Bayesian Species Delimitation Can Be Robust to Guide-Tree Inference Errors

Chi Zhang[1], Bruce Rannala[2], and Ziheng Yang[3,4,*]

[1]*Department of Bioinformatics and Genetics, Swedish Museum of Natural History, Box 50007, SE-104 05 Stockholm, Sweden;* [2]*Genome Center and Section of Evolution and Ecology, University of California at Davis, One Shields Avenue, Davis, CA 95616, USA;* [3]*Department of Genetics, Evolution and Environment, University College London, Darwin Building, Gower Street, London WC1E 6BT, UK; and* [4]*Center for Computational Genomics, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China*
Correspondence to be sent to: Department of Genetics, Evolution and Environment, University College London, Darwin Building, Gower Street, London WC1E 6BT, UK; E-mail: z.yang@ucl.ac.uk.

Species limits are traditionally determined based on morphological, behavioral, and ecological traits. In recent years, genetic sequence data have increasingly been used to delimit species due to the advancement of sequencing technologies and development of statistical methods of data analysis (Wiens 2007; Fujita et al. 2012). Early methods relied on reciprocal monophyly in the reconstructed gene trees, fixed sequence differences between putative species, or simple cut-offs on migration rates or genetic distances between putative species (Sites and Marshall 2004). More recent methods are based on the multispecies coalescent model (Rannala and Yang 2003) and avoid arbitrary cut-offs (Knowles and Carstens 2007). Among the recent methods, the Bayesian method of Yang and Rannala (2010) has a number of advantages over its competitors (Fujita and Leaché 2011). The Bayesian method uses Bayesian model selection to compare different species-delimitation models in the multispecies coalescent framework, and uses reversible-jump Markov chain Monte Carlo (rjMCMC) to estimate the posterior probabilities for different delimitation models. The method accommodates multiple loci, and does not require reciprocal monophyly of inferred gene trees. The underlying multispecies coalescent model accounts for incomplete lineage sorting and species-tree–gene tree conflicts due to ancestral polymorphism. The likelihood calculation on sequence alignments allows the method to make a full use of the information in the data while accounting for the uncertainties in the gene tree topologies and branch lengths. Compared with traditional morphology-based taxonomic practice, which varies widely across taxonomic groups, the Bayesian method infers species status from a genealogical and population genetic perspective and is arguably more objective (Fujita and Leaché 2011; Fujita et al. 2012).

In computer simulations, the Bayesian method was found to have good statistical properties (Leaché and Fujita 2010; Zhang et al. 2011; Camargo et al. 2012), with low false positives (the error of splitting one species into two) and false negatives (the error of failing to recognize distinct species). Simulations also suggest that the method has good power in identifying distinct species in the presence of small amounts of gene flow, and is not misled to infer geographical populations as distinct species when the migration rate is high (Zhang et al. 2011).

To reduce the space of models to be evaluated in the rjMCMC, the implementation of (Yang and Rannala 2010; Rannala and Yang 2013) in the program BPP (for Bayesian Phylogenetics and Phylogeography) requires the user to specify a rooted phylogeny for the populations, called the guide tree. The program then evaluates only those models that can be generated by collapsing nodes on the guide tree. The program currently does not change the relationships among the populations, nor does it split a population into different species.

As a simple evaluation of the impact of the guide tree on species delimitation by BPP, Leaché and Fujita (2010) randomized the populations at the tips of a 10-population guide tree for West African forest geckos and found that the incorrect guide tree caused BPP to over-split. When closely related populations that belong to the same species are incorrectly separated on the guide tree and are grouped with more distant populations, BPP tends to infer all of them as distinct species. However, the analysis of Leaché and Fujita (2010) is on a small scale, and furthermore, the random guide trees generated by permutation may be too wrong, unlikely to be encountered in real data analysis when the guide tree is estimated from real data. Here, we conduct a simulation study to examine the performance

of the method under more realistic scenarios, that is, when the guide tree is inferred from the sequence data.

A number of heuristic methods have been used to construct the guide tree, including:

a) clustering algorithms such as STRUCTURE (Pritchard et al. 2000; Falush et al. 2003), STRUCTURAMA (Huelsenbeck and Andolfatto 2007), or BAPS (Corander et al. 2004), which can assign individuals to populations and even infer a population tree. Those methods are often applied to microsatellite data or single-nucleotide polymorphisms (SNPs).

b) phylogenetic methods such as RAxML (Stamatakis 2006) and MrBayes (Ronquist et al. 2012) applied to either a mitochondrial locus or concatenated nuclear loci.

c) species-tree methods such as BEST (Liu 2008) or *BEAST (Heled and Drummond 2010) applied to multiple nuclear loci.

d) species-discovery methods such as that of O'Meara (2010).

e) empirical population phylogeny based on geographical distributions or morphological and ecological characters.

A useful review of strategies for generating the guide tree used in recent studies of species delimitation by BPP has been provided by Carstens et al. (2013, table 1). Geographical distributions and morphological and ecological features of the populations are always important to defining putative species. However, it is difficult to consider such information in a simulation. In this study, we examine strategies b and c for obtaining a guide tree by analyzing DNA/RNA sequence data. The first approach we examine (strategy b) uses phylogenetic analysis of a mitochondrial locus. Note that in vertebrates, the mitochondrial genome has a much higher mutation rate than the nuclear genome so that the sequence data are more variable and more informative (e.g., Zhou et al. 2012). Furthermore, the effective population size for a mitochondrial locus is only one-fourth that for a nuclear locus, so that incomplete lineage sorting is less likely to occur and the mitochondrial gene tree is more likely to match the species/population phylogeny. This method has been used by Leaché and Fujita (2010), Hamback et al. (2013), Linde et al. (2014), among others. We use the program RAxML (Stamatakis 2006) to infer the unrooted maximum-likelihood (ML) tree and mid-point rooting to generate the rooted tree to be used as the guide tree for BPP. The program is widely used and provides a fast method to infer gene trees using ML. We also used the Bayesian method to infer rooted gene trees for the mitochondrial locus under the molecular clock, using the program BEAST (Drummond and Rambaut 2007), but we expect the results to be similar to the ML method.

TABLE 1. Parameter values used in simulating sequences at the nuclear loci

Species tree	τs	θs
Low-mutation rate		
Tree 1	(0.001, 0.002, 0.003)	0.002
Tree 2	(0.001, 0.001, 0.003)	0.002
High-mutation rate		
Tree 1	(0.01, 0.02, 0.03)	0.02
Tree 2	(0.01, 0.01, 0.03)	0.02

Notes: At the mitochondrial locus, the mutation rate is assumed to be 20 times as high as at a nuclear locus, and the population size is assumed to be one-fourth as large. Thus, $\tau_{mt} = 20\tau_{nuc}$ and $\theta_{mt} = 5\theta_{nuc}$.

The second approach we examine (strategy c) is use of species-tree methods applied to multiple nuclear loci. We use *BEAST (Heled and Drummond 2010) for this purpose. We note that it is possible to apply a traditional phylogenetic method such as ML to the concatenated nuclear data, but concatenation is in general inferior to species-tree methods based on the multispecies coalescent model (see Degnan and Rosenberg [2009] and Edwards [2009] for reviews). The strategy of using *BEAST to infer the guide tree for species delimitation by BPP has been used by Leaché and Fujita (2010), Linde et al. (2014), Satler et al. (2013), among others.

To keep the complexity of our simulation manageable, we do not consider the problem of assignment errors in this study and assume that the individuals are correctly assigned to the populations (see discussions later).

SIMULATION DESIGN

Simulation of Sequence Data

We used two species trees, each of four species, to simulate the sequence data under the multispecies coalescent model (Rannala and Yang 2003). Tree 1 is balanced while tree 2 is unbalanced (Fig. 1). Parameters in the model include three species divergence times (τs) as well as seven population size parameters (θs) for the extant and extinct species on the species tree. Both τ and θ are measured by the expected number of mutations per site. For example, $\tau_{AB} = 0.01$ in species tree 1 of Figure 1a means that the average sequence divergence from the time of the AB ancestor to the present is 1%, whereas $\theta = 0.02$ means that two random sequences sampled from the same population are 2% different on average. We assumed that each of the four species (*A–D*) consisted of two populations (labeled A_1, A_2, B_1, B_2, etc.), so that there are eight populations on the guide tree. We consider two sample sizes, with three or five sequences sampled at each locus from each of the eight populations (i.e., with 24 or 40 sequences in the alignment for each locus). The program MCCOAL in BPP version 2.1c was used to generate gene trees with coalescent times (branch lengths) under the multispecies coalescent model (Rannala and Yang 2003) and to simulate sequence alignments given the gene trees. For each species tree, two sets of parameters

a) Species trees
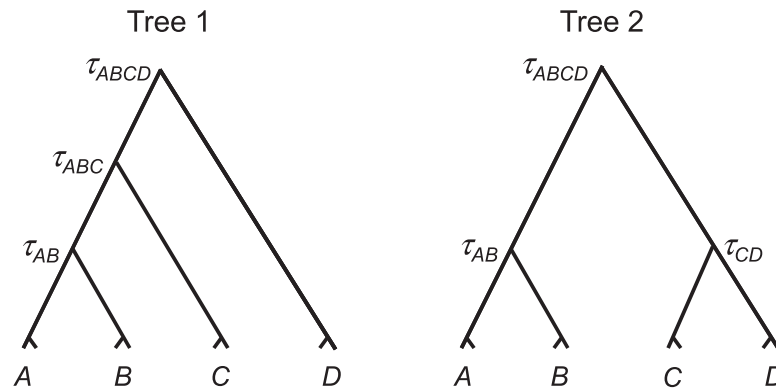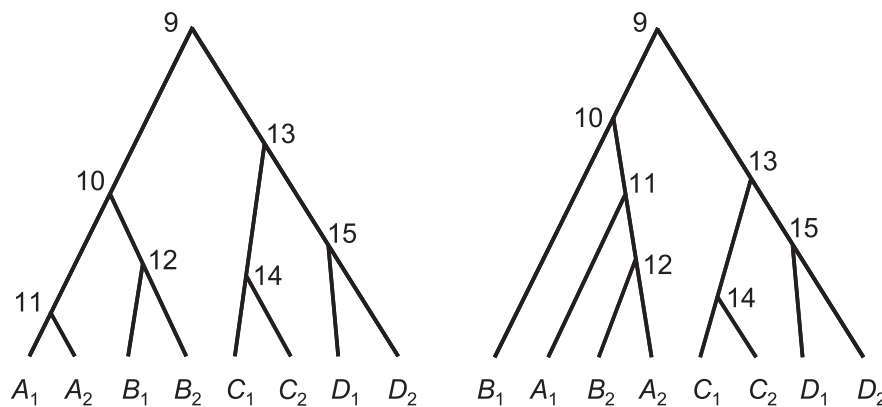

b) Example inferred guide trees

FIGURE 1.   a) Two true species trees used for simulating sequence data under the multispecies coalescent model. Parameters in the model include the three species divergence times (τs) and the population size parameters (θs) for the seven species. Each species is split into two populations in the species-delimitation analysis by BPP, with eight populations on the guide tree. Either three or five sequences are sampled from each populations, with either 24 or 40 sequences in the alignment at each locus. b) Two inferred population (guide) trees, with interior nodes labeled.

were used, mimicking two different mutation rates (Table 1). We simulated either one or five nuclear loci as well as one mitochondrial locus. We assumed that the mutation rate at the mitochondrial locus was 20 times as high as at a nuclear locus and that the population size was one-fourth that for a nuclear locus so that $\tau_{mt} = 20\tau_{nuc}$ and $\theta_{mt} = 5\theta_{nuc}$ (e.g., Zhou et al. 2012). The JC69 mutation model (Jukes and Cantor 1969) was assumed both to generate and to analyze the sequence alignments. Note that the role of the mutation model here is to correct for multiple hits to estimate the gene tree topology and branch lengths, and that JC69 is deemed adequate for analysis of such highly similar sequences (Burgess and Yang 2008); in previous studies, even the infinite sites model produced very similar results (Satta et al. 2004).

The mutation rate was assumed to be constant over lineages (i.e., the molecular clock holds), and across the nuclear loci. The sequence length is 1000 sites for the mitochondrial locus and 500 sites for each nuclear locus. Sequences from the same species are assumed

to coalesce freely (i.e., there is random mating even between individuals from different populations of the same species). We therefore simulated either 6 or 10 sequences from each of the four species and then randomly partitioned them into two populations of 3 or 5 sequences each.

Generation of the Guide Tree from Phylogenetic Analysis Applied to a Mitochondrial Locus(RAxML)

The gene tree was inferred by a ML analysis of the mitochondrial locus using RAxML version 7.4.2. The nuclear loci were not used. The correct population assignment of sequences was assumed. The three or five sequences from each of the eight populations were constrained to be monophyletic. The substitution model used was GTR, since RAxML does not implement the JC69 model. Also, RAxML does not implement the molecular clock and infers unrooted trees instead. We used mid-point rooting to generate a rooted tree, which

was used as the guide tree in later BPP analysis. As an alternative phylogenetic method, we also used the Bayesian method implemented in the program BEAST (Drummond and Rambaut 2007) to infer a rooted gene tree under the molecular clock, to be used as the guide tree in the BPP analysis.

*Generation of the Guide Tree using a Species-Tree Estimation Method Applied to the Nuclear Loci (*BEAST)*

The species tree was inferred from the simulated nuclear loci using *BEAST in BEAST2 (version 2.0.2). Only the nuclear loci were analyzed, and the mitochondrial locus was not used in the *BEAST analysis. Again the correct population assignment of sequences was provided to the program. In other words, the three or five sequences for each of the eight populations (A_1, A_2, B_1, B_2, C_1, C_2, D_1, and D_2) were assigned to the same population, whereas the program estimates the phylogenetic relationships among the eight populations. Note that this is not the same as constraining the three sequences from the same population to be monophyletic on the gene tree. The multispecies coalescent model, while placing constraints on the gene tree, allows non-monophyly of sequences from the same species (see, e.g., figure 1 in Rannala and Yang [2003]). We followed the common practice and used the default improper priors in *BEAST, but note that proper priors may be preferable in real-data analysis. The prior on node ages was specified using a Yule process with an improper prior on the birth rate $f(\lambda) = 1/\lambda$, $0 < \lambda < \infty$. The population size parameters (θs) were assigned improper priors $f(\theta) = 1/\theta$, $0 < \theta < \infty$. The mutation model used was JC69, as in the simulation. The mutation rate was set to 1 so that time is measured by the number of mutations per site. A total of 3000 species trees were collected from the MCMC algorithm by sampling every 2000 iterations (6×10^6 iterations in total). The last 2200 trees were used to generate the maximum *a posteriori* (MAP) tree, to be used as the guide tree in the BPP analysis. In pilot runs, the same analysis was conducted twice to confirm consistency between runs.

Note that as we assume the correct assignment, the only errors that *BEAST and RAxML can make will concern the relationships among the eight populations. Figure 1b shows two possible inferred population (guide) trees. The tree on the left is correct under species tree 2, but the one on the right is wrong regardless of whether species tree 1 or tree 2 is the true tree.

BPP *Analysis*

The guide tree was either the ML tree for the mitochondrial locus inferred by RAxML or the MAP tree inferred from the nuclear loci by *BEAST, as described above. Given the guide tree, the nuclear sequence data (either one locus or five loci) simulated above were analyzed using BPP version 2.2 to delimit species. The mitochondrial locus was not used in the BPP analysis.

The divergence time for the root of the guide tree (τ_{ABCD} or τ_0) and the population size parameters (θs) are assigned diffuse gamma priors G(α, β) with shape parameter $\alpha = 1$ and with the mean of the distribution (α/β) matching the true value used in the simulation. Thus, for the low-mutation rate case, the priors were $\tau_0 \sim$ G(1, 333), with mean 0.003 (3 differences per kb), and $\theta \sim$ G(1, 500), with mean 0.002. For the high-mutation rate case, the priors are $\tau_0 \sim$ G(1, 33) and $\theta \sim$ G(1, 50). Note that the prior for divergence times at other nodes on the guide tree is generated from a uniform Dirichlet distribution (Yang and Rannala 2010). Each analysis was conducted twice, using rjMCMC algorithms 0 (with $\varepsilon = 2$) and 1 (with $\alpha = 2$ and $m = 1$) in BPP. The fine-tuning parameters ε in algorithm 0 and α and m in algorithm 1 are used to propose new parameters in the multispecies coalescent model (θ and τ) when a node on the guide tree is split (Yang and Rannala 2010, equations 3 and 6). The two runs or algorithms were used to check for consistency between runs. For each run, 20,000 samples were collected by sampling every two iterations after a burn-in of 5000 iterations. Samples from the two runs were then combined.

In summary, we considered two species trees (trees 1 and 2 in Fig. 1a), two sample sizes (with three or five sequences from each population), and two mutation rates (Table 1), with a total of eight parameter combinations. For each combination, 1000 replicate data sets were generated. Each replicate data set consisted of one mitochondrial locus and either one or five nuclear loci. Every locus consisted of either 24 or 40 sequences, with either three or five sequences from each of the eight populations. The data were then analyzed using two methods for inferring the guide tree: phylogenetic analysis of the mitochondrial locus using RAxML (and BEAST) and species-tree estimation from the nuclear loci using *BEAST. Those analyses are referred to later as RAxML + BPP (or BEAST + BPP if BEAST was used instead) and *BEAST + BPP. In the RAxML + BPP analysis, RAxML was used to analyze the mitochondrial locus to infer the guide tree, which was then used by BPP to analyze the nuclear data to delimit species. In the BEAST + BPP analysis, BEAST was used as an alternative to RAxML. In the *BEAST + BPP analysis, *BEAST was used to analyze the one or five nuclear loci to infer the guide tree, which is then used by BPP to analyze the same nuclear data to delimit species.

RESULTS

Phylogenetic Errors in Guide-Tree Construction

We will focus on the small sample size, with three sequences sampled from each population, and discuss the results for the large sample size (with five sequences from each population) later. We first examine the phylogenetic errors in the guide-tree construction and then describe their impact on species delimitation.

Given the species trees of Figure 1a and our simulation design, the correct trees for the eight populations (i.e., the

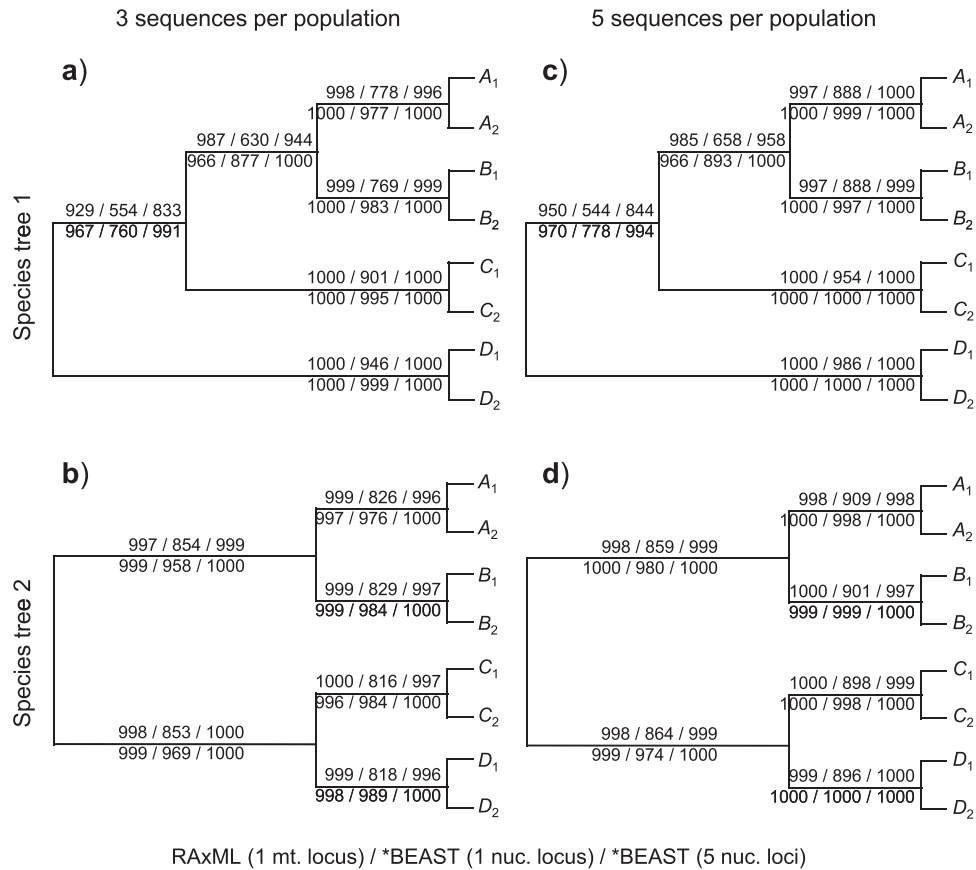3 sequences per population   5 sequences per population

FIGURE 2. Frequency (out of 1000 replicates) at which each clade in the correct population (guide) tree is recovered by RAxML and *BEAST. The numbers above the branch are for the low-mutation rate while those below the branch are for the high rate. In each case, the three numbers are for RAxML for one mitochondrial locus, *BEAST with one nuclear locus, and *BEAST with five nuclear loci.

correct guide trees) are those shown in Figures 2 and 3. The proportion of replicates (out of 1000) in which each clade on the correct guide tree is recovered in the inferred guide tree is also shown (Figs. 2 and 3), calculated using the CONSENSE program in the PHYLIP package version 3.69 (Felsenstein 2005). Note that we used only the population tree topology inferred by the two methods (RAxML/BEAST and *BEAST), and ignored any support measures for clades on the tree, such as the bootstrap support values calculated by RAxML and the posterior clade probabilities calculated by *BEAST. The results show clear effects of the species phylogeny (in particular, the lengths of the internal branches reflecting species divergence times), the mutation rate, and the number of loci. A longer internal branch on the species tree makes the concerned clade easier to recover. A higher mutation rate means that the sequences are more divergent and more informative about the phylogeny (Yang 1998). Similarly, more loci means more data so that the inference is more reliable. Those patterns are easy to understand and are similar to findings from numerous simulation studies that examine the performance of different phylogenetic methods (for review, see Yang [2006, Chapter 6]).

The RAxML analysis of the mitochondrial locus recovered almost all nodes with high probability, except

the difficult clade ABC in tree 1, for which the probability is 93% and 96%, for the low- and high-mutation rates, respectively (Fig. 2a). We also conducted a Bayesian phylogenetic analysis of the same data using the program BEAST (with the same prior settings as for the single nuclear locus), with the results summarized in Figure 3. The probability of recovering the difficult clade ABC in tree 1 is 97% or 98% for the two mutation rates, which are slightly higher than for RAxML (93% and 96%) (Fig. 3a). The slightly poorer performance for RAxML may be due to the fact that the RAxML analysis assumed the more general GTR model with mid-point rooting, which may not be as efficient as the use of the JC69 model and molecular clock rooting (given that the data are simulated under JC69 and the clock). In general, both the ML and the Bayesian analysis of the mitochondrial locus recovered the true clades with very high probability (Figs. 2 and 3). Below we focus on the guide trees inferred using RAxML.

The *BEAST analysis of one nuclear locus performed poorly, especially at the low rate. For example, clade ABC in tree 1 is recovered in only 55% of replicate data sets at the low-mutation rate (Fig. 2a). A single locus at the low-mutation rate does not contain enough information to infer the correct guide tree. However, performance improved dramatically if the mutation rate

3 sequences per population   5 sequences per population

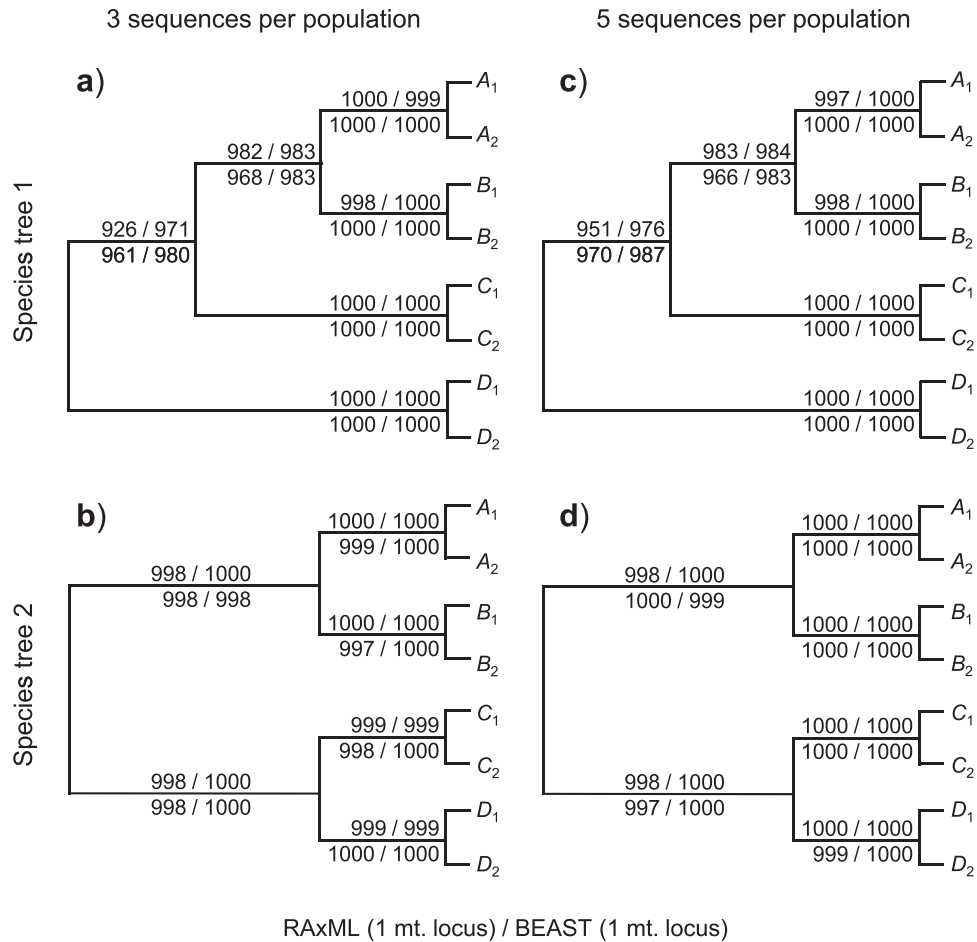RAxML (1 mt. locus) / BEAST (1 mt. locus)

FIGURE 3.   Frequency (out of 1000 replicates) at which each clade in the correct population (guide) tree is recovered by RAxML and BEAST in the analysis of the mitochondrial locus. The numbers above the branch are for the low-mutation rate whereas those below the branch are for the high rate. See legend to Figure 2.

was 10 times higher (with the probability of recovering clade *ABC* in tree 1 to be 76%, Fig. 2a) or if five loci were analyzed (with the probability of recovering clade *ABC* in tree 1 to be 83%, Fig. 2a). The four clades grouping the two populations of each species (A_1A_2, B_1B_2, C_1C_2, and D_1D_2) were recovered with high probabilities on both species trees by both methods except for the *BEAST analysis under the combination of a low rate and one nuclear locus.

False-Positive Rate in Species Delimitation

In the species delimitation analysis by BPP, we considered a split of a node on the guide tree into different species to be well supported only if the posterior probability calculated by BPP was more than or equal to 95%. Thus, we defined the "false-positive rate" as the percentage of data replicates in which two populations of the same species (A_1 and A_2, B_1 and B_2, C_1 and C_2, or D_1 and D_2) are split into different species with posterior probability more than or equal to 95%. For example, if the true species tree is tree 2 of Figure 1a and the inferred guide tree is the tree on the right in

Figure 1b, then we counted a false positive for splitting A_1A_2 if the posterior probability for splitting node 11 was more than or equal to 95%. The results are summarized in Tables 2 and 3 for species trees 1 and 2 of Figure 1a, respectively.

The false-positive errors have contributions from two sources: errors in the inferred guide tree and errors in species delimitation by BPP. In the *BEAST + BPP analysis, the false-positive rate is much lower when five nuclear loci are used than when only one locus is used (Tables 2 and 3). For example, the error rate for splitting clades A_1A_2 and B_1B_2 on species tree 1 at the low-mutation rate was 8.3% for one nuclear locus and approximately 0.7% for five loci. This performance difference is due both to the improved accuracy of guide-tree inference (see Fig. 2) and to increased information content in the BPP analysis. In contrast, in the RAxML + BPP analysis, the performance improvement due to the increased number of nuclear loci is much less dramatic. For example, the error rate for splitting A_1A_2 and B_1B_2 on species tree 1 at the low-mutation rate was 1.2% for one nuclear locus and approximately 0.7% for five loci. In this analysis, there is no reduction in guide-tree

TABLE 2. Percentage of false positives splitting one species into two by BPP with posterior more than or equal to 95% in data simulated using tree 1, with three sequences sampled from each population

Nuclear Loci	Method	A_1A_2	B_1B_2	C_1C_2	D_1D_2
Low-mutation rate					
1 locus	RAxML	0.012	0.012	0.015	0.010
	*BEAST	0.083	0.083	0.047	0.037
5 loci	RAxML	0.009	0.004	0.004	0.004
	*BEAST	0.009	0.006	0.007	0.007
High-mutation rate					
1 locus	RAxML	0.034	0.042	0.035	0.026
	*BEAST	0.044	0.035	0.031	0.027
5 loci	RAxML	0.000	0.000	0.000	0.000
	*BEAST	0.000	0.000	0.000	0.000

Notes: Method refers to two methods for generating the guide tree: phylogenetic method applied to the mitochondrial locus (RAxML), and species-tree inference method applied to the nuclear loci (*BEAST).

TABLE 3. Percentage of false positives splitting one species into two by BPP with posterior more than or equal to 95% in data simulated using tree 2, with three sequences sampled from each population

Nuclear Loci	Method	A_1A_2	B_1B_2	C_1C_2	D_1D_2
Low-mutation rate					
1 locus	RAxML	0.007	0.009	0.003	0.011
	*BEAST	0.053	0.057	0.058	0.064
5 loci	RAxML	0.004	0.006	0.006	0.002
	*BEAST	0.006	0.005	0.009	0.006
High-mutation rate					
1 locus	RAxML	0.030	0.036	0.030	0.030
	*BEAST	0.043	0.037	0.040	0.036
5 loci	RAxML	0.001	0.001	0.002	0.002
	*BEAST	0.000	0.000	0.001	0.000

See note to Table 2.

estimation errors when more nuclear loci are used and the performance improvement is entirely due to the increased information content in the BPP analysis of the nuclear loci. Thus, errors in the guide-tree construction clearly contribute to false-positive errors in species delimitation by BPP.

However, the false-positive rates in those simulations are overall quite low. In all cases except one, the false-positive rates were near or below the nominal rate of 5%. The exception is the case of *BEAST + BPP analysis of one nuclear locus at the low rate for species tree 1, in which BPP splits clades A_1A_2 and B_1B_2 in approximately 8% of replicates, slightly above the nominal 5%. In this case, phylogenetic errors in the guide tree inferred by *BEAST are very common, with clades A_1A_2 and B_1B_2 recovered in only 77% of the replicates (Fig. 2a). To understand why such high errors in the guide-tree inference did not lead to very high false positives in BPP species delimitation, we plot in Figures 4 and 5 the distributions (histograms) of posterior probabilities calculated by BPP (see also Tables 4 and 5 for the medians and quartiles, and online supplementary Figs. S1–S16 for other cases). With one locus (Fig. 4), the posterior probabilities for splitting clades A_1A_2 and B_1B_2 are spread-out. With five

loci (Fig. 5), they shift towards 0 and become highly concentrated. Thus, in the data of a single nuclear locus, the posterior probabilities calculated by BPP did not often reach the 95% cut-off due to the lack of information. With more loci or at the higher mutation rate, the data become far more informative and the posterior probabilities become more extreme. However, in such cases, the guide tree tends to be correctly reconstructed (Fig. 2a) and BPP becomes increasingly accurate with lower rates of false positives and false negatives (Table 2).

The posterior probabilities for splitting clades AB and ABC on species tree 1 reflect the power of BPP to identify distinct species (Figs. 4 and 5 and Tables 4 and 5). Power is high even in the least informative data set of one nuclear locus at the low rate (Fig. 4), and is nearly 100% when five loci are analyzed (Fig. 5).

Note that the false-positive rate we calculate here is a Frequentist property, and there is no theory leading one to expect that the false-positive rate for a Bayesian method (BPP) will be less than 5%. In practice, however, many Bayesian methods are known to also have good Frequentist properties (e.g., Huelsenbeck and Rannala 2004). The BPP method of species delimitation appears to be one such method. Similarly, when the amount of data (e.g., the number of loci) or the amount of information in the data increases, the false-positive rates of BPP for splitting clades A_1A_2, B_1B_2, C_1C_2, and D_1D_2 approach zero, rather than staying at the nominal 5% as in a likelihood ratio test. This is clearly seen from the dramatic reduction in the false-positive rates when the mutation rate was increased by 10-fold or when the number of loci was increased from 1 to 5 (Tables 2 and 3), and from the distribution of the posterior probabilities calculated by BPP for the four clades A_1A_2, B_1B_2, C_1C_2, and D_1D_2 (compare Figs. 4 with 5).

The Impact of the Sample Size

We examined the effect of the sample size by increasing the number of sequences sampled from each population at each locus from 3 to 5, so that there are 40 sequences in the alignment at each locus. The probabilities with which the clades on the correct guide tree are recovered are shown in Figure 2c,d. The recovery probabilities are either very similar to or higher than the corresponding probabilities for the small sample size of Figure 2a,b. For example, in the RAxML analysis of the mitochondrial locus, the probability of recovering clade ABC in tree 1 is 95% and 97% for the low- and high-mutation rates, respectively, when five sequences per population are sampled (Fig. 2c), compared with 93% and 97% for the small sample size of three sequences per population (Fig. 2a). Note that for both the small and large sample sizes, a phylogeny of eight populations is inferred, so that the parameter space (and the number of parameters) of the inference problem remains unchanged even though the gene trees are larger. Thus, a larger sample means more data and more information.

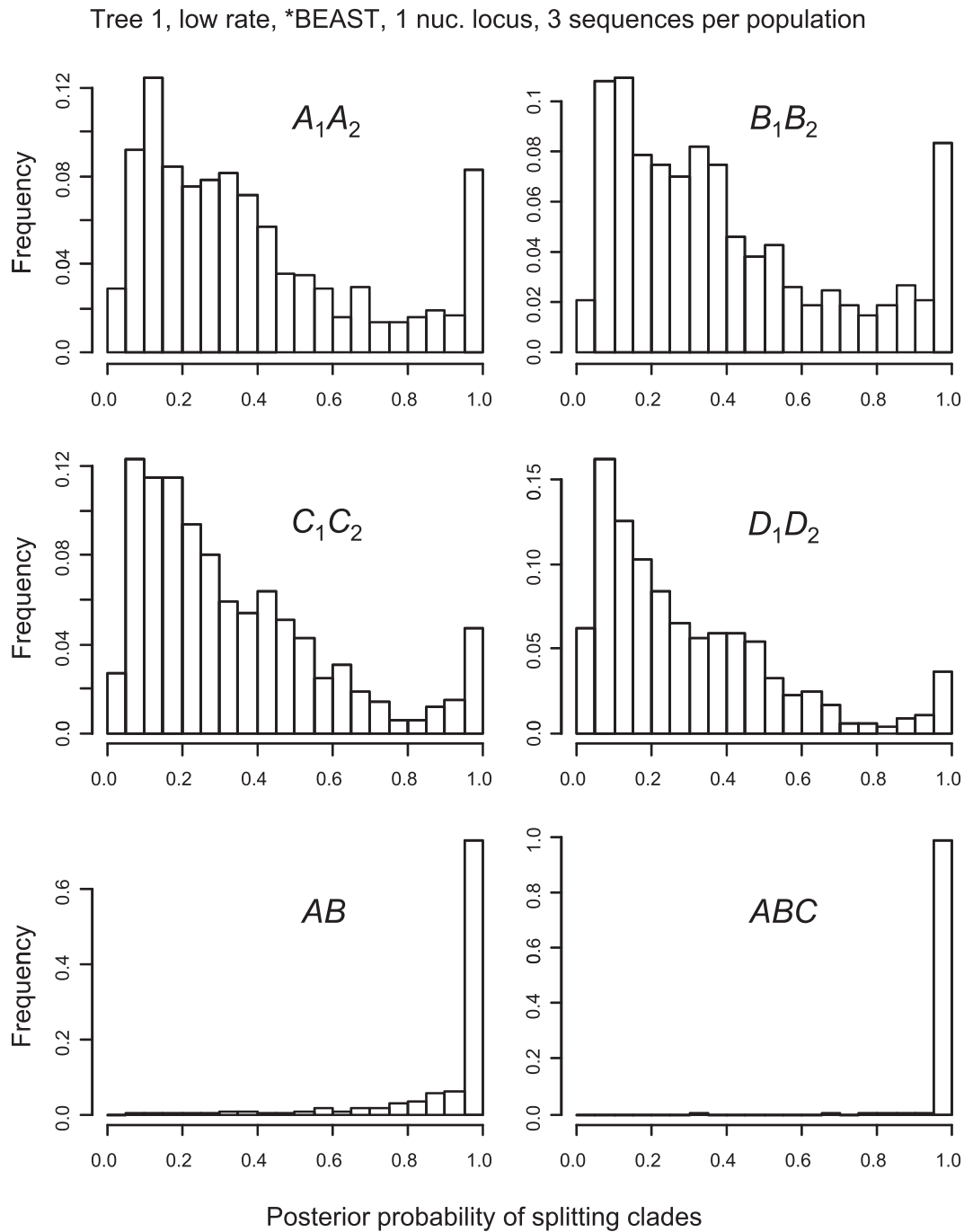Tree 1, low rate, *BEAST, 1 nuc. locus, 3 sequences per population

FIGURE 4.   Histogram of posterior probabilities for splitting clades into different species by BPP in data of one locus, with three sequences sampled from each population at the locus, simulated using tree 1 at the low-mutation rate, when the guide tree was inferred using *BEAST. Each bin is of size 0.05. The frequencies in the last bin for splitting clades A_1A_2, B_1B_2, C_1C_2, and D_1D_2 are the false-positive rates listed in Table 2.

The histograms of posterior probabilities for splitting clades on the correct guide tree for the large sample size are presented in online Supplementary Figures S17–S32. Compared with the corresponding results for the small sample size (Supplementary Figs. S1–S16), species delimitation by BPP performed in general better with the large sample size. For example, in the *BEAST + BPP analysis of one nuclear locus at the low-mutation rate (Supplementary Figs. S1 and S17), the posterior probabilities for splitting clades A_1A_2, B_1B_2, C_1C_2, and D_1D_2 (false positives) are lower in the large sample, indicating lower false positives, whereas the probability for splitting clade AB is higher, indicating higher power. The probability for splitting ABC is approximately 100% for both sample sizes. The better performance of BPP for the large sample size appears to be largely due to the increased information content for species delimitation since the improvement in guide-tree

Tree 1, low rate, *BEAST, 5 nuc. loci, 3 sequences per population
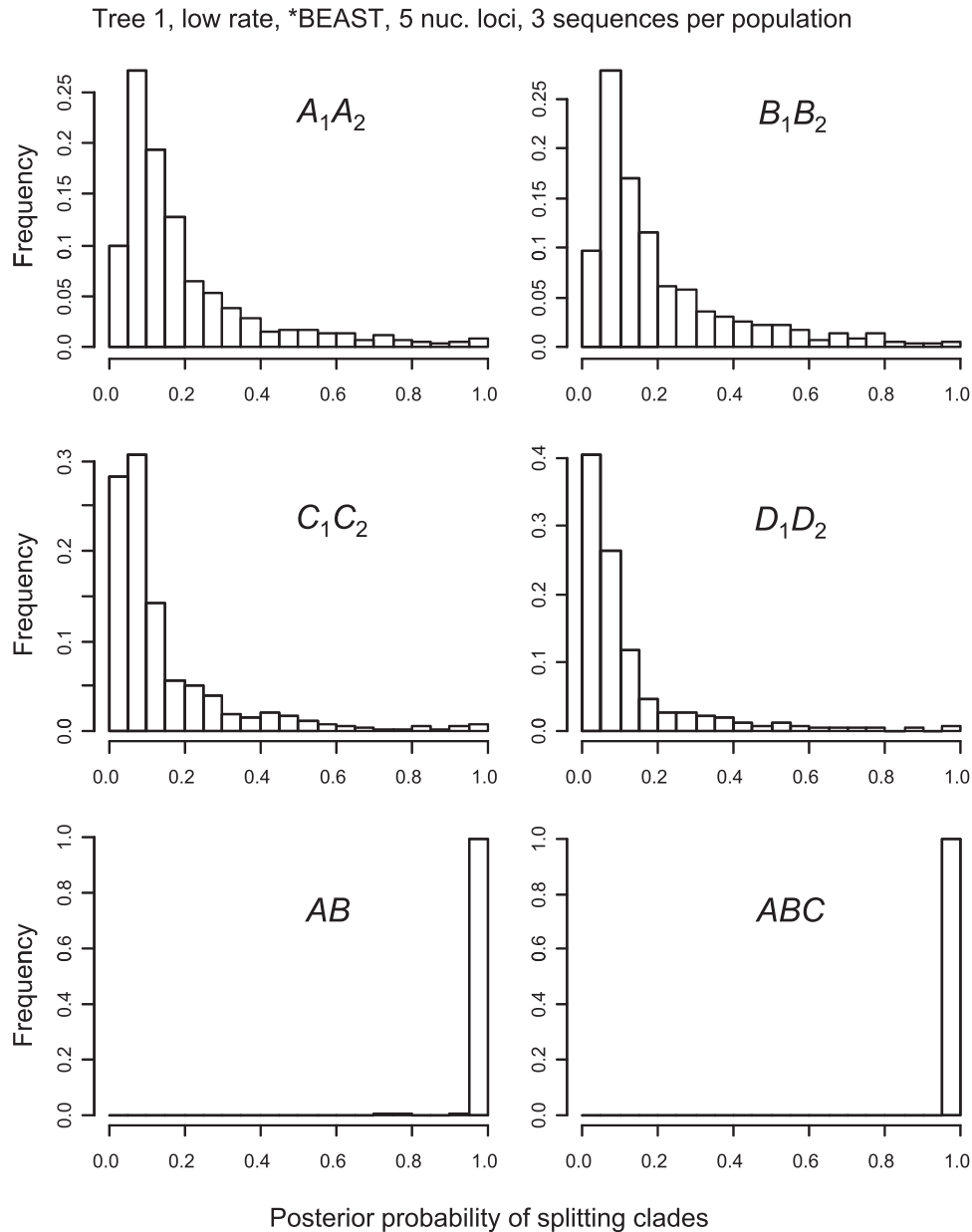
Posterior probability of splitting clades

FIGURE 5. Histogram of posterior probabilities for splitting the clades by BPP in data of five loci, with three sequences per population at each locus, simulated using tree 1 with low-mutation rate, when the guide tree was inferred using *BEAST. See legend to Figure 4.

inference is moderate. A previous simulation found that increasing the number of sequences sampled from the same species improves species delimitation by BPP, leading to both reduction of false positives (over-splitting errors) and increase of power (correctly delimiting distinct species) (Zhang et al. 2011).

DISCUSSION

Impact on Species Delimitation of Errors in the Estimated Guide Tree

We investigated the impact of possible errors in the guide tree on Bayesian species delimitation by BPP,

using two approaches for constructing the guide tree: (i) phylogenetic analysis of a mitochondrial locus using ML and Bayesian methods (RAxML and BEAST) and (iii) species-tree estimation using independent nuclear loci (*BEAST). When the mutation rate was high, both approaches had a good chance of inferring the correct guide tree. When the mutation rate was low, the estimated guide trees might involve considerable errors, especially if only one nuclear locus was used. However, even in this case the false-positive rate in Bayesian species delimitation by BPP was not very high (the highest error rate being ~8% when the nominal value is 5%). This is because when the sequence data lack information, the posterior probabilities calculated by BPP tend to be low

TABLE 4. Median and quartiles (in parentheses) of posterior probabilities for splitting the specified clades by BPP in data simulated using tree 1, with three sequences per population

Nuclear Loci	Method	A_1A_2	B_1B_2	C_1C_2	D_1D_2	AB	ABC
Low-mutation rate							
1 locus	RAxML	0.315	0.308	0.269	0.223	0.990	1.000
		(0.171, 0.480)	(0.164, 0.484)	(0.120, 0.454)	(0.101, 0.407)	(0.915, 1.000)	(1.000, 1.000)
	*BEAST	0.310	0.322	0.262	0.228	0.996	1.000
		(0.153, 0.533)	(0.157, 0.556)	0.142, 0.467)	(0.109, 0.430)	(0.933, 1.000)	(1.000, 1.000)
5 loci	RAxML	0.135	0.135	0.080	0.061	1.000	1.000
		(0.077, 0.258)	(0.074, 0.258)	(0.046, 0.166)	(0.032, 0.132)	(1.000, 1.000)	(1.000, 1.000)
	*BEAST	0.130	0.137	0.082	0.063	1.000	1.000
		(0.077, 0.245)	(0.075, 0.270)	(0.046, 0.161)	(0.033, 0.131)	(1.000, 1.000)	(1.000, 1.000)
High-mutation rate							
1 locus	RAxML	0.042	0.046	0.024	0.018	1.000	1.000
		(0.021, 0.130)	(0.021, 0.150)	(0.011, 0.077)	(0.008, 0.060)	(0.995, 1.000)	(1.000, 1.000)
	*BEAST	0.039	0.039	0.025	0.019	1.000	1.000
		(0.019, 0.124)	(0.019, 0.112)	(0.011, 0.074)	(0.008, 0.065)	(0.993, 1.000)	(1.000, 1.000)
5 loci	RAxML	0.008	0.007	0.003	0.002	1.000	1.000
		(0.003, 0.017)	(0.003, 0.017)	(0.001, 0.009)	(0.000, 0.006)	(1.000, 1.000)	(1.000, 1.000)
	*BEAST	0.007	0.007	0.003	0.001	1.000	1.000
		(0.003, 0.016)	(0.003, 0.018)	(0.001, 0.009)	(0.000, 0.005)	(1.000, 1.000)	(1.000, 1.000)

Notes: The probability for splitting clade *ABCD* (the root) is 1 in every replicate data set.

TABLE 5. Median and quartiles (in parentheses) of posterior probabilities for splitting the specified clades by BPP in data simulated using tree 2, with three sequences per population

Nuclear Loci	Method	A_1A_2	B_1B_2	C_1C_2	D_1D_2	AB	CD
Low-mutation rate							
1 locus	RAxML	0.268	0.275	0.266	0.257	0.986	0.988
		(0.138, 0.427)	(0.144, 0.446)	(0.127, 0.418)	(0.140, 0.443)	(0.896, 0.999)	(0.916, 0.999)
	*BEAST	0.289	0.260	0.291	0.270	0.993	0.993
		(0.146, 0.504)	(0.136, 0.464)	(0.145, 0.503)	(0.136, 0.488)	(0.925, 1.000)	(0.916, 1.000)
5 loci	RAxML	0.106	0.107	0.110	0.109	1.000	1.000
		(0.062, 0.218)	(0.061, 0.215)	(0.062, 0.221)	(0.064, 0.207)	(1.000, 1.000)	(1.000, 1.000)
	*BEAST	0.113	0.110	0.112	0.107	1.000	1.000
		(0.062, 0.201)	(0.062, 0.219)	(0.063, 0.234)	(0.062, 0.204)	(1.000, 1.000)	(1.000, 1.000)
High-mutation rate							
1 locus	RAxML	0.034	0.040	0.036	0.037	1.000	1.000
		(0.016, 0.133)	(0.018, 0.113)	(0.017, 0.120)	(0.018, 0.118)	(0.992, 1.000)	(0.992, 1.000)
	*BEAST	0.036	0.032	0.040	0.038	1.000	1.000
		(0.018, 0.504)	(0.015, 0.464)	(0.018, 0.503)	(0.018, 0.488)	(0.990, 1.000)	(0.991, 1.000)
5 loci	RAxML	0.004	0.005	0.005	0.004	1.000	1.000
		(0.002, 0.012)	(0.002, 0.011)	(0.002, 0.012)	(0.002, 0.010)	(1.000, 1.000)	(1.000, 1.000)
	*BEAST	0.004	0.005	0.004	0.005	1.000	1.000
		(0.002, 0.010)	(0.002, 0.012)	(0.002, 0.011)	(0.002, 0.011)	(1.000, 1.000)	(1.000, 1.000)

Notes: The probability for splitting clade *ABCD* (the root) is 1 in every replicate data set.

and do not reach the 95% threshold. With more data, the posterior probabilities become more extreme, but in that case both guide-tree inference and species delimitation become highly accurate.

For multilocus nuclear data, one could conduct a phylogenetic analysis of the concatenated sequence alignment to generate a guide tree, using for example, RAxML. However, concatenation assumes that the same gene tree underlies all loci and fails to accommodate incomplete lineage sorting due to polymorphism in the ancestral species. We have not examined this alternative method since it is expected to be inferior to species-tree methods (such as *BEAST), which use the multispecies coalescent model to account for gene tree discordance across loci. For the mitochondrial locus, RAxML and

BEAST perform similarly, but RAxML runs several orders of magnitude faster than BEAST. Our discussion has thus focused on RAxML analysis of the mitochondrial locus but we note that Bayesian programs such as BEAST and MrBayes are usable as well. We stress that our objective in this study is not to compare different phylogenetic reconstruction methods (such as RAxML and BEAST) but is instead to evaluate the impact of errors in estimated guide trees on the false-positive and false-negative errors in the downstream species delimitation analysis by BPP. In this regard, our results suggest that the false-positive errors are rather minor when the guide tree is generated using sampled sequence data. Our results complement rather than contradict the previous finding by Leaché and Fujita (2010) that BPP tends to over-split and generate

excessive false positives if a random guide tree, which is most likely to be grossly wrong, is used. Users of BPP should take precautions against using grossly wrong guide trees for species delimitation analysis by BPP. If there are uncertainties concerning the phylogenetic relationships of the populations, the sensitivity of BPP analysis to the guide tree should be examined by using multiple guide trees derived using different strategies (as reviewed early). Furthermore, there is clearly a need to extend the algorithms in BPP to account properly for uncertainties in the guide tree.

The Impact of Gene Flow

In our simulation, we assumed no gene flow (migration, hybridization, or introgression) after species divergence, and conflicts between gene trees from different genomic regions or between mitochondrial and nuclear loci are entirely due to ancestral polymorphism and incomplete lineage sorting. A previous simulation study has examined the impact of gene flow on Bayesian species delimitation by BPP (Zhang et al. 2011). It was found that small amounts of migration (with ≪1 expected immigrant per generation) had little impact on the performance of the method, whereas a single species was inferred if migration between populations was prevalent (say, with ≫1 immigrants per generation). In that study, gene flow was assumed to affect all loci uniformly and the guide tree was assumed to be correct. The effect of migration may be more difficult to predict if migration affects different parts of the genome differently, due to natural selection. For example, the pattern of gene flow may vary considerably across genome regions because some loci are responsible for species adaptations to different ecological habitats and are thus under strong selection whereas other loci are neutral and can cross species boundaries quite freely. As a result, incipient species may show "islands" of divergence between their genomes amidst a sea of gene flow (Ellegren et al. 2012; Martin et al. 2013). Discordance between mitochondrial and nuclear phylogenies may also result from such selective gene flow, which makes the use of the mitochondrial locus to construct the guide tree problematic.

The Impact of Assignment Errors

In this study, we assumed that the population assignments were correct. In a recent simulation study, Olave et al. (2014) used STRUCTURAMA to assign individuals to populations and then used *BEAST to infer the guide tree, to evaluate the impact of errors in the upstream analysis (assignment and guide-tree construction) on the performance of BPP. They found that the error rates may be high when individuals are incorrectly assigned to populations, although BPP had excellent performance when assignment errors were absent. The results highlight the importance of reliable assignments to species delimitation by BPP. They also point to an interesting mismatch in the different steps of the delimitation process: although a few loci appeared to be sufficient for BPP to delimit species given the correct assignment, they were not enough for STRUCTURAMA to assign individuals to populations reliably. Nevertheless, a few issues with the design of the Olave et al. study make their results somewhat difficult to interpret. First, Olave et al. (2014; Fig. 2) used the number of inferred species to measure performance and failed to distinguish between the errors of over-splitting and under-splitting. Over-splitting appears to be a more serious error than under-splitting, as failure to delimit distinct species may simply be due to lack of power of the method or lack of information in the data. Second, Olave et al. (2014) used STRUCTURAMA to analyze the multilocus sequence data (treated as genotypes) to cluster the individuals into populations. The procedure mimics an unrealistic scenario in which multiple sympatric cryptic species exist in a sample with nothing to distinguish them *a priori*. Although the results suggest that a few loci of sequence data are insufficient for STRUCTURAMA to assign individuals to populations reliably, the impact of assignment errors on species delimitation by BPP under more realistic scenarios remains unknown. As discussed by Olave et al., traditional taxonomic boundaries, and morphological and geographical data may be available to determine the number of putative species and assign individuals to populations. Moreover, SNPs and microsatellites across multiple loci may be better suited than sequences for assigning individuals to populations.

SUPPLEMENTARY MATERIAL

Supplementary Figures S1–S32: Distributions or histograms (out of 1000 replicate data sets) of posterior probabilities for splitting the clades on the correct guide tree, calculated by BPP. The 32 figures correspond to 32 simulation parameter settings, with two sample sizes (three or five sequences per population), two species trees (trees 1 or 2 in Fig. 1a), two mutation rates (Table 1), two guide-tree inference methods (RAxML: ML tree for the mitochondrial locus and *BEAST: species-tree inference from the nuclear loci), and two numbers of loci (one or five nuclear loci), with the last factor changing first. Thus, Supplementary Figures S1–S16 are for the small sample size of three sequences per population and Supplementary Figures S17–S32 are for the large sample size of five sequences per population. Supplementary Information related to this article has been deposited at Dryad under http://dx.doi.org/10.5061/dryad.m1r32.

REFERENCES

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

Camargo A., Morando M., Avila L.J., Sites J.W. 2012. Species delimitation with ABC and other coalescent-based methods: a test of accuracy with simulations and an empirical example with lizards of the Liolaemus Darwinii complex (Squamata: Liolaemidae). Evolution 66:2834–2849.

Carstens B.C., Pelletier T.A., Reid N.M., Satler J.D. 2013. How to fail at species delimitation? Mol. Ecol. 22:4369–4383.

Corander J., Waldmann P., Marttinen P., Sillanpaa M.J. 2004. BAPS 2: enhanced possibilities for the analysis of genetic population structure. Bioinformatics 20:2363–2369.

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

Drummond A.J., Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7:214.

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

Ellegren H., Smeds L., Burri R., Olason P.I., Backstrom N., Kawakami T., Kunstner A., Makinen H., Nadachowska-Brzyska K., Qvarnstrom A., Uebbing S., Wolf J.B.W. 2012. The genomic landscape of species divergence in Ficedula flycatchers. Nature 491: 756–760.

Falush D., Stephens M., Pritchard J.K. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567–1587.

Felsenstein J. 2005. Phylip: phylogenetic inference program, Version 3.6. Seattle: University of Washington.

Fujita M.K., Leaché A.D. 2011. A coalescent perspective on delimiting and naming species: a reply to Bauer et al. Proc. R. Soc. Lond. B. Biol. Sci. 278:493–495.

Fujita M.K., Leaché A.D., Burbrink F.T., McGuire J.A., Moritz C. 2012. Coalescent-based species delimitation in an integrative taxonomy. Trends Ecol. Evol. 27:480–488.

Hamback P.A., Weingartner E., Ericson L., Fors L., Cassel-Lundhagen A., Stenberg J.A., Bergsten J. 2013. Bayesian species delimitation reveals generalist and specialist parasitic wasps on Galerucella beetles (Chrysomelidae): sorting by herbivore or plant host. BMC Evol. Biol. 13:92.

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

Huelsenbeck J. P., Andolfatto P. 2007. Inference of population structure under a Dirichlet process model. Genetics 175:1787–1802.

Huelsenbeck J.P., Rannala B. 2004. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. Syst. Biol. 53:904–913.

Jukes T.H., Cantor C.R. 1969. Evolution of protein molecules. In: Munro H.N., editor. Mammalian protein metabolism. New York: Academic Press. p. 21–123.

Knowles L.L., Carstens B.C. 2007. Delimiting species without monophyletic gene trees. Syst. Biol. 56:887–895.

Leaché A.D., Fujita M.K. 2010. Bayesian species delimitation in West African forest geckos (Hemidactylus fasciatus). Proc. R. Soc. Lond. B. Biol. Sci. 277:3071–3077.

Linde C.C., Phillips R.D., Crisp M.D., Peakall R. 2014. Congruent species delineation of Tulasnella using multiple loci and methods. New Phytol. 201:6–12.

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

Martin S.H., Dasmahapatra K.K., Nadeau N.J., Salazar C., Walters J.R., Simpson F., Blaxter M., Manica A., Mallet J., Jiggins C.D. 2013. Genome-wide evidence for speciation with gene flow in Heliconius butterflies. Genome Res. 23:1817–1828.

Olave M., Sola E., Knowles L.L. 2014. Upstream analyses create problems with DNA-based species delimitation. Syst. Biol. 63:263–271.

O'Meara B.C. 2010. New heuristic methods for joint species delimitation and species tree inference. Syst. Biol. 59:59–73.

Pritchard J.K., Stephens M., Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945–959.

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

Rannala B., Yang Z. 2013. Improved reversible jump algorithms for Bayesian species delimitation. Genetics 194:245–253.

Ronquist F., Teslenko M., van der Mark P., Ayres D.L., Darling A., Hohna S., Larget B., Liu L., Suchard M.A., Huelsenbeck J.P. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61:539–542.

Satler J.D., Carstens B.C., Hedin M. 2013. Multilocus species delimitation in a complex of morphologically conserved trapdoor spiders (mygalomorphae, antrodiaetidae, aliatypus). Syst. Biol. 62:805–823.

Satta Y., Hickerson M., Watanabe H., O'hUigin C., Klein J. 2004. Ancestral population sizes and species divergence times in the primate lineage on the basis of intron and BAC end sequences. J. Mol. Evol. 59:478–487.

Sites J.W., Marshall J.C. 2004. Delimiting species: a renaissance issue in systematic biology. Trends Ecol. Evol. 18:462–470.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690.

Wiens J.J. 2007. Species delimitation: new approaches for discovering diversity. Syst. Biol. 56:875–878.

Yang Z. 1998. On the best evolutionary rate for phylogenetic analysis. Syst. Biol. 47:125–133.

Yang Z. 2006. Computational molecular evolution. Oxford (UK): Oxford University Press.

Yang Z., Rannala B. 2010. Bayesian species delimitation using multilocus sequence data. Proc. Natl Acad. Sci. U. S. A. 107:9264–9269.

Zhang C., Zhang D.-X., Zhu T., Yang Z. 2011. Evaluation of a Bayesian coalescent method of species delimitation. Syst. Biol. 60:747–761.

Zhou W.W., Wen Y., Fu J., Xu Y.B., Jin J.Q., Ding L., Min M.S., Che J., Zhang Y.P. 2012. Speciation in the Rana chensinensis species complex and its relationship to the uplift of the Qinghai–Tibetan Plateau. Mol. Ecol. 21:960–973.