10 February 2017

# Bayesian Species Identification under the Multispecies Coalescent Provides Significant Improvements to DNA Barcoding Analyses

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

[1] Department of Genetics, Evolution and Environment, University College London, Gower Street, London WC1E 6BT, UK

[2] Department of Evolution and Ecology, University of California at Davis, One Shields Avenue, Davis, CA 95616, USA

[3] Eco-Beijing, Beijing Normal University, Beijing 00000, China

**Running head**: Next-generation DNA barcoding

**Key words:** Coalescent, DNA barcoding, species delimitation, species identification, BPP.

* Correspondence to
Bruce Rannala <brannala@ucdavis.edu>

# Abstract

DNA barcoding methods use a single locus (usually the mitochondrial COI gene) to assign unidentified specimens to known species in a library based on a genetic distance threshold that distinguishes between-species divergence from within-species diversity. Recently developed species delimitation methods based on the multispecies coalescent (MSC) model offer an alternative approach to individual assignment using either single-locus or multi-loci sequence data. Here we use simulations to demonstrate three features of an MSC method implemented in the program BPP. First, we show that with one locus, MSC can accurately assign individuals to species without the need for arbitrarily determined distance thresholds (as required for barcoding methods). We provide an example in which no single threshold or barcoding gap exists that can be used to assign all specimens without incurring high error rates. Second, we show that BPP can identify cryptic species that may be mis-identified as a single species within the library, potentially improving the accuracy of barcoding libraries. Third, we show that taxon rarity does not present any particular problems for species assignments using BPP, and that accurate assignments can be achieved even when only one or a few loci are available. Thus, concerns that have been raised that MSC methods may have problems analyzing rare taxa (singletons) are unfounded. Currently barcoding methods enjoy a huge computational advantage over MSC methods and may be the only approach feasible for massively large datasets, but MSC methods may offer a more stringent test for species that are tentatively assigned by barcoding.

# Introduction

DNA barcoding has been proposed as a fast and inexpensive approach to species identification. A reference library of sequences for a "universal locus" is constructed using species that are identified *a priori*, and unidentified specimens are then identified by calculating the genetic distance between their query sequence and the sequences in the library (Hebert et al., 2003). The universal locus is usually mitochondrial cytochrome oxidase 1 (CO1) or cytochrome b (cytb) because mtDNA is easier to type than nuclear DNA from highly processed and degraded tissues. One particularly successful application of DNA barcoding is in forensics, where DNA evidence is used to track illegal trade of wildlife (Alacs et al., 2010) or confirm the identity of fish products (Smith et al., 2008).

DNA barcoding has also been used for species discovery or species delimitation (e.g., Rossini et al., 2016). This typically relies on determining a genetic distance threshold or 'barcoding gap'. The query specimen is identified and assigned to an existing species in the library if the shortest pairwise sequence distance from the query to the sequence library is smaller than the pre-specified threshold. If the smallest distance exceeds the threshold, there will be a non-identification, which indicates that the specimen may be a new species not yet represented in the library. The choice of the threshold is a thorny issue and is somewhat arbitrary. For example, the '10× rule' (Hebert et al., 2004) specifies the interspecific divergence to be at least 10 times as large as the intraspecific diversity. Dowton (Dowton et al., 2014) used 4% of CO1 divergence, while Rossini et al. (Rossini et al., 2016) used a Kimura 2-parameter distance of 2%. Other methods use the 95% confidence interval of conspecific distances to determine the threshold, leading to higher thresholds when there is more intraspecific variation (Meier et al., 2006). More sophisticated methods generate the distance threshold by taking a database with a known taxonomy and minimizing the false-positive errors (incorrectly identifying a specimen as a new species) and false-negative errors (incorrectly lumping a specimen into another species) (Meyer and Paulay, 2005). Similar approaches have been used to "optimize" the distance threshold in empirical databases, using programs such as Spider (Brown et al., 2012) and ABGD (Automatic Barcode Gap Discovery, Puillandre et al., 2012). Note that all barcoding methods require a distance threshold, regardless of the method used to determine it.

However, different species have different population sizes and divergence times. As a result, one may expect considerable overlap between intraspecific variation and interspecific divergence among closely-related species (Meyer and Paulay, 2005), so that there may not be a "one-size-fits-all" threshold. In a case study examining DNA barcoding performance in a diverse group of marine gastropods, Meyer and Paulay (Meyer and Paulay, 2005) found that use of one threshold to delineate all species was particularly problematic for closely related species in taxonomically understudied groups.

Another method for species discovery/delimitation is the Generalized Mixed Yule Coalescent (GMYC) method (Pons et al., 2006; Fujisawa and Barraclough, 2013). This uses the reconstructed gene tree for a single locus and fits a mixed model to the estimated divergence times, with the Yule branching process describing species divergences and the coalescent process describe the within-species process of lineage joining. The method is heuristic as it is not based on a fully specified population genetic model and does not accommodate ancestral polymorphism correctly. It also assumes that the gene tree with node ages is known without error and thus does not accommodate phylogenetic errors of gene tree reconstruction.

Species identification and delimitation using genetic sequence data should best be viewed as a statistical inference problem, given the stochastic nature of the coalescent and the process of sequence evolution. The natural framework that describes the process of species divergence and isolation, and ancestral polymorphism and incomplete lineage sorting is the multispecies coalescent (MSC) model, which is a straightforward extension of the standard single-species coalescent (Kingman, 1982; Hudson, 1983; Tajima, 1983) to the case of multiple species (Takahata et al., 1995; Yang, 2002; Rannala and Yang, 2003). The gene genealogies or gene trees have probability distributions specified by parameters in the model, including the species divergence times and the population sizes for both the ancestral and extant species. In theory the MSC framework should allow species delimitation even in extreme cases where the within-species diversity for some species is higher than the between-species divergence between some other species. Furthermore, a full likelihood implementation of the MSC model should be statistically more efficient than heuristic methods for the same inference (Xu and Yang, 2016). Here we demonstrate that this theoretical advantage is realized by the BPP program, which is a Bayesian MCMC implementation of the MSC model and which allows both species delimitation and species tree inference (Rannala and Yang, 2003; Yang and Rannala, 2010; Yang and Rannala, 2014; Yang, 2015).

Another potential benefit of applying the MSC-based approach is its ability to delimit cryptic species (specimens in the database that are distinct species but incorrectly recognised as one species). Although the potential clearly exists, the performance of BPP to delimit cryptic species in practical data analysis has not been carefully examined. Indeed, Collins and Cruickshank (Collins and Cruickshank, 2014) suggested that "it is questionable whether such statistics would be reliable due to the sampling and parameter estimation problems associated with taxon rarity in species delimitation methods." We demonstrate in simulations that BPP can delimit cryptic species with high accuracy even if the species are under-sampled.

The impact of species under-sampling, or the rarity of species, on species delimitation has been extensively discussed (Lim et al., 2012; Collins and Cruickshank, 2014). Rarity indeed appears to be very common. For example, 48.5% of species in the African beetles library examined by Ahrens et al. (Ahrens et al., 2016) were singletons. Many authors have considered species rarity to be a major challenge for species delimitation. For example, Lim et al. (Lim et al., 2012) claim that many newly developed methods either implicitly or explicitly require that all species are well sampled, and argue that delimitation techniques should be modified to accommodate the commonness of rarity. Their conclusions are based on the intuition that species delimitation requires information about the within-species diversity relative to the between-species divergence, and such information will be hard to obtain if some species are under-sampled or are singletons. Furthermore, heuristic methods of species delimitation have indeed been found to suffer from poor species sampling. For example, in a case study of southern African beetles using CO1 sequences from >500 specimens and ~100 species, Ahrens et al. (Ahrens et al., 2016) demonstrated that under-sampling could compromise species delimitation by GMYC, and the difficulty appeared to lie more with the high sensitivity of GMYC to variable population sizes than with high proportions of singletons per se; GMYC appears to have difficulty generating reliable estimates of intra- vs. interspecies evolutionary parameters when some species are under-sampled. Fujisawa and Barraclough (Fujisawa and Barraclough, 2013) also found that the most important factor affecting the accuracy of species delimitation by GMYC is the mean population size relative to divergence times between speciation events. In this paper, we focus on the differences between DNA barcoding and a full likelihood implementation of the MSC model (BPP) and do not include GMYC in our evaluation.

We note that species rarity is naturally accommodated by the MSC model that underlies the BPP program. It is simply a matter of information content and power, and BPP can make reliable inferences using multi-loci data even if a species is represented by a single specimen (singleton). New species have often been described based on rare specimens using morphology and it therefore appears self-evident that genetic sequences should contain enough information to infer the species status of a rare specimen.

In this paper we use simulations and analyses with the species delimitation program BPP to demonstrate that (1) MSC methods can accurately assign individuals to species without the need for arbitrarily determined distance thresholds (as are required for barcoding methods). We provide an example where no single barcoding gap (threshold) exists that can be used to assign all the species in a group without incurring high error rates, yet BPP can accurately assign individuals. (2) BPP can identify cryptic species that are misidentified as a single species within a species library that is being used for assignments, potentially improving the accuracy of barcoding libraries. (3) Taxon rarity does not present problems for species assignments using BPP, and accurate assignments can often be made even with only one or a few loci.

## One barcode gap for identifying all species may not exist

We simulate sequence data using the species tree of figure 1a, with 1 + 10 sequences from *A*, 10 *B*s, 1 + 10 *C*s, 10 *D*s, 10 *E*s, and 1 *F*. The 10 sequences each from species *A*-*E* (50 sequences in total) are used as the 'library', while the 1 *A*, 1 *C* and 1 *F* are used as three 'query' sequences (denoted $A_1$, $C_1$, and *F*). We generate one locus, of 1000bp, by simulating the gene trees under the MSC (Rannala and Yang, 2003) and evolving sequences along the gene tree. The program MCCOAL, which is part of the BPP package (Yang, 2015), was used for the simulation. The species divergence times ($\tau$s) and population size parameters ($\theta$s) used are shown in Fig. 1. Here both $\tau$s and $\theta$s are measured by the mutational distance, so that $\theta = 0.01$ means that two sequences from the population have on average 1 difference per 100 sites, while $\tau = 0.01$ means that the ancestral node in the species tree and the present time are separated by a genetic distance of 1%. Gene tree topologies and branch lengths (coalescent times) are generated from the MSC density (Rannala and Yang, 2003), and are then used to 'evolve' sequences along the gene tree to generate the sequence alignment for the tips of the gene tree. The JC model (Jukes and Cantor, 1969) was used both to simulate and to analyze the data. Each simulated dataset, which consists of 53 sequences for one single locus, was analyzed using either DNA barcoding or the program BPP, with sequences $A_1$, $C_1$ and *F* treated as the query, against the sequence library made up of the remaining 50 sequences. The number of replicate simulated datasets was 100.

**Barcoding analysis.** Rather than using a specific DNA Barcoding program to choose a threshold, we consider all possible sequence distance thresholds. Let $d(A_1, A) = \min\{d(A_1, A_i), i = 2, \ldots, 10\}$ be the smallest distance from the query $A_1$ to species *A*, and define $d(C_1, C)$ and $d(F, E)$ accordingly. $A_1$ is correctly assigned to species *A* if $d(A_1, A)$ is smaller than the distance threshold, while *F* is correctly assigned to be a species distinct from *E* if $d(F, E)$ is larger than the distance threshold. Thus to assign $A_1$ and $C_1$ correctly to species *A* and *C*, respectively, one would prefer a large distance threshold, and to assign *F* correctly into a distinct species from species *E*, one would like a small distance threshold. It will be impossible to assign all three queries correctly if $d(F, E) \le d(A_1, A)$ or if $d(F, E) \le d(C_1, C)$. This happened in 28 out of the 100 replicate datasets. For example in one dataset, $d(A_1, A) = d(A_1, A_2) = 0.001$, $d(C_1, C) = d(C_1, C_5) = 0.004$, and $d(F, E) = d(F, E_2) = 0.003$. With the within-species distance (0.004 for *C*) being greater than the between-species distance (0.003 between *F* and *E*), it is impossible to use one distance threshold to make correct assignments, or to avoid both the false positive error of claiming $A_1$ or $C_1$ as a new species (a non-identification) and the false negative error of lumping *F* into species *E*. In the other 72 datasets, it is theoretically possible to choose a threshold that will allow correct assignment of all three queries, but mis-assignments may still occur if an imperfect threshold is chosen.

Fig. 2 shows the proportion (among 100 replicate datasets) of correct species assignments when each dataset is analyzed using a fixed distance threshold. For example, at the distance threshold of 0.003 (three differences per kb), $A_1$ is correctly assigned to species *A* in 96% of datasets, $C_1$ is correctly assigned to species *C* in 70% of datasets, and *F* is correctly assigned to a species distinct from species *E* in only 39% of datasets.

**BPP analysis.** The same single locus datasets were analyzed using BPP (Yang and Rannala, 2014; Yang, 2015). The BPP program uses reversible-jump Markov chain Monte Carlo (rjMCMC) (Yang and Rannala, 2010) to move between different delimitation models, which correspond to different groupings of populations into the same species, and MCMC to move between different species phylogenies given the same species delimitation (Yang and Rannala, 2014; Rannala and Yang, 2017). The program achieves Bayesian model comparison through MCMC, visiting the competing models with frequencies corresponding to their posterior probabilities. The BPP analysis assumes that individual specimens are assigned to populations. Multiple populations may be merged into one species by the rjMCMC algorithm while one population is never split into two species. Our analysis assumed 8 populations: *A*, *B*, *C*, *D*, and *E*, and the three query sequences. Each of the three query sequences ($A_1$, $C_1$ and *F*) is assigned to its own population so that it can either be merged into one of the existing species in the library (*A*, *B*, *C*, *D*, and *E*) or designated a new species. We use Prior 3 for the species-tree models, which assigns uniform probability (1/8 each) for 1, 2, …, and 8 species (Yang, 2015). Gamma priors are assigned on parameters: $\theta \sim G(2, 200)$, with the mean to be $\alpha/\beta = 2/200 = 0.01$ (one mutation per 100 bp) for the mutation-scaled population size parameters for both modern and ancestral populations, and $\tau_{ABC} \sim G(3, 100)$, with mean 0.03, for the root of the species tree. The values 2 and 3 for the gamma shape parameter ($\alpha$) are relatively small, indicating that the priors are fairly diffuse. The prior means are set to be equal to the true values.

The posterior probabilities for different models calculated using BPP are summarized in Fig. 3. Here P($A_1A$) is the posterior probability that populations $A_1$ and *A* are (correctly) grouped into one species, to the exclusion of all other populations. The average posterior probability was 0.81 for correctly assigning $A_1$ to species *A*, and was 0.17 for recognizing it as a new species (over-splitting). The average posterior probability was 0.71 for correctly assigning $C_1$ to species $C$, with the false positive rate of over-splitting to be 0.29. The average posterior probability for correctly identifying $F$ as a new species was 0.68, with the false negative rate of incorrectly lumping it with species $E$ to be 0.31. High posterior probabilities (>0.9) for an incorrect assignment are very rare (less than 1% on average). Although the information is weak with a single locus, BPP is outperforming the threshold method on average across species: even if an optimal threshold of about 0.0023 were used the average proportion of correct assignment for the threshold method is approximately 0.69, compared with 0.73 for BPP.

Increasing the number of loci from 1 to 10 led to increased posterior probabilities for correct assignments and to reduced error rates. The posterior probabilities for correctly assigning $A_1$ to species $A$, $C_1$ to species $C$, and $F$ to a distinct species from $E$ are shifted towards 1 (Fig. 4a-c), while the posterior probabilities for incorrectly assigning $A_1$ or $C_1$ to distinct species are shifted towards 0 (Fig. 4d-f). Query $F$ is identified as a distinct species with posterior probability 1.0, and the error rate for lumping $E$ and $F$ is 0 in every dataset.

## Identifying cryptic species

To examine the performance of BPP in identifying cryptic species we simulated sequence data using the species tree for three species of figure 1b, with four sequences (two diploid individuals) from each species. The parameter values were $\theta = 0.01$ for all populations, $\tau_{AB} = 0.01$ and $\tau_{ABC} = 0.02$. A and B represent distinct cryptic species, misidentified as one species in the library. Each locus was 1000bp. The number of loci was either 2 or 10, with 12 sequences per locus. The number of replicate simulations was 100. The BPP analysis assumed 5 populations ($A_1$, $A_2$, $B_1$, $B_2$, $C$), with each individual from A and B treated as a separate population. We assign Prior 3 for the species-tree models, which assigns uniform probabilities (1/5 each) for 1, 2, …, 5 species (Yang, 2015). The priors on parameters are $\theta \sim$ G(1, 100) and $\tau_{\mathrm{ABC}} \sim$ G(4, 200). These are diffuse priors with the means equal to the true values.

The true model in this case has 3 species, with the phylogeny (($A$, $B$), $C$), and with $A_1$ and $A_2$ grouped into one species ($A$), and $B_1$ and $B_2$ into another ($B$). With 2 loci, 82% of the simulated datasets produced a maximum *a posteriori* (MAP) model with 3 species that matched the true model. The histogram for the posterior probability for the correct model (which separates $A$ and $B$ into distinct species and also infers the correct species phylogeny) is shown in Fig. 5a. While this probability is >70% in most datasets, it is low in many other datasets, reflecting the low information content in data of one single locus. With 10 loci, 97% of the simulated datasets produced a MAP model with 3 species that matched the true model. The histogram for the posterior probability for the correct model is shown in Fig. 5b. The shift towards 1 relative to that of Fig. 5a reflects the dramatic increase in the information content in data of 10 loci.

If a posterior probability of 0.90 is used as a cut-off to choose a model, the error rate is 1% for 2 loci and 0% for 10 loci. With a 0.90 posterior probability cut-off the power to identify the true model is 17% with 2 loci and 69% with 10 loci. The average posterior probability Pr($A_1A_2$) for grouping $A_1$ and $A_2$ into one species was 0.83 with 2 loci and 0.94 with 10 loci, and the average posterior probability Pr($B_1B_2$) for grouping $B_1$ and $B_2$ into one species was 0.85 with 2 loci and 0.94 with 10 loci. The distributions of posterior probabilities of the $A_1A_2$ and $B_1B_2$ groupings with either 2 or 10 loci are shown in Fig. 6. Additional loci may be needed to infer the true model with almost complete certainty.

## Identifying rare species

To examine the performance of BPP in identifying rare species we simulated data on the species tree of figure 1b, with 1 *A*, 10 *B*s, and 10 *C*s. The one *A* sequence represents one specimen (a singleton) from a haploid species. We are interested in whether BPP can correctly infer *A* to be a distinct species when sequence data from multiple loci are available. Note that having two *A* sequences (as would be available if the species is diploid) will make the task easier. The parameter values are $\theta = 0.01$ for all populations, $\tau_{\text{AB}} = 0.01$ and $\tau_{\text{ABC}} = 0.02$. Each locus was 1000bp. The number of loci was either 2 or 10. The number of replicates was 100. The BPP analysis assumes 3 populations (*A*, *B*, *C*). We assign Prior 3 for the species-tree models, which assigns uniform probability (1/3) for 1, 2, and 3 species (Yang, 2015). The priors on parameters are $\theta \sim$ G(1, 100) and $\tau_{ABC} \sim$ G(4, 200).

The true model in this case is 3 species, with the species tree ((*A*, *B*), *C*). The MAP model was the true model for all simulated datasets with either 2 loci or 10 loci. In all datasets, three species were delimited with posterior probability greater than 0.95, whether 2 or 10 loci are analyzed. The average posterior probability of 3 species was 0.998 with 2 loci and 1.000 with 10 loci. The power of inference is very high in this case compared with the simulation of cryptic species, because multiple individuals (10 sequences) are available from species *B*.

## Discussion

Species assignment by BPP under the MSC efficiently uses information available in the sequence data about between-species divergence versus within-species polymorphism. However, BPP also uses the combined information available from multi-loci gene trees and branch lengths, even though the gene trees at each individual locus may involve considerable uncertainties and sampling errors. The method accommodates the fact that some species have large population sizes (showing greater within-species diversity), and some species diverged recently (so that between-species divergence may not necessarily exceed within-species diversity). By using a formal modeling framework and incorporating information about contemporary and ancestral population sizes available from multi-loci sequence data one avoids the need to specify subjective distance thresholds (as in DNA barcoding). Another advantage of statistical modeling methods such as BPP over heuristic methods (such as DNA barcoding) is that they provide measures of uncertainties in the form of posterior probabilities. By contrast, single-locus high-throughput approaches to discovering species will not work well when population sizes for some species are large and/or divergence events are recent. Our results are consistent with the previous simulation study of Hickerson et al. (Hickerson et al., 2006), who showed that single-gene thresholds for species discovery such as the 10× rule can result in substantial error with recent species divergence times.

Lim et al. (Lim et al., 2012; see also Collins and Cruickshank, 2014) speculate that "in studies using coalescence much of the evidence for species limits comes from coalescence points, which are by definition lacking for rare species…" This intuition is faulty, as can be seen from our simulation results showing that BPP identified the singleton species with higher power, even though the single sequence from the singleton species cannot provide any coalescent points within that species. Indeed, the simulation of Zhang et al. (Zhang et al., 2011, Fig. 3) showed that BPP can assign species correctly with 10 or 50 loci (depending on the mutation rate or sequence divergence level) even if a single sequence is sampled from every population at every locus so that estimation of intra-species diversity is not possible for any species. Lim et al. (Lim et al., 2012) went on to recommend several approaches for identifying statistical 'outliers' for use in species recognition. Those suggestions are not valid or relevant.

Dowton et al. (Dowton et al., 2014) suggested that coalescent-based species delimitation methods can be used to make more accurate specimen identifications than single-locus DNA barcoding. Our simulation results support suggestion. Collins and Cruickshank (Collins and

Cruickshank, 2014) suggest that "to benchmark the efficiency and accuracy of species delimitation methodologies, it should now be a priority to highlight exemplar data sets—empirical and/or simulated—for which MSC methods clearly outperform simpler mtDNA analyses." In this study we have generated several such exemplar cases by simulation and show that a simple distance threshold does not work well.

Bayesian MCMC methods such as BPP involve far more intensive computation than heuristic methods such as DNA Barcoding and the computational requirement increases quickly with an increase in the amount of data (e.g., the number of species/populations, loci, sequences per locus, and sites per sequence — in order of decreasing importance). The current version of the BPP program has been used to analyze data of a few thousand loci, with ~20 sequences per locus and about 10 species/populations. With very few loci, the program can deal with 100-200 sequences per locus. While algorithmic improvements are being made (Rannala and Yang, 2017), the program is not up to the task for very large datasets. We note that BPP recovered the true model (species delimitation and species phylogeny) with near certainty with a moderate number of loci (e.g., 10 or 20). Thus it is not always necessary to use the whole genome to infer species status. Similarly there may not be a need to analyze 500 species, say, in one combined analysis, to delimit species. Reliable results may be obtainable if divergent groups of species are analysed as separate datasets, further reducing the computational burden.

At the same time, simple heuristic methods such as DNA barcoding can be expected to work well if the populations are small and species divergences are ancient so that incomplete lineage sorting is rare. For challenging problems involving large population sizes and recent species divergences, DNA barcoding may be misleading. It is then prudent to try both types of analyses whenever possible.

Meta-genomics is one area in which DNA barcoding is currently the only practical approach. In a meta-genomics analysis, large quantities of sequence data (e.g. >100,000 sequences) are generated for many loci from an environmental sample (e.g., the gut, the permafrost, the ocean, etc). The sample is usually a complex mixture of DNA from hundreds or thousands of individuals and species and it is impossible to assign sequences at different loci to individuals. The computational complexity in the analysis of such data is a serious concern and single-locus methods that are reasonably accurate and computationally efficient are direly needed.

*Acknowledgments*. We thank Robert Cruickshank, Rupert A. Collins, and two anonymous referees for a number of constructive comments. This study is supported by a grant from Biotechnological and Biological Sciences Research Council (BBSRC) to Z.Y.

## References

Ahrens, D., T. Fujisawa, H. J. Krammer, J. Eberle, S. Fabrizi, and A. P. Vogler. 2016. Rarity and incomplete sampling in DNA-based species delimitation. Syst. Biol.

Alacs, E. A., A. Georges, N. N. FitzSimmons, and J. Robertson. 2010. DNA detective: a review of molecular approaches to wildlife forensics. Forensic Sci. Med. Pathol. 6:180-194.

Brown, S. D., R. A. Collins, S. Boyer, M. C. Lefort, J. Malumbres-Olarte, C. J. Vink, and R. H. Cruickshank. 2012. Spider: an R package for the analysis of species identity and evolution, with particular reference to DNA barcoding. Mol. Ecol. Resour. 12:562-565.

Collins, R. A., and R. H. Cruickshank. 2014. Known knowns, known unknowns, unknown unknowns and unknown knowns in DNA barcoding: a comment on Dowton et al. Syst. Biol. 63:1005-1009.

Dowton, M., K. Meiklejohn, S. L. Cameron, and J. Wallman. 2014. A preliminary framework for DNA barcoding, incorporating the multispecies coalescent. Syst. Biol. 63:639-644.

Fujisawa, T., and T. G. Barraclough. 2013. Delimiting species using single-locus data and the generalized mixed yule coalescent approach: a revised method and evaluation on simulated data sets. Syst. Biol. 62:707-724.

Hebert, P. D., A. Cywinska, S. L. Ball, and J. R. deWaard. 2003. Biological identifications through DNA barcodes. Proc. Biol. Sci. 270:313-321.

Hebert, P. D., M. Y. Stoeckle, T. S. Zemlak, and C. M. Francis. 2004. Identification of birds through DNA barcodes. PLoS Biol. 2:1657–1663.

Hickerson, M. J., C. P. Meyer, and C. Moritz. 2006. DNA barcoding will often fail to discover new animal species over broad parameter space. Syst. Biol. 55:729-739.

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

Jukes, T. H., 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. 1982. The coalescent. Stochastic Process Appl. 13:235-248.

Lim, G. S., M. Balke, and R. Meier. 2012. Determining species boundaries in a world full of rarity: singletons, species delimitation methods. Syst Biol 61:165-169.

Meier, R., K. Shiyang, G. Vaidya, P. K. L. Ng, and M. Hedin. 2006. DNA barcoding and taxonomy in *Diptera*: A tale of high intraspecific variability and low identification success. Syst. Biol. 55:715-728.

Meyer, C. P., and G. Paulay. 2005. DNA barcoding: error rates based on comprehensive sampling. PLoS Biol. 3:e422.

Pons, J., T. G. Barraclough, J. Gomez-Zurita, A. Cardoso, D. P. Duran, S. Hazell, S. Kamoun, W. D. Sumlin, and A. P. Vogler. 2006. Sequence-based species delimitation for the DNA taxonomy of undescribed insects. Syst. Biol. 55:595-609.

Puillandre, N., A. Lambert, S. Brouillet, and G. Achaz. 2012. Automatic barcode gap discovery for primary species delimitation. Mol. Ecol. 21:1864-1877.

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

Rannala, B., and Z. Yang. 2017. Efficient Bayesian species tree inference under the multispecies coalescent. Syst. Biol.

Rossini, B. C., C. A. Oliveira, F. A. Melo, V. A. Bertaco, J. M. Astarloa, J. J. Rosso, F. Foresti, and C. Oliveira. 2016. Highlighting *Astyanax* species diversity through DNA barcoding. PLoS One 11:e0167203.

Smith, P. J., S. M. McVeagh, and D. Steinke. 2008. DNA barcoding for the identification of smoked fish products. J. Fish Biol. 72:464-471.

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

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

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

Yang, Z. 2002. Likelihood and Bayes estimation of ancestral population sizes in Hominoids using data from multiple loci. Genetics 162:1811-1823.

Yang, Z. 2015. The BPP program for species tree estimation and species delimitation. Curr. Zool. 61:854-865.

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

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

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

## Figure legends

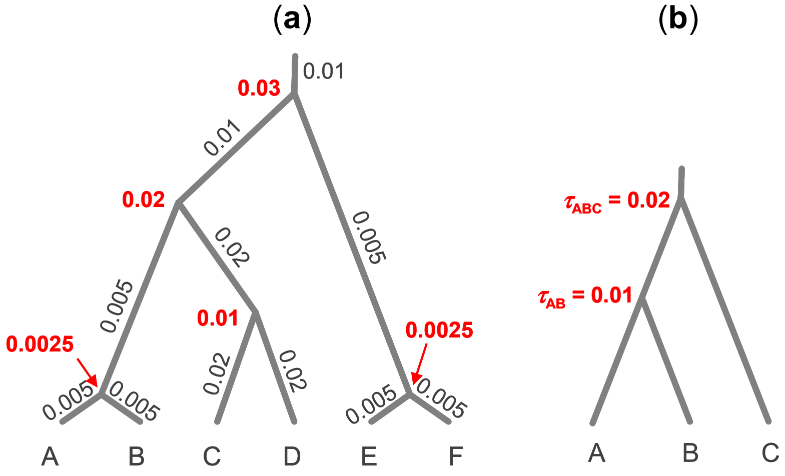

Fig. 1. Species trees used in simulating sequence alignment under the MSC, with branches drawn using the species divergence times ($\tau$s). In (a), the species divergence-time parameters are shown in bold next to the internal nodes: $\tau_{AB} = \tau_{EF} = 0.0025$, $\tau_{CD} = 0.01$, $\tau_{ABCD} = 0.02$, and $\tau_{ABCDEF} = 0.03$, while the population size parameters are shown along the branches: $\theta_A = \theta_B = \theta_{AB} = \theta_E = \theta_F = \theta_{EF} = 0.005$, $\theta_C = \theta_D = \theta_{CD} = 0.02$, and $\theta_{ABCD} = \theta_{ABCDEF} = 0.01$. In (b), the parameters are $\tau_{AB} = 0.01$, $\tau_{ABC} = 0.02$, and $\theta = 0.01$ for all populations. Both $\tau$s and $\theta$s are measured by the expected number of mutations per site.

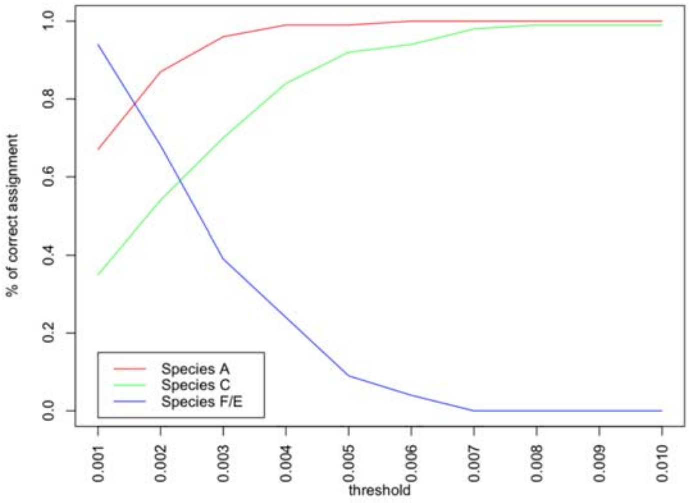

Fig. 2. The proportion of correct species assignments when the distance threshold is fixed at different values, averages over 100 replicate datasets, simulated using the species tree of Fig. 1a.

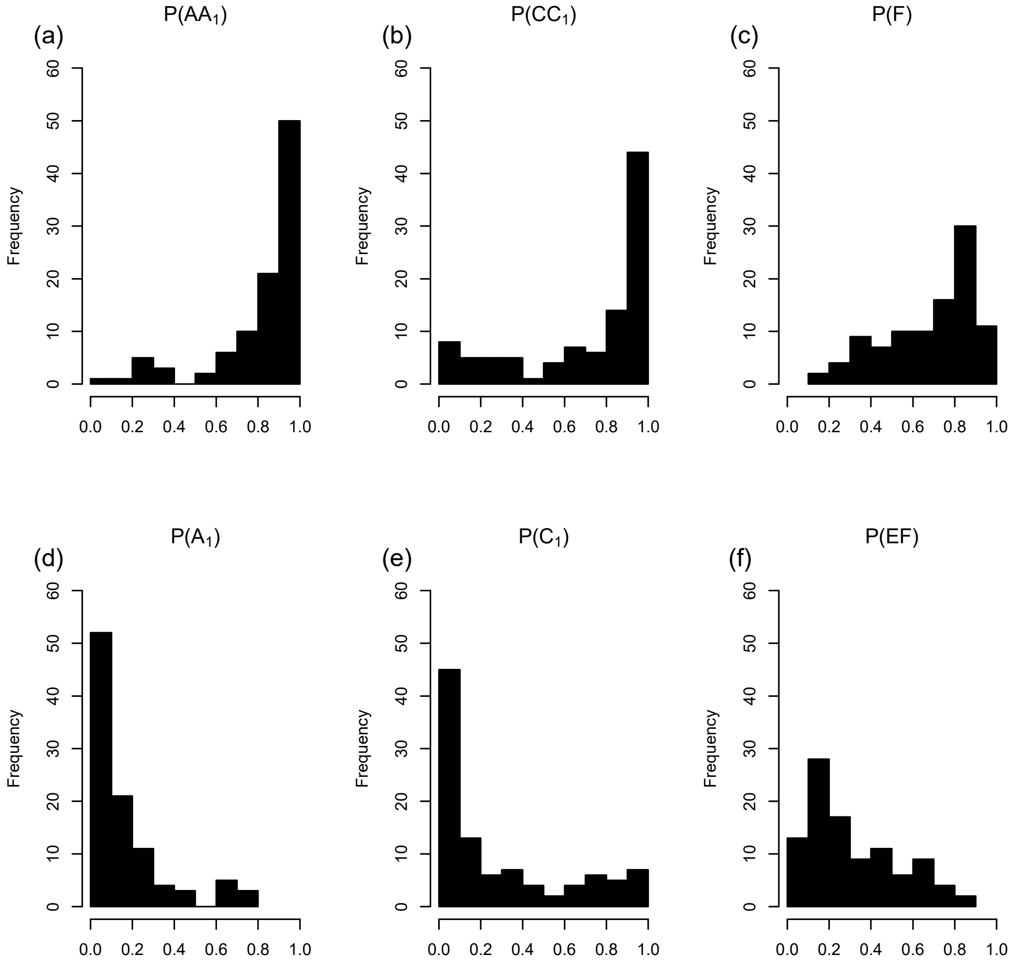

Fig. 3. The histograms of posterior probabilities for correctly assigning $A_1A$, $C_1C$, and $F$ into one species (**a-c**), and for incorrectly assigning $A_1$, $C_1$, $EF$ into one species (**d-f**) by BPP in datasets of one single locus, simulated using the species tree of Fig. 1a. The same data were analyzed using Barcoding in Fig. 2.

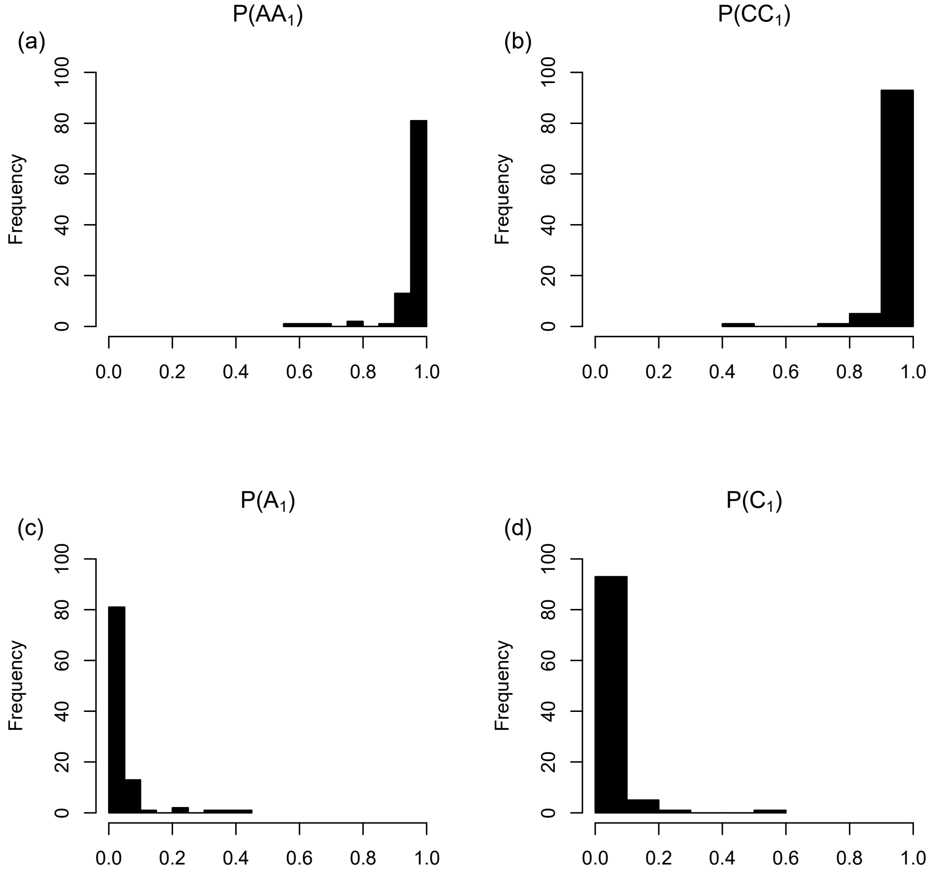

Fig. 4. The histograms of posterior probabilities for species identification and delimitation by BPP in datasets of 10 loci, simulated using the species tree of Fig. 1a. See legend for Fig. 3. P(*F*) = 1 and P(*EF*) = 0 in every dataset so that the plots for the query *F* are not shown.

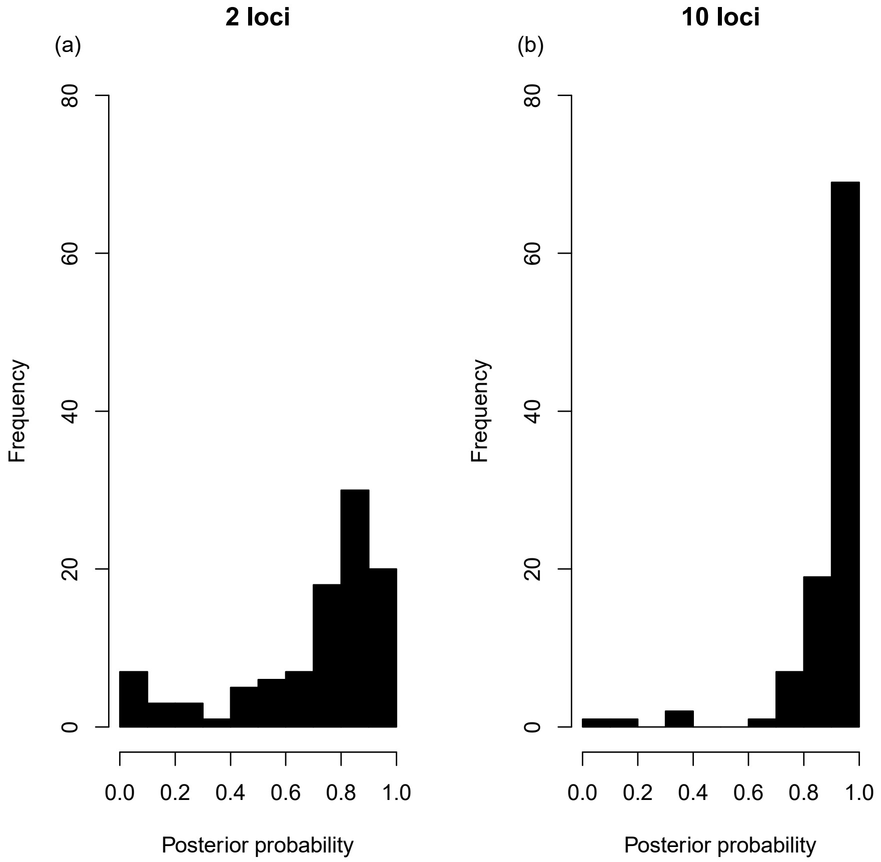

Fig. 5. The histograms of posterior probabilities for correctly inferring the cryptic species status as well as the species phylogeny by BPP using datasets of 2 and 10 loci, simulated using the species tree of Fig. 1b. There are 3 species in the dataset, with *A* and *B* representing two cryptic species.

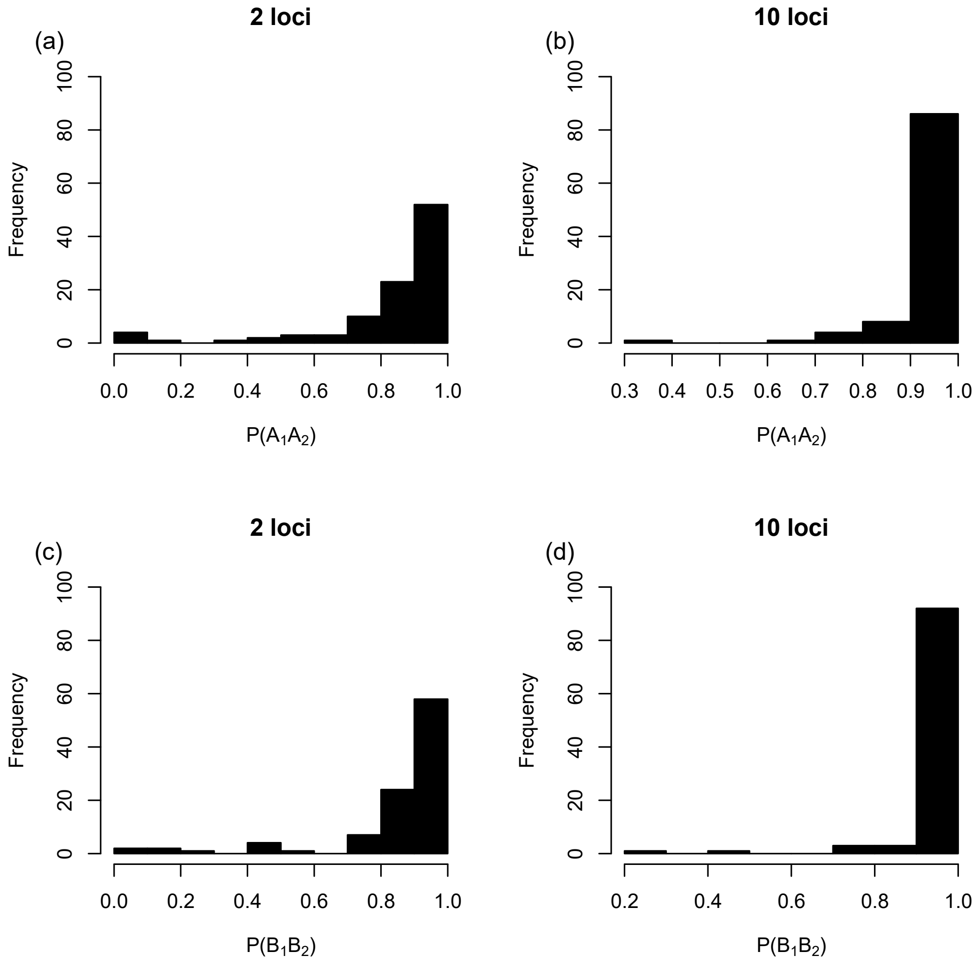

Fig. 6. The histograms of posterior probabilities for correctly identifying cryptic species by BPP using datasets of 2 and 10 loci. See legend to Fig. 5.