 = (\beta + \varepsilon_1)\left[1 - \exp\left(-k\varepsilon_0^2\right)\right] - \beta,$$

we have that

$$\begin{aligned}\mathbb{P}\left[M < \beta|\mathcal{M}_{\mathrm{R1}}|\right] &\leq \mathbb{P}\left[M - \mathbb{E}M < -c_9|\mathcal{M}_{\mathrm{R1}}|\right]\\ &\leq \exp\left(-2c_9^2|\mathcal{M}_{\mathrm{R1}}|\right),\end{aligned}$$

where we used (20) on the first line. Observe moreover that c_9 is strictly positive, provided that k is greater than a constant depending on ε_0 and ε_1. Combining with (18) gives the result. □

We can also control the deviation of $\widehat{p}^{i\downarrow}_{uw}$ around the random variable p^i_{uw}.

Lemma 2 (Deviation of $\widehat{p}^{i\downarrow}_{uw}$). *Fix a pair $u, w \in \mathcal{X}$. For all i and $\varepsilon_0 > 0$, almost surely*

$$\mathbb{P}\left[|\widehat{p}^{i\downarrow}_{uw} - p^i_{uw}| \geq \varepsilon_0 \mid p^i_{uw}\right] \leq 2\exp\left(-k\varepsilon_0^2\right). \tag{21}$$

Proof Conditioned on p^i_{uw}, the random variable $(k/2)\widehat{p}^{i\downarrow}_{uw}$ is $\mathrm{Bin}(k/2, p^i_{uw})$. The result follows again from Hoeffding. □

Let $\mathscr{E}_{\mathrm{qu}}$ be the event that the inequality in Lemma 1, i.e., $p^{(\beta-\varepsilon_1)}_{uw} - \varepsilon_0 \leq \widehat{p}^{(\beta)}_{uw} \leq p^{(\beta+\varepsilon_1)}_{uw} + \varepsilon_0$, holds for $\widehat{p}^{(1/3)}_{xy}$, $\widehat{p}^{(2/3)}_{xz}$, $\widehat{p}^{(5/6)}_{xz}$, $\widehat{p}^{(2/3)}_{yz}$ and $\widehat{p}^{(5/6)}_{yz}$, which occurs with probability at least $1 - 10\exp(-2c_9^2|\mathcal{M}_{\mathrm{R1}}|)$ by a union bound over (17). Let $\mathscr{D}_i$ be the event that the inequality in Lemma 2, i.e., $|\widehat{p}^{i\downarrow}_{uw} - p^i_{uw}| \leq \varepsilon_0$, holds for all pairs (u, w) in $\mathcal{X}$, an event which occurs with probability at least $1 - 6\exp(-k\varepsilon_0^2)$ by a union bound over (21).

Recall that our goal is to show that $\mathscr{Q}_i$ (defined in (16)) implies $\mathscr{E}^i_I$ with high probability. We also condition on $\mathscr{E}_{\mathrm{qu}}$ and $\mathscr{D}_i$, which occur with high probability. Given $\mathscr{E}_{\mathrm{qu}}$, $\mathscr{D}_i$ and $\mathscr{Q}_i$, we have

$$\begin{aligned}p^i_{xy} &\leq \widehat{p}^{i\downarrow}_{xy} + \varepsilon_0\\ &\leq \widehat{p}^{(1/3)}_{xy} + \varepsilon_0\end{aligned}$$

$$
\begin{aligned}
&\le p_{xy}^{(1/3+\varepsilon_1)} + 2\varepsilon_0\\
&\le p_{xy}^{(1/2)},
\end{aligned}
$$

and similarly for the other pairs, where the first inequality follows from $\mathscr{D}_i$, the second inequality follows from $\mathscr{Q}_i$, the third inequality follows from $\mathscr{E}_{\mathrm{qu}}$, and the fourth inequality follows from (15). That is, $\mathscr{E}_I^i$ holds. Finally, the probability that all $i \in I$ satisfy $\mathscr{E}_I^i$ *simultaneously* is at least

$$
\mathbb{P}[\mathscr{E}_I^i, \forall i \in I] \ge 1 - 10\exp(-2c_0^2|\mathcal{M}_{\mathrm{R1}}|) - 6|\mathcal{M}_{\mathrm{R2}}|\exp\left(-k\varepsilon_0^2\right), \tag{22}
$$

where we bounded the probability that all i in I satisfy $\mathscr{D}_i$ with the probability that all i in $\mathcal{M}_{\mathrm{R2}}$ satisfy $\mathscr{D}_i$.

It remains to bound the size of I.

Lemma 3 (Size of I). *There are constants $c_{10}, c'_{10} > 0$ such that*

$$
\mathbb{P}[|I| \ge c'_{10}|\mathcal{M}_{\mathrm{R2}}| \,|\, \mathscr{E}_{\mathrm{qu}}] \ge 1 - 2\exp\left(-2c_{10}^2|\mathcal{M}_{\mathrm{R2}}|\right), \tag{23}
$$

provided k is greater than a constant depending on ε_0.

Proof We show that, under $\mathscr{E}_{\mathrm{qu}}$, the event $\mathscr{Q}_i$ has constant probability and we apply Hoeffding's inequality to $|I|$, which counts the number of i in $\mathcal{M}_{\mathrm{R2}}$ satisfying $\mathscr{Q}_i$.

Observe that, by (15), the events $\mathscr{D}_i$, $\{p_{xy}^i \le p_{xy}^{(7/24)}\}$, and $\mathscr{E}_{\mathrm{qu}}$ (applied in that order) imply

$$
\widehat{p}_{xy}^{i\downarrow} \le p_{xy}^i + \varepsilon_0 \le p_{xy}^{(7/24)} + \varepsilon_0 \le p_{xy}^{(1/3-\varepsilon_1)} - \varepsilon_0 \le \widehat{p}_{xy}^{(1/3)}.
$$

A similar argument shows that, under $\mathscr{E}_{\mathrm{qu}}$, for $i \in \mathcal{M}_{\mathrm{R2}}$ the event $\mathscr{D}_i \cap \mathscr{J}_i$, implies $\mathscr{Q}_i$, where we define

$$
\begin{aligned}
\mathscr{J}_i := \{&p_{xy}^i \le p_{xy}^{(7/24)}, p_{xz}^{(17/24)} \le p_{xz}^i, p_{yz}^{(17/24)} \le p_{yz}^i\}\\
&\cap \{p_{xz}^i \le p_{xz}^{(19/24)} \text{ OR } p_{yz}^i \le p_{yz}^{(19/24)}\}.
\end{aligned}
$$

This leads to the following lower bound on $\mathbb{P}[\mathscr{Q}_i \mid \mathscr{E}_{\mathrm{qu}}]$:

$$
\begin{aligned}
&\ge \mathbb{P}[\mathscr{D}_i \cap \mathscr{J}_i \mid \mathscr{E}_{\mathrm{qu}}]\\
&\ge \mathbb{P}[\mathscr{J}_i \mid \mathscr{E}_{\mathrm{qu}}]\,\mathbb{P}[\mathscr{D}_i \mid \mathscr{J}_i \cap \mathscr{E}_{\mathrm{qu}}]\\
&= \mathbb{P}[\mathscr{J}_i]\,\mathbb{P}[\mathscr{D}_i \mid \mathscr{J}_i \cap \mathscr{E}_{\mathrm{qu}}]\\
&\ge c''_{10}\left[1 - 6\exp\left(-k\varepsilon_0^2\right)\right],
\end{aligned}
$$

for some constant $c''_{10} > 0$, where we used Lemma 2 to bound the conditional probability of $\mathscr{D}_i$ on the last line (recalling that Lemma 2 itself is conditioned on the p_{uw}^i's).

On the third line, we used the fact that $\mathscr{E}_{\mathrm{qu}}$ depends on $\mathcal{M}_{\mathrm{R}1}$, and is therefore independent of p^i_{xy}, p^i_{xz}, p^i_{yz} for $i \in \mathcal{M}_{\mathrm{R}2}$. The existence of the constant c''_{10} follows from an argument similar to that leading up to (27) in the Appendix. The expression on the last line is a strictly positive constant provided k is greater than a constant depending on ε_0.

Finally, applying Hoeffding's inequality to $|I|$, we get the result. □

Combining (22) and (23) concludes the proof. □

Proof (Proposition 4) Fix $x, y \in \mathcal{X}$ and let z be the unique element in $\mathcal{X} - \{x, y\}$. The proof idea is based on Proposition 2. Recall that $\mathscr{I}$ is the event that the conclusion of Proposition 3 holds and that we use $\widetilde{\mathbb{P}}$ and $\widetilde{\mathbb{E}}$ to denote the probability and expectation operators conditioned on $\mathscr{I}$. Let also $\mathscr{G}_I$ be the event that the gene trees in I are $\{G_i\}_{i\in I}$. Similarly to the proof of Proposition 2 we note that, conditioned on $\mathscr{I}$, in all genes in I the coalescences between the lineages of x and z happen in the common ancestral population of x, y and z, irrespective of the species tree topology. The same holds for y and z. That implies that, for $i \in I$, $\delta^i_{xz} = \mu_{rx} + \mu_{rz} + \Gamma^i_{xz}$, and $\delta^i_{yz} = \mu_{ry} + \mu_{rz} + \Gamma^i_{yz}$, where the Γ^is are defined as in the proof of Lemma 1.

Hence

$$\begin{aligned}
\widetilde{\mathbb{E}}\left[\widehat{p}^I_{xz} \,|\, \mathscr{G}_I\right] &= \widetilde{\mathbb{E}}\left[\frac{1}{|I|}\sum_{i\in I}\widehat{p}^i_{xz} \,|\, \mathscr{G}_I\right] \\
&= \frac{1}{|I|}\sum_{i\in I} p^i_{xz} \\
&= \frac{3}{4}\left(1 - \frac{1}{|I|}\sum_{i\in I} e^{-4\delta^i_{xz}/3}\right) \\
&= \frac{3}{4}\left(1 - e^{-4\mu_{rx}/3 - 4\mu_{rz}/3}\left(\frac{1}{|I|}\sum_{i\in I} e^{-4\Gamma^i_{xz}/3}\right)\right),
\end{aligned}$$

and similarly for the pair (y, z). Letting $\ell(x) = -\frac{3}{4}\log\left(1 - \frac{4}{3}x\right)$, we get

$$\begin{aligned}
&\ell\left(\widetilde{\mathbb{E}}\left[\widehat{p}^I_{xz} \,|\, \mathscr{G}_I\right]\right) - \ell\left(\widetilde{\mathbb{E}}\left[\widehat{p}^I_{yz} \,|\, \mathscr{G}_I\right]\right) \\
&= -\frac{3}{4}\log\left(\frac{1 - 4/3\widetilde{\mathbb{E}}\left[\widehat{p}^I_{xz} \,|\, \mathscr{G}_I\right]}{1 - 4/3\widetilde{\mathbb{E}}\left[\widehat{p}^I_{yz} \,|\, \mathscr{G}_I\right]}\right) \\
&= -\frac{3}{4}\log\left(\frac{e^{-4\mu_{rx}/3 - 4\mu_{rz}/3}\left(\frac{1}{|I|}\sum_{i\in I} e^{-4\Gamma^i_{xz}/3}\right)}{e^{-4\mu_{ry}/3 - 4\mu_{rz}/3}\left(\frac{1}{|I|}\sum_{i\in I} e^{-4\Gamma^i_{yz}/3}\right)}\right) \\
&= -\frac{3}{4}\log\left(e^{-4\mu_{rx}/3 + 4\mu_{ry}/3}\right) \\
&= \Delta_{xy},
\end{aligned}$$

where in the third equality we used that, conditioned on $\mathscr{I}$, $\Gamma^i_{xz} = \Gamma^i_{yz}$, for all $i \in I$. Observe that the computation above relies crucially on the conditioning on $\mathscr{G}_I$.

It remains to bound the deviation of $\widehat{\Delta}_{xy} = \ell\left(\widehat{p}^I_{xz}\right) - \ell\left(\widehat{p}^I_{yz}\right)$ around $\Delta_{xy} = \ell\left(\widetilde{\mathbb{E}}\left[\widehat{p}^I_{xz} \,|\, \mathscr{G}_I\right]\right) - \ell\left(\widetilde{\mathbb{E}}\left[\widehat{p}^I_{yz} \,|\, \mathscr{G}_I\right]\right)$, and take expectations with respect to $\mathscr{G}_I$. We do this by controlling the error on $\widehat{p}^I_{xz}$ and $\widehat{p}^I_{yz}$, conditionally on $\mathscr{G}_I$. Indeed, observe that the function ℓ satisfies the following Lipschitz property: for $0 \le x \le y \le M < 3/4$,

$$|\ell(x) - \ell(y)| = \int_x^y \frac{1}{1 - 4t/3}\,\mathrm{d}t \le \frac{|x-y|}{1 - 4M/3}. \tag{24}$$

Hence, to control $|\widehat{\Delta}_{xy} - \Delta_{xy}|$, it suffices to bound $\left|\widehat{p}^I_{uz} - \widetilde{\mathbb{E}}\left[\widehat{p}^I_{uz} \,|\, \mathscr{G}_I\right]\right|$ and $\max\left\{\widehat{p}^I_{uz}, \widetilde{\mathbb{E}}\left[\widehat{p}^I_{uz} \,|\, \mathscr{G}_I\right]\right\}$ for $u = x, y$. □

Lemma 4 (Conditional expectation of $\widehat{p}^I_{uz}$). *Fix $u = x$ or y. There is a constant $c'_{12} \in (0, 3/4)$ small enough such that almost surely*

$$\widetilde{\mathbb{E}}\left[\widehat{p}^I_{uz} \,|\, \mathscr{G}_I\right] \le \frac{3}{4} - c'_{12}.$$

Proof Using $p^i_{uz} = \widetilde{\mathbb{E}}\left[\widehat{p}^i_{uz} \,|\, \mathscr{G}_I\right]$ for $i \in I$, by Proposition 3 (b), we have that

$$\widetilde{\mathbb{E}}\left[\widehat{p}^i_{uz} \,|\, \mathscr{G}_I\right] = p^i_{uz} = \frac{3}{4}\left(1 - e^{-4\delta^i_{uz}/3}\right) \le \frac{3}{4} - c'_{12},$$

for some constant $c'_{12} \in (0, 3/4)$. This constant depends on the parameters μ_U, g' and g. Hence, the result follows from averaging over i. □

Lemma 5 (Conditional deviation of $\widehat{p}^I_{uz}$). *Fix $u = x$ or y. For all $\phi' > 0$, almost surely*

$$\widetilde{\mathbb{P}}\left[\left|\widehat{p}^I_{uz} - \widetilde{\mathbb{E}}\left[\widehat{p}^I_{uz} \,|\, \mathscr{G}_I\right]\right| \ge \phi' \,|\, \mathscr{G}_I\right] \le 2\exp\left(-k|I|(\phi')^2\right).$$

Proof Observe first that, *conditioned on $\mathscr{G}_I$*, the $k\,|I|$ sites that are averaged over in the computation of

$$\widehat{p}^I_{uz} = \frac{2}{k|I|}\sum_{i \in I}\sum_{j=k/2+1}^{k} \mathbb{1}\left\{\xi^{ij}_u \ne \xi^{ij}_z\right\}, \tag{25}$$

are *independent*. Secondly, each random variable in (25) is bounded by 1. Therefore, the result follows from Hoeffding's inequality. □

We set $\phi' = c'_{12}\phi/6$, which is $< c'_{12}$ since $\phi \le 1$. Combining (24) and Lemmas 4 and 5, we get that conditioned on $\mathscr{G}_I$

$$|\widehat{\Delta}_{xy} - \Delta_{xy}| \le \left|\ell\left(\widehat{p}^I_{xz}\right) - \ell\left(\widetilde{\mathbb{E}}\left[\widehat{p}^I_{xz} \,|\, \mathscr{G}_I\right]\right)\right| + \left|\ell\left(\widehat{p}^I_{yz}\right) - \ell\left(\widetilde{\mathbb{E}}\left[\widehat{p}^I_{yz} \,|\, \mathscr{G}_I\right]\right)\right|$$

$$
\begin{aligned}
&\leq \frac{\left|\widehat{p}^{I}_{xz} - \widetilde{\mathbb{E}}\left[\widehat{p}^{I}_{xz} \mid \mathscr{G}_I\right]\right|}{1 - 4/3 \max\left\{\widehat{p}^{I}_{xz}, \widetilde{\mathbb{E}}\left[\widehat{p}^{I}_{xz} \mid \mathscr{G}_I\right]\right\}} + \frac{\left|\widehat{p}^{I}_{yz} - \widetilde{\mathbb{E}}\left[\widehat{p}^{I}_{yz} \mid \mathscr{G}_I\right]\right|}{1 - 4/3 \max\left\{\widehat{p}^{I}_{yz}, \widetilde{\mathbb{E}}\left[\widehat{p}^{I}_{yz} \mid \mathscr{G}_I\right]\right\}} \\
&\leq 2\frac{\phi'}{4/3(c'_{12} - \phi')} \\
&= 2\frac{c'_{12}\phi/6}{4/3(c'_{12} - c'_{12}\phi/6)} \\
&= \frac{1}{4 - (2/3)\phi}\phi \\
&\leq \phi/2,
\end{aligned}
$$

where we used again that $\phi \leq 1$, except with $\widetilde{\mathbb{P}}$-probability

$$
4\exp\left(-k|I|(\phi')^2\right) \leq 4\exp\left(-\frac{1}{8}(c'_{12})^2 c'_{10} k |\mathcal{M}_{\mathrm{R}2}|\phi^2\right) = 4\exp\left(-c_{12} k |\mathcal{M}_{\mathrm{R}2}|\phi^2\right),
$$

by setting $c_{12} = \frac{1}{8}(c'_{12})^2 c'_{10}$. Taking expectations with respect to $\mathscr{G}_I$ gives the result.

6 Concluding remarks

Our main contribution is a novel transformation of sequence data under the MSC-JC that produces a new dataset mimicking a molecular clock. We use this reduction, which is of independent interest, to extend a previous data requirement trade-off for species tree estimation beyond the molecular clock case. Our second contribution is a delicate robustness analysis of the quantile-based triplet test of Mossel and Roch (2017). This represents a step towards the design of practical reconstruction algorithms that avoid gene tree estimation and achieve tight rigorous data requirement guarantees. Further issues must be addressed to achieve this goal. In particular, we have assumed here that the mutation rates and sequence lengths are the same across genes. Relaxing these assumptions is important. Identifiability issues may arise however (Matsen and Steel 2007; Steel 2009).

A Data requirement tradeoff: detailed proofs

In this section we analyze Algorithm 2, which performs a quantile test on the output of Algorithm 1.

Proof (Theorem 2) Let $S_{\mathcal{X}}$ be the species tree S restricted to $\mathcal{X} = [3]$ and let r' denote its root, i.e., the most recent common ancestor of $\mathcal{X}$.

```
Require: Sequence output by Algorithm 1 $\{\xi^{ij}_{x,N} : x \in \mathcal{X} = \{1,2,3\}, i \in \mathcal{M}_{\mathrm{Q}}, j \in [k]\}$.
1: For each $x, y \in \mathcal{X}$ and $i \in \mathcal{M}_{\mathrm{Q}}$, let $\widehat{q}^i_{xy} = \sum_{j=1}^k \mathbb{1}\{\xi^{ij}_{x,N} \neq \xi^{ij}_{y,N}\}$.
2: Set $\alpha \triangleq \max\left\{m^{-1}\log m, k^{-0.5}\sqrt{\log k}\right\}$, and partition $\mathcal{M}_{\mathrm{Q}} = \mathcal{M}_{\mathrm{Q}1} \sqcup \mathcal{M}_{\mathrm{Q}2}$ such that $|\mathcal{M}_{\mathrm{Q}1}|, |\mathcal{M}_{\mathrm{Q}2}|$
   satisfy the conditions in Theorem 2.

   Ultrametric Quantile Test on $\left\{\widehat{q}^i_{xy} : x, y \in \mathcal{X}, i \in \mathcal{M}_{\mathrm{Q}1}\right\}$
3: For each pair of leaves $x, y \in \mathcal{X}$, compute $\widehat{q}^{(c_3\alpha)}_{xy}$, the $c_3\alpha$-th quantile with respect to the data
   $\left\{\widehat{q}^i_{xy} : i \in \mathcal{M}_{\mathrm{Q}1}\right\}$.
     The constant $c_3$ is set as to satisfy Proposition 5. Define $\widehat{q}_* \triangleq \max\{\widehat{q}^{(c_3\alpha)}_{xy} : x, y \in [3]\}$.
4: Next, for $x, y \in \mathcal{X}$, define a similarity measure
     $\widehat{s}_{xy} \triangleq \frac{1}{|\mathcal{M}_{\mathrm{Q}2}|}\left|\left\{i \in \mathcal{M}_{\mathrm{Q}2} : \widehat{q}^i_{xy} \leq \widehat{q}_*\right\}\right|.$
Return Declare that the topology is $xy|z$ if $\widehat{s}_{xy} > \max\left\{\widehat{s}_{xz}, \widehat{s}_{yz}\right\}$.
```

Algorithm 2: Quantile-based triplet test

Define a partition of the set of genes $[m] = \mathcal{M}_{\mathrm{R}1} \sqcup \mathcal{M}_{\mathrm{R}2} \sqcup \mathcal{M}_{\mathrm{Q}1} \sqcup \mathcal{M}_{\mathrm{Q}2}$ such that the following conditions hold:

$$\begin{aligned} &|\mathcal{M}_{\mathrm{R}1}| = c_1 \log \varepsilon^{-1}, \ |\mathcal{M}_{\mathrm{R}2}| = c_1 \left(1 \vee \frac{\log k}{kf^2}\right) \log \varepsilon^{-1}, \\ &|\mathcal{M}_{\mathrm{Q}1}| \geq c_1 \alpha^{-1} \log \varepsilon^{-1}, \ |\mathcal{M}_{\mathrm{Q}2}| \geq c_1 f^{-2}(\alpha + f) \log \varepsilon^{-1}, \end{aligned} \tag{26}$$

for a constant $c_1 > 0$ to be determined later. The reconstruction algorithm on $\mathcal{X}$ has two steps:
Ultrametric reduction: In this step, we invoke Algorithm 1 with sequence data $\{\xi^{ij}_x : x \in \mathcal{X}, i \in [m], j \in [k]\}$. The algorithm outputs new sequences $\{\xi^{ij}_{x,N} : x \in \mathcal{X}, i \in \mathcal{M}_{\mathrm{Q}1} \sqcup \mathcal{M}_{\mathrm{Q}2}, j \in [k]\}$, and Theorem 1 guarantees that these new sequences have the same distribution as a multispecies coalescent process on $S'_{\mathcal{X}}$ with species metric $(\hat{\mu}_{xy})$, where $S'_{\mathcal{X}}$ and S have the same rooted topology, and $(\hat{\mu}_{xy})$ and $(\dot{\mu}_{xy})$ are $\mathcal{O}(f/\sqrt{\log k})$-close with probability at least $1-\varepsilon$.
Quantile test: Now, we invoke Algorithm 2 with the sequence data $\{\xi^{ij}_{x,N} : i \in \mathcal{M}_{\mathrm{Q}1} \sqcup \mathcal{M}_{\mathrm{Q}2}, j \in [k]\}$ output by Step 1. By Propositions 5 and 7 in Sect. 1, it follows that, with probability at least $1-\varepsilon$, Algorithm 2 returns the right topology. □

A.1 Quantile test: robustness analysis

Robustness of quantile test Algorithm 1 produces a new sequence dataset $\left\{\xi^{ij}_{x,N} : x \in \mathcal{X}\right\}$ that appears close to being distributed according to an ultrametric species phylogeny. The next step is to perform a triplet test of Mossel and Roch (2017), as detailed in Algorithm 2. Roughly speaking, this test is based on comparing an appropriately chosen quantile of the gene metrics. In fact, because we do not have direct access to the latter, we use a sequence-based surrogate, the empirical p-distances $\widehat{q}^i_{xy} = \frac{1}{k}\sum_{j=1}^k \mathbb{1}\left\{\xi^{ij}_{x,N} \neq \xi^{ij}_{y,N}\right\}$, for each gene $i \in \mathcal{M}_{\mathrm{Q}}$ in the output of the reduction, whose expectation is a monotone transformation of the corresponding gene metrics. The idea of Algorithm 2 is to use the above p-distances to define a "similarity measure" $\widehat{s}_{xy}$ between each pair of leaves $x, y \in \mathcal{X}$ to reveal the underlying species tree topology on $\mathcal{X}$. It works as follows. The set of genes $\mathcal{M}_{\mathrm{Q}}$ is divided into two disjoint subsets $\mathcal{M}_{\mathrm{Q}1}, \mathcal{M}_{\mathrm{Q}2}$ so that $\left|\mathcal{M}_{\mathrm{Q}1}\right|, \left|\mathcal{M}_{\mathrm{Q}2}\right|$ satisfy the conditions above; this is to avoid unwanted correlations. The set $\mathcal{M}_{\mathrm{Q}1}$ is used to compute the $c_3\alpha$-quantile $\widehat{q}^{(c_3\alpha)}_{xy}$ of $\{\widehat{q}^i_{xy} : i \in \mathcal{M}_{\mathrm{Q}1}\}$, where $c_3 > 0$ is a constant determined in the proofs and $\alpha = \max\left\{\frac{\log m}{m}, \sqrt{\frac{\log k}{k}}\right\}$. Let $\widehat{q}_*$ denote the maximum among $\left\{\widehat{q}^{(c_3\alpha)}_{xy} : x, y \in \mathcal{X}\right\}$. We then use the genes in $\mathcal{M}_{\mathrm{Q}2}$ to define the similarity measure $\widehat{s}_{xy} = \frac{1}{|\mathcal{M}_{\mathrm{Q}2}|}\left|\left\{i \in \mathcal{M}_{\mathrm{Q}2} : \widehat{q}^i_{xy} \leq \widehat{q}_*\right\}\right|$. Whichever pair $x, y \in \mathcal{X}$ produces the largest value of $\widehat{s}_{xy}$ is declared the closest, i.e., the output is $xy|z$ where z is the remaining leaf in $\mathcal{X}$.

As stated in Theorem 1, the output to the ultrametric reduction is almost—but not perfectly—ultrametric. To account for this extra error, we perform a delicate robustness analysis of the quantile-based triplet test. At a high level, the proof follows Mossel and Roch (2017). After (1) controlling the deviation of the quantiles, we establish that (2) the test works in expectation and then (3) finish off with concentration inequalities. All these steps must be updated to account for the error introduced in the reduction step. Step (2) is particularly involved and requires a delicate analysis of the CDF of a mixture of binomials.

For the rest of the proof, we assume that the rooted topology of $S_{\mathcal{X}}$ is 12|3. The other cases are similar.

Control of empirical quantiles In the following proposition, we show that the empirical quantiles are well-behaved, and provide a good estimate of the α-quantile of the underlying MSC random variables. We define the random variables q^i_{xy} and r^i_{xy} associated to a gene tree i:

$$q^i_{xy} = p(\delta^i_{xy} + \widehat{\Delta}_{1x} + \widehat{\Delta}_{1y}), \qquad r^i_{xy} = p(\delta^i_{xy} + \Delta_{1x} + \Delta_{1y}).$$

Also, we need the 0th quantile of these random variables, specifically

$$q^{(0)}_{xy} = p\left(\delta^{(0)}_{xy} + \widehat{\Delta}_{1x} + \widehat{\Delta}_{1y}\right) = p(\mu_{xy} + \widehat{\Delta}_{1x} + \widehat{\Delta}_{1y}), \quad r^{(0)}_{xy} = p(\mu_{xy} + \Delta_{1x} + \Delta_{1y}).$$

We first show that $\widehat{q}_* = \max\{\widehat{q}^{(c_3\alpha)}_{xy} : x, y \in \mathcal{X}\}$ is close to $\max\{r^{(0)}_{xy} : x, y \in \mathcal{X}\}$.

Proposition 5 (Quantile behaviour). *Let* $\alpha = \max\left\{m^{-1}\log m, k^{-0.5}\sqrt{\log k}\right\}$ *and let* ϕ *be as in Theorem 2. Then, there are* $c_3, c_4, c_5 > 0$ *depending on the parameters* μ_U*,* g' *and* g *such that, for each pair of leaves* $x, y \in \mathcal{X}$*, the* $c_3\alpha$*-quantile satisfies the following with probability at least* $1 - 6\exp\left(-c_4\left|\mathcal{M}_{\mathrm{Q}1}\right|\alpha\right)$*, provided Theorem 1* *holds,*

$$\widehat{q}_{xy}^{(c_3\alpha)} \in \left[q_{xy}^{(0)}, q_{xy}^{(0)} + c_5\alpha\right] \subset \left[r_{xy}^{(0)} - c_5\phi, r_{xy}^{(0)} + c_5\phi + c_5\alpha\right].$$

Proof Proposition 4 guarantees that $\widehat{\Delta}_{xy}$ and Δ_{xy} are close. Therefore, using the fact that $p(\cdot)$ is a Lipschitz function, we know that there exists a constant $c_5' > 0$ such that $\left|r_{xy}^{(0)} - q_{xy}^{(0)}\right| \leq c_5'\phi$. The second containment in the statement of the lemma follows from this after adjusting the constant. The first part is proved as in (Mossel and Roch 2017, Claim 12), except for a small change: we use Bernstein's inequality (see e.g. Boucheron et al. 2013) in place of Chebyshev's inequality to obtain an exponential dependence on $\left|\mathcal{M}_{\mathrm{Q}1}\right|$ (ultimately resulting in a *logarithmic dependence in* ε in the condition on $\left|\mathcal{M}_{\mathrm{Q}1}\right|$ in (26)). □

Expected version of quantile test We will use $\bar{\mathbb{P}}$, $\bar{\mathbb{E}}$, and $\overline{\mathrm{Var}}$ to denote probabilities, expectations and variances conditioned on the event that Theorem 1 and Proposition 5 hold. We use the genes in $\mathcal{M}_{\mathrm{Q}2}$, *which are not affected by the conditioning under* $\bar{\mathbb{P}}$, to define a similarity measure among pairs of leaves in $\mathcal{X}$, $\widehat{s}_{xy}$, as above. We next show that this similarity measure has the right behavior in expectation. That is, defining $s_{xy} := \bar{\mathbb{E}}\left[\widehat{s}_{xy}\right]$, we show that $s_{12} > \max\{s_{13}, s_{23}\}$ (proof in Sect. 3).

Proposition 6 (Expected version of quantile test). *For any* $C_2 > 0$, *there exist constants* $c_6, c_7 > 0$ *such that* $s_{12} - \max\{s_{13}, s_{23}\} \geq c_6 p(3f/4) > 0$ *provided*

$$m \geq c_7 \frac{1}{p(3f/4)} \log\left(\frac{1}{p(3f/4)}\right), \qquad k \geq c_7 \left(\frac{\sqrt{\log k}}{p(3f/4)}\right)^{1/C_2},$$
$$\phi \in \mathcal{O}(p(3f/4)/\sqrt{\log k}).$$

Sample version of quantile test We finish the proof by showing that the empirical similarity measures are consistent with the underlying species tree with high probability (proof in Sect. 4).

Proposition 7 (Sample version of quantile test). *There is a* $c_8 > 0$ *such that, provided Proposition* 6 *holds, the* $\bar{\mathbb{P}}$*-probability that Algorithm* 2 *fails is less than* $4\exp\left(-\frac{|\mathcal{M}_{\mathrm{Q}2}|p(3f/4)^2}{c_8(p(3f/4)+\alpha)}\right)$.

A.2 Auxiliary results

We will need a few technical auxiliary claims.

Quantiles We will need an observation about the quantiles of gene tree pairwise distances. For a pair of leaves $a, b \in L$, δ_{ab} is the branch length induced by the random gene tree under the MSC. Notice that, by definition, $\delta_{ab} \geq \mu_{ab}$. We will let $f_{ab}(\cdot)$ and $F_{ab}(\cdot)$ denote respectively the density and the cumulative density function of the random variable $Z_{ab} \triangleq \frac{\delta_{ab} - \mu_{ab}}{2}$. Because of the memoryless property of the exponential, it is natural to think of the distribution of Z_{ab} as a mixture of distributions whose supports are disjoint, corresponding to the different branches that the lineages go through before coalescing. We state this more generally as follows. Suppose that $U_1, U_2, \dots, U_r$ are subsets of $\mathbb{R}$ such that they satisfy $\sup(U_i) \le \inf(U_{i+1})$, $i = 1, 2, \dots, r$. Suppose that f is a probability density function such that $f(x) = \sum_{i=1}^{r} \omega_i f_i(x)$, where $\omega_1, \omega_2, \dots, \omega_r \in (0, 1)$ are such that $\sum_{i=1}^{r} \omega_i = 1$, and the density f_i is supported on U_i for $i = 1, \dots, r$. Then, the quantile function of f is given as follows

$$\begin{aligned}\mathbb{Q}_F(\alpha) &:= \inf\{x \in \mathbb{R} : \alpha \le F(x)\}\\ &= \sum_{i=1}^{r} \mathbb{1}\left\{\alpha \in \left[\sum_{j=0}^{i-1} \omega_j, \sum_{j=0}^{i} \omega_j\right)\right\} \mathbb{Q}_{F_i}\left(\frac{\alpha - \sum_{j=0}^{i-1} \omega_i}{\omega_i}\right),\end{aligned}$$

where, we set $\omega_0 = 0$. Specializing to f_{ab} under the MSC, it follows that there exists a finite sequence of constants $\mu_1, \dots, \mu_r \in [\mu_L, \mu_U]$ and $h_0, \dots, h_{r-1} \in [f', g' + ng]$ such that

$$\omega_i = e^{-\sum_{j=1}^{i-1} \mu_j^{-1}(h_j - h_{j-1})} - e^{-\sum_{j=1}^{i} \mu_j^{-1}(h_j - h_{j-1})}, \qquad f_i(x) = \frac{\mu_i^{-1} e^{-\mu_i^{-1}(x - h_{i-1})}}{1 - e^{-\mu_i^{-1}(h_i - h_{i-1})}}.$$

Cancellations lead to $f_{ab}(x) = \sum_{i=1}^{r} e^{-\sum_{j=1}^{i-1} \mu_j^{-1}(h_j - h_{j-1})} \mu_i^{-1} e^{-\mu_i^{-1}(x - h_{i-1})}$. This formula implies that the density is bounded between positive constants. We will need the following implication. For any $\alpha \in [0, 1)$, we let $\delta_{ab}^{(\alpha)}$ and $p_{ab}^{(\alpha)}$ denote the α-quantile of the δ_{ab} and p_{ab} respectively. Since by definition $p_{ab} = p(\delta_{ab})$, we have that $p_{ab}^{(\alpha)} = p(\delta_{ab}^{(\alpha)})$, where $p(x) = \frac{3}{4}\left(1 - e^{-4x/3}\right)$. Then, for any $0 < \beta' < \beta < 1$, there are constants $0 < c' < c'' < +\infty$ (depending on $\mu_L, \mu_U, g, g', n, \beta', \beta$) such that for any $\xi \in (0, 1 - \beta')$, we have

$$c'\xi \le \delta_{ab}^{(\beta+\xi)} - \delta_{ab}^{(\beta)} \le c''\xi, \qquad c'\xi \le p_{ab}^{(\beta+\xi)} - p_{ab}^{(\beta)} \le c''\xi. \tag{27}$$

CDF lemma Finally, we restate a key bound about the cumulative distribution function from Mossel and Roch (2017).

Lemma 6 (CDF behavior; see Claim 6 in Mossel and Roch (2017)). *There exists a constant* $c_3' > 0$ *depending on the parameters* μ_U, g' *and* g *such that* $\mathbb{P}\left[\widehat{q}_{xy} \le q_{xy}^{(0)}\right] \le \frac{c_3'}{\sqrt{k}} \le c_3'\alpha$.

A.3 Proof of Proposition 6

Proof (Proposition 6) In this proof, we are concerned with behavior in expectation. Hence, we fix an $i \in \mathcal{M}_{Q2}$ and drop the i in $\widehat{q}_{xy}^{i}$, etc. Let $\mathcal{E}_{12|3}$ be the event that there is a coalescence in the internal branch of the species phylogeny and observe that s_{12} can be decomposed as follows

$$s_{12} = \bar{\mathbb{E}}[\widehat{s}_{12}] = \bar{\mathbb{P}}\left[\widehat{q}_{12} \le \widehat{q}_*\right] = \bar{\mathbb{P}}\left[\mathcal{E}_{12|3}\right] \bar{\mathbb{P}}\left[\widehat{q}_{12} \le \widehat{q}_* | \mathcal{E}_{12|3}\right]$$

$$+ \bar{\mathbb{P}}\left[\mathcal{E}_{12|3}^c\right]\bar{\mathbb{P}}\left[\widehat{q}_{12} \leq \widehat{q}_*|\mathcal{E}_{12|3}^c\right] \tag{28}$$

In the proof in Mossel and Roch (2017), instead of $\widehat{q}_{12}$ one deals with $\widehat{r}_{12}$, and it follows from the symmetries of the MSC that $\bar{\mathbb{P}}\left[\widehat{r}_{12} \leq \widehat{q}_*|\mathcal{E}_{12|3}^c\right] = \bar{\mathbb{P}}\left[\widehat{r}_{13} \leq \widehat{q}_*\right]$. This turns out to suffice to establish the expected version of the quantile test in Mossel and Roch (2017). In our setting, we must control quantitatively the difference between these two probabilities due to the slack added by the reduction step. The following lemma is proved below. □

Lemma 7 (Closeness to symmetry). *For $\widehat{q}_*$ as defined in Algorithm 2 and any constant $C_2 > 0$, there exist constants $c_7', c_7'' > 0$ such that*

$$\left|\bar{\mathbb{P}}\left[\widehat{q}_{13} \leq \widehat{q}_*\right] - \bar{\mathbb{P}}\left[\widehat{q}_{12} \leq \widehat{q}_*|\mathcal{E}_{12|3}^c\right]\right| \leq \phi_2 := c_7'\phi\sqrt{\log k}$$

provided

$$m \geq c_7''\frac{1}{\phi\sqrt{\log k}}\log\left(\frac{1}{\phi\sqrt{\log k}}\right), \qquad k \geq \left(\frac{1}{\phi}\right)^{1/C_2}.$$

Using the above lemma in (28), we can now bound s_{12} from below as follows

$$\begin{aligned} s_{12} &\geq \bar{\mathbb{P}}\left[\mathcal{E}_{12|3}\right]\bar{\mathbb{P}}\left[\widehat{q}_{12} \leq \widehat{q}_*|\mathcal{E}_{12|3}\right] + \bar{\mathbb{P}}\left[\mathcal{E}_{12|3}^c\right]\bar{\mathbb{P}}\left[\widehat{q}_{13} \leq \widehat{q}_*\right] - \phi_2 \\ &= \bar{\mathbb{P}}\left[\mathcal{E}_{12|3}\right]\left(\bar{\mathbb{P}}\left[\widehat{q}_{12} \leq \widehat{q}_*|\mathcal{E}_{12|3}\right] - s_{13}\right) + s_{13} - \phi_2. \end{aligned} \tag{29}$$

This implies that

$$s_{12} - s_{13} \geq \bar{\mathbb{P}}\left[\mathcal{E}_{12|3}\right]\left(\bar{\mathbb{P}}\left[\widehat{q}_{12} \leq \widehat{q}_*|\mathcal{E}_{12|3}\right] - s_{13}\right) - \phi_2.$$

The expected version of the quantile test succeeds provided the latter quantity is bounded from below by 0. We establish something a bit stronger, which will be useful in the analysis of the sample version of the quantile test. The following lemma is proved below.

Lemma 8 (Tails) *There are $c_6', c_6'' > 0$ such that $\bar{\mathbb{P}}\left[\widehat{q}_{12} \leq \widehat{q}_*|\mathcal{E}_{12|3}\right] \geq c_6'$ and $s_{13} = \bar{\mathbb{P}}\left[\widehat{q}_{13} \leq \widehat{q}_*\right] \leq c_6''\alpha$.*

The first inequality captures the intuition that, conditioned on coalescence in the internal branch of the species phylogeny, the probability of $\widehat{q}_{12}$ being small is high. The second inequality captures the intuition that since $\widehat{q}_*$ behaves roughly like $q_{13}^{(\alpha)} = p(\delta_{13}^{(\alpha)} + \widehat{\Delta}_{13})$, the event that $\widehat{q}_{13} \leq \widehat{q}_*$ is dominated by the event that the underlying MSC random variable satisfies the same inequality, with the deviations of the JC contribution on top of this being of order $k^{-0.5}$.

Notice that, if we use Lemma 8 in (29), there is a constant $c_6 > 0$ such that $s_{12} - s_{13} \geq c_6\bar{\mathbb{P}}\left[\mathcal{E}_{12|3}\right]$ provided $\phi_2 \leq c_6 p(3f/4)$ for a large enough $c_6 > 0$, where we used that $\bar{\mathbb{P}}\left[\mathcal{E}_{12|3}\right]$ is lower bounded by $p(3f/4)$. This, along with a similar argument for s_{23}, concludes the proof of Proposition 6.

Proof (Lemma 7) First observe that $\bar{\mathbb{P}}\left[\widehat{r}_{13} \le \widehat{q}^* | \mathcal{E}^c_{12|3}\right] = \bar{\mathbb{P}}\left[\widehat{r}_{13} \le \widehat{q}^*\right]$ since $\widehat{r}_{13}$ is unaffected by whether coalescence occurs in the internal branch of the species phylogeny. We prove Lemma 7 by arguing that $\bar{\mathbb{P}}\left[\widehat{q}_{12} \le \widehat{q}^* | \mathcal{E}^c_{12|3}\right]$ is close to $\bar{\mathbb{P}}\left[\widehat{r}_{13} \le \widehat{q}^* | \mathcal{E}^c_{12|3}\right]$, and that $\bar{\mathbb{P}}\left[\widehat{q}_{13} \le \widehat{q}^*\right]$ is close to $\bar{\mathbb{P}}\left[\widehat{r}_{13} \le \widehat{q}^*\right]$.

We do this by analyzing the effect on the cumulative distribution function (CDF) of a perturbation of the mean. In our case, the distribution of interest is a mixture of binomials. Indeed notice that in the latter case, conditioned on the value of δ_{13} and $q_{13} = p(\delta_{13} + \widehat{\Delta}_{13})$, we are seeking to compare two binomial random variables $k\widehat{q}_{13} \sim \mathrm{Bin}(k, q_{xy})$ and $k\widehat{r}_{13} \sim \mathrm{Bin}(k, q_{xy} + \beta)$, for some small β (and similarly for the other inequality). The desired result will be implied by the following bound proved at the end of this subsection. □

Lemma 9 (Mixture of binomials: CDF perturbation). *Let δ_{xy} be the distance between x and y on a random gene tree drawn according to the MSC, and let $p_{xy} = p(\delta_{xy})$ denote the corresponding expected p-distance. Suppose that we have two binomial random variables $J_1 \sim \mathrm{Bin}(k, p_{xy})$ and $J_2 \sim \mathrm{Bin}(k, p_{xy} + \beta)$, for some fixed $\beta \in (0, 1 - p_{xy})$ with $\beta = \Theta(\phi)$, and that we are given constants $c_{14}, \gamma > 0$ such that $\gamma < p_{xy}^{(c_{14}[\alpha \vee \phi])}$. Then, for any constant $C_2 > 0$ there exist constants $c_{13}, c'_{13} > 0$ such that $\left|\bar{\mathbb{P}}\left[J_1 \le k\gamma\right] - \bar{\mathbb{P}}\left[J_2 \le k\gamma\right]\right| \le c'_{13}\beta\sqrt{\log k}$ holds provided*

$$m \ge c_{13}\frac{1}{\beta\sqrt{\log k}}\log\left(\frac{1}{\beta\sqrt{\log k}}\right), \qquad k \ge \left(\frac{1}{\beta}\right)^{1/C_2}.$$

Observe that, although J_1 and J_2 above do not depend on m, the quantity γ—through α—does.

Notice that this result implies that there exists a constant $c'_7 > 0$ such that

$$\left|\bar{\mathbb{P}}\left[\widehat{r}_{13} \le \widehat{q}^*\right] - \bar{\mathbb{P}}\left[\widehat{q}_{13} \le \widehat{q}^*\right]\right| \le \frac{c'_7}{2}\phi\sqrt{\log k},$$
$$\left|\bar{\mathbb{P}}\left[\widehat{q}_{12} \le \widehat{q}^* | \mathcal{E}^c_{12|3}\right] - \bar{\mathbb{P}}\left[\widehat{r}_{13} \le \widehat{q}^* | \mathcal{E}^c_{12|3}\right]\right| \le \frac{c'_7}{2}\phi\sqrt{\log k}.$$

To see why this is true, first observe that $\widehat{q}^* \le q_{13}^{(0)} + c_5\alpha \le q_{13}^{(c'_5\alpha)}$, which follows from Proposition 5 and (27). The bound on $\widehat{q}^*$ also holds in terms of r-quantiles up to an additive term $O(\phi)$ per Proposition 5. So we can take $\gamma = \widehat{q}^*$ in Lemma 9. Using this, and taking β to be $\Theta(\phi)$, we get the above two inequalities. Note that, for both inequalities, we apply Lemma 9 to $\widehat{r}_{13}$, $\widehat{q}_{13}$ or $\widehat{q}_{12}$ so as to keep β positive as required.

Proof (Lemma 8) We start with the second inequality in the statement of the lemma. Notice that, from Proposition 5 (on which $\bar{\mathbb{P}}$ is conditioning) and (27), we know that $\widehat{q}^* \le q_{13}^{(0)} + c_5\alpha \le q_{13}^{(c'_5\alpha)}$.

Hence,

$$\begin{aligned}\bar{\mathbb{P}}\left[\widehat{q}_{13} \leq \widehat{q}_*\right] &\leq \bar{\mathbb{P}}\left[\widehat{q}_{13} \leq q_{13}^{(c_5'\alpha)}\right]\\ &= \bar{\mathbb{P}}\left[\widehat{q}_{13} \leq q_{13}^{(c_5'\alpha)} | q_{13} \leq q_{13}^{(c_5'\alpha)}\right] \bar{\mathbb{P}}\left[q_{13} \leq q_{13}^{(c_5'\alpha)}\right]\\ &\quad + \bar{\mathbb{P}}\left[\widehat{q}_{13} \leq q_{13}^{(c_5'\alpha)} | q_{13} > q_{13}^{(c_5'\alpha)}\right] \bar{\mathbb{P}}\left[q_{13} > q_{13}^{(c_5'\alpha)}\right]\\ &\leq c_5'\alpha + \bar{\mathbb{P}}\left[\widehat{q}_{13} \leq q_{13}^{(c_5'\alpha)} | q_{13} > q_{13}^{(c_5'\alpha)}\right].\end{aligned}$$

The conditional probability on the last line is bounded above by $c_3'\alpha$ by Lemma 6 and the memoryless property of the exponential. This implies the second inequality of the lemma.

To derive the first one, we make a few observations:

(1) from Proposition 5, $\widehat{q}_* \geq \widehat{q}_{13}^{(c_3\alpha)} \geq r_{13}^{(0)} - c_5\phi$;

(2) by definition $q_{12} = p(\delta_{12} + \widehat{\Delta}_{12})$ and, conditioned on δ_{12}, $k\widehat{q}_{12}$ is distributed as $\mathrm{Bin}(k, q_{12})$;

(3) given that $q_{12} = p(\delta_{12} + \widehat{\Delta}_{12})$ and $r_{12} = p(\delta_{12} + \Delta_{12})$ and by Proposition 4, it follows that there is $c_{16} > 0$ such that the event $\{r_{12} \leq r_{13}^{(0)} - c_{16}\phi\}$ implies the event $\{q_{12} \leq r_{13}^{(0)} - c_5\phi\}$ under $\bar{\mathbb{P}}$;

(4) the event $\mathcal{E}_{12|3}$ is equivalent to the condition that $r_{12} = p(\delta_{12} + \Delta_{12}) \leq p(\mu_{13} + \Delta_{13}) = r_{13}^{(0)}$.

We use these facts to bound the desired quantity $\bar{\mathbb{P}}\left[\widehat{q}_{12} \leq \widehat{q}_* | \mathcal{E}_{12|3}\right]$ from below by

$$\begin{aligned}&\overset{(a)}{\geq} \bar{\mathbb{P}}\left[\widehat{q}_{12} \leq r_{13}^{(0)} - c_5\phi | \mathcal{E}_{12|3}\right]\\ &\overset{(b)}{\geq} \bar{\mathbb{P}}\left[\mathrm{Bin}(k, q_{12}) \leq k[r_{13}^{(0)} - c_5\phi] | q_{12} \leq r_{13}^{(0)} - c_5\phi, \mathcal{E}_{12|3}\right] \bar{\mathbb{P}}\left[q_{12} \leq r_{13}^{(0)} - c_5\phi | \mathcal{E}_{12|3}\right]\\ &\overset{(c)}{\geq} \bar{\mathbb{P}}\left[\mathrm{Bin}(k, q_{12}) \leq k[r_{13}^{(0)} - c_5\phi] | q_{12} \leq r_{13}^{(0)} - c_5\phi, \mathcal{E}_{12|3}\right] \bar{\mathbb{P}}\left[r_{12} \leq r_{13}^{(0)} - c_{16}\phi | r_{12} \leq r_{13}^{(0)}\right]\end{aligned}$$

where (a) follows from Observation 1 above, (b) follows after conditioning on the event that $q_{12} \leq r_{13}^{(0)} - c_5\phi$, and (c) follows from Observations 3 and 4. Finally, the last line is greater than some constant $C_j' > 0$ from the Berry-Esséen theorem (see e.g. Durrett 1996), which gives a constant lower bound on the first term, and (27) together with the assumption that $\phi \ll p(3f/4)$ and the fact that the probability of $\mathcal{E}_{12|3}$ is at least $p(3f/4)$, which gives a constant lower bound on the second term. □

Proof (Lemma 9) We need an auxiliary lemma which characterizes the difference between two binomial distributions in terms of the difference of the underlying probabilities. This follows from Roos (2001). □

Lemma 10 (Binomial: CDF perturbation). *For J_1 and J_2 as above and any $\gamma \in (0, 1)$, we have*

$$\left|\bar{\mathbb{P}}\left[J_1 \leq k\gamma | p_{xy}\right] - \bar{\mathbb{P}}\left[J_2 \leq k\gamma | p_{xy}\right]\right| \leq \frac{2\sqrt{2e}\sqrt{k+2}}{\sqrt{(p_{xy} + \beta)(1 - p_{xy} - \beta)}}\beta. \qquad (30)$$

Proof It holds that, if $\theta(p_{xy}+\beta) = \frac{\beta^2(k+2)}{2(p_{xy}+\beta)(1-p_{xy}-\beta)} < 1$,

$$\begin{aligned}\left|\bar{\mathbb{P}}\left[J_1 \le k\gamma | p_{xy}\right] - \bar{\mathbb{P}}\left[J_2 \le k\gamma | p_{xy}\right]\right| &\le \left\|\mathrm{Bin}(k, p_{xy}) - \mathrm{Bin}(k, p_{xy}+\beta)\right\|_1\\ &\le \frac{\sqrt{e\theta(p_{xy}+\beta)}}{\left(1-\sqrt{\theta(p_{xy}+\beta)}\right)^2},\end{aligned}$$

where the above inequality comes from (Roos 2001, (15)), by setting $s = 0$ there, and choosing the Poisson-Binomial distribution to simply be the binomial distribution $\mathrm{Bin}(k, p_{xy})$. If $\beta \le \sqrt{\frac{(p_{xy}+\beta)(1-p_{xy}-\beta)}{2(k+2)}}$, then $1-\sqrt{\theta(p_{xy}+\beta)} \ge 0.5$. In this case, we have (30). On the other hand, if $\beta > \sqrt{\frac{(p_{xy}+\beta)(1-p_{xy}-\beta)}{2(k+2)}}$, since the difference between two probabilities is upper bounded by 2, the upper bound (30) holds trivially. □

We cannot directly apply Lemma 10 to prove Lemma 9 since the $\sqrt{k+2}$ factor in the upper bound is too loose for our purposes. Instead, we employ a more careful argument that splits the domain of p_{xy}.

$p_{xy} \in I_1 = [p_{xy}^{(0)}, p_{xy}^{(2c_{14}\sqrt{\frac{\log k}{k}})}]$ *(low substitution regime for small k)*: In this case, we use the fact that Lemma 10 guarantees that the binomial distributions are $\mathcal{O}(\beta\sqrt{k})$ apart. That is, there exists a constant $c'_{15} > 0$ such that

$$\begin{aligned}&\bar{\mathbb{P}}\left[p_{xy} \in I_1\right]\bar{\mathbb{E}}\left[\left|\bar{\mathbb{P}}\left[J_1 \le k\gamma\right] - \bar{\mathbb{P}}\left[J_2 \le k\gamma\right]\right| | p_{xy} \in I_1\right]\\ &\quad\le \bar{\mathbb{P}}\left[p_{xy} \in I_1\right] c'_{15}\beta\sqrt{k+2} \le c'_{15}\beta\sqrt{\log k},\end{aligned}$$

where the last step follows from the quantile definition, after appropriately increasing the constant c'_{15}.

$p_{xy} \in I_2 = [p_{xy}^{(2c_{14}\sqrt{\frac{\log k}{k}})}, p_{xy}^{(2c_{14}[\alpha\vee\phi])}]$ (*low substitution regime for large k*):

In the case this interval is not empty, there exists a constant $c''_{15} > 0$ such that

$$\begin{aligned}\bar{\mathbb{P}}\left[p_{xy} \in I_2\right]\mathbb{E}\left[\left|\bar{\mathbb{P}}\left[J_1 \le k\gamma\right] - \bar{\mathbb{P}}\left[J_2 \le k\gamma\right]\right| | p_{xy} \in I_2\right] &\le \bar{\mathbb{P}}[p_{xy} \in I_2]\\ &\overset{(a)}{\le} c''_{15}[\alpha \vee \phi]\\ &\overset{(b)}{\le} c''_{15}\frac{\log m}{m} \vee \phi,\end{aligned}$$

where (a) follows from (27), and (b) follows from the definition of α.

$p_{xy} \in I_3 = [p_{xy}^{(2c_{14}[\alpha\vee\phi])}, 0.5]$ *(high substitution regime)*: In this case observe that, since $\gamma < p_{xy}^{(c_{14}[\alpha\vee\phi])}$ (i.e., we are looking at a left tail below the mean), we can apply Chernoff's bound (see e.g. Boucheron et al. 2013) on each of the two terms in the difference individually. For any constant $C_2 > 0$, we can choose $c_{14} > 0$ so that the following inequality holds for some $c'''_{15} > 0$

$$\bar{\mathbb{P}}\left[p_{xy} \in I_3\right]\mathbb{E}\left[\left|\bar{\mathbb{P}}\left[J_1 \le k\gamma\right] - \bar{\mathbb{P}}\left[J_2 \le k\gamma\right]\right| | p_{xy} \in I_3\right]$$

$$\begin{aligned}&\leq 2\exp(-k(p_{xy}^{(2c_{14}[\alpha\vee\phi])} - p_{xy}^{(c_{14}[\alpha\vee\phi])})^2/2)\\ &\leq c_{15}'''k^{-C_2},\end{aligned}$$

where we used that $p_{xy}^{(2c_{14}[\alpha\vee\phi])} - p_{xy}^{(c_{14}[\alpha\vee\phi])} = \Omega(\sqrt{\frac{\log k}{k}})$ by (27) and the definition of α.

Putting the three regimes together, there is a constant $c_7' > 0$ (not depending on f, m, k) such that

$$\left|\bar{\mathbb{P}}\left[J_1 \leq k\gamma\right] - \bar{\mathbb{P}}\left[J_2 \leq k\gamma\right]\right| \leq \frac{c_7'}{3}\left(\beta\sqrt{\log k} + \frac{\log m}{m}\vee\phi + k^{-C_2}\right).$$

There is a c_{13} such that the conditions of the lemma imply $m^{-1}\log m \leq \beta\sqrt{\log k}$ and $k^{-C_2} \leq \beta\sqrt{\log k}$.

A.4 Proof of Proposition 7

Recall that $\mathcal{E}_{12|3}$ is the event that there is a coalescence in the internal branch of the species phylogeny. First we prove:

Lemma 11 (Variance bound). *There is a $c_8' > 0$ such that, $\forall x, y$, $\overline{\mathrm{Var}}(\widehat{s}_{xy}) \leq \frac{c_8'}{|\mathcal{M}_{Q2}|}\left(\bar{\mathbb{P}}\left[\mathcal{E}_{12|3}\right] + \phi + \alpha\right)$.*

Proof Note that the variance of $\widehat{s}_{12}$ is bounded from above by $\frac{1}{|\mathcal{M}_{Q2}|}\bar{\mathbb{P}}\left[\widehat{q}_{12} \leq \widehat{q}_*\right]$.

Observe further that, by Proposition 5, $\widehat{q}_* \leq r_{13}^{(0)} + c_5\phi + c_5\alpha$. Moreover, conditioned on the random distance δ_{12}, $k\,\widehat{q}_{12}$ is distributed according to $\mathrm{Bin}(k, q_{12})$, where $q_{12} = p\left(\delta_{12} + \widehat{\Delta}_{12}\right)$. Finally, from Lemma 6 and the memoryless property of the exponential, we have that

$$\bar{\mathbb{P}}\left[\widehat{q}_{12} \leq r_{13}^{(0)} + c_5\phi + c_5\alpha | q_{12} > r_{13}^{(0)} + c_5\phi + c_5\alpha\right] \leq c_3'\alpha.$$

Therefore, arguing as in the proof of Lemma 8,

$$\bar{\mathbb{P}}\left[\widehat{q}_{12} \leq \widehat{q}_*\right] \leq \bar{\mathbb{P}}\left[q_{12} \leq r_{13}^{(0)} + c_5\phi + c_5\alpha\right] + c_3'\alpha. \tag{31}$$

From (27), it follows that there is a constant $c_8'' > 0$ such that

$$\bar{\mathbb{P}}\left[q_{12} \leq r_{13}^{(0)} + c_5\phi + c_5\alpha\right] \leq c_8''\left(r_{13}^{(0)} + c_5\phi + c_5\alpha - q_{12}^{(0)}\right). \tag{32}$$

Moreover,

$$r_{13}^{(0)} - q_{12}^{(0)} \leq p\left(\delta_{13}^{(0)} + \Delta_{13}\right) - p\left(\delta_{12}^{(0)} + \Delta_{12}\right) + c_8'''\phi \leq c_8'''\left(\bar{\mathbb{P}}\left[\mathcal{E}_{12|3}\right] + \phi\right).$$

The last inequality follows from the fact that $\bar{\mathbb{P}}\left[\mathcal{E}_{12|3}\right]$ is of the order of the length of the internal branch of the species phylogeny; and so is the difference in the middle expression. Notice that this along with (31) and (32) imply that there is a constant $c_8' > 0$ (after changing it appropriately) such that $\bar{\mathbb{P}}\left[\widehat{q}_{12} \leq \widehat{q}_*\right] \leq c_8' \left(\bar{\mathbb{P}}\left[\mathcal{E}_{12|3}\right] + \phi + \alpha\right)$.□

Proof (Proposition 7) An error in the quantile test implies that either $\widehat{s}_{13} > \widehat{s}_{12}$ or $\widehat{s}_{23} > \widehat{s}_{12}$. Therefore, the probability that the algorithm makes an error is at most

$$\begin{aligned}
&\bar{\mathbb{P}}\left[\widehat{s}_{13} \geq \widehat{s}_{12}\right] + \bar{\mathbb{P}}\left[\widehat{s}_{23} \geq \widehat{s}_{12}\right] \\
&\leq \bar{\mathbb{P}}\left[\widehat{s}_{13} - s_{13} \geq \frac{s_{12} - s_{13}}{2}\right] + \bar{\mathbb{P}}\left[s_{12} - \widehat{s}_{12} \geq \frac{s_{12} - s_{13}}{2}\right] \\
&\quad + \bar{\mathbb{P}}\left[\widehat{s}_{23} - s_{23} \geq \frac{s_{12} - s_{23}}{2}\right] + \bar{\mathbb{P}}\left[s_{12} - \widehat{s}_{12} \geq \frac{s_{12} - s_{23}}{2}\right]. \qquad (33)
\end{aligned}$$

Take the second term on r.h.s. (the other terms being similar). We need two ingredients to invoke Bernstein's inequality: (1) a lower bound on the "gap" $\frac{s_{12}-s_{13}}{2}$ (Proposition 6), and (2) an upper bound on the variance of $\widehat{s}_{12}$ (Lemma 11). We get that $\bar{\mathbb{P}}\left[s_{12} - \widehat{s}_{12} \geq \frac{s_{12}-s_{13}}{2}\right]$ is at most

$$\begin{aligned}
&\exp\left(-\frac{0.5\left(\frac{s_{12}-s_{13}}{2}\right)^2}{\mathrm{Var}(\widehat{s}_{12}) + \frac{1}{6}(s_{12} - s_{13})}\right) \\
&\leq \exp\left(-\frac{\left|\mathcal{M}_{Q2}\right| \left(c_6 \bar{\mathbb{P}}\left[\mathcal{E}_{12|3}\right]\right)^2}{c_8' \left(\alpha + \phi + \bar{\mathbb{P}}\left[\mathcal{E}_{12|3}\right]\right) + \frac{c_6}{6}\bar{\mathbb{P}}\left[\mathcal{E}_{12|3}\right]}\right) \\
&\leq \exp\left(-\frac{\left|\mathcal{M}_{Q2}\right| p(3f/4)^2}{c_8\left(p(3f/4) + \alpha\right)}\right), \qquad (34)
\end{aligned}$$

where the last line follows from the fact that $\bar{\mathbb{P}}\left[\mathcal{E}_{12|3}\right]$ is bounded from below by $p(3f/4)$ and the assumption that $\phi \ll p(3f/4)$. □

References

Allman ES, Degnan JH, Rhodes JA (2011) Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. J Math Biol 62(6):833–862

Allman ES, Degnan JH, Rhodes JA (2018) Species tree inference from gene splits by unrooted star methods. IEEE/ACM Trans Comput Biol Bioinf 15(1):337–342

Allman ES, Long C, Rhodes JA (2019) Species tree inference from genomic sequences using the log-det distance. SIAM J Appl Algebra Geom 3(1):107–127 (**Publisher: Society for Industrial and Applied Mathematics**)

Boucheron S, Lugosi G, Massart P (2013) Concentration inequalities: a nonasymptotic theory of independence. Oxford University Press, Oxford

Bayzid MS, Mirarab S, Boussau B, Warnow T (2015) Weighted statistical binning: enabling statistically consistent genome-scale phylogenetic analyses. PLoS ONE 10(6):e0129183–e0129183 (**06**)

Bayzid MS, Warnow T (2013) Naive binning improves phylogenomic analyses. Bioinformatics 29(18):2277–2284 (**07**)

Chifman J, Kubatko L (2014) Quartet inference from SNP data under the coalescent model. Bioinformatics 30(23):3317–3324

Chifman J, Kubatko L (2015) Identifiability of the unrooted species tree topology under the coalescent model with time-reversible substitution processes, site-specific rate variation, and invariable sites. J Theor Biol 374:35–47

DeGiorgio M, Degnan JH (2010) Fast and consistent estimation of species trees using supermatrix rooted triples. Mol Biol Evol 27(3):552–69

DeGiorgio M, Degnan JH (2014) Robustness to divergence time underestimation when inferring species trees from estimated gene trees. Syst Biol 63(1):66

Degnan JH, DeGiorgio M, Bryant D, Rosenberg NA (2009) Properties of consensus methods for inferring species trees from gene trees. Syst Biol 58(1):35–54

Dasarathy G, Nowak R, Roch S (2015) Data requirement for phylogenetic inference from multiple loci: a new distance method. Comput Biol Bioinform IEEE/ACM Trans 12(2):422–432

Degnan JH, Rosenberg NA (2006) Discordance of species trees with their most likely gene trees. PLoS Genetics, 2(5)

Degnan JH, Rosenberg NA (2009) Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol Evol 24(6):332–340

Durrett R (1996) Probability: theory and examples, 2nd edn. Duxbury Press, Belmont, CA

Erdos PL, Steel MA, Székely LA, Warnow TJ (1999) A few logs suffice to build (almost) all trees (i). Random Struct Algorithms 14(2):153–184

Kubatko LS, Degnan JH (2007) Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst Biol 56(1):17–24

Kapli P, Yang Z, Telford MJ (2020) Phylogenetic tree building in the genomic age. Nature Rev Gene

Long C, Kubatko L (2017) Identifiability and reconstructibility of species phylogenies under a modified coalescent. Arxiv publication arXiv:1701.06871

Long C, Kubatko L (2019) Identifiability and reconstructibility of species phylogenies under a modified coalescent. Bull Math Biol 81(2):408–430

Liu L, Yu L, Pearl DK (2010) Maximum tree: a consistent estimator of the species tree. J Math Biol 60(1):95–106

Maddison WP (1997) Gene trees in species trees. Syst Biol 46(3):523–536

Mirarab S, Bayzid Md. S, Boussau B, Warnow T (2014) Statistical binning enables an accurate coalescent-based estimation of the avian tree. Science, 346(6215)

Mirarab S, Bayzid MS, Warnow T (2016) Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. System Biol 65(3):366

Mossel E, Roch S (2010) Incomplete lineage sorting: consistent phylogeny estimation from multiple loci. IEEE/ACM Trans Comput Biol Bioinform 7(1):166–171

Mossel E, Roch S (2017) Distance-based species tree estimation under the coalescent: information-theoretic trade-off between number of loci and sequence length. Ann Appl Probab 27(5):2926–2955

Matsen FA, Steel M (2007) Phylogenetic mixtures on a single tree can mimic a tree of another topology. Syst Biol 56(5):767–775

Nakhleh L (2013) Computational approaches to species phylogeny inference and gene tree reconciliation. Trends Ecol Evol. doi: https://doi.org/10.1016/j.tree.2013.09.004

Rhodes JA (2019) Topological metrizations of trees, and new quartet methods of tree inference. IEEE/ACM Trans Comput Biol Bioinform

Rusinko J, McPartlon M (2017) Species tree estimation using neighbor joining. J Theor Biol 414:5–7

Roch S, Nute M, Warnow T (2019) Long-branch attraction in species tree estimation: inconsistency of partitioned likelihood and topology-based summary methods. System Biol 68(2):281–297

Roch S (2013) An analytical comparison of multilocus methods under the multispecies coalescent: the three-taxon case. In: Biocomputing 2013: proceedings of the pacific symposium, Kohala Coast, Hawaii, USA, January 3-7, 2013, pp 297–306

Roch S (2018) On the variance of internode distance under the multispecies coalescent. In: Comparative genomics - 16th international conference, RECOMB-CG 2018, Magog-Orford, QC, Canada, October 9-12, 2018, Proceedings, pp 196–206

Roos B (2001) Binomial approximation to the poisson binomial distribution: the Krawtchouk expansion. Theory Prob Its Appl 45(2):258–272

Roch S, Steel M (2015) Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. Theor Popul Biol 100:56–62

Roch S, Warnow T (2015) On the robustness to gene tree estimation error (or lack thereof) of coalescent-based species tree methods. Syst Biol 64(4):663–676

Rannala B, Yang Z (2003) Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164(4):1645–1656

Scornavacca C, Delsuc F, Galtier N (2020) Phylogenetics in the Genomic Era. No commercial publisher | Authors open access book

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4(4):406–425

Shekhar S, Roch S, Mirarab S (2017) Species tree estimation using ASTRAL: how many genes are enough? In: RECOMB'17—proceedings of the 21st annual international conference on research in computational molecular biology, pp 393–395

Steel MA, Székely LA (2002) Inverting random functions. II. Explicit bounds for discrete maximum likelihood estimation, with applications. SIAM J. Discrete Math. 15(4):562–575 (**electronic**)

Semple C, Steel M (2003) Phylogenetics, vol 22. Mathematics and its Applications Series. Oxford University Press, Oxford

Semple C, Steel MA (2003) Phylogenetics, vol 24. Oxford University Press, Oxford

Steel M (2009) A basic limitation on inferring phylogenies by pairwise sequence comparisons. J Theor Biol 256(3):467–472

Steel M (2016) Phylogeny—discrete and random processes in evolution, vol 89 CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA

Warnow T (2017) Computational Phylogenetics: an introduction to designing methods for phylogeny estimation. Cambridge University Press, Cambridge