JSE Journal of Systematics and Evolution

doi: 10.1111/jse.12160

Review

Coalescent methods for estimating species trees from phylogenomic data

Liang Liu[1,2]*, Shaoyuan Wu[3], and Lili Yu[4]

[1]Department of Statistics, University of Georgia, Athens, GA 30602, USA
[2]Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA
[3]Department of Biochemistry and Molecular Biology, School of Basic Medical Sciences, Tianjin Medical University, Tianjin 300070, China
[4]Department of Biostatistics, Georgia Southern University, Statesboro, GA 02138, USA
*Author for correspondence. E-mail: lliu@uga.edu. Tel.: 706-542-3309. Fax: 706-542-3391.

Received 20 February 2015; Accepted 4 May 2015; Article first published online 19 June 2015

Abstract Genome-scale sequence data have become increasingly available in the phylogenetic studies for understanding the evolutionary histories of species. However, it is challenging to develop probabilistic models to account for heterogeneity of phylogenomic data. The multispecies coalescent model describes gene trees as independent random variables generated from a coalescence process occurring along the lineages of the species tree. Since the multispecies coalescent model allows gene trees to vary across genes, coalescent-based methods have been popularly used to account for heterogeneous gene trees in phylogenomic data analysis. In this paper, we summarize and evaluate the performance of coalescent-based methods for estimating species trees from genome-scale sequence data. We investigate the effects of deep coalescence and mutation on the performance of species tree estimation methods. We found that the coalescent-based methods perform well in estimating species trees for a large number of genes, regardless of the degree of deep coalescence and mutation. The performance of the coalescent methods is negatively correlated with the lengths of internal branches of the species tree.

Key words: coalescent methods, incomplete lineage sorting, phylogenomic data, species tree.

Molecular sequences have been predominantly used to understand the evolutionary history of species (Hillis et al., 1993; Swofford et al., 1996). With the ancestral information encoded in the genetic material of contemporary species, researchers are able touncover some of the evolutionary mysteries that occurred millions of years ago (Song et al., 2012; Jarvis et al., 2014). Yet, a large number of evolution-related questions remain unresolved. Adequately addressing those questions demands phylogenetically informative datasets and computational approaches that can effectively extract information contained in the molecular data (Liu et al., 2015b). Over the past few years, genome-scale sequence data have become increasingly available for phylogenetic studies (Edwards et al., 2007; Casci, 2011; Kumar et al., 2012; Song et al., 2012; Jarvis et al., 2014; Xi et al., 2014). Meanwhile, the complexity of genome-scale data imposes a tremendous challenge on developing probabilistic models to account for heterogeneity of phylogenomic data (Edwards, 2009). The challenge arises from the observation that phylogenomic data analyses often produce highly incongruent gene trees, i.e., genes may have quite different histories (Degnan & Rosenberg, 2009; Liu et al., 2015a). Traditional phylogenetic approaches, which infer species trees from the alignments concatenated across genes (De Queiroz & Gatesy, 2007), cannot handle heterogeneity among gene trees estimated from phylogenomic data.

A number of biological processes can produce incongruent gene trees embedded in the species tree (Maddison, 1997; Avise, 2000; Ma et al., 2000). It is desirable to integrate those biological processes into mathematical models that describe the probability distribution of gene trees generated from the species tree. Those mathematical models, which allow gene trees to vary across genes, are the generalization to the concatenation model by relaxing the assumption that all genes have the same history. A variety of mathematical models have been developed along this line (Liu et al., 2008; Kubatko, 2009; Bloomquist & Suchard, 2010; Rasmussen & Kellis, 2012), among which the multispecies coalescent model (Rannala & Yang, 2003) has become most popular for phylogenomic data analysis. In this paper, we describe the multispecies coalescent model, and summarize the statistical and computational properties of the coalescent-based methods for estimating species trees. Previous simulation studies produced mixed results for the performance of the coalescent and concatenation methods in estimating species trees (Leache & Rannala, 2011; Mirarab et al., 2014a). We investigate the effects of deep coalescence and mutation on the performance of species tree estimation methods. We find that the coalescent-based methods perform well, regardless of the degree of incomplete lineage sorting (ILS). In contrast, high ILS may positively mislead the concatenation method (Kubatko & Degnan, 2007; Roch & Steel, 2015). In the presence of a high degree of deep coalescence and mutation, the coalescent methods can accurately estimate the species tree with a high probability, when there are a large number of genes. The performance of the coalescent methods is negatively correlated with the lengths of internal branches of the species tree.

The Multispecies Coalescent Model

The multispecies coalescent model was developed to deal with gene tree heterogeneity observed in multilocus sequence data (Rannala & Yang, 2003; Edwards et al., 2007; Liu & Pearl, 2007). The multispecies coalescent model extends the classical Kingman coalescent (Kingman, 1982) to multiple populations with evolutionary relationships described by a species tree. As discussed above, incongruent gene trees may result from many biological processes, including ILS, horizontal gene transfer (HGT), hybridization, recombination, and gene duplication/loss. The biological process that is primarily responsible for the incongruent gene trees may vary across species. Nevertheless, the coalescence process often serves as the null hypothesis (Liu et al., 2015b), and the evolution of individual genes is commonly modeled as a coalescence process occurring along the lineages of a species tree (Wakeley, 2008). When the effects of hybridization, HGT, and recombination are not negligible, the coalescent model can be generalized by integrating those biological processes into the model. Gene tree reconciliation methods estimate species trees by minimizing the distance between the species tree and gene trees caused by deep coalescence, HGT, and gene duplication/loss (Pamilo & Nei, 1988; Powell, 1991; Baum, 1992; Doyle, 1992; Hudson, 1992; Brower et al., 1996; Page & Charleston, 1997; Cao et al., 1998; Nichols, 2001; Pollard et al., 2006). It has been shown that the method for minimizing deep coalescence (Maddison & Knowles, 2006) is statistically inconsistent in estimating species trees under the multispecies coalescent model (Than & Rosenberg, 2011). Alternatively, gene tree reconciliation can be achieved by minimizing a tree distance metric that is not defined upon any biological event (i.e., deep coalescence, HGT, or gene duplication/loss). For example, species trees are estimated from incongruent multicopy gene trees using the Robinson–Foulds distance (mulRF, Chaudhary et al., 2015). The mulRF method can take unresolved gene trees or gene trees with multiple alleles per species as input data to infer species trees. For binary gene trees with one allele per species, mulRF is equivalent to the majority rule consensus method. Moreover, the Accurate Species TRee ALgorithm (ASTRAL) (Mirarab et al., 2014b) estimates species trees by minimizing the quartet distance between gene trees and the species tree.

In the multispecies coalescent model, incongruent gene trees $\mathbf{G} = (g_1, \ldots, g_n)$, are assumed to be independently generated from a coalescence process occurring along the lineages of the species tree S (Fig. 1), in which n denotes the number of genes (Liu & Pearl, 2007). The assumptions of the multispecies coalescent model include that (i) incongruent gene trees are caused by deep coalescence, (ii) there is no gene flow after speciation, (iii) there is no recombination within genes but free recombination between genes, (iv) mating is random, and (v) there is no selection. These assumptions are sufficient for modeling gene trees as independent random variables given the species tree S. Moreover, the multigene alignments $\mathbf{D} = (d_1, \ldots, d_n)$ are assumed to evolve independently along the branches of individual gene trees under the substitution models (Fig. 1).

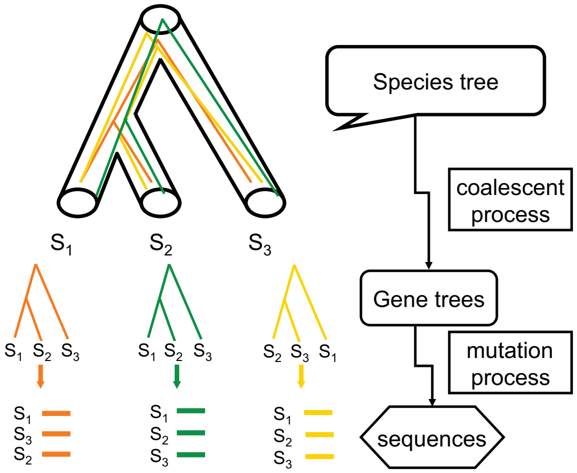

Fig. 1. The hierarchical model for multigene sequence data. The model consists of three components—alignments, gene trees, and the species tree. This hierarchical model involves two layers; the sequences-and-genetree layer and the genetree-and-speciestree layer. The species tree involes three species S1, S2, and S3. The model assumes that the gene trees are generated from a coalescent process occurring along the lineages of the species tree, while the sequences are generated from a mutation process occuring on the branches of the gene trees.

The evolution of multigene alignments involves two stochastic processes—the coalescence process and the mutation process (Fig. 1). Under the multispecies coalescent model, gene trees are generated from a coalescence process occurring along the lineages of the species tree (Fig. 1). Meanwhile, molecular sequences evolve on gene trees, following a mutation process described by a substitution model (Felsenstein, 1981). The two stochastic processes characterize the relationships among molecular sequences **D**, gene trees **G**, and the species tree S. The probability distribution of the gene tree given the species tree can be derived from the coalescent process (Rannala & Yang, 2003; Degnan & Salter, 2005). The probability density of gene trees **G** given the species tree is

$$f(\mathrm{G}|\mathrm{S}) = \Pi_i f(g_i|\mathrm{S}). \tag{1}$$

With the assumption of free recombination between genes, individual gene trees $(g_1, \ldots, g_n)$ are treated as independent random variables conditional on the species tree S. Because gene trees are random quantities, the multispecies coalescent model allows genes to have distinct histories. Additionally, the probability distribution of the alignments given a gene tree can be derived from the mutation process (Fig. 1). The likelihood function derived from the mutation process is one used for calculating maximum likelihood (ML) gene trees by traditional phylogenetic methods, i.e.,

$$f(\mathrm{D}|\mathrm{G}, \gamma) = \Pi_i f(d_i|g_i, \gamma_i). \tag{2}$$

In Equation (2), γ_i represents the parameters of the substitution model for gene *i*. Combining Equations (1) and (2), the likelihood function of the species tree S is given by

$$\iota(\mathrm{S}|\mathrm{D}) = \int_{\mathrm{G}} f(\mathrm{D}|\mathrm{G},\gamma) f(\mathrm{G}|\mathrm{S}) d\mathrm{G}. \qquad (3)$$

From Equation (3), genetic variation of multigene sequence data is the consequence of the combination of coalescence and mutation processes. The coalescence process results in genetic variation between genes; the mutation process results in genetic variation within genes. Both coalescence and mutation variation can influence the accuracy of species tree estimation.

Coalescent-Based Methods for Estimating Species Trees

A number of coalescent-based methods have been developed for estimating species trees from multigene sequence data. The Bayesian coalescent approaches estimate species trees from alignments (single or multiple alleles) (Table 1) using both the likelihood function and the prior distribution of the species tree (Liu & Pearl, 2007; Liu et al., 2008; Heled & Drummond, 2009). Bayesian inference is based on the posterior probability distribution approximated by a sample of species trees generated from a Markov Chain Monte Carlo (MCMC) algorithm. The Bayesian approach involves intensive computation. Thus, it is not practical to apply the Bayesian approach to genome-scale sequence data. To reduce computational cost, various coalescent-based methods were developed to estimate species trees in two steps—estimating gene trees from multigene sequences and then estimating the species tree from the estimated gene trees. Carstens & Knowles (2007) proposed a coalescent-based approach for estimating species trees from a collection of estimated gene trees. Given the estimated gene trees, this approach calculates the likelihood scores of all possible species trees (Degnan & Salter, 2005). The best tree is selected by a likelihood ratio test with the correction for multiple comparisons (Anisimova & Gascuel, 2006). Because this approach needs to calculate the likelihood scores of all possible species trees, it cannot be used to reconstruct phylogenies that involve a large number of taxa.

Numerous phylogenetic methods estimate species trees using the summary statistics of a set of gene trees (Table 1). Due to their computational advantages, gene-tree-based coalescent methods have been primarily adopted in phylogenomic data analysis. However, the performance of gene-tree-based methods can be significantly affected by the fact that those methods do not utilize all phylogenetic information contained in molecular sequence data (Liu et al., 2015b). As one of the gene-tree-based methods, Global LAteSt Split (GLASS) (Mossel & Roch, 2007), which is also called the Maximum tree (Liu et al., 2010b), clusters species using minimum coalescence times. Given true gene trees, GLASS and the Maximum tree are statistically consistent under the multispecies coalescent model, as the number of genes goes to infinity. If population size parameter θ is constant across populations on the species tree, the Maximum tree is the ML estimate of the species tree (Liu et al., 2010b). However, when gene trees are estimated from DNA sequences, the minimum coalescence time across gene trees converges to 0 as the number of genes grows because the probability that two arbitrary sequence shave exactly the same nucleotides is positive. Since the species tree is estimated by the minimum coalescent times, the biased minimum coalescence times can consistently produce the wrong estimate of the species tree. Thus, when gene trees are estimated from DNA sequences, GLASS and the Maximum tree are statically inconsistent (Degiorgio & Degnan, 2014). The principle of clustering species by minimum coalescence times is also implemented in the software Species Tree Estimation using ML, or STEM (Kubatko et al., 2009). In contrast, the STEAC method estimates species trees using average coalescence times, which is more robust to the estimation error of coalescence times. Moreover, Liu et al. (2009b) proposed to estimate species trees using average ranks of gene coalescence times (STAR). The STAR method estimatesthe species tree by a neighbor-joining tree built from a distance matrix, in which the entries are twice the average ranks across gene trees. Simulation studies suggest that STAR outperforms STEAC, when the estimation error of coalescence times is large. The STAR and STEAC methods can quickly infer phylogenies even for large-scale phylogenomic data (Liu et al., 2009a). When the true gene trees are given, STAR and STEAC are statistically consistent in estimating species trees (Liu et al., 2009b; Allman et al., 2013; Degnan, 2013). Both methods are robust to a limited amount of

Table 1 Coalescent methods for estimating species trees

	Input	Output	Method	Website	Speed
ASTRAL	Gene trees	T	Summary statistics	https://github.com/smirarab/ASTRAL	Fast
*BEAST	Alignments	T and B	Bayesian method	http://beast.bio.ed.ac.uk/Main_Page	Slow
BEST	Alignments	T and B	Bayesian method	http://www.stat.osu.edu/∼dkp/BEST/introduction/	Slow
BUCKy	Gene trees	T	Bayesian method	http://www.stat.wisc.edu/∼ane/bucky/	Slow
GLASS	Gene trees	T and B	Summary statistics	http://code.google.com/p/phybase/downloads/list	Fast
MP-EST	Gene trees	T and B	Likelihood method	http://bioinformatics.publichealth.uga.edu/	Fast
NJst	Gene trees	T	Summary statistics	http://bioinformatics.publichealth.uga.edu/	Fast
STAR	Gene trees	T	Summary statistics	http://bioinformatics.publichealth.uga.edu/	Fast
STEAC	Gene trees	T	Summary statistics	http://bioinformatics.publichealth.uga.edu/	Fast
STELLS	Gene trees	T and B	Likelihood method	http://www.engr.uconn.edu/∼ywu/STELLS.html	Slow
STEM	Gene trees	T	Likelihood method	http://www.stat.osu.edu/∼lkubatko/software/STEM/	Fast

B, branch lengths of the species tree; T, topology of the species tree.

horizontal transfer as well as deviations from the molecular clock assumption, because some small values of coalescence times due to horizontal transfer or rate variation in a small number of genes do not have major effects on the average ranks and average coalescence times when the number of genes is moderate or large. STELLS (Wu, 2012) estimates the species tree from a set of gene trees by maximizing the probability of gene trees given the species tree under the multispecies coalescent model (Degnan & Salter, 2005). Liu et al. (2010a) introduced a maximum pseudo-likelihood method for estimating species trees (MP-EST). The MP-EST method estimates the species tree by maximizing the pseudo-likelihood of the triplets in the species tree. Unlike most summary-statistics-based methods, the MP-EST method is able to estimate both the topology and branch lengths (in coalescent units) of the species tree. Given the true gene trees, MP-EST is statistically consistent, as the number of genes increases to infinity. Moreover, if the sequence length also goes to infinity, MP-EST based on the estimated gene treescan consistently recover the true species tree under the multispecies coalescent model (Liu et al., 2010a).

The coalescent methods described above require that the input trees must be rooted gene trees. A distance method, NJst, can infer species trees from unrooted gene trees (Liu & Yu, 2011). In the NJst method, the distance between two species is defined as the average number of internal nodes between two species across gene trees. The species tree is estimated by the neighbor-joining tree built from the distance matrix. ASTRAL can estimate species trees from unrooted gene trees, because it minimizes the quartet distance between gene trees and the species tree(Mirarab et al., 2014b). When gene trees are accurately estimated, both NJst and ASTRAL methods are statistically consistent under the multispecies coalescent model. BUCKy is a phylogenetic program for Bayesian concordance analysis (Ane et al., 2007). BUCKy estimates concordance trees (or population trees) by calculating concordance factors (or quartet concordance factors) of a group of estimated gene trees. Both ASTRAL and BUCKy do not rely on any biological process to explain discordant gene trees. Thus, ASTRAL and BUCKy are not coalescent methods per se. However, because both methods are statistically consistent under the multispecies coalescent model, they are categorized as coalescent methods in this paper (Table 1).

Coalescent versus Concatenation

Probabilistic models are the foundations of statistical phylogenetic inference. Thus, model comparison is critical in phylogenetic analysis. The concatenation model implicitly assumes that all genes evolved with the same history (Edwards, 2009). Under the concatenation model, gene trees **G** are identical with the species tree S, i.e., $g_i = \mathrm{S}$ for $i = 1,\ldots,n$. Since the sites evolve independently, the likelihood function of the species tree S is the product of the likelihoods of individual sites, i.e.,

$$\iota_{\mathrm{con}}(\mathrm{S}|\mathrm{D}) = \Pi_i f(\mathrm{d}_i|g_i,\gamma_i) = \Pi_i f(\mathrm{d}_i|\mathrm{S},\gamma_i) \qquad (4)$$

In contrast, gene trees in the multispecies coalescent model are independent random variables conditional on the species tree. To simply the multispecies coalescent model, it is assumed that one allele is sampled from each species. When the species tree has long internal branches (in coalescent units), the gene trees generated under the multispecies coalescent model have the same or similar topology. As the population size parameter $\theta = 4\mu\mathrm{N_e}$ (μ is the mutation rate and $\mathrm{N_e}$ is the effective population size) goes to 0, the coalescence times of the gene trees converge to the species divergence times of the species tree. Thus, when the species tree has long internal branches and small θ, the multispecies coalescent model reduces to the concatenation model (Liu et al., 2015b). The likelihood function of the species tree under the multispecies coalescent model reduces to

$$\begin{aligned}\iota_{\mathrm{coal}}(\mathrm{S}|\mathrm{D}) &= \int_{\mathrm{G}} f(\mathrm{D}|\mathrm{G},\gamma)f(\mathrm{G}|\mathrm{S})d\mathrm{G} = f(\mathrm{D}|\mathrm{G},\gamma)\\ &= \Pi_i f(d_i|\mathrm{S},\gamma_i)\end{aligned} \qquad (5)$$

The likelihood function (5) of the multispecies coalescent model equals the likelihood function (4) of the concatenation model. Thus, when the species tree has long internal branches and small θ, the Bayesian coalescent approaches and the concatenation methods perform similarly in estimating species trees because they use the same likelihood function to infer species trees. However, gene-tree-based coalescent methods do not use the full information of the sequence data. It is expected that the Bayesian approaches outperform the gene-tree-based coalescent methods in estimating species trees. Thus, when ILS is low and gene trees are poorly estimated, the concatenation method may outperform gene-tree-based coalescent methods. However, the performance of the gene-tree-based coalescent methods can be improved by increasing the number of genes.

In contrast, the concatenation analyses may consistently estimate wrong species trees when the internal branches of species trees are short (Roch & Steel, 2015). The poor performance of the concatenation method cannot be improved by increasing the number of genes. Previous studies have compared by simulation the performance of the concatenation and coalescent methods for a finite number of genes. In general, gene-tree-based coalescent methods outperform concatenation methods when there is a high degree of ILS, i.e., a large variation among gene trees. Since high ILS is most likely to occur on the short branches of the species tree, it indicates that coalescent methods are more accurate than the concatenation methods in estimating the short branches of the species tree. On the other hand, when there is high uncertainty in gene tree estimation, the concatenation methods may outperform the coalescent methods in estimating the long branches of the species tree. When gene trees are poorly estimated, we would expect that both concatenation and coalescent methods produce poorly supported estimates of the species tree. However, when increasing the number of genes, coalescent methods can accurately estimate the species tree with a high probability.

The coalescent-based methods have been applied to estimate species trees in a number of empirical studies. Song et al. (2012) used both coalescent and concatenation methods and incorporated 447 nuclear genes from 33 mammalian species and four outgroup species to address the effect of ILS on estimating deep-level phylogeny of mammals. The mammalian genomic data exhibit considerable gene tree heterogeneity, as all 447 gene trees differ from the estimated species tree in topology. Such high level of ILS in data exerted different influence on the performance of coalescent and concatenation methods in estimating species phylogenies. Coalescent methods, which take gene tree heterogeneity into account, were able to estimate a reliable and consistent species phylogeny for mammals, and showed a positive correlation between the number of genes and the nodal support values for nodes that exhibit high amount of ILS. In contrast, concatenation methods, which assume homogeneous gene tree across loci, can result in conflicting but strongly support phylogenies for mammals from different subsamples of loci. The mammalian genomic data will be used in simulation to compare the performance of coalescent-based and concatenation methods. In addition, gene-tree-based coalescent methods were employed to estimate the phylogenetic placement of *Amborella* (Xi et al., 2014). Their results showed that rate variation among sites can mislead the results of concatenation methods, as fast and slow evolving sites support conflict and strongly supported phylogenies. Coalescent methods, on the other hand, were able to infer a consistent placement for *Amborella* from either fast or slow evolving sites. These empirical studies demonstrated that coalescence-based methods are able to estimate accurate species trees from genome-scale data with high level of ILS or rate variation, whereas this complexity can lead concatenation methods to generate misleading results (Xi et al., 2014).

Performance of Species Tree Estimation Methods

Simulation for the 5-taxon species tree

The amount of phylogenetic signal in the DNA sequence data is determined by genetic variation generated from the coalescence and mutation processes. We used simulation to evaluate the effects of deep coalescence and mutation on species tree estimation. Gene tree variation due to ILS is positively correlated with the population size parameter θ. Two values of θ were chosen to simulate high and low ILS. To simulate high uncertainty in gene tree estimation, the branch lengths of gene trees were shorten by multiplying with a small constant. Specifically, gene trees were generated from a five-taxon species tree with one allele per species under the multispecies coalescent model using an R package Phybase (Liu & Yu, 2010). The population size parameter θ was constant across the internal branches of the species tree. Two values (0.02, 0.0002) of θ were considered to simulate high and low ILS. When $\theta = 0.02$, 90% of the simulated gene trees are incongruent with the species tree (i.e., high ILS). As the distribution of gene trees is almost flat, gene trees have little coalescent signal for accurately estimating species trees. When $\theta = 0.0002$, all simulated gene trees are congruent with the species tree (i.e., no ILS), which satisfies the assumption of the concatenation model. In this case, the multispecies coalescent model reduces to the concatenation model. The simulated gene trees were used as the input data to estimate the species tree by the gene-tree-based coalescent methods. Each simulation was repeated 100 times. To evaluate the effect of uncertainty in gene tree estimation, the branch lengths of gene trees were multiplied with a scale parameter $= 0.2$, 0.4, 0.6, 0.8, or 1.0. Gene trees with short branches (scale $= 0.2$) tend to have large estimation errors. DNA sequences were then simulated from the rescaled gene trees using Seq-Gen (Rambaut & Grassly, 1997) with the Jukes–Cantor model (Jukes & Cantor, 1969). We estimated gene trees from the simulated sequences using PhyML (Guindon & Gascuel, 2003) with the Jukes–Cantor model. Finally, the estimated gene trees were used to infer the species tree. Each simulation was repeated 100 times. Since all gene trees simulated with $\theta = 0.0002$ have the same topology, topological variation among the estimated gene trees was caused solely by the mutation process.

For high ILS ($\theta = 0.02$), gene-tree-based coalescent methods can accurately estimate the species tree with a high probability as the number of genes increases (Fig. 2). The coalescent methods outperform greedy consensus and concatenation for all simulations with high ILS (Fig. 2). Given the true gene trees, NJst and STAR appear to outperform MP-EST and ASTRAL when the number of genes is small (Fig. 2). In our simulations, there is no significant difference in the performance of STAR, MP-EST, ASTRAL, and NJst. When gene trees are estimated from DNA sequences, the probability of estimating the true species tree for all coalescent methods decreases as the scale parameter μ decreases from 1.0 to 0.2 (Fig. 2). Because small μ results in large uncertainty in estimating gene trees, this result suggests that gene tree estimation error reduces the performance of gene-tree-based coalescent methods. The negative effect of gene tree uncertainty becomes more severe when the number of genes is small (Fig. 2). In the presence of high uncertainty of gene tree estimation, all gene-tree-based coalescent methods perform similarly in recovering the true species tree.

When the sequence data were simulated with no ILS (i.e., $\theta = 0.0002$), gene-tree-based coalescent methods can accurately estimate the species tree for all simulations as the number of genes increases (Fig. 2). Moreover, the species tree with no ILS is more likely to be recovered by gene-tree-based coalescent methods than the species tree with high ILS, because high ILS further increases the variability among the estimated gene trees. All gene-tree-based coalescent methods perform equally well in estimating the true species tree, but concatenation appears to outperform all gene-tree-based coalescent methods for scale $= 0.2$ and a small number of genes (ngene $= 10$) (Fig. 2). This result is consistent with our expectation discussed in the previous section. Moreover, the performance of all species tree estimation methods can be dramatically improved by increasing the number of genes (Fig. 2).

Simulation for the Mammalian tree

We simulated DNA sequences from the MP-EST tree estimated from a mammalian dataset. The mammalian dataset contained DNA sequences from 447 loci for 37 species (Song et al., 2012). We used the MP-EST tree built from the mammalian dataset as the true species tree to generate gene trees under the multispecies coalescent model. Mirarab et al. (2014a) used the same mammalian tree to compare the performance of greedy consensus, concatenation, and MP-EST. We re-estimated the branch lengths of the true

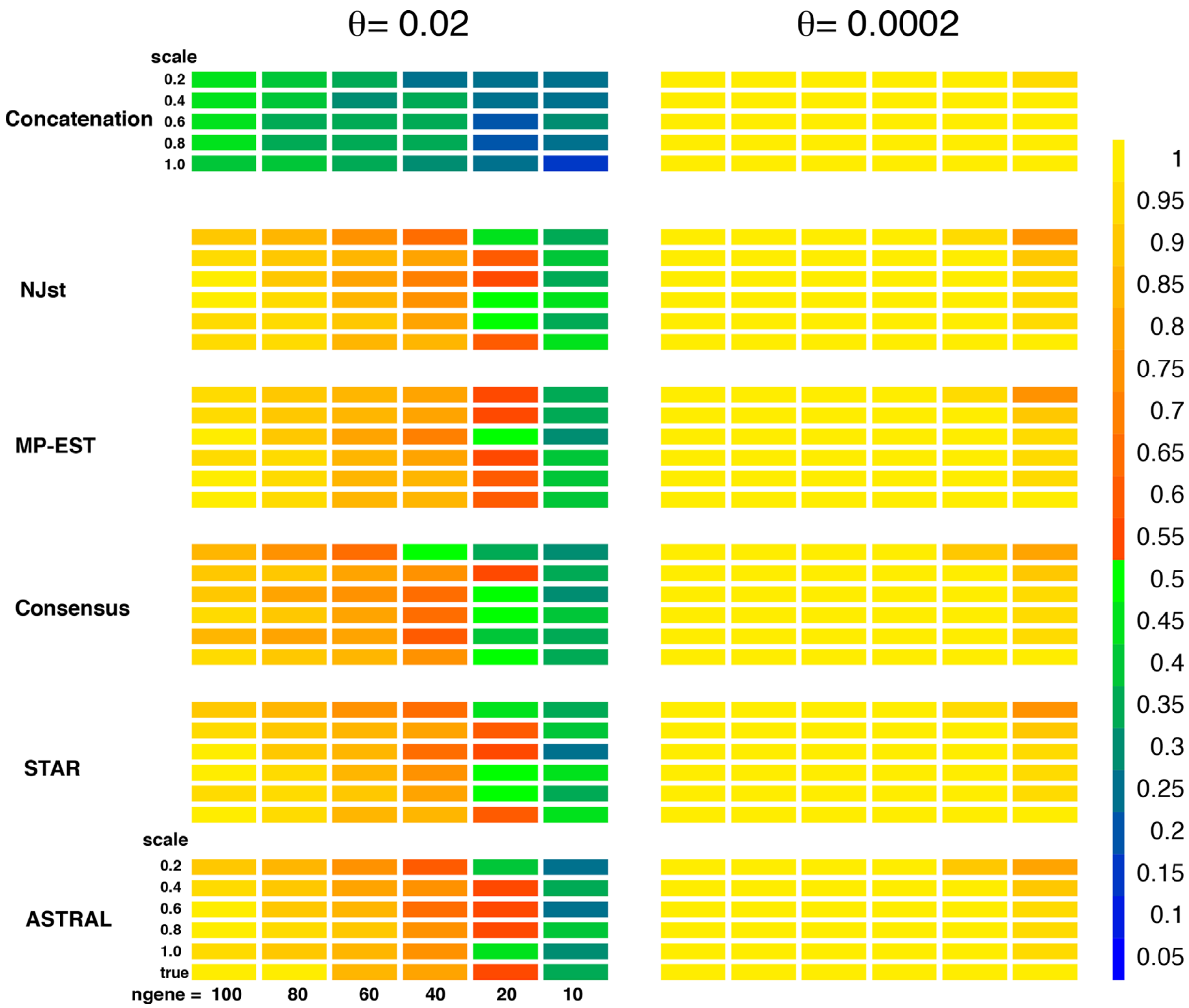

Fig. 2. The heat map for the performance of species tree estimation methods. Gene trees (10, 20, 40, 60, 80, and 100) were generated from a five-taxon species tree with the population size parameter θ = 0.02 or 0.0002. High incomplete lineage sorting (ILS) for θ = 0.02 and no ILS for θ = 0.0002. The branch lengths of the gene trees were multiplied by a scale parameter (scale = 0.2, 0.4, 0.6, 0.8, 1.0). Gene trees were then used to simulate DNA sequences under the Jukes–Cantor model. The colors represent the proportions of the correct species tree estimated by ASTRAL, STAR, greedy consensus, MP-EST, NJst, and concatenation.

species tree from the sequences concatenated across 447 loci using PhyML with the GTR+Γ model (Tavaré, 1986; Yang, 1994). We converted the species tree to an ultrametric tree by equalizing the distances from the tips to the root of the tree (Fig. 3a). The population size parameter θ = 0.05 was constant on the entire tree. In this simulation, we compared the performance of different species tree estimation methods. Moreover, we investigated the correlation between the performance of species tree estimation methods and the lengths of internal branches of the species tree. We simulated 100 and 500 gene trees from the true species tree under the multispecies coalescent model using Phybase. Then, we simulated DNA sequences of 1000 base pairs from gene trees using Seq-Gen with the Jukes–Cantor model. We estimated gene trees from the simulated sequence data using PhyML with the Jukes–Cantor model. The estimated gene trees were then used as the input data of STAR, MP-EST, ASTRAL, greedy consensus, and NJst to estimate the species tree. In addition, we built ML trees for the concatenated sequences using PhyML with the Jukes–Cantor model. Each simulation was repeated 100 times. To evaluate the performance of species tree estimation methods, we calculated the positive branch rate for each branch of the species tree. The positive branch rate for branch *i* is the proportion of the estimated species trees in which branch *i* is successfully recovered. Moreover, we calculated the correlation between the positive branch rate and the length of the internal branch of the species tree.

The true species tree has short and long internal branches (in mutation units, Fig. 3a). As the population size parameter θ = 0.05 is constant across internal branches of the species tree, the branch lengths in coalescent units are equal to the branch lengths in mutation units divided by the population size parameter θ = 0.05 (Fig. 3b). According to coalescent theory, short branches (in coalescent units) are associated

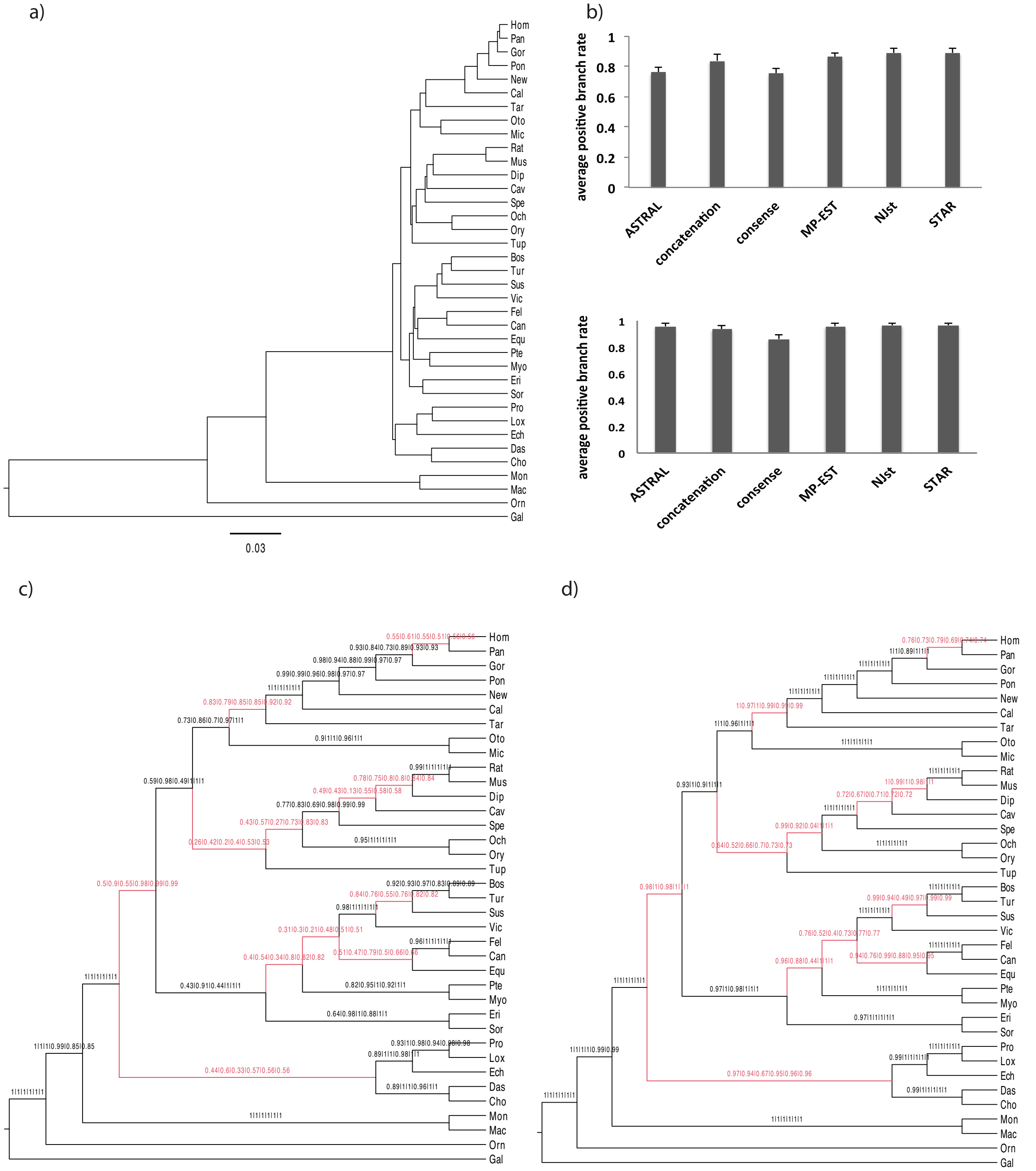

Fig. 3. Simulation results for the Mammalian tree. **a,** The species tree used for simulating DNA sequence data. There are short and long internal branches in the species tree. **b,** The average positive branch rates for species tree estimation methods. The average positive branch rate is equal to the average of the positive branch rates across all internal branches of the species tree. DNA sequences were simulated from 100 and 500 genes. **c,** The positive branch rate on each internal branch of the species tree for the 100 genes simulation. On each branch, the numbers (from left to right) are the positive branch rates of ASTRAL, concatenation, greedy consensus, MP-EST, NJst, and STAR, respectively. The short branches (branch length < 0.005 in mutation units) are highlighted. **d,** The positive branch rate for the 500 genes simulation. The positive branch rates of all species tree estimation methods increase when the number of genes increases to 500. On each branch, the numbers (from left to right) are the positive branch rates of ASTRAL, concatenation, greedy consensus, MP-EST, NJst, and STAR, respectively.

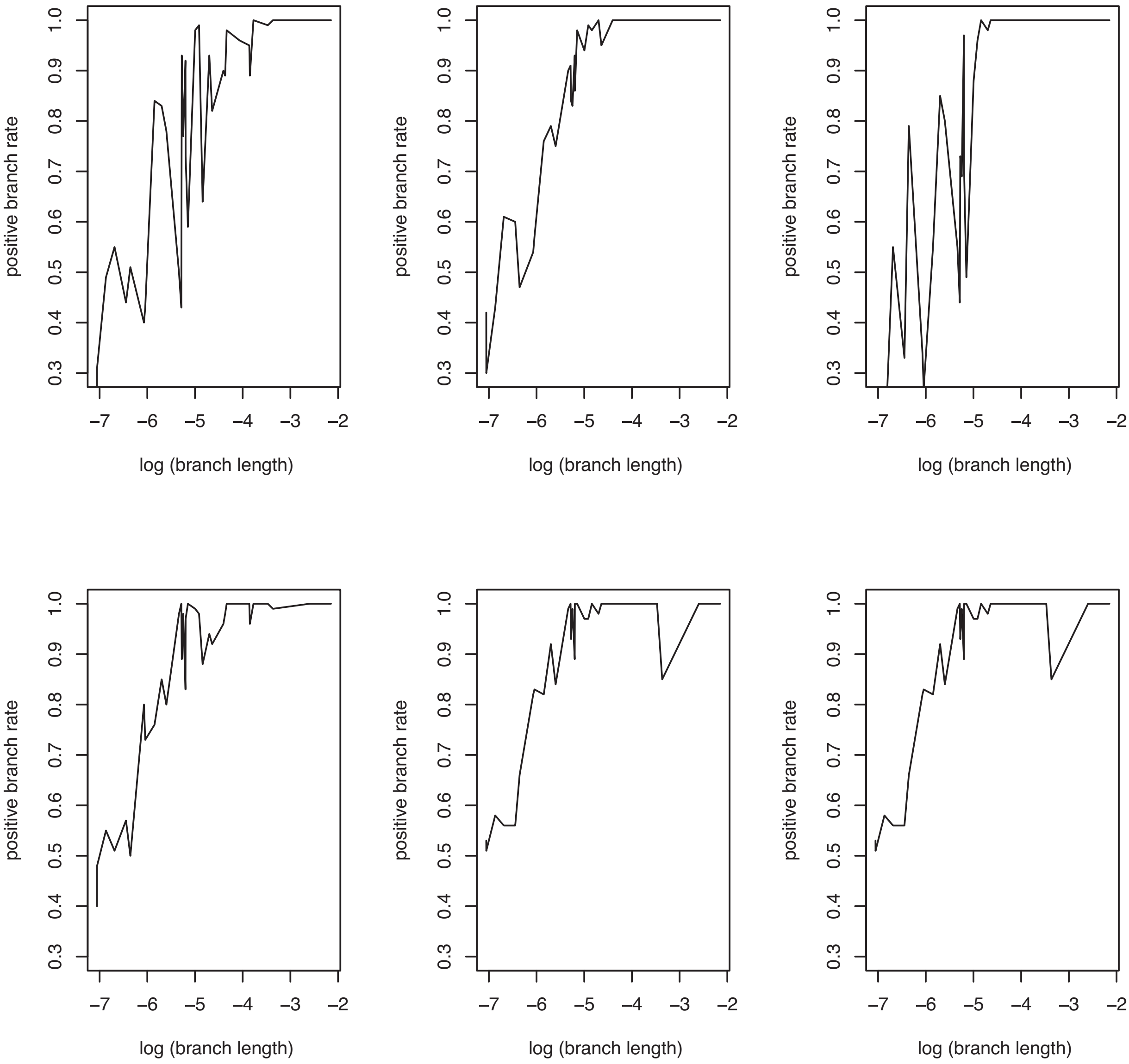

Fig. 4. The correlation between the positive branch rate and the length of the internal branch of the species tree.

with high ILS. Overall, the average positive branch rates of STAR and NJst are higher than those of other species tree estimation methods (Fig. 3b). In contrast, ASTRAL and consensus receive the lowest average positive branch rate (Fig. 3b). Increasing the number of genes can improve the performance of species tree estimation methods. For 500 genes, the positive branch rates of the species tree estimation methods except concatenation (0.94) and greedy consensus (0.86) are greater than 0.95. The improvement is more significant for the coalescent-based methods than for concatenation (Fig. 3b). When the number of gene trees increases to 500, ASTRAL performs better than the concatenation and consensus methods (Fig. 3b). Moreover, the positive branch rates for short internal branches are less than those for long internal branches of the species tree (Figs. 3c, 3d). The positive branch rates of MP-EST, STAR, and NJst are higher than those of ASTRAL, concatenation, and greedy consensus (Figs. 3c, 3d). In general, the poor performance of ASTRAL and consensus is more significant for short internal branches of the species tree (Fig. 3c). In addition, the positive branch rate is positively correlated with the length of the internal branch of the species tree for all species tree estimation methods (Fig. 4).

Conclusions

Genome-scale data have become increasingly available in phylogenetic studies. Empirical analyses have demonstrated strong evidence of heterogeneous gene trees. There are a number of biological processes that can cause discrepancy between gene trees and species trees. A reasonable mathematical model for analyzing genome-scale data should treat gene trees as random variables that may vary across genes. Since the multispecies coalescent model allows gene trees to vary across genes, it is more realistic than the concatenation model that assumes the same tree for all genes. In fact, under certain conditions, the multispecies coalescent model reduces to the concatenation model. Thus, the coalescent methods perform well under the concatenation model, but the concatenation methods may perform poorly under the multispecies coalescent model, especially when the internal branches of the species tree are short.

The evolution of multigene sequences involves two random processes; the coalescent and mutation processes. The performance of a species tree estimation method (coalescent or concatenation) is determined bythe sampling error from the combination of the two processes. When gene trees have the same topology, the Bayesian coalescent methods and the concatenation methods perform equally well, because they use the same likelihood function to make inference about the species tree. In contrast, the gene-tree-based coalescent methods estimate species trees using only the topologies of gene trees. Thus, when gene tree variation is primarily caused by mutation errors, our simulation suggests that concatenation outperform the gene-tree-based coalescent methods. However, the performance of the gene-tree-based coalescent methods can be greatly improved by increasing the number of genes.

The coalescent-based methods are promising for accurately estimating species trees from phylogenomic data. The coalescent inference is based on the assumption that the multispecies coalescent model is a good approximation to the real biological process that causes incongruent gene trees. There are preliminary attempts to assess the goodness of fit of the multispecies coalescent model, but they were either limited to small data sets (Reid et al., 2014) or they were based on the distance between gene trees and the species tree (Song et al., 2012), which may not have the power to reject the multispecies coalescent model. Reid et al. (2014) evaluated the multispecies coalescent model in a Bayesian framework using posterior predictive simulation (PPS), in which the estimated gene trees were compared with the predictive distribution of gene trees. Since the predictive distribution of gene trees is generated from a Bayesian coalescent approach (i.e., *BEAST), PPS is not able to evaluate the multispecies coalescent model for genome-scale sequence data. A powerful goodness of fittest is needed to validate the multispecies coalescent model for phylogenomic data.

Acknowledgements

This work was supported by the United States National Science Foundation DMS-1222745 to L.L.

References

Allman ES, Degnan JH, Rhodes JA. 2013. Species tree inference by the STAR method and its generalizations. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 20: 50–61.

Ane C, Larget B, Baum DA, Smith SD, Rokas A. 2007. Bayesian estimation of concordance among gene trees. *Molecular Biology and Evolution* 24: 412–426.

Anisimova M, Gascuel O. 2006. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Systematic Biology* 55: 539–552.

Avise JC. 2000. *Phylogeography: The history and formation of species.* Cambridge: Harvard University Press.

Baum BR. 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon* 41: 3–10.

Bloomquist EW, Suchard MA. 2010. Unifying vertical and nonvertical evolution: A stochastic ARG-based framework. *Systematic Biology* 59: 27–41.

Brower AVZ, Desalle R, Vogler A. 1996. Gene trees, species trees, and systematics: A cladistic perspective. *Annual Review of Ecology and Systematics* 27: 423–450.

Cao Y, Janke A, Waddell PJ, Westerman M, Takenaka O, Murata S, Okada N, Paabo S, Hasegawa M. 1998. Conflict among individual mitochondrial proteins in resolving the phylogeny of eutherian orders. *Journal of Molecular Evolution* 47: 307–322.

Carstens BC, Knowles LL. 2007. Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: An example from Melanoplus grasshoppers. *Systematic Biology* 56: 400–411.

Casci T. 2011. Phylogenomics: improving our family tree. *Nature Reviews Genetics* 12: 298–299.

Chaudhary R, Fernandez-Baca D, Burleigh JG. 2015. MulRF: A software package for phylogenetic analysis using multi-copy gene trees. *Bioinformatics* 31: 432–433.

De Queiroz A, Gatesy J. 2007. The supermatrix approach to systematics. *Trends in Ecology & Evolution* 22: 34–41.

Degiorgio M, Degnan JH. 2014. Robustness to divergence time underestimation when inferring species trees from estimated gene trees. *Systematic Biology* 63: 66–82.

Degnan JH. 2013. Evaluating variations on the STAR algorithm for relative efficiency and sample sizes needed to reconstruct species trees. *Pacific Symposium on Biocomputing* 262–272.

Degnan JH, Rosenberg NA. 2009. Gene tree discordance, phylogenetic inference, and the multispecies coalescent. *Trends in Ecology & Evolution* 24: 332–340.

Degnan JH, Salter LA. 2005. Gene tree distributions under the coalescent process. *Evolution* 59: 24–37.

Doyle JJ. 1992. Gene trees and species trees—molecular systematics as one-character taxonomy. *Systematic Botany* 17: 144–163.

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

Edwards SV, Liu L, Pearl DK. 2007. High-resolution species trees without concatenation. *Proceedings of the National Academy of Sciences USA* 104: 5936–5941.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution* 17: 368–376.

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

Heled J, Drummond AJ. 2009. Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution* 27: 570–580.

Hillis DM, Allard MW, Miyamoto MM. 1993. Analysis of DNA sequence data: Phylogenetic inference. *Methods in Enzymology* 224: 456–487.

Hudson RR. 1992. Gene trees, species trees and the segregation of ancestral alleles. *Genetics* 131: 509–512.

Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SY, Faircloth BC, Nabholz B, Howard JT, Suh A, Weber CC, Da Fonseca RR, Li J, Zhang F, Li H, Zhou L, Narula N, Liu L, Ganapathy G, Boussau B, Bayzid MS, Zavidovych V, Subramanian S, Gabaldon T, Capella-Gutierrez S, Huerta-Cepas J, Rekepalli B, Munch K, Schierup M, Lindow B, Warren WC, Ray D, Green RE, Bruford MW, Zhan X, Dixon A, Li S, Li N, Huang Y, Derryberry EP, Bertelsen MF, Sheldon FH, Brumfield RT, Mello CV, Lovell PV, Wirthlin M, Schneider MP, Prosdocimi F, Samaniego JA, Vargas Velazquez AM, Alfaro-Nunez A, Campos PF, Petersen B, Sicheritz-Ponten T, Pas A, Bailey T, Scofield P, Bunce M, Lambert DM, Zhou Q, Perelman P, Driskell AC, Shapiro B, Xiong Z, Zeng Y, Liu S, Li Z, Liu B, Wu K, Xiao J, Yinqi X, Zheng Q, Zhang Y, Yang H, Wang J, Smeds L, Rheindt FE, Braun M, Fjeldsa J, Orlando L, Barker FK, Jonsson KA, Johnson W, Koepfli KP, O'brien S, Haussler D, Ryder OA, Rahbek C, Willerslev E, Graves GR, Glenn TC, Mccormack J, Burt D, Ellegren H, Alstrom P, Edwards SV, Stamatakis A, Mindell DP, Cracraft J, Braun EL, Warnow T, Jun W, Gilbert MT, Zhang G. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* 346: 1320–1331.

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. *Mammalian protein metabolism*. New York: Academic Press.

Kingman J. 1982. On the genealogy of large populations. *Journal of Applied Probability* 19A: 27–34.

Kubatko LS. 2009. Identifying hybridization events in the presence of coalescence via model selection. *Systematic Biology* 58: 478–488.

Kubatko LS, Carstens BC, Knowles LL. 2009. STEM: Species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics* 25: 971–973.

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

Kumar S, Filipski AJ, Battistuzzi FU, Kosakovsky Pond SL, Tamura K. 2012. Statistics and truth in phylogenomics. *Molecular Biology and Evolution* 29: 457–472.

Leache AD, Rannala B. 2011. The accuracy of species tree estimation under simulation: A comparison of methods. *Systematic Biology* 60: 126–137.

Liu L, Pearl DK. 2007. Species trees from gene trees: Reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Systematic Biology* 56: 504–514.

Liu L, Pearl DK, Brumfield RT, Edwards SV. 2008. Estimating species trees using multiple-allele DNA sequence data. *Evolution* 62: 2080–2091.

Liu L, Xi Z, Davis CC. 2015a. Coalescent methods are robust to the simultaneous effects of long branches and incomplete lineage sorting. *Molecular Biology and Evolution* 32: 791–805.

Liu L, Xi Z, Wu S, Davis C, Edwards S. 2015b. Estimating phylogenetic trees from genome-scale data. *Annals of the New York Academy of Sciences* doi: 10.1111/nyas.12747

Liu L, Yu L. 2010. Phybase: An R package for species tree analysis. *Bioinformatics* 26: 962–963.

Liu L, Yu L. 2011. Estimating species trees from unrooted gene trees. *Systematic Biology* 60: 661–667.

Liu L, Yu L, Edwards SV. 2010a. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evolutionary Biology* 10: 302–320.

Liu L, Yu L, Kubatko LS, Pearl DK, Edwards SV. 2009a. Coalescent methods for estimating multilocus phylogenetic trees. *Molecular Phylogenetics and Evolution* 53: 320–328.

Liu L, Yu L, Pearl DK. 2010b. Maximum tree: A consistent estimator of the species tree. *Journal of Mathematical Biology* 60: 95–106.

Liu L, Yu L, Pearl DK, Edwards SV. 2009b. Estimating species phylogenies using coalescence times among sequences. *Systematic Biology* 58: 468–477.

Ma B, Li M, Zhang LX. 2000. From gene trees to species trees. *SIAM Journal of Computation* 30: 729–752.

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

Maddison WP, Knowles LL. 2006. Inferring phylogeny despite incomplete lineage sorting. *Systematic Biology* 55: 21–30.

Mirarab S, Bayzid MS, Warnow T. 2014a. Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. *Systematic Biology* doi:10.1093/sysbio/syu063

Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson MS, Warnow T. 2014b. ASTRAL: Genome-scale coalescent-based species tree estimation. *Bioinformatics* 30: 541–548.

Mossel E, Roch S. 2007. Incomplete lineage sorting: consistent phylogeny estimation from multiple Loci. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 7: 166–171.

Nichols R. 2001. Gene trees and species trees are not the same. *Trends in Ecology & Evolution* 16: 358–364.

Page RDM, Charleston MA. 1997. From gene to organismal phylogeny: Reconciled trees and the gene tree species tree problem. *Molecular Phylogenetics and Evolution* 7: 231–240.

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

Pollard DA, Iyer VN, Moses AM, Eisen MB. 2006. Widespread discordance of gene trees with species tree in Drosophila: Evidence for incomplete lineage sorting. *Plos Genetics* 2: 1634–1647.

Powell JR. 1991. Monophyly/paraphyly/polyphyly and gene/species trees—an example from *Drosophila*. *Molecular Biology and Evolution* 8: 892–896.

Rambaut A, Grassly NC. 1997. Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer Applications in the Biosciences: CABIOS* 13: 235–238.

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

Rasmussen MD, Kellis M. 2012. Unified modeling of gene duplication, loss, and coalescence using a locus tree. *Genome Research* 22: 755–765.

Reid NM, Hird SM, Brown JM, Pelletier TA, Mcvay JD, Satler JD, Carstens BC. 2014. Poor fit to the multispecies coalescent is widely detectable in empirical data. *Systematic Biology* 63: 322–333.

Roch S, Steel M. 2015. Likelihood-based tree reconstruction on a concatenation of alignments can be positively misleading. *arXiv:1409.2051 [q-bio.PE]*.

Song S, Liu L, Edwards SV, Wu S. 2012. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proceedings of the National Academy of Sciences USA* 109: 14942–14947.

Swofford DL, Olsen GJ, Waddell PJ, Hillis DM. 1996. Phylogenetic inference. In Hillis DM, Moritz C, Mable BK eds. *Molecular Systematics*. Sunderland, MA: Sinauer. 407–514.

Tavaré S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. *American Mathematical Society: Lectures on Mathematics in the Life Sciences* 17: 57–86.

Than CV, Rosenberg NA. 2011. Consistency properties of species tree inference by minimizing deep coalescences. *Journal of* *computational biology: A Journal of Computational Molecular Cell Biology* 18: 1–15.

Wakeley J. 2008. *Coalescent Theory: An Introduction*. Greenwood Village: Roberts & Company Publishers.

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

Xi Z, Liu L, Rest JS, Davis CC. 2014. Coalescent versus concatenation methods and the placement of *Amborella* as sister to water lilies. *Systematic Biology* 63: 919–932.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *Journal of Molecular Evolution* 39: 306–314.