Syst. Biol. 63(2):263–271, 2014

DOI:10.1093/sysbio/syt106
Advance Access publication December 20, 2013

Upstream Analyses Create Problems with DNA-Based Species Delimitation

MELISA OLAVE[1], EDUARD SOLÀ[2], AND L. LACEY KNOWLES[3]

[1]*Centro Nacional Patagónico – Consejo Nacional de Investigaciones Científicas y Técnicas (CENPAT-CONICET), Puerto Madryn, Chubut U 9120 ACD, Argentina,* [2]*Department de Genètica, Facultat de Biologia and Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Av. Diagonal, 643, 08028, Barcelona, Catalonia, Spain and* [3]*Department of Ecology and Evolutionary Biology, The University of Michigan, Ann Arbor, MI 41809-1029, USA*
**Correspondence to be sent to: Department of Ecology and Evolutionary Biology, The University of Michigan, Ann Arbor, MI 41809-1029, USA; E-mail: knowlesl@umich.edu.*

Received 30 July 2013; reviews returned 18 September 2013; accepted 10 December 2013
Associate Editor: Robb Brumfield

Molecular data are expanding rapidly as a primary data source for species delimitation owing to both the availability of DNA sequences and recent analytical developments based upon the multispecies coalescent (Rannala and Yang 2003; Degnan and Rosenberg 2009). With such methodologies, species can be recognized despite genealogical discord across loci and incomplete lineage sorting (i.e., before reciprocal monophyly has been achieved) (Knowles and Carstens 2007). Nevertheless, we show that some of the zeal bestowed by theoretical ideals needs to be tempered by the practical problems associated with the implementation of coalescent-based approaches to species delimitation because of the potential for errors to be compounded across the multiple steps involved with analyzing DNA sequences.

Genetic approaches to species delimitation generally involve three separate steps: 1) assigning individuals to species, 2) estimating species relationships, and 3) in the case of Bayesian approaches to species delimitation (e.g., Yang and Rannala 2010), estimating the posterior probability that assigned groups are distinct (see O'Meara 2010 for a heuristic approach that does not require a priori assignment of individuals to species). The accuracy of approaches used for delimiting species in the latter two portions of this framework has received considerable attention (e.g., Liu 2008; Knowles 2009; Kubatko et al. 2009; Heled and Drummond 2010; Yang and Rannala 2010; Huang et al. 2010; Leaché and Rannala 2011; Camargo et al. 2012a; Knowles et al. 2012; Rannala and Yang 2013). In contrast, the assignment of individuals to putative species—the first step in species delimitation and pre-requisite in the increasingly popular Bayesian method implemented in the program bpp (Yang and Rannala 2010)—and how it impacts the accuracy of systematic studies that rely exclusively on genetic data for species delimitation has not been studied. Here, we specifically examine how the accuracy of assigning individuals to putative species impacts the downstream delimitation of species from the Bayesian program bpp (Yang and Rannala 2010).

The aim of this study is not an evaluation of bpp *per se*. In fact, previous studies have shown very good performance of bpp when the correct guide tree is provided, even with small datasets (Yang and Rannala 2010; Zhang et al. 2011; Camargo et al. 2012b; Rannala and Yang 2013). Our focus is on the input to bpp, and specifically, how errors and uncertainty with the assignment of individuals to species (i.e., determining individual-species associations) affect the accuracy of species delimitation. Our study focuses on the accuracy of delimited species from the Bayesian program bpp (Yang and Rannala 2010) when using the program STRUCTURAMA (Huelsenbeck and Andolfatto 2007), which like the program STRUCTURE (Pritchard et al. 2000; Falush et al. 2003), is advocated and typically used for the required a priori assignment of individuals to species in bpp, including for datasets with as few as six to eight loci (e.g., Leaché and Fujita 2010; Burbrink et al. 2011; Fujita et al. 2012).

Using simulated data, we chose a small set of conditions that differ with respect to the level of incomplete lineage sorting and conducted analyses with the goal of identifying which factors are driving the errors in the delimitation of species in downstream analyses with bpp (as opposed to characterizing the probability of errors in delimited species by simulating datasets across a broad range of divergence histories and sampling efforts). Nevertheless, the results are directly relevant to empiricists. For example, we focus on the first steps in the DNA sequence-based species delimitation process because of a mismatch between the theoretical recommendations for sampling (e.g., number of loci and individuals) for each of the separate components of

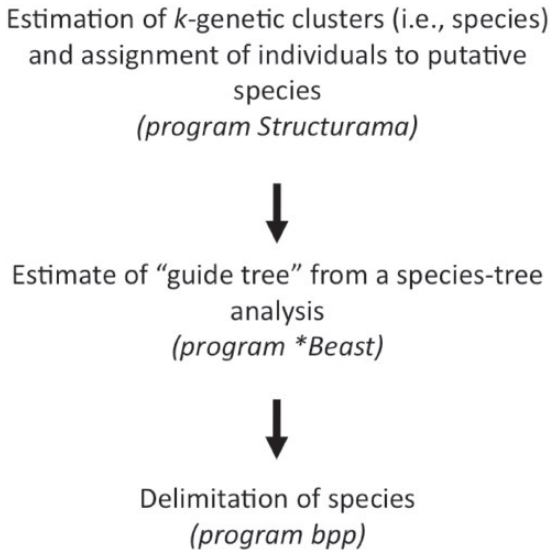

FIGURE 1. Steps involved in genetic-based species delimitation, which involve a series of analyses using different programs (which in this study involved STRUCTURAMA, *BEAST, and bpp). Note that bpp analyses were run with the following conditions: 1) set $k = 8$ with individuals assigned to species a priori (as opposed to estimating them), 2) set $k = 16$ for datasets with two individual sampled per species (i.e., assume that each individual is potentially a different species), and 3) set $k = 8$ and $k = 10$, and estimate individual-species associations with STRUCTURAMA.

analysis, which have gone largely overlooked in practice (Fig. 1). In particular, although bpp may provide accurate estimates of the number of species with a sample of fewer than 10 sequenced loci (Yang and Rannala 2010), estimates of putative species numbers (i.e., k genetic clusters) and assignments of individuals to species with programs like STRUCTURE and STRUCTURAMA (Pritchard et al. 2000; Huelsenbeck and Andolfatto 2007) may not be accurate without large datasets (i.e., datasets approaching 100 independent loci; Rittmeyer and Austin 2012). This raises the concern that the results from empirical studies may be compromised by errors incurred during the estimation of the number of putative species and/or assignments of individuals to species in upstream analyses, even when the practices advanced for users of programs like bpp are followed (see Fujita et al. 2012). Moreover, by using simulated datasets that mirror empirical data collected for species delimitation and species-tree analysis, and in this specific case, a group of South American lizards (genus *Liolaemus*), the estimates can be compared with the known history to assess accuracy using sample sizes currently advocated as best practices. Not only do our results confirm that errors in the upstream analyses used to estimate individual-species association have a significant impact on the accuracy of delimited species but they also call into question current practices with species delimitation based solely on DNA sequences, despite the potential of such approaches in theory.

MATERIALS AND METHODS

Datasets

Simulated datasets, with respect to both the number of taxa and loci, correspond to many representative empirical datasets (reviewed in Fujita et al. 2012). Specifically, eight-taxon symmetric and asymmetric species trees were generated in Mesquite v2.74 (Maddison and Maddison 2010) under total tree depths of 0.4N and 4.0N, representing more and less difficult conditions for species delimitation, respectively (e.g., Knowles and Carstens 2007; Yang and Rannala 2010; Rittmeyer and Austin 2012). Note that we do not consider older species divergences because such scenarios are not particularly challenging and such data would not typically be analyzed with the coalescent-based approaches used here. Coalescent genealogies were generated for five individuals per species for each species tree using the program *ms* (Hudson 2002) under a model of constant population size, no migration, and no recombination within loci. DNA sequences were simulated with the program Seq-Gen (Rambaut and Grassly 1997). All nucleotide datasets were simulated under an HKY model of nucleotide substitution, with a transition–transversion ratio of 3.0, a gamma distribution with shape parameter of 0.8, and nucleotide frequencies of A = 0.3, C = 0.2, T = 0.3, and G = 0.2. Specifically, 1000 base pairs were generated, with $\theta = 0.07$, which was estimated from an actual empirical lizard dataset (genus *Liolaemus*) (Olave et al. in review) using Lamarc v2.1.8 (Kuhner 2006). Similar results were observed with smaller theta values for simulating nucleotide datasets (results not shown). Datasets were simulated with 4, 8, and 14 loci, which cover the range of loci used in the majority of published datasets that apply this approach to delimit species (Fig. 1; reviewed in Fujita et al. 2012).

We also analyzed an empirical dataset with eight *Liolaemus* species of the *boulengeri* and *rothi* complexes (five individuals per species, 14 loci; details of the markers are shown in Supplementary Table S1; doi:10.5061/dryad.3hc8s). These taxa are a subset of those used in a large phylogenetic study of the genus (Olave et al. in review).

Analyses

For each species tree and sample design, 50 replicate datasets were analyzed (following the three steps summarized in Fig. 1; these are the same steps that an empiricist would follow). A total of 2400 bpp analyses were conducted across the 50 replicates of each simulated dataset under the different scenarios.

Individual-species associations.—During the standard practice of species delimitation (Fig. 1), the number of genetic groups (or putative species in this case, and hereafter referred to as species) and individual-species associations would be estimated, for example, using the software STRUCTURAMA 2.0 (Huelsenbeck and

Andolfatto 2007). However, because of difficulties with accurately estimating the number of k genetic clusters using STRUCTURAMA (i.e., the number of species was significantly underestimated under a Dirichlet process prior, with an average of $k = 4$ across datasets), the total number of distinct species was not estimated. Instead of estimating the number of k genetic clusters (i.e., species), individual-group associations were determined assuming eight distinct genetic groups (i.e., a $k = 8$, which corresponded to the actual conditions used to simulate the data). Because k was set to the known value, issues over how to estimate k (see Evanno et al. 2005) do not confound the interpretations of our results from the analyses. However, note that our results are conservative with respect to the errors introduced to downstream analyses involved in delimiting species because we set the number of putative species to the actual value k, as opposed to estimating k. To confirm that the difficulties with estimating the number of k genetic clusters reflect limited amounts of sequence data (see also Rittmeyer and Austin 2012), rather than a sensitivity to the number of taxa used in the simulations, we also simulated and analyzed 50 replicate datasets for species trees with two and four taxa, instead of eight, with five individuals per species under the same parameter settings described earlier. Inaccuracy of the estimated number of clusters was also observed for these datasets, with k significantly overestimated (e.g., $k > 10$ with the four-taxon datasets). Hence, only the eight-taxon datasets were considered for further analyses and discussion. All STRUCTURAMA analyses were run for a total of one million generations for each diploid dataset, sampling every 100 generations; 10% of the data were discarded as burn-in.

Because we are interested in ways that might improve the accuracy of DNA sequence-based species delimitation, we also used a slightly larger number than the actual number of species (i.e., set $k = 10$) for estimating individual-group associations. This decision was made because the maximum number of species a program like bpp can identify is set by the user based on the input of the guide tree. By using a larger number of genetic groups (e.g., $k = 10$ when the data were simulated under a $k = 8$), we can therefore evaluate whether downstream analyses of the species delimitation process are robust to divisions of genetic groupings that are slightly finer than the actual species boundaries. This issue has never been investigated in bpp.

We also considered an alternative approach in which each individual is treated as a potential species, thereby skipping the first step of estimating the number of putative species and assigning individuals to putative species with a program like STRUCTURAMA (and likewise, by passing the potential errors in individual-species associations). For these analyses, only two individuals per species (for a total of 16 potential species) were considered because of computational constraints with bpp; the input into bpp was the tree estimated from *BEAST (Heled and Drummond 2010).

Generating a guide tree of the relationships among putative species.—A species tree was estimated for each dataset using *BEAST (Heled and Drummond 2010) for individual-species assignments based on estimates made with either $k = 8$ or $k = 10$ in STRUCTURAMA, or considering each individual as a potential species (i.e., $k = 16$ in this case with two individuals sampled per species). Each *BEAST analysis was run for 50 million generations with samples taken every 5000 generations and 10% of the data discarded as burn-in, with a model of nucleotide evolution that matched the simulated data (as detailed earlier). Effective sample size (ESS) values were checked and for the few cases where ESS were <200, we ran the Markov chain Monte Carlo (MCMC) until every ESS parameter was >200.

Species delimitation with the program bpp.—The program bpp samples from the posterior distribution of models of species limits using reversible-jump MCMC. That is, given a starting guide tree, the program sequentially collapses internal nodes in the guide tree, evaluating the posterior distribution for each of fewer and fewer putative species. The program assumes no recombination within a locus, free recombination between loci, no gene flow between species, and that the DNA sequences evolved neutrally.

The simulated data were analyzed with bpp v2.0 (Yang and Rannala 2010) with algorithm 1 and the finetune parameter ε set to 15. For species trees with a depth of 4N, we set θ and τ (the timing of species divergence) priors to values that encompass those used to simulate the data, specifically G (7, 100), which results in a mean = 0.07. For the more recent divergence history of 0.4N, the priors on θ and τ were adjusted accordingly to G (0.7, 100), which results in a mean = 0.007. The step lengths for proposals in the MCMC were automatically adjusted to obtain optimal acceptance rates during the analysis that consisted of a burn-in phase of 10 000 steps and 100 000 posterior samples sampled every two steps. Runs were checked to make sure values were between 0.15 and 0.7, as well as ESS values were >200 to assure convergence.

Accuracy of analyses

The accuracy of species delimitation at different steps in the process was evaluated using a number of metrics. This included measures of errors associated with upstream analyses that might impact the downstream bpp analysis (see Fig. 1).

Accuracy of individual-species associations.—A simple index (I_s) was used to examine errors in upstream analyses involving the assignment of individuals to their respective putative species (i.e., individual-species associations). This index, (I_s), measures how many times actual species lineages (i.e., known species lineages) were split as

$$\frac{\sum_{i=1}^{k} \frac{n_{g_i}-1}{n_{r_i}-1}}{k_r}$$

where the numbers of different species (or genetic clusters) that software recognized within the *i*th actual species minus one (n_{gi} – 1), is calculated relative to the maximum number of splits possible, which is the actual number of individuals that are part of the *i*th species minus one (n_{ri} – 1). The index ranges from zero (perfect assignment) to one (species maximally oversplit). Additionally, a mean was calculated among *k* species.

Number of putative species recovered by bpp.—A mean and standard deviation of number of species delimited per dataset among the 50 replicated analyses were calculated for each scenario and combination of different number of loci.

Type I error estimation (failure to reject the wrong hypothesis).—We calculate the proportion of analyses that led to well supported, but nonetheless incorrect inferences about the number of putative species (i.e., under- or overestimates of the number of species with posterior probabilities >0.95). We also used the R statistical software environment (R Core Team 2013) to test for an association between datasets with incorrectly delimited numbers of species (i.e., |the actual number of species – the estimated number of species from bpp|) and the estimated posterior probabilities from the bpp analyses, using linear regression analyses and correlation tests.

Results and Discussion

For the number of loci considered here (which span those typically used in empirical studies that delimited species with bpp to date), there were frequent errors in the delimitation of species for both divergence times (Fig. 2). Particularly disconcerting is the high error rates in the delimitation of species even when the correct number of species is set in STRUCTURAMA (i.e., *k* = 8), with almost all datasets showing errors in the delimitation of species with four loci (i.e., >90% of datasets) and most datasets showing errors with eight loci (i.e., >60% of datasets). Surprisingly in some cases (i.e., when the number of putative species is incorrectly set at *k* = 10), the support for the wrong number of species actually gets stronger with the addition of loci (i.e., the frequency of species delimited with posterior probability of >0.95, shown in black, increases disproportionately relative to the total frequency of errors).

These errors in the delimited species do not reflect the inherent difficultly (i.e., the recency of diversification) of the scenarios represented in the simulations, such that the datasets are simply intractable with respect to analysis with bpp. In almost every case where the number of putative species and individual-species associations were input into bpp (i.e., when they are not estimated with STRUCTURAMA), the number of species was accurately delimited (see Fig. 2 for the few exceptions), which is consistent with studies investigating the performance of bpp by itself (e.g., Yang and Rannala 2010; Zhang et al. 2011; Camargo et al. 2012b). Likewise, the high errors in delimitation do not apparently reflect recalcitrant species-tree estimates. If this was the case, when each individual sampled was treated as a putative species (i.e., *k* = 16 in this case, where two individuals per species were simulated), we would expect pervasive high error rates in the delimitation of species with bpp because of errors in the guide tree. Yet, instead much lower error rates were observed when each individual was treated as a putative species (Fig. 2).

Considered together, these results highlight that a primary source of error in the upstream analysis involves the assignment of individuals to putative species (discussed below). Moreover, the large impact of upstream analyses on the accuracy of downstream analyses used to delimit species (Fig. 2) not only

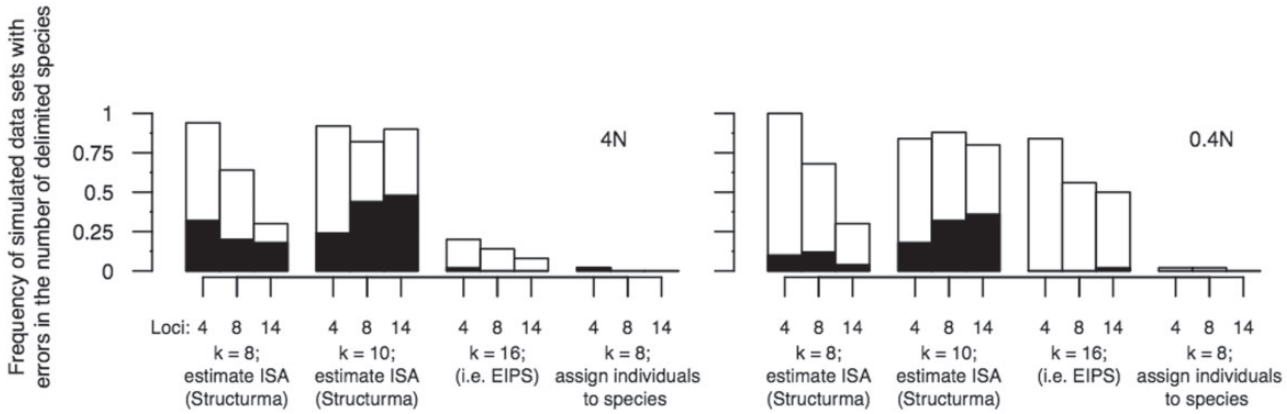

FIGURE 2. The frequency of incorrect inferences with bpp about the number of species delimited across simulated datasets for different sampling efforts and when individual-species associations (ISA) were estimated with STRUCTURAMA with different settings for numbers of putative species (i.e., *k* = 8, *k* = 10), or when the species were correctly assigned to the known species, or when each individual was treated as a potential species in bpp (i.e., EIPS). In some cases, support for the wrong number of species gets stronger with additional loci (i.e., the number of species delimited with posterior probability of > 0.95, shown in black, increases disproportionately). Only the results for the simulations under an asymmetric species tree are shown, and see Supplementary Figure S6 for similar results under a symmetric species tree.

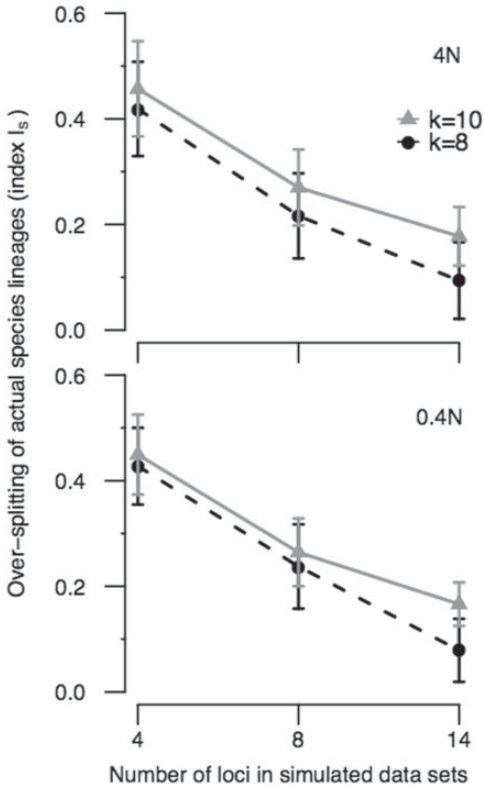

FIGURE 3. Measure of the over-splitting of actual species lineages by the index I_s for different numbers of putative species, k, used for assigning individuals to species and for estimating the guide tree for bpp, for simulated datasets with four, eight, or 14 loci; only the results for the simulations under an asymmetric species tree (at 4N and 0.4N total tree depth) are shown, and see Supplementary Figure S7 for similar results under a symmetric species tree. The index ranges from zero (perfect assignment) to one (species maximally over split).

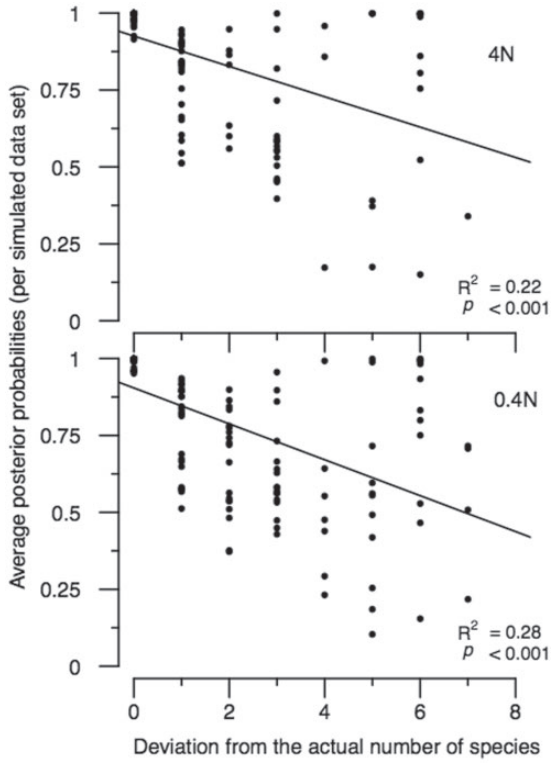

FIGURE 4. Negative association between the posterior probabilities of species delimited with bpp and the deviation from the actual number of species (i.e., 8) when the putative number of species, k, is set to 8, and individual-species associations are estimated with STRUCTURAMA. Note that the correlation was only significant when the putative number of species k is set as 8 (see Supplementary Fig. S9 for k set as 10). Only the results for the simulations under an asymmetric species tree (at 4N and 0.4N total tree depth) are shown given the similarity of results under a symmetric species tree (see Supplementary Fig. S8).

highlights a significant problem in current practices (summarized in Fig. 1) but also suggests an alternative approach for delimiting species with genetic data that may prove more accurate (discussed below).

Impact of errors with upstream analyses on the accuracy of bpp output..—The analyses show that there were always errors with the assignment of individuals to species (Fig. 3), even when the correct number of species (i.e., k = 8) and largest number of loci were used (14 loci, which is consistent with the sample sizes used in empirical datasets; see Fujita et al. 2012). Larger numbers of loci can certainly reduce the errors with upstream STRUCTURAMA (or STRUCTURE) analyses (see Rittmeyer and Austin 2012), as might lower haplotype diversity within loci, given that information about coancestry among individuals from k putative species are characterized by a set of allele frequencies at each locus with these programs (Huelsenbeck and Andolfatto 2007). Nevertheless, our results highlight the problems that can arise because of the mismatch in the data types required at different steps in the delimitation process (Fig. 1) and the high error rates that may accompany studies that rely exclusively on limited numbers of DNA sequences to delimit species (Fig. 2) without some additional data for improving the accuracy of assigning individuals to putative species for recently diverged taxa.

Because we used simulations with a known history, we are able to explore the cause of errors in the downstream analyses (i.e., we can show it is not a function of an intractable history with respect to estimating a guide tree or the delimitation process implemented in bpp, as discussed earlier; Figs. 2 and 3). We can also show that although the posterior probabilities from bpp analyses may be negatively correlated with the number of incorrectly delimited species (Fig. 4), the high variance among replicate datasets (at both k = 8 and k = 10) means that it is possible to get strong support for incorrect estimates of the number of putative species (see also Fig. 2).

These findings have direct relevance to observations from empirical studies regarding the delimitation of species using DNA sequences exclusively. For example, consistent with the high errors in the detection of putative species and assignment of individuals to taxa observed in the simulations here, different empirical studies have also shown that genetic data alone did not detect the same number of putative species recognized in traditional taxonomic treatments in upstream analyses

TABLE 1. Results of the analysis of empirical lizard dataset (genus *Liolaemus*) with properties that corresponded to the simulated datasets (with respect to sampling effort and models of nucleotide variation; see Supplementary Table S1) when the number of species and individual-species associations are set according to traditional taxonomic criteria (primarily morphological features), as opposed to using estimates from STRUCTURAMA

Number of putative species, *k*	Individual-species associations	I_s	Delimited species with bpp	Posterior probabilities
Set at $k = 8$ according to traditional taxonomy	Set according to traditional taxonomy	*na*	8	0.995584
Estimated with STRUCTURAMA, $k = 9$	Estimated with STRUCTURAMA	0.196	7	0.551412
Set at $k = 8$ according to traditional taxonomy	Estimated with STRUCTURAMA	0.053	7	0.987816
Set to at $k = 10$	Estimated with STRUCTURAMA	0.071	7	0.478044

Notes: The I_s-index is a measure of the over-splitting of actual species lineages; I_s is not applicable (na) when the number of putative species is set at $k = 8$ according to traditional taxonomy.

(e.g., Harrington and Near 2012; Edwards and Knowles 2014). With the simulated datasets analyzed here, a much lower number of species was estimated with STRUCTURAMA as well, with an average of $k = 4$ (as noted in the section Materials and Methods, this is why we set k, rather than estimated k). As a consequence, an underestimation of taxa compared with traditional taxonomic treatments would result from estimates with bpp without alternative approaches for establishing individual-species associations. Because the number of estimated species can only decrease, and not increase, from the number of putative species identified in the guide tree used by bpp (Yang and Rannala 2010), underestimates of number of putative species in upstream analyses will always have a significant impact. Such underestimates are certainly not obvious based on an examination of the support values accompanying delimited species. High posterior probability support is associated with many simulated datasets in which the number of species is underestimated with bpp (Fig. 4; and Supplementary Fig. S8). Likewise, when analyzed following the standard protocol advocated for species delimitation (Fig. 1), the actual DNA sequences collected in the *Liolaemus* lizards also provide what appears to be an underestimate of the putative species with high posterior support compared with recognized taxa based on morphology (Table 1).

Alternative procedures in the delimitation of species

Interestingly, treating each individual as a possible species—that is, bypassing the steps of estimating putative species and assigning individuals to these taxa with STRUCTURAMA—produced fewer errors in the delimitation of species with bpp (Fig. 2). Although the correct number of species was frequently delimited with this approach, unfortunately the support for the delimited species was consistently quite low (Fig. 5). This means it is unlikely that the analysis would be interpreted as supporting the correct number of species (which in this case was 8). Note that when the empirical data from *Liolaemus* were analyzed using this strategy, indeed very low posterior probabilities were observed (an average of 0.3411). The low posterior probabilities from the bpp analyses probably reflect the limited information contained in the data about the effective population size of species, a key parameter in bpp, considering that only two individuals were sampled per species (i.e., setting $k = 16$). Adding more individuals sampled per species would provide more information for estimating population parameters (see Yang and Rannala 2010). However, this strategy of considering each individual as a putative species would also have the undesirable effect of increasing the number of parameters to be estimated in bpp, as well as introducing additional errors in the guide tree because of incomplete lineage sorting.

Of course the approach discussed here (Fig. 1), and the program bpp in particular, is just one of many different methods available for species delimitation based on genetic data (reviewed in Carstens et al. 2013). Moreover, despite the failure to accurately delimit species for the set of conditions simulated here, we are not suggesting that researchers should avoid bpp and adopt a different program. Given differences in the assumptions and algorithms employed across methods, the accuracy of the delimited species from the simulated datasets could very well differ depending upon the method used. Instead, our aim is to draw attention to what are potentially compounded problems when the properties of the genetic datasets are sufficient for one, but not all steps in the practice of species delimitation.

Applying multiple genetic markers, such as single nucleotide polymorphisms or microsatellites across multiple loci for the estimation individual-species associations for k putative species and multilocus DNA sequence data for downstream bpp analyses, could provide one obvious potential solution. Another alternative and efficient approach, and perhaps the most cost effective, would be to use more than one data type for species delimitation. For example, traditional taxonomic boundaries might be used in cases where such information is available to determine the number of putative species and assign individuals to species, thereby bypassing the high errors associated with using DNA sequences from a limited number of loci to perform such steps. This alone should greatly enhance the accuracy of species delimitation (i.e., compare the results when putative species and individual-species

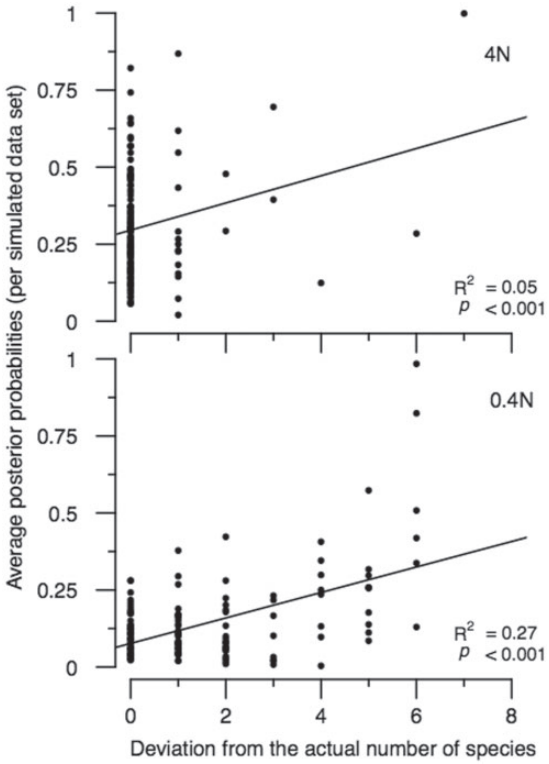

FIGURE 5. Positive association between the posterior probabilities of species delimited with bpp and the deviation from the actual number of species when each individual is treated as potentially a different species (i.e., k = 16, given that two individuals were sampled per species); only results for the simulations under an asymmetric species tree are shown; see Supplementary Figure S10 for similar results under a symmetric species tree.

associations are estimated from genetic data to when they are set, Fig. 2). Morphological and geographic data can also provide valuable information in delimiting species (e.g., Zapata and Jiménez 2012), especially in the identification of putative species and establishing individual-species associations needed for downstream analyses with bpp, even in cases with cryptic species are involved (e.g., Barley et al. 2013).

There is also arguably inherent merit in incorporating multiple data types when delimiting species, which extends beyond the aim of avoiding potential errors in upstream analyses that impact DNA-based estimates of putative species. These pertain to the interpretation of our DNA-based putative species. Depending upon the genetic markers and sampling strategy employed, there is no theoretical reason why the "minimal diagnostic genetic unit" would not extend below species boundaries. As such, it is important to recognize that the issues surrounding DNA-based species delimitation are certainly broader than decisions about what particular analytical approach to use to analyze the genetic data or whether different approaches produce congruent results (see discussion in Carstens et al. 2013). Efforts toward developing methods to accommodate multiple data types in a single quantitative framework, as opposed to the sequential analyses used to integrate information from different data types, are critically needed (Yeates et al. 2011). If such model-based approaches could be extended to multiple data types, such as morphology (i.e., a program that considers not only neutral markers but also morphological characters, including those undergoing selective divergence, for evaluating hypotheses about putative species), we could accommodate taxa where divergence might be more evident along axes of differentiation other than neutral genetic divergence. Moreover, it would bring the field of species delimitation one step closer to identifying boundaries that reflect the accumulation of differences associated with reproductive isolation, as opposed to the ephemeral boundaries only evident in the patterns of neutral genetic markers (i.e., differentiation below the species level).

CONCLUSIONS

Our study highlights how errors in upstream analyses, and specifically, the estimation of individual-species associations, impact the accuracy of downstream analyses with the program bpp. Contrasts in the accuracy of delimited species when individual-species associations are estimated versus setting them to conditions used in the simulations demonstrate that the errors encountered in the bpp analyses are not simply a byproduct of recalcitrant species histories. The errors associated with assignment of individuals to species reflect the mismatch in data requirements at different steps in the process—in fact, the frequency of error estimates reported here is an underestimate given that we set the number of putative species when estimating individual-species associations (e.g., k = 8), rather than estimating both the number of putative species and individual-species associations with the program STRUCTURAMA (see Evanno et al. 2005). Interestingly, higher accuracy of delimitation with bpp was achieved when treating each individual sampled as a putative species, but the low posterior probabilities from such analyses mean it is unlikely that this alternative approach of bypassing the errors in upstream analyses will be useful practically. Overall, these results raise significant questions about current advocated practices for DNA sequence-based species delimitation (note the number of loci used in the simulations, albeit limited, covers the range from the majority of published papers to date).

We suggest that complementing DNA-based approaches for delimitation with other data types, such as morphology, especially for the assignment of individuals to putative species, may be one of the best ways to increase the accuracy of species delimited with programs like bpp (as also noted by Yang and Rannala 2010), which by themselves are accurate with limited genetic data. Moreover, the integration of data types might be necessary given that increasing the number of loci, for example, by applying next-generation sequencing technologies, is unlikely to provide a simple

solution because here too lies a mismatch between data requirements for the programs used in the delimitation process. That is, the short sequence reads from next-generation sequencing platforms (e.g., those from Illumina), while compatible for estimating individual-species associations based on allele frequencies at each locus with programs like STRUCTURAMA, are not ideal for gene-tree based approaches like bpp. Finally, without integrating across data types, interpreting what our DNA-based approaches actually delimit (i.e., putative species, populations, or kin groups) will remain ambiguous, reflecting the resolution of the genetic markers and sampling strategy of the researcher.

Supplementary Material

Data files and/or other supplementary information related to this paper have been deposited at Dryad under doi:10.5061/dryad.3hc8s.

Funding

This work was supported by a doctoral fellowship from Consejo Nacional de Investigaciones Científicas y Técnicas from Argentina (CONICET) (to M.O.); a Fulbright-Bunge y Born fellowship (to M.O.); doctoral fellowships BES-2009-022530 and EEBB-1-2012-05462 from the Ministerio de Ciencia e Innovación from Spain (to E.S.); and a NSF grant [DEB 11-18815] (to L.L.K.).

Acknowledgments

Thanks to members of the Knowles lab, especially to D. Alvarado, C. Muñoz, Q. He, and H. Lanier for helpful comments and suggestions. We also thank all members of Grupo de Herpetología Patagónica (CENPAT-CONICET, Argentina), especially Dr. M. Morando and Dr. L.J. Avila, Dr. M. Riutort group (Facultat de Biologia and IRBio, Universitat de Barcelona), and Dr. J.W. Sites Jr. for generously allowing us to use the marylou5 supercomputer cluster in the Fulton Supercomputing Lab at Brigham Young University (BYU). This research benefitted from valuable comments from three anonymous reviewers.

References

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.

Burbrink F.T., Yao H., Ingrasci M., Bryson R.W. Jr., Guiher T.J., Ruane S. 2011. Speciation at the Mogollon Rim in the Arizona Mountain Kingsnake (*Lampropeltis pyromelana*). Mol. Phylogenet. Evol. 60:445–454.

Camargo A., Avila L.J., Morando M., Sites J.W. Jr. 2012a. Accuracy and precision of species trees: effects of locus, individual, and base pair sampling on inference of species trees in lizards of the *Liolaemus darwinii* group (Squamata, Liolaemidae). Syst. Biol. 61:272–288.

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

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

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

Edwards D.L, Knowles L.L. 2014. Species detection and individual assignment in species delimitation: can integrative data increase efficacy? Proc. R. Soc. B 20132765. http://dx.doi.org/10.1098/rspb.2013.2765

Evanno G., Regnaut S., Goudet J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol. Ecol. 14:2611–2620.

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

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

Harrington R.C., Near T.J., 2012. Phylogenetic and coalescent strategies of species delimitation in Snubnose Darters (Percidae: *Etheostoma*). Syst. Biol. 61:63–79.

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

Huang H., He Q., Kubatko L.S., Knowles L.L. 2010. Sources of error for species-tree estimation: impact of mutational and coalescent effects on accuracy and implications for choosing among different methods. Syst. Biol. 59:573–583.

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

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

Knowles L.L. 2009. Statistical phylogeography. Annu. Rev. Ecol. Syst. 40:593–612.

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

Knowles L.L., Lanier H., Klimov P.B., He Q. 2012. Method choice and species-tree accuracy: full modeling versus summarizing gene-tree uncertainty Mol. Phylogenet. Evol. 65:501–509.

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.

Kuhner M.K. 2006. LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. Bioinformatics 22: 768–770.

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

Leaché A.D., Rannala B. 2011. The accuracy of species tree estimation under simulation: a comparison of methods. Syst. Biol. 60:126–137.

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

Maddison W.P., Maddison D.R. 2010. Mesquite: a modular system for evolutionary analysis. Version 2.74. Available from: http://mesquiteproject.org.

Olave M., Avila L.J., Sites J.W. Jr., Morando M. Model-based approach to test hard polytomies in the Eulaemus clade of the most diverse South American lizard genus Liolaemus (Liolaemini, Squamata). J Evol. Biol. (in review)

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

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

R Core Team. 2013. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing. Available from: http://www.R-project.org/.

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.

Rannala B., Yang Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164:1645–1656.
Rannala B., Yang Z. 2013. Improved reversible jump algorithms for Bayesian species delimitation. Genetics 194:245–253.
Rittmeyer E.N., Austin C.C. 2012. The effects of sampling on delimiting species from multi-locus sequence data. Mol. Phylogenet. Evol. 65:451–463.
Yang Z., Rannala B. 2010. Bayesian species delimitation using multilocus sequence data. Proc. Natl. Acad. Sci. USA. 107:9264–9269.
Yeates D.K., Seago A., Nelson L., Cameron S.L., Joseph L., Trueman J.W.H. 2011. Integrative taxonomy, or iterative taxonomy? Syst. Entomol. 36:209–217.
Zapata F., Jiménez I. 2012. Species delimitation: inferring gaps in morphology across geography. Syst. Biol. 61:179–194.
Zhang C., Zhang D.X., Zhu T., Yang Z. 2011. Evaluation of a Bayesia coalescent method of species delimitation. Syst. Biol. 60: 747–761.