Complexity of the simplest species tree problem

Tianqi Zhu,[1,2] and Ziheng Yang[3,1,*]

[1]Institute of Applied Mathematics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

[2]Key Laboratory of Random Complex Structures and Data Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

[3]Department of Genetics, University College London, Gower Street, London WC1E 6BT, UK

Corresponding author: E-mail: z.yang@ucl.ac.uk.

Associate Editor: xxx xxx

Abstract

The multispecies coalescent (MSC) model provides a natural framework for species tree estimation accounting for gene-tree conflicts. While a number of species tree methods under the MSC have been suggested and evaluated using simulation, their statistical properties remain poorly understood. Here we use mathematical analysis aided by computer simulation to examine the identifiability, consistency, and efficiency of different species tree methods in the case of three species and three sequences under the molecular clock. We consider four major species-tree methods including concatenation, two-step, independent-sites maximum likelihood (ISML) and maximum likelihood (ML). We develop approximations that predict that the probit transform of the species tree estimation error decreases linearly with the square root of the number of loci. Even in this simplest case major differences exist among the methods. Full-likelihood methods are considerably more efficient than summary methods such as concatenation and two-step. They also provide estimates of important parameters such as species divergence times and ancestral population sizes while these parameters are not identifiable by summary methods. Our results highlight the need to improve the statistical efficiency of summary methods and the computational efficiency of full likelihood methods of species tree estimation.

Key words: Concatenation, efficiency, molecular clock, MSC, multispecies coalescent, species tree

Introduction

The multispecies coalescent (MSC) model (Rannala and Yang 2003) combines the phylogenetic process of species divergences with the population genetic process of coalescent and naturally accommodates *delayed coalescence* (also known as *incomplete lineage sorting*, Maddison 1997), the phenomenon in which gene sequences fail to coalesce in their most recent common ancestor but do so only in more ancient ancestors. Delayed coalescence causes the gene tree for a gene or genomic region to differ from the species tree and is the most important factor for gene-tree–species-tree discordance (Maddison 1997; Nichols 2001; Szollosi *et al.* 2015). The MSC provides a natural framework for estimating species trees accounting for genealogical heterogeneity among genes or across the genome (Edwards 2009; Xu and Yang 2016; Kubatko 2019; Rannala *et al.* 2020).

Two lines of research into the MSC have provided the foundation for species tree methods. The first concerns the probabilities of different gene tree topologies (Hudson 1983; Pamilo and Nei 1988) and algorithms for their efficient calculation given the species tree (Degnan and Salter 2005; Degnan and Rosenberg 2006). The gene tree distribution can be used in the two-step method of species tree estimation, by inferring gene trees for the individual loci and then applying maximum likelihood (ML) to counts of gene tree topologies (as in STELLS, Wu 2012). Nevertheless, widely used two-step methods, including ASTRAL (Mirarab *et al.* 2014) and MP-EST (Liu *et al.* 2010), are simpler, and estimate species trees for species triplets (assuming the molecular clock) or quartets (without the clock) and then assemble the subtrees to produce a species-tree estimate for all species. Studies of gene-tree probabilities led to the discovery of the *anomaly zone*, the region of the parameter space in which the most probable gene tree has a different topology from the species tree (Degnan and Salter 2005; Degnan and Rosenberg 2006). In the anomaly zone, the two-step method, which uses the most common gene tree as the species tree estimate, will be inconsistent.

The second line of research into MSC is the development of the joint probability distribution of the gene

tree and coalescent times (Rannala and Yang 2003). This forms the basis for exact methods of inference, including ML (Yang 2002; Dalquen *et al.* 2017) and Bayesian methods (Liu and Pearl 2007; Heled and Drummond 2010; Yang and Rannala 2014; Ogilvie *et al.* 2017; Rannala and Yang 2017). While heuristic methods use summaries of the data, exact methods use the multi-locus sequence alignments directly and naturally accommodate phylogenetic reconstruction errors and uncertainties (Xu and Yang 2016; Kubatko 2019; Rannala *et al.* 2020).

Simulation has been used to examine the performance of different species-tree methods (e.g., Leaché and Rannala 2011; Mirarab *et al.* 2014; Chou *et al.* 2015; Xu and Yang 2016). A limitation of simulation is that it can examine only a small portion of the parameter space and the results often have limited applicability. Analytical results on the efficiency of different methods have been lacking. Here we analyze species tree estimation under the MSC in the case of three species, with one sequence from each species per locus. We focus on closely related species and assume the JC mutation model (Jukes and Cantor 1969) and the molecular clock. We are in particular interested in the efficiency of the various methods, measured by the probability of recovering the correct species tree.

We consider four inference methods: (i) ML (a full likelihood method under the MSC applied to the multilocus sequence alignments), (ii) 2-STEP (or majority-vote), (iii) concatenation (CONCAT), and (iv) independent-sites ML (ISML, also known as coalescent-aware concatenation or CONCAT) (Xu and Yang 2016). ML is the full-likelihood method and calculates the likelihood function using the multi-locus sequence alignments or a sufficient summary. The 2-STEP method estimates the gene tree at each locus and then uses the most common gene tree as the species tree estimate. It does not account for the uncertainties in the estimated gene trees. For the case of three species considered here, 2-STEP is equivalent to the maximum pseudo-likelihood method (MP-EST) (Liu *et al.* 2010). Concatenation applies ML to the concatenated sequences, assuming that the same tree underlies all sites in the super alignment. In the case considered here, concatenation is equivalent to STEAC (Liu *et al.* 2009), which uses average coalescent times over loci as data to infer a gene tree, which is the species tree estimate. ISML (or CONCAT) estimates the species tree by ML under the assumption that all sites, both from the same locus and from different loci, have independent gene trees (Xu and Yang 2016). This was suggested as an improvement to SVDQUARTETS of Chifman and Kubatko (2014). All four methods considered here use ML, but the likelihood function is applied to different summaries of the same data. Here we refer to the full-likelihood or full-data method as the ML method, while all other methods (2-STEP, concatenation, and ISML)

are considered heuristic summary methods: 2-STEP uses the (estimated) gene tree topologies, while concatenation and ISML use the site-pattern counts pooled across loci. We derive approximations to the error rate of species tree estimation by the different methods and assess their accuracy. We use the theory to characterize the differences in the use of information in the data by different methods.

Results

Multispecies coalescent in the case of three species

For three species A, B, and C, there are three possible species trees: $S_1 = ((AB)C)$, $S_2 = ((BC)A)$, and $S_3 = ((CA)B)$, each with two divergence times (τ_0 and τ_1) and two population sizes (θ_0 and θ_1) (fig. 1a). Both τs and θs are measured by the expected number of mutations per site. For each species, the population size parameter is $\theta = 4N\mu$, where N is the (effective) population size and μ is the mutation rate per site per generation. We consider only one sequence from each species, so that θs for the modern species are not considered. The parameters have different interpretations in different species trees: in S_1, the two ancestral species are AB and ABC so the parameters are $\boldsymbol{\theta}_1 = \{\tau_0, \tau_1, \theta_0, \theta_1\} = \{\tau_{ABC}, \tau_{AB}, \theta_{ABC}, \theta_{AB}\}$.

At each locus, three sequences (a, b, and c) are sampled, one from each species. They are related through a gene tree. The three possible gene trees are $G_1 = ((ab)c)$, $G_2 = ((bc)a)$, and $G_3 = ((ca)b)$, with probabilities

$$\begin{aligned} \mathbb{P}(G_1|S_1, \boldsymbol{\theta}_1) &= 1 - \tfrac{2}{3}\phi, \\ \mathbb{P}(G_2|S_1, \boldsymbol{\theta}_1) &= \mathbb{P}(G_3|S_1, \boldsymbol{\theta}_1) = \tfrac{1}{3}\phi, \end{aligned} \quad (1)$$

where $\phi = \mathrm{e}^{-2(\tau_{ABC} - \tau_{AB})/\theta_{AB}}$ is the probability that sequences a and b do not coalesce in population AB so that all three sequences enter the ancestor ABC and the three gene trees occur with equal probability (fig. 1b) (Hudson 1983). Here $2(\tau_{ABC} - \tau_{AB})/\theta_{AB}$ is known as the internal branch length in coalescent units, as the average coalescent time in population AB is $2N_{AB}$ generations or $\theta_{AB}/2$ mutations per site.

For locus i, let $\boldsymbol{t}_i = \{t_{i0}, t_{i1}\}$ be the coalescent times (node ages) on the gene tree (fig. 1b). The joint MSC density for the gene tree and coalescent times given species tree S_1 and parameters $\boldsymbol{\theta}_1$ is then

$$f(G_{1a}, \boldsymbol{t}_i|S_1, \boldsymbol{\theta}_1) = \tfrac{2}{\theta_1} \mathrm{e}^{-\frac{2}{\theta_1}(t_{i1} - \tau_1)} \cdot \tfrac{2}{\theta_0} \mathrm{e}^{-\frac{2}{\theta_0}(t_{i0} - \tau_0)},$$

$$\tau_1 < t_{i1} < \tau_0, t_{i0} > \tau_0,$$

$$f(G_k, \boldsymbol{t}_i|S_1, \boldsymbol{\theta}_1) = \mathrm{e}^{-\frac{2}{\theta_1}(\tau_0 - \tau_1)} \qquad (2)$$

$$\times \tfrac{2}{\theta_0} \tfrac{2}{\theta_0} \mathrm{e}^{-\frac{6}{\theta_0}(t_{i1} - \tau_0) - \frac{2}{\theta_0}(t_{i0} - t_{i1})},$$

$$t_{i1} > \tau_0, t_{i0} > t_{i1},$$

for $k = 1b, 2, 3$ (Takahata *et al.* 1995; Yang 2002). The probability densities for S_2 and S_3 are given similarly.

The data consist of sequence alignments at m loci. Under the JC mutation model, the data at locus i can be summarized as counts of five site patterns: xxx, xxy, yxx, xyx, and xyz, where x, y, z are any three distinct nucleotides. Let those counts be $\boldsymbol{x}_i = \{x_{i0}, x_{i1}, x_{i2}, x_{i3}, x_{i4}\}$, with $\sum_{j=0}^{4} x_{ij} = n$ to be the number of sites (sequence length) at each locus. Let $f_{ij} = x_{ij}/n$ be the frequencies. Let data at all m loci be $\boldsymbol{x} = \{\boldsymbol{x}_i\}$.

Given the gene tree and coalescent times at locus i, the probability of the sequence data, $f(\boldsymbol{x}_i|G_i, \boldsymbol{t}_i)$, is given by the multinomial distribution for the five site patterns. For example, given gene tree G_1 with node ages t_{i0} and t_{i1} (fig. 1b), the site-pattern probabilities, $\boldsymbol{p}_i = \{p_{i0}, p_{i1}, p_{i2}, p_{i3}, p_{i4}\}$, are

$$p_{i0} = \mathbb{P}(xxx|G_1, \boldsymbol{t}_i) = \tfrac{1}{16}(1 + 3v^2 + 6u + 6uv),$$
$$p_{i1} = \mathbb{P}(xxy|G_1, \boldsymbol{t}_i) = \tfrac{1}{16}(3 + 9v^2 - 6u - 6uv),$$
$$p_{i2} = \mathbb{P}(yxx|G_1, \boldsymbol{t}_i) = \tfrac{1}{16}(3 - 3v^2 + 6u - 6uv), \quad (3)$$
$$p_{i3} = \mathbb{P}(xyx|G_1, \boldsymbol{t}_i) = p_2,$$
$$p_{i4} = \mathbb{P}(xyz|G_1, \boldsymbol{t}_i) = \tfrac{1}{16}(6 - 6v^2 - 12u + 12uv),$$

where $u = \mathrm{e}^{-8t_{i0}/3}$ and $v = \mathrm{e}^{-4t_{i1}/3}$ (Yang 1994b). Note that $p_{i1} > p_{i2} = p_{i3}$ as $t_{i0} > t_{i1}$. The probabilities for gene trees G_2 or G_3 are given by symmetry. Then the sequence data or the five site-pattern counts at the locus have the multinomial probabilities

$$f(\boldsymbol{x}_i|G_1, \boldsymbol{t}_i) = p_{i0}^{x_{i0}} p_{i1}^{x_{i1}} p_{i2}^{x_{i2}+x_{i3}} p_{i4}^{x_{i4}},$$
$$f(\boldsymbol{x}_i|G_2, \boldsymbol{t}_i) = p_{i0}^{x_{i0}} p_{i1}^{x_{i2}} p_{i2}^{x_{i3}+x_{i1}} p_{i4}^{x_{i4}}, \quad (4)$$
$$f(\boldsymbol{x}_i|G_3, \boldsymbol{t}_i) = p_{i0}^{x_{i0}} p_{i1}^{x_{i3}} p_{i2}^{x_{i1}+x_{i2}} p_{i4}^{x_{i4}}.$$

The ML method of species tree estimation

The log-likelihood function for species tree S_1 with parameters $\boldsymbol{\theta}_1$ is given by summing over the gene trees and integrating over the coalescent times.

$$\ell_1(\boldsymbol{\theta}_1) = \sum_{i=1}^{m} \log f(\boldsymbol{x}_i|S_1, \boldsymbol{\theta}_1)$$
$$= \sum_{i=1}^{m} \log \left\{ \sum_{G_i} \int f(G_i, \boldsymbol{t}_i|S_1, \boldsymbol{\theta}_1) f(\boldsymbol{x}_i|G_i, \boldsymbol{t}_i) \mathrm{d}\boldsymbol{t}_i \right\}, \quad (5)$$

where $f(G_i, \boldsymbol{t}_i|S_1, \boldsymbol{\theta}_1)$ is the MSC density for the gene tree and coalescent times at locus i (eq. 2), and $f(\boldsymbol{x}_i|G_i, \boldsymbol{t}_i)$ is the probability of the sequence data at locus i given the gene tree (eq. 4). The log likelihood functions, $\ell_2(\boldsymbol{\theta}_2)$ and $\ell_3(\boldsymbol{\theta}_3)$, for S_2 (with parameters $\boldsymbol{\theta}_2$) and S_3 (with $\boldsymbol{\theta}_3$) are defined similarly.

Maximizing the log-likelihood function (eq. 5) with respect to the parameters will lead to a log-likelihood value for the given species tree, and the species tree that achieves the highest ℓ is the ML species tree. This is not analytically tractable. The program 3s implements the method by explicitly summing over the gene trees (G_i)

and by using Gaussian quadrature to calculate the 2-D integrals over $\boldsymbol{t}_i$ (eq. 5) (Yang 2002; Zhu and Yang 2012; Dalquen *et al.* 2017). This is used in simulations.

We present two theorems for approximating the error in species tree estimation.

THEOREM 1. *(a) Suppose* $z_i = (z_{i1}, z_{i2}, z_{i3})^T$, $i = 1, \cdots, m$, *are an independent and identically distributed (i.i.d.) sample of size m from a distribution with means* $\mu = (\mu_1, \mu_2, \mu_2)^T$, *with* $\Delta\mu = \mu_1 - \mu_2 > 0$, *and variances* $\Sigma = \{\sigma_{jk}\}$, *where* $\sigma_{11} = \sigma_1^2$, $\sigma_{12} = \sigma_{13} = \rho_{12}\sigma_1\sigma_2$, $\sigma_{22} = \sigma_{33} = \sigma_2^2$ *and* $\sigma_{23} = \rho_{23}\sigma_2^2$. *Let* $\bar{z} = \{\bar{z}_1, \bar{z}_2, \bar{z}_3\}^T$ *be the sample means, with* $\bar{z}_j = \frac{1}{m}\sum_{i=1}^{m} z_{ij}$, $j = 1, 2, 3$. *For large* m, $\bar{z} \sim \mathbb{N}_3(\mu, \frac{1}{m}\Sigma)$. *Let* $\zeta = \mathbb{P}\{\bar{z}_1 < \max(\bar{z}_2, \bar{z}_3)\}$. *Then*

$$\zeta \approx \Phi\left(\frac{-\Delta\mu\sqrt{m} + \sqrt{\frac{1}{\pi}(\sigma_2^2 - \sigma_{23})}}{\sqrt{\sigma_1^2 - 2\sigma_{12} + \sigma_2^2 - \frac{1}{\pi}(\sigma_2^2 - \sigma_{23})}} \right) \quad (6)$$
$$\equiv \zeta_N(m, \Delta\mu, \sigma_1^2, \sigma_2^2, \sigma_{12}, \sigma_{23}),$$

where Φ is the cumulative distribution function (CDF) for the normal distribution $\mathbb{N}(0, 1)$. We also write ζ_N as $\zeta_N(m, \Delta\mu, \Sigma)$.

(b) Let $a = s_2/s_1$ and $b = \Delta\mu\sqrt{m}/s_1$, with $s_1^2 = \sigma_1^2 - 2\sigma_{12} + \sigma_2^2 - \frac{1}{2}(\sigma_2^2 - \sigma_{23})$ and $s_2^2 = \frac{1}{2}(\sigma_2^2 - \sigma_{23})$. Then ζ is bounded by

$$\Phi(-h)\left(1 + \tfrac{2}{\pi}\tan^{-1}a\right) \leq \zeta < 2\Phi(-h)$$
$$= \Phi\left(-h + \tfrac{1}{h}\log 2 + o(\tfrac{1}{h})\right), \quad (7)$$

where $h = \frac{b}{\sqrt{1+a^2}}$. The equality for the lower bound holds when $h = b = 0$. We write those bounds as $\zeta_{L1} < \zeta < \zeta_{U1}$, so that $\Phi(-h) \leq \zeta_{L1} \leq \zeta < 2\Phi(-h)$.

Proof. A proof is given in Appendix A, in which we discuss alternative approximations and also give a tighter pair of bounds (ζ_{L2}, ζ_{U2}) in eq. A27, with $\zeta_{L1} < \zeta_{L2} < \zeta < \zeta_{U2} < \zeta_{U1}$. □

In this paper, ζ represents the error probability of species tree estimation. Thus the bounds $\Phi(-h) \leq \zeta < 2\Phi(-h)$ suggest that when $m \to \infty$, the probit transform of the species-tree error probability, $\Phi^{-1}(\zeta)$, where Φ^{-1} is the inverse CDF of $\mathbb{N}(0, 1)$, decreases linearly with $\sqrt{m}$. For practical calculations for finite m in this paper, eq. 6 is more accurate (see Appendix A) and will be used later.

COROLLARY 2. *Let (y_0, y_1, y_2, y_3) be random variables from the multinomial distribution $MN(m, q_0, q_1, q_2, q_3)$, with $q_0 = 1 - q_1 - q_2 - q_3$, $q_1 > q_2 = q_3$, and $\Delta q = q_1 - q_2 > 0$. Then $\mathbb{P}\{y_1 < \max(y_2, y_3)\}$ can be approximated*

by

$$\zeta(m, q_1, q_2) = \Phi\left(\frac{-\Delta q\sqrt{m} + \sqrt{\frac{q_2}{\pi}}}{\sqrt{q_1 + q_2 - (\Delta q)^2 - \frac{q_2}{\pi}}}\right), \quad (8)$$

$$\zeta_{\text{ZLY}}(m, q_1, q_2) = \Phi\left(\frac{-\Delta q\sqrt{m - \frac{1}{\Delta q}} + \sqrt{\frac{q_2}{\pi}}}{\sqrt{q_1 + q_2 - \frac{q_2}{\pi}}}\right). \quad (9)$$

Proof. Let $\bar{z}_j = y_j/m$, $j = 1, 2, 3$ be the observed frequencies. We have $\sigma_{jj} = q_j(1 - q_j)$ and $\sigma_{jk} = -q_j q_k$ for $j \neq k$. Then eq. 8 follows from eq. 6 in Theorem 1. The form ζ_{ZLY}, an alternative to eq. 8, is from Yang (1996, eq. 3), based on Zharkikh and Li (1992, eq. 20). This applies the term $1/\Delta q$ to correct for discontinuity (Fleiss *et al.* 2003) and ignores correlations between y_1, y_2, and y_3 as well as some terms of small probabilities. The discontinuity correction does not appear to be useful. If $m \gg 1/\Delta q$, both forms, with and without the discontinuity correction, are very close. □

The error rate for the ML method (eq. 5) is analyzed in Appendix B. When the number of loci $m \to \infty$, the MLE $\hat{\boldsymbol{\theta}}_j \to \boldsymbol{\theta}_j^*$ in species tree S_j, $j = 1, 2, 3$. Note that S_1 represents the true model and $\boldsymbol{\theta}_1^*$ are the true parameter values, while S_2 and S_3 are misspecified models and $\boldsymbol{\theta}_2^*$ and $\boldsymbol{\theta}_3^*$ are the *best-fitting or pseudo-true parameter values*. The Kullback-Leibler distance D_{12} from S_2 to S_1 is

$$D_{12} = \int f(x|S_1, \boldsymbol{\theta}_1^*) \log \frac{f(x|S_1, \boldsymbol{\theta}_1^*)}{f(x|S_2, \boldsymbol{\theta}_2^*)} \, dx \quad (10)$$
$$= \mathbb{E}(l_1(\boldsymbol{\theta}_1^*)) - \mathbb{E}(l_2(\boldsymbol{\theta}_2^*)),$$

where $l_j(\boldsymbol{\theta}_j^*) \equiv \log f(x|S_j, \boldsymbol{\theta}_j^*)$, with x to be one data point (or site pattern counts at one locus), and where the integral means summation over all possible data outcomes at a locus. We use the per-locus log-likelihood values to compare the three species trees: $\bar{z}_j \equiv \frac{1}{m} l_j(\hat{\boldsymbol{\theta}}_j)$, $j = 1, 2, 3$. When m is large, these have the means $\mathbb{E}(\bar{z}_j) \approx \mathbb{E}(l_j(\boldsymbol{\theta}_j^*)) \equiv \mu_j$, with $\mu_1 - \mu_2 = D_{12}$, and the variance matrix $\frac{1}{m}\Sigma$, where $\Sigma = \{\sigma_{jk}\}$ and $\sigma_{jk} \equiv \text{Cov}(l_j(\boldsymbol{\theta}_j^*), l_k(\boldsymbol{\theta}_k^*))$. The error of the ML method, $e_{\text{ML}} = \mathbb{P}\{l_1(\hat{\boldsymbol{\theta}}_1) < \max(l_2(\hat{\boldsymbol{\theta}}_2), l_3(\hat{\boldsymbol{\theta}}_3))\}$, is then given by Theorem 1 as

$$e_{\text{ML}} = \mathbb{P}\{\bar{z}_1 < \max(\bar{z}_2, \bar{z}_3)\} \approx \zeta_N(m, D_{12}, \Sigma). \quad (11)$$

Eq. 11 cannot be used to calculate the error rate for ML as D_{12} and σ_{jk} are not easily computable. It predicts a linear relationship between $\Phi^{-1}(e_{\text{ML}})$ and $\sqrt{m}$. This is confirmed by simulation (fig. 2a'-c').

Precise results may be obtained in special cases. In the case of one locus ($m = 1$), the ML gene tree is the ML species tree except for rare datasets: the true species tree S_1 is recovered if $x_{i1} > \max(x_{i2}, x_{i3})$. In rare datasets of extreme divergence, even if $x_{i1} > \max(x_{i2}, x_{i3})$, ties for gene trees are possible, with the star tree being as good

as the binary trees (Yang 2000), while ML under MSC favors S_1. One such dataset is $x_i = (4, 13, 12, 11, 50)$, in which case the three gene trees as well as the star tree achieve the same likelihood, while ML under MSC favors S_1. However, such datasets involve sequences more divergent than random sequences have vanishingly small probability when n is large. Thus we ignore them and consider all methods to be equivalent when $m = 1$. With one locus, it is impossible to identify all parameters in the MSC model: there are four parameters and only three independent site-pattern frequencies ($f_{i0}, f_{i1}, f_{i2} + f_{i3}$ for S_1, for example).

The case of one site per locus ($n = 1$) is analyzed later in the section on ISML. Numerical calculations on a model species tree are presented in table 1. They will be discussed later in comparison with other methods.

In the case of $n \to \infty$, the gene tree (including the coalescent times) at each locus is given without errors. The likelihood is then the product of MSC densities of gene trees across the loci (eq. 2). This likelihood has singularities, with one or more species trees achieving infinite likelihood (Liu *et al.* 2010; Yang 2014). In the case of three species considered here, only one species tree (given by the smallest coalescent time) achieves infinite likelihood and will be the unique species-tree estimate, so that the estimation can proceed despite the singularity (Yang 2014, p.360, Problem 9.4). Let the smallest coalescent/divergence time between species across all loci be t_{ab}, t_{bc} and t_{ca}. If t_{ab} is the smallest among the three, then species tree S_1 achieves infinite likelihood, by collapsing on the coalescent time t_{ab}; that is, $\ell_1(\hat{\boldsymbol{\theta}}_1) \to \infty$ as $\hat{\tau}_0 = \hat{\tau}_1 = t_{ab}$ and $\hat{\theta}_1 \to 0$ (see eq. 2) (Yang 2014, p.338-9), while the other two species trees have finite likelihood.

Given S_1 as the true species tree, both t_{bc} and t_{ca} are $> \tau_{ABC}$ (fig. 1b). If sequences a and b coalesce in population AB at any of the m loci, t_{ab} will be smaller than both t_{bc} and t_{ca}, and S_1 will be the ML species tree. Thus an incorrect species tree is inferred only if a and b do not coalesce in AB at any of the m loci and are not the first to coalesce in the root population ABC. Thus

$$e_{\text{ML},\infty} = \phi^m \times \frac{2}{3}, \quad (12)$$

where $\phi = e^{-\frac{2}{\theta_{AB}}(\tau_{ABC} - \tau_{AB})}$ is the probability that a and b do not coalesce in population AB. This equation is exact and applies to both small and large m (fig. 3b).

Concatenation

Sequence alignments at the m loci are merged into a super-alignment of length nm, and the data are the site-pattern counts pooled across loci: $\boldsymbol{x}. = \{x_{.j}\}$, with $x_{.j} = \sum_i x_{ij}$, $j = 0, 1, \cdots, 4$. The likelihood function is given by the multinomial probability of eq. 4 except that $x_{.j}$ is used instead of x_{ij}. The ML tree is G_1 if $x_{.1} > \max(x_{.2}, x_{.3})$ (Yang 1994b; 2000). We discuss the error

rate of concatenation below in the section on the ISML method.

We also examine biases in parameter estimation using concatenation. We use species tree S_1 with $\tau_{ABC} = 0.02$, $\tau_{AB} = 0.01$, $\theta_{ABC} = 0.02$ and $\theta_{AB} = 0.01$ to simulate $m = 10^4$ loci each with $n = 250$ sites. We obtain MLEs $\hat{t}_0$ and $\hat{t}_1$ on gene tree G_1 from the concatenated data for comparison with the MLEs $\hat{\tau}_0$ and $\hat{\tau}_1$ on species tree S_1 in the MSC model (eq. 5). With so much data, both concatenation and ML recover the true tree with near certainty. The MLEs under the MSC (obtained using the 3S program) are very close to the true values, while concatenation (BASEML in PAML, Yang 2007) produced seriously biased estimates (table 2). Even the relative age, $\hat{t}_0/\hat{t}_1 = 1.92$, differs from $\tau_{ABC}/\tau_{AB} = 2$, which means that molecular clock dating analysis using concatenated data will produce biased time estimates (Angelis and dos Reis 2015; Ogilvie *et al.* 2017; Tiley *et al.* 2020).

ISML

The ISML method assumes that all sites in the super-alignment are i.i.d. Like concatenation, the data are summarized as pooled site-pattern counts, $\boldsymbol{x}. = \{x._0, x._1, x._2, x._3, x._4\}$. However, ISML is coalescent-aware and uses the MSC model to calculate the probabilities for the site patterns. By averaging the conditional site-pattern probabilities of eq. 3 over the MSC density of gene trees and coalescent times of eq. 2, we derive the marginal site-pattern probabilities, $\bar{\boldsymbol{p}} = (\bar{p}_0, \cdots, \bar{p}_4)$, as

$$\bar{p}_0 = \tfrac{1}{16}\left(1 + 18a_0 + 54a_0b + 54a_0c_0 + 9c_1 + 9a_1\right),$$

$$\bar{p}_1 = \tfrac{3}{16}\left(1 - 6a_0 - 18a_0b - 18a_0c_0 + 9c_1 + 9a_1\right),$$

$$\bar{p}_2 = \tfrac{3}{16}\left(1 + 6a_0 - 18a_0b - 18a_0c_0 - 3c_1 - 3a_1\right), \quad (13)$$

$$\bar{p}_3 = \bar{p}_2,$$

$$\bar{p}_4 = \tfrac{6}{16}\left(1 - 6a_0 + 18a_0b + 18a_0c_0 - 3c_1 - 3a_1\right),$$

where $a_0 = \frac{e^{-8\tau_0/3}}{3+4\theta_0}$, $a_1 = \frac{e^{-8\tau_1/3}}{3+4\theta_1}$, $b = \frac{e^{-4\tau_1/3}}{3+2\theta_1}$, $c_0 = 2\phi \cdot$ $(\theta_1 - \theta_0) \cdot \frac{e^{-4\tau_0/3}}{(3+2\theta_0)(3+2\theta_1)}$, and $c_1 = 4\phi \cdot (\theta_1 - \theta_0) \cdot \frac{a_0}{3+4\theta_1}$, with $\phi = e^{-2(\tau_0-\tau_1)/\theta_1}$. Note that $\{\bar{p}_j\}$ are functions of a_0, $b + c_0$ and $a_1 + c_1$, although these do not appear to permit simple biological interpretations. The cases for S_2 and S_3 are given by symmetry.

The likelihood function (or the probability for the pooled site-pattern counts) for each species tree is

$$f(\boldsymbol{x}.|S_1, \boldsymbol{\theta}_1) = \bar{p}_0^{x._0} \bar{p}_1^{x._1} \bar{p}_2^{x._2+x._3} \bar{p}_4^{x._4},$$

$$f(\boldsymbol{x}.|S_2, \boldsymbol{\theta}_2) = \bar{p}_0^{x._0} \bar{p}_1^{x._2} \bar{p}_2^{x._3+x._1} \bar{p}_4^{x._4}, \quad (14)$$

$$f(\boldsymbol{x}.|S_3, \boldsymbol{\theta}_3) = \bar{p}_0^{x._0} \bar{p}_1^{x._3} \bar{p}_2^{x._1+x._2} \bar{p}_4^{x._4}.$$

THEOREM 3. *(a) If the true species tree is S_1 with parameters $\boldsymbol{\theta}_1$, then $\bar{p}_1 > \bar{p}_2 = \bar{p}_3$. (b) ISML infers the species tree S_1 if $x._1 > \max\{x._2, x._3\}$.*

Proof. (a) Each of the marginal site pattern probabilities $\bar{p}_j$, $j = 0, \cdots, 4$, is a sum over the four gene trees of figure 1b: G_{1a}, G_{1b}, G_2 and G_3. The three gene trees G_{1b}, G_2, and G_3 have the same densities (eq. 2). Together their contribution to the site pattern *xxy* is the same as that to the pattern *yxx* or pattern *xyx*. If the gene tree is G_{1a} (with any coalescent times $t_0 > t_1$), site pattern *xxy* will have a higher probability than *yxx* or *xyx*, with $p_1 > p_2 = p_3$. Averaging over all the four gene trees, we have $\bar{p}_1 > \bar{p}_2 = \bar{p}_3$.

(b) We show that if $x._1 > x._2$, then $\ell(S_1, \hat{\boldsymbol{\theta}}_1) > \ell(S_2, \hat{\boldsymbol{\theta}}_2)$, where $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ are the MLEs under each species tree. First note that if $x._1 > x._2$ and $q_1 > q_2 > 0$, then $q_1^{x._1} q_2^{x._2} > q_1^{x._2} q_2^{x._1}$. Let $q_1 = \bar{p}_1(S_1, \hat{\boldsymbol{\theta}}_2)$ and $q_2 = \bar{p}_2(S_1, \hat{\boldsymbol{\theta}}_2)$, and we have $\ell(S_1, \hat{\boldsymbol{\theta}}_2) > \ell(S_2, \hat{\boldsymbol{\theta}}_2)$. In other words, even if we use $\hat{\boldsymbol{\theta}}_2$ (the MLE for S_2) to calculate the likelihood for species tree S_1, tree S_1 will have a higher likelihood than S_2. Since $\hat{\boldsymbol{\theta}}_2$ may not be optimal for S_1, it follows that $\ell(S_1, \hat{\boldsymbol{\theta}}_1) \geq \ell(S_1, \hat{\boldsymbol{\theta}}_2) > \ell(S_2, \hat{\boldsymbol{\theta}}_2)$. $\square$

Theorem 3 means that ISML infers species tree S_j if $x._j$ is the greatest among $x._1$, $x._2$, and $x._3$, just like concatenation.

To study the error rate for ISML (or CONCAT), let p_{ij}, $j = 0, \cdots, 4$ be the site-pattern probabilities at any locus i. Data at each locus are represented by the site-pattern frequencies $f_{ij} = x_{ij}/n$. Let $f_i = \{f_{ij}\}$ be the data at locus i. The f_i are i.i.d. among loci from a common distribution with mean $\mathbb{E}(f_{ij}) = \bar{p}_j$ and variance/covariance $\sigma_{jj} \equiv \mathbb{V}(f_{ij})$ and $\sigma_{jk} \equiv \mathrm{Cov}(f_{ij}, f_{ik})$. Let $\bar{f}_j = \frac{1}{m}\sum_{i=1}^m f_{ij} = x._j/m$ be the means over loci. Here $\{f_{ij}\}$ constitute the full data while $\{\bar{f}_j\}$ are summaries used by ISML: the species tree estimate is S_j if $\bar{f}_j$ is the largest among $(\bar{f}_1, \bar{f}_2, \bar{f}_3)$. Thus $e_{\mathrm{ISML}} = \mathbb{P}(\bar{f}_1 < \max\{\bar{f}_2, \bar{f}_3\}) \approx \zeta_N(m, \bar{p}_1 - \bar{p}_2, \Sigma)$, where $\Sigma = \{\sigma_{jk}\}$. Below we derive the variances.

At $n = 1$, they are given by the multinomial distribution as

$$\sigma_{jj}^{(1)} = \bar{p}_j(1 - \bar{p}_j), \quad \sigma_{jk}^{(1)} = -\bar{p}_j\bar{p}_k, \quad 1 \leq j, k \leq 3. \quad (15)$$

At $n = \infty$, we have $f_{ij} = p_{ij}$, given by eq. 3. The variances, denoted $\sigma_{jk}^{(\infty)}$, can be generated by simulating gene trees with coalescent times and calculating the site-pattern probabilities (eq. 3) (table S1). This distribution is 3-dimensional (for f_{i0}, f_{i1}, and $f_{i2} = f_{i3}$ under S_1), indexed by four parameters ($\boldsymbol{\theta}_1$ in S_1), and is a mixture distribution with 4 components corresponding to the four gene trees of figure 1b. It reflects the coalescent fluctuation in gene genealogies.

For any finite $1 \leq n < \infty$, the variances are given by

$$\sigma_{jj} = \mathbb{V}(\mathbb{E}(f_{ij}|p_{ij})) + \mathbb{E}(\mathbb{V}(f_{ij}|p_{ij}))$$
$$= \mathbb{V}(p_{ij}) + \mathbb{E}(p_{ij}(1 - p_{ij})/n)$$
$$= \mathbb{V}(p_{ij}) + \tfrac{1}{n}[\mathbb{E}(p_{ij})(1 - \mathbb{E}(p_{ij})) - \mathbb{V}(p_{ij})]$$

$$= \tfrac{1}{n}\sigma_{jj}^{(1)} + \tfrac{n-1}{n}\sigma_{jj}^{(\infty)}, \tag{16}$$

$$\sigma_{jk} = \mathrm{Cov}(p_{ij}, p_{ik}) + \mathbb{E}(\mathrm{Cov}(f_{ij}, f_{ik}|p_{ij}, p_{ik}))$$
$$= \mathrm{Cov}(p_{ij}, p_{ik}) + \tfrac{1}{n}[-\mathbb{E}(p_{ij})\mathbb{E}(p_{ik}) - \mathrm{Cov}(p_{ij}, p_{ik})]$$
$$= \tfrac{1}{n}\sigma_{jk}^{(1)} + \tfrac{n-1}{n}\sigma_{jk}^{(\infty)},$$

where $\mathbb{E}(p_{ij}) \equiv \bar{p}_j$ (eq. 13), while $\mathbb{V}(p_{ij}) = \sigma_{jj}^{\infty}$ and $\mathrm{Cov}(p_{ij}, p_{ik}) = \sigma_{jk}^{\infty}$ are the variances/covariances over the coalescent process. These are calculated for a set of parameter values in table S1. The variances of f_{ij} are thus weighted averages of variances at $n = 1$ and ∞.

The approximation $e_{\mathrm{ISML}} \approx \zeta_N(m, \bar{p}_1 - \bar{p}_2, \Sigma)$ is very accurate, with errors < 0.002 in the simulation of table 1. At large n, accommodating correlation is useful as ζ_{N0} which ignores correlation is less accurate (see fig. 4 for the case of $n = \infty$). For example, the correlation $\rho(f_{i1}, f_{i2}) = \frac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}}$ is $-0.124, -0.153$, and -0.181 at $n = 1, 1000$, and ∞, respectively (table S1).

We now consider parameter estimation by ISML. Theorem 3 allows species tree estimation by ISML without knowledge of the MLE of the parameters. With data of $x_{\cdot j}, j = 0, \cdots, 4$, there are only three observations (three free proportions $\bar{f}_0, \bar{f}_1$, and $\bar{f}_2 + \bar{f}_3$ in the case of S_1). As there are four parameters in the MSC model, it is impossible to identify all of them.

If we assume $\theta_0 = \theta_1 = \theta$ (as in Tian and Kubatko 2016), all three parameters (τ_0, τ_1, θ) will be identifiable. As $c_0 = c_1 = 0$, eq. 13 simplifies to

$$\bar{p}_0 = \tfrac{1}{16}(1 + 18a_0 + 54a_0 b + 9a_1),$$
$$\bar{p}_1 = \tfrac{3}{16}(1 - 6a_0 - 18a_0 b + 9a_1),$$
$$\bar{p}_2 = \tfrac{3}{16}(1 + 6a_0 - 18a_0 b - 3a_1) = \bar{p}_3, \tag{17}$$
$$\bar{p}_4 = \tfrac{6}{16}(1 - 6a_0 + 18a_0 b - 3a_1),$$

where a_0, a_1 and b are defined in eq. 13 with $\theta_0 = \theta_1 = \theta$. By equating the observed site-pattern frequencies to their expected probabilities (eq. 17), we have

$$\tfrac{1}{4}(9a_1 + 1) = \bar{f}_0 + \bar{f}_1 \equiv h_1,$$
$$\tfrac{1}{4}(9a_0 + 1) = \bar{f}_0 + \tfrac{1}{2}(\bar{f}_2 + \bar{f}_3) \equiv h_2, \tag{18}$$
$$\tfrac{3}{8}(-18a_0 b + 3a_1 + 1) = \bar{f}_1 + \tfrac{1}{2}(\bar{f}_2 + \bar{f}_3) \equiv h_3.$$

Thus we have a quadratic equation in $\hat{\theta}$:

$$4(4h_3 - 2h_1 - 1)^2 \hat{\theta}^2 + [3(4h_3 - 2h_1 - 1)^2$$
$$- (4h_1 - 1)(4h_2 - 1)^2](4\hat{\theta} + 3) = 0. \tag{19}$$

This always has a unique positive root. Given $\hat{\theta}$, the estimates $\hat{\tau}_0$ and $\hat{\tau}_1$ are given by eq. 18, which are guaranteed to be positive.

Thus under the assumption $\theta_0 = \theta_1$, the ISML method provides estimates of the three parameters in the model: θ, τ_0 and τ_1. As there is a one-to-one correspondence between the parameters and the multinomial proportions, the estimates are consistent and approach the true values

when $m \to \infty$ for any $n \geq 1$ if the assumption of $\theta_0 = \theta_1$ is correct (table 3c&d). However, the pooled site-pattern counts or average site-pattern frequencies are summaries of the original data and are not sufficient statistics. It then follows that the ISML estimates will be less efficient and have larger asymptotic variances than the MLEs obtained from the full data under the same model assumption of $\theta_0 = \theta_1$ (table 3, case c). Furthermore, if $\theta_0 \neq \theta_1$, assuming $\theta_0 = \theta_1$ will lead to biased and inconsistent parameter estimates even if the same species tree estimate is produced. In other words if $\theta_0 \neq \theta_1$, the ISML method assuming $\theta_0 = \theta_1$ will produce a consistent estimate of the species tree and inconsistent estimates of the model parameters (table 3, cases e&f).

Two-step method (majority vote)

In the 2-STEP method we estimate gene trees at individual loci and then use the most common gene tree topology as the species tree estimate. Under JC, the ML gene tree for locus i (which is also the UPGMA tree) is tree G_j if x_{ij} is the largest among x_{i1}, x_{i2}, and x_{i3} (Yang 1994b; 2000); site patterns xxy, yxx and xyx 'support' gene trees G_1, G_2, and G_3, respectively. There is no need for numerical optimization to obtain the ML tree at each locus.

Let g_1, g_2 and g_3 be the probabilities that the estimated gene tree is G_1, G_2 and G_3, respectively; that is, $g_1 = \mathbb{P}\{x_{i1} > \max(x_{i2}, x_{i3})\}$, and so on. These are functions of all four parameters in the MSC model ($\boldsymbol{\theta}_1$) as well as the sequence length n, and can be computed numerically (Yang 2002, eq. 12) or by simulation. Under JC and the clock, $g_2 = g_3 < g_1 < \mathbb{P}(G_1|S_1, \boldsymbol{\theta}_1)$ (Yang 2002). This result has several implications. First, $g_1 < \mathbb{P}(G_1)$ means that phylogenetic errors inflate gene-tree–species-tree discordance and lead to underestimation of the internal branch length in the species tree (Yang 2002). Second $g_1 < \mathbb{P}(G_1)$ also means that use of estimated (rather than true) gene trees leads to reduced probability for recovering the correct species tree. Third, $g_1 > g_2 = g_3$ means that the 2-STEP estimate of the species tree is consistent even if estimated gene trees are used.

Let the number of loci at which G_1 is the ML tree be $m_1 = \sum_{i=1}^{m} \mathbb{I}_{x_{i1} > \max(x_{i2}, x_{i3})}$, where the indicator function $\mathbb{I}_a = 1$ if statement a is true and 0 otherwise. Similarly define m_2 and m_3 to be the counts for the two mismatching gene trees. The correct species tree is inferred if and only if $m_1 > \max(m_2, m_3)$. Thus the error rate can be approximated by $e_{2\text{-STEP}} \approx \zeta(m, g_1, g_2)$ (eq. 8).

The accuracy of this approximation is assessed in table 1 at different values of n with $m = 1000$ and with parameter values $\tau_0 = 0.02$, $\tau_1 = 0.019$, $\theta_0 = 0.01$, and $\theta_1 = 0.05$. Consider first the case of $n = 1$. The gene tree is resolved if the single site at the locus has site patterns 1, 2, or 3, but is unresolved if the site has patterns 0 or 4. Whether we ignore loci with ties (with

site patterns 0 or 4) or break ties evenly (assigning $\frac{1}{3}$ to each gene tree) does not affect the species tree estimate. Thus $g_1(1) = \bar{p}_1$ and $g_2(1) = \bar{p}_2$ (eq.13) and the error is $e_{2\text{-STEP}} \approx \zeta(m, \bar{p}_1, \bar{p}_2)$. This is equivalent to $e_{\text{ISML}} \approx \zeta_N(m, \bar{p}_1 - \bar{p}_2, \Sigma)$ for ISML, consistent with the fact that at $n = 1$ all methods considered here are equivalent.

If $n = \infty$, the estimated gene trees will be the true gene trees so that $g_1 = \mathbb{P}(G_1)$ and $g_2 = \mathbb{P}(G_2)$. The error rate is then $\zeta(m, \mathbb{P}(G_1), \mathbb{P}(G_2)) = \zeta(1000, 0.3594737, 0.3202631) = 0.1132$, close to 0.114 from simulation. At $n = 1000$, the proportions of estimated gene trees are $g_1 = 0.33273$ and $g_2 = 0.30811$, so that $\zeta(m, g_1, g_2) = 0.264$, close to 0.260 by simulation (table 1). These are much larger than 0.114 at $n = \infty$, suggesting that with $n = 1000$ sites in the sequence, the estimated gene trees have substantial errors and uncertainties.

The approximations ζ_{ZLY} and ζ (eq. 8) give nearly identical results. The error rate is found to be very sensitive to the precise values of g_1 and g_2. Overall, the approximation is good, with errors within or close to 1%.

Numerical comparison of different methods

We use simulation to compare the different species-tree estimation methods and to assess the reliability of our approximations. We use a challenging species tree with parameters $\tau_0 = 0.02$, $\tau_1 = 0.019$, $\theta_0 = 0.01$, and $\theta_1 = 0.05$. The error is plotted against the number of loci (m) when the number of sites per locus is fixed at $n = 1, 2$, or 1000 (fig. 2).

In the case of one site per sequence ($n = 1$), all four methods considered in this study are equivalent, with the species tree given by the most frequent pooled site pattern (i.e., the greatest of $x._1$, $x._2$, and $x._3$). With one site, the independent-sites assumption is correct, and ML and ISML are exactly the same. As discussed earlier, concatenation and 2-STEP also select the species tree according to the pooled site patterns. Treatment of ties among $x._1, x._2, x._3$ has very minor effects on the error rate. For $n = 1$ and $m = 1000$, simulation gave the error estimate $e = 0.642$ if ties are broken evenly (table 1) or 0.641 if datasets with ties are ignored. As predicted by our theory, the probit transform of the error, $\Phi^{-1}(e)$, shows a linear relationship with $\sqrt{m}$ (fig. 2a', $R^2 = 0.9994$).

In the case of $n = 2$ sites per locus, ISML (= concatenation), 2-STEP and ML are all distinct. To see that concatenation and 2-STEP may produce different species trees, consider the case of $m = 3$ loci and $n = 2$ sites. If the dataset at the three loci are 11, 02, and 00, where 0–4 represent the five site patterns, concatenation will infer the correct species tree S_1 (as $x._1 = 2, x._2 = 1, x._3 = 0$), while 2-STEP will have a tie between S_1 and S_2 (as $m_1 = 1, m_2 = 1, m_3 = 0$). If the dataset at the three loci are 33, 01, and 14, concatenation will have a tie between S_1 and S_3 (as $x._1 = 2, x._2 = 0, x._3 = 2$) while 2-STEP will infer the correct species tree (as

$m_1 = 2, m_2 = 0, m_3 = 1$). We also confirm that at $n = 2$ ML differs from all three summary methods and can identify and consistently estimate all four parameters in the MSC model. Indeed ML is far more efficient for species tree estimation than the summary methods when $n = 2$ (fig. 2b&b'). While the summary methods improve only slightly when n changes from 1 to 2, there is a major performance boost for ML (fig. 3a). This may be due to the fact that the model is fully identifiable with $n = 2$ but not when $n = 1$. The predicted linear relationship between $\Phi^{-1}(e)$ and $\sqrt{m}$ holds well for the three summary methods (fig. 2b'). For ML, if we remove the first two points (for $m = 10$ and 20), the relationship is nearly linear, with $y = -0.0022x + 0.0391$, with $R^2 = 0.97$.

The most interesting case is with $n \gg 1$, since in real datasets n may be in the range 50–5000, say. We used $n = 1000$ in fig. 2c&c'. As in the case of $n = 2$, there is a large performance divide between ML and the three summary methods (ISML = CONCAT and 2-STEP), while the summary methods have similar performance. The approximate linear relationship between $\Phi^{-1}(e)$ and $\sqrt{m}$ holds well for all methods.

The superior performance of ML persists in the limit of $n = \infty$ (fig. 3b). For example, $e_{\text{ML},\infty} = 0.45$ and 0.01 for ML at $m = 10$ and 100, respectively, compared with $e_{2\text{-STEP},\infty} = 0.60$ and 0.46 for 2-STEP or $e_{\text{ISML},\infty} = 0.62$ and 0.51 for ISML. The differences between ML and 2-STEP reflect the information in the coalescent times or gene-tree branch lengths. The differences between ML and ISML reflect the information in the variation of site-pattern frequencies among loci, as ISML uses only the averages across loci.

Fig. 3c examines the error rates of different methods while $nm = 10^4$ is fixed. At the two ends ($n = 1$ or $m = 1$), all four methods are equivalent, with $e = 0.587$ at $n = 1$ and $m = 10^4$, and $e = 0.646$ at $m = 1$ and $n = 10^4$. Note that when $n = 1$ and $m \to \infty$, the error $e \to 0$, whereas if $m = 1$ and $n \to \infty$, the error $e = 1 - g_1(n) \to 1 - \mathbb{P}(G_1) = 0.6405$. The high error at $m = 1$ even when $n = \infty$ is because a single gene tree (with coalescent times), even if known with certainty, does not contain much information about the MSC process. Away from the two ends ($n > 1$ or $m > 1$), ML is considerably more efficient than the summary methods (fig. 3c). The case of $m = 10^4$ ($n = 1$), at which $e_{\text{ML}} = 0.587$, and the case of $m = 2$ ($n = 5000$), at which $e_{\text{ML}} = 0.487$, make an interesting contrast. In the first case all sites are i.i.d., while in the second there are only two independent genes, each of 5000 sites in complete linkage. One might expect data of independent sites to be more informative than two loci with correlated sites at the same locus (e.g., Long and Kubatko 2018), but the opposite is true. With $n = 1$, not all model parameters are identifiable, and this non-identifiability issue appears to impact species tree

estimation as well (Shi and Yang 2018, p.172). With nm fixed, the smallest error e_{ML} occurs at intermediate values of n and m, around $n = m = 100$, although performance is similar over a large range of n (fig. 3c).

In table 1, we calculated the species-tree error probability using eqs. 6 and 8, as well as two pairs of bounds (ζ_{L1}, ζ_{U1}) and (ζ_{L2}, ζ_{U2}) (Theorem 1, Appendix A), for comparison with the simulation results. The asymptotic results are expected to apply when the sequence length n is fixed while the number of loci $m \to \infty$. Here m is fixed at 1000, so that $b < 2$ for all cases (table 1), and is too small for the asymptotic approximations to be reliable. As a result, eqs. 6 and 8 are more accurate.

Discussion

Errors of species tree estimation by different methods

Under the MSC model, data at different loci are i.i.d., so that the number of loci (m) constitutes the sample size in the statistical model. Thus we have derived approximations to the error rate for different methods when m increases, with the sequence length n fixed. For large m, the error can be approximated by $\Phi(-c\sqrt{m})$, where c is a constant. This is seen to apply to all four methods considered in this study (ML, ISML = concatenation, and 2-STEP) (see table 4 for a summary).

The theory for ML in Appendix B applies generally to ML selection of nonnested models, whether one model (which may and may not be the true model) fits the data better than the others, judged by the K-L divergence to the true data-generating model. In particular, the theory applies to conventional phylogenetic reconstruction without the MSC model. For example, figure 5 applies the same prediction to simulation results on four-taxa trees from Yang (1997). Previously Susko (2011) developed a large-sample approximation to the log-likelihood difference between two trees and to the probability that each tree will be the ML tree in the case of four-species without the molecular clock. It was assumed that the internal branch length in the tree is small and approaches 0 at the rate of $n^{-\frac{1}{2}}$ or faster when the number of sites n increases. In our analysis, we take the conventional approach of fixing the parameters when the data size increases.

We note that in problems of parameter estimation, the standard error for the parameter estimate or the width of the confidence interval typically decreases at the rate of $n^{-\frac{1}{2}}$, so that quadrupling the data size halves the interval. In contrast, the probability of recovering the best-fitting model approaches 1 much faster. As the probit transform of the error decreases linearly with $\sqrt{n}$, it will soon reach a point beyond which the precise error probability is of no practical significance: for example, $\Phi^{-1}(e) = -3$ means $e = 0.0013$ while $\Phi^{-1}(e) = -5$ means $e = 2.9 \times 10^{-7}$. The different dynamics between model selection

and parameter estimation when the data size grows is consistent with the fact that we tend to obtain extreme support for phylogenies inferred in large datasets (Yang and Zhu 2018).

Implications of our study to species tree methods

While the species tree problem studied here is the simplest, it has the complexities of the general problem. Furthermore, we have represented all major species tree methods in our analysis. We expect ML to be asymptotically similar to Bayesian inference as both are full-data methods.

We have assumed the JC mutation model and the molecular clock. Our results are thus applicable to shallow species phylogenies and may not apply to distantly related species for which the JC model may be inadequate for multiple-hit correction and the molecular clock may be seriously violated. In the case of three species examined in this paper, concatenation and ISML always produce the same species tree estimate. However, in more general settings with four or more species and when the clock is violated and unrooted trees are used, concatenation and ISML are known to be different. In particular, concatenation (as well as 2-STEP) can be inconsistent (Roch and Steel 2015), while ISML is a coalescent-aware method and is always consistent.

The ISML method considered here is similar to SVDQUARTETS (Chifman and Kubatko 2014). Both are summary methods based on pooled site-pattern counts. SVDQUARTETS is sometimes described as a site pattern-based method (e.g. Kubatko 2019). This is not a helpful description. Site-pattern counts for different loci ($\{f_{ij}\}$) are sufficient statistics under the model and carry the same amount of information as the sequence alignments at the same loci so that it makes no difference whether site patterns or sequences are used. Indeed virtually all methods involving likelihood calculation on sequences operate on site patterns instead of sites. Instead what matters is whether site patterns are pooled across loci. In the original data, the sites of the same locus share the same gene tree and the variation among loci provides information about parameters of the coalescent process such as the ancestral population sizes. Pooling sites across loci means that such information is lost (Shi and Yang 2018). As a result, the pooled site-pattern counts are unable to identify all parameters of the MSC model even if they can identify the species tree topology. Previously Long and Kubatko (2018) found in simulations that SVDQUARTETS performed better in datasets of 600 coalescent-independent sites ($m = 600, n = 1$ in the notation of this paper) than in data of two genes each of 300 bp ($m = 2, n = 300$), and suggested that this is because "[t]he 600 sites observed from 600 distinct gene trees give independent genealogical information about the species tree, though indirectly, while the 300 sites for each of the two

genes can give a reasonable indication of the individual gene trees, but still provide only two observed gene genealogies." Our analysis suggests that this is not a correct interpretation. When the information in the data is used properly (as in the ML method), there is in fact more information in two genes each of 300bp than in 600 independent sites (fig. 3c).

To understand the issue of parameter unidentifiability and the potential information loss for species tree estimation due to the pooling of sites across loci in SVDQUARTETS, consider the simple random-effects model

$$y_{ij} = \mu + \alpha_i + e_{ij}, \quad i = 1, \cdots, m; j = 1, \cdots, n, \quad (20)$$

where the treatment effect $\alpha_i \sim \mathbb{N}(0, \sigma_a^2)$ and the error $e_{ij} \sim \mathbb{N}(0, \sigma_e^2)$. Parameters in the model include the grand mean μ and the variance components σ_a^2 and σ_e^2. It is obvious that if there are no replications within treatment ($n = 1$) or if the observations (y_{ij}) are pooled across treatments, the between-treatment variation and within-treatment errors will be confounded so that σ_a^2 and σ_e^2 will not be identifiable even though μ still is. In species tree estimation, pooling site patterns across loci (as in ISML and SVDQUARTETS) causes some parameters of the MSC model to become unidentifiable even though the species tree still is. This issue of information loss due to averaging over the whole genome may be even more serious for methods designed for data of single nucleotide polymorphisms (SNPs) (Leaché and Oaks 2017), such as SNAPP (Bryant et al. 2012), because the removal of constant sites in the SNP data causes further loss of information (even if the ascertainment bias is accounted for in the method).

An important difference between ISML and SVDQU-ARTETS is that ISML applies ML to the pooled site-pattern counts while SVDQUARTETS uses a criterion based on linear invariants to avoid the ML optimization (Xu and Yang 2016). Use of a non-ML criterion is expected to lead to further reduction in efficiency, in addition to information loss due to the pooling of sites across loci (Chou et al. 2015; Xu and Yang 2016; Shi and Yang 2018).

The MSC model analyzed in this paper assumes free recombination among loci and no recombination between sites of the same locus. Data for such analysis are typically loosely linked short genomic segments that are far apart from each other so that recombination within a locus is rare while different loci are nearly independent (e.g. Takahata et al. 1995; Burgess and Yang 2008; Lohse et al. 2016). Both assumptions of free recombination among loci and no recombination within locus are expected to be violated in real data analysis, and the impact of within-locus recombination is of particular concern. The ML method considered in this paper assumes no recombination (with $r = 0$) while ISML (and SVDQUARTETS) assumes free

recombination ($r = \infty$). The relative performance of the methods will depend on the true recombination rate: ML may be expected to perform better than ISML if r is close to 0, while ISML may be superior if r is large. At very high recombination rates, it may even be possible for ML (assuming $r = 0$) to be inconsistent since the method is similar to concatenation and merges sites of the same locus with different histories into one sequence. In contrast ISML is consistent for all values of r. Previously Lanier and Knowles (2012) found in a computer simulation that species-tree estimation was robust to moderate levels of within-locus recombination (see also discussions in Xu and Yang 2016; Edwards et al. 2016). It will be interesting to evaluate the relative performance of modern species-tree estimation methods (including ISML and SVDQUARTETS) under realistic recombination rates.

Materials and Methods

Simulation

We use a challenging species tree with parameters $\tau_0 = 0.02$, $\tau_1 = 0.019$, $\theta_0 = 0.01$, and $\theta_1 = 0.05$ (fig. 1a). A C program is written to simulate gene trees and sequence alignments for the case of three species/sequences, under the JC model (Jukes and Cantor 1969) with the clock. To simulate the gene tree and the sequence alignment for each locus, we generate an exponential coalescent waiting time (s_1) with mean $\theta_1/2$. If $s_1 < \tau_1$, the gene tree is G_{1a}, and another exponential waiting time s_0 is generated with mean $\theta_0/2$ to get $t_0 = \tau_0 + s_0$ and $t_1 = s_1$. If $s_1 > \tau_1$, the gene tree is one of G_{1b}, G_2, G_3, chosen at random, and two coalescent waiting times (s_1 and s_0) are generated with means $\theta_0/6$ and $\theta_0/2$, respectively, so that $t_1 = \tau_0 + s_1$ and $t_0 = t_1 + s_0$ (fig. 1b). The gene tree and node ages (t_0, t_1) are then used to calculate the site-pattern probabilities for the locus (eq. 3), and the site-pattern counts are generated from multinomial sampling (eq. 4). Each dataset consists of m loci with the sequence length of n sites. We use a large number of replicates (typically $R = 10^6$ or 10^8) so that sampling errors due to a limited number of replicates is not a concern. Species tree estimation by concatenation (= ISML) and 2-STEP is done by counting site patterns.

For the ML method (eq. 5), we used the simulation program MCCOAL, which is part of the BPP program (Yang 2015), to simulate the gene trees and sequence alignments. The data are then analyzed using the ML program 3S (Yang 2002; Dalquen et al. 2017). The JC model is used to simulate and analyze data.

Acknowledgments

References

Angelis, K. and dos Reis, M. 2015. The impact of ancestral population size and incomplete lineage sorting on Bayesian estimation of species divergence times. *Curr. Zool.*, 61: 874–885.

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. A., and RoyChoudhury, A. 2012. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Mol. Biol. Evol.*, 29: 1917–1932.

Burgess, R. and Yang, Z. 2008. Estimation of hominoid ancestral population sizes under Bayesian coalescent models incorporating mutation rate variation and sequencing errors. *Mol. Biol. Evol.*, 25(9): 1979–1994.

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

Chou, J., Gupta, A., Yaduvanshi, S., Davidson, R., Nute, M., Mirarab, S., and Warnow, T. 2015. A comparative study of SVDquartets and other coalescent-based species tree estimation methods. *BMC Genomics*, 16: S2.

Dalquen, D., Zhu, T., and Yang, Z. 2017. Maximum likelihood implementation of an isolation-with-migration model for three species. *Syst. Biol.*, 66: 379–398.

Dawid, A. 2011. Posterior model probabilities. In P. S. Bandyopadhyay and M. Forster, editors, *Philosophy of Statistics*, pages 607–630. Elsevier, New York.

Degnan, J. H. and Rosenberg, N. A. 2006. Discordance of species trees with their most likely gene trees. *PLoS Genet.*, 2: e68.

Degnan, J. H. and Salter, L. A. 2005. Gene tree distributions under the coalescent process. *Evolution*, 59: 24–37.

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

Edwards, S. V., Xi, Z., Janke, A., Faircloth, B. C., McCormack, J. E., Glenn, T. C., Zhong, B., Wu, S., Lemmon, E. M., Lemmon, A. R., Leaché, A. D., Liu, L., and Davis, C. C. 2016. Implementing and testing the multispecies coalescent model a valuable paradigm for phylogenomics. *Mol. Phylogenet. Evol.*, 94(Pt A): 447–462.

Fleiss, J. L., Levin, B., and Palk, M. C. 2003. *Statistical Methods for Rates and Proportions*. John Wiley and Sons, New York, 3 edition.

Heled, J. and Drummond, A. J. 2010. Bayesian inference of species trees from multilocus data. *Mol. Biol. Evol.*, 27: 570–580.

Hudson, R. 1983. Testing the constant-rate neutral alele model with protein sequence data. *Evolution*, 37: 203–217.

Jukes, T. and Cantor, C. 1969. Evolution of protein molecules. In H. Munro, editor, *Mammalian Protein Metabolism*, pages 21–123. Academic Press, New York.

Kubatko, L. 2019. The multispecies coalescent. In D. Balding, I. Moltke, and J. Marioni, editors, *Handbook of Statistical Genomics*, pages 219–245. Wiley, New York, 4th edition.

Lanier, H. C. and Knowles, L. L. 2012. Is recombination a problem for species-tree analyses? *Syst. Biol.*, 61(4): 691–701.

Leaché, A. D. and Oaks, J. 2017. The utility of single nucleotide polymorphism (SNP) data in phylogenetics. *Annu. Rev. Ecol. Evol. Syst.*, 48(1): 69–84.

Leaché, A. D. and Rannala, B. 2011. The accuracy of species tree estimation under simulation: a comparison of methods. *Syst. Biol.*, 60(2): 126–137.

Liu, L. and Pearl, D. K. 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Syst. Biol.*, 56(3): 504–514.

Liu, L., Yu, L., Pearl, D. K., and Edwards, S. V. 2009. Estimating species phylogenies using coalescence times among sequences. *Syst. Biol.*, 58: 468–477.

Liu, L., Yu, L., and Edwards, S. V. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evol. Biol.*, 10: 302.

Lohse, K., Chmelik, M., Martin, S. H., and Barton, N. H. 2016. Efficient strategies for calculating blockwise likelihoods under the coalescent. *Genetics*, 202(2): 775–786.

Long, C. and Kubatko, L. 2018. The effect of gene flow on coalescent-based species-tree inference. *Syst. Biol.*, 67(5): 770–785.

Maddison, W. 1997. Gene trees in species trees. *Syst. Biol.*, 46: 523–536.

Mirarab, S., Reaz, R., Bayzid, M. S., Zimmermann, T., Swenson, M. S., and Warnow, T. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics*, 30(17): i541–548.

Nichols, R. 2001. Gene trees and species trees are not the same. *Trends Ecol. Evol.*, 16: 358–364.

Ogilvie, H. A., Bouckaert, R. R., and Drummond, A. J. 2017. StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. *Mol. Biol. Evol.*, 34: 2101–2114.

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

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

Rannala, B. and Yang, Z. 2017. Efficient Bayesian species tree inference under the multispecies coalescent. *Syst. Biol.*, 66: 823–842.

Rannala, B., Edwards, S., Leaché, A. D., and Yang, Z. 2020. The multispecies coalescent model and species tree inference. In C. Scornavacca, F. Delsuc, and N. Galtier, editors, *Phylogenetics in the Genomic Era*, book section 3.3, page 3.3:1–20. No Commercial Publisher.

Roch, S. and 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.

Shi, C. and Yang, Z. 2018. Coalescent-based analyses of genomic sequence data provide a robust resolution of phylogenetic relationships among major groups of gibbons. *Mol. Biol. Evol.*, 35: 159–179.

Susko, E. 2011. Large sample approximations of probabilities of correct evolutionary tree estimation and biases of maximum likelihood estimation. *Stat. Appl. Genet. Mol. Biol.*, 10.

Szollosi, G. J., Tannier, E., Daubin, V., and Boussau, B. 2015. The inference of gene trees with species trees. *Syst. Biol.*, 64(1): e42–62.

Takahata, N., Satta, Y., and Klein, J. 1995. Divergence time and population size in the lineage leading to modern humans. *Theor. Popul. Biol.*, 48: 198–221.

Tian, Y. and Kubatko, L. S. 2016. Distribution of coalescent histories under the coalescent model with gene flow. *Mol. Phylogenet. Evol.*, 105: 177–192.

Tiley, G. P., Poelstra, J. P., dos Reis, M., Yang, Z., and Yoder, A. D. 2020. Molecular clocks without rocks: new solutions for old problems. *Trends Genet.*, 36(11): 845–856.

White, H. 1982. Maximum likelihood estimation of misspecified models. *Econometrica*, 50: 1–25.

Wu, Y. 2012. Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. *Evolution*, 66: 763–775.

Xu, B. and Yang, Z. 2016. Challenges in species tree estimation under the multispecies coalescent model. *Genetics*, 204: 1353–1368.

Yang, Z. 1994a. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.*, 39: 306–314.

Yang, Z. 1994b. Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. *Syst. Biol.*, 43: 329–342.

Yang, Z. 1996. Phylogenetic analysis using parsimony and likelihood methods. *J. Mol. Evol.*, 42: 294–307.

Yang, Z. 1997. How often do wrong models produce better phylogenies? *Mol. Biol. Evol.*, 14: 105–108.

Yang, Z. 2000. Complexity of the simplest phylogenetic estimation problem. *Proc. R. Soc. B: Biol. Sci.*, 267: 109–116.

Yang, Z. 2002. Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. *Genetics*, 162(4): 1811–1823.

Yang, Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.*, 24: 1586–1591.

Yang, Z. 2014. *Molecular Evolution: A Statistical Approach*. Oxford University Press, Oxford, England.

Yang, Z. 2015. The BPP program for species tree estimation and species delimitation. *Curr. Zool.*, 61(5): 854–865.

Yang, Z. and Rannala, B. 2014. Unguided species delimitation using DNA sequence data from multiple loci. *Mol. Biol. Evol.*, 31(12): 3125–3135.

Yang, Z. and Rodríguez, C. E. 2013. Searching for efficient markov chain monte carlo proposal kernels. *Proc. Natl. Acad. Sci. U.S.A.*, 110(48): 19307–19312.

Yang, Z. and Zhu, T. 2018. Bayesian selection of misspecified models is overconfident and may cause spurious posterior probabilities for phylogenetic trees. *Proc. Natl. Acad. Sci. U.S.A.*, 115(8): 1854–1859.

Zharkikh, A. and Li, W.-H. 1992. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. i. four taxa with a molecular clock. *Mol. Biol. Evol.*, 9: 1119–1147.

Zhu, T. and Yang, Z. 2012. Maximum likelihood implementation of an isolation-with-migration model with three species for testing speciation with gene flow. *Mol. Biol. Evol.*, 29: 3131–3142.

**

Appendix A. Proof of Theorem 1

(a) Define the random variable

$$y = \bar{z}_1 - \max(\bar{z}_2, \bar{z}_3) = \bar{z}_1 - \tfrac{1}{2}(\bar{z}_2 + \bar{z}_3) - \tfrac{1}{2}|\bar{z}_2 - \bar{z}_3|$$
$$= y_1 - |y_2|, \tag{A1}$$

where $y_1 = \bar{z}_1 - \tfrac{1}{2}(\bar{z}_2 + \bar{z}_3) \sim \mathbb{N}(\Delta\mu, s_1^2/m)$ and $y_2 = \tfrac{1}{2}(\bar{z}_2 - \bar{z}_3) \sim \mathbb{N}(0, s_2^2/m)$, with

$$\tfrac{1}{m}s_1^2 = \mathbb{V}(\bar{z}_1) + \tfrac{1}{4}\mathbb{V}(\bar{z}_2 + \bar{z}_3) - 2\mathrm{Cov}(\bar{z}_1, \bar{z}_2)$$
$$= \tfrac{1}{m}\big[\sigma_1^2 + \tfrac{1}{2}(\sigma_2^2 + \sigma_{23}) - 2\sigma_{12}\big], \tag{A2}$$
$$\tfrac{1}{m}s_2^2 = \tfrac{1}{4}\mathbb{V}(\bar{z}_2 - \bar{z}_3)^2 = \tfrac{1}{2m}(\sigma_2^2 - \sigma_{23}).$$

Here we treat $\bar{z}_1, \bar{z}_2$ and $\bar{z}_3$ as normal variables, according to the central limit theorem as $m \to \infty$. As $\mathrm{Cov}(y_1, y_2) = 0$ and both y_1 are y_2 are normal variables, they are independent. Then

$$\zeta = \mathbb{P}\{y_1 < |y_2|\}$$
$$= \mathbb{P}\{y_2 < 0, y_1 < -y_2\} + \mathbb{P}\{y_2 > 0, y_1 < y_2\}$$
$$= 2\mathbb{P}\{y_2 > 0, y_1 < y_2\}$$
$$= 2\int_0^\infty \phi(y_2; 0, \tfrac{1}{m}s_2^2)\Phi\big(\tfrac{y_2 - \Delta\mu}{s_1/\sqrt{m}}\big)\,\mathrm{d}y_2 \tag{A3}$$
$$= 2\int_0^\infty \phi(t)\Phi(at - b)\,\mathrm{d}t,$$

where $a = s_2/s_1$, $b = \Delta\mu\sqrt{m}/s_1$, and $\phi(x; \mu, \sigma^2)$ is the probability density function (PDF) for $\mathbb{N}(\mu, \sigma^2)$ while $\phi(x)$ is the PDF for $\mathbb{N}(0, 1)$. The last integral has been studied by Yang and Rodríguez (2013, SI) in a different context and can be written as

$$\zeta = \frac{1}{\pi}\int_{-\frac{\pi}{2}}^{\tan^{-1}a} \exp\left\{-\frac{b^2}{2(\sin\theta - a\cos\theta)^2}\right\}\,\mathrm{d}\theta, \tag{A4}$$

or, by letting $t = a - \tan\theta$, with $\mathrm{d}\theta = -\frac{1}{t^2+1}\mathrm{d}t$, as

$$\zeta = \frac{1}{\pi}\int_0^\infty \frac{1}{(t-a)^2+1}\,\mathrm{e}^{-\frac{b^2[(t-a)^2+1]}{2t^2}}\,\mathrm{d}t. \tag{A5}$$

Eqs. A4 and A5 can be calculated using Gaussian quadrature and match direct calculations using the CDF for the bivariate normal distribution for $(\bar{z}_1 - \bar{z}_2, \bar{z}_1 - \bar{z}_3)$. When $\Delta\mu = 0$, we have $b = 0$ and

$$\zeta = 2\int_0^\infty \phi(t)\Phi(at)\,\mathrm{d}t = \tfrac{1}{2} + \tfrac{1}{\pi}\tan^{-1}a. \tag{A6}$$

In the symmetrical case of $\Delta\mu = 0$, $\sigma_1^2 = \sigma_2^2$, and $\sigma_{12} = \sigma_{23}$ (with $a = \tfrac{1}{\sqrt{3}}$, $b = 0$), this gives $\tfrac{1}{2} + \tfrac{1}{\pi}\tan^{-1}\big(\tfrac{1}{\sqrt{3}}\big) = \tfrac{2}{3}$, as expected. In this case the three variables $\bar{z}_1, \bar{z}_2$ and $\bar{z}_3$ have the same probability of being the greatest so that the error is $\tfrac{2}{3}$.

To avoid numerical integration, we note that $y_2 \sim \mathbb{N}\big(0, \tfrac{1}{2m}(\sigma_2^2 - \sigma_{23})\big)$, and $|y_2|$ is a folded normal variable

with mean and variance

$$\mathbb{E}(|y_2|) = \sqrt{\tfrac{1}{m\pi}(\sigma_2^2 - \sigma_{23})}, \tag{A7}$$
$$\mathbb{V}(|y_2|) = \big(\tfrac{1}{2m} - \tfrac{1}{m\pi}\big)(\sigma_2^2 - \sigma_{23}).$$

Thus

$$\mathbb{E}(y) = \Delta\mu - \sqrt{\tfrac{1}{m\pi}(\sigma_2^2 - \sigma_{23})}.$$
$$\mathbb{V}(y) = \mathbb{V}(\bar{z}_1) + \tfrac{1}{4}\mathbb{V}(\bar{z}_2 + \bar{z}_3) + \tfrac{1}{4}\mathbb{V}(|\bar{z}_2 - \bar{z}_3|)$$
$$\quad - \mathrm{Cov}(\bar{z}_1, \bar{z}_2 + \bar{z}_3) - \mathrm{Cov}(\bar{z}_1, |\bar{z}_2 - \bar{z}_3|)$$
$$\quad + \tfrac{1}{2}\mathrm{Cov}(\bar{z}_2 + \bar{z}_3, |\bar{z}_2 - \bar{z}_3|) \tag{A8}$$
$$= \mathbb{V}(\bar{z}_1) + \tfrac{1}{4}\mathbb{V}(\bar{z}_2 + \bar{z}_3) + \tfrac{1}{4}\mathbb{V}(|\bar{z}_2 - \bar{z}_3|)$$
$$\quad - 2\mathrm{Cov}(\bar{z}_1, \bar{z}_2) - \mathrm{Cov}(\bar{z}_1, |\bar{z}_2 - \bar{z}_3|)$$
$$\quad + \mathrm{Cov}(\bar{z}_2, |\bar{z}_2 - \bar{z}_3|).$$

We have

$$\mathbb{V}(\bar{z}_2 + \bar{z}_3) + \mathbb{V}(|\bar{z}_2 - \bar{z}_3|) = \mathbb{E}(\bar{z}_2 + \bar{z}_3)^2 + \mathbb{E}(\bar{z}_2 - \bar{z}_3)^2$$
$$\quad - \mathbb{E}^2(\bar{z}_2 + \bar{z}_3) - \mathbb{E}^2(|\bar{z}_2 - \bar{z}_3|)$$
$$= 4\mathbb{E}(\bar{z}_2^2) - 4\mu_2^2 - \tfrac{4}{m\pi}(\sigma_2^2 - \sigma_{23}),$$
$$= \tfrac{4}{m}\sigma_2^2 - \tfrac{4}{m\pi}(\sigma_2^2 - \sigma_{23}),$$
$$\mathrm{Cov}(\bar{z}_1, |\bar{z}_2 - \bar{z}_3|) = 0, \tag{A9}$$
$$\mathrm{Cov}(\bar{z}_2, |\bar{z}_2 - \bar{z}_3|) = 0.$$

Collecting all terms in eq. A8, we get

$$\mathbb{V}(y) = \tfrac{1}{m}\big[\sigma_1^2 - 2\sigma_{12} + \sigma_2^2 - \tfrac{1}{\pi}(\sigma_2^2 - \sigma_{23})\big] \tag{A10}$$

If we assume that y is approximately normally distributed, as in Zharkikh and Li (1992) and Yang (1996), then eq. 6 follows. Note that eq. 6 can also be written as $\zeta_N = \Phi\left(\frac{-b + a\sqrt{2/\pi}}{\sqrt{1 + a^2(1 - \frac{2}{\pi})}}\right)$. Because $|y_2|$ has a folded normal distribution and is not a normal variable, the error of approximation of eq. 6 does not approach zero when $m \to \infty$. For instance, in the symmetrical case ($\Delta\mu = 0$, $\sigma_1^2 = \sigma_2^2$, and $\sigma_{12} = \sigma_{23}$), eq. 6 gives $\Phi\big(\frac{1}{\sqrt{2\pi}-1}\big) = 0.66824$, not $\tfrac{2}{3}$. This level of accuracy is acceptable for our calculations for finite m in this paper, as the precise value of the error is unimportant if the error is nearly zero, but eq. 6 may not give the correct asymptotic error rate when $m \to \infty$ (fig. 6, $a = 10$).

(b) To study the asymptotic behavior of the error probability ζ when $m \to \infty$, we derive bounds on ζ. From eq. A3,

$$\zeta = 2\int_0^\infty \phi(t)\int_{-\infty}^{at-b}\phi(x)\,\mathrm{d}x\,\mathrm{d}t$$
$$= 2\int_{-\infty}^\infty \int_{-\infty}^{at-b}\phi(t)\phi(x)\,\mathrm{d}x\,\mathrm{d}t - 2\int_{-\infty}^0 \int_{-\infty}^{at-b}\phi(t)\phi(x)\,\mathrm{d}x\,\mathrm{d}t$$
$$= 2S - 2A, \tag{A11}$$

where the first integral is $S = \Phi(-h)$, with $h = \frac{b}{\sqrt{1+a^2}} = \frac{\Delta\mu\sqrt{m}}{\sqrt{\sigma_1^2 - 2\sigma_{12} + \sigma_2^2}}$ to be the distance from the origin $(0,0)$ to the line $x = at - b$ (fig. 7), and the second integral is

$$2A = 2\int_{-\infty}^{0}\int_{-\infty}^{at-b} \phi(t)\phi(x)\,dx\,dt$$
$$= \int_{-\infty}^{-b} \phi(x)\int_{(x+b)/a}^{-(x+b)/a} \phi(t)\,dt\,dx. \tag{A12}$$

By considering the area of integration (fig. 7), it is obvious that

$$0 < 2A \le \Phi(-b)\left[1 - \tfrac{2}{\pi}\tan^{-1}a\right], \tag{A13}$$

where the equality holds when $b = 0$. Let

$$\zeta_{L2} = 2\Phi(-h) - \Phi(-b)\left[1 - \tfrac{2}{\pi}\tan^{-1}a\right].$$

As $\Phi(-b) > \Phi(-h)$, we have

$$\zeta_{L2} > \Phi(-h)\left[1 + \tfrac{2}{\pi}\tan^{-1}a\right] \equiv \zeta_{L1}, \tag{A14}$$

or

$$\Phi(-h) \le \zeta_{L1} \le \zeta_{L2} \le \zeta < 2\Phi(-h) \equiv \zeta_{U1}, \tag{A15}$$

as in eq. 7. The equality in the lower bounds is achieved at $b = 0$. Note that the bounds apply to all $a > 0$ and $b > 0$. We use the bounds (ζ_{L1}, ζ_{U1}) in Theorem 1 and in the calculation of table 1. The width of the interval is $\Phi(-h)\left[1 - \tfrac{2}{\pi}\tan^{-1}a\right] \le \Phi(-h) \le \zeta$, so that using any value inside the interval as the estimate will give an error of approximation that is smaller than the error probability ζ.

Note that the bounds $\Phi(-h) < \zeta < 2\Phi(-h)$ are also given by the definition $\zeta = \mathbb{P}\{\bar{z}_1 < \bar{z}_2 \cup \bar{z}_1 < \bar{z}_3\}$, since

$$\mathbb{P}(\bar{z}_1 < \bar{z}_2) < \zeta < \mathbb{P}(\bar{z}_1 < \bar{z}_2) + \mathbb{P}(\bar{z}_1 < \bar{z}_3) = 2\mathbb{P}(\bar{z}_1 < \bar{z}_2), \tag{A16}$$

with $\mathbb{P}(\bar{z}_1 < \bar{z}_2) = \Phi(-h)$.

Next we consider the upper bound in eq. A15 when h or b is large. Note that

$$\Phi(-h) = \int_{h}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\,dy$$
$$= \int_{0}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x+h)^2}\,dx$$
$$= \frac{1}{\sqrt{2\pi}} e^{-h^2/2} \int_{0}^{\infty} e^{-(hx+\frac{1}{2}x^2)}\,dx \tag{A17}$$
$$= \frac{1}{h\sqrt{2\pi}} e^{-h^2/2} \int_{0}^{\infty} e^{-t}\, e^{-\frac{1}{2h^2}t^2}\,dt$$
$$= \frac{1}{h\sqrt{2\pi}} e^{-h^2/2} B,$$

where $B = \int_{0}^{\infty} e^{-t} e^{-\frac{1}{2h^2}t^2}\,dt < \int_{0}^{\infty} e^{-t}\,dt = 1$. For large h,

$$B > \int_{0}^{\sqrt{h}} e^{-t} e^{-\frac{1}{2h^2}t^2}\,dt > e^{-\frac{1}{2h}} \int_{0}^{\sqrt{h}} e^{-t}\,dt \tag{A18}$$
$$= e^{-\frac{1}{2h}}\left(1 - e^{-\sqrt{h}}\right) = 1 - \tfrac{1}{2h} + o\left(\tfrac{1}{h}\right).$$

Thus for large h, $\Phi(-h)$ is bounded by

$$\left(1 - \tfrac{1}{2h} + o\left(\tfrac{1}{h}\right)\right)\frac{1}{h\sqrt{2\pi}} e^{-h^2/2} < \Phi(-h) < \frac{1}{h\sqrt{2\pi}} e^{-h^2/2}, \tag{A19}$$

or

$$\Phi(-h) = \frac{1}{h\sqrt{2\pi}} e^{-h^2/2} + O\left(\tfrac{1}{h^2} e^{-h^2/2}\right). \tag{A20}$$

Let $\varepsilon > 0$ such that $\Phi(-(h+\varepsilon)) = \alpha\Phi(-h)$ for $0 < \alpha < 1$; in other words, ε is the offset at the probit level to reduce the probability by a fraction. From eq. A20,

$$\frac{1}{(h+\varepsilon)\sqrt{2\pi}} e^{-\frac{1}{2}(h^2 + 2\varepsilon h + \varepsilon^2)} + O\left(\frac{1}{(h+\varepsilon)^2} e^{-\frac{1}{2}(h+\varepsilon)^2}\right)$$
$$= \frac{\alpha}{h\sqrt{2\pi}} e^{-\frac{1}{2}h^2} + O\left(\tfrac{1}{h^2} e^{-\frac{1}{2}h^2}\right). \tag{A21}$$

Thus

$$\frac{1}{h+\varepsilon} e^{-\frac{1}{2}(h^2 + 2\varepsilon h + \varepsilon^2)} = \frac{\alpha}{h} e^{-\frac{1}{2}h^2} + O\left(\tfrac{1}{h^2} e^{-\frac{1}{2}h^2}\right), \tag{A22}$$

which gives $\varepsilon = -\tfrac{1}{h}\log\alpha + o\left(\tfrac{1}{h}\right)$ or

$$\Phi\left(-h + \tfrac{1}{h}\log\alpha + o\left(\tfrac{1}{h}\right)\right) = \alpha\Phi(-h). \tag{A23}$$

In particular, for $\alpha = \tfrac{1}{2}$, we have

$$\Phi\left(-\left(h + \tfrac{1}{h}\log 2 + o\left(\tfrac{1}{h}\right)\right)\right) = \tfrac{1}{2}\Phi(-h). \tag{A24}$$

Thus for large h, we have

$$2\Phi(-h) = \Phi\left(-h + \tfrac{1}{h}\log 2 + o\left(\tfrac{1}{h}\right)\right), \tag{A25}$$

as in eq. 7. It may be noteworthy that for large h, a very small change at the probit level, of about $\tfrac{1}{h}\log 2$, changes the probability by a factor of 2.

A tighter lower bound for $2A$ than zero of eq. A13 is

$$2A > \frac{\varphi}{\pi}\exp\left\{-\frac{b^2(a^2k^2 + 1)}{2a^2(k-1)^2}\right\}, \tag{A26}$$

where $\varphi = \tan^{-1}\frac{1}{ka}$ with $k > 1$ (fig. 7). Thus we have a tighter pair of bounds on ζ,

$$2\Phi(-h) - \Phi(-b)\left[1 - \tfrac{2}{\pi}\tan^{-1}a\right] \le \zeta <$$
$$2\Phi(-h) - \tfrac{1}{\pi}\tan^{-1}\tfrac{1}{ka}\exp\left\{-\frac{b^2(a^2k^2 + 1)}{2a^2(k-1)^2}\right\}, \tag{A27}$$

where $k > 1$. We write this pair of bounds as $\zeta_{L2} < \zeta < \zeta_{U2}$. We have $\Phi(-b) \le \Phi(-h) \le \zeta_{L1} \le \zeta_{L2} \le \zeta < \zeta_{U2} < \zeta_{U1} = 2\Phi(-h)$. These bounds, as well as the exact value and eq. 6, are plotted against b in fig. 7 for $a = 0.01, 0.1, 1$ and 10.

Appendix B. The asymptotics of ML species tree estimation

The proof below borrows heavily from White (1982), Dawid (2011) and Yang and Zhu (2018). Let $S_j, j = 1, 2, 3$ be the three species trees with parameters $\boldsymbol{\theta}_j$. Note that S_1 is the true model while S_2 and S_3 are mis-specified models. Let the data at m loci be $\boldsymbol{x} = \{\boldsymbol{x}_i\}, i = 1, \cdots, m$. The log-likelihood function is $\ell_j(\boldsymbol{\theta}_j) = \log f(\boldsymbol{x}|S_j, \boldsymbol{\theta}_j)$. We also define $l_j(\boldsymbol{\theta}_j) = \log f(x|S_j, \boldsymbol{\theta}_j)$ for one data point (that is, site-pattern counts at any single locus), $x \equiv (x_{i0}, x_{i1}, x_{i2}, x_{i3}, x_{i4})$. When the number of loci $m \to \infty$, the MLE $\hat{\boldsymbol{\theta}}_j \to \boldsymbol{\theta}_j^*$. We assume that both $\hat{\boldsymbol{\theta}}_j$ and $\boldsymbol{\theta}_j^*$ are inner points in the parameter space. Whether $\hat{\boldsymbol{\theta}}_j$ is inside the parameter space or at its boundary should not affect the asymptotic rate of convergence. Here $\boldsymbol{\theta}_1^*$ for the true species tree S_1 is the true parameter value, while $\boldsymbol{\theta}_2^*$ for S_2 (as well as $\boldsymbol{\theta}_3^*$ for S_3) is the *pseudo-true parameter value*, which minimizes the Kullback-Leibler distance from the misspecified model S_2 to the true model S_1.

$$D_{12} = \int f(x|S_1, \boldsymbol{\theta}_1^*) \log \frac{f(x|S_1, \boldsymbol{\theta}_1^*)}{f(x|S_2, \boldsymbol{\theta}_2^*)} \, \mathrm{d}x$$
$$= \mathbb{E}\{l_1(\boldsymbol{\theta}_1^*) - l_2(\boldsymbol{\theta}_2^*)\}, \quad (A28)$$

where the expectation is over the true distribution $f(x|S_1, \boldsymbol{\theta}_1^*)$. D_{13} is defined similarly, with $D_{13} = D_{12}$.

We consider the log-likelihood ratio, $\ell_j(\hat{\boldsymbol{\theta}}) - \ell_j(\boldsymbol{\theta}^*)$, given the data ($\boldsymbol{x}$) for any of the species tree j. We drop the subscript j for clarity. As in White (1982) and Dawid (2011), we define two matrices:

$$I(\boldsymbol{\theta}) = \mathbb{E}\{\nabla \log f(x|\boldsymbol{\theta}) \cdot \nabla \log f(x|\boldsymbol{\theta})^T\}$$
$$= \mathbb{E}\{l'(\boldsymbol{\theta})(l'(\boldsymbol{\theta}))^T\}, \quad (A29)$$

$$J(\boldsymbol{\theta}) = \mathbb{E}\{-\nabla^2 \log f(x|\boldsymbol{\theta})\} = \mathbb{E}\{-l''(\boldsymbol{\theta})\},$$

where the superscript T stands for transpose and where the expectation is over the true distribution, and ∇ and ∇^2 are the first and second derivatives with respect to $\boldsymbol{\theta}$.

Apply Taylor expansion to the log likelihood around the MLE $\hat{\boldsymbol{\theta}}$:

$$\ell(\boldsymbol{\theta}) \approx \ell(\hat{\boldsymbol{\theta}}) + \ell'(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + \tfrac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \ell''(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}), \quad (A30)$$

where both the gradient and the Hessian are evaluated at the MLE ($\hat{\boldsymbol{\theta}}$), with $\ell'(\hat{\boldsymbol{\theta}}) = 0$. Setting $\boldsymbol{\theta} = \boldsymbol{\theta}^*$, we have

$$\ell(\hat{\boldsymbol{\theta}}) \approx \ell(\boldsymbol{\theta}^*) + \tfrac{1}{2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^T(-\ell''(\hat{\boldsymbol{\theta}}))(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*). \quad (A31)$$

Apply Taylor expansion to the derivative $\ell'(\boldsymbol{\theta})$ around the MLE $\hat{\boldsymbol{\theta}}$ and let $\boldsymbol{\theta} = \boldsymbol{\theta}^*$, and we have

$$\ell'(\boldsymbol{\theta}) \approx \ell''(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}), \quad (A32)$$

and

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \approx -\ell''(\hat{\boldsymbol{\theta}})^{-1}\ell'(\boldsymbol{\theta}^*), \quad (A33)$$

Each of $\ell'(\hat{\boldsymbol{\theta}})$ and $\ell''(\hat{\boldsymbol{\theta}})$ is a sum of m i.i.d. elements. When $m \to \infty$, $-\ell''(\hat{\boldsymbol{\theta}}) \approx m\mathbb{E}\{-l''(\boldsymbol{\theta}^*)\} = mJ^*$, with

$J^* = J(\boldsymbol{\theta}^*)$ (eq. A29). Furthermore,

$$\mathbb{E}\{\ell'(\boldsymbol{\theta}^*)\} = 0,$$
$$\mathbb{V}\{\ell'(\boldsymbol{\theta}^*)\} = m\mathbb{V}\{l'(\boldsymbol{\theta}^*)\} = mI^*, \quad (A34)$$

where $I^* = I(\boldsymbol{\theta}^*)$ (eq. A29). Thus

$$\sqrt{m}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \xrightarrow{P} \mathbb{N}\left(0, (J^{*-1})^T I^*(J^{*-1})\right). \quad (A35)$$

Thus $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^* + O_p(m^{-1/2})$. Eq. A31 becomes

$$\ell(\hat{\boldsymbol{\theta}}) \approx \ell(\boldsymbol{\theta}^*) + \tfrac{1}{2}\{\sqrt{m}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\}^T J^*\{\sqrt{m}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\}$$
$$= \ell(\boldsymbol{\theta}^*) + O_p(1). \quad (A36)$$

Eqs. A29–A36 apply to all three species trees. In the case of S_1 (the true model), $J^* = I^*$, the Fisher information matrix, and $\ell(\hat{\boldsymbol{\theta}}) - \ell(\boldsymbol{\theta}^*) \sim \tfrac{1}{2}\chi_d^2$. For S_2 or S_3, $\ell(\hat{\boldsymbol{\theta}}) - \ell(\boldsymbol{\theta}^*)$ is a quadratic form of normal variates and is a mixture of noncentral χ^2 variables with mean $\tfrac{1}{2}\mathrm{tr}(I^*J^{*-1})$ and variance $\tfrac{1}{2}\mathrm{tr}((I^*J^{*-1})^2)$, both of $O(1)$.

Now consider using $\bar{z}_j \equiv \frac{1}{m}\ell_j(\hat{\boldsymbol{\theta}}_j), j = 1, 2, 3$, to compare species trees S_1, S_2 and S_3. We have

$$\mathbb{E}(\bar{z}_j) \approx \mathbb{E}(l_j(\boldsymbol{\theta}_j^*)) \equiv \mu_j,$$
$$\mathbb{V}(\bar{z}_j) \approx \tfrac{1}{m}\mathbb{V}(l_j(\boldsymbol{\theta}_j^*)) \equiv \tfrac{1}{m}\sigma_{jj}, \quad (A37)$$
$$\mathrm{Cov}(\bar{z}_j, \bar{z}_k) \approx \tfrac{1}{m}\mathrm{Cov}(l_j(\boldsymbol{\theta}_j^*), l_k(\boldsymbol{\theta}_k^*)) \equiv \tfrac{1}{m}\sigma_{jk}.$$

Thus when the number of loci $m \to \infty$, $\{\bar{z}_j\} = \{\frac{1}{m}\ell_j(\hat{\boldsymbol{\theta}}_j)\}$ have means (μ_1, μ_2, μ_2) and variance/covariance matrix $\frac{1}{m}\Sigma$, where $\Sigma = \{\sigma_{jk}\}$ is $O(1)$ and independent of m. The error of the ML method, $\mathbb{P}\{\ell_1(\hat{\boldsymbol{\theta}}_1) > \max(\ell_2(\hat{\boldsymbol{\theta}}_2), \ell_3(\hat{\boldsymbol{\theta}}_3))\} = \mathbb{P}\{\bar{z}_1 > \max(\bar{z}_2, \bar{z}_3)\}$, is then given by Theorem 1 as eq. 11.

Table 1. Probabilities (g_1, g_2, g_3) of estimated gene trees at different sequence lengths (n) and the error rates for the summary methods 2-STEP and ISML with $m = 1000$ loci, each with n sites

n	1	2	10	100	1000	∞
			2-STEP (MP-EST)			
$\mathbb{P}(\text{tie})$	0.92948	0.8673	0.57015	0.22159	0.05105	0
$g_1(n)$	0.02378	0.04474	0.14515	0.26646	0.33273	0.35947
$g_2(n) = g_3(n)$	0.02337	0.04398	0.14235	0.25598	0.30811	0.32026
$e_{\text{2-STEP}}$	0.642	0.633	0.597	0.470	0.260	0.114
$\zeta(m, g_1, g_2)$	0.644	0.635	0.600	0.472	0.264	0.113
$\zeta_{\text{ZLY}}(m, g_1, g_2)$	NA	NA	0.613	0.482	0.271	0.117
(ζ_{L1}, ζ_{U1})	(0.635, 0.953)	(0.623, 0.935)	(0.578, 0.869)	(0.430, 0.647)	(0.219, 0.331)	(0.087, 0.132)
(ζ_{L2}, ζ_{U2})	(0.637, 0.729)	(0.626, 0.714)	(0.585, 0.668)	(0.446, 0.561)	(0.242, 0.328)	(0.103, 0.132)
ζ (mean2)	0.683	0.670	0.627	0.504	0.285	0.118
a	0.574051	0.574056	0.573612	0.569708	0.562911	0.555962
b	0.0678913	0.0930368	0.190376	0.527747	1.11658	1.72268
			ISML (CACAT)			
e_{ISML}	0.642	0.632	0.590	0.438	0.246	0.196
ζ_N	0.644	0.634	0.592	0.443	0.254	0.194
ζ_{N0}	0.643	0.633	0.591	0.437	0.234	0.166
(ζ_{L1}, ζ_{U1})	(0.635, 0.953)	(0.622, 0.934)	(0.568, 0.854)	(0.397, 0.598)	(0.211, 0.318)	(0.157, 0.237)
(ζ_{L2}, ζ_{U2})	(0.637, 0.728)	(0.625, 0.713)	(0.576, 0.659)	(0.416, 0.536)	(0.233, 0.316)	(0.177, 0.237)
ζ (mean2)	0.683	0.669	0.618	0.476	0.275	0.207
a	0.574029	0.573971	0.57356	0.569747	0.558232	0.553151
b	0.067892	0.0958963	0.21228	0.607057	1.14253	1.35035

Note.— $\mathbb{P}(\text{tie})$ is the probability for ties in gene trees, with $\mathbb{P}(\text{tie}) + g_1 + 2g_2 = 1$. The probabilities of estimated gene trees (g_1, g_2, g_3) as well as the error rates ($e_{\text{2-STEP}}$ and e_{ISML}) are estimated by simulation using a C program, with $\geq 10^6$ replicates. Ties are broken evenly in the error calculation. The parameter values used are $(\tau_0, \tau_1, \theta_0, \theta_1) = (0.02, 0.019, 0.01, 0.05)$. The marginal (pooled) site pattern probabilities are $\bar{p} = (\bar{p}_0, \bar{p}_1, \bar{p}_2, \bar{p}_3, \bar{p}_4) = (0.92831926, 0.023777106, 0.023372801, 0.023372801, 0.001158033)$, given by eq. 13. For 2-STEP, at $n = 1$, the estimated gene tree is determined by the single site so that $g_1(1) = \bar{p}_1$ and $g_2(1) = \bar{p}_2$, while at $n = \infty$ the estimated gene tree is the true gene tree, so that $g_1(\infty) = \mathbb{P}(G_1)$ and $g_2(\infty) = \mathbb{P}(G_2)$ (eq. 1). For 2-STEP, ζ_{ZLY} (eq. 9) is inapplicable at $n = 1$ or 2 as $m = 1000$ is too small. For ISML, $\zeta_{N0} = \zeta_N(m, \Delta\mu, \sigma_1^2, \sigma_2^2, 0, 0)$ ignores the correlation (eq. 6), while ζ_N accounts for the correlation. The bounds (ζ_{L1}, ζ_{U1}) and (ζ_{L2}, ζ_{U2}) are calculated using eqs. 7 and A27, with $k = 2$ used in ζ_{U2}. "mean2" is the average of the tight bounds: $(\zeta_{L2} + \zeta_{U2})/2$.

Table 2. Estimates of divergence times (true values in parentheses) by ML under the MSC (3S) and by concatenation (BASEML) in two simulated datasets, each of $m = 10^4$ loci and $n = 250$ sites

data/method	τ_{ABC} (0.02)	τ_{AB} (0.01)	θ_{ABC} (0.02)	θ_{AB} (0.01)
dataset 1, 3S	0.0201	0.0096	0.0199	0.0101
dataset 2, 3S	0.0196	0.0100	0.0201	0.0100
dataset 1, BASEML	0.0298	0.0155		
dataset 2, BASEML	0.0298	0.0156		

Table 3. Characterization of the ISML method

|	true model	assumption	data size	parameters	ISML vs. ML
(a)	$\theta_0 \neq \theta_1$	$\theta_0 \neq \theta_1$	$n > 1$	3 out of 4 identifiable	ISML $\neq$ ML
(b)	$\theta_0 \neq \theta_1$	$\theta_0 \neq \theta_1$	$n = 1$	3 out of 4 identifiable	ISML $=$ ML
(c)	$\theta_0 = \theta_1$	$\theta_0 = \theta_1$	$n > 1$	all 3 identifiable	ISML $\neq$ ML
(d)	$\theta_0 = \theta_1$	$\theta_0 = \theta_1$	$n = 1$	all 3 identifiable	ISML $=$ ML
(e)	$\theta_0 \neq \theta_1$	$\theta_0 = \theta_1$	$n > 1$	3 out of 4 identifiable, inconsistent	ISML $\neq$ ML
(f)	$\theta_0 \neq \theta_1$	$\theta_0 = \theta_1$	$n = 1$	3 out of 4 identifiable, inconsistent	ISML $=$ ML

Note.— In all cases, the species tree topology is identifiable and consistently estimated by ISML when the number of loci $m \to \infty$. If the parameters are identifiable, their estimates will be consistent. When ISML differs from ML and the assumed model is correct, ISML is less efficient than ML for parameter estimation (case c).

Table 4. Summary of analytical approximations to species-tree estimation error by different methods

Method	$n = 1$	$n \geq 2$	$n = \infty$
ML		eq. 11	eq. 12
2-STEP	$\zeta(m, \bar{p}_1, \bar{p}_2)$	$\zeta(m, g_1, g_2)$	$\zeta(m, \mathbb{P}(G_1), \mathbb{P}(G_2))$
ISML/concatenation	$\zeta_N(m, \Delta p, \Sigma^{(1)})$	$\zeta_N(m, \Delta p, \Sigma^{(n)})$	$\zeta_N(m, \Delta p, \Sigma^{(\infty)})$

Note.— For ISML/concatenation, $\Delta p = \bar{p}_1 - \bar{p}_2$, and the variance-covariance matrix at n is $\Sigma^{(n)} = \frac{1}{n}\Sigma^{(1)} + \frac{n-1}{n}\Sigma^{(\infty)}$

(eq. 16). In the case of $n = 1$, $\zeta(m, \bar{p}_1, \bar{p}_2) = \zeta_N(m, \Delta p, \Sigma^{(1)})$, and 2-STEP, ISML, concatenation, and ML are all equivalent.
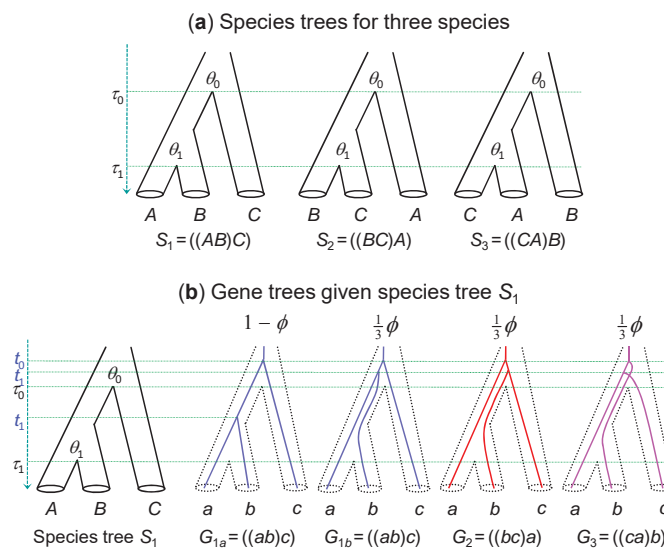
FIG. 1. (**a**) The three species trees (S_1, S_2, S_3) for three species (A, B, C) and the parameters in each MSC model. (**b**) The possible gene trees with coalescent times (t_0, t_1) for a locus with three sequences (a, b, c) given the species tree S_1. The probabilities for the gene trees are shown above them, where $\phi = e^{-\frac{2}{\theta_1}(\tau_0 - \tau_1)}$ is the probability that a and b do not coalesce in population AB or over the time interval (τ_1, τ_0). Note that if the species tree is S_2 (or S_3), it will be possible for sequences b and c (or c and a) to coalesce in the time interval (τ_1, τ_0).
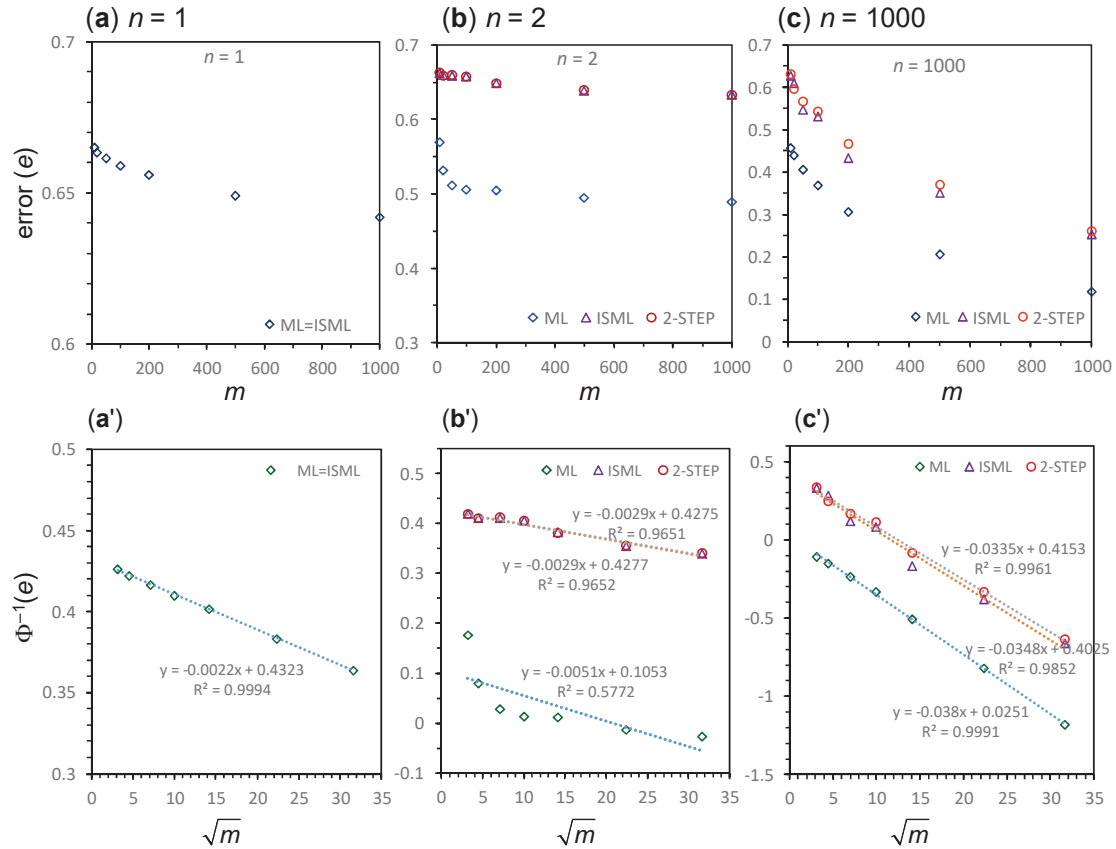
FIG. 2. (**a–c**) Species-tree estimation error (e) at three sequence lengths ($n = 1, 2, 1000$) plotted against the number of loci (m) for different methods. (**a′–c′**) The probit transform of the species-tree error, $\Phi^{-1}(e)$, plotted against $\sqrt{m}$. The parameters used in the simulation are $\tau_0 = 0.02$, $\tau_1 = 0.019$, $\theta_0 = 0.01$, and $\theta_1 = 0.05$. When $n = 1$, all four methods (ML, 2-STEP, concatenation, and ISML) give the same species tree estimate, while concatenation and ISML are equivalent in all cases considered in this paper. The number of replicates is $R \geq 10^4$ for ML and $\geq 10^6$ for the other methods.
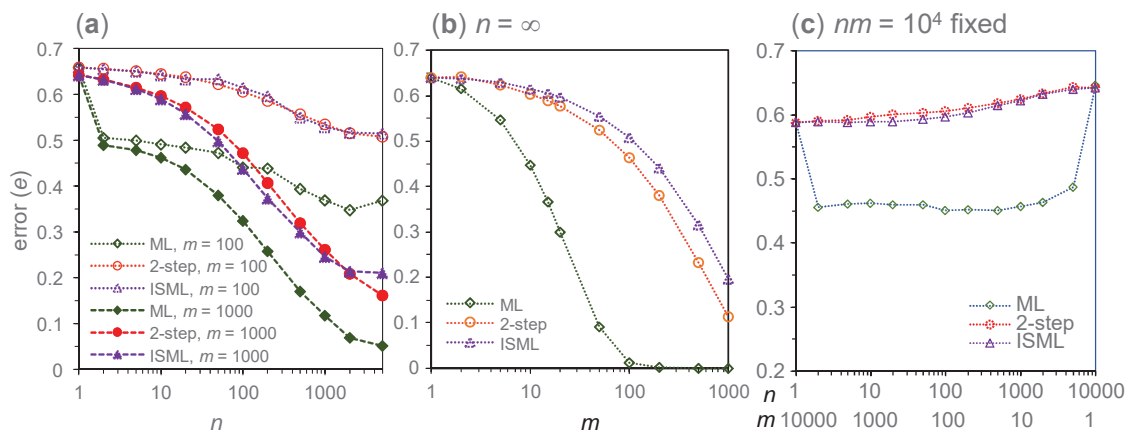
FIG. 3. Error rates in species-tree estimation by ML, 2-STEP and ISML (= concatenation). (**a**) Error plotted against sequence length n when the number of loci m is fixed at 100 or 1000, generated by simulation. (**b**) Error plotted against m when $n = \infty$. Error for ML is given by eq. 12, while those for ISML and 2-STEP are generated by simulation. (**c**) Error plotted against n when $nm = 10^4$ is fixed, generated by simulation. Note that all four methods are equivalent when $n = 1$ or $m = 1$, while concatenation and ISML are equivalent in all cases. Parameters used in the simulation are $\tau_0 = 0.02$, $\tau_1 = 0.019$, $\theta_0 = 0.01$, and $\theta_1 = 0.05$. The number of replicates is $R \geq 10^4$.
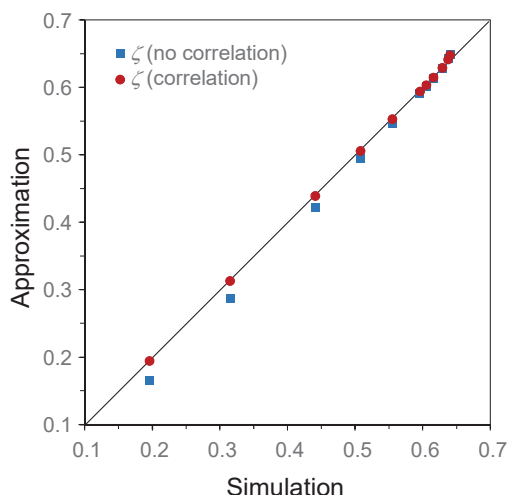
FIG. 4. Species tree error for ISML at $n = \infty$ generated by simulation (10^8 replicates) and by approximation based on ζ_N either with and without accounting for correlations. The error goes from 0.64 (at $m = 1$)) to 0.19 (at $m = 1000$). Results for other methods for the same parameter settings are in figure 3b.
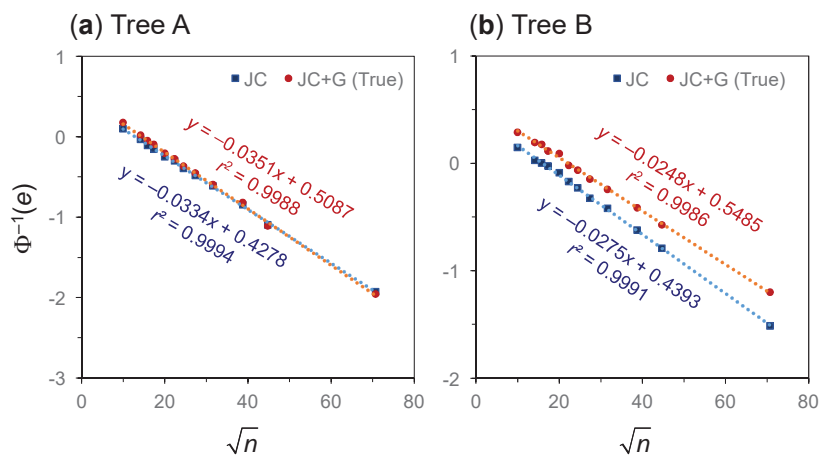
FIG. 5. The probit transform of the phylogenetic reconstruction error, $\Phi^{-1}(e)$, is a linear function of the square root of the number of sites in the alignment ($\sqrt{n}$). Simulation results from Yang (1997, fig. 1A & B) are used in the plot. The trees used in the simulation have four taxa, with branch lengths ((0.5, 0.5):0.1, 0.5, 0.5) for tree A and ((0.5, 0.5):0.1, 0.6, 1.4) for tree B. Data are simulated under the JC+G model (Yang 1994a) and analyzed under both JC and JC+G (Jukes and Cantor 1969; Yang 1994b). Note that in (B), ML under the incorrect model (JC) is more efficient than ML under the correct model (JC+G).
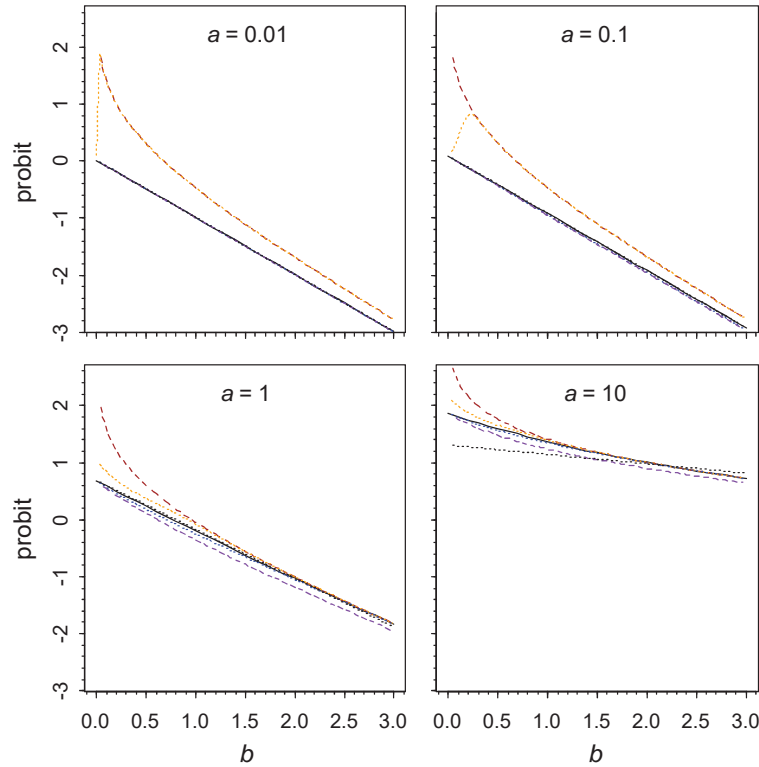
FIG. 6. Probit of error, $\Phi^{-1}(\zeta)$, plotted against b for different values of a. Six methods for calculated ζ are shown. The first five are, from top to bottom, ζ_{U1} (brown dashed line), ζ_{U2} (orange dotted), Exact (black solid line), ζ_{L2} (blue dotted), and ζ_{L1} (purple dashed). Eq. 6 (black dotted) is included as well.
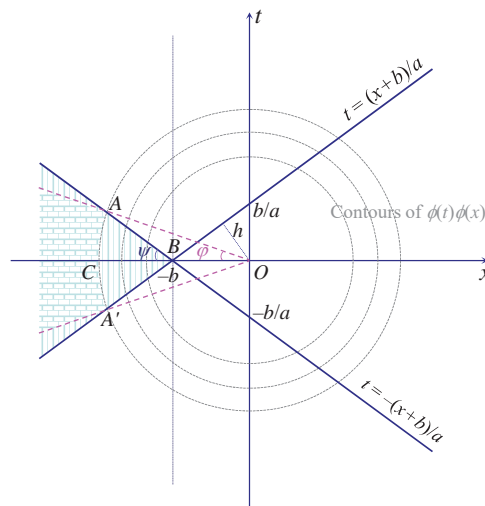
FIG. 7. The areas of integration for integrals in eqs. A11 and A12. The two angles are $\psi = \tan^{-1}\frac{1}{a}$ and $\varphi = \tan^{-1}\frac{1}{ka}$, $k > 1$, with $\varphi < \psi < \frac{\pi}{2}$. The integral over the half-plane $x < -b$ is $\Phi(-b)$, while the integral over the half-plane $t > (x+b)/a$ is $S = \Phi(-h) = \mathbb{P}\{\bar{z}_1 < \bar{z}_2\}$. The integral over the arc ABA' (the shaded area) is $2A = \mathbb{P}\{\bar{z}_1 < \bar{z}_2, \bar{z}_1 < \bar{z}_3\}$. This is smaller than $\Phi(-b) \cdot \psi / \frac{\pi}{2}$ and greater than the integral over the area shaded with the brick pattern: these give the bounds (ζ_{L2}, ζ_{U2}) in Appendix A. The pink dashed lines are $t = (x+b)/(ka)$ and $t = -(x+b)/(ka)$. They cross the blue lines at A and A', with the length of the line segment OA to be $r = \frac{b\sqrt{a^2k^2+1}}{a(k-1)}$. Note that the integral over the circle $x^2 + t^2 < r^2$ is $1 - e^{-\frac{1}{2}r^2}$.

Table S1. Variances-covariance matrix of site-pattern frequencies (f_{jk}) among loci (with n sites per locus)

pattern	$0 : xxx$	$1 : xxy$	$2 : yxx$	$3 : xyx$	$4 : xyz$
mean ($\bar{p}_j$)	0.92831926	0.023777106	0.023372801	0.023372801	0.001158033
$n = 1$					
0: xxx	0.0665371525	-0.0220737114	-0.0216961587	-0.0216906657	-0.00107661674
1: xxy	-0.0220737114	0.0232125909	-0.000555721922	-0.000555581223	$-2.7576288 \times 10^{-5}$
2: yxx	-0.0216961587	-0.000555721922	0.022825064	-0.000546078464	$-2.71046183 \times 10^{-5}$
3: xyx	-0.0216906657	-0.000555581223	-0.000546078464	0.0228194232	$-2.70977559 \times 10^{-5}$
4: xyz	-0.00107661674	$-2.7576288 \times 10^{-5}$	$-2.71046183 \times 10^{-5}$	$-2.70977559 \times 10^{-5}$	0.00115839534
$n = 2$					
0: xxx	0.0333193524	-0.0110518301	-0.010863736	-0.0108646269	-0.000539159418
1: xxy	-0.0110518301	0.0116302167	-0.000282670114	-0.000282578449	$-1.31380675 \times 10^{-5}$
2: yxx	-0.010863736	-0.000282670114	0.0114355946	-0.000276246353	$-1.2942121 \times 10^{-5}$
3: xyx	-0.0108646269	-0.000282578449	-0.000276246353	0.0114364545	$-1.30027622 \times 10^{-5}$
4: xyz	-0.000539159418	$-1.31380675 \times 10^{-5}$	$-1.2942121 \times 10^{-5}$	$-1.30027622 \times 10^{-5}$	0.000578242368
$n = 10$					
0: xxx	0.00674696401	-0.00223806143	-0.00219929448	-0.00219926798	-0.000110340106
1: xxy	-0.00223806143	0.00236735227	$-6.38114276 \times 10^{-5}$	$-6.36625594 \times 10^{-5}$	$-1.81685014 \times 10^{-6}$
2: yxx	-0.00219929448	$-6.38114276 \times 10^{-5}$	0.00232577673	$-6.08615665 \times 10^{-5}$	$-1.80925447 \times 10^{-6}$
3: xyx	-0.00219926798	$-6.36625594 \times 10^{-5}$	$-6.08615665 \times 10^{-5}$	0.00232560997	$-1.81786191 \times 10^{-6}$
4: xyz	-0.000110340106	$-1.81685014 \times 10^{-6}$	$-1.80925447 \times 10^{-6}$	$-1.81786191 \times 10^{-6}$	0.000115784072
$n = 100$					
0: xxx	0.000767781846	-0.000254691861	-0.000249506713	-0.000249726749	$-1.38565232 \times 10^{-5}$
1: xxy	-0.000254691861	0.000282913562	$-1.45054073 \times 10^{-5}$	$-1.44492196 \times 10^{-5}$	$7.32926136 \times 10^{-7}$
2: yxx	-0.000249506713	$-1.45054073 \times 10^{-5}$	0.000275629759	$-1.23445851 \times 10^{-5}$	7.2694645×10^{-7}
3: xyx	-0.000249726749	$-1.44492196 \times 10^{-5}$	$-1.23445851 \times 10^{-5}$	0.000275789815	$7.30738276 \times 10^{-7}$
4: xyz	$-1.38565232 \times 10^{-5}$	$7.32926136 \times 10^{-7}$	7.2694645×10^{-7}	$7.30738276 \times 10^{-7}$	$1.16659123 \times 10^{-5}$
$n = 1000$					
0: xxx	0.000169701052	$-5.65782078 \times 10^{-5}$	$-5.44757242 \times 10^{-5}$	$-5.44615584 \times 10^{-5}$	$-4.18556146 \times 10^{-6}$
1: xxy	$-5.65782078 \times 10^{-5}$	$7.46936699 \times 10^{-5}$	$-9.51058048 \times 10^{-6}$	$-9.60077773 \times 10^{-6}$	$9.95896135 \times 10^{-7}$
2: yxx	$-5.44757242 \times 10^{-5}$	$-9.51058048 \times 10^{-6}$	$7.05311669 \times 10^{-5}$	$-7.5138745 \times 10^{-6}$	$9.69012304 \times 10^{-7}$
3: xyx	$-5.44615584 \times 10^{-5}$	$-9.60077773 \times 10^{-6}$	$-7.5138745 \times 10^{-6}$	$7.06080886 \times 10^{-5}$	$9.68122061 \times 10^{-7}$
4: xyz	$-4.18556146 \times 10^{-6}$	$9.95896135 \times 10^{-7}$	$9.69012304 \times 10^{-7}$	$9.68122061 \times 10^{-7}$	$1.25253095 \times 10^{-6}$
$n = \infty$					
0: xxx	0.000103268649	$-3.42669472 \times 10^{-5}$	$-3.29431183 \times 10^{-5}$	$-3.29485916 \times 10^{-5}$	$-3.10999147 \times 10^{-6}$
1: xxy	$-3.42669472 \times 10^{-5}$	5.1215364×10^{-5}	$-8.98272353 \times 10^{-6}$	$-8.98309369 \times 10^{-6}$	$1.01740041 \times 10^{-6}$
2: yxx	$-3.29431183 \times 10^{-5}$	$-8.98272353 \times 10^{-6}$	$4.78926166 \times 10^{-5}$	$-6.96552523 \times 10^{-6}$	$9.98750549 \times 10^{-7}$
3: xyx	$-3.29485916 \times 10^{-5}$	$-8.98309369 \times 10^{-6}$	$-6.96552523 \times 10^{-6}$	4.7898289×10^{-5}	$9.98921539 \times 10^{-7}$
4: xyz	$-3.10999147 \times 10^{-6}$	$1.01740041 \times 10^{-6}$	$9.98750549 \times 10^{-7}$	$9.98921539 \times 10^{-7}$	$9.49189765 \times 10^{-8}$

Note.— The parameter values used are $(\tau_0, \tau_1, \theta_0, \theta_1) = (0.02, 0.019, 0.01, 0.05)$. The means are calculated using eq. 13, confirmed by simulation. The variances at $n = \infty$ are estimated by simulating gene trees with coalescent times and calculating p_j (eq. 3). Those for other n are estimated by simulating gene trees, calculating site-pattern probabilities p_j (eq. 3), and then using them to sample the site-pattern counts from the multinomial distribution (eq. 4). The number of replicates ranges from $R = 10^6$ to 5×10^9.