*Syst. Biol.* 60(6):747–761, 2011

DOI:10.1093/sysbio/syr071
Advance Access publication on August 29, 2011

# Evaluation of a Bayesian Coalescent Method of Species Delimitation

CHI ZHANG[1,2,3], DE-XING ZHANG[1,2], TIANQI ZHU[4], AND ZIHENG YANG[1,5,*]

[1]*Center for Computational and Evolutionary Biology, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China;* [2]*State Key Laboratory of Integrated Management of Pest Insects and Rodents, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China;* [3]*Graduate University of Chinese Academy of Sciences, Beijing 100049, China;* [4]*School of Mathematical Sciences, Peking University, Beijing 100871, China; and* [5]*Department of Biology, University College London, Darwin Building, Gower Street, London WC1E 6BT, UK;*
**Correspondence to be sent to: Department of Biology, University College London, Darwin Building, Gower Street, London WC1E 6BT, UK; E-mail: z.yang@ucl.ac.uk.*

*Received 16 November 2010; reviews returned 10 February 2011; accepted 20 April 2011*
*Associate Editor: Thomas Buckley*

*Abstract*.—A Bayesian coalescent-based method has recently been proposed to delimit species using multilocus genetic sequence data. Posterior probabilities of different species delimitation models are calculated using reversible-jump Markov chain Monte Carlo algorithms. The method accounts for species phylogenies and coalescent events in both extant and extinct species and accommodates lineage sorting and uncertainties in the gene trees. Although the method is theoretically appealing, its utility in practical data analysis is yet to be rigorously examined. In particular, the analysis may be sensitive to priors on ancestral population sizes and on species divergence times and to gene flow between species. Here we conduct a computer simulation to evaluate the statistical performance of the method, such as the false negatives (the error of lumping multiple species into one) and false positives (the error of splitting one species into several). We found that the correct species model was inferred with high posterior probability with only one or two loci when 5 or 10 sequences were sampled from each population, or with 50 loci when only one sequence was sampled. We also simulated data allowing migration under a two-species model, a mainland-island model and a stepping-stone model to assess the impact of gene flow (hybridization or introgression). The behavior of the method was diametrically different depending on the migration rate. Low rates at $< 0.1$ migrants per generation had virtually no effect, so that the method, while assuming no hybridization between species, identified distinct species despite small amounts of gene flow. This behavior appears to be consistent with biologists' practice. In contrast, higher migration rates at $\geqslant 10$ migrants per generation caused the method to infer one species. At intermediate levels of migration, the method is indecisive. Our results suggest that Bayesian analysis under the multispecies coalescent model may provide important insights into population divergences, and may be useful for generating hypotheses of species delimitation, to be assessed with independent information from anatomical, behavioral, and ecological data. [Species delimitation; coalescent; Bayesian inference; simulation; stepping-stone model; Lindley's paradox.]

Species have traditionally been identified based on morphological and behavioral traits, such as plumage, mating behavior, reproductive incompatibility, etc. The practice has much subjectivity and can vary widely among taxonomists working on different species. For example, Issac et al. (2004) observed that ant taxonomists tend to be "splitters," while butterfly taxonomists are "lumpers," so that species counts in those two groups are not comparable. Genetic sequence data can provide valuable information about processes related to speciation and species delimitation, such as gene flow (Hey 2010). Much recent interest has focused on the use of genetic sequence data to infer the species tree despite considerable gene tree conflicts, caused for example by ancestral polymorphism and lineage sorting. A number of programs have been developed for this purpose, including BUCKy (Ane et al. 2007), BEST (Liu 2008; Liu et al. 2009), STEM (Kubatko et al. 2009), and *BEAST (Heled and Drummond 2010). They assume that individuals are already correctly assigned to species although the species phylogeny is yet to be estimated.

Genetic data have also been used to delimit species. Compared with traditional morphological characters, genetic data have a clear advantage in delimiting cryptic species, which may be indistinguishable morphologically. Nevertheless, analyzing genetic data in their proper population genetic and genealogical framework is a challenging task. Some studies use arbitrary cutoffs on certain indicators of species status such as the amount of sequence divergence and the migration rate. For example, the "10× rule" requires the between-species divergence to be at least 10 times as large as the within-species polymorphism (Hebert et al. 2004). Another common strategy is to reconstruct gene trees at individual loci and then use the inferred gene trees for further analysis without accommodating errors in phylogeny reconstruction. Such errors may be substantial due to the high sequence similarity and low information content of the data and may have a large impact on inference concerning ancestral processes (Yang 2002). For example, delimitation of species using the genealogical species concept (Baum and Shaw 1995) has often relied on the gene trees at all loci (or the consensus of the gene trees at all loci) showing reciprocal monophyly. In addition to ignoring phylogenetic errors, the requirement for reciprocal monophyly is unnecessarily stringent because the expected time to achieve reciprocal monophyly at a neutral locus is very long (Neigel and Avise 1986; Hickerson et al. 2006) and because conflicting gene trees can be generated by the stochastic fluctuation of the coalescent process in the ancestral species (Hudson and Coyne 2002; Rannala and Yang 2003). Inferred gene trees are also used by Knowles and Carstens (2007; see also O'Meara 2010) as observed data to construct a likelihood ratio test (LRT) to compare the one-species

and two-species models. This method accommodates species tree–gene tree conflicts due to ancestral polymorphism and lineage sorting but ignores phylogenetic errors in gene tree reconstruction. Also, the authors' use of the $\chi^2$ with one degree of freedom for the LRT appears to be incorrect.

A Bayesian method for species delimitation using multilocus genetic sequence data has recently been developed by Yang and Rannala (2010). This uses Bayesian model selection to calculate posterior probabilities of different species delimitation models. For example, the one-species model assumes that the gene trees among loci are generated by the standard coalescent with one single population size parameter $\theta$ (Kingman 1982a,1982b; Hudson 1983; Tajima 1983). In contrast, a two-species model may involve three population size parameters (two $\theta$s for the two extant species and a $\theta$ for the common ancestor) and a parameter for the divergence time of the two species ($\tau$), with the multispecies coalescent model specifying the distribution of gene trees at different loci (Takahata et al. 1995; Yang 2002; Rannala and Yang 2003). Calculation of the Bayesian posterior probabilities for the two models allows one to assess whether the sequence data are compatible with the one-species model, or the two-species model has to be invoked to explain the data. The method makes use of concordance of gene trees across multiple loci as evidence for existence of multiple species but does not rely on reciprocal monophyly. It accounts for the species phylogeny, random fluctuations in the coalescent process, and uncertainties in the gene tree topology and branch lengths. The method has been used to delimit new species of African forest geckos by Leache and Fujita (2010; see Bauer et al. 2011; Fujita and Leache 2011 for discussions).

In this paper, we conduct a computer simulation to examine the statistical properties of the method. Sequence data at multiple loci are simulated assuming either the one-species model or the two-species model and are analyzed using the reversible-jump Markov chain Monte Carlo (rjMCMC) algorithms implemented in the program BPP (Yang and Rannala 2010) to calculate the posterior model probabilities. This part of the simulation extends the small-scale simulation of Yang and Rannala (2010) to include more parameter settings and to use more realistic priors on parameters. We are interested in two kinds of errors: the false positives (or the error of splitting the same species into two) and the false negatives (or the error of lumping two species into one). We also simulate data under several models involving migration (hybridization) to assess the impact of migration on the Bayesian inference. Whereas the current implementation of Yang and Rannala (2010) assumes no gene flow, it is interesting to know how much gene flow is sufficient to cause the Bayesian method to infer one species. A real data set of butterfly nuclear loci is analyzed to evaluate the impact of priors and to understand the similarities and differences between the rjMCMC algorithms and the $\tau$-threshold method suggested by Yang and Rannala (2010).

## METHODS

### *Simulation of Data*

*Generation of Gene Trees.*—Data at multiple loci were simulated using the program MCCOAL in the BPP package (Rannala and Yang 2003; Yang and Rannala 2010). A random genealogical tree with branch lengths was generated for each locus and used to "evolve" sequences along the branches of the tree. Sequences at the tips of the tree constitute the data to be analyzed. The simulation program MCCOAL allows migration, even though the inference program BPP assumes no migration.

If no migration is assumed in the simulation model, the gene trees (topology and branch lengths) follow distributions specified by the multispecies coalescent model (Rannala and Yang 2003) and are generated by simulating the coalescent process in each population (Hudson 2002).

Here we describe the simulation procedure under the migration model. In this study, we do not distinguish among different forms of gene flow, such as migration, hybridization, and introgression, and use those terms interchangeably. Migration rates are specified using the matrix $\mathbf{M} = \{M_{ij}\}$, where the (scaled) migration rate $M_{ij} = N_j m_{ij}$ is the expected number of migrants from population $i$ to population $j$ per generation and where $m_{ij}$ is the migration rate from populations $i$ to $j$ or the proportion of individuals in population $j$ that are immigrants from population $i$. The gene tree is generated by tracking the genealogy backwards in time in different time epochs, defined by the species/population tree, so that within each epoch the number of populations is fixed, as is the per-lineage migration rate. In each time epoch, the waiting time until the next event is sampled from an exponential distribution with the intensity parameter (total rate) to be the sum of the coalescent rates and the migration rates. Consider population $i$, with size $N_i$ and with $n_i$ lineages ancestral to the sample. With time measured in generations, the coalescent rate is $n_i(n_i-1)/2\times 1/(2N_i)$, whereas the migration rate from population $j$ (to population $i$) is $n_i m_{ji}$. Divide both coalescent and migration rates by $\mu$, so that time is measured by distance or the expected number of mutations per site. Then the coalescent rate becomes $n_i(n_i-1)/2\times 2/\theta_i$, whereas the migration rate from population $j$ (into population $i$) becomes $n_i m_{ji}/\mu = n_i M_{ji}/\theta_i \times 4$. Here $\theta_i = 4N_i\mu$ is the population size parameter for population $i$. The coalescent and migration rates are summed over all populations for the time epoch, and the total rate is used to sample the waiting time until the next event. Given the occurrence of the event, the event type (coalescent or migration) is sampled in proportion to their rates. This process is repeated until the time epoch is exhausted or until the most recent common ancestor for the whole sample is reached.

For example, for the species tree of Figure 1, there are 5 populations (3 extant and 2 ancestral), so the migration matrix $\mathbf{M}$ is of size $5\times 5$. Simulation is done in three time epochs. During the first epoch, which runs from time 0 (present) to $\tau_{12}$, there exist three populations (1, 2, and

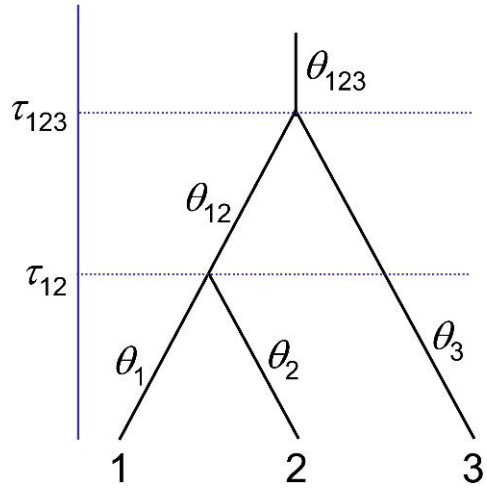


FIGURE 1. Species tree and parameters used to illustrate the algorithm for simulating gene trees under the multispecies coalescent model with migration. There are five populations, referred to as 1, 2, 3, 12 (the ancestor of 1 and 2), and 123 (the ancestor of 1, 2, and 3). This figure is available in black and white in print and in color at *Systematic Biology* online.

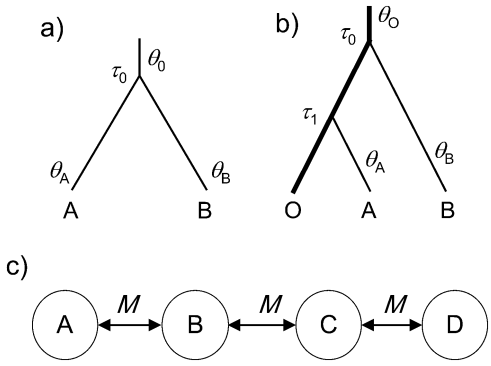


FIGURE 2. Species trees and parameters used in the simulation to generate sequences at multiple loci. a) A two-species model may have up to four parameters: the divergence time $\tau_0$ and three $\theta$ parameters for the three populations $(\theta_A, \theta_B, \theta_0)$. b) In the mainland-island model, the mainland population (O) has a large constant size with parameter $\theta_O$, whereas the two island populations B and A arose at times $\tau_0$ and $\tau_1$, respectively, and have since been receiving immigrants from the mainland. The population size parameters are assumed to be $\theta_A = \theta_B = 0.1\theta_O = 0.001$. c) In the stepping-stone model, migration occurs between two neighboring populations only, at the rate $M = Nm$. Samples are then taken from two populations for Bayesian analysis.

3), and migration may be possible between them. During the second epoch (from $\tau_{12}$ to $\tau_{123}$), there exist two populations: 12 and 3, and migration may be possible if $M_{12\to3} > 0$ or $M_{3\to12} > 0$. During the third epoch (from $\tau_{123}$), only population 123 exists, so that only coalescent events are possible.

Coalescent events create new nodes in the gene tree. The branch length is calculated as the difference between the ages of the two nodes at the ends of the branch.

We confirmed the correctness of the simulation program by comparison with the theoretical results of Wilkinson-Herbots (2008), who gives the expectations of the coalescent times between two sequences under several models of population subdivision and migration. The results for this program validation are in Supplementary material (available from http://www.sysbio.oxfordjournals.org).

*Simulation of Sequence Alignments.*—After the gene tree with branch lengths is generated for each locus, sequences at the tips of the gene tree are simulated by "evolving" sequences along the branches. The JC69 model (Jukes and Cantor 1969) is used in both the simulation and analysis of the data. We assume no recombination between sites within each locus and free recombination between loci, so that gene trees are independent across loci. The sequence length at each locus is 1000 sites.

*The One-Species and Two-Species Models.*—We examined the simplest case of comparing the one-species and two-species models (Fig. 2a). The sequence data were simulated by fixing the species divergence time $\tau_0$ at 0 (one-species model), or at 0.001 or 0.01 (two-species model), whereas all $\theta$ parameters were either 0.01 or 0.001. Both $\tau_0$ and $\theta$ are measured by distance: $\tau_0 = 0.01$ means that the sequence at the root of the two-species tree of Figure 2a is about 1% different from the sequences at the tips A or B, whereas $\theta = 0.001$ means that two random sequences drawn from the population are ~0.1% different. Note that at such low sequence divergences, correction for multiple hits has negligible effect. Previous estimates of $\theta$ for extant species include 0.0006 for humans (Rannala and Yang 2003), ~0.01 for the mangroves (Zhou et al. 2007), and a broad range (0.0005–0.02) for a variety of animal and plant species (Zhang and Hewitt 2003). Estimates of $\theta$ for ancestral species tend to be much larger than for modern species, but it is unclear whether the pattern reflects true biological processes (such as population subdivision) or is due to analytical artifact caused by gene flow at the time of speciation creating variable divergence times among loci (Wu and Ting 2004; Yang 2010). At any rate, the values 0.001 and 0.01 for $\theta$ may be representative of many species. Estimates for $\tau$ include 0.004 for the human–chimpanzee divergence (Rannala and Yang 2003) but may be much smaller for recently diverged species. For the butterfly data analyzed below, $\theta \approx 0.005$ and $\tau \approx 0.0013$.

We considered three sample configurations: (1, 1), (1, 5), and (5, 5), where $(n_1, n_2)$ means sampling $n_1$ sequences from species A and $n_2$ sequences from species B. In a few cases, larger samples were used as well to examine the impact of the sample size on the inference. The data were then analyzed using the rjMCMC algorithms in the program BPP (Yang and Rannala 2010) to compare the two models. We used the gamma priors $\theta \sim G(1, 100)$ for all $\theta$s and $\tau_0 \sim G(1, 100)$ for the root of the species tree. For data simulated under the two-species model, the guide tree used had the correct assignment of the individuals to species. For the data simulated under the one-species model, the guide tree

was generated by random assignment of individuals to the two potential species. This part of the simulation without migration extends the simulation of Yang and Rannala (2010) and provides a basis for comparison with simulations that involve migration.

We then simulated data sets under the two-species models assuming migration between the two species/populations. The migration rate is assumed to be the same in the two directions, and migration rate is measured by the expected number of immigrants in one generation in each population. The data sets are then analyzed using the rjMCMC algorithms as above to calculate the posterior probabilities for the one-species and two-species models. The same priors $\theta \sim G(1, 100)$ and $\tau_0 \sim G(1, 100)$ are assumed. Note that the BPP analysis assumes no gene flow.

*The Mainland-Island Model.*—The species tree is shown in Figure 2b. The mainland population O has the constant size with parameter $\theta_O$, and the two island populations A and B were much smaller with $\theta_A$ and $\theta_B$. We assumed $\theta_O = 0.01$ and $\theta_A = \theta_B = 0.001$. Populations B and A diverged from population O at times $\tau_0$ and $\tau_1$, respectively. We used $\tau_0 = 0.01$ and considered three values for $\tau_1$: 0.001, 0.005, and 0.009. We considered a case of no migration as well as a case of migration from the mainland to the islands, with $M_{OA} = M_{OB} = M$, whereas migration from the islands A or B to the mainland O, or between A and B, was absent. The priors used in data analysis were $\theta \sim G(1, 100)$ for all $\theta$s and $\tau_0 \sim G(1, 100)$.

*The Stepping-stone Model.*—Data were simulated using a linear stepping-stone model with four populations of equal size ($\theta$) (Fig. 2c). Migration was allowed to occur at the same rate $M = Nm$ in both directions between any two adjacent populations, whereas migration between other populations was absent. Our simulation program MCCOAL assumes a population/species tree. The stepping-stone model is equivalent to the isolation-migration model with infinite species divergence times. Thus, we simulated this model by assuming an arbitrary species tree, for example ((A, B), (C, D)), with the three divergence times ($\tau$s) being much greater than the $\theta$s so that the root of the gene tree was expected to be much younger than all the divergence times. The sample configuration was (5, 5), with five sequences each taken from two populations: (a) A and B, (b) A and C, or (c) A and D, whereas no sample was taken from the other two populations. The data were then analyzed to compare the one-species and two-species models. The parameter values were $\theta = 0.01$ for all four populations and $M = 0.001$, 0.01, 0.1, 1, 10, and 100. The priors used in the analysis were $\theta \sim G(1, 100)$ for all $\theta$s and $\tau_0 \sim G(1, 100)$.

*Variable Rates among Loci.*—To examine the impact of the variable mutation rates among loci on posterior probability for different species models, we simulated sequence alignments assuming that the mutation rate for each locus is a random variable from the gamma distribution $G(1, 1)$. The parameters used were $\theta = 0.01$ and $\tau_0 = 0$ (one species) or 0.01 (two species). These are defined using the average mutation rate over all loci. The data were analyzed using the rjMCMC to compare the one-species and two-species models assuming either a constant rate for all loci or a model of variable rates among loci, described using a Dirichlet distribution with $\alpha = 2$ (Burgess and Yang 2008, equation 4). Note that the gamma and Dirichlet models for variable rates among loci are equivalent except for a slight difference in Bayesian parametrization: the former assumes that the rate for each of the $L$ loci has expectation 1: $E(r_i) = 1$, whereas the latter assumes that the average rate across the $L$ loci is 1: $(r_1 + r_2 + \ldots + r_L)/L = 1$.

*Running the rjMCMC Algorithms.*—For each parameter setting, 1000 replicate data sets were simulated. Each data set was analyzed using the two rjMCMC algorithms described in Yang and Rannala (2010) to ensure that the results were stable between runs. The two MCMC samples were then merged to calculate the posterior probabilities for the different species delimitation models. Good fine-tune parameters were obtained by trial and error according to the manual for BPP and they differ for different simulation conditions such as different data sizes.

### *Analysis of an Empirical Data Set of* Heliconius *Butterflies*

Sequence data at four nuclear autosomal loci for two sibling butterfly species *Heliconius demeter* and *H. eratosignis* were kindly provided by James Mallet and Kanchon Dasmahapatra. Those species were identified as *H. demeter ucayalensis* and *H. demeter* ssp. nov. in Dasmahapatra et al. (2010), but there now seems to be little doubt that they are "good" separate species, referable to the already published names *Heliconius demeter* and *Heliconius eratosignis* (K. Dasmahapatra and J. Mallet, personal communication). The two cryptic species are largely allopatric or parapatric, but they overlap in sympatry without evidence of hybridization at Tarapoto, Peru, from where the data were obtained (Dasmahapatra et al. 2010). The four loci are *Ef1a* (18 sequences, 766 bp), *Mpi* (9 sequences, 496 bp), *Rp15* (15 sequences, 713 bp), and *Tektin* (9 sequences, 733 bp).

We have two objectives with analysis of this data set. First, we are interested in the impact of the priors on the Bayesian comparison of the species models. Thus, the data were analyzed using the rjMCMC algorithms with different priors for $\theta$ and $\tau_0$ to calculate the posterior model probabilities. Second, we used the data to examine the similarities and differences between the rjMCMC algorithm and the $\tau$-threshold method (Yang and Rannala 2010). The $\tau$-threshold method runs the ordinary MCMC (instead of the rjMCMC) under the two-species model and then evaluates the posterior probability that the divergence time $\tau$ is less than a

prespecified threshold value ($\tau_T$). This was implemented by Yang and Rannala (2010) as an alternative to the rjMCMC algorithms, which often have mixing problems in large data sets. Here the data set is small enough for both methods to be applicable. Mathematically both the rjMCMC method and the $\tau$-threshold method are just different priors on $\tau_0$ under the same two-species model. The rjMCMC uses a mixture prior on $\tau_0$: a component of 0 and another component from the gamma distribution, each with probability 50%, whereas the $\tau$-threshold uses a simple gamma prior. The case is similar to the use of the gamma model versus the invariable sites plus gamma model to accommodate variable rates among sites in phylogenetic analysis (i.e., the "+Γ" and "I + Γ" models) (Yang 1996). However, in a phylogenetic analysis, the focus is on the phylogeny and branch lengths with the rate distribution to be of secondary importance. Here $\tau_0$ is the focus.

## RESULTS

### *Simulation Comparing the One-Species and Two-Species Models*

We simulated data under either the one-species model ($\tau_0 = 0$) or the two-species model ($\tau_0 > 0$) and analyzed them using the rjMCMC algorithms to calculate the posterior probabilities for the two models. Let these be $P_1$ and $P_2$, with $P_1 + P_2 = 1$. The results are shown in Figure 3.

First we consider Figure 3a–c, which shows $P_1$ for data simulated with $\tau_0 = 0$ (one species). $P_1$ is quite high even with one locus. Note that without data, the two models have probability 1/2 each from the prior. If we would like to avoid species inflation and consider the false-positive error (the error of incorrectly selecting the two-species model) to be serious and claim that there are two species only if $P_2 > 95\%$, we may calculate the false-positive error rate, akin to the type I error rate in frequentist hypothesis testing. For the simulations of Figure 3a–c, the false-positive rates are very low, at $\leq$0.5% for $\theta = 0.001$ and $\leq$1% for $\theta = 0.01$. Note that the probability $P_1$ for the correct one-species model can reach ~1 even with one locus when five sequences were sampled from the population. Indeed $P_1$ was higher for the sample configuration (5, 5) with only one locus (with a total of 10 sequences) than for the sample configuration (1, 1) with 10 loci (with a total of 20 sequences) (cf. Fig. 3a with Fig. 3c).

Figure 3d–i shows $P_2$ for simulation with $\tau_0 = 0.001$ and 0.01 so that the true model is the two-species model, and incorrectly selecting the one-species model may be considered a false-negative error. $P_2$ was much lower when $\tau_0 = 0.001$ than when $\tau_0 = 0.01$: the two species must be much harder to identify if their divergence is more recent. For the sample configurations (1, 1) and (1, 5), $P_2$ is close to 1 when 50 loci are available, whereas for the configuration (5, 5), $P_2$ is close to 1 when five or more loci are available. The "power" of the method appeared to be quite high. To see the impact of sampling,

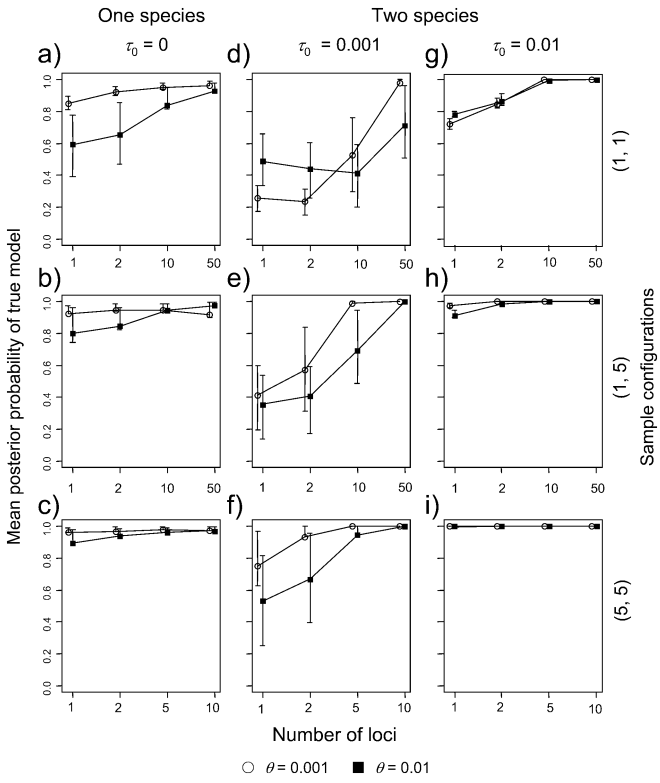


FIGURE 3. Mean posterior probability for the correct model when the data are simulated under the one-species model ($\tau_0 = 0$) or under the two-species model ($\tau_0 > 0$) (Fig. 2a) without gene flow. The quartiles (25% and 75% points) are shown as error bars. The priors used in the analysis are $\theta \sim G(1, 100)$ and $\tau_0 \sim G(1, 100)$. Note that the *x*-axis is not to scale and that for clarity the points are slightly off position.

we also simulated data sets of one locus for the sample configurations (10, 10), (15, 15), and (20, 20) for the case of $\tau_0 = 0.001$ (cf. Fig. 3f). For $\theta = 0.001$, $P_2 = 0.91$, 0.96, and 0.98 for the three configurations, respectively, whereas for $\theta = 0.01$, the corresponding values are $P_2 = 0.83$, 0.97, and 0.99, respectively. It is noteworthy that the method can infer the correct two-species model with posterior probability close to 1 with just one locus, as long as a large sample is taken from each population. Similarly, $P_2$ is higher for the sample configuration (5, 5) with only one locus than for the sample configuration (1, 1) with 10 loci (cf. Fig. 3d with Fig. 3f). This pattern is in contrast to the estimation of $\theta$ from a single population, in which inclusion of more sequences adds very little information when three to five sequences are already sampled, because coalescent events occur extremely quickly near the tips of the gene tree (e.g., Felsenstein 1992). The information used in the comparison of species tree models is clearly different from that used for estimating a single $\theta$: for example, reciprocal monophyly of the gene tree for two large samples from the two populations will be strong evidence for two distinct species.

In smaller data sets, for example, with one or two loci for the sample configurations (1, 1) or (1, 5), $P_2$ can be rather low, with substantial false-negative errors. For some parameter combinations with only one or two loci

(Fig. 3d,e), $P_2 < 1/2$, so that the method performed more poorly than without data. This is because of the impact of the priors in small data sets. The prior means for $\tau_0 \sim G(1, 100)$ and $\theta \sim G(1, 100)$ are much larger than the true values, so that the priors are somewhat in conflict with the data, leading to reduced support for the two-species model (see Discussion section). Similar results were observed in the simulation of Yang and Rannala (2010), where the priors used were even more extreme and unrealistic.

### *The Impact of Migration*

*Two Populations with Migration.*—Migration should have the effect of homogenizing the populations and cause the Bayesian analysis, which ignores migration, to favor the one-species model. Figure 4 shows $P_2$ for data simulated under the two-species model with migration but analyzed assuming no migration. If the migration rate is low, with <0.1 migrants per generation, the posterior probability for the correct two-species model $P_2$ was nearly identical to those when there is no migration (cf. Fig. 4a with Fig. 3g and Fig. 4b with Fig. 3i for the case of $\theta = 0.01$ with 1 locus or 10 loci). Migration at this level appeared to have little impact on Bayesian species delimitation.

If the migration rate is high, with ≥10 migrants per generation, $P_2$ is near zero for most settings except for the very small data sets with sample configuration (1, 1) and 1 locus, in which $P_2$ was moderate, influenced by the prior. At this level of migration, the method strongly favors the one-species model.

If the migration rate is moderate with one migrant per generation $(M = 1)$, $P_2$ is neither very high nor very low in small data sets, although in large data sets, with the configuration (10, 10) or (20, 20) and 10 loci, $P_2 \approx 1$.

Note that if $M = 0$, the two-species model is true, whereas if $M \to \infty$, the one-species model is true. For Bayesian species delimitation under the parameter settings used here, the "phase change" appeared to occur around $M = 1$ migrant per generation or in the range $0.1 < M < 10$. We extended the simulation of Figure 4b for 10 loci with the configuration (5, 5), to explore further the impact of the migration rate $M$ and the divergence time $\tau_0$. The same priors were used as before: $\theta \sim G(1, 100)$ and $\tau_0 \sim G(1, 100)$. The results are shown in Figure 5. The effect of migration was striking and diametrically different depending on the migration rate: when the migration rate was low with <0.1 immigrants per generation, the method behaved as if there was no migration. At high migration rate with 5 or 10 immigrants per generation, the method inferred one species. In comparison, the effect of the divergence time $\tau_0$ was minor. When $M = 5$ or 10, $P_2$ did not increase with the increase of $\tau_0$.

The apparent peaks in $P_2$ around $\tau_0 = 0.0005$–$0.001$ for moderate levels of migration (with $M = 1$ and 5) were apparently due to the impact of the prior: $P_2$ tends to be high when the prior on $\tau_0$ is consistent with the data (Discussion and unpublished results).

Overall, the results appear to be consistent with the theories in population genetics that examine the impact of migration on population differentiation at neutral loci measured by $F_{ST}$ (e.g.,Takahata 1983): if $M = Nm \ll 1$, the populations will be strongly differentiated, whereas if $M \gg 1$ (e.g., if there are more than 10 or so migrants),

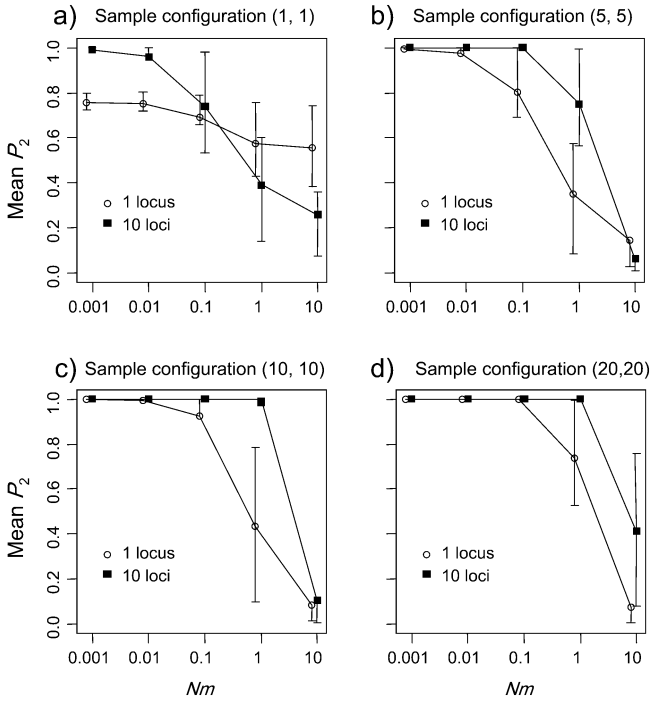


FIGURE 4. Mean posterior probability for the correct model ($P_2$) when the data are simulated under the two-species model with migration (Fig. 2a). The parameters are $\theta = \tau_0 = 0.01$. The priors used in the analysis are $\theta \sim G(1, 100)$ and $\tau_0 \sim G(1, 100)$. The migration rate is measured by $M = Nm$, the expected number of migrants per generation. Note that the $x$-axis is on the logarithmic scale and the points are shifted off position for clarity.

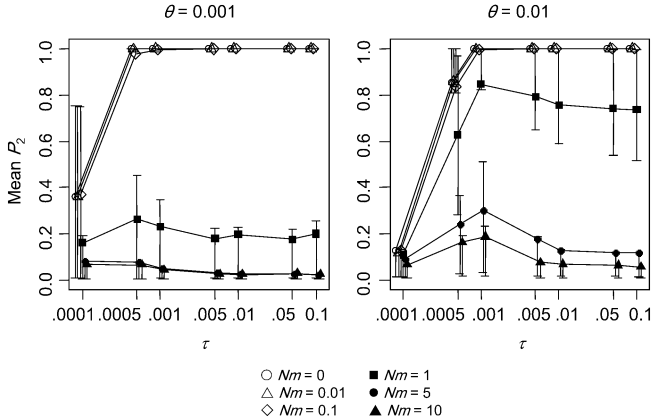


FIGURE 5. Mean posterior probability for the two-species model ($P_2$) when the data are simulated under the two-species model with and without migration (with $Nm = 0$ or $> 0$). The data consist of 10 loci with sample configuration (5, 5). The priors used in the analysis are $\theta \sim G(1, 100)$ and $\tau_0 \sim G(1, 100)$. Parts of the results are shown in Figure 4b.

migration will swamp the population and the population will behave as a panmictic unit. It is interesting that Bayesian comparison of species models led to the same conclusion from a very different perspective.

*The Mainland-Island Model.*—In this simulation, the mainland population (O) has had a large and constant size with parameter $\theta_O = 0.01$, and it gave rise to two island populations B and A through dispersal, with parameters $\theta_A = \theta_B = 0.001$ (Fig. 2b). We fix the divergence time at the root of the species tree at $\tau_0 = 0.01$, whereas three values are used for $\tau_1$: 0.001, 0.005, and 0.009. Migration is always from the mainland to the islands with the scaled migration rate to be $M = 0, 0.001, \ldots, 10$. In the analysis of the data, the true species tree, ((O, A), B), was used as the guide tree to run the rjMCMC algorithms. There are three species models, generated by collapsing none, one, or both of the two internal nodes in the guide tree, respectively (Fig. 2b): the three-species model, the two-species model (with two species B and OA) and the one-species model. Let the posterior probabilities for them be $P_3, P_2$, and $P_1$. Figure 6a–c shows $P_3$, whereas Figure 6d–f shows $P_3, P_2$, and $P_1$ for the sample configuration (5, 2, 2) with 1 locus. As in the analysis of the one-species and two-species models, including more sequences from the same population increased the power of the method considerably, so that $P_3$ for (5, 2, 2) with 1 locus was higher than $P_3$ for (2, 1, 1) with five loci. Migration at rates lower than 0.01 migrants per generation had little impact, whereas one migrant per generation tended to lead to the inference of one species. Migration appeared to be more important here than in the comparison between the one-species and two-species models in Figure 4. For example, the results for $M = 0.1$ and 0 were quite different in Figure 6 but were similar in Figure 4. This appears to be due to the fact that here $\theta_A = \theta_B = 0.001$ is 10 times smaller than in Figure 4. The impact of migration is affected not only by the number of immigrants ($Nm$) but also by the

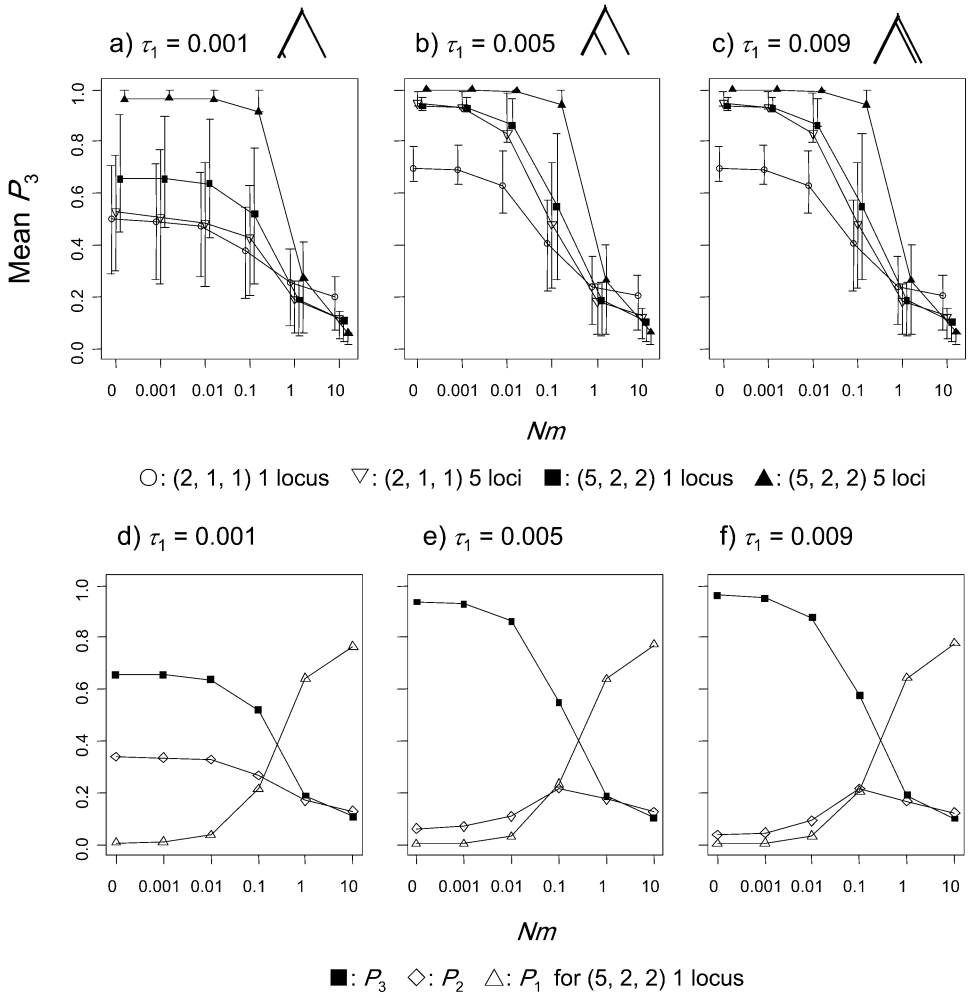


FIGURE 6. Mean posterior probabilities for different species-tree models when the data are simulated under the mainland-island model (Fig. 2b). In (a)–(c), the mean posterior probability $P_3$ for the three-species model is shown, whereas in (d)–(f), the mean posterior probabilities for all three species models $(P_3, P_2, P_1)$ are shown for the sample configuration (5, 2, 2) and 1 locus. The parameters used are $\theta_O = 0.01, \theta_A = \theta_B = 0.001, \tau_0 = 0.01$, whereas $\tau_1 = 0.001$ (a,d), 0.005 (b,e), and 0.009 (c,f). The priors used in the analysis are $\theta_i \sim G(1, 100)$ and $\tau_0 \sim G(1, 100)$.

population size: at the same $Nm$, a smaller population size ($N$) means a larger proportion of immigrants ($m$).

*The Stepping-stone Model.*—The data were simulated under the equilibrium migration model of Figure 2c, but samples were taken from two populations only. We refer to the data sets as AB, AC, and AD data, respectively, depending on the two populations sampled. This simulation is intended to mimic geographical isolation, where a species has a very broad geographical distribution with migration occurring between close localities only. The concern for the Bayesian method of species delimitation is that with samples taken from distant localities, the method may be misled to infer two species because the migration rate between them is very low. We took two samples from localities that are close by (AB data), intermediate (AC data), or very distant (AD data), and then run the rjMCMC algorithm to compare the one-species and two-species models.

Figure 7 shows the mean posterior probability for the two-species model ($P_2$). First, we consider the results for the AB data. If the migration rate is low, at $\leq$0.1 migrants per generation between any two adjacent populations, $P_2$ is near 1. With 10 or more migrants per generation, $P_2$ is near 0 so that BPP will support the one-species model with probability near 1. The results are similar to those for the two-species case of Figure 4b.

The results for the AD data (Fig. 7) show that $P_2$ for the AD data was not much higher than for the AB and AC data. At $M > 1$, the populations appear to have been homogenized by migration, so that the one-species model is strongly supported by the Bayesian method. This result may be surprising, as intuitively one may expect the migration rate between A and D to be close to $m^3$ if the rate between A and B is $m$. However, this intuition is incorrect. It is known from analysis of similar stepping-stone models in population genetics that the migration rate between A and D is in the order of $m/3$ instead of $m^3$ (Strobeck 1987; Slatkin 1991). The results of Figure 7 are consistent with this theory. As this faulty intuition appears to be common, we include a more detailed version of Slatkin's proof in the Appendix.

The results of Figure 7 suggest that the Bayesian inference may be quite robust to complex population structures or ghost populations (Beerli 2004; Wakeley and Aliacar 2001). When sequences are sampled from A and D and used for species delimitation, the method is not misled to infer two species even though the intermediate populations (B and C) are not sampled.

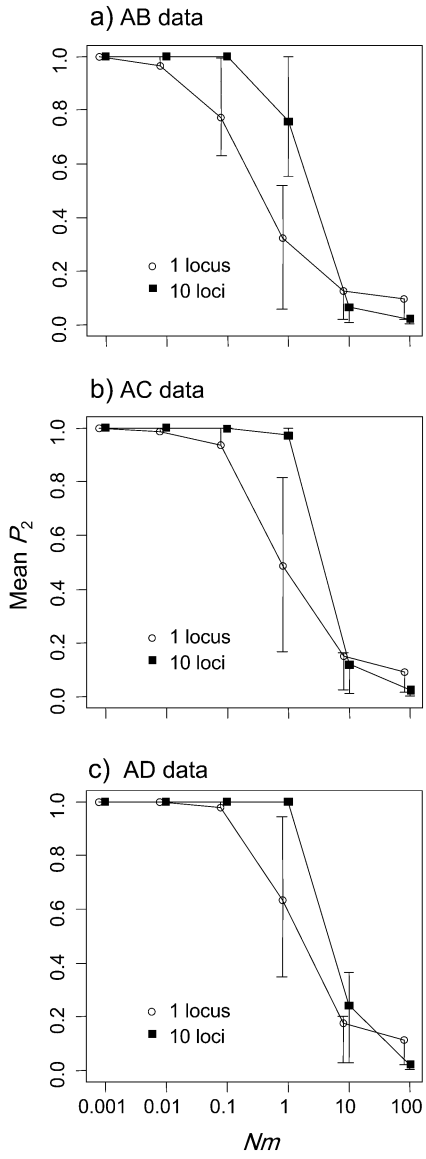


FIGURE 7. Mean posterior probability for the correct model ($P_2$) when the data are simulated under the stepping-stone model of Figure 2c. The size of each of the four populations is $\theta = 0.01$, whereas $M = Nm$ is the migration rate between two adjacent populations. The sample configuration is (5, 5), with samples taken (a) from A and B, (b) from A and C, and (c) from A and D. The data were then analyzed to compare the one-species and two-species models. The priors used in the analysis are $\theta \sim G(1, 100)$ and $\tau_0 \sim G(1, 100)$.

### *The Impact of Mutation Rate Variation among Loci*

Inference of ancestral population parameters relies to some extent on the stochastic fluctuation of the coalescent process among loci, generating different topologies and branch lengths in the gene trees. It may thus be a concern that such inference may be sensitive to mutation rate variation among loci. To examine the impact of the variable mutation rates on the posterior probability for different species models, we simulated sequence alignments under the one-species and two-species models assuming that the mutation rate for each locus is a random variable from the gamma distribution $G(1, 1)$. The shape parameter $\alpha = 1$ may be too small if the multiple loci represent noncoding genomic regions but reasonable if coding regions are used as well (Yang 1996). The parameters used were $\theta = 0.01$ and $\tau_0 = 0$ (one species) or 0.01 (two species). These are defined as averages over all loci.

The data were analyzed using the rjMCMC algorithms to compare the one-species and two-species

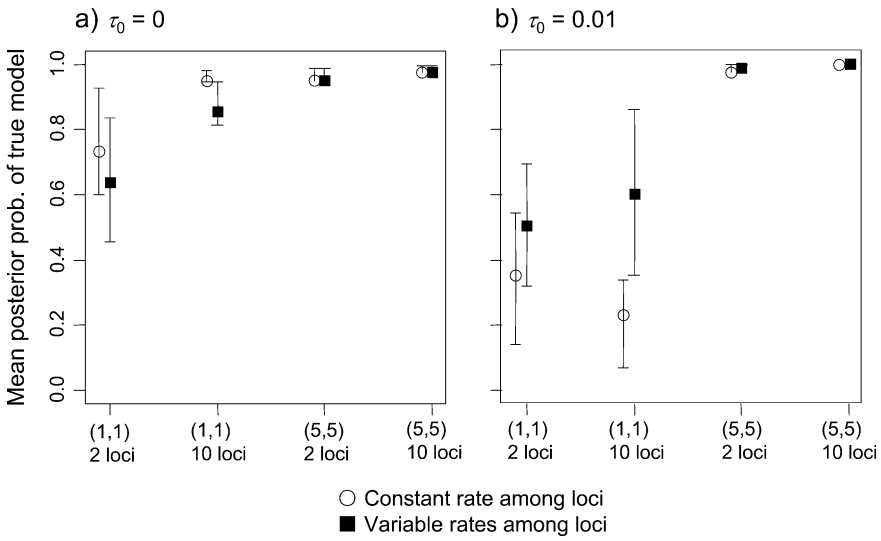


FIGURE 8. Mean posterior probability for the correct species model when the data are simulated under a model of variable rates among loci. The sequence data were simulated under the one-species ($\tau_0 = 0$) and two-species ($\tau_0 > 0$) models, with variable mutation rates among loci drawn from $G(1, 1)$. The parameters used were $\theta = 0.01$ and $\tau_0 = 0$ (a) or 0.01 (b). The data were analyzed under both the model of a constant rate for loci (○) and the model of variable rates among loci (■).

models assuming either a constant rate for all loci or a model of variable rates among loci, modeled using a Dirichlet distribution with $\alpha = 2$ (Burgess and Yang 2008). The results are shown in Figure 8. In small data sets, that is, with the sample configuration (1, 1) and 2 or 10 loci, the posterior probability for the one-species model ($P_1$) is higher when the rate variation among loci is ignored than when it is accommodated in the model. In other words, incorrectly ignoring rate variation among loci causes the method to unduly favor the one-species model, whether the true model used for data generation is one species or two species. The reasons for this effect are not well understood. However, we examined the posterior parameter estimates. Under the one-species model, the single parameter $\theta$ is very slightly overestimated when rate variation is ignored. Under the two-species model, parameter $\theta_{AB}$ for the ancestor is seriously overestimated and $\tau$ is seriously underestimated when rate variation is ignored. For example, for the case of $\tau_0 = \theta = 0.01$ and when rate variation is ignored, the average posterior means were 0.0217 for $\theta_{AB}$ and 0.0019 for $\tau$ for data configuration (1, 1) with 10 loci, and were 0.0254 for $\theta_{AB}$ and 0.0039 for $\tau$ for the configuration (5, 5) with 10 loci. When rate variation was accommodated, the posterior means were close to the true values: 0.0125 for $\theta_{AB}$ and 0.0081 for $\tau$ for configuration (1, 1) with 10 loci, and 0.0092 for $\theta_{AB}$ and 0.0103 for $\tau$ for configuration (5, 5) with 10 loci. The underestimated $\tau$ and overestimated $\theta$ when rate variation among loci is ignored should make the two-species model look similar to the one-species model, which may explain the increased $P_1$. The overestimation of ancestral $\theta$ in the case of two species when rate variation among loci is ignored was discussed extensively by Yang (1997).

In large data sets, that is, with the sample configuration (5, 5) and with 2 or 10 loci, the posterior probability for the correct model ($P_1$ in Fig. 8a and $P_2$ in Fig. 8b) is ~1 whether rate variation among loci is ignored or accommodated, so there is little difference between the two analyses.

### *Analysis of the Butterfly Data Set*

Sequence data at four nuclear loci from the butterfly species *H. demeter* and *H. eratosignis* were analyzed. First, we apply the two-species model to obtain some basic parameter estimates. Use of the priors $\theta \sim G(2, 500)$ and $\tau_0 \sim G(2, 2000)$ led to the following posterior estimates (mean and 95% confidence interval): 0.0058 (0.0027, 0.0108) for $\theta_D$, 0.0048 (0.0023, 0.0086) for $\theta_E$, 0.0078 (0.0041, 0.0130) for $\theta_O$, and 0.0013 (0.0006, 0.0022) for $\tau_0$. The priors were noted to have some impact on the estimates of the $\theta$ parameters. For example, with the priors $\theta \sim G(2, 2000)$ and $\tau_0 \sim G(2, 200)$, the estimates were 0.0035 (0.0019, 0.0057) for $\theta_D$, 0.0032 (0.0018, 0.0051) for $\theta_E$, 0.0043 (0.0022, 0.0069) for $\theta_0$, and 0.0014 (0.0006, 0.0024) for $\tau_0$. Note that the estimates of $\tau_0$ (at ~0.0013) were quite stable. If we use a mutation rate of $10^{-9}$ substitutions per site per year, this $\tau_0$ estimate will translate into 1.4 myr of divergence between the two species.

We then use the rjMCMC algorithms to calculate the posterior probability for the two-species model ($P_2$) with different priors for $\theta$s and $\tau_0$. The one-species model involves a single parameter $\theta$, whereas the two-species model involves four parameters: $\theta_D$ and $\theta_E$ for the two extant species, $\theta_0$ for the ancestor, and $\tau_0$. Figure 9a and b plots $P_2$ against the parameters in the

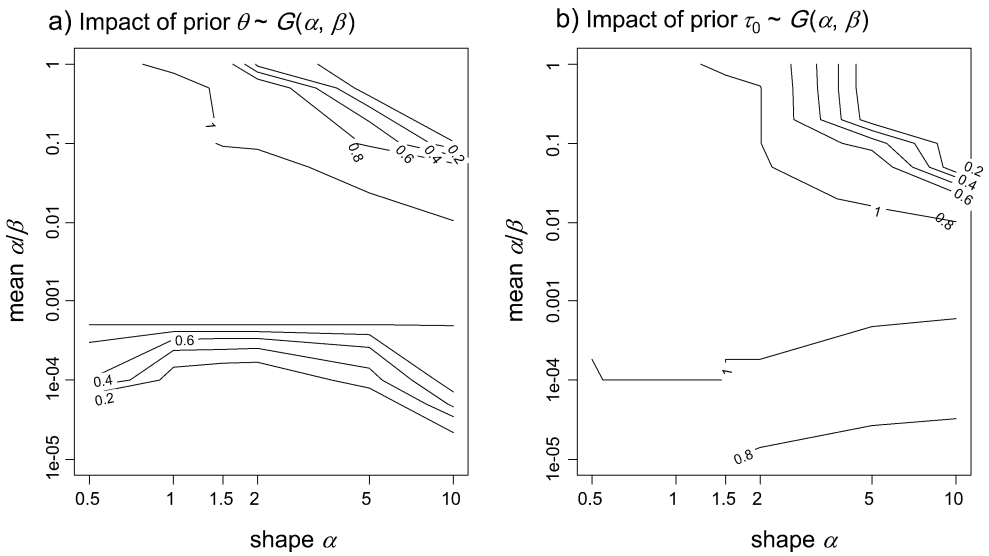


FIGURE 9. Contour plot of the posterior probability ($P_2$) for the two-species model plotted (a) against the parameters in the prior $\theta \sim G(\alpha, \beta)$, with $\tau_0 \sim G(2, 2000)$ fixed; and (b) against the parameters in the prior $\tau_0 \sim G(\alpha, \beta)$, with $\theta \sim G(2, 500)$ fixed. In each case, $P_2$ is plotted against the shape parameter $\alpha$ and the mean $\alpha/\beta$ in the gamma prior. The four-loci data from the butterfly species *Heliconius demeter* and *H. eratosignis* are analyzed using the rjMCMC algorithms. Note that both axes are on the logarithmic scale.

prior for the $\theta$s and the prior for $\tau_0$, respectively. Note that the gamma prior $G(\alpha, \beta)$ has mean $E = \alpha/\beta$, mode $(\alpha - 1)/\beta$ (if $\alpha \geqslant 1$), and variance $V = \alpha/\beta^2$. Thus $\alpha = E^2/V$ determines whether the gamma prior is diffuse or informative: $\alpha = 1$ may be considered a diffuse prior, whereas $\alpha = 5$ or 10 highly informative priors. The impact of the prior on $P_2$ is largely determined by two factors: how informative the prior is $(\alpha)$ and whether the prior is in conflict with the data, with the latter being indicated by how close the prior mean or mode is to the maximum likelihood estimate (MLE). From the Bayesian estimates under different priors (see above), the MLEs of parameters under the two-species model may be close to 0.005–0.01 for the $\theta$s and 0.0013 for $\tau_0$. Thus, we may consider the priors to be reasonable if the prior means are in the range 0.001–0.05 for $\theta$s and 0.0001–0.01 for $\tau_0$. For priors in those ranges, $P_2 \approx 1$, consistent with the species status of the two species. Our theory (see Discussion section) predicts that $P_2$ tends to be lower when the prior (in particular, the prior on $\tau_0$) is informative and in conflict with the data. In Figure 9, the smallest $P_2$ was found when the prior is highly informative ($\alpha = 5$ or 10) and the prior mean is orders of magnitude too large or too small for the data.

We then analyzed the data using the $\tau$-threshold method (Yang and Rannala 2010). An ordinary MCMC algorithm was used to generate the posterior distribution of parameters under the two-species model (Rannala and Yang 2003) and the probability that the divergence time $\tau_0$ is greater than a threshold value $\tau_T$ was calculated from the posterior; this was taken as the posterior probability for the two-species model: $P_2 = \Pr\{\tau_0 > \tau_T\}$. For this method, the choice of the threshold value $\tau_T$ and of the prior on $\tau_0$ requires care.

Here we consider two ideas. The first is to try to match up the means and the variances between the two priors on $\tau_0$. As the two prior distributions look very different even with the same mean and variance, they may still produce very different posterior model probabilities. Suppose the gamma prior is $\tau_0 \sim G(\alpha, \beta)$ in the $\tau$-threshold method, whereas the rjMCMC method assigns the prior probability 0.5 to the point value 0 and probability 0.5 to $G(a, b)$. Equating the means and variances between the two priors leads to $\alpha = 2a$, and $\beta = b$ while $\tau_T$ is chosen so that there is 50% prior probability left and right of $\tau_T$, that is, $\tau_T$ is the median of $G(\alpha, \beta)$. With the rjMCMC algorithms, we used the priors $\theta \sim G(2, 500)$ and $\tau_0 \sim G(2, \beta)$, and found $P_2 \approx 1$ for $\beta = 100$, 1000, and 10,000. For the $\tau$-threshold method, the matching priors were $\theta \sim G(2, 500)$ and $\tau_0 \sim G(1, \beta)$, with $\tau_T = \log(2)/\beta$, which gave $P_2 = 0.00$, 0.95, and 1.00 for $\beta = 100$, 1000, and 10,000. Although the two analyses agreed with each other for $\beta$=1000 and 10,000, they were very different for $\beta = 100$.

The second idea is to decide on the threshold $\tau_T$ based on a species definition. Here, we chose $\tau_T = 0.0002$ based on $10^6$ generations of divergence for a butterfly species, with 4–6 generations per year, and a mutation rate of $10^{-9}$ mutations per site per year. We then fixed the shape parameter $\alpha = 1$ for the gamma prior for $\tau_0 \sim G(1, \beta)$ and use $\Pr(\tau_0 > \tau_T) = 0.5$ to derive the scale parameter $\beta$, giving $\beta = \log(2)/\tau_T = 3466$. In other words, $\tau_0 \sim G(1, 3466)$. The posterior probability under this prior was calculated to be $P_2 \approx 1$, with a strong support for the two-species model.

In summary, both the rjMCMC and $\tau$-threshold methods provide strong support for the distinct species status of the two butterfly species. Although both methods may give compatible results for the same data, we

note that the $\tau$-threshold method is very sensitive to the divergence threshold $\tau_T$. We thus suggest that the rjMCMC method of calculating posterior model probabilities should be preferred.

## Discussion

### *Barcoding Dap, Reciprocal Monophyly, and DNA Taxonomy*

DNA barcoding uses a genetic marker (often a single gene) to assign an individual to a particular known species. It has also been suggested that barcoding can be used to identify unknown species based on the expectation that interspecific genetic divergence considerably exceeds intraspecific variation to form a clear "barcode gap." The "10× rule" requires a 10-fold difference in the within- and between-species divergence (Hebert et al. 2004). DNA barcoding has gained popularity and provoked much discussion (e.g., Hebert et al. 2004; Hickerson et al. 2006; Nielsen and Matz 2006; Spooner 2009). In a simulation study under a model of speciation resulting from Dobzhansky–Muller incompatibilities (Dobzhansky 1937), the 10× divergence threshold failed miserably in discovering recently divergent species (Hickerson et al. 2006). Indeed, a cutoff on sequence divergence appears neither necessary nor sufficient for species delimitation. In some of our simulations, the within-species variation and between-species divergence are similar. For example, under the two-species model (Fig. 2a), the within-species divergence is $\theta$, whereas the between-species divergence is $2\tau + \theta_0$. These are 0.01 and 0.012, with only a 20% difference, for the case $\tau = 0.001$ and $\theta = 0.01$ (Fig. 3d–f). Yet, BPP inferred the correct two-species model with probability near 1 when 10–50 loci were used or when large samples (15 or 20 from each species) were taken at a single locus.

We also compared Bayesian species delimitation with the criterion of reciprocal monophyly of gene trees, with one locus used. The parameters are $\theta = 0.01$ or 0.001 and $\tau_0 = 0.01$ or 0.001 under the two-species model, and we considered the following sample configurations: (5, 5), (10, 10), (15, 15), and (20, 20). The results are shown in Table 1 (see also Fig. 3f, i). The proportion of data sets in which the gene tree at one locus shows reciprocal monophyly is calculated by two approaches: (i) using the true simulated gene tree, and (ii) using the estimated gene tree by the UPGMA method. Because errors in gene tree reconstruction tends to destroy reciprocal monophyly, use of the inferred gene trees leads to reduced power compared with use of the true gene tree. When the population is small and divergence is ancient ($\theta = 0.001, \tau_0 = 0.01$), both BPP and reciprocal monophyly have power close to 1. Otherwise, the power for reciprocal monophyly is much lower than for BPP. In particular, with a large population size and recent divergence (i.e., $\theta = 0.01$ and $\tau_0 = 0.001$), the proportion of gene trees showing reciprocal monophyly is nearly 0, whereas BPP still identifies the two species with high posterior probability, reaching 99.2% for the large data set of 20 sequences from each species.

TABLE 1. The power of delimiting two species by BPP and by reciprocal monophyly of gene trees

| Data configuration | $P_2$ (BPP) | Reciprocal monophyly |
|---|---|---|
| $\theta = 0.001, \tau_0 = 0.001$ | | |
| (5, 5) | 0.774 | 0.673 (0.425) |
| (10, 10) | 0.912 | 0.635 (0.403) |
| (15, 15) | 0.957 | 0.577 (0.338) |
| (20, 20) | 0.975 | 0.567 (0.292) |
| $\theta = 0.001, \tau_0 = 0.01$ | | |
| (5, 5) | 0.999 | 1.000 (1.000) |
| (10, 10) | 1.000 | 1.000 (1.000) |
| (15, 15) | 1.000 | 1.000 (1.000) |
| (20, 20) | 1.000 | 1.000 (1.000) |
| $\theta = 0.01, \tau_0 = 0.001$ | | |
| (5, 5) | 0.500 | 0.019 (0.011) |
| (10, 10) | 0.831 | 0.001 (0.000) |
| (15, 15) | 0.965 | 0.000 (0.000) |
| (20, 20) | 0.992 | 0.000 (0.000) |
| $\theta = 0.01, \tau_0 = 0.01$ | | |
| (5, 5) | 0.997 | 0.671 (0.631) |
| (10, 10) | 1.000 | 0.623 (0.582) |
| (15, 15) | 1.000 | 0.596 (0.561) |
| (20, 20) | 1.000 | 0.568 (0.527) |

The two numbers in each cell for reciprocal monophyly are calculated using the true gene tree and the estimated gene tree (in parentheses), respectively. The UPGMA method was used to infer rooted gene trees, with sequence distances calculated under JC69 (Jukes and Cantor 1969) using the programs DNADIST and NEIGHBOR in the PHYLIP package (Felsenstein 2005).

Nonmonophyletic gene trees for well-established species are quite common in real data sets (see Funk and Omland 2003 for a summary based on animal mitochondrial DNA). In general, the criterion of gene tree reciprocal monophyly is too stringent to be useful for species delimitation. The power will be even lower if we require all gene trees at multiple loci to be reciprocally monophyletic.

Another question is whether a single DNA segment is sufficient for species delimitation and whether the current recommendation of sampling 5–10 individuals (Hajibabaei et al. 2007) is adequate. Our results suggest that one single gene locus may indeed contain enough information to delimit species. However, 15 or more individuals from each species seem necessary if the species divergence is recent (e.g., $\tau_0 = 0.001$), whereas five individuals may be enough for identifying well-diverged species. When it is unfeasible to sample multiple individuals, as with rare or protected species, multiple loci should be used for effective species delimitation.

### *Species Delimitation and Statistics*

Although there are fundamental philosophical disagreements between frequentist and Bayesian statistics, the two methodologies most often produce *numerically* similar results when applied to real-world problems. A major exception, however, is the problem of hypothesis testing or model selection. Unfortunately, species delimitation as formulated in Yang and Rannala (2010; see also Carstens and Richards, 2007) is exactly one such

problem, so that the controversies in statistics affect our use and interpretation of BPP. The species-delimitation models are nested statistical hypotheses: the one-species model may be considered the null hypothesis with $\tau_0 = 0$, whereas the two-species model is the alternative with $\tau_0 > 0$. Here, we consider a simple case based on a normal sample to illustrate the differences between frequentist hypothesis testing and Bayesian model selection. The analysis will also provide insights into the effect of the priors for model parameters on the Bayesian inference and explanations for some of the simulation results observed earlier.

Suppose we take an independent sample of size $n$ from the normal distribution $N(\mu, 1)$ with unknown mean and known variance to compare model $H_1$: $\mu = 0$ with model $H_2$: $\mu \neq 0$. The sample mean $\bar{x}$ is the sufficient statistic, with $\bar{x} \sim N(0, 1/n)$ under $H_1$ and $\bar{x} \sim N(\mu, 1/n)$ under $H_2$. The $P$ value for the test of $H_1$ against $H_2$ is $\Phi(-\sqrt{n}|\bar{x}|)$, where $\Phi(\cdot)$ is the cumulative density function (CDF) of the standard normal distribution. Note that this is also the LRT because the likelihood under $H_1$: $\mu = 0$ is

$$L_1 = \frac{1}{\sqrt{2\pi/n}} \exp\left\{-\frac{n}{2}\bar{x}^2\right\}, \quad (1)$$

the (maximized) likelihood under $H_2$: $\mu \neq 0$ is

$$L_2 = L_2(\hat{\mu}) = \frac{1}{\sqrt{2\pi/n}} \exp\left\{-\frac{n}{2}(\bar{x} - \hat{\mu})^2\right\} = \frac{1}{\sqrt{2\pi/n}}, \quad (2)$$

evaluated at the MLE $\hat{\mu} = \bar{x}$, and the LRT statistic is

$$2\Delta\ell = -2 \log \frac{L_1}{L_2} = n\bar{x}^2. \quad (3)$$

Note that if $\sqrt{n}|\bar{x}| \sim N(0, 1)$, then $n\bar{x}^2 \sim \chi^2_1$, so that the $P$ values based on the two test statistics are identical.

In the Bayesian framework, we assign the prior $\pi_1 = \pi_2 = 1/2$ for the two models, and $\mu \sim N(\mu_0, \sigma^2)$ under $H_2$. As $H_1$ does not involve any unknown parameters, the marginal likelihood under $H_1$ is $M_1 = L_1$. The marginal likelihood under $H_2$ is an average of the likelihood $L_2(\mu)$ over the prior on $\mu$:

$$\begin{aligned} M_2 &= E(L_2(\mu)) \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi/n}} \exp\left\{-\frac{n}{2}(\bar{x} - \mu)^2\right\} \\ &\quad \times \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(\mu - \mu_0)^2\right\} \mathrm{d}\mu \\ &= \frac{1}{\sqrt{2\pi(1/n + \sigma^2)}} \exp\left\{-\frac{1}{2(1/n + \sigma^2)}(\bar{x} - \mu_0)^2\right\}. \end{aligned} \quad (4)$$

Thus, the posterior probability for model $H_2$ is

$$\begin{aligned} P_2 &= \frac{\pi_2 M_2}{\pi_1 M_1 + \pi_2 M_2} = \frac{1}{1 + M_1/M_2} \\ &= \frac{1}{1 + \sqrt{1 + n\sigma^2} \exp\left\{\frac{n(\mu_0^2 - 2\mu_0\bar{x} - n\sigma^2\bar{x}^2)}{2(1+n\sigma^2)}\right\}}. \end{aligned} \quad (5)$$

If one uses $\mu_0 = 0$ in the prior, a case considered by Yang (2006: Eq. 5.21 on page 157), we have

$$P_2 = \frac{1}{1 + \sqrt{1 + n\sigma^2} \exp\left\{-\frac{n\bar{x}^2}{2[1+(1/(n\sigma^2))]}\right\}}. \quad (6)$$

Now suppose that in a particular data set $\bar{x}$ is quite different from 0, so that we reject $H_1$ at the significance level $\alpha$, that is, $\sqrt{n}|\bar{x}| = z_{\alpha/2}$, the $\alpha/2$ quantile of the standard normal distribution. However, if $n\sigma^2$ is large, we may have $P_2 \approx 0$. Thus, although the test rejects $H_1$, the Bayesian analysis of the same data strongly supports $H_1$ with $P_1 = 1 - P_2 \approx 1$. This contradiction between methods, known as Lindley's paradox (Lindley 1957; see also Jeffreys 1939), is highly controversial. Nevertheless a few relevant remarks can be made. First, one should exercise caution in applying one's intuition based on hypothesis testing to interpret the results obtained from BPP. Second, compared with the Bayesian analysis, the LRT does not penalize parameter-rich models (such as $H_2$) enough, especially in large data sets (Schwarz 1978). Third, Bayesian model comparison may be very sensitive to priors on parameters that are in one model but not in the other.

Indeed, $P_2$ in equations 5 and 6 can be made as close to 0 as one likes by choosing a very diffuse prior (i.e., by using a large enough $\sigma^2$). For the upper bound on $P_2$, note that the marginal likelihood $M_2$ of equation 4 is no larger than the maximized likelihood $L_2$ of equation 2

$$M_2 = E(L_2(\mu)) \leqslant L_2(\hat{\mu}), \quad (7)$$

and thus

$$P_2 = \frac{M_2}{M_1 + M_2} \leqslant \frac{L_2}{L_1 + L_2}. \quad (8)$$

This upper bound can be achieved if the prior on $\mu$ is very close to the MLE, that is, if $\mu_0 = \bar{x}$ and $\sigma^2 \to 0$. In other words, $P_2$ will be large if the prior is highly concentrated around the MLE and is thus highly consistent with the data. When the LRT is significant at the 5% level (i.e., when $L_2/L_1 = e^{1.92}$), the highest $P_2$ achievable is 0.872. In such a data set, $P_2$ may go from $\sim 0$ to 0.872 by changes to the prior on $\mu$.

In the species delimitation problem, we compare the null one-species model $S_1$: $\tau_0 = 0$ against the alternative two-species model $S_2$: $\tau_0 > 0$. This is noted to have a few extra complications relative to the normal example above. First, both models $S_1$ and $S_2$ have unknown parameters. Second, $\tau_0 = 0$ is at the boundary of the parameter space in $S_2$ as $\tau_0$ is nonnegative. Third, when $\tau_0 = 0$, some parameters in $S_2$ ($\theta_A$ and $\theta_B$) are undefined. The last two complications invalidate the use of the $\chi^2$ distribution for the LRT. However, none of those complications makes a qualitative difference to the Bayesian analysis and the patterns we identified above from the simple example largely apply to the species delimitation problem. For example, the posterior probability for the two-species model $P_2$ will be larger if the priors on

the parameters unique to $H_2$ ($\tau_0$ and the $\theta$s for the modern species) are concentrated around their MLEs, and $P_2$ will be smaller if those priors are highly incompatible with the data. This theory provides an explanation for the results of Figure 9, in which the effect of the priors on $\theta$s and $\tau$ does not have a fixed direction and the lowest $P_2$ is for prior means that are orders of magnitude away from the MLEs.

The sensitivity of posterior probabilities for the species models to the priors on $\theta$ and $\tau_0$ appears to be a feature of the problem. Biologically, it is difficult to specify universal criteria, such as the number of generations, the level of genetic sequence divergence, etc., that can convincingly define species. The controversies surrounding the species concepts will no doubt have an impact on species delimitation using genetic data, as BPP attempts to do.

### *The Utility of Bayesian Species Delimitation*

Compared with traditional taxonomic practices for species delimitation, which may vary widely among taxa, the Bayesian method is arguably more objective as all its model assumptions are explicit and can be tested (Fujita and Leache 2011). An important feature of this method is that it infers species status from a genealogical and population genetic perspective, relaxing the requirement of reciprocal monophyly of gene trees followed in current DNA taxonomy and barcoding practice. It should also be superior to methods that analyze estimated gene trees without accommodating phylogenetic errors (e.g., Knowles and Carstens 2007; O'Meara 2010).

The current implementation of Bayesian species delimitation in BPP is based on the biological species concept, assuming complete cessation of gene flow following species divergence (Yang and Rannala 2010). This simulation study, however, suggests that the behavior of BPP when there is gene flow is consistent with the practice of taxonomists. Low levels of migration, with the expected number of immigrants per generation at $M = Nm < 0.1$, have virtually no impact: the method infers different species despite small amounts of gene flow. This appears to be consistent with biologists' common practice of identifying distinct species despite occasional hybridizations. Although Mayr (1963) initially defined a biological species as "groups of interbreeding natural populations that are reproductively isolated from other such groups," Coyne and Orr (2004, p. 30) revised the definition so that "distinct species are characterized by *substantial but not necessarily complete reproductive isolation*."

When the migration rate is high, with 10 or more migrants per generation, the method infers one species. By any sensible species concept, the two populations should be considered one species at such high levels of hybridization. Our simulation also demonstrates that the method is very unlikely to be misled to infer separate species if samples are taken from distant localities of one species with a wide geographical distribution and experiencing isolation by distance. For the purpose of delimiting species, there does not appear to be a need to explicitly incorporate migration in BPP. However, a model of migration (e.g., Hey 2010) is useful for estimating parameters such as migration rates when such migrations are known to occur.

The Bayesian method of Yang and Rannala (2010) is designed for analyzing multilocus genomic sequences that evolve neutrally. Although protein-coding genes under similar purifying selection in different species may be used in the analysis, perhaps with the different mutation rates among loci accommodated in the model, genes undergoing species-specific selection such as those involved in the establishment of reproductive isolation are not suitable for analysis by this method. Similarly, the method does not take into account whether the lack of gene flow or the low migration rate is due to geographical barriers or to intrinsic reproductive isolation. Two allopatric populations that diverge due to neutral drift without the establishment of reproductive barriers may be inferred to be two species by the method if the divergence time is long enough or if a sufficiently large data set is analyzed. The species status of allopatric populations is often debatable, and we expect this ambiguity to affect the Bayesian analysis. We suggest that Bayesian inference by BPP not be used as the sole criterion for species delimitation, and instead the results from the Bayesian analysis be integrated with other sources of information, such as information on morphological, behavioral, and ecological traits. We note that this ambiguity of interpretation does not exist if sympatric populations are analyzed.

For the present, it is unclear how large the data set has to be for BPP to infer two species even when divergence is relatively recent. Computational problems in the current rjMCMC algorithms make it impossible to analyze very large data sets. It is thus important to improve the algorithms, perhaps by integrating some parameters analytically rather than through the Markov chain (e.g., Hey 2010). Furthermore, Leache and Fujita (2010) have demonstrated that the use of an incorrect guide tree can have adverse effects on the inference, causing BPP to infer multiple species. It is thus important to remove the reliance on the guide tree or to accommodate possible errors in the guide tree topology.

Finally, we hope that the development and application of coalescent-based statistical methods such as BPP may have the effect of prompting taxonomists and speciation biologists to formulate their models and concepts precisely, which may be tested using the ever-increasing genomic sequence data.

## SUPPLEMENTARY MATERIAL

Supplementary material, including data files and/or online-only appendices, can be found at http://www.sysbio.oxfordjournals.org/.

## FUNDING

D.-X.Z. is supported by the Natural Science Foundation of China (grant no. 30730016) and the Ministry of Science and Technology of China (grant no. 2006CB805901). Z.Y. acknowledges the support of Biotechnological and Biological Sciences Research Council (grant no. BB/G006431/1) and of K.C. Wong Education Foundation, Hong Kong.

## ACKNOWLEDGMENTS

We thank two anonymous reviewers and the editors for many constructive comments. We thank Adam Leaché, Bruce Rannala, Monty Slatkin, and Weiwei Zhai for discussions and comments. We gratefully acknowledge extensive discussions with Jim Mallet throughout this project.

## REFERENCES

Ane C., Larget B., Baum D.A., Smith S.D., Rokas A. 2007. Bayesian estimation of concordance among gene trees. Mol. Biol. Evol. 24:412–426.

Bauer A.M., Parham J.F., Brown R.M., Stuart B.L., Grismer L., Papenfuss T.J., Bohme W., Savage J.M., Carranza S., Grismer J.L., Wagnerm P., Schmitz A., Ananjeva N.B., Inger R.F. 2011. Availability of new Bayesian-delimited gecko names and the importance of character-based species descriptions. Proc. R. Soc. Lond. B. Biol. Sci. 278:490–492.

Baum D.A., Shaw K.L. 1995. Genealogical perspectives on the species problem. In: Hoch P.C., Stephenson A.G., editors. Molecular and experimental approaches to plant biosystematics. St Louis (MO): Missouri Botanical Garden. p. 289–303.

Beerli P. 2004. Effect of unsampled populations on the estimation of population sizes and migration rates between sampled populations. Mol. Ecol. 13:827–836.

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.

Carstens B.C., Richards C.L. 2007. Integrating coalescent and ecological niche modeling in comparative phylogeography. Evolution. 61:1439–1454.

Coyne J.A., Orr H.A. 2004. Speciation. Sunderland (MA): Sinauer Associates.

Dasmahapatra K.K., Iamas G., Simpson F., Mallet J. 2010. The anatomy of a 'suture zone' in Amazonian butterflies: a coalescent-based test for vicariant geographic divergence and speciation. Mol. Ecol. 19:4283–4301.

Dobzhansky T.G. 1937. Genetics and the origin of species. New York: Columbia University Press.

Feller W. 1968. An introduction to probability theory and its applications. 3rd ed. New York: Wiley.

Felsenstein J. 1992. Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. Genet. Res. 59:139–147.

Felsenstein J. 2005. Phylip: phylogenetic inference program. Version 3.6. Seattle (WA): University of Washington.

Fujita M.K., Leache 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.

Funk D.J., Omland K.E. 2003. Species-level paraphyly and polyphyly: frequency, causes, and consequences, with insights from animal mitochondrial DNA. Annu. Rev. Ecol. Syst. 34:397–423.

Hajibabaei M., Singer G.A., Hebert P.D., Hickey D.A. 2007. DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. Trends Genet. 23:167–172.

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

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

Hey J. 2010. Isolation with migration models for more than two populations. Mol. Biol. Evol. 27:905–920.

Hickerson M.J., Meyer C.P., Moritz C. 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.

Hudson R.R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics. 18:337–338.

Hudson R.R., Coyne J.A. 2002. Mathematical consequences of the genealogical species concept. Evolution. 56:1557–1565.

Issac N.J.B., Mallet J., Mace G.M. 2004. Taxonomic inflation: its influence on macroecology and conservation. Trends Ecol. Evol. 19:464–469.

Jeffreys H. 1939. Theory of probability. Oxford: Clarendon Press.

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.

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

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

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

Kubatko L.S., Carstens B.C., Knowles L.L. 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics. 25:971–973.

Leache 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.

Lindley D.V. 1957. A statistical paradox. Biometrika. 44:187–192.

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

Liu L., Yu L., Kubatko L., Pearl D.K., Edwards S.V. 2009. Coalescent methods for estimating phylogenetic trees. Mol. Phylogenet. Evol. 53:320–328.

Mayr E. 1963. Animal species and evolution. Cambridge (MA): Belknap Press.

Neigel J.E., Avise J.C. 1986. Phylogenetic relationships of mitochondrial DNA under various demographic models of speciation. In: Karlin S., Nevo E., editors. Evolutionary processes and theory. New York: Academic Press. p. 515–534.

Nielsen R., Matz M. 2006. Statistical approaches for DNA barcoding. Syst. Biol. 55:162–169.

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

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.

Schwarz G. 1978. Estimating the dimension of a model. Ann. Stat. 6:461–464.

Slatkin M. 1991. Inbreeding coefficients and coalescence times. Genet. Res. 58:167–175.

Spooner L.J. 2009. DNA barcoding will frequently fail in complicated groups: a example in wild potatoes. Am. J. Bot. 96:1177–1189.

Strobeck K. 1987. Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. Genetics. 117:149–153.

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

Takahata N. 1983. Gene identity and genetic differentiation of populations in the finite island model. Genetics. 104:497–512.

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

Wakeley J., Aliacar N. 2001. Gene genealogies in a metapopulation. Genetics. 159:893–905.

Wilkinson-Herbots H.M. 2008. The distribution of the coalescence time and the number of pairwise nucleotide differences in the "isolation with migration" model. Theor. Popul. Biol. 73:277–288.

Wu C.I., Ting C.T. 2004. Genes and speciation. Nat. Rev. Genet. 5:114–122.
Yang Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol. Evol. 11:367–372.
Yang Z. 1997. On the estimation of ancestral population sizes. Genet. Res. 69:111–116.
Yang Z. 2002. Likelihood and Bayes estimation of ancestral population sizes in Hominoids using data from multiple loci. Genetics. 162:1811–1823.
Yang Z. 2006. Computational molecular evolution. Oxford: Oxford University Press.
Yang Z. 2010. A likelihood ratio test of speciation with gene flow using genomic sequence data. Genom. Biol. Evol. 2:200–211.
Yang Z., Rannala B. 2010. Bayesian species delimitation using multilocus sequence data. Proc. Natl. Acad. Sci. U.S.A. 107:9264–9269.
Zhang D.X., Hewitt G.M. 2003. Nuclear DNA analyses in genetic studies of populations: practice, problems and prospects. Mol. Ecol. 12:563–584.
Zhou R., Zeng K., Wu W., Chen X., Yang Z., Shi S., Wu C.-I. 2007. Population genetics of speciation in nonmodel organisms: I. ancestral polymorphism in mangroves. Mol. Biol. Evol. 24:2746–2754.

## APPENDIX

### *Proof of Slatkin's (1991) Result Concerning the Effective Migration Rate in the Circular Stepping-stone Model*

*Slatkin's result.*This is stated right above equation 16 in Slatkin (1991): "The average time until two genes $i$ steps apart initially are first found in the same deme is $(d - i)i/2m$." The model is a circular stepping-stone model with $d$ demes. In our simulation, we considered a linear stepping-stone model with four demes: A $\leftrightarrow$ B $\leftrightarrow$ C $\leftrightarrow$ D. If A and D are linked the model will be circular. The linear and circular models have qualitatively the same behavior, but the circular model is slightly easier to analyze. Slatkin's result is that if the migration rate between two adjacent demes in each direction is $m/2$, so that in each generation, a proportion $m$ of alleles in each deme are immigrants, then the expected waiting time for two alleles drawn from two demes $i$ steps apart in the circle to be found in the same deme is

$$T_i = i(d - i)/(2m). \tag{A.1}$$

If the number of demes $d$ is large and $i$ is small, this is nearly linear with $i$, so that the expected waiting time for $i = 3$ is about three times as long as it is for $i = 1$. The reciprocal of the expected waiting time gives the "effective migration rate." Thus, the result can be stated as follows: two alleles from populations three demes apart in the stepping-stone model with migration rate $m$ are as divergent as two alleles taken from two neighboring populations with the migration rate at $m/3$ (instead of $m^3$ as the faulty intuition mentioned in the text has).

Below is a more detailed version of Slatkin's proof. A proof based on difference equations is given by Strobeck (1987, equation 7).

1. Duration of the Gambler's Ruin game. Suppose a gambler has $i$ pounds and bets against a machine which holds $d - i$ pounds. He tosses a fair coin and either wins or loses a pound depending on whether it lands heads or tails. The game ends when the gambler has either 0 or $d$ pounds. The gambler's fortune constitutes a symmetric random walk on the states $0, 1, \ldots, d$, with 0 and $d$ to be the absorbing states. The expected duration of the game is $i(d - i)$. This result is well known in theories of random walks (see, e.g., Feller 1968, equation 3.5). Its common proof is through solving a difference equation, constructed by considering the outcome of the first coin toss:

$$T_i = \frac{1}{2}T_{i-1} + \frac{1}{2}T_{i+1} + 1, \tag{A.2}$$

under the boundary condition $T_0 = T_d = 0$.
2. A slight extension of the above model includes a nonzero probability of no state change. Suppose the coin lands on its edge with probability $1 - c$, and when that happens, the gambler's fortune does not change. Suppose $c$ does not depend on $i$. Then the expected duration of the game is $i(d - i)/c$. If an event occurs with probability $c$, the average time to wait until such an event is $1/c$. Here the expected waiting time until a state change is $1/c$.
3. Waiting time $T_i$ in the circular stepping-stone model. Imagine $d + 1$ demes on a line, but with deme $d+1$ to be deme 1. The distance between two alleles can be $0, 1, \ldots,$ or $d$, with $d$ being the same as 0 or with both 0 and $d$ to mean that the two alleles are in the same deme. Thus the distance between the two alleles form a symmetric random walk on $0, 1, \ldots, d$, with 0 and $d$ to be the absorbing states. When we trace back the genealogy in each generation, the state (the distance between the two alleles) changes by 0, 1, and 2, but changes of 2 can be ignored as they occur with rates of order $m^2$. The probability of change (by +1 or $-1$) is $c = 2m(1 - m) \approx 2m$. Thus, the expected waiting time until absorption or until the two alleles are in the same deme is $i(d - i)/(2m)$.