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

CHI ZHANG1, BRUCE RANNALA2, AND ZIHENG YANG3,4,*

1Department of Bioinformatics and Genetics, Swedish Museum of Natural History, Box 50007, SE-104 05 Stockholm, Sweden; 2Genome Center and Section of Evolution and Ecology, University of California at Davis, One Shields Avenue, Davis, CA 95616, USA; 3Department of Genetics, Evolution and Environment, University College London, Darwin Building, Gower Street, London WC1E 6BT, UK; and 4Center 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.

Received 4 December 2013; reviews returned 14 May 2014; accepted 12 July 2014

Associate Editor: Thomas Buckley

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 (Hubisz et al. 2009), and BAPS (Corander et al. 2004), which can assign individuals to populations and even infer a population tree. These 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 effective population size is much smaller. Furthermore, the effective population size for a mitochondrial locus is only one-quarter as large. Thus, \( \theta = \frac{2N_{mt}}{2N_{nuc}} \).

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), Sattler 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 (\( \tau \)) as well as seven population size parameters (\( \theta \)) for the extant and extinct species on the species tree. Both \( r \) and \( \theta \) 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 \( A_1, A_2, B_1, B_2, E_1, E_2, F_1, F_2 \)), 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 McCoMIL 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

<table>
<thead>
<tr>
<th>Species tree</th>
<th>( \tau )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tree 1</td>
<td>(0.001, 0.002, 0.003)</td>
</tr>
<tr>
<td>Tree 2</td>
<td>(0.001, 0.001, 0.003)</td>
</tr>
</tbody>
</table>

**High-mutation rate**

<table>
<thead>
<tr>
<th>Species tree</th>
<th>( \theta )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tree 1</td>
<td>0.02</td>
</tr>
<tr>
<td>Tree 2</td>
<td>0.02</td>
</tr>
</tbody>
</table>

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, \( \theta_{mt} = \frac{20}{2}\theta_{nuc} \) and \( \theta_{mt} = \frac{5}{2}\theta_{nuc} \).
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 ($\tau$) and the population size parameters ($\theta$) 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 population, 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 $\tau_{\text{mt}} = 20\tau_{\text{nuc}}$ and $\theta_{\text{mt}} = 5\theta_{\text{nuc}}$ (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* (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 combination 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 \( \theta \) 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. In each analysis, BEAST was used as an alternative to RAxML in the *BPP* analysis. 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
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
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 (A1A2, B1B2, C1C2, and D1D2) 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 (A1 and A2, B1 and B2, C1 and C2, or D1 and D2) were 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 A1A2 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 A1A2 and B1B2 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 A1A2 and B1B2 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
two by BPP with posterior more than or equal to 95% in data simulated
increased information content in the BPP analysis of the 
estimation errors when more nuclear loci are used and
See note to Table 2.

<table>
<thead>
<tr>
<th>Nuclear Loci</th>
<th>Method</th>
<th>A1,A2</th>
<th>B1,B2</th>
<th>C1,C2</th>
<th>D1,D2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low-mutation rate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 locus</td>
<td>RAxML</td>
<td>0.012</td>
<td>0.012</td>
<td>0.015</td>
<td>0.013</td>
</tr>
<tr>
<td></td>
<td>*BEAST</td>
<td>0.063</td>
<td>0.063</td>
<td>0.047</td>
<td>0.037</td>
</tr>
<tr>
<td>5 loci</td>
<td>RAxML</td>
<td>0.009</td>
<td>0.004</td>
<td>0.004</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td>*BEAST</td>
<td>0.009</td>
<td>0.006</td>
<td>0.007</td>
<td>0.007</td>
</tr>
<tr>
<td>High-mutation rate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 locus</td>
<td>RAxML</td>
<td>0.034</td>
<td>0.042</td>
<td>0.035</td>
<td>0.026</td>
</tr>
<tr>
<td></td>
<td>*BEAST</td>
<td>0.044</td>
<td>0.035</td>
<td>0.031</td>
<td>0.027</td>
</tr>
<tr>
<td>5 loci</td>
<td>RAxML</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>*BEAST</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

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 locus (*BEAST).

<table>
<thead>
<tr>
<th>Nuclear Loci</th>
<th>Method</th>
<th>A1,A2</th>
<th>B1,B2</th>
<th>C1,C2</th>
<th>D1,D2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low-mutation rate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 locus</td>
<td>RAxML</td>
<td>0.007</td>
<td>0.009</td>
<td>0.003</td>
<td>0.011</td>
</tr>
<tr>
<td></td>
<td>*BEAST</td>
<td>0.053</td>
<td>0.057</td>
<td>0.058</td>
<td>0.064</td>
</tr>
<tr>
<td>5 loci</td>
<td>RAxML</td>
<td>0.004</td>
<td>0.006</td>
<td>0.006</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td>*BEAST</td>
<td>0.006</td>
<td>0.005</td>
<td>0.009</td>
<td>0.006</td>
</tr>
<tr>
<td>High-mutation rate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 locus</td>
<td>RAxML</td>
<td>0.030</td>
<td>0.036</td>
<td>0.030</td>
<td>0.030</td>
</tr>
<tr>
<td></td>
<td>*BEAST</td>
<td>0.043</td>
<td>0.037</td>
<td>0.040</td>
<td>0.036</td>
</tr>
<tr>
<td>5 loci</td>
<td>RAxML</td>
<td>0.001</td>
<td>0.001</td>
<td>0.002</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td>*BEAST</td>
<td>0.000</td>
<td>0.000</td>
<td>0.001</td>
<td>0.000</td>
</tr>
</tbody>
</table>

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 A1,A2 and B1,B2 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 A1,A2 and B1,B2 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 A1,A2 and B1,B2 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
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 A1,A2, B1,B2, C1,C2, and D1,D2
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 A1,A2, B1,B2, C1,C2,
and D1,D2 (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.
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
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 (ii) 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 A1A2 B1B2 C1C2 D1D2 AB ABC

Low-mutation rate
1 locus RAxML 0.315 0.308 0.269 0.223 0.990 1.000
  *BEAST 0.310 0.322 0.262 0.228 0.996 1.000
5 loci RAxML 0.135 0.135 0.080 0.061 1.000 1.000
  *BEAST 0.130 0.137 0.082 0.063 1.000 1.000

High-mutation rate
1 locus RAxML 0.042 0.046 0.024 0.018 1.000 1.000
  *BEAST 0.039 0.039 0.025 0.019 1.000 1.000
5 loci RAxML 0.008 0.007 0.003 0.002 1.000 1.000
  *BEAST 0.007 0.007 0.003 0.001 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 A1A2 B1B2 C1C2 D1D2 AB CD

Low-mutation rate
1 locus RAxML 0.268 0.275 0.266 0.257 0.986 0.988
  *BEAST 0.289 0.261 0.291 0.270 0.993 0.993
5 loci RAxML 0.106 0.107 0.110 0.109 1.000 1.000
  *BEAST 0.113 0.110 0.107 0.100 1.000 1.000

High-mutation rate
1 locus RAxML 0.034 0.040 0.036 0.037 1.000 1.000
  *BEAST 0.036 0.032 0.036 0.039 1.000 1.000
5 loci RAxML 0.004 0.005 0.004 0.004 1.000 1.000
  *BEAST 0.004 0.005 0.004 0.004 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 \( \leq 1 \) 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–S52: 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–S52 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.

FUNDING

This study is supported by a grant from the Biotechnological and Biological Sciences Research Council (BBSRC) and a Royal Society-Wolfson Merit Award, both to Z.Y. C.Z. wishes to acknowledge UCL CoMPLEX for a Visiting Researcher Award, which allowed him to visit London.
REFERENCES


