Consistency of SVDQuartets and Maximum Likelihood for Coalescent-based Species Tree Estimation

Matthew Wascher[1*] and Laura Kubatko[1,2]

[1]*Department of Statistics, The Ohio State University, Columbus, OH, USA*
[2]*Department of Evolution, Ecology, and Organismal Biology, The Ohio State University, Columbus, OH, USA*

[*]*To whom correspondence should be addressed;*

E-mail: wascher.1@osu.edu

Author to receive proofs:

Matthew Wascher

1958 Neil Ave.

Columbus, OH 43210

FAX: 614-292-3648

E-mail: wascher.1@osu.edu

Abtract– Numerous methods for inferring species-level phylogenies under the coalescent model have been proposed within the last 20 years, and debates continue about the relative strengths and weaknesses of these methods. One desirable property of a phylogenetic estimator is that of statistical consistency, which means intuitively that as more data are collected, the probability that the estimated tree has the same topology as the true tree goes to 1. To date, consistency results for species tree inference under the multispecies coalescent have been derived only for summary statistics methods, such as ASTRAL and MP-EST. These methods have been found to be consistent given true gene trees, but may be inconsistent when gene trees are estimated from data for loci of finite length (Roch et al., 2019). Here we consider the question of statistical consistency for four taxa for SVDQuartets for general data types, as well as for the maximum likelihood (ML) method in the case in which the data are a collection of sites generated under the multispecies coalescent model such that the sites are conditionally independent given the species tree (we call these data Coalescent Independent Sites (CIS) data). We show that SVDQuartets is statistically consistent for all data types (i.e., for both CIS data and for multilocus data), and we derive its rate of convergence. We additionally show that ML is consistent for CIS data under the JC69 model, and discuss why a proof for the more general multilocus case is difficult. Finally, we compare the performance of maximum likelihood and SDVQuartets using simulation for both data types.

Advances in sequencing technology over the last 20 years have led to widespread availability of large-scale sequence data sets from multiple loci for which the goal is to obtain an estimate of the species-level phylogenetic relationships among the taxa under consideration. Analysis of such data has presented significant computational challenges, however, because inference methods must include models that capture variation at two distinct scales. First, a model for the process by which the phylogenetic histories of individual loci vary given the overall species tree must be developed. The coalescent process (Kingman, 1982b,c,a) is usually used for this purpose. Second, the mutation process arising along the locus-specific phylogenies, typically called gene trees, must be modeled. This is usually accomplished using standard nucleotide substitution models (Liò and Goldman, 1998). Together, these two model components are often referred to as the multispecies coalescent (MSC). Numerous methods for inference of species trees under the MSC have been developed (reviews of these methods can be found in several places, e.g., Liu et al. (2009) and Kubatko (2019)).

Inference of the species phylogeny under the MSC is challenging because the gene trees are not directly observed, and must therefore be integrated over when computing probabilities associated with the DNA sequence data. Consider a species tree with M species labeled $1, 2, \ldots, M$, and suppose that m_j individuals are sampled within each species j. Thus, $\mathcal{M} = \sum_{j=1}^{M} m_j$ is the total number of sequences in the data set. Using the framework of the MSC, we denote the probability density of gene tree history h and associated vector of coalescent times $\mathbf{t}_h$, conditional on species tree topology S and vector of speciation times τ, by $f_{(h,\mathbf{t}_h)|(S,\tau)}$ (see Rannala and Yang (2003) for a description of how to compute this density). We further define a site pattern to be an assignment of states $i_1 i_2 \cdots i_{\mathcal{M}}$ to the $\mathcal{M}$ tips of the tree, such that $i_k \in \{A, C, G, T\}$ for $k = 1, 2, \ldots, \mathcal{M}$, and we denote the probability of site pattern $p^h = i_1 i_2 \cdots i_{\mathcal{M}}$ arising from *gene tree history* $(h, \mathbf{t}_h)$ by $p^h_{(i_1 i_2 \cdots i_{\mathcal{M}})|(h,\mathbf{t}_h)}$. This probability is the usual phylogenetic likelihood along a gene tree, computed assuming one of the standard nucleotide substitution models. The probability of observing site pattern

$p = i_1 i_2 \cdots i_{\mathcal{M}}$ from the *species tree* is then given by

$$p_{i_1 i_2 \cdots i_{\mathcal{M}}|(S,\tau)} = \sum_{h \in \mathcal{H}} \int_{\mathbf{t}_h} p^h_{(i_1 i_2 \cdots i_n)|(h,\mathbf{t}_h)} f_{(h,\mathbf{t}_h)|(S,\tau)} d\mathbf{t}_h \tag{1}$$

where the sum is taken over all gene tree histories $\mathcal{H}$ with corresponding branch lengths $\mathbf{t}_h$ appropriately integrated out. See Chifman and Kubatko (2015) for full details of the calculations.

Note that Equation (1) implies that each site in the sequence alignment is an independent observation from the model; that is, each site represents a draw from the distribution of gene trees given the species tree as specified by the MSC, with subsequent mutation along the sampled gene tree according to one of the standard nucleotide substitution models. We use the term *coalescent independent sites* (CIS) to distinguish data of this type from SNP data, which do not usually include invariable sites. Under this model, a sample of N CIS can be viewed as a sample from the multinomial distribution, where the number of categories is the number of possible sites patterns, $4^{\mathcal{M}}$, and the category probabilities are given by the site pattern probabilities. Thus, assuming that the sites are independent conditional on the species tree, the log likelihood of species tree (S, τ) is given by

$$\mathcal{L}\Big((S,\tau)\Big) = \sum_{q=1}^{4^{\mathcal{M}}} x_q \log(p_q) \tag{2}$$

where x_q is the observed number of sites with pattern q, p_q is the probability of site pattern q under the model, $q = 1, 2, \ldots, 4^{\mathcal{M}}$, and $\sum_{q=1}^{4^{\mathcal{M}}} x_q = N$. We note that the site pattern probabilities are functions of the parameters in the MSC model, including both the branch lengths and the effective population sizes along each branch. This likelihood has been mentioned earlier by Xu and Yang (2016).

The likelihood for multilocus data is more complicated, because in that case sites within a locus share the same gene tree and are thus correlated with one another, unless we condition on the gene tree. Suppose that there are G loci and that locus g has length n_g. Let ${p^h}_{j|(h,\mathbf{t}_h)}$ denote the probability that the site pattern observed for site j within a particular locus arises

from gene tree history $(h, \mathbf{t}_h)$. Then, the multilocus likelihood of species tree $(\mathcal{S}, \tau)$ is

$$\mathcal{L}\Big((\mathcal{S},\tau)\Big) = \prod_{g=1}^{G}\Big(\sum_{\mathcal{H}}\int_{\mathbf{t}_h}\Big(\prod_{j=1}^{n_g} {p^h}_{j|(h,\mathbf{t}_h)}\Big) f_{(h,\mathbf{t}_h)|(S,\tau)} d\mathbf{t}_h\Big) \tag{3}$$

The outermost product is taken over the G loci, and assumes that the loci are independent, conditional on the species tree. Comparing the terms inside this outer product to the expression in Equation (1), we see that within the integral over the gene tree branch lengths, the product over the n_g sites within each gene must be taken. These sites are conditionally independent given the gene tree and branch lengths, but are not independent when conditioning only on the species tree because they share a common gene tree. The product appearing inside the integral makes it difficult to apply standard asymptotic arguments to this expression. Even taking the log of this likelihood, which allows the likelihood based on CIS data (Equation (2)) to be handled in a straightforward way, does not resolve the problem of the product appearing inside the integral. This also makes clear why computation of the species tree likelihood for multilocus data under the MSC model is challenging. In fact, we do not know of any direct implementations that compute this likelihood for trees larger than the four-taxon case we consider here.

To study the convergence properties of SVDQuartets and of maximum likelihood (ML), we consider the case in which $M = 4$ and $m_j = 1$ for all j, that is, we consider four-taxon trees with one sequence sampled in each species. In this case, there are $4^4 = 256$ possible site patterns, 15 rooted species trees, and 3 unrooted species trees. When considering ML to estimate the species tree, we restrict our attention to CIS data and use the likelihood given in Equation (2), given the difficulty in handling the multilocus likelihood discussed above. To find the ML estimate of the species tree for CIS data, one needs to be able to compute the true site pattern probabilities for each possible species tree. Formulas for these site pattern probabilities were given by Chifman and Kubatko (2015) for simple substitution models (e.g., JC69 (Jukes and Cantor, 1969)). Under the JC69 model and using these formulas with a single value of the effective population size parameter, θ, specified for the entire tree, we can find the ML estimate of the species tree by considering each of the 15 rooted species trees

and finding the set of speciation times that maximize the likelihood for each. The tree with the highest likelihood is the ML estimate. We have implemented this method in R using the optim function to carry out the optimization for each topology. Our code can be found at `https://github.com/lkubatko/SpeciesTreeConsistency`.

To obtain an estimate of the four-taxon species tree for SVDQuartets for any data type (both CIS and multilocus data) and for the GTR+I+Γ model or any sub-model, let L denote the set of four taxa under consideration, and suppose that L is partitioned into two sets, L_1 and L_2, such that $|L_1| = |L_2| = 2$. We say that $L_1|L_2$ is a *split*. The split $L_1|L_2$ is *valid* for tree S if the subtrees containing the taxa in L_1 and in L_2 do not intersect; otherwise the split is not valid. For example, consider the tree $((1,2),(3,4))$. The split $12|34$ is valid, while the splits $13|24$ and $14|23$ are not valid.

For each of the three possible splits, the 256 possible site patterns can be arranged into a 16×16 matrix in which the rows of the matrix correspond to possible states for the taxa in L_1 and the columns correspond to possible states for the taxa in L_2. Such a matrix is called a *flattening matrix*, and is denoted $Flat_{L_1|L_2}$. For an empirical data set, the entries of the matrix are the observed frequencies of the site pattern that corresponds to the row and column indices, i.e.,

$$Flat_{12|34} = \begin{pmatrix} & [AA] & [AC] & [AG] & [AT] & [CA] & \cdots & [TT] \\ [AA] & p_{AAAA} & p_{AAAC} & p_{AAAG} & p_{AAAT} & p_{AACA} & \cdots & p_{AATT} \\ [AC] & p_{ACAA} & p_{ACAC} & p_{ACAG} & p_{ACAT} & p_{ACCA} & \cdots & p_{ACTT} \\ [AG] & p_{AGAA} & p_{AGAC} & p_{AGAG} & p_{AGAT} & p_{AGCA} & \cdots & p_{AGTT} \\ [AT] & p_{ATAA} & p_{ATAC} & p_{ATAG} & p_{ATAT} & p_{ATCA} & \cdots & p_{ATTT} \\ [CA] & p_{CAAA} & p_{CAAC} & p_{CAAG} & p_{CAAT} & p_{CACA} & \cdots & p_{CATT} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ [TT] & p_{TTAA} & p_{TTAC} & p_{TTAG} & p_{TTAT} & p_{TTCA} & \cdots & p_{TTTT} \end{pmatrix}$$

For example, the $(3,2)$ entry, p_{AGAC}, is the probability of observing nucleotide A for taxon 1, G for taxon 2, A for taxon 3, and C for taxon 4. When the rows and columns of the matrix correspond to a valid split, the matrix will have rank 10 for data observed perfectly from the model. When the rows and columns correspond to a split that is not valid, the

matrix will be rank 16. The SVDQuartets method constructs three matrices (one for each of the three possible splits for four taxa), and computes the *SVD score* for each matrix,

$$SVD(L_1|L_2) = \sqrt{\sum_{k=11}^{16} \hat{\sigma}_k^2} \tag{4}$$

where $\hat{\sigma}_k$ is the k^{th} singular value computed for the matrix of observed site pattern frequencies. For observed data, the magnitudes of the 11^{th} through 16^{th} singular values are expected to be small when the matrix corresponds to the valid split, and thus the split $L_1|L_2$ with the lowest $SVD(L_1|L_2)$ is selected. Note that in the case of four taxa, identifying the valid split is equivalent to inferring the unrooted species tree.

Under either criterion for estimation, we denote the estimator of the species tree by S^* and the true species tree by S. Intuitively, consistency means that as more data are used to form the species tree estimate, the probability that $S^* = S$ goes to 1. SVDQuartets has been assumed to be statistically consistent, but a formal proof has not been provided. ML is known to be consistent when used to estimate gene trees, but consistency of ML has not been formally examined in the species tree case. In the sections below, we prove that SVDQuartets is consistent for both CIS and multilocus data and that ML is consistent for CIS data. We derive bounds for the error probability of SVDQuartets, and compare both methods using both theory and simulations.

Consistency Results

We first define a generative model for multilocus data.

Definition 0.1. We assume that data are generated from the following statistical model:

1. Population and genome sizes are large enough that the fact that genes and sites are sampled without replacement can be ignored.

2. Define $\mathbf{p}$ to be a vector of multinomial probabilities such that if we select a gene at

random and sample one nucleotide at random from that gene, the unconditional site pattern distribution $\mathbf{X} \sim Multinomial(1, \mathbf{p})$.

3. Define $\{\mathbf{D}\}_{i=1}^{N} \overset{iid}{\sim} F$ such that $E(\mathbf{D_i}) = \mathbf{0}$ and if we select N genes at random, each of the N genes will have multinomial site pattern probabilities $\mathbf{p_i}$ where $\mathbf{p_i} \overset{d}{=} \mathbf{p} + \mathbf{D_i}$.

4. Conditional on $\{\mathbf{D}\}_{i=1}^{N}$, if $\mathbf{X_i}$ are the observed site pattern counts for a sample of n_i nucleotides from gene i, then $\mathbf{X_i} \sim Multinomial(n_i, \mathbf{p_i})$ and the collection $\{\mathbf{X_i}\}_{i=1}^{N}$ is independent.

Note that when $n_i = 1$ for all i, this model generates CIS data, which may thus be considered a special case of multilocus data.

Consistency for Maximum Likelihood for CIS Data

While the literature contains numerous proofs of consistency of ML for estimation of gene trees (e.g., Yang (1994); Rogers (1997); RoyChoudhury et al. (2015); Truszkowski and Goldman (2016), described further below), no such proofs have been given for the case of ML estimation of the species tree, in part because it is not computationally feasible to use ML for species tree estimation under the coalescent model for trees of arbitrary size as discussed above. Some recent attention has also been given to evaluating the consistency of methods other than ML for estimating the species tree, but such work has focused primarily on the case in which multilocus data are collected and summary statistics methods are used to form estimators (Roch et al., 2019) or on the concatenation method (Roch and Steel, 2015). In this section, we formally prove that for CIS data ML estimation of the species tree under the multispecies coalescent described above is statistically consistent for four-taxon trees. We follow the proof of Truszkowski and Goldman (2016) for the case of gene trees, as most of their proof generalizes directly to the species tree case and their proof corrects the omissions of earlier proofs. We refer the reader to Truszkowski and Goldman (2016) for many of the

details.

We first review related work for the case of ML estimation of gene trees, i.e., trees estimated using data from a single locus under one of the standard models of nucleotide substitution. Early proofs of the consistency of ML estimation for gene trees were given by Yang (1994) and Rogers (1997), but more recent examinations by RoyChoudhury et al. (2015) and Truszkowski and Goldman (2016) have found that these proofs are incomplete. RoyChoudhury et al. (2015) explains the problems with these proofs succinctly; we outline their argument here as it will apply to our proof for the species tree case given below. First, note that, assuming identifiability of the gene tree topology, which requires non-zero internal edges (for a proof, see, e.g., Allman et al. (2008), and note condition (2) in Section 2.1), the following proposition results from a straightforward application of the Strong Law of Large Numbers.

Proposition 0.2. Suppose T_0 is the true tree and T_j is any other tree. Then there exists N such that for all $n \geq N$

$$L(T_0) > L(T_j) \ \ a.s. \tag{5}$$

Though it is tempting to use Proposition (0.2) to claim consistency of the ML estimate of the gene tree topology, as noted by RoyChoudhury et al. (2015) and Truszkowski and Goldman (2016), this result is not sufficient to conclude that ML estimation is consistent. To see why, consider the typical definition of consistency of the maximum likelihood estimator (MLE) that states that if $\hat{T}$ is the MLE then

$$\hat{T} \xrightarrow{P} T_0 \tag{6}$$

under a metric $D(\cdot, \cdot)$, where $\xrightarrow{p}$ denotes convergence in probability. In order to guarantee that (6) holds, we either need to show that for any $\epsilon > 0$ there exists some constant $C_\epsilon > 0$ such that

$$\sup_{T_j : D(T_j, T_0) > \epsilon} \{L(T_0) - L(T_j)\} \geq C_\epsilon, \tag{7}$$

so that we are assured there cannot be trees of arbitrarily high likelihood far away from the true tree, or that the parameter space is compact. Under any reasonable metric, it is easy to see that the parameter space is not compact because it does not include trees with branches of length 0, as noted above. Truszkowski and Goldman (2016) provide a corrected proof by defining the following metric and showing that (7) holds for this metric (see Lemma 3 of Truszkowski and Goldman (2016)).

Definition 0.3 (Distance between two trees, Truszkowski and Goldman (2016))**.** For two taxa a and b in tree S, define their distance, $d_S(a, b)$, to be the sum of the lengths of all edges on the path from a to b. Further, define the distance between two trees S_1 and S_2 to be $D(S_1, S_2) = \max_{a,b \in L} |d_{S_1}(a, b) - d_{S_2}(a, b)|$. Note that $D(\cdot, \cdot)$ is a metric as long as all branch lengths are positive.

We now state and prove a modified version of Truszkowski and Goldman (2016)'s gene tree consistency result for the case of species trees estimated from a sample of CIS obtained under the multispecies coalescent.

Theorem 0.4 (Consistency of the ML estimator of the species tree for CIS data)**.** *Let S_N^* denote the MLE of species tree S for a sample of N CIS obtained under the multispecies coalescent. Then $D(S_N^*, S) \to 0$ with probability 1 as $N \to \infty$.*

Our proof follows the general outline given by Truszkowski and Goldman (2016) for the gene tree case. Two crucial steps in their proof must be verified for the species tree case. First, the species tree must be identifiable, which has been established by Chifman and Kubatko (2015) for species trees that satisfy the molecular clock and by Long and Kubatko (2019) for non-clock species trees and for trees in which the effective population sizes vary throughout the tree. Second, a particular function of the pairwise distribution of states at the tips must satisfy a concavity condition. We state and verify this condition in the following proof.

Proof. Because the site pattern counts for a random sample of N CIS follow a multinomial distribution with probabilities given in Equation (1) above, the likelihood function for the ML estimate of the species tree is similar in form to that in the case of a gene tree. Thus Proposition (0.2) and most steps in the consistency proof given by Truszkowski and Goldman (2016) can be verified in a straightforward manner. The only non-trivial condition to be verified in the species tree case is that Lemma 3 of Truszkowski and Goldman (2016) still holds for the particular site pattern probabilities that arise in the species tree setting. This lemma involves some conditions on the pairwise site pattern probabilities, which we define below.

Following the notation of Truszkowski and Goldman (2016), let f_{xy}^{ab} denote the frequency with which taxon a is observed to have state x and taxon b is observed to have state y, where $x, y \in \{A, C, G, T\}$. Let $p_{xy}^{d'}$ denote the probability that taxon a has state x and taxon b has state y, where $x, y \in \{A, C, G, T\}$, when $d_S(a, b) = d'$. To verify Lemma 3 of Truszkowski and Goldman (2016) it is sufficient to verify that the function

$$\sum_{x,y \in \{A,C,G,T\}} f_{xy}^{ab} \log(p_{xy}^{d'}) \tag{8}$$

is concave in d'. Under the JC69 model (Jukes and Cantor, 1969) and the multspecies coalescent model, Chifman and Kubatko (2015) (see their Supplement A) gave explicit formulas for the site patterns probabilities on four-taxon trees. Using these with $\mu = 4/3$ as specified by the JC69 model and θ, the effective population size parameter, set to 0.01, we can sum over pairs of taxa to find

$$p_{xy}^{d'} = \begin{cases} \frac{1}{4} + \frac{225}{304} e^{-4d'/3}, & x = y \\ \frac{3}{4} - \frac{225}{304} e^{-4d'/3}, & x \neq y \end{cases} \tag{9}$$

Ruskino (2018, personal communication) has derived more general expressions as a function the θ parameter that follow the same general form. Using these expressions, it is straightforward to verify that the expression in Equation (8) is concave in d' and thus that Lemma 3 of Truszkowski and Goldman (2016) holds. This establishes Theorem 0.4, and thus the ML estimate of the species tree for four taxa is statistically consistent for CIS data. □

We next consider consistency for SVDQuartets.

Consistency and Error Rate for SVDQuartets

Recall that for the SVDQuartets method, we choose the tree with split $\text{argmin}_{L_1|L_2} SVD(L_1|L_2)$ as our estimate of the species tree. In this section, we prove that this estimator is consistent for multilocus data in the following sense and give its rate of convergence.

Theorem 0.5 (Consistency of SVDQuartets)**.** *Suppose that the conditions of the model proposed by Chifman and Kubatko (2015) are satisfied, and $L_1|L_2^*$ is the true valid split among splits with $|L_1| = |L_2| = 2$. Fix $\epsilon > 0$. Assume $\lim_{N\to\infty} \max_{i=1\ldots N}\{n_i\} = K < \infty$, and that all of the entries of the vector $\mathbf{p}$ are strictly between 0 and 1. Then $\exists N_\epsilon$ such that $\forall N \geq N_\epsilon$, $\mathbb{P}\left(argmin_{L_1|L_2} SVD(L_1|L_2) \neq L_1|L_2^*\right) < \epsilon$.*

We give the details of the proof of Theorem 0.5 in the remainder of this section. The result follows from consistency of the $\hat{p}_{ij}$ in the flattening matrix and the fact that singular values of a matrix satisfy a Lipschitz condition with respect to perturbations of the matrix (see Golub and VanLoan (2013)). The assumption that all of the entries of the vector $\mathbf{p}$ are strictly between 0 and 1 may seem arbitrary, but if this is not true, then we are considering the problem of estimating site pattern probabilities for sites that either always or never occur, and such cases are neither realistic nor interesting.

Lemma 0.6. *[Corollary 8.6.2 of Golub and VanLoan (2013)] Let $A, E \in \mathbb{R}^{m\times n}$ with $m \geq n$, and let σ_i, $i \in \{1,\ldots n\}$, denote the singular values in descending order. Then for $i \in \{1,\ldots n\}$, $|\sigma_i(A+E) - \sigma_i(A)| \leq ||E||_2 = \sigma_1(E)$.*

We first establish that the p_{ij} are consistently estimated (for two reasonable uses of multilocus data) and give their asymptotic error. The estimator $\hat{\mathbf{p}}_2$ is currently used by SVDQuartets as implemented in PAUP* (Swofford, 2019).

Lemma 0.7. *Suppose data are generated as in Definition 0.1 with N and n_i, $i = 1 \dots N$, such that $\lim_{N\to\infty} \max_{i=1\dots N}\{n_i\} = K < \infty$, and that all of the entries of the vector $\mathbf{p}$ are strictly between 0 and 1. Consider the estimators*

$$\hat{\mathbf{p}}_1 = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathbf{X}_i}{n_i} \qquad \text{and} \qquad \hat{\mathbf{p}}_2 = \frac{1}{\sum_{i=1}^{N} n_i}\sum_{i=1}^{N}\mathbf{X}_i.$$

The following hold:

1. *Let $\hat{p}_1^j$ be the j^{th} entry of $\hat{\mathbf{p}}_1$, and let p^j be the j^{th} entry of $\mathbf{p}$ in Definition 0.1. Then for any $\epsilon > 0$,*

$$\mathbb{P}\left(|\hat{p}_1^j - p^j| > \epsilon\right) \leq 2\exp(-2N\epsilon^2).$$

2. *Let $\hat{p}_2^j$ be the j^{th} entry of $\hat{\mathbf{p}}_2$. Then for the K defined in Theorem 0.5 and $\epsilon > 0$,*

$$\mathbb{P}\left(|\hat{p}_2^j - p^j| > \epsilon\right) \leq 2\exp(-2N(\frac{\epsilon}{K})^2).$$

Proof. In both cases we will apply Hoeffding's inequality.

1. Let $X_{i,j}$ denote the j^{th} entry of $\mathbf{X}_i$, and let $W_{i,j} = \frac{X_{i,j}}{n_i}$. Then the $W_{i,j}$ are bounded between 0 and 1 and independent with respect to the i index for any given j. Thus we can apply Hoeffding's inequality to conclude

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N} W_{i,j} - E\left(\frac{1}{N}\sum_{i=1}^{N} W_{i,j}\right)\right| > \epsilon\right) \leq 2\exp(-2N\epsilon^2). \tag{10}$$

The first term in the expression above is $\hat{p}_1^j$. The second term is equal to $\frac{1}{N}\sum_{i=1}^{N} E(W_{i,j})$. Now let $D_{i,j}$ be the j^{th} entry of $\mathbf{D}_i$. Then

$$E(W_{i,j}) = E(E(W_{i,j}|D_{i,j})) = E(p^j + D_{i,j}) = p^j$$

since our generative model assumes that $E(\mathbf{D}_i) = \mathbf{0}$. Thus, 10 states that $\mathbb{P}\left(|\hat{p}_1^j - p^j| > \epsilon\right) \leq 2\exp(-2N\epsilon^2)$ as desired.

2. Again let $X_{i,j}$ denote the j^{th} entry of $\mathbf{X}_i$, and now let $W_{i,j} = \frac{X_{i,j}}{K}$ where $K = \lim_{N\to\infty} \max_{i=1\ldots N}\{n_i\}$ which we have assumed is finite as stated in Theorem 0.5. Then the $W_{i,j}$ are bounded between 0 and 1 and independent with respect to the i index for any j. Note that

$$\hat{p}_2^j = \left(\frac{NK}{\sum_{i=1}^N n_i}\right)\frac{1}{N}\sum_{i=1}^N W_{i,j}$$

and

$$\begin{aligned} E\left[\left(\frac{NK}{\sum_{i=1}^N n_i}\right)\frac{1}{N}\sum_{i=1}^N W_{i,j}\right] &= \left(\frac{NK}{\sum_{i=1}^N n_i}\right)\frac{1}{N}\sum_{i=1}^N E(W_{i,j}) \\ &= \left(\frac{NK}{\sum_{i=1}^N n_i}\right)\sum_{i=1}^N \frac{n_i p^j}{K} = p^j \end{aligned}$$

where the second equality above holds because $E(W_{i,j}) = E(E(W_{i,j})|D_{i,j}) = E(\frac{n_i}{K}(p^j + D_{i,j})) = \frac{n_i p^j}{K}$.

Thus, noting that $\frac{NK}{\sum_{i=1}^N n_i} \leq K$, we have

$$\begin{aligned} \mathbb{P}\left(|\hat{p}_2^j - p^j| > \epsilon\right) &= \mathbb{P}\left(\left|\left(\frac{NK}{\sum_{i=1}^N n_i}\right)\frac{1}{N}\sum_{i=1}^N W_{i,j} - E\left[\left(\frac{NK}{\sum_{i=1}^N n_i}\right)\frac{1}{N}\sum_{i=1}^N W_{i,j}\right]\right| > \epsilon\right) \\ &= \mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^N W_{i,j} - E\left(\frac{1}{N}\sum_{i=1}^N W_{i,j}\right)\right| > \frac{\epsilon}{NK/\sum_{i=1}^N n_i}\right) \\ &< \mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^N W_{i,j} - E\left(\frac{1}{N}\sum_{i=1}^N W_{i,j}\right)\right| > \frac{\epsilon}{K}\right) \\ &\leq 2\exp(-2N(\epsilon/K)^2) \end{aligned}$$

where the last inequality is the result of applying Hoeffding's inequality.

It might appear from the above bounds that $\hat{p}_1$ should be preferred to $\hat{p}_2$ because the K term does not appear in the bound for $\hat{p}_1$, resulting in a smaller bound in that case. However, the bound in part (2) of the lemma is not tight. Rather, allowing the K term to appear in the exponent is simply a convenient way of dealing with the heterogeneity

arising from the term $\frac{1}{\sum_{i=1}^{N} n_i}$. We discuss the relative merits of $\hat{p}_1$ and $\hat{p}_2$ in the later section "SVDQuartets for Multilocus Data."

If $\lim_{N\to\infty} \max_{i=1\ldots N}\{n_i\} = K < \infty$ does not hold, $\hat{\mathbf{p}}_1$ and $\hat{\mathbf{p}}_2$ are still consistent estimators of $\mathbf{p}$, but the deviations may have thinner tails than the bounds given above. Since it seems unrealistic that this assumption would be violated as in practice genes are finite in length, we do not provide a proof for the case where it does not hold. □

Lemma 0.8. *For any split $L_1|L_2$, $SVD(L_1|L_2) \xrightarrow{p} \sqrt{\sum_{i=11}^{16} \sigma_i^2}$ where σ_i are the descending ordered singular values of the 16×16 matrix $Flat_{L_1|L_2}(P)$.*

Proof. Because $SVD(L_1|L_2)$ is a continuous function of the vector $(\sigma_1, \ldots, \sigma_{16})$, it suffices to show that $\hat{\sigma}_i \xrightarrow{p} \sigma_i$ uniformly in i. Fix $\epsilon, \delta > 0$. We will show $\exists N_{\epsilon,\delta}$ such that $\forall N \geq N_{\epsilon,\delta}$, $\mathbb{P}\left(\sup_i |\hat{\sigma}_i - \sigma_i| > \delta\right) < \epsilon$.

We index the vector of site pattern probabilities $\mathbf{p}$ as $\{p_{ij}\}$ to match their locations in the flattening matrix. Note that this is a modification of the notation in Lemma 0.7 which used vectors to denote the site pattern probabilities. Likewise, we index $\hat{\mathbf{p}}$ as $\{\hat{p}_{ij}\}$. Define $e_{ij} := \hat{p}_{ij} - p_{ij}$, and observe that Lemma 0.7 implies that for any i, j, $\mathbb{P}\left(|e_{ij}| > \delta\right) \leq 2\exp(-2N(\frac{\delta}{K})^2)$. Now choose $N_{\epsilon,\delta}$ large enough so that when $N \geq N_{\epsilon,\delta}$, $\mathbb{P}\left(|e_{ij}| > \frac{\delta}{64}\right) \leq 2\exp(-2N(\frac{\delta}{64K})^2) < \frac{\epsilon}{256}$. Using a union bound, we have

$$\begin{aligned}
\mathbb{P}\left(\max_{i,j}|e_{ij}| > \delta\right) &\leq \sum_{i,j} \mathbb{P}\left(|e_{ij}| > \frac{\delta}{64}\right) \\
&\leq \sum_{i,j} \frac{\epsilon}{256} \\
&< \epsilon.
\end{aligned}$$

Now choose E in Lemma 0.6 to be $E = \{e_{ij}\}$. Then $\sup_i |\hat{\sigma}_i - \sigma_i| \leq ||E||_2$. It is well-known that for any matrix $E \in \mathbb{R}^{k\times k}$, $||E||_2 \leq \sqrt{k}||E||_1 \leq k\sqrt{k}|\max_{i,j}(e_{ij})|$. Applying this

fact with $k = 16$, we have

$$\begin{aligned} P(\sup_i |\hat{\sigma}_i - \sigma_i| < \delta) &> P(16(4)|\max(e_{ij})| < \delta) \\ &= 1 - P\left(|\max(e_{ij})| > \frac{\delta}{64}\right) \\ &> 1 - \epsilon \end{aligned}$$

which gives $\mathbb{P}\left(\sup_i |\hat{\sigma}_i - \sigma_i| > \delta\right) < \epsilon$, as desired. □

We can now prove Theorem 0.5:

Proof. Theorem 1 of Chifman and Kubatko (2015) implies that $\sqrt{\sum_{i=11}^{16} \sigma_i^2} = 0$ if and only if we choose the split $L_1|L_2^*$. Then because we have finitely many (3) splits to choose from, we can find some $c > 0$ such that $\sqrt{\sum_{i=11}^{16} \sigma_i^2} > c$ for any split $L_1|L_2 \neq L_1|L_2^*$. Fix $\epsilon > 0$. Choose $\epsilon^* = \frac{\epsilon}{3}$ and $\delta = \frac{c}{2}$. Then for the $N_{\epsilon^*,\delta}$ that satisfies Lemma 0.8 using ϵ^* and δ, for $N \geq N_{\epsilon^*,\delta}$, we will have $\mathbb{P}\left(SVD(L_1|L_2^*) > c/2\right) < \epsilon/3$ and $\mathbb{P}\left(SVD(L_1|L_2) < c/2\right) < \epsilon/3$ for every $L_1|L_2 \neq L_1|L_2^*$. Then, using the union bound,

$$\begin{aligned} & \mathbb{P}\left(\text{argmin}_{L_1|L_2} SVD(L_1|L_2) \neq L_1|L_2^*\right) \\ \leq\ & \mathbb{P}\left(SVD(L_1|L_2^*) > c/2\right) + \sum_{L_1|L_2 \neq L_1|L_2^*} \mathbb{P}\left(SVD(L_1|L_2) < c/2\right) \\ <\ & \epsilon \end{aligned}$$

which completes the proof and establishes that SVDQuartets is a statistically consistent method for species tree estimation under the MSC. □

We emphasize that this result proves consistency of SVDQuartets for both CIS data and for multilocus data using either of the estimators in Lemma 0.7 above. In both of these cases, the above result also gives a bound on the error rate, as described below.

Corollary 0.9. *When estimating the split with SVDQuartets using a sample of n_i, $i = 1 \ldots N$, loci from each of N genes, there exists a constant $\sigma^* > 0$ such that for large N the probability of choosing an incorrect split is bounded by*

$$\mathbb{P}\left(Error_{SVD}\right) \leq (2)(256)\exp(-2N(\frac{\sigma^*}{128K})^2) \tag{11}$$

Proof. Let σ^* be the smallest in absolute value of the 11^{th} -16^{th} nonzero singular values among all possible splits $L_1|L_2$. Note that as a consequence of Lemma 0.6, for any split and each $\hat{\sigma}_i$,

$$|\hat{\sigma}_i - \sigma_i| \leq 64 \max |e_{ij}|. \tag{12}$$

Let σ_i^F denote the i^{th} singular value for an incorrect split for any $i = 11, \ldots, 16$, and let $\hat{\sigma}_i^F$ denote the corresponding observed value. Applying (12) and assuming that $64 \max |e_{ij}| < |\sigma^*/2|$ gives

$$\begin{aligned} |\hat{\sigma}_i^F| &\geq |\sigma_i^F| - 64 \max |e_{ij}| \\ &\geq |\sigma^*| - 64 \max |e_{ij}| \\ &\geq 2(64 \max |e_{ij}|) - 64 \max |e_{ij}| = 64 \max |e_{ij}| \end{aligned} \tag{13}$$

for each $i = 11, \ldots, 16$.

Applying (12) again to singular values from the true split gives $|\hat{\sigma}_i^T - 0| \leq 64 \max |e_{ij}|$, and we have

$$\begin{aligned} SVD(L_1|L_2) &= \sqrt{\sum_{i=11}^{16} (\hat{\sigma}_i)^2} \\ &\geq \sqrt{\sum_{i=11}^{16} (64 \max |e_{ij}|)^2} \\ &\geq \sqrt{\sum_{i=11}^{16} (\hat{\sigma}_i^T)^2} \\ &= SVD(L_1|L_2^*) \end{aligned} \tag{14}$$

where (14) follows from (13). This establishes that the correct split will be selected by SVDQuartets whenever $64 \max |e_{ij}| < |\sigma^*/2|$.

The probability that SVDQuartets makes an error in selecting the split can thus be given by

$$\mathbb{P}\left(Error_{SVD}\right) \leq \mathbb{P}\left(64 \max |e_{ij}| > \sigma^*/2\right) \leq \sum_{i=1}^{16} \sum_{j=1}^{16} \mathbb{P}\left(|e_{ij}| > \sigma^*/128\right).$$

Recall from Lemmas 0.7 and 0.8 that $\mathbb{P}\left(|e_{ij}| > \epsilon\right) \leq 2\exp(-2N(\frac{\epsilon}{K})^2)$, so

$$\mathbb{P}\left(Error_{SVD}\right) \leq \sum_{i=1}^{16}\sum_{j=1}^{16}\mathbb{P}\left(|e_{ij}| > \sigma^*/128\right) \leq (2)(256)\exp(-2N(\frac{\sigma^*}{128K})^2).$$

□

Note that our proof of consistency and error bound derivation depend on the structure of a four-taxon species tree only insofar as Theorem 1 of Chifman and Kubatko (2015) has only been proven for trees of four taxa and our choice of constants. Should that result be extended to trees with a larger number of taxa, our arguments above imply that the estimator based on the SVD score in such cases would also be consistent for multilocus data and would have an error rate bound of $O(\exp(-2N(\frac{|\sigma^*|}{128K})^2))$.

Comparison of Asymptotic Properties of ML and SVDQuartets

Theoretical Comparison

Shi and Yang (2018) conjecture that SVDQuartets is inefficient compared to ML when both are applied to multilocus data, as measured by the probability of recovering the correct species tree. Note that this is a different notion of efficiency than that which is applied in typical statistical settings, raising the question of whether classical statistical results concerning asymptotic efficiency of ML estimators (see, e.g., Lehmann and Casella (1998)) apply in this case. As mentioned earlier in our discussion of consistency, it is not clear whether ML for the species tree estimation problem satisfies the general conditions of Wald (1949) for consistency. Nonetheless, we have been able to show that both ML for CIS data and SVDQuartets for multilocus and CIS data give statistically consistent estimators of the species tree. We next try to summarize what is know about the asymptotic error probabilities of the methods, as a way of addressing the claim made by Shi and Yang (2018) about the relative efficiency of the two methods.

To our knowledge, error rate bounds for ML when applied to multilocus data have been rigorously derived in only a few special cases. Xu and Yang (2016) showed that in the case of a three-taxon species tree, the probability of choosing the wrong topology when using ML for data consisting of rooted gene trees is approximately

$$\mathbb{P}\left(Error_{ML}\right) \approx C_1 \Phi(-C_2\sqrt{N})$$

for explicit constants C_1 and C_2 that depend on the probabilities of the three possible gene trees that can arise within the species tree (which in turn can be computed from the other parameters.) It is important to note, however, that this result is an approximation rather than a bound. It does not account for the rate at which $\mathbb{P}\left(Error_{ML}\right)$, which is not exactly normal, converges to a normal distribution, and this rate could potentially be slower than the decay of the normal tail given by the approximating expression $C_1\Phi(-C_2\sqrt{N})$. An equivalent result for four-taxon species trees has not been derived.

Another partial result about the error rate of ML estimation comes from the following idea. Suppose that rather than sample n_i sites from each of N loci, we are able to sample gene trees directly, so that we in fact know the topology and branch lengths of each of the N sampled gene trees, $G_1 \ldots G_N$. Letting $\mathbf{p}_l$ denote the observed site patterns (i.e., the alignment) for gene l, we note that in this case,

$$\begin{aligned} \mathcal{L}\Big((S,\tau)|(G_1,p_1),\ldots(G_N,p_N)\Big) &= \prod_{\ell=1}^{N} f(G_\ell|(S,\tau)) \\ &= \mathcal{L}\Big((S,\tau)|G_1,\ldots G_N\Big) \end{aligned} \tag{15}$$

where f is the gene tree density under the MSC. In words, if we observe the gene trees directly, the alignments give no additional information and the likelihood of interest is the species tree likelihood based on the sampled gene trees. Furthermore, since this sampling scheme uses strictly more information than sampling only finite-length alignments, it seems reasonable to assume that its estimation power should be at least as high, (i.e., its error rate

no worse than that of ML based on multilocus sampling) although this also requires a proof to be made fully rigorous.

Results about error rates for trees of any size have been derived for the problem of estimating the species tree topology using gene trees directly. Liu et al. (2010) showed that for their maximum tree method, the probability of choosing the wrong topology is bounded by an expression of the form

$$\mathbb{P}\left(Error_{MT}\right) \leq C_1 \exp(-C_2 N),$$

and that if all populations have the same size, the maximum tree estimator is also the ML estimator. This result is a rigorous upper bound rather than an approximation. We note that it is comparable to our result for SVDQuartets insofar as both bounds take the form $C_1 \exp(-C_2 N)$, albeit likely for different values of C_1 and C_2.

A rigorous comparison of the performance of ML and SVDQuartets is inconclusive, in large part because not enough is known about the performance of ML. Since the result of Xu and Yang (2016) comes from multinomial probabilities, it is likely that applying Hoeffding's inequality in that case would also yield a bound of the form $C_1 \exp(-C_2 N)$ in addition to the approximation given in their work, although we have not rigorously verified this. One might additionally conjecture that such a result holds for species trees with arbitrary numbers of taxa, rather than just the three-taxon species tree. If this is true, then we could say that ML and SVDQuartets both have error rate bounds of the form $C_1 \exp(-C_2 N)$, where the constants C_1 and C_2 likely differ between the methods, but we cannot compare beyond this statement. We hope that scholars interested in comparing the performance of ML and SVDQuartets will derive more complete rigorous results that will allow for a more comprehensive theoretical comparison.

Comparison via Simulation

We conducted several simulation studies to comparatively evaluate the performance of ML and SVDQuartets. In the first simulation study, CIS data were simulated along the four-taxon symmetric and asymmetric species trees by first simulating gene trees using the package COAL (Degnan and Salter, 2005) and then simulating sequence data under the JC69 model (Jukes and Cantor, 1969) using Seq-Gen (Rambaut and Grassly, 1997). The JC69 model was used because Chifman and Kubatko (2015) provided explicit formulas for the site pattern probabilities for four-taxon trees under the coalescent for this model, allowing us to implement the maximum likelihood method in this case. We considered three species trees with all internal branch lengths and all external branch lengths leading to cherries set to the same value, either 0.5, 1.0, or 2.0 in coalescent units. We also consider three species trees with varying branch lengths. In these three cases, all branch lengths were either 0.5 or 1.0, and the placement of the shorter branches was varied between internal and external branches. The precise trees used are given in the captions to Figures 1 and 2, which show the simulation results. For all of the model trees, we set the effective population size parameter $\theta = 4N\mu$ to 0.001, 0.005, or 0.01 for all branches. In addition to examining the performance of SVDQuartets for CIS data, we examined its performance on SNP data by re-running it on each simulated dataset after removing all of the constant sites. Since SNP data are more commonly collected, this will provide an indication of how much information is lost in moving from CIS to SNP data when using SVDQuartets for inference.

For each of the three methods (SVDQuartets for CIS data, SVDQuartets for SNP data, and ML), we examined the proportion of times out of the 500 replicates that each of the methods correctly estimated the unrooted species tree when the total number of sites sampled ranged from 1,000 to 10,000. In some cases, particularly those in which the overall mutation rate is low, as often results from both small effective population size and short branches, the ML algorithm will not converge and/or singular values cannot be computed accurately

enough to infer the tree with SVDQuartets. When this occurred, we discarded that replicate from the summary of that method's performance. If a particular simulation setting had fewer than 100 replicates in which estimation was completed without error, we did not include the result for that setting in the relevant figure.

Our second simulation study considers multilocus data. We applied SVDQuartets as implemented in PAUP*, which ignores information about loci and treats the data as CIS data (this is the common and recommended practice for SVDQuartets at present). Because the multilocus likelihood is not computationally tractable, we approximated ML inference by running the BPP software (Yang and Rannala, 2014; Yang, 2015; Rannala and Yang, 2017; Flouris et al., 2018) with the prior for τ set to IG(3,0.015) and the prior for θ set to IG(3,0.01). We hereafter refer to these as the default priors. We discarded the first 400 samples as burnin, and recorded every other sampled tree for a total of 1,500 samples. For four-taxon trees, the species tree with the highest posterior probability will be generally equivalent to the ML tree. We consider the same model trees as in the first simulation study and the same choices of θ. We used 5, 10, 15, 20, 25, 35, or 50 loci, with 200bp per locus, and replicated each simulation condition 500 times. Replicates in which SVDQuartets failed to return an estimate of the species tree due to numerical imprecision of the singular value computation were discarded, as described above.

For the third simulation study, we considered more difficult species trees, namely those found in the anomaly zone. In particular, we considered the tree found in Xu and Yang (2016), which is given by (Species1:0.48,(Species2:0.44,(Species3:0.4,Species4:0.4):0.04):0.04) when branch lengths are reported in coalescent units. For the analysis in BPP, we considered both default priors, as we did in our second simulation study, and priors suggested by the simulation study in Xu and Yang (2016). Because the current version of BPP uses the inverse gamma (IG) distribution, rather than the gamma distribution used by Xu and Yang (2016), we use inverse gamma prior for θ and τ that have $\alpha = 3$ and mean set to the true value. Another difference from our simulation conditions and the simulation carried

out by Xu and Yang (2016) is the number of sites per locus. Thus, we include one set of simulations with 200bp per loci (as above) and another with 1000bp per loci as in Xu and Yang (2016). Our preliminary simulations with these settings indicated that many more sites were needed for accurate inference for both ML and SVQuartets, and thus for the CIS simulations, we considered the number of sites ranging from 10,000 to 1,000,000. For the multilocus simulations, we considered 50, 100, 150, 200, 300, and 400 loci. Because of the additional computational cost associated with the large number of sites, we used only 100 replicates of each simulation condition. Finally, we considered θ values that differed somewhat from those used above, in order to reproduce the results of Xu and Yang (2016) reported in their Figure 7. In particular, we considered $\theta = 0.05$ (the value used by Xu and Yang (2016)) and $\theta = 0.01$. As mentioned above, we considered both informative and default priors for BPP. For the informative priors, we assumed τ is IG(3.0,0.024), and θ is IG(3,0.1) when the true value of $\theta = 0.05$ and θ is IG(3,0.02) when the true value of $\theta = 0.01$.

Figure 1 shows the results of the first simulation for the symmetric species tree, and Figure 2 shows the results for the asymmetric species tree. In general, both methods are able to accurately infer the unrooted four-taxon species tree with sufficient data. When the model species tree is symmetric (Figure 1), both methods are very accurate when the branch lengths within the model species tree are equal, though shorter branch lengths and lower values of θ (settings which correspond to lower overall mutation rates) are more difficult for both methods. When branch lengths vary within the tree for the symmetric case, the first varying lengths setting, which corresponds to short internal branch lengths, was most difficult for both methods, though ML in general showed higher error than SVDQuartets for all three choices of θ. Results for the asymmetric model species tree (Figure 2) were likewise similar for both methods, with shorter internal branch lengths corresponding to lower accuracy for both methods. An important observation is that SVDQuartets does not decrease in accuracy when applied to SNP data as compared to CIS data. This can be explained by the observation that constant site patterns do not play a role in the reduced

rank result of Chifman and Kubatko (2015) that is the basis for the SVDQuartets method.

Figure 3 shows the results of the second simulation for the symmetric species tree, and Figure 4 shows the results for the asymmetric species tree. In most cases, the accuracy of SVDQuartets is lower than that of BPP, which is not surprising given that BPP is designed explicitly for multilocus data and SVDQuartets is designed for CIS data. It is clear that as the number of loci increases and the branches become longer, BPP accurately infers the true four-taxon species tree, while the performance of SVDQuartets lags behind. This suggests that the SDVDQuartets method may be most useful for genome-scale multilocus data, a setting in which the asymptotic consistency result suggests good performance and in which Bayesian methods become computationally expensive, while Bayesian methods such as BPP may be more appropriate when a more limited number of loci are available. We further compare the two frameworks (i.e., a likelihood-based framework such as BPP and the SVDQuartets methods) in the Discussion.

The results of the third simulation study are shown in Figure 5 for both CIS data (first row) and multilocus data (second row). As expected, for species trees for which there are anomalous gene trees, estimation of the correct species tree is more difficult for both methods, and more data are required to achieve reasonable accuracy. In the case of CIS data, SVDQuartets performs well with accuracy near 100% once a sufficient amount of data are available, in this case more than 500,000 sites. It is again worth noting that the accuracy of the method applied to SNP data is nearly identical to the CIS case, suggesting that SVDQuartets will be a very effective method when genome-scale SNP data are available. The performance of ML lags behind, likely due to the rather low number of informative sites available when species tree branch lengths are extremely short. For example, when branch length are short and θ is not very large, most of the site patterns generated will be constant (i.e., invariable) site patterns. These site patterns are not informative for either SVDQuartets or ML. The next most frequently occurring site patterns will be those with a different nucleotide in only one species. Such site patterns provide no information about

topology for ML, since they don't provide information about which two taxa are most closely related. However, these site patterns are informative for SDVQuartets, because the reduced rank result on which the method is based uses the relationship that site patterns $xxxy$ and $xxyx$ should occur in equal frequency. Thus it is reasonable that SVDQuartets performs better than ML for CIS and SNP data in low-information settings such as this.

For the multilocus setting, we compared the performance of BPP with both informative and default priors, and we note that the choice of prior has an impact on the resulting inference. This is particularly apparent when $\theta = 0.05$ (Figure 5(d)), where we see that the proportion correct decreases by $\sim 10\%$ when default priors are used rather than priors centered at the true value of θ. This is because this particular choice of θ is larger than the typically-observed empirical values over which the default priors are centered. We selected this value in attempt to reproduce the high accuracy of BPP reported by Xu and Yang (2016) for this species tree. However, we also considered the more realistic value of $\theta = 0.01$ (Figure 5(c)), where we see that the effect of the prior is less substantial, likely because the default prior puts more weight close to this value. However, BPP's accuracy decreases for this value of θ, which can be attributed to the fact that the mutation rate is lower, resulting in fewer informative sites and therefore less information available for inference. Also notable is the effect of locus length in Figure 5(c) and (d), with shorter loci resulting in lower accuracy. Only with informative priors and the larger value of θ is accuracy substantially above 60% achieved for BPP when loci are 200bp in length. The strong performance of BPP noted by Xu and Yang (2016) for this tree is only achieved with a large value of θ, informative priors, and long loci (1000bp each).

We note that the performance of SVDQuartets is poor overall for the multilocus setting, with accuracy only around 50% for most conditions examined. Though this is similar to BPP's accuracy for default priors, short loci, and a lower value of θ, all of which reflect common empirical conditions, it is clear that an increase in information available, either through longer loci or a larger value of θ, benefits BPP more than SVDQuartets. This

result, together with the encouraging results for the CIS and SNP cases in Figure 5(a) and (b), further support our assertion that SVDQuartets may be most effective when the amount of data available, whether multilocus or SNP, is very large – precisely the situations in which Bayesian methods become more computationally expensive.

SVDQuartets for Multilocus Data: $\hat{\mathbf{p}}_1$ vs. $\hat{\mathbf{p}}_2$

SVDQuartets was originally formulated for CIS data, and is easily applied to SNP data, as the constant patterns present in CIS data and absent in SNP data do not impact the reduced-rank results that form the theoretical basis of the method (see Chifman and Kubatko (2015) for details). However, in many cases, multilocus data have been sequenced and are already available. Recall that Theorem 0.5 showed that SVDQuartets is consistent when using either of

Estimator 1: $\hat{\mathbf{p}}_1 = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathbf{X}_i}{n_i}$

Estimator 2: $\hat{\mathbf{p}}_2 = \frac{1}{\sum_{i=1}^{N} n_i}\sum_{i=1}^{N}\mathbf{X}_i$

to estimate site pattern probabilities when multilocus data are generated via Definition 0.1. A natural question is then whether one of these estimators should be preferred. **?** examine this question and find that neither is uniformly better; rather the relative performance depends on the distribution F in Definition 0.1. Very generally, when F is concentrated around some value, $\hat{\mathbf{p}}_2$ is better while when F is spread out, $\hat{\mathbf{p}}_1$ is better. For further discussion, we refer readers to **?**, noting that $\hat{\mathbf{p}}_1$ corresponds to their arithmetic average while $\hat{\mathbf{p}}_2$ corresponds to their weighted average.

Discussion

Our work gives the first consistency results for four-taxon species tree inference under the coalescent model for SVDQuartets for both CIS and multilocus data and for maximum likelihood for CIS data. Previous consistency results for maximum likelihood were only derived in the case of gene trees. In addition, we have proved that the SVDQuartets estimator has asymptotic error probability $O(\exp(-CN))$ for CIS and multilocus data, where N is the number of loci. The constant C probably depends on the structure of the tree being estimated, but our simulations show that it does not appear to be particularly unreasonable in a variety of scenarios. We compare the performance of SVDQuartets and ML theoretically, and find that what is known rigorously is not sufficient to confirm the conjecture of Shi and Yang (2018) that ML is more efficient than SVDQuartets; rather the comparison is inconclusive, in large part because not enough is known about the performance of ML.We can only note that our error bounds for SVDQuartets and those conjectured from partial theoretical results for ML both take the form $O(\exp(-CN))$ where the constant C may differ between the methods.

In our simulations, we assumed that the effective population size, θ, was constant throughout the tree. However, for empirical data, θ may vary from branch to branch, or even along branches within the tree. It is therefore important to note that our proofs of consistency did not rely on the assumption of constant effective population size. In the case of consistency of ML for CIS data, identifiability is known to hold when θ varies through the tree (Long and Kubatko, 2019) and expressions analogous to that in Equation (9) can be obtained for varying effective population sizes (Rusinko, 2018). In the case of the consistency of SVDQuartets, recent work (Long and Kubatko, 2019) has established that the method holds in the case of varying θs, as well as in the absence of a molecular clock. Thus, the consistency result for SVDQuartets applies to a wide variety of mechanisms for data generation.

Our simulations demonstrate comparable performance for both ML and SVDQuartets

for CIS data, while ML (as implemented in BPP) generally performs better with multilocus data. Importantly, our first simulation shows that SVDQuartets can be applied to SNP data without any loss of power to infer the true species tree, making it a good choice for computationally efficient analysis of SNP data under the MSC. Examination of the performance of these methods in the anomaly zone indicates that BPP can be sensitive to the choice of prior and to the number of sites within the loci, while SVDQuartets may require a large number of loci to obtain high accuracy. We also note that, at present, BPP implements only the JC69 model, while the theoretical results underlying SVDQuartets hold for the general time reversible (GTR) model and all submodels (Chifman and Kubatko, 2015), as well as for species trees that violate the molecular clock (Long and Kubatko, 2019), making the method quite generally applicable. Given the consistency results derived here, we suggest that for multilocus data, SVDQuartets will be a useful alternative to Bayesian methods such as BPP when the size of the data, in terms of the number of loci and/or the number of species, makes MCMC-based methods computationally prohibitive, i.e., our results indicate that SVDQuartets can be used to achieve consistent estimates of the species tree topology in precisely the cases in which Bayesian methods are currently computationally expensive.

Acknowledgements

We thank Edward Susko, Ziheng Yang, and an anonymous reviewer for helpful comments on earlier drafts of this manuscript that led to its improvement. We are particularly grateful to Dr. Susko for suggesting a correction to our proof of consistency of SVDQuartets, and for several helpful comments about our overall approach.

References

Allman, E. S., C. Ané, and J. A. Rhodes. 2008. Identifiability of a Markovian model of molecular evolution with gamma-distributed rates. Advances in Applied Probability 40:228–249.

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. Journal of Theoretical Biology 374:35–47.

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

Flouris, T., X. Jiao, B. Rannala, and Z. Yang. 2018. Species tree inference with bpp using genomic sequences and the multispecies coalescent. Molecular Biology and Evolution 35:2585–2593.

Golub, G. H. and C. F. VanLoan. 2013. Matrix Computations. Johns Hopkins University Press.

Jukes, T. and C. R. Cantor. 1969. Evolution of protein molecules. Pages 21–123 *in* Mammalian protein metabolism (H. N. Munro, ed.). Academic Press, New York.

Kingman, J. F. C. 1982a. Exchangeability and the evolution of large populations. Pp. 97-112 *in* G. Koch and F. Spizzichino, eds. Exchangeability in probability and statistics. North-Holland: Amsterdam.

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

Kingman, J. F. C. 1982c. The coalescent. Stoch. Proc. Appl. 13:235–248.

Kubatko, L. 2019. The multispecies coalescent. Pages 219–246 *in* Handbook of Statistical Genetics (D. J. Balding, I. Moltke, and J. Marioni, eds.) 4 ed. Wiley.

Lehmann, E. L. and G. Casella. 1998. Theory of Point Estimation. Springer Texts in Statistics Springer-Verlag New York.

Liò, P. and N. Goldman. 1998. Models of molecular evolution and phylogeny. Genome Research 8:1233–1244.

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

Liu, L., L. Yu, L. Kubatko, D. K. Pearl, and S. V. Edwards. 2009. Coalescent methods for estimating multilocus phylogenetic trees. Molecular Phylogenetics and Evolution 53:320–328.

Long, C. and L. Kubatko. 2019. Identifiability and reconstructibility of species phylogenies under a modified coalescent. Bulletin of Mathematical Biology 81:408–430.

Rambaut, A. and N. Grassly. 1997. SeqGen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Computer Applications in Biosciences 13:235–238.

Rannala, B. and Z. Yang. 2003. Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. Genetics 164:1645–1656.

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

Roch, S., M. Nute, and T. Warnow. 2019. Long-branch attraction in species tree estimation: Inconsistency of partitioned likelihood and topology-based summary methods. Systematic Biology 68:281?297.

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

Rogers, J. 1997. On the consistency of maximum likelihood estimation of phylogenetic trees from nucleotide sequences. Systematic Biology 46:354–357.

RoyChoudhury, A., A. Willis, and J. Bunge. 2015. Consistency of a phylogenetic tree maximum likelihood estimator. Journal of Statistical Planning and Inference 161:73–80.

Shi, C.-M. and Z. Yang. 2018. Coalescent-based analyses of genomic sequence data provide a robust resolution of phylogenetic relationships among major groups of gibbons. Molecular Biology and Evolution 35:159–179.

Swofford, D. L. 2019. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Available at https://paup.phylosolutions.com.

Truszkowski, J. and N. Goldman. 2016. Maximum likelihood phylogenetic inference is consistent on multiple sequence alignments, with or without gaps. Systematic Biology 65:328–333.

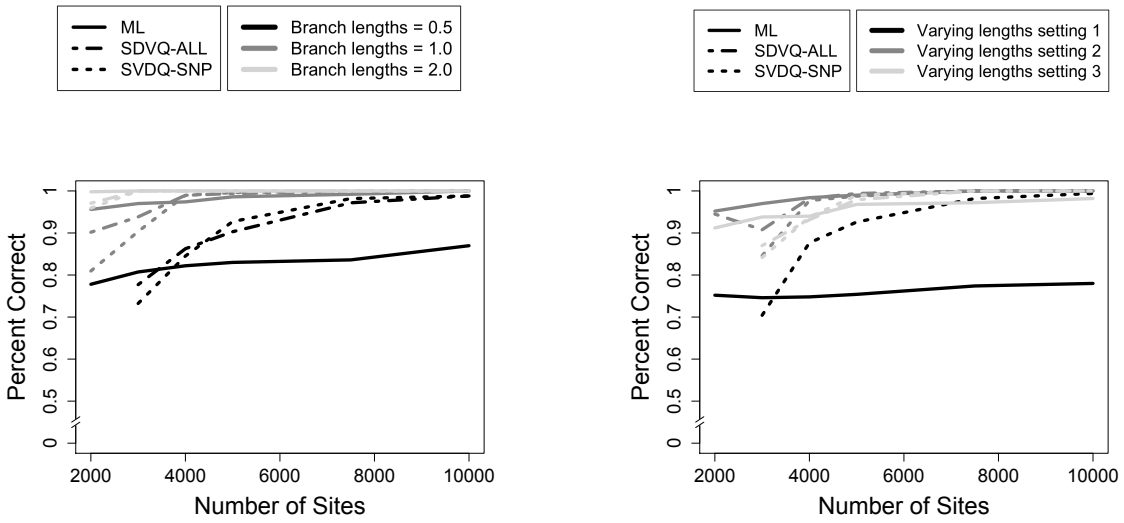
Wald, A. 1949. Note on the consistency of the maximum likelihood estimate. Ann. Math. Statist. 20:595–601.

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

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

Yang, Z. 2015. The bpp program for species tree estimation and species delimitation. Current Zoology 61:854–865.

Yang, Z. and B. Rannala. 2014. Unguided species delimitation using dna sequence data from multiple loci. Molecular Biology and Evolution 31:3125–3135.



(a) $\theta = 0.001$

(b) $\theta = 0.001$

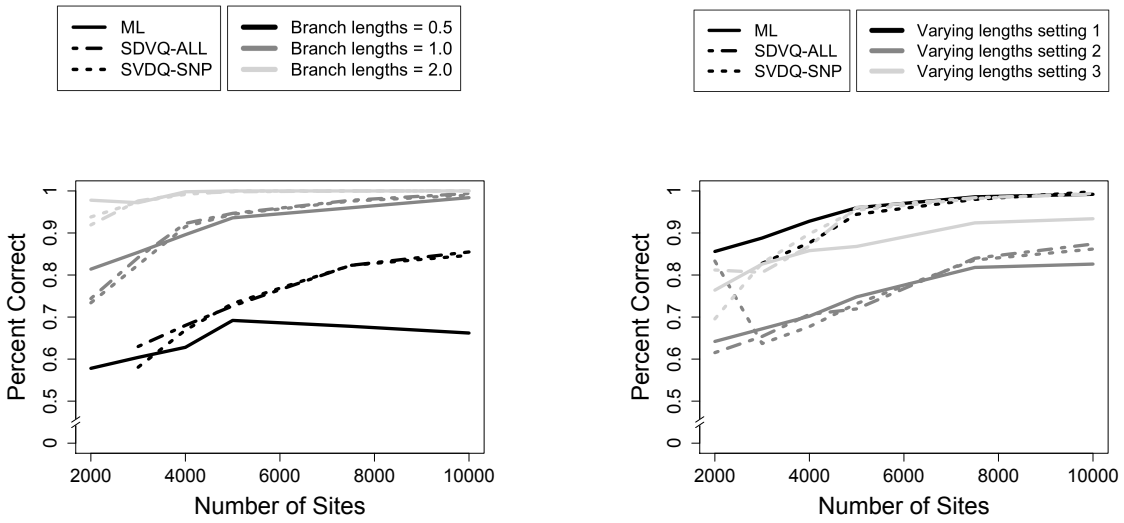
(c) $\theta = 0.005$

(d) $\theta = 0.005$

(e) $\theta = 0.01$

(f) $\theta = 0.01$

Figure 1: Results of the simulation study for the symmetric species tree for CIS data. The x-axis shows the number of CIS, and the y-axis shows the proportion of correctly-estimated unrooted species trees for each method. (a) $\theta = 0.001$, all branch lengths equal to value given in the legend; (b) $\theta = 0.001$, varying branch lengths; (c) $\theta = 0.005$, all branch lengths equal to the value given in the legend; (d) $\theta = 0.005$, varying branch lengths; (e) $\theta = 0.01$, all branch lengths equal to the value given in the legend; (f) $\theta = 0.01$, varying branch lengths. For (b), (d), and (f), setting 1 refers to tree ((Species1:1.0,Species2:1.0):0.5,(Species3:1.0,Species4:1.0):0.5); setting 2 refers to tree ((Species1:0.5,Species2:0.5):1.0,(Species3:0.5,Species4:0.5):1.0); and setting 3 refers to tree ((Species1:1.0,Species2:1.0):0.5,(Species3:0.5,Species4:0.5):1.0).

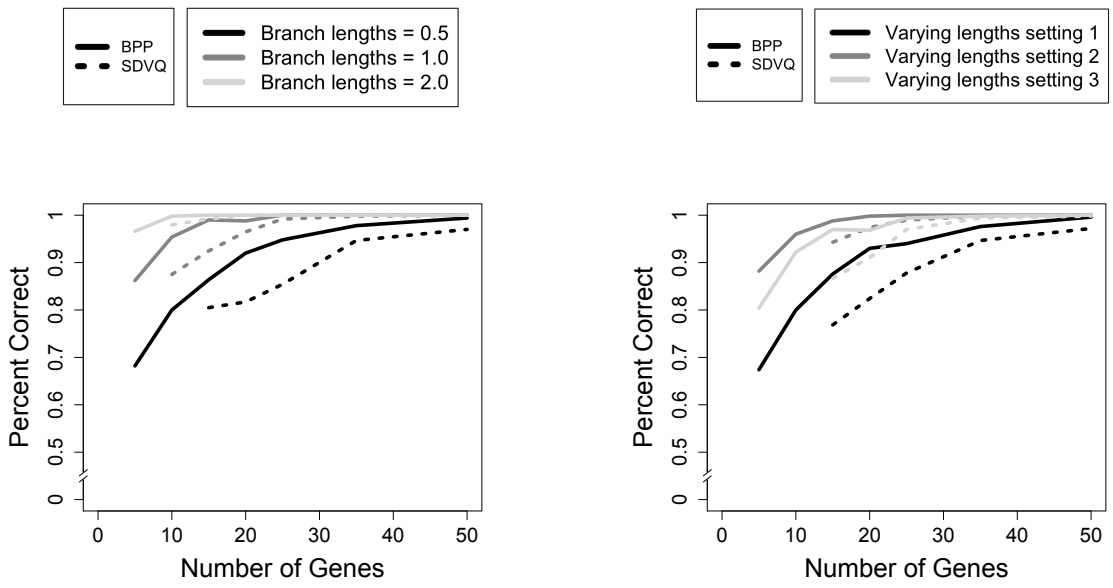


(a) $\theta = 0.001$ (b) $\theta = 0.001$

(c) $\theta = 0.005$ (d) $\theta = 0.005$

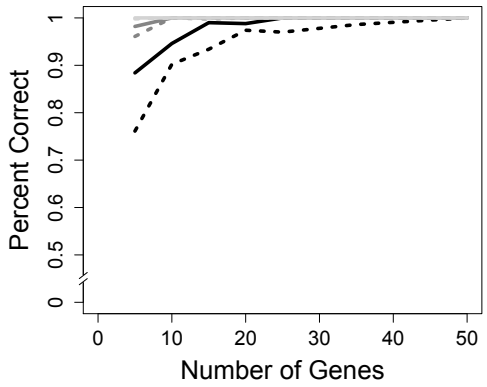
(e) $\theta = 0.01$ (f) $\theta = 0.01$

Figure 2: Results of the simulation study for the asymmetric species tree for CIS data. The x-axis shows the number of CIS, and the y-axis shows the proportion of correctly-estimated unrooted species trees for each method. (a) $\theta = 0.001$, all branch lengths equal to value given in the legend; (b) $\theta = 0.001$, varying branch lengths; (c) $\theta = 0.005$, all branch lengths equal to the value given in the legend; (d) $\theta = 0.005$, varying branch lengths; (e) $\theta = 0.01$, all branch lengths equal to the value given in the legend; (f) $\theta = 0.01$, varying branch lengths. For (b), (d), and (f), setting 1 refers to tree (Species4:2.5,(Species3:1.5,(Species2:0.5,Species1:0.5):1.0):1.0); setting 2 refers to tree (Species4:2.0,(Species3:1.0,(Species2:0.5,Species1:0.5):0.5):1.0); and setting 3 refers to tree (Species4:2.5,(Species3:2.0,(Species2:1.0,Species1:1.0):1.0):0.5).

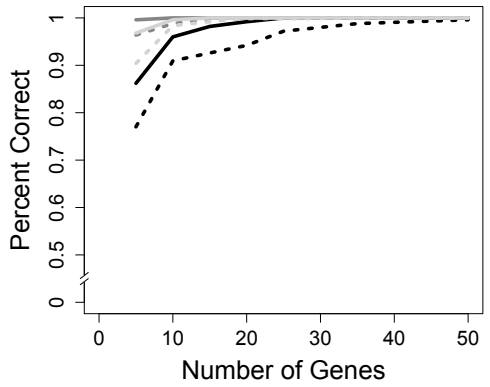


(a) $\theta = 0.001$

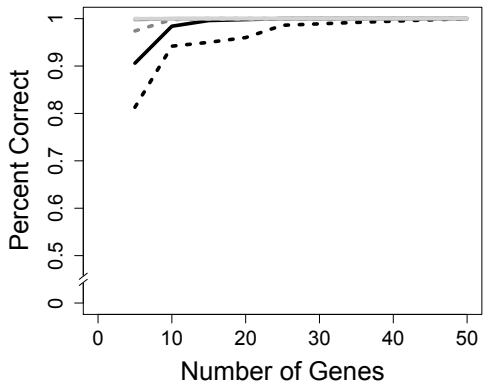
(b) $\theta = 0.001$



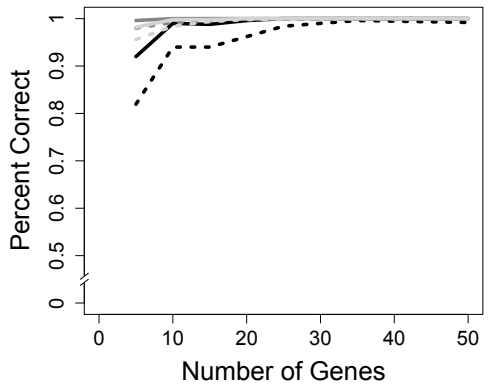
(c) $\theta = 0.005$



(d) $\theta = 0.005$

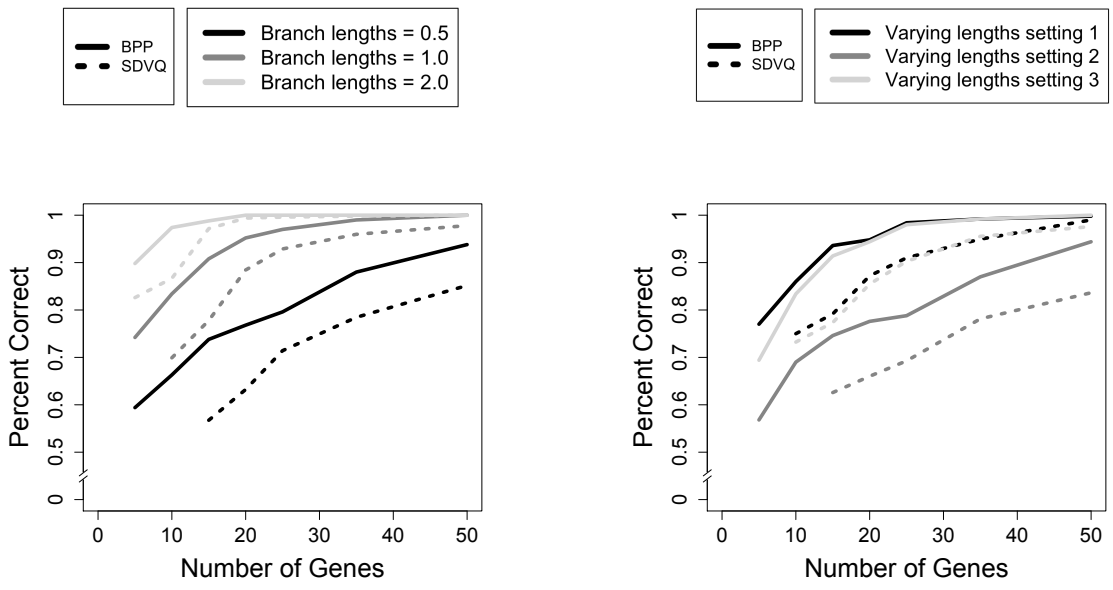


(e) $\theta = 0.01$



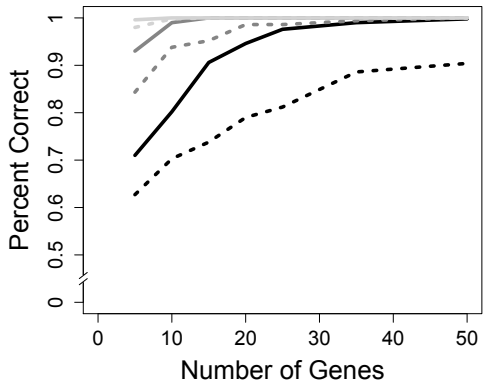
(f) $\theta = 0.01$

Figure 3: Results of the simulation study for the symmetric species tree for multilocus data. The x-axis shows the number of genes, and the y-axis shows the proportion of correctly-estimated unrooted species trees for each method. (a) $\theta = 0.001$, all branch lengths equal to value given in the legend; (b) $\theta = 0.001$, varying branch lengths; (c) $\theta = 0.005$, all branch lengths equal to the value given in the legend; (d) $\theta = 0.005$, varying branch lengths; (e) $\theta = 0.01$, all branch lengths equal to the value given in the legend; (f) $\theta = 0.01$, varying branch lengths. For (b), (d), and (f), setting 1 refers to tree ((Species1:1.0,Species2:1.0):0.5,(Species3:1.0,Species4:1.0):0.5); setting 2 refers to tree ((Species1:0.5,Species2:0.5):1.0,(Species3:0.5,Species4:0.5):1.0); and setting 3 refers to tree ((Species1:1.0,Species2:1.0):0.5,(Species3:0.5,Species4:0.5):1.0).

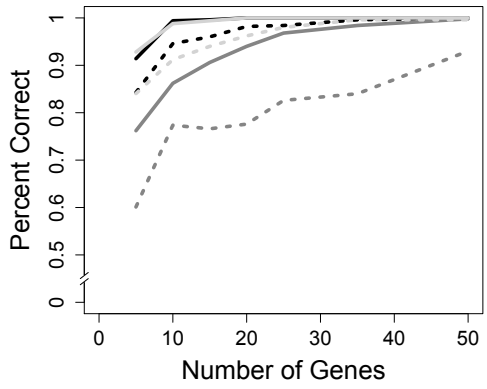


(a) $\theta = 0.001$

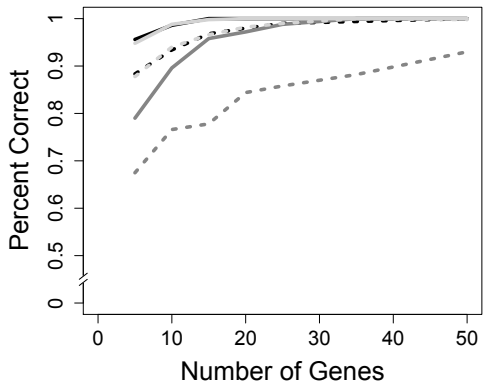
(b) $\theta = 0.001$



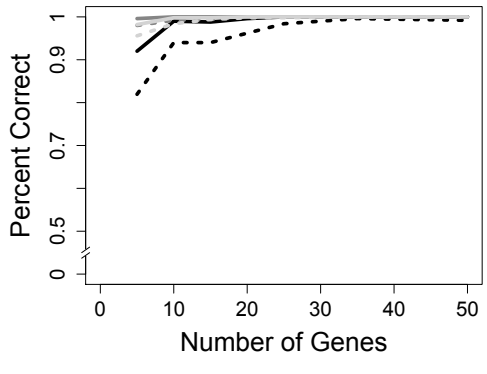
(c) $\theta = 0.005$



(d) $\theta = 0.005$

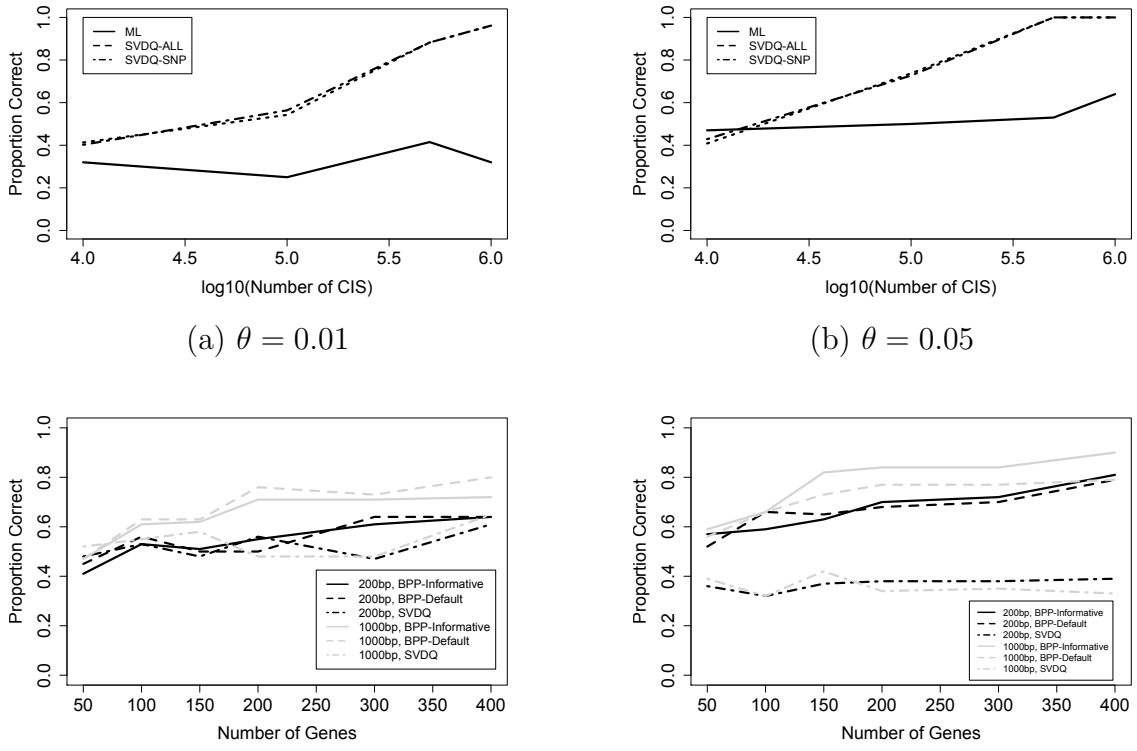


(e) $\theta = 0.01$



(f) $\theta = 0.01$

Figure 4: Results of the simulation study for the asymmetric species tree for multilocus data. The x-axis shows the number of genes, and the y-axis shows the proportion of correctly-estimated unrooted species trees for each method. (a) $\theta = 0.001$, all branch lengths equal to value given in the legend; (b) $\theta = 0.001$, varying branch lengths; (c) $\theta = 0.005$, all branch lengths equal to the value given in the legend; (d) $\theta = 0.005$, varying branch lengths; (e) $\theta = 0.01$, all branch lengths equal to the value given in the legend; (f) $\theta = 0.01$, varying branch lengths. For (b), (d), and (f), setting 1 refers to tree ((Species1:1.0,Species2:1.0):0.5,(Species3:1.0,Species4:1.0):0.5); setting 2 refers to tree ((Species1:0.5,Species2:0.5):1.0,(Species3:0.5,Species4:0.5):1.0); and setting 3 refers to tree ((Species1:1.0,Species2:1.0):0.5,(Species3:0.5,Species4:0.5):1.0).



(a) $\theta = 0.01$ (b) $\theta = 0.05$

(c) $\theta = 0.01$ (d) $\theta = 0.05$

Figure 5: Results of the third simulation study which considers the anomalous species tree of Xu and Yang (2016) given by (Species1:0.48,(Species2:0.44, (Species3:0.4,Species4:0.4):0.04):0.04). In each plot, the x-axis shows the amount of data, and the y-axis shows the proportion of correctly-estimated unrooted species trees for each method.The first row shows the results for CIS data for (a) $\theta = 0.01$ and (b) $\theta = 0.05$. The second row shows the results for multilocus data for (c) $\theta = 0.01$ and (d) $\theta = 0.05$.