Challenges in Species Tree Estimation Under the Multispecies Coalescent Model

Bo Xu* and Ziheng Yang*,†,1

*Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China and †Department of Genetics, Evolution and Environment, University College London, London WC1E 6BT, United Kingdom

ORCID ID: 0000-0003-3351-7981 (Z.Y.)

ABSTRACT The multispecies coalescent (MSC) model has emerged as a powerful framework for inferring species phylogenies while accounting for ancestral polymorphism and gene tree-species tree conflict. A number of methods have been developed in the past few years to estimate the species tree under the MSC. The full likelihood methods (including maximum likelihood and Bayesian inference) average over the unknown gene trees and accommodate their uncertainties properly but involve intensive computation. The approximate or summary coalescent methods are computationally fast and are applicable to genomic datasets with thousands of loci, but do not make an efficient use of information in the multilocus data. Most of them take the two-step approach of reconstructing the gene trees for multiple loci by phylogenetic methods and then treating the estimated gene trees as observed data, without accounting for their uncertainties appropriately. In this article we review the statistical nature of the species tree estimation problem under the MSC, and explore the conceptual issues and challenges of species tree estimation by focusing mainly on simple cases of three or four closely related species. We use mathematical analysis and computer simulation to demonstrate that large differences in statistical performance may exist between the two classes of methods. We illustrate that several counterintuitive behaviors may occur with the summary methods but they are due to inefficient use of information in the data by summary methods and vanish when the data are analyzed using full-likelihood methods. These include (i) unidentifiability of parameters in the model, (ii) inconsistency in the so-called anomaly zone, (iii) singularity on the likelihood surface, and (iv) deterioration of performance upon addition of more data. We discuss the challenges and strategies of species tree inference for distantly related species when the molecular clock is violated, and highlight the need for improving the computational efficiency and model realism of the likelihood methods as well as the statistical efficiency of the summary methods.

KEYWORDS anomaly zone; BPP; concatenation; gene trees; incomplete lineage sorting; maximum likelihood; multispecies coalescent; species trees

In comparisons of genomic sequences from multiple species, it is often observed that different genes or genomic regions may produce conflicting phylogenetic trees. A number of biological processes may cause such gene tree-species tree discordance, including (i) gene duplications and losses combined with misidentification of orthologs, (ii) hybridization, introgression or horizontal gene transfer across species boundaries, and (iii) incomplete lineage sorting (ILS) due to polymorphism in ancestral species (Maddison 1997; Nichols 2001; Degnan and Rosenberg 2009; Szollosi *et al.* 2014). There is increasing empirical evidence that gene flow (introgression or hybridization) occurs commonly between sister or even non-sister species, especially during radiative speciations (Turelli *et al.* 2014; Mallet *et al.* 2016), and indeed many studies have highlighted gene flow and ILS as two major challenges to inference of shallow species phylogenies (Fontaine *et al.* 2015; Pease *et al.* 2016). Currently full likelihood methods that deal with both gene flow and ILS are lacking (Dalquen *et al.* 2016). In this article, we focus on ILS only. ILS occurs when the coalescent process in ancestral species causes the gene tree to differ from the species tree. This is important whenever species divergences are close in time (as occurs in radiative speciations) and the population sizes of the ancestral species are large. The significance of ILS to species tree inference was highlighted by the characterization of the anomaly zone, regions of the parameter space

doi: 10.1534/genetics.116.190173
Manuscript received April 6, 2016; accepted for publication September 25, 2016
[1]Corresponding author: Department of Genetics, Evolution and Environment, University College London, Gower St., London WC1E 6BT, United Kingdom. E-mail: z.yang@ucl.ac.uk

(species tree with associated parameters) with short internal branches and large population sizes, in which the majority-vote approach of using the most common gene tree as an estimate of the species tree is statistically inconsistent (Degnan and Salter 2005; Degnan and Rosenberg 2006). Intuitively one might think that coalescent is a population genetics process and irrelevant to species tree estimation. However, the issue lies with the *length* rather than the *depth* of the internal branches on the species phylogeny (Edwards *et al.* 2005). For example, the species tree of Figure 1A is hard to reconstruct due to the short internal branches (this will be discussed in detail later), but the task is not made any easier if the tip branches (*A*, *B*, *C*, and *D*) are extended by 300 MY of evolution.

The multispecies coalescent (MSC) model provides a natural framework for estimating the species tree in presence of ILS, and indeed the use of the coalescent model in species tree inference has been described as a paradigm shift in molecular phylogenetics (Edwards 2009). Nearly all coalescent-based species tree estimation methods were developed within the last 10 years. Two classes of methods have been developed side by side: the full likelihood methods (including maximum likelihood, ML, and Bayesian inference, BI) and approximate or summary methods. Full likelihood methods involve averaging over the gene trees and are computationally intensive. The ML method integrates over the coalescent times numerically so that the computation is possible for three species only but tens of thousands of loci can be handled (Yang 2002; Dalquen *et al.* 2016). Bayesian methods such as BEST (Liu and Pearl 2007; Edwards *et al.* 2007; Liu 2008), *BEAST (Heled and Drummond 2010), and BPP (Yang and Rannala 2014; Rannala and Yang 2016) use Markov chain Monte Carlo algorithms to average over gene trees (topologies and branch lengths) and parameters. They can deal with more species but are currently impractical for large datasets with >1000 loci. Furthermore, current Bayesian implementations do not deal with the violation of the molecular clock adequately and do not appear to work very well for deep species phylogenies (Ogilvie *et al.* 2016). At the same time, over a dozen approximate methods have been proposed for estimating the species tree despite gene tree conflicts. Most of them take a two-step approach of estimating the gene trees at the individual loci using phylogenetic methods and then treating the estimated gene trees as observed data. These are also called summary methods, as gene trees may be viewed as summary statistics derived from the original sequence data. However traditional summary statistics are features or observations of the data, while gene trees are estimates and may differ from the true unobserved gene trees. Approximate or summary methods are fast and applicable to genomic datasets with hundreds or thousands of loci and have been more commonly used than likelihood methods. For recent reviews on species tree methods, the reader may consult Yang (2014, Chap. 9), Liu *et al.* (2015), Edwards *et al.* (2016), and Mallo and Pasada (2016).

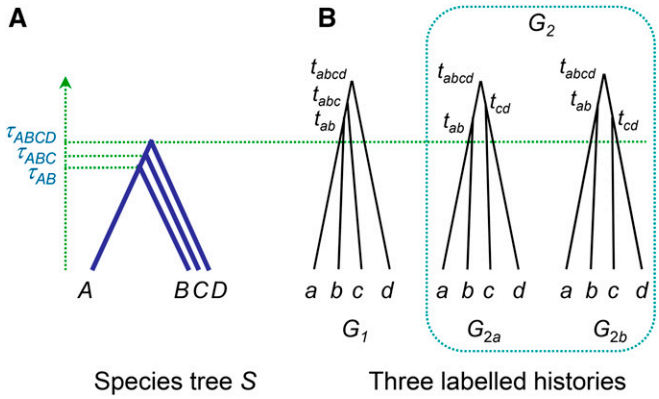

Figure 1 (A) Asymmetrical species tree (*S*) for four species and (B) symmetrical and asymmetrical gene trees (G_1 and G_2). When the two internal branch lengths in the species tree are ≈0, all three coalescent events on the gene tree occur in the common ancestor *ABCD*, so that all 18 labeled histories have equal probabilities ($\frac{1}{18}$) (Figure. 2), with $P(G_2) \approx 2P(G_1)$. When the internal branch lengths are nonzero but very small, it is possible to have $P(G_2) > P(G_1)$, in which case the species tree *S* is in the anomaly zone.

In this article we review the statistical nature and conceptual issues of species tree estimation under the MSC. We focus mainly on simple cases involving three or four species, because they are more likely to be tractable and because most summary methods are based on insights from small species trees (*e.g.*, triplet and quartet trees). We assume that the species are closely related so that the molecular clock approximately holds, but will discuss the additional challenges for inferring deep species trees. We demonstrate that counterintuitive behaviors of species tree methods discussed in the literature, such as the existence of the anomaly zone, the singularity of the likelihood surface, and worse performance with more data, apply to the two-step summary methods, but not to the full-likelihood methods. We first study the case of three species and three sequences per locus, which is the simplest species tree problem. Then we discuss the anomaly zone, regions of the parameter space within which the most frequent gene tree differs from the species tree, so that the simple majority-vote method of species tree estimation is inconsistent. Lastly we use simulation to confirm that more data can indeed cause summary methods such as MP-EST (Liu *et al.* 2010a) to perform worse, but show that this is due to inefficient use of information in the data by summary methods and does not occur with the full likelihood methods (ML and BI).

The Multispecies Coalescent

The coalescent describes the stochastic process of lineage joining when one traces the genealogical history of a sample of sequences from a population backward in time until their most recent common ancestor (MRCA) (Kingman 1982a,b; Hudson 1983; Tajima 1983). The theory provides a powerful framework for inference of population genetic processes using genetic sequence data (Hudson 1990; Hein *et al.* 2005; Nordborg 2007; Wakeley 2009). The coalescent waiting time

for two randomly sampled sequences from a diploid species with population size N has an exponential distribution with the mean of $2N$ generations, or equivalently coalescent between two sequences occurs at the Poisson rate of $1/(2N)$ per generation. In analysis of sequence data, it is convenient to measure time by the mutational distance, so that one time unit is defined as the amount of time taken to accumulate one mutation per site in the sequence. Then coalescent between two sequences occurs at the rate of $\frac{2}{\theta}$ per time unit, and the average coalescent waiting time is $\theta/2$, where $\theta = 4N\mu$ is the expected number of mutations per site between two randomly drawn sequences and μ is the mutation rate per site per generation. For example, $\theta_{\mathrm{H}} \approx 0.0006$ for the human means that two human genomes have on average ~0.6 differences per kilobase.

For a sample of n sequences, the genealogical history is described by a succession of coalescent lineage-joining events. With j sequences in the sample, each of the $\binom{j}{2}$ pairs coalesces at the rate $\frac{2}{\theta}$, so that the total rate is $\frac{2}{\theta} \times \binom{j}{2}$, and the waiting time until the next coalescent event (which reduces the number of lineages from j to $j-1$) is an exponential random variable with mean $\frac{\theta}{2}/\binom{j}{2}$. This process generates $H = \binom{n}{2} \cdot \binom{n-1}{2} \cdots \binom{2}{2}$ possible genealogical trees, depending on the order of pairs of sequences to coalesce, and each of them has the same probability $(1/H)$. Here the genealogical tree is a rooted tree with the internal nodes ordered by time, called a *labeled history* (LH) by Edwards (1970). Figure 2 shows all the 18 LHs for four sequences. The joint probability distribution of the genealogical tree (G) and the coalescent times, $\boldsymbol{t} = \{t_j\}$, is thus

$$f(G, \boldsymbol{t}|\theta) = f(G) f(\boldsymbol{t}|G, \theta) = \frac{1}{H} \times \prod_{j=2}^{n} \frac{2}{\theta} \binom{j}{2} \exp\left\{ -\frac{2}{\theta} \binom{j}{2} t_j \right\}$$
$$= \prod_{j=2}^{n} \frac{2}{\theta} \exp\left\{ -\frac{2}{\theta} \binom{j}{2} t_j \right\}. \quad (1)$$

The multispecies coalescent (MSC) is a simple extension of the single-population coalescent to multiple species (Figure 3). The different species are related by a phylogeny (species tree), and they may have different population sizes. There are two sets of parameters in the MSC model: the species divergence times (τs) and the population size parameters (θs) on the species tree, with both τs and θs measured by the number of mutations or substitutions per site. For example, for the species tree of Figure 3A, $\Theta = \{\tau_{AB}, \tau_{ABC}, \theta_{AB}, \theta_{ABC}\}$. Within each species (either modern or ancestral), sequences coalesce at random, at the rate $\frac{2}{\theta_i}$ for each pair in population i, as in the standard coalescent, independently of other populations. However at the time of species divergence, two or more lineages may leave the population and enter the ancestral population. Different aspects of the MSC model have been discussed by a number of authors, including Gillespie and Langley (1979), Hudson (1983), Pamilo and Nei (1988), Felsenstein (1988), and Takahata (1989). The probability density of the gene tree and branch lengths (coalescent times) for an arbitrary species tree and arbitrary sampling of sequences at a locus is given by Rannala and Yang (2003). An important feature of the process is that *the sequence divergence time is greater than the species divergence time*, or more intuitively, that *gene trees must "fit inside" the species tree*.

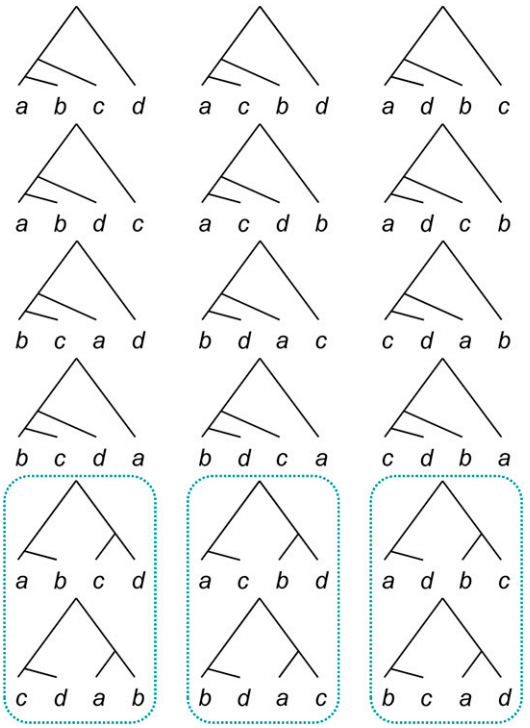

Figure 2 The 18 labeled histories for four sequences sampled from a population (*a*, *b*, *c*, *d*), with the node ages drawn to reflect the expectations of the coalescent times. A labeled history is a rooted tree with the interior nodes rank-ordered by age. Thus the rooted tree ((*a*, *b*), (*c*, *d*)) corresponds to two labeled histories, depending on whether sequences *a* and *b* coalesce before or after sequences *c* and *d*. Under the coalescent model, all possible labeled histories (but not the rooted trees) have equal probabilities. For four sequences, each of the 12 asymmetrical rooted trees is compatible with only one labeled history and has probability 1/18, while each of the three symmetrical rooted trees is compatible with two labeled histories and has probability 2/18.

Full Likelihood Methods of Species Tree Estimation Under the MSC

Full likelihood methods of species tree estimation under the MSC follow standard statistical theory. Let S be the species tree topology, and Θ be the parameters in the MSC model on the species tree (τs and θs). S is a statistical model while Θ are its parameters. The data consist of sequence alignments at L loci. The ideal loci for this kind of analysis are 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 1986; Burgess and Yang 2008; Lohse *et al.* 2011). For distantly related species, a locus may correspond to a gene or exon. The MSC model assumes that the gene trees at different loci are independent while all sites at the same locus share the same history (gene

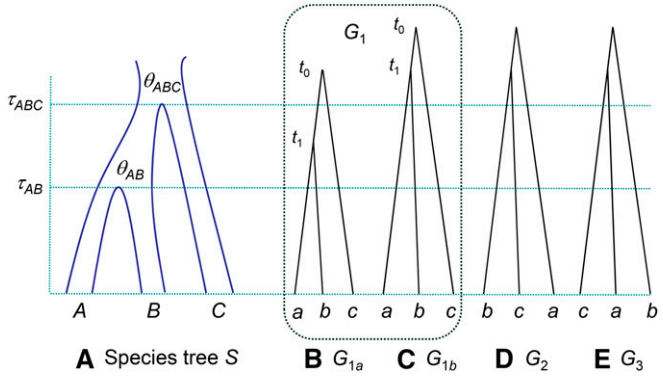

Figure 3 (A) The species tree ((*A*, *B*), *C*) for three species, showing the parameters in the MSC model, $\Theta = \{\tau_{ABC}, \tau_{AB}, \theta_{ABC}, \theta_{AB}\}$. Both τs and θs are measured by the expected number of mutations per site. If multiple sequences are sampled for the same locus from the same species (say, *A*), the population size parameter for that species (say, θ_A) will also be a parameter. (B–E) The possible gene trees for a locus with three sequences (*a*, *b*, *c*), one sequence from each species. Under the MSC, gene trees G_{1b}, G_2, and G_3 have the same probability, so that the species tree-gene tree mismatch probability is $P_{SG} = P(G_2) + P(G_3) = 1 - P(G_1)$.

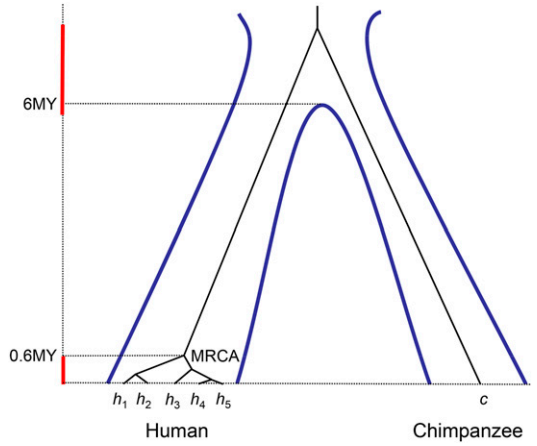

Figure 4 The MSC assumes that all sites at the same locus share the same gene tree (topology and branch lengths). This assumption is valid if there is no recombination around the time periods when coalescent events occur (highlighted by thick bars on the time axis), even though recombination may occur in other parts of the gene tree, when there is only one sequence ancestral to the sample in a population. In the example, humans and chimpanzees diverged at 6 MA, while the MRCA for the human sample is at 0.6 MA. Recombination events over the time period (0.6, 6) do not affect the MSC density of gene trees.

tree topology and coalescent times). The assumption of shared history at the same locus is commonly considered to be equivalent to the assumption of no recombination throughout the gene tree. For distantly related species, a locus without recombination throughout the gene tree (called a coalescent-gene or c-gene by Springer and Gatesy 2016) must be very short. For example, for a eutherian mammal dataset, Springer and Gatesy (2016) used empirical estimates of primate recombination rates to calculate the *c*-gene size to be ~12 bp. However, this calculation is unnecessarily stringent. All sites at the locus will have the same gene tree topology and coalescent times so that the MSC density of Rannala and Yang (2003) will be valid as long as there is no recombination during the parts of the gene tree where coalescent events occur (Lanier and Knowles 2012; Edwards *et al.* 2016). See Figure 4 for an illustration of the assumption using the human and chimpanzee as an example.

Let X_i be the alignment of n_i sequences at locus i, with $X = \{X_i\}$. The sequences are assumed to be neutrally evolving. Let G_i be the gene tree (or more precisely, the labeled histories, Edwards 1970) and coalescent times ($\boldsymbol{t}_i$) at the locus. The gene trees (G_i and $\boldsymbol{t}_i$) are unobserved random variables (latent variables) with well-specified distributions given the species tree and parameters. The log likelihood function for estimating S and Θ averages over the gene trees

$$\ell(S,\Theta) = \sum_i \log f(X_i|S,\Theta) = \sum_i \log\left\{\sum_{G_i}\left[\int f(G_i,\boldsymbol{t}_i|S,\Theta) f(X_i|G_i,\boldsymbol{t}_i) \mathrm{d}\boldsymbol{t}_i\right]\right\}, \quad (2)$$

where $f(G_i, \boldsymbol{t}_i \mid S, \Theta)$ is the MSC density for the gene tree and coalescent times at locus i (Rannala and Yang 2003) and $f(X_i \mid G_i, \boldsymbol{t}_i)$ is the probability of the sequence data given the gene tree, known as the phylogenetic likelihood (Felsenstein 1981). If the mutation/substitution model assumed in the likelihood calculation involves parameters (such as the transition/transversion rate ratio κ), they should be included in Θ as well; in this paper, we use the simple JC mutation model (Jukes and Cantor 1969). The ML method estimates S and Θ by maximizing $\ell(S, \Theta)$. The difficulty with ML lies in the sum over all possible gene trees and the integral over the coalescent times, because the number of possible gene trees is huge and the integral over $\boldsymbol{t}_i$ for each gene tree is $(n_i - 1)$-dimensional. The only ML implementation appears to be that of Yang (2002), which is limited to only three species and three sequences per locus (one sequence from each species).

With the Bayesian approach, we assign a prior on the species trees, $f(S)$, and a prior on the parameters, $f(\Theta \mid S)$. The posterior of S and Θ averages over the gene tree topologies (G_i) and coalescent times ($\boldsymbol{t}_i$):

$$f(S,\Theta|X) = \frac{1}{C} f(S) f(\Theta|S) \times \prod_i f(X_i|S,\Theta) = \frac{1}{C} f(S) f(\Theta|S) \times \prod_i \left\{\sum_{G_i}\left[\int f(G_i,\boldsymbol{t}_i|S,\Theta) f(X_i|G_i,\boldsymbol{t}_i)\, \mathrm{d}\boldsymbol{t}_i\right]\right\}, \quad (3)$$

where C is the normalizing constant, to ensure that the posterior integrates to 1. Equation 3 is not practical because both the normalizing constant C and the marginal likelihood $f(X_i \mid S, \Theta)$ involve huge sums and high-dimensional integrals. Instead, Bayesian programs such as BEST (Liu and Pearl 2007), *BEAST (Heled and Drummond 2010) and BPP (Yang and Rannala 2014; Rannala and Yang 2016) use Markov chain Monte Carlo (MCMC) algorithms to generate a sample from the

joint posterior distribution of species tree (S and Θ) and gene trees (G_i and $\boldsymbol{t}_i$ for each locus i):

$$f(S, \Theta, \{G_i, \boldsymbol{t}_i\} \mid X) \propto f(S) f(\Theta|S) \prod_i \{f(G_i, \boldsymbol{t}_i|S, \Theta) f(X_i|G_i, \boldsymbol{t}_i)\}. \quad (4)$$

Then by ignoring the gene trees $\{G_i, \boldsymbol{t}_i\}$ in the MCMC sample, we obtain the marginal posterior of the species tree: $f(S \mid X)$ as well as the within-tree parameter posterior $f(\Theta \mid X, S)$. While the Bayesian method is commonly described as a joint estimation of both the species tree and the gene trees, it should be noted that the maximum *a posteriori* (MAP) species tree, which is the Bayesian point estimate, maximizes the marginal posterior $f(S \mid X)$, instead of the joint posterior $f(S, \Theta, \{G_i, \boldsymbol{t}_i\} \mid X)$ or $f(S, \Theta \mid X)$. Recall that given a table of counts representing a joint distribution, $f(X, Y)$, with rows for X and columns for Y, the "marginal" distribution for X is generated by summing the counts along each row and writing the sum in the right margin, with the columns ignored. In effect, the summation and integration over the gene trees of Equation 3 is achieved numerically in the MCMC algorithm.

The statistical properties of the ML and BI methods when the number of loci (L) and/or the number of sites per locus (n) increases have yet to be carefully investigated. Under the MSC model, the data (sequence alignments) at the different loci are independently and identically distributed. Statistical consistency of the ML and Bayesian estimates of the species tree when L increases then follows automatically as long as the MSC model is identifiable (*e.g.*, Dawid 2011). In other words, when $L \to \infty$, the estimate will converge in probability to the true species tree and true parameter values. Identifiability under commonly used substitution models has been discussed by Steel (2013) and Chifman and Kubatko (2015). The efficiency of the methods, while very important, appears intractable analytically.

Approximate Methods of Species Tree Estimation

The MSC makes simple predictions about different aspects of the gene trees. If the gene trees are known or estimated, those properties can be used to devise methods for species tree estimation. Almost all approximate methods take this two-step approach and treat gene trees inferred using phylogenetic methods as observed data. We review some of them here to illustrate the strategies taken, but do not attempt an exhaustive list.

Methods that use gene tree topologies

A number of authors have studied the probabilities of gene tree topologies under the MSC. The case of three species (A, B, and C), with the species tree $((AB)C)$, and of three sequences per locus (a, b, and c) was considered by Hudson (1983), who derived the mismatch probability that the species tree (S) and the gene tree (G) differ to be $\frac{2}{3}$ the probability that sequences a and b do not coalesce in the ancestral species AB (Figure 3)

$$P_{SG} = \frac{2}{3} e^{-\frac{(T_{ABC} - T_{AB})}{2N_{AB}}} = \frac{2}{3} e^{-\frac{2(\tau_{ABC} - \tau_{AB})}{\theta_{AB}}}. \quad (5)$$

Note that for a Poisson process with rate λ (or with exponential waiting time of mean $1/\lambda$), the probability of no event over time interval t is $e^{-\lambda t}$. Here the exponent, $(T_{ABC} - T_{AB})/(2N_{AB})$, is known as the internal branch length of the species tree in coalescent time units, since a coalescent time unit is $2N$ generations. Equation 5 was used to estimate the ancestral population size of the human and chimpanzee in the so-called "trichotomy" or "tree-mismatch" method (Takahata *et al.* 1995; Chen and Li 2001; Yang 2002). Equation 5 also gives the most common gene tree as the estimate of the species tree topology. This underlies the rooted triples (Ewing *et al.* 2008) and MP-EST (Liu *et al.* 2010a) methods of species tree estimation.

Gene tree probabilities for four or more species have been studied by Pamilo and Nei (1988). Degnan and Salter (2005) and Degnan and Rosenberg (2006) designed general algorithms that apply to arbitrary numbers of species. Given a collection of gene trees, the likelihood function — the probability of the gene tree topologies given the species tree S and internal branch lengths in coalescent units ($\boldsymbol{\tau}$') — will simply be the product of the probabilities for the gene tree topologies at the multiple loci.

$$\ell(S, \tau') = \sum_{i=1}^{L} \log P(G_i), \quad (6)$$

where the probability of gene tree G_i at locus i can be calculated using the algorithms of Degnan and Salter (2005) and Degnan and Rosenberg (2006) (see also Wu 2012). This likelihood method treating gene tree topologies as data are implemented in the STELLS program (Wu 2012) and is a special case of the network method of Wen *et al.* (2016), with gene flow disallowed. The multinomial likelihood leads to a statistically consistent method for species tree estimation, as the distribution of rooted gene tree topologies determines the rooted species tree topology and internal branch lengths (in coalescent units) (Allman *et al.* 2011).

The method of minimizing deep coalescence (MDC) (Maddison 1997) is based on gene tree topologies but does not use all information in them. Maddison (1997) defines *deep coalescence* as the phenomenon that two or more lineages pass through an ancestral species when one traces the gene genealogy backward in time. If ancestral populations are very small, all lineages should coalesce as soon as they enter an ancestral species, and deep coalescence does not occur. Then the gene trees will track the species tree faithfully. A parsimony-like argument suggests that given a collection of gene trees, the species tree that minimizes the number of deep coalescence events (Figure 5) is likely to be the true species tree (Maddison 1997; Than and Nakhleh 2009). Than and Rosenberg (2011) demonstrate that anomaly zones or inconsistency zones exist for MDC for asymmetric species trees for four species, and for all species trees with five or more species.

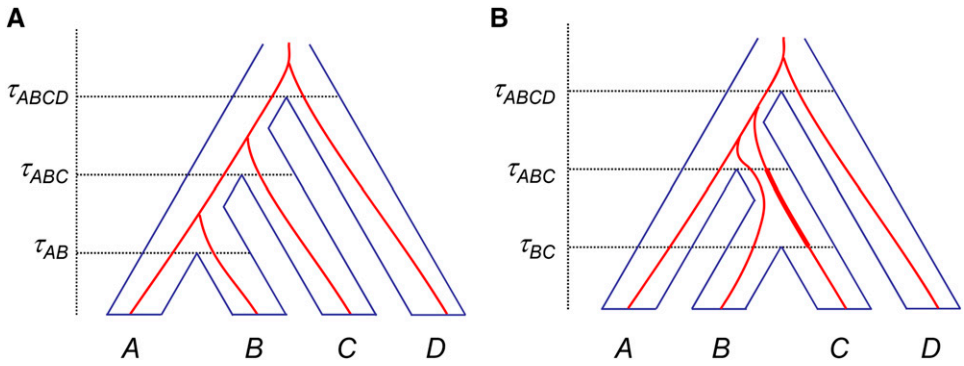

Figure 5 Deep coalescence, marked by thick segment of a gene-tree branch, means that two or more lineages pass through an ancestral species when one traces the gene genealogy backward in time (Maddison 1997). The given rooted gene tree, (((*a*, *b*), *c*), *d*), is fitted to two species trees: (((*A*, *B*), *C*), *D*) in (A) and ((*A*, (*B*, *C*)), *D*) in (B). In (A), at most one lineage leaves each ancestral species so that the number of deep coalescence is 0. In (B), two lineages (sequences *b* and *c*) pass ancestral species *BC* and one of them is counted as a deep coalescence. The method of minimum deep coalescence for species tree estimation (MDC) minimizes the total number of deep coalescence over all gene trees.

Many gene tree-based summary methods make use of properties for small gene trees with three or four sequences and then assemble the triplet or quartet trees into a big tree, which is the estimate of the species tree for all species. The rooted triples method (Ewing *et al.* 2008) and the MP-EST method (Liu *et al.* 2010a) use different strategies to assemble triplet trees. In the rooted triples method, rooted triplet trees are extracted from all gene trees and their consensus tree is constructed as an estimate of the species tree for all species. In the MP-EST method, the probability of the three alternative gene trees given a triplet species tree can be calculated from the trinomial distribution (see Equation 5 and our discussion below on the three-species case). For each species tree (for all species), a likelihood is then calculated by multiplying the trinomial probabilities across all species triplets induced by the specie tree. As the species triplets are not independent, this likelihood function is not the correct one and the method is thus a pseudo-likelihood approach. Both the rooted triples and the MP-EST methods are statistically consistent, as the probability distribution of rooted triplet gene trees determines the species tree topology and internal branch lengths (in coalescent units) (Allman *et al.* 2011). The program ASTRAL (Mirarab *et al.* 2014; Mirarab and Warnow 2015) assembles quartet trees. It takes a collection of unrooted gene trees, collects all their quartets into a set, which may have many identical quartet trees, and then judges each species tree by how well its quartet trees match those in the set. Thus ASTRAL produces the maximum quartet support species tree (MQSST) (Mirarab *et al.* 2014).

Some gene tree-based summary methods use gene tree topologies to define a distance between species, such that the distance tracks the species tree under the MSC. Liu *et al.* (2009) used the rank of the interior nodes on the gene trees as a measure of distance between species, and developed the method of species tree estimation using average ranks of coalescences (STAR) (Figure 6A). This is based on the observation that the expected ranks of the coalescences among sequences tracks the order of species divergences in the species tree. The method is consistent when the correct gene trees are given as data. For unrooted gene trees, Liu and Yu (2011) defined the distance between two species as the number of internodes between the two species on the gene tree, and use the average gene-tree internode distance to construct the NJ tree (Saitou and Nei 1987) (Figure 6B). This is the NJ_{st} method of species tree estimation. Liu and Yu (2011) show that for any four species, the expected gene-tree internode distance satisfies the four-point condition, and thus NJ_{st} is statistically consistent in estimating the unrooted species tree.

In theory if the data are the gene tree topologies, the ML method (Equation 6) uses all information in the data about the species tree, and summary methods such as MDC, STAR, ASTRAL, and NJ_{st} should suffer from an information loss and reduced efficiency. Not much effort has been made to quantify the information loss.

Methods that use gene tree branch lengths

A class of summary methods use branch lengths (or coalescent times) in gene trees, perhaps in addition to the gene tree topologies. A major feature of the MSC model is that *genes split before species* or *gene trees run inside the species tree*: the sequence divergence between any two species must be greater than the species divergence, or $t_{ab} > \tau_{AB}$ for any two species A and B and any two sequences a and b from the two species. As a result, the expected coalescent time between species tracks the species phylogeny. If species A and C diverged earlier than species A and B, with $\tau_{AB} < \tau_{ABC}$ (Figure 3), the expected sequence divergences will track that relationship: $E(t_{ab}) < E(t_{ac})$. Indeed, if all species have the same population size (θ), the expected sequence divergences are simply $E(t_{ab}) = \tau_{AB} + \frac{1}{2}\theta$ and $E(t_{ac}) = \tau_{ABC} + \frac{1}{2}\theta$, so that $E(t_{ab}) < E(t_{ac})$ follows from $\tau_{AB} < \tau_{ABC}$. If we define the distance between two species as twice the average coalescent time (node age) between the species, the resulting distance matrix can be used to construct a NJ tree, which will be a consistent estimate of the species tree. This is species tree estimation using average coalescence times (STEAC) of Liu *et al.* (2009).

Takahata (1989) considered the case of three species with multiple sequences sampled per species, and argued that the *minimum* sequence divergence between species (over

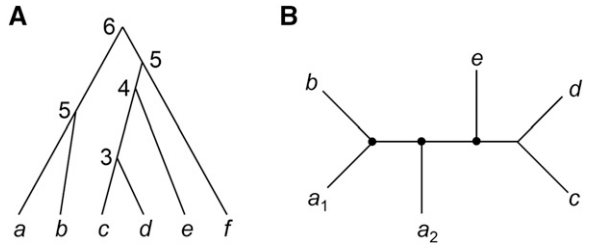

Figure 6 Gene tree topologies can be used to define a distance between two species and the resulting distance matrix can be used to construct the species tree using, *e.g.*, NJ (Saitou and Nei 1987). (A) In the STAR method (for species tree estimation using average ranks of coalescences, Liu *et al.* 2009), the distance between two species is defined as the rank of the ancestral node for the two species on the rooted gene tree. In the example tree, species *A* and *B* have the distance or rank 5, while species *A* and *C* have the rank 6. The rank for the root is the number of sequences, and the rank decreases from the root to the tips of the gene tree. Note that distantly related species tend to have large distances or ranks. A distance matrix is constructed by averaging the ranks across all gene trees, and then analyzed using NJ (Liu *et al.* 2009). (B) The NJ_{st} method (Liu and Yu 2011) uses the gene-tree internode distance, defined as the number of internal nodes in the unrooted gene tree between the two species. If multiple sequences are sampled from the same species, the internode distance is averaged across all pairs from the two species. In the example, the internode distance is 3 between species *B* and *E* and is 1.5 between *A* and *B*. The matrix of average internode distances between species (averaged across loci) is used to construct a species tree using NJ.

all sequence pairs and over all loci) may be more informative about the species tree than the *average* sequence divergence between species. Maddison and Knowles (2006) proposed to cluster species by the shallowest (minimum) coalescences occurring between two species. The method is formalized as the ML method of species tree estimation when the data are rooted gene trees with branch lengths (node ages). This is maximum tree (MT) method of Liu *et al.* (2010b), implemented in the STEM program (Kubatko *et al.* 2009), or the GLASS algorithm (for Global LAteSt Split) (Mossel and Roch 2010). Under the assumption of equal population size, MT, STEM, and GLASS are equivalent. It is remarkable that the ML solution is analytically tractable in this case, and here we use the case of two species to illustrate the result.

In the case of two species with one sequence from each species at every locus, the MSC model involves two parameters: $\Theta = \{\tau_{AB}, \theta_{AB}\}$. Suppose the sequence divergence times at L loci are given: $X = \{x_i\}$, $i = 1, \ldots, L$. The likelihood is then the product of the exponential densities:

$$f(X|\tau_{AB}, \theta_{AB}) = \prod_{i=1}^{L} \frac{2}{\theta_{AB}} \mathrm{e}^{-\frac{2}{\theta_{AB}}(x_i - \tau_{AB})}, \quad \theta_{AB} > 0, \tau_{AB} < \min(x_j). \tag{7}$$

The MLEs are easily found to be $\hat{\tau}_{AB} = \min(x_i)$ and $\hat{\theta}_{AB} = 2(\bar{x} - \hat{\tau}_{AB})$, where $\bar{x}$ is the mean sequence divergence across loci. The MLE of the species divergence is the minimum sequence divergence.

This holds true in general with an arbitrary number of species. Given a collection of rooted gene trees with node ages, the ML species tree under the assumption of equal population size is the one in which the species divergence between any two species is equal to the minimum sequence divergence across all gene trees for the two species (Liu *et al.* 2010b). The ML species tree or the maximum tree achieves the maximum species divergences allowed by the gene trees. However, if the different species are allowed to have independent population sizes (θs), the ML method is noted to encounter singularities on the likelihood surface (Liu *et al.* 2010a). This will be discussed later.

While in theory the MT method uses information in both the branch lengths and the gene tree topologies, it did not perform particularly well in simulations (Liu *et al.* 2009; Leaché and Rannala 2011), apparently because of its sensitivity to estimation errors in branch lengths. Estimation of branch lengths or node ages in the ultrametric rooted gene tree may be affected by the mutation model, by rate variation among loci and among sites of the same locus, and by violation of the molecular clock. Thus methods that use branch lengths may not necessarily perform better than topology-only methods. Similarly, Liu *et al.* (2009) found that STAR (which ignores branch lengths) consistently outperforms STEAC (which uses branch lengths) when the molecular clock is seriously violated. Nevertheless, for closely related species, the clock should be adequate, and the impact of the mutation model should be minor. In this case, the information gain from the use of branch lengths relative to the sensitivity to sampling errors in branch length estimation may be an interesting topic for further research, as are possible strategies for accommodating sampling errors in branch length estimation.

The two-step summary methods suffer from a few weaknesses. First, they use only part of the data such as the gene tree topology, resulting in information loss. Second, they ignore the phylogenetic errors in gene tree reconstruction due to the finite number of sites at each locus (called the mutation variance by Huang and Knowles 2009). For closely related species, the sequences may be highly similar, with very limited phylogenetic information so that gene tree topologies may be unresolved or highly uncertain. Simulations using both the true and estimated gene trees often suggest large differences, highlighting the importance of phylogenetic reconstruction errors (Mirarab and Warnow 2015). However, accounting for uncertainties in the gene trees using multilocus bootstrap did not lead to consistent improvement to species tree estimation and sometimes made things worse (Mirarab *et al.* 2014). This appears to be due to the fact that the bootstrap proportions are not appropriate weights for alternative gene trees at the same locus. Even the Bayesian posterior probabilities may not be the appropriate weights, if they are not calculated under the MSC model with a species tree underlying all gene trees. From Equation 2, the gene tree density $f(G_i, \boldsymbol{t}_i \mid S, \Theta)$ depends on the species tree and parameters, and the gene trees should be correlated among loci.

The invariance method SVDquartets

SVDquartets (for Singular Value Decomposition for quartets) is a quartet-based method recently developed by Chifman and

Kubatko (2014), and uses a quartet assembly algorithm to generate a species tree estimate. However, it differs from the two-step summary methods in that it generates quartet trees from the 256 observed site pattern counts, not from estimated gene trees. SVDquartets assumes that different sites in the sequence data have independent histories given the species tree, and is thus similar to the SNAPP method for single nucleotide polymorphism data (SNPs) (Bryant *et al.* 2012).

Let the observed site pattern counts be o_k, with $k = 1, \ldots, 256$. The standard statistical approach to analysis of such data is to use a goodness-of-fit criterion such as $\ell = \sum_{k=1}^{256} o_k \log \frac{o_k}{np_k}$ or $X^2 = \sum_{k=1}^{256} \frac{(o_k - np_k)^2}{np_k}$ to estimate the parameters for each species tree, where the expected site pattern probability p_k is calculated as a function of the species tree and parameters by summing over the 18 labeled histories and integrating over three coalescent times in each (Figure 2). In other words, p_k is $f(X_i|S, \Theta)$ of Equation 2 with X_i being an alignment of only one site. This calculation can be achieved analytically under the JC model (Chifman and Kubatko 2014). Maximization of ℓ to obtain the MLE of Θ or minimization of X^2 to obtain the minimum X^2 estimate of Θ also leads to the optimized ℓ or X^2 as a score for species tree estimation.

Chifman and Kubatko (2014) did not use such optimization, and instead relied on phylogenetic invariance to generate the quartet tree. The expected site pattern probabilities (p_k), when arranged into a 16×16 matrix according to the true species tree, has rank ≤ 10, while the rank is >10 if the matrix is arranged according to an incorrect species tree. Note that the rank of a matrix is equal to the number of nonzero singular values or eigenvalues, and that a nonsingular 16×16 matrix has rank 16, but linear relationships among rows or columns reduce its rank. In other words, the site pattern probabilities generated by a species tree should satisfy a number of linear relationships, depending on the assumed mutation or substitution model. The criterion used by the method examines whether the 11th–16th eigenvalues are close to 0. The SVDquartets method for generating the quartet tree is thus similar to the evolutionary parsimony method by Lake (1987), which is also an invariance-based method for quartet data. Evolutionary parsimony makes use of only a few site patterns and is known to be inefficient and sensitive to the details of the substitution model (Jin and Nei 1990). In contrast, SVDquartets uses all 256 site patterns. Its statistical performance has yet to be carefully evaluated. In one simulation study, SVDquartets did not show the expected advantage over competing methods for very short alignments (Chou *et al.* 2015).

The Case of Three Species

In the case of three species (A, B, and C) and three sequences per locus (a, b, and c) the mismatch probability of Equation 5 can be used to estimate the species tree and the parameters using a collection of rooted gene trees. Let (x_1, x_2, x_3) be the numbers of gene trees having topologies G_1, G_2, and G_3. Their probabilities given the species tree are $(p_1, p_2, p_3) = (1 - P_{SG}, \frac{1}{2}P_{SG}, \frac{1}{2}P_{SG})$, with $p_1 > p_2 = p_3$, and with G_1 to be the gene tree that matches the species tree. The likelihood is then the trinomial probability $p_1^{x_1} p_2^{x_2} p_3^{x_3}$. Maximizing this likelihood simply gives the most common gene tree as the estimate of the species tree, with the internal branch length (in coalescent units) given by Equation 5. While the MSC model involves at least four parameters (τ_{AB}, τ_{ABC}, θ_{AB}, θ_{ABC}) (Figure 3), use of the gene tree topologies identifies only one. With the gene tree topologies given as data, this appears to be the only sensible solution, and it is the solution by the rooted triples method of Ewing *et al.* (2008) and the MP-EST method of Liu *et al.* (2010a) in this case.

When the gene trees are unknown and reconstructed using phylogenetic methods, two issues arise. First, if two or more gene trees are equally best at a locus, that locus is discarded. Second, phylogenetic errors of gene tree reconstruction inflate the mismatch probability, so that the probability that the species tree differs from the estimated gene tree ($\hat{G}$) is $P_{S\hat{G}} > P_{SG}$ (Yang 2002). For example, for the human-chimpanzee-gorilla trio, $P_{S\hat{G}} \approx 0.4$ for loci of 500 bp, while $P_{SG} \approx 0.3$ (Burgess and Yang 2008; Scally *et al.* 2012). The inflated species tree-gene tree mismatch erodes our chance of inferring the true species tree, and furthermore, the internal branch length in the species tree (in coalescent units) will be inconsistently estimated. We thus have a highly unusual estimation problem, in which the species tree is identifiable and consistently estimated, the internal branch length is identifiable but inconsistently estimated, while the other three parameters of the MSC model are unidentifiable and inestimable.

Given that the species tree (S) is consistently estimated by the method, it is of interest to know how fast the estimation error approaches 0 when the amount of data (the number of loci) increases. Liu *et al.* (2010a) provided a proof that the error drops to 0 at least at the rate of $1/L$. This result may be strengthened. For the case of three species, the estimated species tree is correct if and only if $x_1 > x_2$ and $x_1 > x_3$, where x_1, x_2, x_3 are the counts of gene trees G_1, G_2, and G_3, with G_1 being the gene tree that matches the species tree. Let (p_1, p_2, p_3) be the probabilities for those gene trees, with $p_2 = p_3$. We also consider x_0 as the number of loci with ties and p_0 the probability of ties. From the fact that (x_0, x_1, x_2, x_3) have a multinomial distribution, the species tree estimation error may be approximated by

$$P_{\text{error}} = 1 - \Pr\{x_1 > x_2, x_1 > x_3\} \approx 1 - \Phi\left(\frac{(p_1 - p_2)\sqrt{L - \frac{1}{p_1 - p_2}} - \sqrt{\frac{p_2}{\pi}}}{\sqrt{p_1 + \left(1 - \frac{1}{\pi}\right)p_2}}\right), \quad (8)$$

where $\Phi(\cdot)$ is the cumulative distribution function (CDF) for the standard normal distribution $N(0, 1)$ (Yang 1996). The approximation assumes $L \gg 1/p_2$. For very large L, Equation 8 is approximately $\Phi(-a\sqrt{L})$, which approaches 0 much

faster than $1/L$. For example, with parameters $\tau_{ABC} = 0.06$, $\tau_{AB} = 0.05$, $\theta_{ABC} = \theta_{AB} = 0.02$ for the species tree of Figure 3A, and with $n = 1000$ sites per locus, we have, through simulation, $p_0 = 0.0197$ for the proportion of loci with ties, $p_1 = 0.6915$ for the matching gene tree, and $p_2 = p_3 = 0.1444$ for the mismatching gene trees. Then the error probability P_{error} is 0.0296, 0.0023, and 0.000034 for 10, 20, and 40 loci, respectively, according to simulation, while Equation 8 gives 0.0398, 0.0030, and 0.000020. The discrepancy appears quite large, relative to P_{error} itself. Better approximations will be desirable. Note also that for those parameter values, Equation 5 gives the species tree-gene tree mismatch probability as $P_{SG} = \frac{2}{3}e^{-1} = 0.24525$, while with 1000 sites per locus, the mismatch probability for the estimated gene tree is $P_{S\hat{G}} = 0.3010$ by simulation (with ties for gene trees at the locus broken evenly). Phylogenetic errors cause considerable inflation of the mismatch probability.

The Anomaly Zone

Degnan and Salter (2005) and Degnan and Rosenberg (2006) derived the probabilities for gene tree topologies under the MSC model, and highlighted the fact that the most probable rooted gene tree topology may not match the species tree when the species tree has short internal branches and large ancestral populations. Thus the simple majority-vote method of using the most commonly observed gene tree as the estimate of the species tree will be statistically inconsistent and will converge in probability to a wrong species tree when the number of loci increases. The species tree and parameters that lead to such *anomalous gene trees* are said to be in the *anomaly zone* (Degnan and Rosenberg 2006). In other words, the anomaly zone is the inconsistency zone for the majority-vote method. Computer simulation (Kubatko and Degnan 2007) and mathematical analysis (Roch and Steel 2015) suggest that concatenation may similarly be inconsistent in certain regions of the parameter space; concatenation has its own anomaly or inconsistency zone.

The case of four species is simple and illuminating (Figure 1). Given the asymmetrical species tree $S_1 = (((AB)C)D)$, the probabilities of the gene trees $G_1 = (((ab)c)d)$ and $G_2 = ((ab)(cd))$ for four sequences (a, b, c, d) under the MSC model are

$$\begin{aligned} P(G_1) &= 1 - \frac{2}{3}e^{-x} - \frac{2}{3}e^{-y} + \frac{1}{3}e^{-(x+y)} + \frac{1}{18}e^{-(3x+y)}, \\ P(G_2) &= \frac{1}{3}e^{-x} - \frac{1}{6}e^{-(x+y)} - \frac{1}{18}e^{-(3x+y)}, \end{aligned} \tag{9}$$

where $x = 2(\tau_{ABCD} - \tau_{ABC})/\theta_{ABC}$ and $y = 2(\tau_{ABC} - \tau_{AB})/\theta_{AB}$ are the lengths of the deeper and shallower internal branches (in coalescent units) in the species tree (Rosenberg 2002; Degnan and Rosenberg 2006). Now if $x \to 0$ and $y \to 0$, we have $P(G_1) \to 1/18$ and $P(G_2) \to 2/18$. When the internal branches on the species tree disappear and the species tree becomes a star tree, all three coalescent events for the four sequences will occur in the ancestral species *ABCD*, and the process is simply the single-population coalescent (Figure 2). As the asymmetrical gene tree G_1 is compatible with only one labeled history while the symmetrical gene tree G_2 is compatible with two labeled histories (G_{2a} and G_{2b}), we have $P(G_1) = 1/18$ and $P(G_2) = 2/18$. Now if x and y are nonzero but very small, $P(G_1) < P(G_2)$ may still hold even if $P(G_1) > P(G_2)/2$. Degnan and Salter (2005) and Degnan and Rosenberg (2006) designed algorithms for determining such boundaries in the parameter space (values of x and y). When $P(G_1) < P(G_2)$, the majority-vote estimate of the species tree will be inconsistent and will converge to the mismatching topology of G_2.

Thus the coalescent process favors symmetrical trees, and the effect is greater for larger trees. The number of compatible labeled histories is 1 for a completely asymmetrical rooted tree. For any other rooted tree, this number is given as the product of $\binom{x+y}{x}$ over the internal nodes on the rooted tree, where x and y are the numbers of descendent internal nodes on the left and right part of the internal node, respectively (Yang and Rannala 2014). For example, for the rooted tree G_2 of Figure 1, $x = y = 1$ for the root node and $x = y = 0$ for the two other internal nodes, so that the number of compatible labeled histories is $\binom{1+1}{1} = 2$, with the convention $\binom{k}{0} = \binom{0}{0} = 1$. An alternative method to calculate this number, suggested by one of the referees, is to divide $(n-1)!$ by the product of $n(k)-1$ over all the interior nodes k of the tree, where n is the number of tips on the tree and $n(k)$ is the number of tips below node k. If there are eight species (and sequences) and if the true species tree is completely asymmetrical with the six internal branch lengths nearly 0, the completely asymmetrical matching gene tree has only about 1/80 the probability of the completely symmetrical mismatching gene tree (Table 1). It may thus be easier to find anomaly with larger species trees, or, as Rosenberg and Tao (2008) put it, the anomaly zone seems to expand when the number of species increases. There is no anomaly zone for the case of three species. For four species, the anomaly zone exists for the asymmetrical species tree only. For five or more species, it exists for any species tree (Degnan and Rosenberg 2006).

Figure 7 shows the behavior of species tree estimation in the anomaly zone for the case of four species. Gene trees and sequence alignments at multiple loci are simulated using MCcoal, which is part of the bpp package (Yang 2015), using the species tree, $(((A, B), C), D)$, with $\tau_{AB} = 0.01$, $\tau_{ABC} = 0.011$, and $\tau_{ABCD} = 0.012$, and with $\theta = 0.05$ for all species. Those parameter values give $P(G_1) = 0.0740$ and $P(G_2) = 0.1191$ (Equation 9), so the species tree is in the anomaly zone. The JC model (Jukes and Cantor 1969) is used for both simulation and analysis. Each replicate dataset consists of L loci, of $n = 1000$ sites, and the number of replicates is $R = 100$. We consider three methods: the majority vote, concatenation, and BI (bpp) (Yang and Rannala 2014). With the majority vote, baseml (which is part of the paml package, Yang 2007) is used to construct the rooted gene tree at each locus under the molecular clock, and the most common gene

Table 1 The number of labeled histories for the most balanced rooted tree for a given number of taxa

Number of species (*s*)	Total number of rooted trees (*T*)	Total number of labeled histories (*H*)	Number of labeled histories for the most balanced rooted tree (*L*)
3	1	1	1
4	15	18	2
5	105	180	3
6	945	2,700	6
7	10,395	56,700	20
8	135,135	1,587,600	80
9	2,027,025	57,153,600	210
10	34,459,425	2,571,912,000	630

In the single-population coalescent, the labeled histories have uniform probabilities, so that the asymmetrical gene tree has probability 1/*H*, while any other gene tree has the probability *L*/*H*, where *L* is the number of compatible labeled histories for the rooted tree.

tree is taken as the estimate of the species tree. Loci with ties (that is, with two or more binary rooted trees to be equally best) are discarded. With concatenation, the concatenated sequences are analyzed as one supergene using BASEML under the clock and the resulting rooted gene tree is taken as the estimate. Both majority-vote and concatenation are inconsistent so that when the number of loci (L) increases, the probability of recovering the true species tree approaches 0 (Figure 7). Both methods converge to the incorrect symmetrical species tree, $((A, B), (C, D))$.

In the BPP analysis (Analysis A01 in Yang 2015), the uniform prior is assigned on the rooted trees (Prior 1 in Yang and Rannala 2014), while gamma priors with shape parameter 2 and with the means to be the true values are assigned to the parameters: $\tau_{ABCD} \sim G(2, 2/0.012)$ for the age of the root and $\theta \sim G(2, 40)$. The shape parameter of 2 means the gamma priors are fairly diffuse. As the Bayesian method of model selection is consistent, the probability of recovering the true species tree approaches 1 when the number of loci increases.

The issue of anomalous gene trees has been discussed extensively. Degnan and Rosenberg (2006) wrote that "the use of multiple genomic regions for species tree inference is subject to a surprising new difficulty, the problem of "anomalous gene trees."" We emphasize that the anomaly zone is not an intrinsic difficulty of the estimation problem. It exists because of the heuristic nature of the majority-vote method, and vanishes if the data are analyzed using full likelihood methods. If the data consist of a collection of gene tree topologies, the likelihood function is given in Equation 6. As discussed earlier, this ML method is consistent. Intuitively if the true species tree is asymmetrical and if the two internal branches are very short, one expects the symmetrical mismatching gene tree G_2 to be nearly twice as frequent as the asymmetrical matching gene tree G_1 (Figure 1). Thus the slightly higher proportion of G_2 than G_1 may be seen to be compatible with an asymmetrical species tree (with short internal branch lengths) and may not be evidence against it.

The analysis of Degnan and Salter (2005); Degnan and Rosenberg (2006) assumes that the gene trees are known without error. In real data analysis, gene trees reconstructed using phylogenetic methods may involve substantial errors and uncertainties. Huang and Knowles (2009) evaluated the impact of phylogenetic reconstruction errors on the anomaly zone and observed an expansion of the anomaly zone due to phylogenetic errors: there is a larger space of species tree parameters within which anomalous gene trees are observed when the gene trees are estimated than when they are given. Nevertheless, Huang and Knowles (2009) suggest that in the anomaly zone, phylogenetic analysis of typical datasets tends to return unresolved gene trees rather than resolved incorrect trees, so that the difficulty posed by anomalous gene trees to practical phylogenetic analysis may be limited. At any rate, ML and BI methods for analyzing real datasets are based on the sequence likelihood (Equations 2 and 3), so that reliable inference of the species tree is possible even if the estimated gene tree at every locus is unreliable and involves substantial uncertainties.

Can More Data Make Things Worse?

Liu *et al.* (2015, Figure 4) reported a simulation study showing that adding "weak" genes (short genes with few sites) to a set of "strong" genes (long genes with many sites) might cause MP-EST to perform worse. Here we confirm this counterintuitive behavior by conducting two simulation experiments, using species trees of five and three species, respectively. We find that the effect is subtle. We then discuss two analogous simple examples to explain the result, offering an argument that it does not occur with full likelihood methods.

Our first simulation largely follows Liu *et al.* (2015). The species tree is ((((*A*: 0.002, *B*: 0.002): 0.002, (*C*: 0.002, *D*: 0.002): 0.002): 0.002, *E*: 0.006): 0.01, *F*: 0.016), with $\theta = 0.008$ for all populations (Figure 8A). Species *F* is used as the outgroup. There are two kinds of loci: strong genes with 1000 bp and weak genes with 100 bp. Gene trees and sequence alignments were generated using MCCOAL (Yang 2015). The mutation model used in data generation and analysis is JC (Jukes and Cantor 1969). For the MP-EST analysis, PHYML (Guindon and Gascuel 2003) was used to infer the unrooted ML tree for each locus, with the sequence from species *F* used to root the tree. The rooted gene trees for the five ingroup species are then processed using MP-EST (Liu *et al.* 2010a) to infer the species tree. We consider the use of strong genes only, weak genes only, and a mixture of 20 strong genes

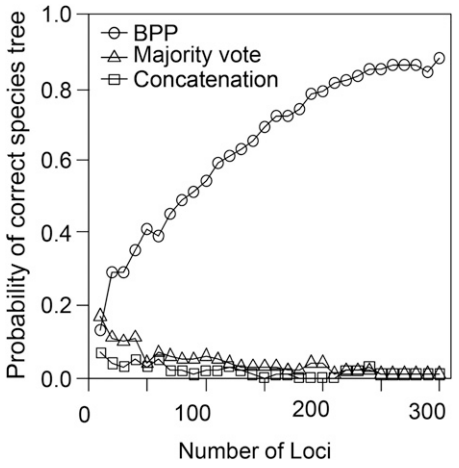

Figure 7 Species tree inference in the anomaly zone. The probability of inferring the correct species tree by majority-vote, concatenation, and BI (BPP), plotted against the number of loci (L).

followed by addition of weak genes (Figure 8). MP-EST performed much worse in the simulation of Liu *et al.* (2015, Figure 4) than here, as those authors used bootstrap gene trees while we used ML gene trees (L. Liu, personal communication). At any rate, Figure 8 shows a similar pattern to that in Liu *et al.* (2015, Figure 4): adding weak genes to the dataset of 20 strong genes led to deteriorated performance by MP-EST. The effect, however, is much weaker than in Liu *et al.* (2015). As the method is consistent even with weak genes only, the probability of correct species tree will eventually approach 1 when the number of loci $L \to \infty$.

We also included BPP for comparison. As BPP assumes the clock and works with rooted trees, we removed the sequence from species F at every locus. We assigned a uniform prior on rooted trees (Yang and Rannala 2014), and gamma priors on parameters in the MSC: $\tau_{ABCDE} \sim$ G(2, 2/0.006) with mean 0.006 for the root age, and $\theta \sim$ G(2, 250) with mean 0.008 for θ. BPP performed much better than MP-EST, with large differences when the data include only weak genes or only a few strong genes. BPP did not show a deterioration of performance upon adding weak genes, but this is hard to discern as the probability of recovering the true species tree is already 98% with 20 strong genes alone.

Our second simulation uses the species tree $((AB)C)$ for three species (Figure 3). Note that for three species, many summary methods such as MP-EST (Liu *et al.* 2010a), MDC (Maddison 1997), and STAR (Liu *et al.* 2009) are equivalent, so that our results for MP-EST apply to the other equivalent methods as well. We used $\tau_{AB} = 0.05$, $\tau_{ABC} = 0.06$, and $\theta_{AB} = \theta_{ABC} = 0.02$. We again simulated strong genes (1000 sites) and weak genes (50 sites). Gene tree reconstruction under the clock and the JC model for three sequences is tractable analytically (Yang 2000), so that we can analyze a huge number of replicates for the MP-EST method without the need to use the MP-EST program. Each alignment of three sequences (a, b, c) can be summarized as counts (n_0, n_1, n_2, n_3, n_4) of five site patterns: *xxx*, *xxy*, *yxx*, *xyx*, and *xyz*, where x, y, and z are any distinct nucleotides. We take the ML tree as $((ab)c)$ if and only if $n_1 > \max(n_2, n_3)$. If two trees are equally best, we discard the locus. The most common gene tree topology is then the estimated species tree. Ties for species trees are broken evenly: if two species trees have the same maximum number of matching gene trees, each is given 50%, and if all three species trees are equally good, each is given 1/3. The performance deterioration with MP-EST upon addition of weak genes is real but the effect is so small that a huge number of replicates are necessary to demonstrate its presence. With five strong genes, the probability of recovering the correct species tree by MP-EST is 0.8883, and this probability drops to 0.8867–0.8875 when 1–3 weak genes are added to the dataset, and rises back to 0.8883 when four weak genes are included. With 10 strong genes, the probability is 0.9705, and it drops to 0.9690–0.9699 when 1–7 weak genes are added to the dataset. The effect exists but is too small to have any biological significance.

For comparison, we included ML (3S) (Yang 2002) and BI (BPP). Note that 3S optimizes the parameters (Θ) while BPP averages over them, so that the two methods are not equivalent even for this simple case of three species. For BPP, we used the uniform prior on the three species trees, and gamma priors on parameters: $\tau_0 =$ G(2, 2/0.06) for the root age, and $\theta =$ G(2, 100) for θs. The two methods had very similar performance (Figure 9). As expected neither 3S nor BPP shows the counterintuitive behavior. When only weak genes or a few strong genes are analyzed, 3S and BPP recovered the true species tree with higher probabilities than MP-EST.

We note that the performance differences between the summary method (MP-EST) and the full-likelihood methods (3S and BPP) are smaller for the three-species case (Figure 9) than for the five-species case (Figure 8). This may be expected from the simplicity of the three-species case. Indeed as far as the point estimate of the species tree is concerned, MP-EST, ML (3S) and BI (BPP) are expected to be equivalent if the data consist of only one locus. Let p_0 be the probability that a locus shows a tie (with two or more gene trees being equally best), p_1 be the probability of the matching gene tree, and p_2 and p_3 be the probabilities of the two mismatching gene trees (Figure 3). For the case of Figure 9A, these are $p_0 = 0.1989$, $p_1 = 0.4013$, $p_2 = p_3 = 0.2000$ for the weak genes (50 bp) and $p_0 = 0.0197$, $p_1 = 0.6915$, $p_2 = p_3 = 0.01444$ for the strong genes (1000 bp). The probability of recovering the correct species tree for one locus is $p_1 + p_0/3$. The three methods are not the same when two or more loci are analyzed. For MP-EST, the probability of recovering the species tree with two loci can be calculated, by considering the different outcomes of gene tree reconstruction at the two loci, as $p_1^2 + 2p_0p_1 + 2p_1(p_2 + p_3)/2 + p_0^2/3 = p_1(1 + p_0) + p_0^2/3$. If the outcome is 11 (both loci producing the matching gene tree G_1), or 01 or 10, the correct species tree is recovered. If the outcome is 12, 21, 13, or 31, the species tree is half correct, and so on. If $p_0 \approx 0$ (as in the case of strong genes), MP-EST recovers the correct species tree with nearly the same probability ($\approx p_1$) for one locus and two loci (Figure 9A). This is not the case for 3S and BPP.

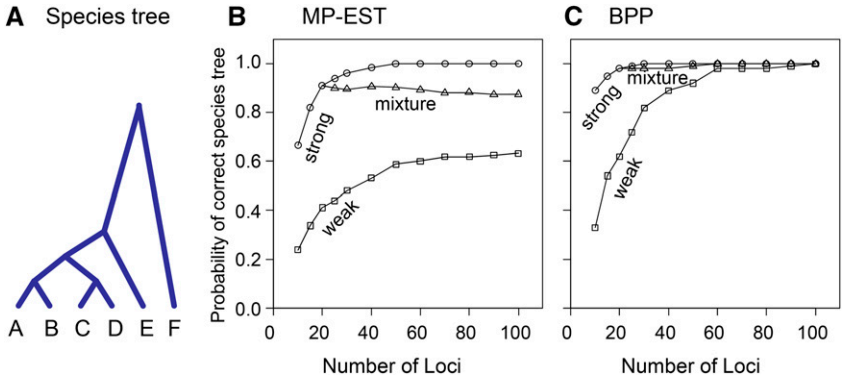

Figure 8 More data for worse performance for five species? The probability of inferring (A) the correct species tree for five species by (B) MP-EST and (C) BPP using weak genes only, strong genes only, and a mixture of 20 strong genes plus a number of weak genes. The divergence times on the species tree are $\tau_{AB} = \tau_{CD} = 0.002$, $\tau_{ABCD} = 0.004$, $\tau_{ABCDE} = 0.006$, and $\tau_{ABCDEF} = 0.016$, with $\theta = 0.008$ for all populations. The number of replicates is 1000 for MP-EST and 100 for BPP.

To gain insights into the counterintuitive behavior of MP-EST in the cases of three species (Figure 9A), we study two small analogous problems. The first is a binomial example of coin toss in which a coin is tossed $n = 100$ times to estimate the probability of heads, p. If the observation is x heads, the estimate $\hat{p} = x/n$ has variance $p(1-p)/n$. Now we add a second experiment, with x' heads out of $n' = 10$ tosses. The second experiment alone would give us the estimate $\hat{p}' = x'/n'$, with variance $p(1-p)/n'$. A simple method of combining the two experiments is to take the average $\hat{p}_{\text{dumb}} = \frac{1}{2}\left(\frac{x}{n} + \frac{x'}{n'}\right)$, which has the variance $\frac{1}{4}p(1-p)\left(\frac{1}{n} + \frac{1}{n'}\right)$ and effective sample size $4/\left(\frac{1}{n} + \frac{1}{n'}\right) = 36.4$, which is even smaller than $n = 100$. Thus adding data (from the second experiment) made things worse, because we ignore the fact that the two experiments have very different precisions. A better method is to use the likelihood function to combine the two experiments:

$$\begin{aligned} L(p) &= \left[p^x(1-p)^{n-x}\right] \times \left[p^{x'}(1-p)^{n'-x'}\right] \\ &= p^{x+x'}(1-p)^{(n+n')-(x+x')}. \end{aligned} \tag{10}$$

The MLE is then $\hat{p}_{\text{smart}} = (x + x')/(n + n')$, with an effective sample size of $n + n'$, as expected. The MLE is a weighted average of $\hat{p}$ and $\hat{p}'$, with the precision (the reciprocal of the variance) used as weights.

The second analogous example is similar to the MP-EST estimation of the species tree, but with two instead of three possible trees to consider. The gene tree that matches the species tree will be called a "good" gene tree, while the mismatching gene tree a "bad" one. Suppose a strong gene and a weak gene produce a good gene tree with probability p and p', respectively, with $p > p' > \frac{1}{2}$. With n strong genes only, the number of good gene trees $x \sim \text{Bin}(n, p)$, so that the correct species tree is recovered with probability

$$P(n) = \Pr\left\{x > \frac{n}{2}\right\} \approx \Phi\left(\frac{np - \frac{n}{2}}{\sqrt{np(1-p)}}\right), \tag{11}$$

if n is large. Now consider adding n' weak genes, which will produce x' good gene trees, with $x' \sim \text{Bin}(n', p')$. A simple method for combining the data recovers the correct species tree if and only if more than half of the gene trees are good. This occurs with probability

$$\begin{aligned} P_{\text{dumb}}(n, n') &= \Pr\left\{x + x' > \frac{n + n'}{2}\right\} \\ &\approx \Phi\left(\frac{np + n'p' - \frac{n+n'}{2}}{\sqrt{np(1-p) + n'p'(1-p')}}\right), \end{aligned} \tag{12}$$

if both n and n' are large. Note that x and x' are independent binomial variables, so that $x + x'$ has mean $np + n'p'$ and variance $np(1-np) + n'p'\ (1-p')$.

Is it possible for $P_{\text{dumb}}(n, n') < P(n)$ for certain values of n and n'? The answer is Yes. Consider for example $p = 0.6$ for strong genes and $p' = 0.525$ for weak genes. Thus $P(50) = 0.9255$ with $n = 50$ strong genes only, and $P_{\text{dumb}}(n, n') < P(n)$ if $1 \le n' < 432$ (Figure 10). Adding weak genes made things worse, and performance did not recover to the level of $n = 50$ strong genes until 432 weak genes were added. The simple method (Equation 12) does not account for the fact that the gene trees from the weak genes involve larger sampling errors than those from the strong genes.

Liu *et al.* (2015) suggest that "adding weak genes may actually reduce the performance of species tree estimation methods, negating the old adage that "more data is always better."" The authors further suggest that "[a]n important rule of thumb that has emerged from both simulation and empirical studies is that species trees are only as good as the gene trees on which they are built. This maxim applies both to two-step species tree methods, in which gene trees are used as input data, and to single-step approaches, such as Bayesian methods, in which gene and species trees are estimated simultaneously." Those statements need to be qualified. Our analysis above suggests that the counterintuitive result of performance deterioration upon addition of weak genes may be more easily explained by the inefficient use of information in the data by the summary methods. For full likelihood methods, the likelihood calculation on the sequence alignment should automatically accommodate the fact that a strong gene with more sites is more informative about the gene tree than a weak gene with fewer sites. The counterintuitive behavior will then not occur. We expect similar patterns if the different loci have different mutation rates and different information contents, if the rate variation among loci is accommodated appropriately in the likelihood model. Thus species trees can be good even if all gene trees are bad. In the

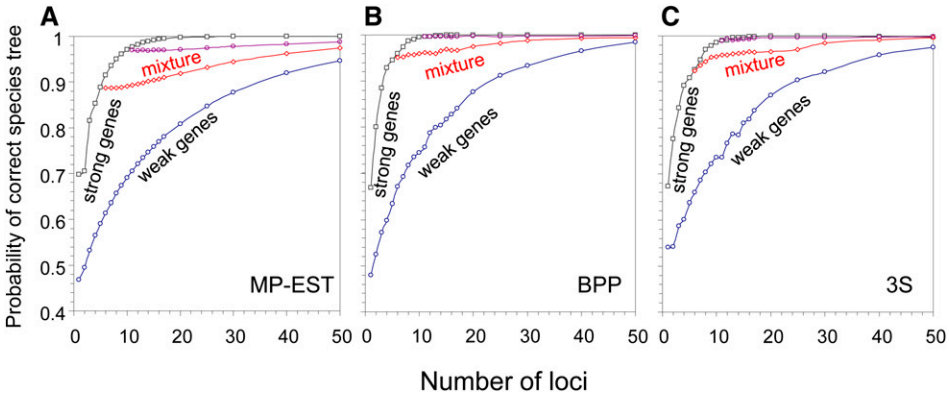

Figure 9 More data for worse performance for three species? The probability of inferring the correct species tree for three species by MP-EST, ML (3S), and BI (BPP) using weak genes only, strong genes only, and a mixture of 5 or 10 strong genes plus a number of weak genes. The true species tree is the one in Figure 3A, with $\tau_{ABC} = 0.06$ and $\tau_{AB} = 0.05$, and with $\theta = 0.02$ for all populations. The number of replicates is 1000 for 3S and BPP, and ranges from 10^5 to 10^7 for MP-EST.

extreme, adding weak genes which have only two sequences (and thus no phylogenetic information) should also help, by providing information on parameters in the model.

The performance deterioration for MP-EST upon adding the weak genes is so small that the effect does not have any biological impact. Thus the conclusions to be drawn from this analysis are largely conceptual, in that one should not be surprised at counterintuitive behaviors when heuristic methods are used. Note that large performance differences are possible between likelihood methods and summary methods when the data are not very informative (Figure 8).

Singularity on the Likelihood Surface

As discussed earlier, given that the data are a collection of rooted gene trees with node ages (branch lengths on ultrametric gene trees), the ML estimate of the species tree under the assumption of equal population size for all species is given by the maximum tree algorithm (Liu *et al.* 2010b; see also Mossel and Roch 2010). For the more general problem with different species having independent population size parameters (θs), however, Liu *et al.* (2010a) pointed out that the likelihood may become infinite for certain species tree and parameter values. The likelihood function in this case is the MSC density of gene trees and coalescent times given the species tree, or $f(G_i, \boldsymbol{t}_i \mid S, \Theta)$ in Equation 2 (Rannala and Yang 2003). The observation motivated the development of the summary method MP-EST, which ignores branch lengths and uses gene tree topologies only. Here we illustrate how such singularity can occur, and point out that it is not a problem if the sequence data (rather than the estimated gene trees) are analyzed using full likelihood methods. We also discuss ways of avoiding the singularity in the summary methods (Yang 2014, p.338–9).

Suppose that the data consist of one "observed" gene tree, which is G_{1a} of Figure 3, with branch lengths (node ages) t_0 and t_1, and consider the calculation of the likelihood for the species tree S of Figure 3a, with parameters $\Theta = \{\theta_{AB}, \theta_{ABC}, \tau_{AB}, \tau_{ABC}\}$. The likelihood is given as the product of the two exponential densities for the two coalescent events (Rannala and Yang 2003):

$$f(G_{1a}, t_0, t_1 | S, \Theta) = \frac{2}{\theta_{AB}} e^{-\frac{2}{\theta_{AB}}(t_1 - \tau_{AB})} \times \frac{2}{\theta_{ABC}} e^{-\frac{2}{\theta_{ABC}}(t_0 - \tau_{ABC})},$$
$$\text{with } \tau_{AB} < t_1 < \tau_{ABC} < t_0. \quad (13)$$

Note that this is a function of θ_{AB}, θ_{ABC}, τ_{AB}, and τ_{ABC}, with t_0 and t_1 fixed. Now let $\tau_{AB} \to t_1$, $\tau_{ABC} \to t_1$, and let $\theta_{AB} = c\,(t_1 - \tau_{AB}) \to 0$, with $c > 0$ to be a constant. Then the first term $\frac{2}{\theta_{AB}} e^{-\frac{2}{\theta_{AB}}(t_1 - \tau_{AB})} = \frac{2}{\theta_{AB}} e^{-2/c} \to \infty$, and the likelihood function becomes infinite. In other words, when we shrink the internal branch length $(\tau_{ABC} - \tau_{AB})$ in the species tree to zero, collapse the two internal nodes onto the coalescent event, and increase the coalescent rate $(2/\theta_{AB})$ to infinity, the likelihood becomes infinite, with the parameter estimates $\hat{\tau}_{AB} = \hat{\tau}_{ABC} = t_1$ and $\hat{\theta}_{AB} = 0$ (with $\hat{\theta}_{ABC}$ indefinite). There are thus rays of singularity on the likelihood surface. In this case, there is no singularity if both coalescent events are assumed to occur in the common ancestor *ABC*. In other words, the species tree compatible with the given gene tree has infinite likelihood while the mismatching species trees have finite likelihood. In the general case of more than three species and sequences, multiple species trees can have infinite likelihood (Liu *et al.* 2010a). Singularity can occur for an arbitrary species tree and multiple gene trees as long as one can collapse an internal branch on the species tree onto a coalescent event on a gene tree.

Here we point out that such singularity does not occur when full likelihood methods (ML or BI) are applied to sequence data, with the likelihood calculated using the sequence alignments (Equations 2 and 3). Approaches using sequences also have the advantage of accommodating uncertainties in the gene trees and branch lengths. If the gene trees with branch lengths are treated as observed data, singularity disappears if one uses the Bayesian method, as the prior shrinks the parameters away from extreme values (such as $\theta_{AB} = 0$, and $\tau_{AB} = \tau_{ABC}$), and if one uses the ML method under the assumption that all populations have the same size (same θ) (Liu *et al.* 2010a), as in the maximum tree method or the STEM program (Kubatko *et al.* 2009).

Perspectives and Practical Data Analysis

We have discussed a suite of counterintuitive behaviors of summary methods for species tree estimation, such as

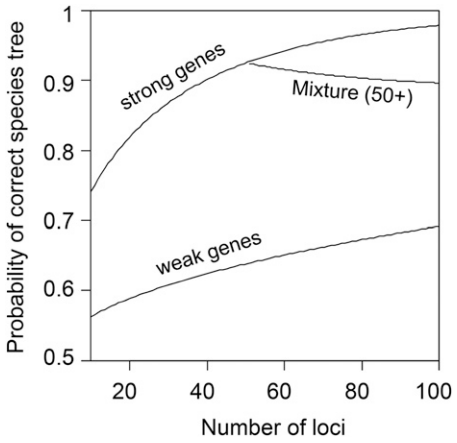

Figure 10 More data for worse performance for a toy example of binary data? The species tree is estimated using a summary method that mimics MP-EST.

unidentifiability, inconsistency (anomaly zone), singularity on the likelihood surface, and deteriorated performance upon inclusion of more data. Those behaviors are due to the information loss when the summary methods use estimated gene tree topologies only, ignoring information in the branch lengths, and due to their failure to account for the uncertainties in the estimated gene trees. They do not occur when the sequence data are analyzed using full likelihood methods. Our analyses demonstrate that large performance differences may exist between full likelihood methods and summary methods even if both are based on the MSC.

While other factors may also cause gene tree-species tree conflicts (Maddison 1997; Nichols 2001; Szollosi *et al.* 2014), we have taken an idealized viewpoint assuming that the MSC is the true data-generating model. A number of challenges exist with current implementations of the Bayesian inference methods. A major problem is the intensive computation involved and the inefficient mixing of the transmodel MCMC algorithms used by Bayesian programs, although improvements are being made (*e.g.*, Rannala and Yang 2016).

We suggest that the utility of summary *vs.* full-likelihood methods will depend on the nature of the species tree estimation problem. For easy problems with long internal branches in the species tree and little incomplete lineage sorting, different methods are likely to produce the same results, and simple methods such as concatenation may have even higher statistical efficiency than coalescent-based full likelihood methods. For shallow species phylogenies, characterized by recent divergences and short internal branches (as occurs in radiative speciation), full likelihood methods may have a big advantage over summary methods or simple methods such as concatenation. Genomic datasets are being generated from a variety of species, such as mosquitos (Fontaine *et al.* 2015), butterflies (Martin *et al.* 2013), hares (Melo-Ferreira *et al.* 2012), bears (Liu *et al.* 2014), and gibbons (Carbone *et al.* 2014). As the species are closely related, the molecular clock holds approximately and can be used to root the tree and to provide information about the node ages (coalescent times). The high sequence similarity and low phylogenetic information content at each locus may be problematic to summary methods that are sensitive to phylogenetic reconstruction errors but should be ideal for full-likelihood methods (Ogilvie *et al.* 2016; Rannala and Yang 2016).

Deep species phylogenies characterized by ancient rapid divergences pose the greatest challenge. Two strategies are possible to account for the violation of the molecular clock in deep phylogenies: (i) the use of relaxed-clock models in Bayesian analysis to root the tree and to extract information about coalescent times in gene-tree branch lengths, and (ii) the use of outgroups to root the tree, ignoring information in branch lengths (or coalescent times) in gene trees, as in ASTRAL (Mirarab *et al.* 2014; Mirarab and Warnow 2015) and NJ_{st} (Liu and Yu 2011). In theory relaxed-clock models allow the inference of the root and the node ages in both the species tree and the gene trees. However, several difficulties may arise. First, current relaxed-clock models of rate drift, especially when applied to data of multiple genetic loci, may be highly unrealistic (dos Reis *et al.* 2016). Relaxed-clock models developed for dating species divergences (Thorne *et al.* 1998; Drummond *et al.* 2006; Rannala and Yang 2007) allow the substitution rate to drift over time or among branches, but they do not appear appropriate for analysis under the MSC. A simple modification may be to assign rates at a locus to branches of the species tree (rather than to branches of the gene tree), so that gene-tree branches residing in the same species share the same rate. A gene tree branch may be broken into several segments with different rates, and the branch length or the total amount of evolution along the branch is calculated by summing over the segments. More worrying are the impacts of among-loci heterogeneity in the substitution rate and in the process of substitution rate drift among branches (Zhu *et al.* 2015). Current models assume that rates drift over time independently among loci. This assumption is unrealistic but allows rates at different loci to be treated as independent realizations of the same process and be teased apart from the species divergence times. Allowing for rate correlation among loci or lineage effects may render the model unidentifiable, since both the lineage rates and the species divergence times have genome-wide effects and are thus seriously confounded. Relaxed-clock rooting appears to be less reliable than outgroup rooting when the clock is seriously violated. It is an open question how Bayesian relaxed-clock models compare with summary methods that use unrooted gene trees with outgroup rooting in inferring deep phylogenies.

Acknowledgments

We thank Liang Liu for extensive discussions, especially concerning MP-EST. We are grateful to Scott Edwards, Laura Kubatko, Liang Liu, Mike Steel, and three anonymous referees for many constructive comments on earlier versions of the manuscript.

Literature Cited

Allman, E. S., J. H. Degnan, and J. A. Rhodes, 2011 Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. J. Math. Biol. 62: 833–862.

Bryant, D., R. Bouckaert, J. Felsenstein, N. A. Rosenberg, A. RoyChoudhury *et al.*, 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 Z. Yang, 2008 Estimation of hominoid ancestral population sizes under Bayesian coalescent models incorporating mutation rate variation and sequencing errors. Mol. Biol. Evol. 25: 1979–1994.

Carbone, L., R. A. Harris, S. Gnerre, K. R. Veeramah, B. Lorente-Galdos *et al.*, 2014 Gibbon genome and the fast karyotype evolution of small apes. Nature 513: 195–201.

Chen, F.-C., and W.-H. Li, 2001 Genomic divergences between humans and other Hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am. J. Hum. Genet. 68: 444–456.

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

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

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

Dalquen, D., T. Zhu, and Z. Yang, 2016 Maximum likelihood implementation of an isolation-with-migration model for three species. Syst. Biol. DOI: 10.1093/sysbio/syw063.

Dawid, A. P., 2011 Posterior model probabilities, pp. 607–630 in *Philosophy of Statistics*, edited by P. S. Bandyopadhyay, and M. Forster. Elsevier, New York.

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

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

Degnan, J. H., and N. A. Rosenberg, 2009 Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24: 332–340.

dos Reis, M., P. C. J. Donoghue, and Z. Yang, 2016 Bayesian molecular clock dating of species divergences in the genomics era. Nat. Rev. Genet. 17: 71–80.

Drummond, A. J., S. Y. W. Ho, M. J. Phillips, and A. Rambaut, 2006 Relaxed phylogenetics and dating with confidence. PLoS Biol. 4: e88.

Edwards, A. W. F., 1970 Estimation of the branch points of a branching diffusion process (with discussion). J. R. Stat. Soc. B 32: 155–174.

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

Edwards, S. V., W. B. Jennings, and A. M. Shedlock, 2005 Phylogenetics of modern birds in the era of genomics. Proc. Biol. Sci. 272: 979–992.

Edwards, S. V., L. Liu, and D. K. Pearl, 2007 High-resolution species trees without concatenation. Proc. Natl. Acad. Sci. USA 104: 5936–5941.

Edwards, S. V., Z. Xi, A. Janke, B. C. Faircloth, J. E. McCormack *et al.*, 2016 Implementing and testing the multispecies coalescent model: a valuable paradigm for phylogenomics. Mol. Phylogenet. Evol. 94: 447–462.

Ewing, G. B., I. Ebersberger, H. A. Schmidt, and A. von Haeseler, 2008 Rooted triple consensus and anomalous gene trees. BMC Evol. Biol. 8: 118.

Felsenstein, J., 1981 Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17: 368–376.

Felsenstein, J., 1988 Phylogenies from molecular sequences: inference and reliability. Annu. Rev. Genet. 22: 521–565.

Fontaine, M. C., J. B. Pease, A. Steele, R. M. Waterhouse, D. E. Neafsey *et al.*, 2015 Mosquito genomics. Extensive introgression in a malaria vector species complex revealed by phylogenomics. Science 347: 1258524.

Gillespie, J. H., and C. H. Langley, 1979 Are evolutionary rates really variable? J. Mol. Evol. 13: 27–34.

Guindon, S., and O. Gascuel, 2003 A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52: 696–704.

Hein, J., M. H. Schieriup, and C. Wiuf, 2005 *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford University Press, Oxford.

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

Huang, H., and L. L. Knowles, 2009 What is the danger of the anomaly zone for empirical phylogenetics? Syst. Biol. 58: 527–536.

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

Hudson, R. R., 1990 Gene genealogies and the coalescent process, pp. 1–44 in *Oxford Surveys in Evolutionary Biology*, edited by D. J. Futuyma, and J. D. Antonovics. Oxford University Press, New York.

Jin, L., and M. Nei, 1990 Limitations of the evolutionary parsimony method of phylogenetic analysis. Mol. Biol. Evol. 7: 82–102 (erratum: Mol. Biol. Evol. 7: 201).

Jukes, T. H., and C. R. Cantor, 1969 Evolution of protein molecules, pp. 21–123 in *Mammalian Protein Metabolism*, edited by H. N. Munro. Academic Press, New York.

Kingman, J. F. C., 1982a The coalescent. Stochastic Process. Appl. 13: 235–248.

Kingman, J. F. C., 1982b On the genealogy of large populations. J. Appl. Probab. 19A: 27–43.

Kubatko, L. S., and J. H. Degnan, 2007 Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst. Biol. 56: 17–24.

Kubatko, L. S., B. C. Carstens, and L. L. Knowles, 2009 STEM: species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics 25: 971–973.

Lake, J. A., 1987 A rate-independent technique for analysis of nucleic acid sequences: evolutionary parsimony. Mol. Biol. Evol. 4: 167–191.

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

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

Liu, L., 2008 BEST: Bayesian estimation of species trees under the coalescent model. Bioinformatics 24: 2542–2543.

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

Liu, L., and L. Yu, 2011 Estimating species trees from unrooted gene trees. Syst. Biol. 60: 661–667.

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

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

Liu, L., L. Yu, and D. K. Pearl, 2010b Maximum tree: a consistent estimator of the species tree. J. Math. Biol. 60: 95–106.

Liu, L., Z. Xi, S. Wu, C. Davis, and S. V. Edwards, 2015 Estimating phylogenetic trees from genome-scale data. Ann. N. Y. Acad. Sci. 1360: 36–53. doi: 10.1111/nyas.12747.

Liu, S., E. D. Lorenzen, M. Fumagalli, B. Li, K. Harris *et al.*, 2014 Population genomics reveal recent speciation and rapid evolutionary adaptation in polar bears. Cell 157: 785–794.

Lohse, K., R. J. Harrison, and N. H. Barton, 2011 A general method for calculating likelihoods under the coalescent process. Genetics 189: 977–987.

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

Maddison, W. P., and L. L. Knowles, 2006 Inferring phylogeny despite incomplete lineage sorting. Syst. Biol. 55: 21–30.

Mallet, J., N. Besansky, and M. W. Hahn, 2016 How reticulated are species? BioEssays 38: 140–149.

Mallo, D., and D. Posada, 2016 Multilocus inference of species trees and DNA barcoding. Philos. Trans. R. Soc. Lond. B Biol. Sci. 371: 20150335.

Martin, S. H., K. K. Dasmahapatra, N. J. Nadeau, C. Salazar, J. R. Walters *et al.*, 2013 Genome-wide evidence for speciation with gene flow in Heliconius butterflies. Genome Res. 23: 1817–1828.

Melo-Ferreira, J., P. Boursot, M. Carneiro, P. J. Esteves, L. Farelo *et al.*, 2012 Recurrent introgression of mitochondrial DNA among hares (Lepus spp.) revealed by species-tree inference and coalescent simulations. Syst. Biol. 61: 367–381.

Mirarab, S., and T. Warnow, 2015 ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics 31: i44–i152.

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

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

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

Nordborg, M., 2007 Coalescent theory, pp. 843–877 in *Handbook of Statistical Genetics*, edited by D. Balding, M. Bishop, and C. Cannings. Wiley, New York.

Ogilvie, H. A., J. Heled, D. Xie, and A. J. Drummond, 2016 Computational performance and statistical accuracy of *BEAST and comparisons with other methods. Syst. Biol. 65: 381–396.

Pamilo, P., and M. Nei, 1988 Relationships between gene trees and species trees. Mol. Biol. Evol. 5: 568–583.

Pease, J. B., D. C. Haak, M. W. Hahn, and L. C. Moyle, 2016 Phylogenomics reveals three sources of adaptive variation during a rapid radiation. PLoS Biol. 14: e1002379.

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

Rannala, B., and Z. Yang, 2007 Inferring speciation times under an episodic molecular clock. Syst. Biol. 56: 453–466.

Rannala, B., and Z. Yang, 2016 Efficient Bayesian species tree inference under the multispecies coalescent. Syst. Biol.

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

Rosenberg, N. A., 2002 The probability of topological concordance of gene trees and species trees. Theor. Popul. Biol. 61: 225–247.

Rosenberg, N. A., and R. Tao, 2008 Discordance of species trees with their most likely gene trees: the case of five taxa. Syst. Biol. 57: 131–140.

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

Scally, A., J. Y. Dutheil, L. W. Hillier, G. E. Jordan, I. Goodhead *et al.*, 2012 Insights into hominid evolution from the gorilla genome sequence. Nature 483: 169–175.

Springer, M. S., and J. Gatesy, 2016 The gene tree delusion. Mol. Phylogenet. Evol. 94: 1–33.

Steel, M., 2013 Consistency of Bayesian inference of resolved phylogenetic trees. J. Theor. Biol. 336: 246–249.

Szollosi, G. J., E. Tannier, V. Daubin, and B. Boussau, 2014 The inference of gene trees with species trees. Syst. Biol. 64: e42–e62.

Tajima, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics 105: 437–460.

Takahata, N., 1986 An attempt to estimate the effective size of the ancestral species common to two extant species from which homologous genes are sequenced. Genet. Res. 48: 187–190.

Takahata, N., 1989 Gene genealogy in three related populations: consistency probability between gene and population trees. Genetics 122: 957–966.

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

Than, C., and L. Nakhleh, 2009 Species tree inference by minimizing deep coalescences. PLOS Comput. Biol. 5: e1000501.

Than, C. V., and N. A. Rosenberg, 2011 Consistency properties of species tree inference by minimizing deep coalescences. J. Comput. Biol. 18: 1–15.

Thorne, J. L., H. Kishino, and I. S. Painter, 1998 Estimating the rate of evolution of the rate of molecular evolution. Mol. Biol. Evol. 15: 1647–1657.

Turelli, M., J. R. Lipkowitz, and Y. Brandvain, 2014 On the Coyne and Orr-igin of species: effects of intrinsic postzygotic isolation, ecological differentiation, X chromosome size, and sympatry on Drosophila speciation. Evolution 68: 1176–1187.

Wakeley, J., 2009 *Coalescent Theory: An Introduction*. Roberts & Company, Greenwood Village, Colorado.

Wen, D., Y. Yu, M. W. Hahn, and L. Nakhleh, 2016 Reticulate evolutionary history and extensive introgression in mosquito species revealed by phylogenetic network analysis. Mol. Ecol. 25: 2361–2372.

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

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

Yang, Z., 2000 Complexity of the simplest phylogenetic estimation problem. Proc. 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: 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: 854–865.

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

Zhu, T., M. dos Reis, and Z. Yang, 2015 Characterization of the uncertainty of divergence time estimation under relaxed molecular clock models using multiple loci. Syst. Biol. 64: 267–280.

Communicating editor: M. Turelli