3-1-2018

Impact of Model Violations on the Inference of Species Boundaries under the Multispecies Coalescent

Anthony J. Barley
University of Hawai'i at Mānoa

Jeremy M. Brown
Louisiana State University

Robert C. Thomson
University of Hawai'i at Mānoa

Recommended Citation

Barley, A., Brown, J., & Thomson, R. (2018). Impact of Model Violations on the Inference of Species Boundaries under the Multispecies Coalescent. *Systematic Biology, 67* (2), 269-284. https://doi.org/10.1093/sysbio/syx073

Syst. Biol. 67(2):269–284, 2018

DOI:10.1093/sysbio/syx073
Advance Access publication September 4, 2017

Impact of Model Violations on the Inference of Species Boundaries Under the Multispecies Coalescent

ANTHONY J. BARLEY[1,*], JEREMY M. BROWN[2], AND ROBERT C. THOMSON[1]

[1]*Department of Biology, University of Hawai'i, 2538 McCarthy Mall, Edmondson Hall 216, Honolulu, HI 96822, USA;* [2]*Department of Biological Sciences and Museum of Natural Science, Louisiana State University, 202 Life Sciences Building, Baton Rouge, LA 70803, USA*
**Correspondence to be sent to: Department of Biology, University of Hawaii, 2538 McCarthy Mall, Edmondson Hall 216, Honolulu, HI 96822, USA; E-mail: ajbarley@hawaii.edu.*

Received 15 March 2017; reviews returned 21 June 2017; accepted 31 August 2017
Associate Editor: Laura Kubatko

Abstract.—The use of genetic data for identifying species-level lineages across the tree of life has received increasing attention in the field of systematics over the past decade. The multispecies coalescent model provides a framework for understanding the process of lineage divergence and has become widely adopted for delimiting species. However, because these studies lack an explicit assessment of model fit, in many cases, the accuracy of the inferred species boundaries are unknown. This is concerning given the large amount of empirical data and theory that highlight the complexity of the speciation process. Here, we seek to fill this gap by using simulation to characterize the sensitivity of inference under the multispecies coalescent (MSC) to several violations of model assumptions thought to be common in empirical data. We also assess the fit of the MSC model to empirical data in the context of species delimitation. Our results show substantial variation in model fit across data sets. Posterior predictive tests find the poorest model performance in data sets that were hypothesized to be impacted by model violations. We also show that while the inferences assuming the MSC are robust to minor model violations, such inferences can be biased under some biologically plausible scenarios. Taken together, these results suggest that researchers can identify individual data sets in which species delimitation under the MSC is likely to be problematic, thereby highlighting the cases where additional lines of evidence to identify species boundaries are particularly important to collect. Our study supports a growing body of work highlighting the importance of model checking in phylogenetics, and the usefulness of tailoring tests of model fit to assess the reliability of particular inferences. [Populations structure, gene flow, demographic changes, posterior prediction, simulation, genetics.]

The multispecies coalescent (MSC) model has become an important and widely used tool in the field of systematics (Degnan and Rosenberg 2009; Edwards 2009). This advance stems from the recognition that gene trees (which represent the evolutionary history of individual alleles) are conceptually distinct from (and can differ from) species trees (which represent the evolutionary relationships among species) (Maddison 1997). Accordingly, the MSC was developed to estimate species tree topologies, while accounting for the coalescent process that can lead to incongruence among gene trees (i.e., incomplete lineage sorting; Liu and Pearl 2007; Heled and Drummond 2010). Additional species tree methods have also been developed that accommodate other sources of gene tree discordance, including gene flow, as well as gene duplication and loss (Gerard et al. 2011; Rasmussen and Kellis 2012; Boussau et al. 2013). These sources of gene tree discordance arise out of distinct biological scenarios and require distinct approaches to account for them in phylogenetic analyses. A mismatch between the biological scenario that has shaped a particular data set and the methods that are used to model that biological scenario can lead to inaccurate and biased inferences (i.e., systematic error). The importance of systematic error has long been recognized, however, the field of phylogenetics has recently experienced renewed interest in methods for assessing the fit of phylogenetic models to empirical data sets (Brown 2014b; Lewis et al. 2014; Doyle et al. 2015; Duchêne et al. 2015). The benefit of these approaches is that they allow researchers to ask, in an absolute sense, whether or not a particular model adequately describes an individual empirical data set. Ideally, by tailoring these tests to assess different aspects of model performance, researchers can identify when particular inferences drawn under the model are likely to be biased by poor fit of the model to the data. For example, previous work has demonstrated poor fit of the MSC to a variety of empirical data sets, and that this poor fit can mislead inference of phylogeny (Reid et al. 2014).

Several recent studies have extended the MSC so that it can be used for inference of species boundaries (Yang and Rannala 2010; Grummer et al. 2014; Jones 2017). The benefits of placing species delimitation in an explicit statistical context have been discussed extensively (Fujita et al. 2012; Carstens et al. 2013). In addition to increasing accuracy, previous studies have argued that coalescent models help to remove some of the perceived "subjectivity" associated with the practice of species delimitation. Because the MSC model links population-level coalescent processes and species-level phylogenetic processes, it provides a natural framework for understanding the process of lineage divergence and speciation. At present, all of the available methods for species delimitation using the MSC assume that lineage sorting, alone, is the source of all gene tree heterogeneity. Other assumptions of the model include no gene flow following speciation, random mating (panmixia) within species, no selection, no recombination within a locus, and no linkage (or free recombination) between loci.

Many aspects of the way speciation occurs in nature, and our conception of species as metapopulation

lineages (de Queiroz 1998), may not always match these assumptions. For example, the assumption of panmixia is more consistent with our conceptual understanding of geographically separated subpopulations of demes, whereas metapopulations frequently exhibit population structuring that reflects demographic history or patterns of isolation by distance (IBD) (Rousset 1997; Hanski and Gaggiotti 2004). Alternatively, an increasing number of studies are identifying divergence with gene flow as a common mode of speciation across the tree of life (Nosil 2008; Payseur and Rieseberg 2016). Many evolutionary radiations are also characterized by strong ecological or sexual selection (e.g., Wagner et al. 2012; Schrider et al. 2016). Although these processes are unaccounted for in the standard MSC model, they influence the distribution of gene trees, and therefore directly influence the information that the MSC uses (potentially impacting our ability to accurately delimit species). Analyses of empirical data suggest that the assumptions of the MSC are not always met in empirical systems, including some systems that are of particular interest in terms of species delimitation (Carstens and Dewey 2010; Barley et al. 2013; Gratton et al. 2016). Although this model has been applied extensively in research aimed at identifying the outcome of the complex process of speciation, comparatively little work has addressed the implications of these simplifying assumptions on the practice of species delimitation, or the performance of various implementations of the MSC. Previous tests of the sensitivity of methods that assume the MSC to violations of model assumptions have found sensitivity to gene flow and population structure under a limited set of demographic scenarios (Zhang et al. 2011; Camargo et al. 2012; Jackson et al. 2016; Sukumaran and Knowles 2017). Other studies have detected poor model fit for some data sets (Reid et al. 2014), however the connection between model fit and violations of specific assumptions has rarely been explored (but see Gruenstaeudl et al. 2016). Finally, the connection between model fit and accuracy of species delimitation has not yet been explored in even a rudimentary sense, and therefore deserves further characterization.

In order to understand whether these concerns are important in practice, we undertook an assessment of model performance. We used two different simulation approaches to evaluate the performance of MSC methods for species delimitation. First, we simulated data under a variety of different demographic scenarios that depart from the assumptions of the MSC (but may be common in empirical systems) and assessed the performance of inference under the model in these circumstances (Fig. 1). These departures included population structure, isolation by distance, population size changes within species, and gene flow between species. These simulations allowed us to characterize the sensitivity of inference under the model to violations of its assumptions. In doing so, we are knowingly misusing the model by pushing it outside the boundaries of what it is intended to do. We take this approach in order to mimic the ways in which these methods are actually used in practice, as researchers frequently do not know the extent to which the assumptions of the model are violated in any particular empirical system. An important consideration for species delimitation is that different violations of model assumptions could impact inference in fundamentally different ways. For example, population structure and IBD could potentially lead to inference based on the MSC to incorrectly identify multiple lineages, whereas gene flow between species may cause MSC inference to incorrectly identify a single species when in fact multiple exist. However, we recognize that speciation is a continuous process, and there are not specific thresholds of isolation or migration that distinguish species-level lineages. Therefore, we focused on understanding the behavior of these methods given different biological scenarios and leave it up to individual researchers to decide if the method is accurately identifying the number of species and the associated uncertainty.

We also used posterior predictive simulation to assess the fit of the MSC to several exemplar empirical data sets, including examples that are expected to match the model well and those where a more complex speciation process is thought to have occurred. We took an "inference-based" approach to assessing model fit (Brown 2014a), since we were primarily interested in determining whether model violations affect the species delimitation (and the other model parameter estimates) that are inferred under the MSC. Finally, in order to assess the sensitivity of the posterior predictive tests to violations of model assumptions, we performed posterior prediction on several exemplar simulated data sets. Our motivation was to develop an understanding of how accurate estimates of species boundaries under the MSC are likely to be, and identify if certain aspects of the model are in need of elaboration. In exploring these issues, we develop approaches for model assessment that should be useful for other researchers interested in assessing model fit for their data.

Methods

Model Sensitivity

For the sensitivity simulation tests we evaluated the performance of two different implementations of the MSC: Bayesian Phylogenetics and Phylogeography (BPP; Yang and Rannala 2010) and STACEY (Jones 2017) in BEAST2 (Bouckaert et al. 2014). Whereas BPP employs reversible-jump Markov chain Monte Carlo (rjMCMC) to directly sample the posterior distribution of species trees and species delimitations, STACEY uses an alternative parameterization on node heights in the species tree known as the "birth-death-collapse" model that allows for species delimitation to be performed. By including a spike in the prior probability density on node heights near 0, populations can be lumped into a single species when the node height estimates are very small.

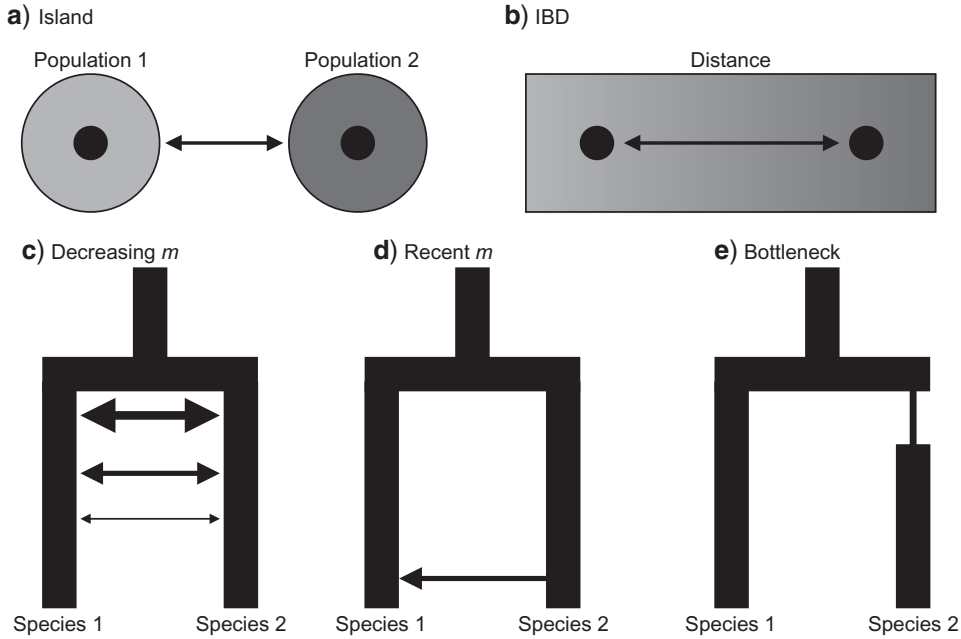

FIGURE 1. Illustration of the simulation scenarios used to test for sensitivity of inference under the MSC to model violations; arrows indicate migration. a) An island model of population structure in which two populations are connected by migration. b) IBD in a single, continuous population sampled at two disjunct geographic locations. c) Species divergence with gene flow that decreases through time. d) Recent migration (or secondary contact) between two species following a period of isolation. e) A population bottleneck experienced by one species following lineage separation.

We evaluated the effect of population structure within species on the inference of species boundaries under the MSC by simulating data under an island population model (Fig. 1a). We simulated two populations connected by varying levels of symmetrical migration (Nm = 0.1, 1, 5, 10, 20) to determine the sensitivity of inference under the MSC to population structure. We simulated gene trees under the coalescent using *ms* (Hudson 2002), and then simulated sequences from the genealogies using seq-gen (Rambaut and Grassly 1997). Each data set consisted of 10 sequences for 20 loci that were 1000 base pairs long and evolved under a Jukes–Cantor (JC69) substitution model (Jukes and Cantor 1969). Values of θ per site in natural populations have been shown to vary between 0.0005 and 0.1 (Rannala and Yang 2003; Zhang and Hewitt 2003; Carling and Brumfield 2007; Zhou et al. 2007). Therefore, we simulated data sets under several different values across this range (0.0005, 0.001, 0.01, 0.1). We simulated 100 data sets for each combination of parameters and analyzed them using both BPP and STACEY.

We evaluated the impact that model violations stemming from IBD within a species may have by simulating a population under IBD using IBDSim and then sampling individuals from across the continuous range (Fig. 1b; Leblois et al. 2009). For each simulation, we drew two separate samples from one of three configurations : 1) a single area at the center of the matrix, 2) two distinct, but adjacent areas near the center, or 3) from opposite ends of the matrix (see Supplementary Material available on Dryad at http://dx.doi.org/10.5061/dryad.h6s2k for exact sampling points). We drew these samples from a single pseudocontinuous population, constant across space and time, and the datasets were identical in size to those described in the population structure section above. We assumed a JC69 substitution model and simulated data on a 5000 × 5000 node lattice with absorbing boundaries, a stepping stone dispersal distribution with an emigration rate of 1/2, and used mutation rates of 1×10^{-7} or 1×10^{-8}. We simulated 100 data sets for each scenario.

For analyses of the island and IBD simulations, we performed species delimitation and species tree inference using BPP v3.1, with the default species tree model prior, setting the mean of the gamma distribution for the prior on θ equal to the value under which the data were simulated, and setting the prior on τ equal to G(2,1000), which is representative of a recent divergence between populations and should result in a more conservative species delimitation estimate (Yang and Rannala 2010). Each analysis was run for 500,000 iterations, sampling every 500 iterations after discarding the first 50,000 iterations as burnin. For STACEY, we performed species tree inference in BEAST v2.4.0. We analyzed the simulated data sets using default priors, except where changes were suggested by the software documentation, including assigning an exponential distribution with a mean of 0.1 to the scaling factor for the population sizes, a lognormal distribution with a mean of 5 and a standard deviation of 2 to the growth rate prior, and a uniform distribution from 0 to 1 for the collapse weight. Each data set was partitioned by gene, with independent strict clock models and rates drawn

from a lognormal prior distribution with a mean of 0 and a standard deviation of 1. We ran each analysis for 50 million generations, sampling every 5000 generations. We analyzed the posterior distribution of species delimitations using SpeciesDelimitationAnalyzer v1.8.0, with a burnin of 5000 and a collapse height of 0.0001. Given the simple nature of estimating a two-species phylogeny, we did not expect mixing/convergence problems in our analyses. We confirmed this in BEAST by randomly choosing 10 analyses for each simulation scenario and manually checking each for convergence using Tracer v1.6 (Rambaut et al. 2014). The BPP documentation suggests running multiple analyses to ensure consistency among results as a check for convergence, which we have done here by performing each of the simulations twice and ensuring the results were qualitatively similar.

We also assessed the performance of inference under the MSC in scenarios in which gene flow had occurred between species descended from a common ancestor or population size had fluctuated within one of the descendant species. The impact of gene flow on species delimitation was assessed under: 1) a partial isolation model where gene flow persists through the time of lineage divergence (or speciation), and decreases gradually through time (Fig. 1c), and 2) a recent admixture model that would correspond to two species that came into secondary contact and experienced introgression (Fig. 1d). We evaluated the effect of population size changes on species delimitation by simulating data under a population bottleneck scenario, where one species experienced a reduction in population size immediately following lineage splitting (Fig. 1e). This scenario could represent speciation by dispersal due to the founder effect, as has been suggested to occur in many island systems. Because these scenarios encompass a huge area of potential parameter space, exhaustive simulations are impractical. Instead, we explored several different parameter values that are biologically realistic and likely to highlight areas of interest.

Data sets for these three scenarios were simulated using CoMuS v2.0 (Papadantonakis et al. 2016) and were identical in size to those described above. All simulations were performed using a two species guide tree with either a relatively old species divergence time of 0.01 or a more recent species divergence time of 0.001 (phylogenetic units), and assuming a JC69 substitution model. The partial isolation (or speciation with gene flow) model is described further in (Papadantonakis et al., 2016), but in each case the migration rate at the time of speciation was Nm=20, which decreased gradually through time until some point in the past (which varied among simulations—see Results section for details). In the recent gene flow or secondary contact scenario, we simulated a population joining event where, at a recent time in the past (0.009 or 0.0009 phylogenetic units, depending on the tree length), one population of species 1 (which varied in size) merged with species 2. For the population size change simulation, we varied the magnitude of the population size change and assumed a bottleneck length of 0.002 phylogenetic units (or 0.0002 phylogenetic units, depending on the tree length) before the population size returned to normal. The simulated data sets were analyzed in BPP and STACEY as described above, except with the mean of the gamma distribution for the priors on θ and τ in BPP equal to the values under which the data was simulated. For each scenario, we assessed the accuracy of the inferred species delimitation and the estimated divergence time between species.

TABLE 1. Results of the IBD simulations

Sampling scheme	Mutation rate	BPP	STACEY
Single	1×10^{-7}	0.80	1.0
Single	1×10^{-8}	1.0	1.0
Adjacent	1×10^{-7}	0.61	1.0
Adjacent	1×10^{-8}	0.80	1.0
Separated	1×10^{-7}	0.0	0.98
Separated	1×10^{-8}	0.0	0.98

Notes: Data were simulated for a single, continuous population, and then individuals were sampled in different spatial locations. In the "single" sampling scheme, all individuals were sampled from a single location. In the "adjacent" scheme, individuals were sampled from two distinct, but directly adjacent locations. In the "separated" scheme, individuals were sampled from two locations on opposite ends of the simulation region. Results are shown for simulations done under two different mutation rates. Species delimitation was performed on the simulated data sets in BPP and STACEY and the mean posterior probability for the one species model (vs. a two species model) is shown across 100 replicates.

Posterior Predictive Simulation

For the model fit tests, we used the MSC implementation available in BPP. Because this is the only implementation that directly samples the posterior distribution of species trees and species delimitations (as well as the other parameters of the MSC), it is an ideal framework for using posterior predictive simulation to assess model fit from the perspective of species delimitation. We gathered a set of empirical genetic data sets previously used for species delimitation (Supplementary Material Table 1 available on Dryad). These included the fence lizard (*Sceloporus*) data set from (Yang and Rannala, 2010), the southern cavefish (*Typhlichthys*) and coast horned lizard (*Phrynosoma*) data sets from (Yang and Rannala, 2014), and the brown frog (*Rana*) example from (Yang, 2015). We included several additional data sets from previous studies in which authors suggested that violations of model assumptions were potentially impacting inferences from coalescent models. These data sets included the spruce-fir moss spider (*Microhexura*) data set from (Hedin et al., 2015) and the sun skink (*Eutropis*) data set from (Barley et al., 2013), which cited concerns about population structure. We also included the little brown bat (*Myotis*) data set from (Carstens and Dewey, 2010), who cited gene flow as a potential problem (though this study used an alternative coalescent approach for delimiting species).

Finally, we analyzed the human (*Homo sapiens*) data set from (Jackson et al., 2016), who found that BPP delimited multiple species using the rjMCMC algorithm. For each data set, we first inferred species trees and delimitations using BPP (Yang and Rannala 2014). We used the same parameter settings that were employed in each of the original studies (except for the bat data set which is from a study that did not use BPP). We also removed the *S. woodi* sample from the fence lizard data set because BPP cannot estimate θ when only one sequence is available from a population/species (Yang 2015). We also used the automatic adjustment option for the MCMC step lengths to help ensure proper mixing and estimated the species tree topology in our analyses rather than fixing it using a "guide tree." In some of the original studies, the authors explored the impact of different combinations of priors in their analyses. We repeated these tests and found similar results in all cases. For clarity of presentation, here we focus on a single combination of priors for each data set (choosing the "preferred" combination if one was identified by the authors). Each analysis was run for 500,000 iterations, sampling every 50 iterations after discarding the first 50,000 iterations as burnin. We ran each analysis twice using different starting trees and checked for similarity between runs and the results from the original publications to help ensure convergence and stability of all analyses.

We performed posterior predictive simulation using custom python scripts and MCcoal (Yang and Rannala 2014). For each data set, we randomly sampled one hundred species trees from the post-burnin posterior distribution of each analysis. We then used MCcoal to simulate gene trees from the species trees under the MSC and sequence data on the gene trees, resulting in data sets with the same numbers of individuals, genes, and gene lengths as the empirical data set. We performed species tree inference and species delimitation using BPP for each of these simulated data sets under identical settings as the initial empirical analyses. In preliminary analyses, we allowed each individual in the data set to potentially be considered a distinct species. However, this approach was not computationally feasible given the size of the data sets. Therefore, the number of species under which the data was simulated was set as the maximum number of potential species, with individual sequences being correctly assigned to potential species in the BPP analyses based on the simulation conditions. Because we were interested in determining the impact of poor model fit on species delimitation itself, we used the number of species as an inference-based test statistic, which we previously found to be useful when assessing model fit in DNA barcoding (Barley and Thomson 2016). We also compared the posterior distributions of divergence times and θ from the empirical and simulated data sets. To quantitatively compare these distributions, we used a variety of different values as test statistics to identify those that are the most useful, including: the mean, median, standard deviation, minimum, maximum, quartiles, skewness, and kurtosis. For each test statistic, we calculated the two-tailed posterior predictive *P*-value (Brown 2014a) (which are distinct from frequentist *P*-values and tend to cluster near 1.0 for two-tailed tests (Gelman et al. 2013). We also calculated the posterior predictive effect size as the absolute value of the difference between the empirical test statistic value and the mean of the test statistic values for the posterior predictive distribution, divided by the standard deviation (Doyle et al., 2015).

We used the same posterior predictive framework to analyze several exemplar simulated data sets from the different biological scenarios described above, in order to assess the power of the posterior predictive tests to detect violations of different model assumptions. Results of the empirical posterior predictive analyses suggested that the test statistics based on the number of species might suffer from a lack of power when the posterior contains only a small number of discrete values (see Discussion). Because this situation would necessarily be true for our simulations [all simulations involved 2 species (or populations)], we performed these analyses only for the gene flow (Fig. 1c,d) and bottleneck (Fig. 1e) scenarios, and assessed the sensitivity of the divergence time and θ test statistics.

Results

Model Sensitivity

The island population model simulations showed sensitivity of both implementations of the MSC to population structure that was dependent on θ (Fig. 2). Unless two populations exchange multiple migrants per generation, both methods often identified the two-species model as preferred, with the one-species model generally favored as migration rates increased. Our results also show, however, that even under high migration rates, both methods favored a two-species model at higher values of θ (with the STACEY implementation being less sensitive). Results of our IBD simulations suggested that IBD can cause BPP to prefer a two-species model over a single-species model (Table 1). STACEY, however, did not show this same sensitivity, always preferring the single-species model in the set of scenarios that we investigated.

Under demographic scenarios where species descended recently from a common ancestor and violations of model assumptions were minor, both methods performed well in identifying the correct number of species, and in estimating divergence times (Figs. 3–5, Tables 2–4). As model violations became more extreme, the methods began to show sensitivity, though again the sensitivity varied between implementations (and to a lesser extent was exacerbated by smaller values of θ). Under a partial isolation model, divergence times in both STACEY and BPP became increasingly biased towards the present as the length of time the two species experienced migration increased (Fig. 3). BPP always preferred the two-species model regardless of the contact time, as did STACEY except when complete isolation had occurred relatively recently (Table 2).

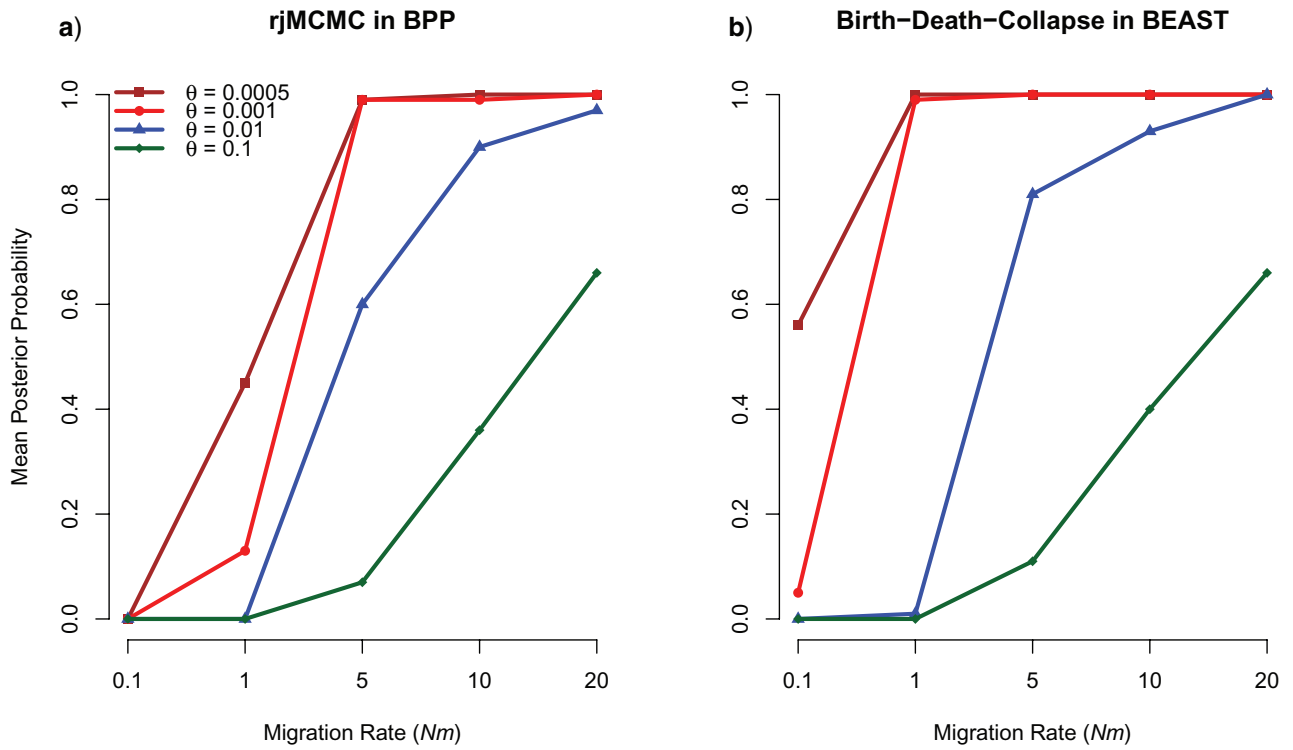

FIGURE 2. Results of the population structure simulation analyses. Data were simulated for two populations connected by varying levels of symmetrical migration and species delimitation was performed using a) BPP and b) STACEY. Each point represents the mean posterior probability for 1 species model across 100 replicates under a wide range of parameter values for θ per site and the migration rate (expressed as *Nm*).

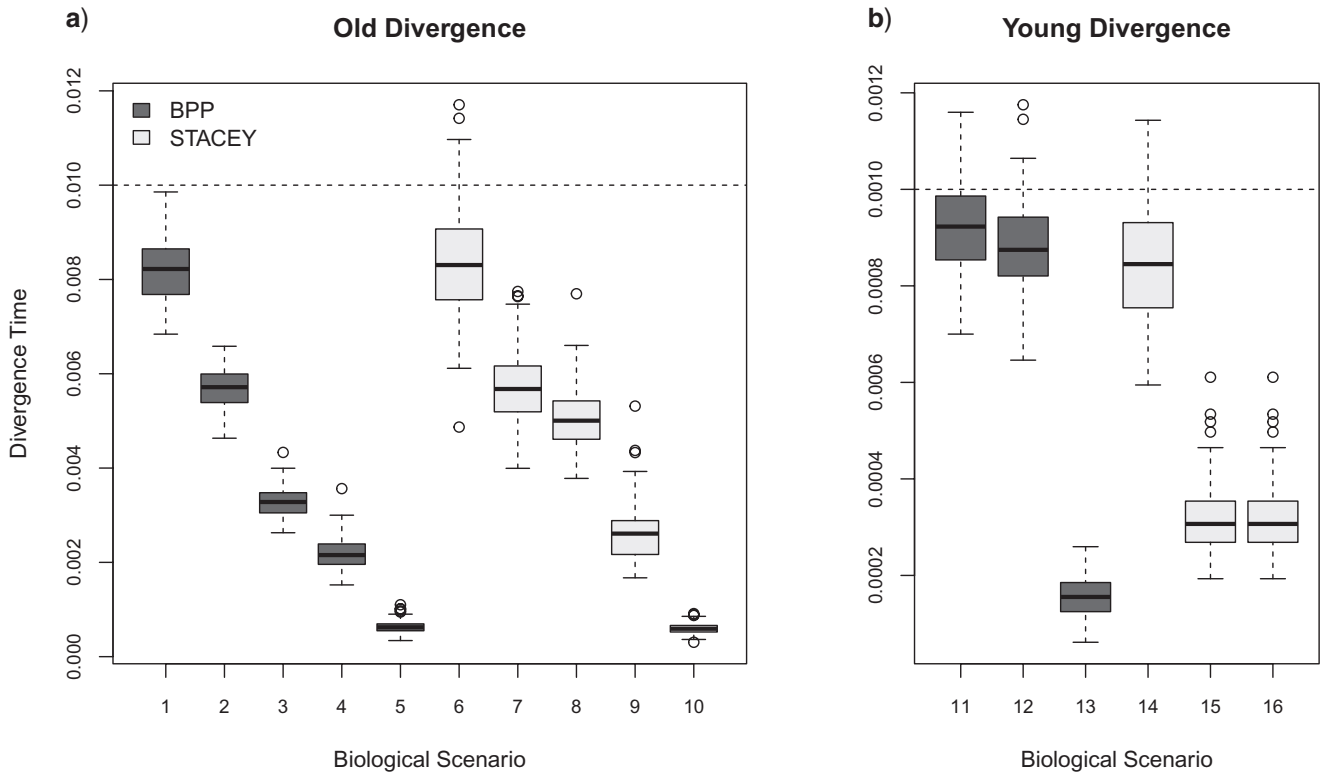

FIGURE 3. Box plots showing mean divergence time estimates (across 100 replicates) for speciation with gene flow simulations analyzed in BPP and STACEY. a) Simulations under a relatively old species divergence time. b) Simulations under a relatively recent speciation time. Scenario numbers correspond to those in Table 2, which shows the simulation parameters. The dotted lines represent the true divergence times.

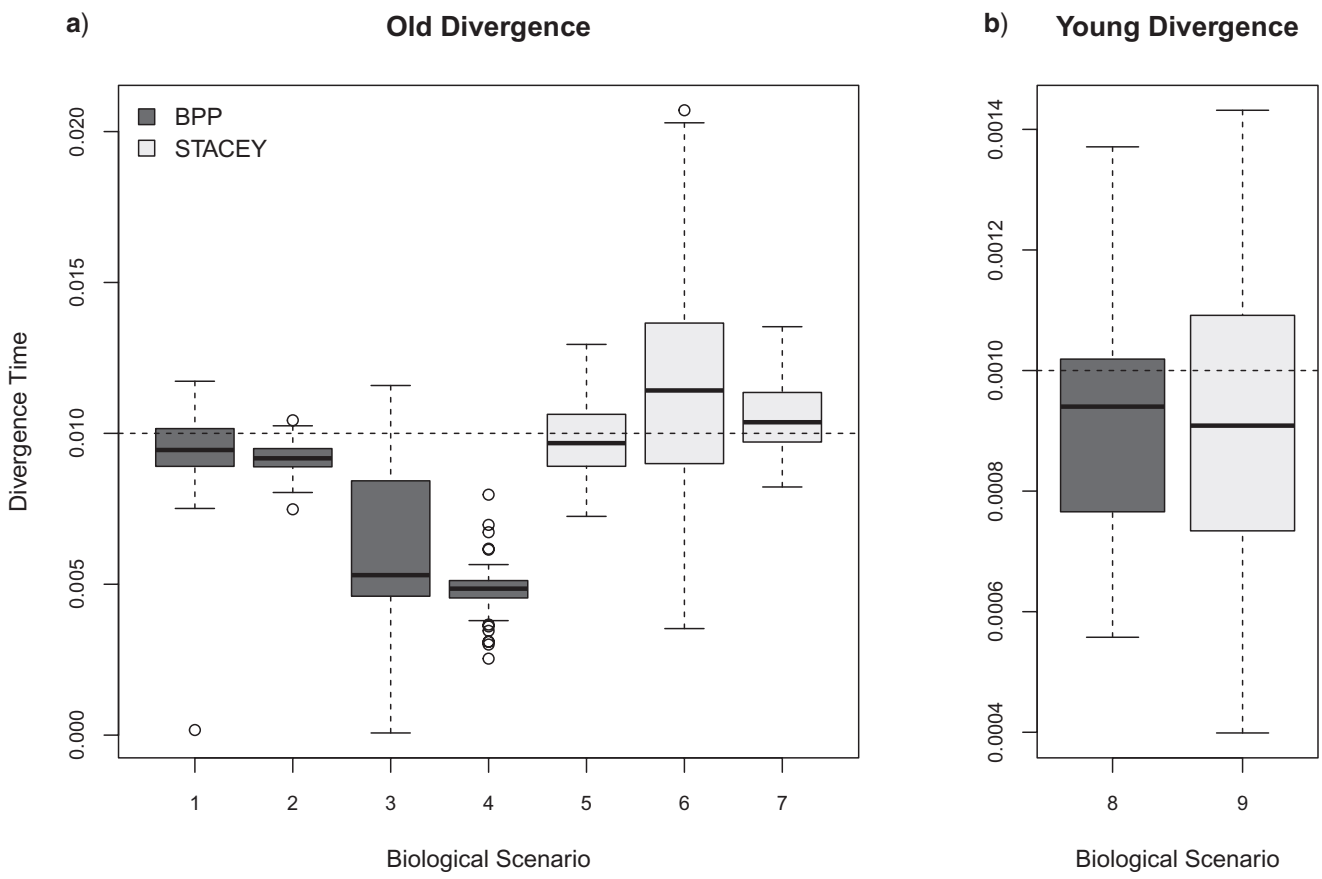

FIGURE 4. Box plots showing mean divergence time estimates (across 100 replicates) for bottleneck simulations analyzed in BPP and STACEY. a) Simulations under a relatively old species divergence time. b) Simulations under a relatively recent speciation time. Scenario numbers correspond to those in Table 3, which shows the simulation parameters. The dotted lines represent the true divergence times.

When a large population bottleneck had occurred in the past for one species, divergence times in BPP were biased towards the present in some circumstances (Fig. 4). By contrast, divergence time estimates in STACEY were not significantly biased by population size changes in any of the scenarios we examined (Fig. 4). Both methods consistently preferred the two-species model, regardless of the bottleneck magnitude (Table 3). Divergence time estimates were quite biased towards the present for both methods under the secondary contact scenario (Fig. 5). BPP always preferred the two-species model regardless of the amount of migration that occurred or the time of lineage divergence in our simulations (Table 4). As the number of migrants from species 1 into species 2 increased, however, the STACEY model increasingly identified the one species model as the preferred model (Table 4).

Posterior Predictive Simulation

Model performance varied widely across empirical data sets and test statistics (Table 5). Since model fit is inherently a continuum, we view these values as relative measures of model performance, rather than outcomes of hypothesis tests (*sensu* Gelman et al. 2013). Therefore, we are more interested in comparing the relative magnitudes of *P*-values and effect sizes to understand if certain inferences should be treated with caution. Researchers generally will differ in the strength of evidence they require to accept alternative interpretations of their results, which might also depend on other knowledge they have about the particular biological system being investigated.

The posterior mean and median number of species were generally similar between empirical and posterior predictive data sets. The largest effect size for the mean number of species (1.30) was in the spider data set. Here, seven species had the highest posterior probability in the empirical analysis, which had a probability that was nearly twice as large as that for the number of species with the next highest probability. In the analyses of the predictive data sets, six species usually had the highest posterior probability, which, on average was also twice as large as that for the number of species with the next highest probability. In some cases, we saw large effect sizes using the minimum, maximum, and standard deviation as test statistics (Supplementary Material Table 2 available on Dryad shows the values of these summary statistics for the empirical and posterior predictive data sets). In these instances, test statistic

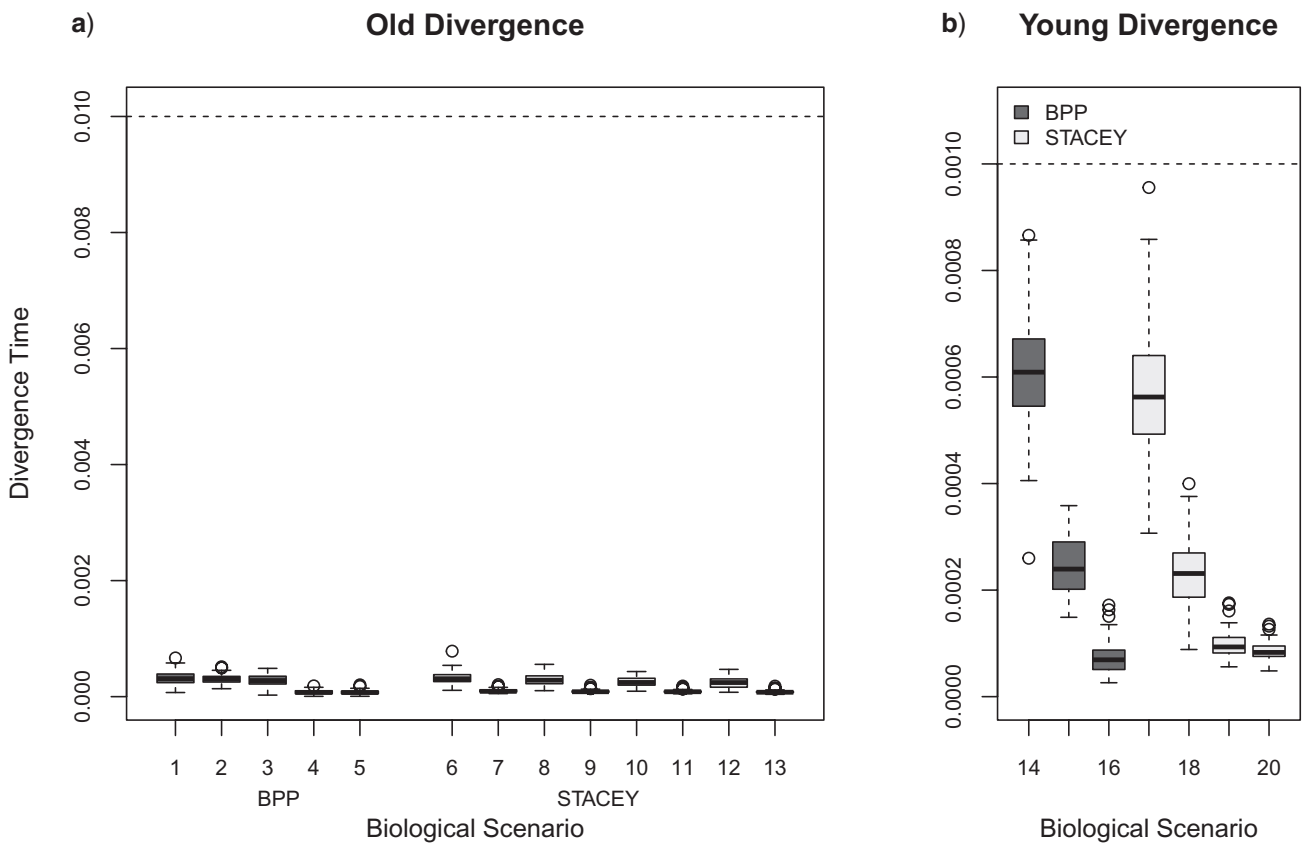

FIGURE 5. Box plots showing mean divergence time estimates (across 100 replicates) for secondary contact simulations analyzed in BPP and STACEY. a) Simulations under a relatively old species divergence time. b) Simulations under a relatively recent speciation time. Scenario numbers correspond to those in Table 4, which shows the simulation parameters. The dotted lines represent the true divergence times.

values were generally larger in the empirical analysis than in analyses of the posterior predictive data sets, indicating that the inferred number of species, or the uncertainty in this estimate, was larger than expected under the assumed model. The largest effect size for the standard deviation of the number of species occurred in the spider data set (2.08). Here, there was a larger range of values (5–8) sampled across all analyses of the predictive data sets than in the empirical analysis (6–8). However, the majority of the probability was usually concentrated over two values (6 and 7) in the analyses of the predictive data sets, whereas it was spread more evenly across the three values in the empirical analysis.

Despite the presence of gene flow among lineages (Carstens and Dewey 2010), we found all 11 species/subspecies in the bat data set to be strongly supported as distinct. Several test statistics indicated poor model fit, and this is also in line with our simulation results that show limited sensitivity of species delimitation in BPP to gene flow between species. As in (Jackson et al., 2016) we found that BPP delimited multiple species of humans. Several test statistics based on the posterior distribution of species number had large posterior predictive effect sizes, and many analyses of the posterior predictive data sets found at least some probability for a single species model (unlike the analysis of the empirical data set).

Of the three test statistic types, the divergence times showed the least evidence of poor model fit (Table 5). Most of the *P*-values suggested good model fit, though several summary statistics had large effect sizes in several data sets (including the bat data set, which we expected to perform poorly in terms of divergence time estimation). The poorest model fit for the divergence time test statistics was seen in the spider data set, where the effect size for the difference in the mean divergence time was 1.12. From a biological perspective, this meant that the mean divergence time was, on average, ~30% older in the analyses of the predictive data sets than in the empirical analysis. When using test statistics based on θ, model performance varied substantially across data sets (Table 5), with the sun skink and cavefish showing the poorest fit. The effect sizes for the mean test statistic were 5.98 and 2.06 for the cavefish and skinks, respectively. Assuming a per generation mutation rate of 1×10^{-8}, this would equate to differences in the mean effective population size estimates of 45,000 (cavefish) and 192,500 (skinks) between empirical analyses and the analyses of the predictive data sets. The sun skink, cavefish, spider, bat, and to a lesser extent horned lizard data sets all showed substantial differences in several test statistic

TABLE 2. Results of speciation with gene flow simulations

Model	θ	Tree length	Contact Time	Posterior Probability
BPP				
Scenario 1	10	0.01	0.25	1.0
Scenario 2	10	0.01	0.5	1.0
Scenario 3	10	0.01	0.75	1.0
Scenario 4	1	0.01	0.75	1.0
Scenario 5	1	0.01	0.95	1.0
Scenario 11	10	0.001	0.5	1.0
Scenario 12	10	0.001	0.75	1.0
Scenario 13	1	0.001	0.95	1.0
STACEY				
Scenario 6	10	0.01	0.25	1.0
Scenario 7	10	0.01	0.5	1.0
Scenario 8	1	0.01	0.5	1.0
Scenario 9	1	0.01	0.75	1.0
Scenario 10	10	0.01	0.95	0.99
Scenario 14	10	0.001	0.75	1.0
Scenario 15	1	0.001	0.75	0.99
Scenario 16	1	0.001	0.95	0.34

Notes: Data were simulated for two species experiencing decreasing rates of gene flow through time after divergence. The contact time refers to the proportion of the total simulation time the two species experienced some migration. Species delimitation was performed in BPP and STACEY and the mean posterior probability for a two species model (vs. a one species model) across 100 replicates is shown. Tree length refers to the length of species tree under which data was simulated (or the true divergence time, expressed as expected substitutions per site). θ is the population mutation rate (expressed as $4N_e\mu$). Refer to Figure 3 for divergence time results.

TABLE 3. Results of the bottleneck simulations

Model	θ	Tree length	Size reduction	Posterior probability
BPP				
Scenario 1	10	0.01	0.01	1.0
Scenario 2	1	0.01	0.01	1.0
Scenario 3	10	0.01	0.001	1.0
Scenario 4	1	0.01	0.001	1.0
Scenario 8	1	0.001	0.001	1.0
STACEY				
Scenario 5	1	0.01	0.01	1.0
Scenario 6	10	0.01	0.001	1.0
Scenario 7	1	0.01	0.001	1.0
Scenario 9	1	0.001	0.001	1.0

Notes: Data was simulated for two species, one of which experienced a reduction in population size immediately following lineage splitting that extended for 20% of the total simulation time. Size reduction refers to the factor by which the population size was reduced during the bottleneck. Species delimitation was performed in BPP and STACEY and the mean posterior probability for a two species model across 100 replicates is shown. Tree length refers to the length of the species tree under which data was simulated (or the true divergence time expressed as expected substitutions per site). θ is the population mutation rate (expressed as $4N_e\mu$). Refer to Figure 4 for divergence time results.

values between empirical and posterior predictive data sets. The fence lizard and brown frog data sets had large *P*-values and small effect sizes for virtually all summary statistics.

Posterior predictive analyses of the simulated data sets were generally in line with our expectations and

TABLE 4. Results of the secondary contact simulations

Model	θ	Tree length	Proportion	Posterior probability
BPP				
Scenario 1	10	0.01	0.09	0.99
Scenario 2	10	0.01	0.17	0.99
Scenario 3	10	0.01	0.23	0.99
Scenario 4	1	0.01	0.23	1.0
Scenario 5	1	0.01	0.29	1.0
Scenario 14	10	0.001	0.09	1.0
Scenario 15	1	0.001	0.09	1.0
Scenario 16	1	0.001	0.33	1.0
STACEY				
Scenario 6	10	0.01	0.09	0.95
Scenario 7	1	0.01	0.09	0.27
Scenario 8	10	0.01	0.17	0.90
Scenario 9	1	0.01	0.17	0.19
Scenario 10	10	0.01	0.23	0.88
Scenario 11	1	0.01	0.23	0.16
Scenario 12	10	0.01	0.29	0.81
Scenario 13	1	0.01	0.29	0.14
Scenario 17	10	0.001	0.09	0.99
Scenario 18	1	0.001	0.09	0.90
Scenario 19	1	0.001	0.23	0.25
Scenario 20	1	0.001	0.29	0.14

Notes: Data were simulated for two species that experienced a population joining event in the recent past (in which one population of species 1 merged with species 2). Proportion refers to the proportion of the population sampled in species 2 composed of migrants following the population joining event. Species delimitation was performed in BPP and STACEY and the mean posterior probability for a two species model (vs. a one species model) across 100 replicates is shown. Tree length refers to the length of the species tree under which the data was simulated (or the true divergence time expressed as expected substitutions per site). θ is the population mutation rate (expressed as $4N_e\mu$). Refer to Figure 5 for divergence time results.

our analyses of the empirical data sets: simulated data sets from scenarios that had stronger violations of model assumptions (and more biased parameter estimates) tended to have test statistics with more extreme *P*-values and larger effect sizes (Table 6). This varied across test statistics and simulation scenarios, however, and suggests that these test statistics may only detect poor model fit if model violations are quite strong. Additionally, however, many of the largest effect sizes from analyses of the empirical data sets were larger than any of those observed in the analyses of the simulated data sets (Table 6), despite the simulation conditions being sufficient to generate substantial inference problems. In combination, these results highlight both the conservative nature of these tests, and the strength of the model violations that seem to affect these empirical data sets.

DISCUSSION

The MSC offers a biologically motivated framework for extracting information about lineage divergence from genetic data (Rannala and Yang 2003), which has been recognized in a number of studies that developed methods for species delimitation under the MSC

TABLE 5. Results of model performance assessments for number of species, divergence time, and θ test statistics across eight empirical data sets analyzed in BPP

	Sun skinks		Cavefish		Fence lizards		Horned lizards		Brown frog		Spiders		Bats		Humans	
	P	ES	*P*	ES	*P*	ES	*P*	ES	*P*	ES	*P*	ES	*P*	ES	*P*	ES
Species																
Mean	0.32	0.25	0.06	0.69	—	—	0.82	0.54	0.70	0.32	0.14	1.30	0.66	0.26	0.62	0.28
Median	1.0	0.29	1.0	0.04	—	—	1.0	0.83	1.0	0.27	0.74	1.10	-	-	1.0	0.55
Standard deviation	0.20	1.11	0.14	1.82	—	—	0.68	0.87	0.60	0.41	0.0	2.08	0.66	0.02	0.16	1.47
Minimum	0.78	1.25	0.94	0.85	—	—	1.0	0.27	1.0	0.42	1.0	0.26	0.80	1.22	0.48	1.07
0.25 Quantile	1.0	0.29	1.0	0.00	—	—	0.90	1.08	1.0	0.50	0.60	1.34	-	-	1.0	0.62
0.75 Quantile	1.0	0.29	1.0	0.10	—	—	1.0	0.69	1.0	0.25	0.10	2.43	-	-	1.0	0.53
Maximum	1.0	0.25	0.38	2.06	—	—	1.0	0.50	1.0	0.23	0.32	1.72	-	-	1.0	0.50
Skewness	0.46	0.37	0.26	0.69	—	—	0.80	0.24	0.72	0.15	0.52	0.08	0.14	1.83	0.32	0.42
Kurtosis	0.46	0.14	0.94	0.45	—	—	0.96	0.37	0.72	0.34	0.96	0.23	0.14	1.64	0.36	0.07
Divergence times																
Mean	0.92	0.10	0.76	0.27	0.76	0.29	0.68	0.46	0.98	0.24	0.28	1.12	0.68	0.31	0.96	0.30
Median	0.54	0.48	0.54	0.57	0.96	0.15	0.72	0.35	0.94	0.14	0.36	0.85	0.72	0.37	0.86	0.30
Standard deviation	0.96	0.13	0.80	0.29	0.74	0.30	0.84	0.31	1.0	0.16	0.46	0.64	0.32	1.02	0.98	0.28
0.025 Quantile	0.64	0.48	0.22	0.92	0.60	0.72	0.55	0.31	0.62	0.44	0.22	1.05	0.18	1.23	0.94	0.16
0.25 Quantile	0.96	0.12	0.80	0.21	0.84	0.10	0.66	0.56	0.94	0.28	0.08	1.73	0.88	0.21	0.98	0.23
0.75 Quantile	0.90	0.22	0.90	0.18	0.70	0.42	0.62	0.46	0.98	0.22	0.38	0.81	0.86	0.09	0.90	0.35
0.975 Quantile	0.94	0.14	0.68	0.36	0.70	0.42	0.84	0.27	0.94	0.22	0.28	0.94	0.34	0.91	0.86	0.22
Skewness	1.0	0.09	0.84	0.28	0.68	0.03	0.28	0.92	0.96	0.02	0.94	0.20	0.16	1.34	0.98	0.36
Kurtosis	0.88	0.03	1.0	0.13	0.78	0.07	0.18	1.39	0.58	0.26	0.52	0.70	0.12	1.53	0.98	0.05
θ																
Mean	0.02	2.06	0.00	5.98	0.38	0.99	0.34	1.02	0.82	0.25	0.60	0.61	0.60	0.41	0.74	0.16
Median	0.20	1.27	0.40	0.87	0.52	0.82	0.20	1.27	0.92	0.13	0.86	0.39	0.92	0.03	0.86	0.31
Standard deviation	0.04	1.93	0.00	7.45	0.68	0.28	0.74	0.38	0.84	0.11	0.68	0.53	0.08	1.67	0.30	0.90
0.025 Quantile	0.86	0.15	0.92	0.03	0.96	0.01	0.60	0.48	0.58	0.52	0.40	0.82	0.10	1.03	0.92	0.19
0.25 Quantile	0.58	0.70	0.28	0.99	0.46	0.78	0.36	1.04	0.68	0.40	0.90	0.16	0.22	1.28	0.80	0.37
0.75 Quantile	0.04	1.86	0.62	0.37	0.36	0.92	0.44	0.81	0.90	0.09	0.58	0.73	0.66	0.34	0.68	0.17
0.975 Quantile	0.02	1.92	0.00	7.52	0.64	0.42	0.74	0.42	0.76	0.30	0.74	0.48	0.12	1.37	0.24	0.93
Skewness	0.04	2.25	0.04	1.73	0.78	0.28	0.60	0.37	0.50	0.52	0.50	0.72	0.24	0.86	0.52	0.63
Kurtosis	0.04	2.31	0.04	1.54	0.92	0.33	0.76	0.35	0.54	0.37	0.62	0.54	0.36	0.73	0.64	0.33

Notes: Values show two-tailed posterior predictive *P*-values (*P*) and posterior predictive effect sizes (ES). Values not shown indicate that the posterior and posterior predictive distributions consisted of a single value for that test statistic.

TABLE 6. Results of model performance assessments for divergence time, and θ test statistics across eight simulated data sets analyzed in BPP

	MSC		Part. Iso. 4		Bottleneck 4		Sec. Cont. 5		Part. Iso. 11		Bottleneck 1		Sec. Cont. 15	
	P	ES	*P*	ES	*P*	ES	*P*	ES	*P*	ES	*P*	ES	*P*	ES
Divergence times														
Mean	0.90	0.16	0.88	0.32	0.34	0.94	0.28	1.09	0.88	0.21	0.96	0.11	0.84	0.20
Median	0.92	0.16	0.90	0.30	0.38	0.83	0.28	1.09	0.92	0.19	0.96	0.09	0.86	0.18
Standard deviation	0.76	0.20	0.70	0.52	0.28	1.19	0.78	0.49	0.0	1.77	0.48	0.55	0.68	0.53
0.025 Quantile	0.82	0.20	0.96	0.07	0.34	0.84	0.30	1.19	0.76	0.24	0.96	0.10	0.90	0.11
0.25 Quantile	0.84	0.18	0.84	0.33	0.36	0.18	0.22	1.24	0.92	0.05	0.98	0.03	0.90	0.17
0.75 Quantile	0.98	0.15	0.84	0.35	0.36	1.01	0.38	0.94	0.70	0.35	0.96	0.15	0.82	0.23
0.975 Quantile	0.96	0.18	0.76	0.40	0.28	1.29	0.40	0.97	0.48	0.58	0.82	0.32	0.82	0.32
Skewness	0.82	0.13	0.96	0.35	0.28	1.11	0.98	0.00	0.86	0.03	0.38	0.43	0.16	1.42
Kurtosis	0.82	0.20	0.82	0.55	0.54	0.08	0.06	1.05	0.34	0.71	0.70	0.16	0.16	1.11
θ														
Mean	0.90	0.11	0.36	0.78	0.94	0.13	0.40	0.75	0.92	0.10	0.86	0.13	0.96	0.12
Median	1.0	0.05	0.80	0.06	0.86	0.25	0.18	1.17	0.98	0.04	0.74	0.36	0.84	0.10
Standard deviation	0.08	0.93	0.38	0.68	0.96	0.10	0.34	0.94	0.30	1.05	0.90	0.27	0.58	0.63
0.025 Quantile	0.92	0.24	0.54	0.60	0.26	0.68	0.22	1.26	0.42	0.74	0.86	0.13	0.70	0.47
0.25 Quantile	0.66	0.55	0.66	0.46	0.34	0.81	0.26	1.20	0.56	0.58	1.0	0.05	0.62	0.39
0.75 Quantile	0.56	0.46	0.36	0.79	0.84	0.22	0.64	0.50	0.60	0.55	0.72	0.32	0.78	0.44
0.975 Quantile	0.82	0.49	0.38	0.39	0.86	0.05	0.70	0.33	0.52	0.74	0.90	0.19	0.68	0.48
Skewness	1.0	0.14	0.76	0.39	0.22	0.80	0.64	0.74	0.86	0.35	0.92	0.26	0.88	0.14
Kurtosis	0.62	0.57	0.46	0.72	0.24	0.62	0.60	0.28	0.82	0.31	0.52	0.54	0.94	0.04

Notes: Values show two-tailed posterior predictive *P*-values (*P*) and posterior predictive effect sizes (ES). "MSC" refers to a data set simulated under the multispecies coalescent model itself. Simulation descriptions refer to the scenarios described in the text and the numbers refer to the parameterizations described in Tables 2 (Partial Isolation), 3 (Bottleneck), or 4 (Secondary Contact).

(Yang and Rannala 2010; Grummer et al. 2014; Jones 2017). However, empirical data from across the tree of life suggest that a number of processes not accounted for by the MSC (such as population structure, IBD, isolation by environment, gene flow, selection, and population demographic changes) are common in empirical systems (Nosil 2008; Barton 2010; Sexton et al. 2014; Wang and Bradburd 2014; Payseur and Rieseberg 2016; Tigano and Friesen 2016). The field of systematics currently has a limited understanding of the extent to which ignoring these complexities impacts species delimitation. This disconnect is concerning given that support from coalescent models is increasingly being used as the primary evidence for making many taxonomic decisions (Burbrink and Guiher 2015; Fennessy et al. 2016; Hotaling et al. 2016; Weir et al. 2016). In this study, we took a step towards addressing this knowledge gap by using simulation to characterize the sensitivity of inference methods that use the MSC to several violations of model assumptions thought to be common in nature, and that might impact species delimitation. We also used posterior predictive simulation to assess the fit of the MSC to empirical data in order to determine if we can identify individual data sets in which violations of model assumptions might lead to biases in species delimitation.

Model Sensitivity

Our simulations suggest varying sensitivity of inferences under the MSC to different violations of model assumptions. This sensitivity can manifest as different types of bias, depending on which assumptions of the model are violated. Both implementations of the MSC we examined showed a tendency to identify two populations connected by migration as distinct species except under low values of θ and/or high rates of migration. This result builds on previous research that demonstrated that BPP was unable to distinguish population structure from speciation (Sukumaran and Knowles 2017). However, the simulations in Sukumaran and Knowles (2017) only explored a limited area of parameter space (e.g., fixing haploid population sizes to 100,000). Here, we characterized the demographic context (in terms of θ and Nm) in which the MSC transitions between strongly supporting a single species model versus a multispecies model. The sensitivity of inference under the MSC to population structure is not surprising given that the model assumes random mating (or panmixia) within species. However, few examples of truly panmictic populations likely exist in nature. Thus, researchers should be cognizant of the fact that MSC inference may identify subpopulations as "species" in some circumstances (as opposed to larger, inclusive populations made up of these subpopulations that would correspond to metapopulation lineages or species *sensu* de Queiroz (1998). Our results also speak to the inherent difficulty in identifying species based purely on migration rate among lineages/populations. While the lowest migration rates we simulated data under are within the range of values that have been described between populations of the same species (e.g. Wang 2009; Hey 2010; Strasburg and Rieseberg 2010), other described species likely exhibit rates of gene flow that are higher than these values.

The impact of gene flow between species on species delimitation under the MSC has previously been investigated using simulation (Zhang et al. 2011; Camargo et al. 2012). These studies demonstrated that the BPP implementation of the MSC generally splits species when migration rates between them are low, and lumps species when migration rates are high. However, this transition appears to depend on the particular parameterization of the simulations (Jackson et al. 2016). These studies assumed constant, bidirectional migration among species over the course of divergence, which may not reflect the way gene flow often occurs in natural populations. We extend this previous work by investigating the impact of gene flow in additional scenarios that have commonly been described in empirical systems: divergence with gene flow where migration decreases over the course of speciation, and recent gene flow during secondary contact. We found that both the partial isolation and the secondary contact scenarios can bias divergence time estimates across a broad range of parameter space. However, BPP was largely robust to these model violations from the perspective of species delimitation, always preferring a two species model to a one species model in all simulations. STACEY usually recognized two species under the partial isolation model, except when species had become completely genetically isolated relatively recently. STACEY was also more sensitive to recent gene flow, often lumping species under the secondary contact model (Table 4). Taken together, these results suggest that recent gene flow is more likely to bias species delimitation under the MSC than gene flow early in the process of speciation.

The two implementations of the MSC also exhibit varying sensitivity to model violations, with BPP being more prone to identify structured populations as species, and STACEY being more prone to lump species that have experienced gene flow. Given the similarity between our analysis conditions for the two methods, we suspect that these differences might result from the distinct approaches they use to delimit species (i.e., the birth-death-collapse model using node heights in STACEY vs. rjMCMC in BPP). Sensitivity of inference under the MSC to population size changes also appears to vary among implementations, although this sensitivity only impacted divergence time estimation. Both STACEY and BPP preferred the correct species delimitation model in all population size change simulations. Under the more extreme population bottleneck scenarios examined, divergence time estimates from BPP became biased towards the present in some circumstances. Although both models assume each branch in the species tree has a constant population size parameter that is independent and identically distributed, divergence time estimates in

STACEY were not substantially biased in our bottleneck simulations.

Model Fit and Species Delimitation

As in previous studies (Reid et al. 2014), our model performance assessments identify poor fit of the MSC to some empirical data sets (Table 5). Because our test statistics are inference-based, our results directly demonstrate that this poor fit sometimes impacts species delimitation and may lead to biased estimates of species demographic parameters (divergence times and θ). Encouragingly, and in concordance with our expectations, we noted the poorest model performance in data sets that previous authors suggested were impacted by MSC model violations (the skink, spider, bat, and human data sets). We also identified poor model fit in the bat data set using the divergence time test statistics, as should be expected given that gene flow among species appears to be common in this system and previous studies have shown that divergence time estimates under the MSC can be biased by gene flow (Leaché et al. 2014b).

The posterior predictive *P*-value has a straightforward definition: it is the posterior probability that another data set generated by the assumed model will have a more extreme quality of interest than the empirical data. The quality of interest is defined by the test statistic. In the case of inference-based test statistics, these values should be related to the potential for our estimate of a particular biological quantity (such as the number of species) to be biased. Effect sizes describe the position of the empirical test statistic value relative to the posterior predictive distribution (effectively capturing the magnitude of any discrepancy between inferences produced by the observed and predicted data sets). Because they are not bounded at 0, effect sizes might be more amenable to identifying whether the observed differences are biologically important. Since delimited species are assigned value as biological entities of intrinsic interest, we need tools that can warn us of challenges in our ability to accurately define their boundaries. Posterior predictive *P*-values and effect sizes, based on inferred delimitations, may offer some hope.

Using test statistics based on the posterior distribution of the number of species is theoretically appealing, because it is of primary interest in species delimitation. However, our results also suggest that these types of test statistics may suffer from a lack of power when the posterior predictive distribution is invariant (or nearly so) and the empirical data also has the same value. For statistics based on species number, such invariance most often occurs when all of the posterior distributions (for both the empirical and posterior predictive data sets) are themselves invariant, placing all probability on one delimitation. Unfortunately, this situation is encountered most frequently when the data contain the most information. In this case, more nuanced approaches to quantifying information content (e.g., Bayes factors) about species number could be more useful as test statistics (Brown and Thomson 2017). Alternative test statistics based on posterior distributions of continuous parameters may also be useful for large or informative data sets.

While test statistics based on posterior distributions of discrete values all suffer, to some extent, from the challenge of resolution described above, the strategy we used to ensure computational tractability of species delimitation for all the posterior predictive data sets likely exacerbated this effect. As is usually done in BPP, we assigned individuals *a priori* to populations when conducting species delimitation on a posterior predictive data set (effectively setting the simulated number of species as the maximum and ensuring individuals were correctly assigned to species). Thus, the number of species estimated from the posterior predictive data sets could not be larger than the number of species estimated from the empirical data (because the simulated data sets are derived from parameter draws from the posterior). The maximum number of possible species for each data set often comprised a substantial proportion of the posterior probability: 1.0 for the fence lizard data set, 0.99 in the bat data set, 0.98 in the skink data set, 0.95 in the frog data set, 0.84 in the human data set, 0.76 in the horned lizard data set, 0.27 in the spider data set, and 0.15 in the cavefish data set. These limitations may partially explain why the mean number of species was usually so similar between the empirical and posterior predictive distributions and the limited power of this particular test.

Although posterior predictive tests are known to be conservative in general, our results demonstrate that these tests of model fit can be useful for characterizing the suitability of the MSC. Posterior predictive *P*-values do not follow frequentist expectations, so values $\leq$ 0.1 occur with a frequency much less than 0.1 when assumptions of the model are met (Gelman et al. 2013). Therefore, the occurrence of values in this range warrants close attention. When small posterior predictive *P*-values or large effect sizes are consistently associated with certain data sets or test statistics, we should be wary of corresponding inferences. For instance, the cavefish and spider data sets that we analyzed show concerning deviation from expectations under the model regarding the number of delimited species (Table 5). Test statistics based on the inferred number of species produced posterior predictive *P*-values $\leq$0.10 and effect sizes $>$2.0 in both cases. Since we are using inference-based test statistics, these discrepancies directly indicate that either the inferred number of species, or the uncertainty associated with that inference, is surprising given the model. While some of these cases were selected based on prior suspicion that they may not match the assumptions of the MSC well, such results could (and should) motivate greater scrutiny for species whose biology is less well understood. One outcome of these tests could be greater reticence to make formal taxonomic changes when poor fit is identified.

Biases in species delimitation due to poor model fit could manifest themselves in two different ways, and these might be distinguishable using different test statistics. For example, biases in the accuracy of our species delimitation estimate might be most effectively detected using the mean, minimum, and maximum value summary statistics for the number of species. Alternatively, biases in the precision of our estimate might be more effectively identified using summary statistics such as the variance, standard deviation, or the range. However, biases in precision and accuracy may often occur together, and since some test statistics are more powerful than others, only one may be "detected." Therefore, we suggest that best practice is to treat any inference with caution if discrepancies are found for either class of test statistic.

Clearly, more work is needed to fully understand the behavior of the test statistics that we have started exploring here. Different approaches to quantifying a data set's information about species delimitation may ultimately prove more efficient or powerful. Nonetheless, these statistics have some appealing properties, since they focus so directly on the goal of a species delimitation analysis—understanding species boundaries. There may even be some appeal in their conservative nature, since they are unlikely to call attention to model problems when they do not exist. For example, our empirical analyses (Table 5) produced more extreme *P*-values and effect sizes than did analyses of the simulated data sets (Table 6), despite the fact that we know the simulated conditions can affect inferred species boundaries. This result reinforces the need for caution in interpreting these species delimitation results and others that give similar results.

Model Elaboration and Future Directions

The poor fit of the MSC to some data sets and the sensitivity of inference under the model to more substantial violations of model assumptions suggests that implementing more elaborate coalescent models for species delimitation that account for these complex demographic processes would be valuable for the field. We should point out that this poor fit and sensitivity do not represent problems with the MSC *per se*. Rather, they highlight biologically realistic scenarios that are not good matches to the scenario that this model was meant to address. We hope that this work will be instructive for those applying these methods to empirical questions, and provide direction for the field going forward. For example, we encourage systematists to use posterior predictive approaches in empirical studies, especially if MSC-based species delimitation will be used as key support for revising taxonomy. Even though it is widely recognized that the process of speciation is largely continuous and species boundaries are sometimes "fuzzy," taxonomy itself is inherently categorical and the field historically has not focused on characterizing this uncertainty. Researchers ultimately also have to determine if a particular entity constitutes a species, which has important real-world implications for conservation, and can impact results of subsequent studies that depend on accurate estimates of species boundaries (e.g., diversification rate studies in the field of systematic biology). These issues may be compounded going forward as coalescent models are used to analyze large, complex genomic data sets, resulting in parameter estimates that have small variances and high statistical confidence, and which are increasingly sensitive to model fit (Kumar et al. 2012). Approaches for model selection and model checking are inherently complementary, and both should be conducted when analyzing genetic data sets to ensure the resulting inferences are accurate. Increasing the use of model checking in species delimitation will help highlight both individual data sets in which inferences should be considered suspect and biological processes that are important in empirical systems, but not well-accommodated by current coalescent models. Ideally, this will also lead to deeper insights into how the process of speciation occurs in nature.

Our results suggest that the structured coalescent (Notohara 1990) and the coalescent with migration (Nath and Griffiths 1993) are candidates for model elaboration. Although the current implementations of the MSC for species delimitation can be computationally demanding (Leaché et al. 2014a), recently developed algorithms are improving on this (Jackson et al. 2016; Jones 2017; Rannala and Yang 2017). Additionally, the influx of genomic data into the field of systematics, and the ease with which large data sets can now be collected for nearly any organism will provide increased power for inferring parameters of coalescent models, and a wealth of data for increasing our understanding of how the speciation process impacts patterns of genetic diversity. Whether or not parameters of more complex models will be identifiable will need to be determined. However, it is encouraging that many recent studies have been successful in estimating parameters of complex demographic models using genomic data sets in the field of population genetics (Excoffier et al. 2013; Harris et al. 2013; Kuhlwilm et al. 2016; Malaspinas et al. 2016). The complexity of the MSC model and the biological processes that lead to speciation can also make it difficult to identify which particular model inadequacies lead to poor model fit. By evaluating the fit of a variety of coalescent models to empirical data sets, however, researchers could potentially identify which biological processes are the most important to accommodate.

Conclusions

Our findings suggest much optimism about the potential for using coalescent models for species delimitation, as methods that assume the MSC appear to fit several empirical data sets well and are robust to minor violations of model assumptions. However, given the complexity of the speciation process in nature,

species delimitation studies should also routinely assess the fit of coalescent models to empirical data sets, and seek additional data to confirm results of coalescent species delimitation analyses. Here, we have developed a framework for assessing model fit in one implementation of the MSC. Future work should expand tools for posterior predictive simulation to all implementations for species delimitation, and start to assess the statistical behavior and performance of different test statistics. Our results also highlight the fact that developing new models that better match the complexity of the speciation process could be valuable for the practice of species delimitation in many systems. These models could provide a more nuanced understanding of species boundaries in systems where species limits are not as distinct, and allow species delimitation analyses to be tailored to individual empirical systems and the processes that are likely to be most important.

Supplementary material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.h6s2k.

Acknowledgements

We used HPC resources provided by the University of Hawaii in conducting many analyses presented here. We thank Amber Wright and Genevieve Mount for helpful discussions that improved this manuscript. We thank Laura Kubatko, Bryan Carstens, and two anonymous reviewers for additional suggestions that improved this manuscript. We also thank Bryan Carstens and Nathan Jackson for providing the human data set.

Funding

This work was supported by an Arnold O. Beckman Postdoctoral Fellowship to A.J.B.; NSF awards [DBI 1356796 to R.C.T.], [DEB 1354506 to R.C.T.], and [DEB 1355071 to J.M.B.].

References

Barley A.J., Thomson R.C. 2016. Assessing the performance of DNA barcoding using posterior predictive simulations. Mol. Ecol. 25:1944–1957.

Barley, A.J., White, J., Diesmos, A.C., Brown, R.M. 2013. The challenge of species delimitation at the extremes: diversification without morphological change in Philippine sun skinks. Evolution 67:3556–3572.

Barton, N. H. 2010. What role does natural selection play in speciation? Philos. Trans. R. Soc. B 365:1825–1840.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M.A., Rambaut, A., Drummond, A.J., Drummond, A., Rambaut, A., Drummond, A., Suchard, M., Xie, D., Rambaut, A., Drummond, A., Ho, S., Phillips, M., Rambaut, A., Drummond, A., Rambaut, A., Shapiro, B., Pybus, O., Minin, V., Bloomquist, E., Suchard, M., Heled, J., Drummond, A., Lemey, P., Rambaut, A., Drummond, A., Suchard, M., Lemey, P., Rambaut, A., Welch, J., Suchard, M., Bouckaert, R., Suchard, M., Rambaut, A., Ayres, D., Darling, A., Zwickl, D., Beerli, P., Holder, M., Baele, G., Lemey, P., Bedford, T., Rambaut, A., Suchard, M., Alekseyenko, A., Lee, C., Suchard, M., Heled, J., Drummond, A., Kuo, L., Mallick, B., Wu, C., Suchard, M., Drummond, A., Kimura, M., Felsenstein, J., Hasegawa, M., Kishino, H., Yano, T., Tamura, K., Nei, M., Tavaré, S., Hayasaka, K., Gojobori, T., Horai, S., Stadler, T., Kühnert, D., Bonhoeffer, S., Drummond, A., Kühnert, D., Stadler, T., Vaughan, T., Drummond, A., Vaughan, T., Kühnert, D., Popinga, A., Welch, D., Drummond, A., Beerli, P., Felsenstein, J., Beerli, P., Felsenstein, J., Beerli, P., van de Laar, T., Pybus, O., Bruisten, S., Brown, D., Nelson, M., Bouckaert, R., Alvarado-Mora, M., R. Pinho, J., Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N., RoyChoudhury, A., Vaughan, T., Drummond, A. 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comput. Biol. 10:e1003537.

Boussau, B., Szöllosi, G.J., Duret, L., Gouy, M., Tannier, E., Daubin, V. 2013. Genome-scale coestimation of species and gene trees. Genome Res. 23:323–330.

Brown, J. M. Thomson, R.C. 2017. Bayes factors unmask highly variable information content, bias, and extreme influence in phylogenomic analyses. Syst. Biol. 66:517–530.

Brown, J.M. 2014a. Detection of implausible phylogenetic inferences using posterior predictive assessment of model fit. Syst. Biol. 63:334–348.

Brown, J.M. 2014b. Predictive approaches to assessing the fit of evolutionary models. Syst. Biol. 63:289–292.

Burbrink, F.T. Guiher, T.J. 2015. Considering gene flow when using coalescent methods to delimit lineages of North American pitvipers of the genus *Agkistrodon*. Zool. J. Linn. Soc. 173:505–526.

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

Carling, M.D. Brumfield, R.T. 2007. Gene sampling strategies for multi-locus population estimates of genetic diversity (θ). PLoS One 2:e160.

Carstens, B.C. Dewey, T.A. 2010. Species delimitation using a combined coalescent and information-theoretic approach: an example from North American *Myotis* bats. Syst. Biol. 59:400–414.

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

de Queiroz, K. 1998. The general lineage concept of species, species criteria, and the process of speciation: a conceptual unification and terminological recommendations. In: Howard D., Berlocher S., editors. Endless forms: species and speciation. Oxford, UK: Oxford University Press. p. 57–75.

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

Doyle, V.P., Young, R.E., Naylor, G.J.P., Brown, J.M. 2015. Can we identify genes with increased phylogenetic reliability? Syst. Biol. 64:824–837.

Duchêne, D.A., Duchêne, S., Holmes, E.C., Ho, S.Y.W. 2015. Evaluating the adequacy of molecular clock models using posterior predictive simulations. Mol. Biol. Evol. 32:2986–2995.

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

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V.C., Foll, M. 2013. Robust demographic inference from genomic and SNP data. PLoS Genet. 9:e1003905.

Fennessy, J., Bidon, T., Reuss, F., Kumar, V., Elkan, P., Nilsson, M., Vamberger, M., Fritz, U., Janke, A., Lydekker, R., Dagg, A., Foster, J., Brown, D., Brenneman, R., Koepfli, K.-P., Pollinger, J., Milá, B., Georgiadis, N., Louis, E., Grether, G., Jacobs, D., Wayne, R., Bock, F., Fennessy, J., Bidon, T., Tutchings, A., Marais, A., Deacon, F., Janke, A., Linnaeus, C., Fennessy, J., Bock, F., Tutchings, A., Brenneman, R., Janke, A., Krumbiegel, I., Evanno, G., Regnaut, S., Goudet, J., Gallus, S., Kumar, V., Bertelsen, M., Janke, A., Nilsson, M., Hassanin, A., Delsuc, F., Ropiquet, A., Hammer, C., B.J. van Vuuren, Matthee, C., M. Ruiz-Garcia, Catzeflis, F., Areskoug, V., Nguyen, T., Couloux, A., Harper, F., Lydekker, R., Matschie, P., ICZN,

Coyne, J., Orr, H., Queiroz, K.D., Avise, J., Ball, R., Hassanin, A., Ropiquet, A., A.-L. Gourmand, Chardonnet, B., Rigoulet, J., Flanagan, S., Brown, M., Fennessy, J., Bolger, D., Lackey, L., Hailer, F., Kutschera, V., Hallström, B., Klassert, D., Fain, S., Leonard, J., Arnason, U., Janke, A., Baker, R., Bradley, R., Wernersson, R., Schierup, M., Jørgensen, F., Gorodkin, J., Panitz, F., Staerfeldt, H.-H., Christensen, O., Mailund, T., Hornshøj, H., Klein, A., et Al., Kuramoto, T., Nishihara, H., Watanabe, M., Okada, N., Agaba, M., Ishengoma, E., Miller, W., McGrath, B., Hudson, C., Reina, O.B., Ratan, A., Burhans, R., Chikhi, R., Medvedev, P., et Al., Elsik, C., Tellam, R., Worley, K., Gibbs, R., Muzny, D., Weinstock, G., Adelson, D., Eichler, E., Elnitski, L., Guigó, R., Consortium, B.G.S., Analysis, et Al., Archibald, A., Cockett, N., Dalrymple, B., Faraut, T., Kijas, J., Maddox, J., McEwan, J., Oddy, V.H., Raadsma, H., Wade, C., I.S.G. Consortium, et Al., Drummond, A., Suchard, M., Xie, D., Rambaut, A., Darriba, D., Taboada, G., Doallo, R., Posada, D., Stamatakis, A., Shimodaira, H., Shimodaira, H., Hasegawa, M., Dmitriev, D., Rakitov, R., Mirarab, S., Reaz, R., Bayzid, M., Zimmermann, T., Swenson, M., Warnow, T., Yang, Z., Librado, P., Rozas, J., Clement, M., Posada, D., Crandall, K., Pritchard, J., Stephens, M., Donnelly, P., Jakobsson, M., Rosenberg, N., Earl, D., VonHoldt, B., Jombart, T., Yang, Z., Excoffier, L., Lischer, H., Bolger, A., Lohse, M., Usadel, B., Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., et Al., Li, W., Godzik, A., Kofler, R., Schlötterer, C., Lelley, T., Rozen, S., Skaletsky, H. 2016. Multi-locus analyses reveal four giraffe species instead of one. Curr. Biol. 26:2543–2549.

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

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, B.D. 2013. Bayesian data analysis. Boca Raton (FL): Chapman and Hall/CRC.

Gerard, D., Gibbs, H.L., Kubatko, L. 2011. Estimating hybridization in the presence of coalescence using phylogenetic intraspecific sampling. BMC Evol. Biol. 11:291.

Gratton, P., Trucchi, E., Trasatti, A., Riccarducci, G., Marta, S., Allegrucci, G., Cesaroni, D., Sbordoni, V. 2016. Testing classical species properties with contemporary data: how 'Bad Species' in the brassy ringlets (*Erebia tyndarus* complex, Lepidoptera) turned good. Syst. Biol. 65:292–303.

Gruenstaeudl, M., Reid, N.M., Wheeler, G.L., Carstens, B.C. 2016. Posterior predictive checks of coalescent models: P2C2M, an R package. Mol. Ecol. Resour. 16:193–205.

Grummer, J.A., Bryson, R.W., Reeder, T.W. 2014. Species delimitation using Bayes factors: simulations and application to the *Sceloporus scalaris* species group (Squamata: Phrynosomatidae). Syst. Biol. 63:119–133.

Hanski, I.A., Gaggiotti, O.E., editors. 2004. Ecology, genetics and evolution of metapopulations. Cambridge, MA: Academic Press.

Harris, K., Nielsen, R., Slatkin, M., Madison, W., Templeton, A., Tajima, F., Slatkin, M., Hudson, R., Wakeley, J., Hey, J., Griffths, R., Tavaré, S., Griffths, R., Tavaré, S., Kuhner, M., Yamato, J., Felsenstein, J., Nielsen, R., Nielsen, R., Beerli, P., Felsenstein, J., Nielsen, R., Wakeley, J., Yang, Z., Rannala, B., Gronau, I., Hubisz, M., Gulko, B., Danko, C., Siepel, A., Schierup, M., Hein, J., Strasburg, J., Rieseberg, L., Green, R., Krause, J., Briggs, A., Maricic, T., Stenzel, U., Rasmussen, M., Guo, X., Wang, Y., Lohmueller, K., Rasmussen, S., Tavaré, S., Balding, D., Griffths, R., Donnelly, P., Pritchard, J., Seielstad, M., Perez-Lezun, A., Feldman, M., Beaumont, M., Zhang, W., Balding, D., Nielsen, R., Williamson, S., Hernandez, R., Fledel-Alon, A., Zhu, L., Nielsen, R., Gutenkunst, R., Hernandez, R., Williamson, S., Bustamante, C., Wiuf, C., Hobolth, A., Christensen, O., Mailund, T., Schierup, M., Li, H., Durbin, R., Steinrücken, M., Paul, J., Song, Y., Sheehan, S., Harris, K., Song, Y., Wiuf, C., Hein, J., McVean, G., Cardin, N., Mailund, T., Halager, A., Westergaard, M., Dutheil, J., Munch, K., Miller, W., Schuster, S., Welch, A., Ratan, A., Bedoya-Reina, O., Browning, B., Browning, S., Purcell, S., Neale, B., K. Todd-Brown, Thomas, L., Ferreira, M., Moltke, I., Albrechtsen, A., Hansen, T., Nielsen, F., Nielsen, R., Gusev, A., Lowe, J., Stoffel, M., Daly, M., Altshuler, D., Hayes, B., Visscher, P., McPartlan, H., Goddard, M., MacLeod, I., Meuwissen, T., Hayes, B., Goddard, M., Palamara, P., Lencz, T., Darvasi, A., I. Pe'er, Ralph, P., Coop, G., Pool, J., Nielsen, R., Gravel, S., Moorjani, P., Patterson, N., Hirschhorn, J., Keinan, A., Hao, L., Marjoram, P., Wall, J., Pritchard, J., Schaffner, S., Foo, C., Gabriel, S., Reich, D., Daly, M., Gravel, S., Henn, B., Gutenkunst, R., Indap, A., Marth, G., Hodgkinson, A., Ladoukakis, E., Eyre-Walker, A., Kong, A., Gudbjartsson, D., Sainz, J., Jonsdottir, G., Gudjonsson, S., Paul, J., Steinrücken, M., Song, Y., Scally, A., Durbin, R., Kong, A., Frigge, M., Masson, G., Besenbacher, S., Sulem, P., Cox, M., Woerner, A., Wall, J., Hammer, M., Noonan, J., Coop, G., Kudarvalli, S., Smith, D., Krause, J., Sankararaman, S., Patterson, N., Li, H., Pääblo, S., Reich, D., Browning, S., Browning, B., Li, Y., Willer, C., Ding, J., Scheet, P., Abecasis, G., Sabeti, P., Reich, D., Higgins, J., Levine, H., Richter, D., Pickerell, J., Coop, G., Novembre, J., Kudarvalli, S., Li, J., Charlesworth, D., Charlesworth, B., Morgan, M., McVicker, G., Gordon, D., Davis, C., Green, P., Lohmueller, K., Albrechtsen, A., Li, Y., Kim, S., Korneliussen, T., Hudson, R. 2013. Inferring demographic history from a spectrum of shared haplotype lengths. PLoS Genet. 9:e1003521.

Hedin, M., Carlson, D., Coyle, F. 2015. Sky island diversification meets the multispecies coalescent - divergence in the spruce-fir moss spider (*Microhexura montivaga*, Araneae, Mygalomorphae) on the highest peaks of southern Appalachia. Mol. Ecol. 24:3467–3484.

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

Hey, J. 2010. The divergence of chimpanzee species and subspecies as revealed in multipopulation isolation-with-migration analyses. Mol. Biol. Evol. 27:921–933.

Hotaling, S., Foley, M.E., Lawrence, N.M., Bocanegra, J., Blanco, M.B., Rasoloarison, R., Kappeler, P.M., Barrett, M.A., Yoder, A.D., Weisrock, W. D. 2016. Species discovery and validation in a cryptic radiation of endangered primates: coalescent-based species delimitation in Madagascar's mouse lemurs. Mol. Ecol. 25:2029–2045.

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

Jackson, N.D., Carstens, B.C., Morales, A.E., O'Meara, B.C. 2016. Species delimitation with gene flow. Syst. Biol. 66:799–812.

Jones, G. 2017. Algorithmic improvements to species delimitation and phylogeny estimation under the multispecies coalescent. J. Math. Biol. 74:447–467.

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

Kuhlwilm, M., Gronau, I., Hubisz, M.J., de Filippo, C., Prado-Martinez, J., Kircher, M., Fu, Q., Burbano, H.A., Lalueza-Fox, C., de la Rasilla, M., Rosas, A., Rudan, P., Brajkovic, D., Kucan, Ž., Gušic, I., Marques-Bonet, T., Andrés, A.M., Viola, B., Pääbo, S., Meyer, M., Siepel, A., Castellano, S. 2016. Ancient gene flow from early modern humans into Eastern Neanderthals. Nature 530:429–433.

Kumar, S., Filipski, A.J., Battistuzzi, F.U., Kosakovsky Pond, S.L., Tamura, K. 2012. Statistics and truth in phylogenomics. Mol. Biol. Evol. 29:457–472.

Leaché, A.D., Fujita, M.K., Minin, V.N., Bouckaert, R.R. 2014a. Species delimitation using genome-wide SNP data. Syst. Biol. 63:534–542.

Leaché, A.D., Harris, R.B., Rannala, B., Yang, Z. 2014b. The influence of gene flow on species tree estimation: a simulation study. Syst. Biol. 63:17–30.

Leblois, R., Estoup, A., Rousset, F. 2009. IBDSim: a computer program to simulate genotypic data under isolation by distance. Mol. Ecol. Resour. 9:107–109.

Lewis, P.O., Xie, W., Chen, M.-H., Fan, Y., Kuo, L. 2014. Posterior predictive Bayesian phylogenetic model selection. Syst. Biol. 63:309–321.

Liu, L., Pearl, D.K. 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Syst. Biol. 56:504–514.

Maddison, W.P. 1997. Gene trees in species trees. Syst. Biol. 46:523–536.

Malaspinas, A.-S., Westaway, M.C., Muller, C., Sousa, V.C., Lao, O., Alves, I., Bergström, A., Athanasiadis, G., Cheng, J.Y., Crawford, J.E., Heupink, H, T., Macholdt, E., Peischl, S., Rasmussen, S., Schiffels, S., Subramanian, S., Wright, J.L., Albrechtsen, A., Barbieri, C.,

Dupanloup, I., Eriksson, A., Margaryan, A., Moltke, I., Pugach, I., Korneliussen, T.S., Levkivskyi, I.P., Moreno-Mayar, J.V., Ni, S., Racimo, F., Sikora, M., Xue, Y., Aghakhanian, F.A., Brucato, N., Brunak, S., Campos, P.F., Clark, W., Ellingvåg, S., Fourmile, G., Gerbault, P., Injie, D., Koki, G., Leavesley, M., Logan, B., Lynch, A., A, E. Matisoo-Smith, McAllister, P.J., Mentzer, A.J., Metspalu, M., Migliano, A.B., Murgha, L., Phipps, M.E., Pomat, W., Reynolds, D., F.-X. Ricaut, Siba, P., Thomas, G, M., Wales, T., Wall, C.M., Oppenheimer, S.J., C. Tyler-Smith, Durbin, R., Dortch, J., Manica, A., Schierup, M.H., Foley, R.A., Lahr, M.M., Bowern, C., Wall, J.D., Mailund, T., Stoneking, M., Nielsen, R., Sandhu, M.S., Excoffier, L., Lambert, D.M., Willerslev, E. 2016. A genomic history of Aboriginal Australia. Nature 538:207–214.

Nath, H.B., Griffiths, R.C. 1993. The coalescent in two colonies with symmetric migration. J. Math. Biol. 31:841–851.

Nosil, P. 2008. Speciation with gene flow could be common. Mol. Ecol. 17:2103–2106.

Notohara, M. 1990. The coalescent and the genealogical process in geographically structured population. J. Math. Biol. 29:59–75.

Papadantonakis, S., Poirazi, P., Pavlidis, P. 2016. CoMuS: Simulating coalescent histories and polymorphic data from multiple species. Mol. Ecol. Resour. 16:1435–1448.

Payseur, B.A., Rieseberg, L.H. 2016. A genomic perspective on hybridization and speciation. Mol. Ecol. 25:2337–2360.

Rambaut, A., Grassly, N.C. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13:235–238.

Rambaut, A., Suchard, M.A., Xie, D., Drummond, A. 2014. Tracer v1.6, Available from http://tree.bio.ed.ac.uk/software/tracer/

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

Rannala, B., Yang, Z. 2017. Efficient Bayesian species tree inference under the multi-species coalescent. Syst. Biol. 66: 823–842.

Rasmussen, M.D., Kellis, M. 2012. Unified modeling of gene duplication, loss, and coalescence using a locus tree. Genome Res. 22:755–765.

Reid, N.M., Hird, S.M., Brown, J.M., Pelletier, T.A., McVay, J.D., Satler, D. J., Carstens, B.C. 2014. Poor fit to the multispecies coalescent is widely detectable in empirical data. Syst. Biol. 63:322–333.

Rousset, F. 1997. Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. Genetics 145:1219–1228.

Schrider, D., Shanku, A.G., Kern, A.D. 2016. Effects of linked selective sweeps on demographic inference and model selection. Genetics 204:1207–1223.

Sexton, J.P., Hangartner, S.B., Hoffmann, A.A. 2014. Genetic isolation by environment or distance: which pattern of gene flow is most common? Evolution 68:1–15.

Strasburg, J.L., Rieseberg, L.H. 2010. How robust are "isolation with migration"; analyses to violations of the IM model? A simulation study. Mol. Biol. Evol. 27:297–310.

Sukumaran, J., Knowles, L.L. 2017. Multispecies coalescent delimits structure, not species. Proc. Natl. Acad. Sci. USA 114:1607–1612.

Tigano, A., Friesen, V.L. 2016. Genomics of local adaptation with gene flow. Mol. Ecol. 25:2144–2164.

Wagner, C.E., Harmon, L.J., Seehausen, O. 2012. Ecological opportunity and sexual selection together predict adaptive radiation. Nature 487:366–369.

Wang, I.J. 2009. Fine-scale population structure in a desert amphibian: landscape genetics of the black toad (*Bufo exsul*). Mol. Ecol. 18:3847–3856.

Wang, I.J., Bradburd, G.S. 2014. Isolation by environment. Mol. Ecol. 23:5649–5662.

Weir, J.T., Haddrath, O., Robertson, H.A., Colbourne, R.M., Baker, A.J. 2016. Explosive ice age diversification of kiwi. Proc. Natl. Acad. Sci. USA 113:E5580–E5587.

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

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

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

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

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.