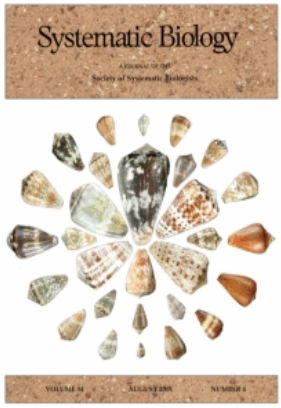

Systematic Biology
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713658732

Delimiting Species without Monophyletic Gene Trees

L. Lacey Knowles [a]; Bryan C. Carstens [a]
[a] Department of Ecology and Evolutionary Biology, Museum of Zoology, 1109 Geddes Avenue, University of Michigan, Ann Arbor, MI, USA

First Published on: 01 December 2007

To cite this Article: Knowles, L. Lacey and Carstens, Bryan C. (2007) 'Delimiting Species without Monophyletic Gene Trees', Systematic Biology, 56:6, 887 — 895

To link to this article: DOI: 10.1080/10635150701701091
URL: http://dx.doi.org/10.1080/10635150701701091

Syst. Biol. 56(6):887–895, 2007

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150701701091

Delimiting Species without Monophyletic Gene Trees

L. LACEY KNOWLES AND BRYAN C. CARSTENS

Department of Ecology and Evolutionary Biology, Museum of Zoology, 1109 Geddes Avenue, University of Michigan, Ann Arbor, MI 48109-1079, USA; E-mail: knowlesl@umich.edu (L.L.K.)

Abstract.—Genetic data are frequently used to delimit species, where species status is determined on the basis of an exclusivity criterium, such as reciprocal monophyly. Not only are there numerous empirical examples of incongruence between the boundaries inferred from such data compared to other sources like morphology—especially with recently derived species, but population genetic theory also clearly shows that an inevitable bias in species status results because genetic thresholds do not explicitly take into account how the timing of speciation influences patterns of genetic differentiation. This study represents a fundamental shift in how genetic data might be used to delimit species. Rather than equating gene trees with a species tree or basing species status on some genetic threshold, the relationship between the gene trees and the species history is modeled probabilistically. Here we show that the same theory that is used to calculate the probability of reciprocal monophyly can also be used to delimit species despite widespread incomplete lineage sorting. The results from a preliminary simulation study suggest that very recently derived species can be accurately identified long before the requisite time for reciprocal monophyly to be achieved following speciation. The study also indicates the importance of sampling, both with regards to loci and individuals. Withstanding a thorough investigation into the conditions under which the coalescent-based approach will be effective, namely how the timing of divergence relative to the effective population size of species affects accurate species delimitation, the results are nevertheless consistent with other recent studies (aimed at inferring species relationships), showing that despite the lack of monophyletic gene trees, a signal of species divergence persists and can be extracted. Using an explicit model-based approach also avoids two primary problems with species delimitation that result when genetic thresholds are applied with genetic data—the inherent biases in species detection arising from when and how speciation occurred, and failure to take into account the high stochastic variance of genetic processes. Both the utility and sensitivities of the coalescent-based approach outlined here are discussed; most notably, a model-based approach is essential for determining whether incompletely sorted gene lineages are (or are not) consistent with separate species lineages, and such inferences require accurate model parameterization (i.e., a range of realistic effective population sizes relative to potential times of divergence for the purported species). It is the goal (and motivation of this study) that genetic data might be used effectively as a source of complementation to other sources of data for diagnosing species, as opposed to the exclusion of other evidence for species delimitation, which will require an explicit consideration of the effects of the temporal dynamic of lineage splitting on genetic data. [Coalescence; genealogical discord; genealogical species concept; gene trees; incomplete lineage sorting.]

Gene trees are often used to infer species boundaries, where some genetic threshold is used to delimit species. For example, conclusions about species boundaries may be based on some level of genetic exclusivity, such as complete reciprocal monophyly or degree of genetic clustering; reviewed in Sites and Marshall, 2004a). Such exclusivity criteria may provide unambiguous definitions of species (e.g., Baum and Shaw, 1995; Herbert et al., 2003), in contrast to the difficult (and sometimes unfeasible) tasks of assessing reproductive compatibility (Mayr, 1963). Such ease of application has propelled the use of genetic markers in species delimitation, as exemplified by the high-throughput screening approaches of DNA sequence variation (i.e., DNA barcoding) for species discovery that provide rapid assessments of biodiversity—as opposed to the time-intensive endeavor of traditional species description and taxonomy (Janzen, 2004). However, well-established population genetic theory (discussed below) raises significant concerns about the use of thresholds applied to genetic data (Takahata and Nei, 1985; Hudson and Coyne, 2002; Hudson and Turelli, 2003; Moritz and Ciero, 2004; Matz and Nielsen, 2005). Indeed, species identified from exclusivity criteria are often incongruent with species delimited from other sources of data (Sites and Marshall, 2004b), raising questions about the accuracy of purported species boundaries (Balakrishan, 2005).

Exclusivity criteria for species delimitation are intuitively appealing for a variety of reasons. Perhaps most notably, they provide a utilitarian approach of broad applicability across disparate taxa. For example, all loci undergo a transition from an initial state of polyphyly upon lineage splitting towards monophyly as the time since speciation increases (Avise and Ball, 1990), so eventually all taxa might be delimited with a criteria such as reciprocal monophyly, assuming no gene flow. Yet, the inherent disconnect between the exclusivity criterion used to delimit species and the actual process of speciation creates a variety of problems. Analytical expectations derived from population genetic theory (Hudson, 1992; Rosenberg, 2002; Wakeley, 2006) indicate that a substantial amount of time is required after the initial divergence of species before there will be a high probability of observing reciprocal monophyly at a sample of multiple loci (Hudson and Coyne, 2002; Hudson and Turelli, 2003). For example, under a strict reciprocal monophyly criterion, it would take more than 1 million years after speciation before species would be delimited if 15 loci were sampled in species with an effective population size (N_e) of 100,000, assuming one generation a year. In larger populations, the number of years that must pass before the species would be recognized increases proportionally (Hudson and Coyne, 2002). Consequently, recently derived species will tend to go undiscovered under a

reciprocal monophyly criterion since species boundaries are not faithfully reflected in a gene tree until ancestral polymorphism has fully sorted (e.g., Hickerson et al., 2006). Similarly, the use of general rules as the basis of species delimitation, such as the 10× rule used in DNA barcoding (Herbert et al., 2003) or a dichotomous key rather than an absolute categorical property (Wiens and Penkrot, 2002), may be problematic. Despite the intent to accommodate aspects of the process of species divergence (e.g., the lack of complete concordance between gene trees and species boundaries due to gene flow or retention of ancestral polymorphism), the large stochastic variance of genetic processes limits the utility of general rules for historical inference (e.g., Knowles and Maddison, 2001; Hudson and Turelli, 2003; Panchal and Beaumont, 2007). Effective use of genetic data for species delimitation requires that (i) the process of species divergence is taken into account, and (ii) the potential contribution of random genetic processes to discordance between gene trees and species boundaries is also considered.

Here we focus specifically on the challenge posed by the retention of ancestral polymorphism to species delimitation to illustrate how the difficulties caused by the lack of monophyletic gene trees can be overcome. One pervading notion (and perhaps a reason for the overreliance on exclusivity principles) is that species delimitation will necessarily be misled by discordance if gene lineages within a species coalesce below the species divergence (i.e., below the speciation event), also known as the species-tree gene-tree discordance problem (Maddison, 1997). A gene tree should not be equated with a species tree (Fig. 1)—clearly the two may differ in topology. However, it is also a misconception to believe that discordant gene genealogies do not provide information about species boundaries. In fact, recent work clearly demonstrates that the gene genealogies provide information about the history of species splitting (i.e., the species tree), despite widespread incomplete lineage sorting in a gene tree (Degnan and Salter, 2005; Maddison and Knowles, 2006; Carstens and Knowles, 2007a; Knowles and Carstens, 2007; Liu and Pearl, 2006). This finding suggests that species lineages can be delimited long before reciprocal monophyly has been reached (most species concepts agree fundamentally that species are lineages; Mayden, 1997; de Queiroz, 2005a, 2005b). We explore this intriguing possibility with a preliminary simulation study of a coalescent-based approach to species delimitation.

The key feature of the approach presented here is that the species history is modeled probabilistically, which differs from previous approaches that use genetic data to infer species boundaries (reviewed in Sites and Marshall, 2004a). In this framework, interpretation of the genetic data is based on explicit reference to the process underlying patterns of genetic differentiation, thereby taking into account the impact of biologically significant events involved in reproductive isolation on patterns of genetic divergence (Orr and Orr, 1996; Hudson and Coyne, 2002; Gavrilets 2003; de Queiroz this issue). Specifically, the approach uses a coalescent framework to estimate gene-tree probabilities under a particular history (Degnan and Salter, 2005) to evaluate the likelihood of lineage splitting (i.e., that speciation has occurred). The same coalescent theory forms the basis for estimating the probability of reciprocal monophyly of gene trees (Hudson, 1992; Rosenberg, 2003; Hudson and Coyne, 2002; Hudson and Turelli, 2003). However, the approach proposed here can be applied to species for which there has not been sufficient time for the full sorting of ancestral polymorphism by genetic drift. As with modeling evolutionary relationships probabilistically (e.g., Rannala and Yang, 2002; Rosenberg, 2002; Liu and Pearl, 2006; Carstens and Knowles, 2007a; Edwards et al., 2007), this approach focuses attention on the relationship between gene trees and the divergence of species lineages rather than equating gene trees with the species history (Maddison, 1997).

The results of this preliminary study are promising in two regards: (i) because very recently originated species can be delimited with the approach it captures the biologically relevant event of lineage splitting (in contrast to relying on a reciprocal-monophyly criterion), and (ii) because genetic differentiation (or lack thereof) is interpreted based on the likelihood of observing such data under a specific historical context of lineage splitting (see also Knowles, 2004), the accuracy of species delimitation can be evaluated. Although this study focuses exclusively on the stochastic loss of gene lineages by genetic drift, additional processes could in principle be incorporated into the model (e.g., gene flow and the effect of mutational processes). We discuss some of the challenges and future developments with coalescent-based approaches for species delimitation (see also Pons et al., 2006, for a coalescent-based application when gene trees are fully sorted within species to infer species boundaries based on the transition from a speciation/extinction dynamic to a process of lineage coalescence). One of the primary goals of this paper is to show how genetic data might be used effectively to delimit species and how individual researchers can evaluate whether such inferences are likely to be accurate. However, we are not advocating the use of genetic data to the exclusion of other evidence for species delimitation. To the contrary, corroboration of species boundaries via independent lines of evidence is very important for diagnosing species (e.g., Payne and Sorenson, 2007), and it is this perspective that motivates the investigation of a coalescent-based approach for species delimitation. By considering how patterns of genetic data reflect the dynamic of species divergence (e.g., the amount of time required for differentiation between species to become apparent; de Queiroz, 2005b) and taking into account the inherent stochasticity of genetic processes (Hudson, 1992), congruence (of lack thereof) of species boundaries among independent data sets might be interpreted (e.g., Masta and Maddison, 2002). Such an endeavor is not possible when species status is diagnosed from genetic data using exclusivity criteria.

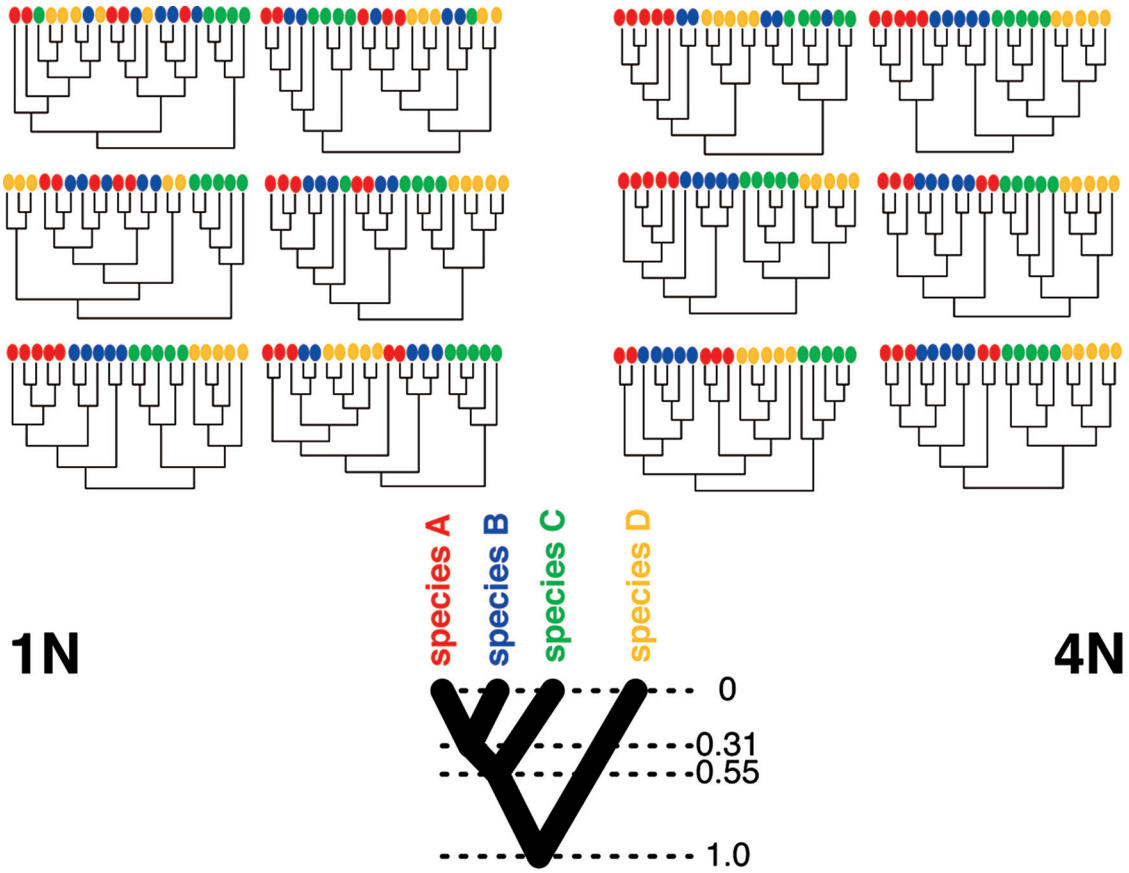

FIGURE 1. Examples of gene trees for recently diverged species showing the level of incomplete lineage sorting expected for any single locus and the degree of discord among sampled loci. Gene trees were simulated with five gene copies (i.e., individuals) per species by neutral coalescence within the species tree (shown in red) with a total tree depth from root to tip of IN (on the left) and 4N (on the right), where N = 100,000.

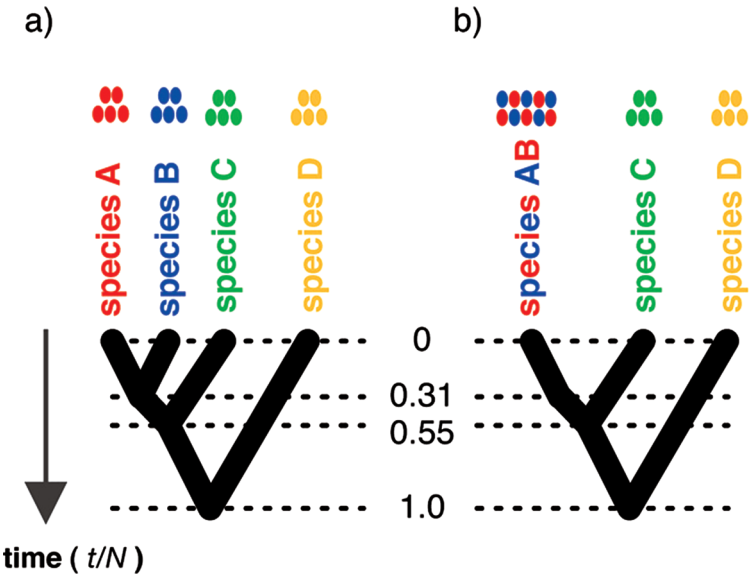

FIGURE 2. Models of the histories used in the simulations to evaluate the coalescent-based approach to species delimitation, where the focus of the study is on whether (a) the history of species divergence of the A and B species lineages can be distinguished from (b) the lack of divergence of the AB lineage.

METHODS

A simulation study was used to evaluate the accuracy of a coalescent-based approach for delimiting recently derived species (i.e., species which are not monophyletic at sampled loci). To examine the effect of the timing of species divergence on the ability to accurately recover the signal of lineage splitting, gene genealogies were simulated under a range of total tree depths (i.e., different species divergence times). These different tree depths correspond to conditions in which the level of incomplete lineage sorting differs, which makes species delimitation based on genetic data increasingly difficult for recent compared to deeper divergence times because of the higher levels of incomplete lineage sorting in the former. Only relatively shallow species histories (i.e., tree depths less than 6N generations) are examined, because the ambiguity associated with species delimitation considered here is caused by incomplete lineage sorting—i.e., the gene trees are not reciprocally monophyletic (Fig. 1). For each of the different time depths, the number of loci sampled in each individual was also varied to examine how increasing the number of loci sampled per species affected the ability to delimit species.

The history of lineage splitting (i.e., the species tree) used in the simulation was simulated under a Yule model, where the species lineages A and B were chosen as the target taxa for evaluating the performance of the coalescent-based approach (Fig. 2a). Gene trees were simulated under a neutral-coalescent process without gene flow (Kingman, 1982; Hudson, 1990). Although the gene trees were simulated across a range of total species-tree depths, the relative branch lengths remained constant. Varying the total species-tree depth resulted in differing amounts of topological discordance among loci and differing levels of incomplete lineage sorting in the simulated gene trees (see also Maddison and Knowles, 2006).

Tests of Separate Species Lineages

Coalescent theory can provide the probability that gene lineages would coalesce to yield a particular gene tree under a specific history (i.e., given the number of generations since divergence and the effective population size of the species; Pamilo and Nei, 1988; Takahata, 1989). Here we consider the topology of a gene tree in relation to its probability under different models of a species history—that is, a history of separate species lineages versus a single species lineage (Fig. 2) is modeled probabilistically from a set of gene tree. The product of the probabilities from the gene trees of each locus under a specific history is then used to evaluate the likelihood of whether species A and B are separate species lineages. This coalescent-based approach involves (1) computing the probability of the gene tree for each specified species tree (i.e., a model where A and B are [are not] separate lineages) using the program COAL (Degnan and Salter, 2005); (2) calculating the likelihood of lineage splitting from the products of the probabilities of the gene trees given the species history (i.e., the species tree); and (3) using a likelihood-ratio test tests (with 1 degree of freedom) to assess whether the likelihood of the model of lineage splitting is significantly higher than a model of no speciation.

To address the question of whether the recently derived species could be effectively delimited (i.e., to estimate the false-negative error rate), 100 replicate data sets were simulated for each time of species divergence and number of sampled loci (i.e., for each of 50 different configurations; see below for details), thereby taking into account the effects of the inherent stochasticity of the coalescent process on the ability to recovery the known history of lineage splitting. Accuracy of the coalescent-based approach was evaluated by recording the proportion of data sets in which the true species history had been identified.

Number of Individuals and Loci Used in Simulations

Gene trees with five gene copies per species were simulated by a neutral-coalescent process, representing five individuals sampled in each species, and multiple gene trees were simulated for each replicate, representing sampled unlinked loci, under the history of lineage splitting (Fig. 2a). Multiple individuals were sampled in each species following the recommendations about sampling design for estimating population relationships with incomplete gene lineage sorting (Maddison and Knowles, 2006; see also Takahata, 1989). Although sampling of individuals might provide some additional information, sampling of multiple loci is critical for providing independent realizations of the process of allele coalescence for a given species history (Felsenstein 2006; Wakeley 2006). The chosen sampling design reflects consideration of this tradeoff; the potential gain in information through the sampling of more individuals will decrease as the time of divergence increases because of the coancestry among individuals (Hudson, 1990; Donnely and Tavaré, 1995). Moreover, increasing the number of individuals sampled per species (as opposed to loci) dramatically increases the number of possible gene trees. Therefore, the information gained through the additional samples would have to offset the lower probabilities of each individual gene tree, and consequently, the reduced ability to distinguish among the models of lineage splitting versus no speciation (Fig. 2).

The gene trees were simulated across a range of total species-tree depths (i.e., different divergence times), specifically at depths of 1N, 2N, 3N, 4N, and 6N, where time is expressed in generations; a depth of 1N would be the equivalent of a total tree depth of 100,000 generations for species with an effective population size of 100,000, and a divergence of 31,000 generations between species A and B (Fig. 2), whereas at a depth of 6N, species divergence is six times greater than the effective population size (i.e., the split between A and B would have occurred 186,000 generations ago with an effective population size of 100,000). A high probability of reciprocal monophyly is not expected for any given locus at any of these depths (see Table 1), with the greatest to the least amount of incomplete lineage sorting at 1N and 6N, respectively.

TABLE 1. Probability of reciprocal monophyly across the different species tree depths when 5 versus 10 individuals were sampled per species, and either a single locus or 10 loci were sequenced in each individual.

	5 Individuals sampled per species		10 Individuals sampled per species	
Depth	1 locus	10 loci	1 locus	10 loci
1N	9.81×10^{-3}	8.30×10^{-21}	6.22×10^{-4}	8.75×10^{-33}
2N	3.29×10^{-3}	1.47×10^{-15}	6.91×10^{-3}	2.47×10^{-22}
3N	7.11×10^{-2}	3.31×10^{-12}	2.55×10^{-2}	1.15×10^{-16}
4N	0.123	7.73×10^{-10}	5.90×10^{-2}	5.08×10^{-13}
5N	0.184	4.48×10^{-8}	0.106	1.86×10^{-10}
6N	0.252	1.02×10^{-6}	0.165	1.48×10^{-8}

Using MESQUITE (Maddison and Maddison, 2004), 100 replicate data sets were simulated for each time of divergence and different sampling efforts. Gene trees were generated using Mesquite's Neutral Coalescence module, which uses an exponential approximation to avoid fully explicit modeling of individuals; mutational variance was not included in this study (see Maddison and Knowles, 2006, for details on how to use MESQUITE to model both coalescent and mutational stochasticity). A constant effective population size (N_e) of 100,000 was used for all species (i.e., ancestral and descendant) in all simulations. Although this value is on the same order of magnitude as observed in many empirical studies (Milot et al., 2000; Jennings and Edwards, 2005; Won et al., 2005; DeChaine and Martin, 2005; Carstens and Knowles, 2007b), scaling all species trees by N allows the results of this study to apply across species which differ in their effective population size. For example, the results at 1N would apply to a species with an effective population size of 500,000 that diverged 500,000 years ago, assuming one generation per year. Similarly, an inheritance scalar can be used to consider how the results might differ between mitochondrial versus nuclear loci; results plotted for mitochondrial DNA (mtDNA) can be scaled to generate expectations for nuclear loci by decreasing the time of divergence (e.g., results for a divergence 4N generations ago for mtDNA would correspond to a 1N divergence for a nuclear locus).

RESULTS AND DISCUSSION

Species Delimitation with or without Monophyly?

At the level of species divergence examined here, the probability of reciprocal monophyly of species A and B for any single locus is very low, and incredibly small if multiple loci are considered (Table 1). Concordance across independent loci is therefore highly unlikely (Hudson and Coyne, 2002; Hudson and Turelli, 2003; Rosenberg, 2002). With this level of incomplete lineage sorting (Fig. 1), species cannot be diagnosed based on visual inspection of the gene trees—it is not clear whether the clustering of gene copies for any one locus, or the degree of concordance (or lack thereof) across loci necessarily constitutes two separate species lineages (see also Maddison and Knowles, 2006).

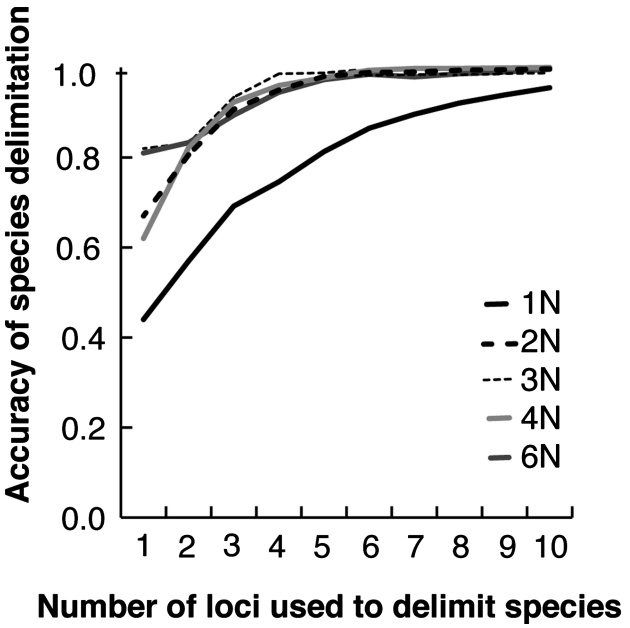

FIGURE 3. Accuracy of the coalescent-based approach for delimiting species A and B with different sampling efforts, showing the false-negative error rate decreases as the number of loci sampled increases from 1 to 10 loci; each line represents a set of simulations for a specific divergence time, ranging from a total species tree depth (see Fig. 2) of 1N to 6N.

The A and B lineages were successfully delimited with the coalescent-based approach across all the different times of divergence examined. Although the separate species were recovered with a high probability, the proportion of replicate data sets in which the species were correctly delimited differed depending on both the timing of species divergence (Fig. 3) and number of loci examined in each species (Fig. 4). Across all divergence times, increased sampling of loci resulted in a decrease in false-negatives (i.e., failures to delimit the separate species) (Fig. 5). Interestingly, even at the very shallow species tree depths (e.g., 1N, which corresponds to a divergence of the A and B lineages of just 31,000 generations ago, with an effective population size of 100,000 for the species), the species were delimited in virtually all the data sets with 10 loci sampled per species. In fact, the species were delimited in almost 90% of the data sets with just three loci per species (Fig. 3). With such recent speciation events, the probability of reciprocal monophyly at any single locus is very low, and the probability of monophyly at multiple loci is effectively zero (Table 1). Even with a relaxed exclusivity criterion of 50% of sampled loci showing reciprocal monophyly, a high probability of species delimitation would not be possibly until 3.76N generations has passed (Hudson and Coyne, 2002). In other words, it would take more than 10 times longer (i.e., beyond the actual time of lineage splitting) before species could be delimited with the relaxed exclusivity criterion compared to the coalescent-based approach. Using a probabilistic model, as presented here, represents a significant advance in species delimitation by reducing the false-negative error rate (i.e., failure to

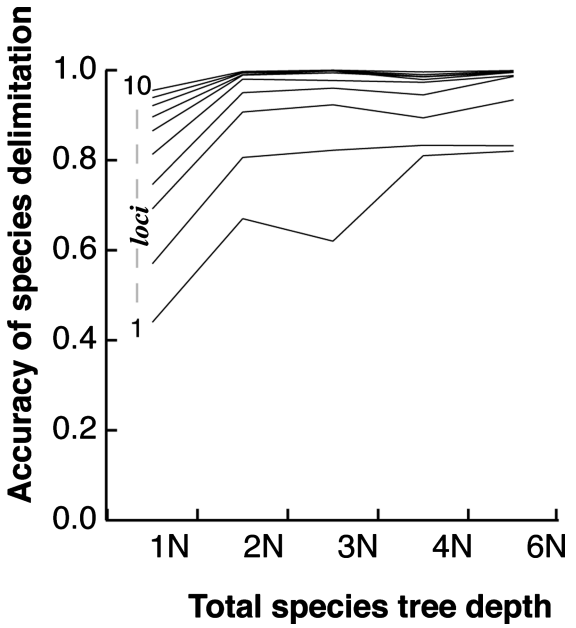

FIGURE 4. Effects of the timing of divergence on accurate delimitation of species A and B, showing the false-negative error rate decreases as the time since species divergence increases for all sampling efforts; the different lines represent a given number of loci sampled.

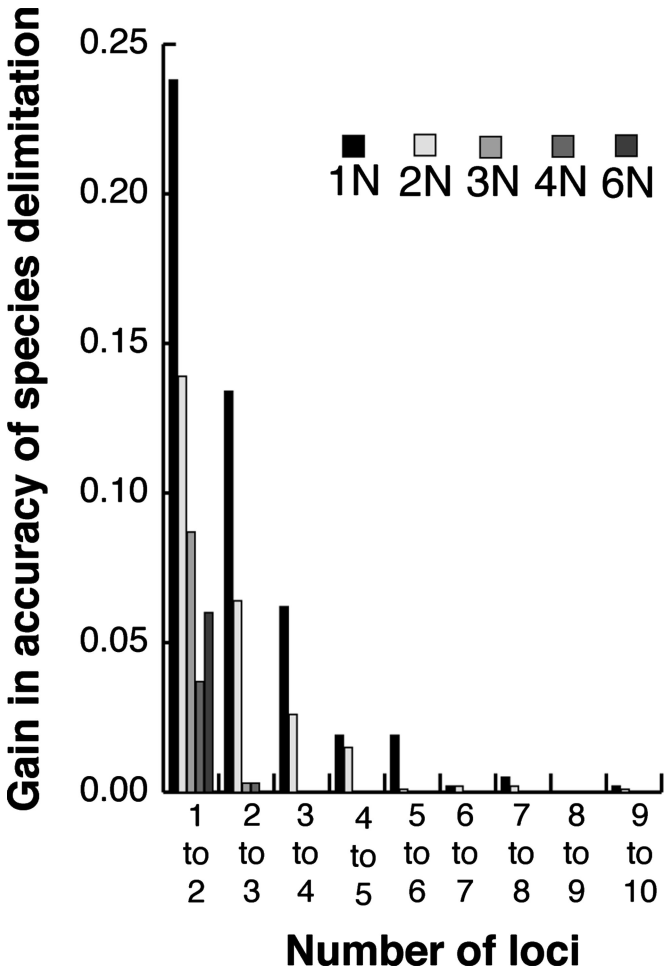

FIGURE 5. Examination of how the additional sampling of loci affects the ability to delimit species. There are incremental gains in accuracy (i.e., decrease in the number of false-negatives, or failures to delimit species) as more loci are added (shown on the *x*-axis), where the different species divergence times are marked with the different shaded bars.

recognize species boundaries), which empirical studies suggest can be quite high (e.g., Gompert et al., 2006; Hickerson et al., 2006; Meier et al., 2006).

The preliminary simulation study, although promising, does not provide a guide to the parameter space in which species may be delimited. This will not be possible without an extensive study. This was not the goal of this study; rather we were interested in exploring the possibility of species delimitation from nonmonophyletic gene trees. This question bears directly on the general utility of genetic data for species delimitation when such boundaries may not be obvious. Stochasticity of the mutation process could also impact the results presented here (and any approach to species delimitation based on genetic data). For example, genetic data may not be effective for delimiting species because of unresolved gene trees when there is insufficient genetic variation or when there are errors with the estimated gene trees.

Probabilistic Models for Species Delimitation

The approach to species delimitation presented here overcomes some of the primary problems with applying exclusivity criteria (e.g., reciprocal monophyly, 10× rule, general rules, and dichotomous keys): the interpretation of the genetic data is based on an explicit consideration of the processes underlying patterns of genetic differentiation, as well as the inherent stochasticity of genetic processes. The probabilistic models considered here focused on the difficulty of species delimitation when there is widespread incomplete sorting due to the retention of ancestral polymorphism. There is, of course, a diversity of processes that may be involved in lineage splitting (Wiens and Penkrot, 2002; Wiens, 2004; Marshall et al., 2006; Omland et al., 2006; Riesberg et al., 2006), each of which could (in principle) be modeled probabilistically (e.g., divergence with gene flow). Deciding which processes might be involved is a separate, but nonetheless important issue. Inferred species boundaries will only be reliable to the extent that the model used is an accurate account of the process of speciation (for a discussion on model selection see Knowles, 2004).

With probabilistic models for species delimitation, sampling of multiple loci and individuals not only has a significant impact on whether species can be delimited in the face of extensive incomplete lineage sorting (Fig. 5), but sampling will also be critical to accurately estimating the parameters used in the model for delimiting species (see also Pons et al., 2006). For example, critical parameters such as estimates of the effective population size (N_e) relative to the timing of divergence (T) between species will determine whether gene trees are more (or less) probable with a history of speciation versus

no lineage splitting (Maddison, 1997). In any empirical application of this approach, an exploration of the sensitivity of the conclusions to a potential mismatch between the actual species' histories and the model used to calculate the gene tree probabilities needs to be carefully examined. For example, rather than relying on point estimates of N_e and T, a range of parameter space that spans the confidence intervals surrounding estimates of these critical population-genetic parameters can be examined (see Carstens and Knowles, 2007a).

It is noteworthy that the proposed approach provides a framework for exploring both the statistical power for delimiting species and the robustness of the conclusions. The false-negative error rate (failure to detect the separate species) depends on the sampling effort (Fig. 3) and the specific context of species divergence (Fig. 4)—the ability to delimit species is context dependent. For a particular set of historical conditions (e.g., specific range of effective population sizes relative to divergence times), any individual investigator can use simulations to examine whether a sampling strategy will provide sufficient power to delimit the putative species of interest, as illustrated by our preliminary study (e.g., Fig. 5). Moreover, the robustness of the conclusions—whether the species status would change by adding more data (Hudson and Coyne, 2002)—can similarly be explored. For example, the empirical data collected (e.g., gene trees from two loci) might be augmented with gene trees estimated from nucleotide data simulated under similar models of molecular evolution to investigate whether the historical signature of species divergence might become apparent with the addition of loci (i.e., without additional sampling, the signal of the separate species lineages is not sufficiently strong to overcome the noise in the gene trees caused by the stochasticity of the coalescent and mutation).

The Use of Genetic Data for Species Delimitation

Even with the limited scope of parameter space considered in this preliminary study, the results highlight some general considerations that are important to any endeavor aimed at delimiting species with genetic data. The power of the test clearly depends on the number of sampled loci (Fig. 3; see also Matz and Nielsen, 2005; Hickerson et al., 2006). It is worth noting that the condition with the poorest performance—a single locus—is what typically is relied upon and is currently the approach taken in DNA barcoding efforts (Fig. 5). The smaller effective population size of mitochondrial DNA is not sufficient to overcome this problem—species were correctly delimited in less than 50% of the replicates for a recent species divergence of 1N (Fig. 3). This implies that unfortunately an accurate assessment of species boundaries will not be possible for the majority of studies that rely on genetic data for delimiting species when they are recently diverged if these studies rely only on mitochondrial sequences. However, our results suggest that good results are possible with a modest number of loci (Fig. 4).

The information contained in independent loci provides valuable information for delimiting species, even though the gene trees are not completely concordant—which is similar to recent studies on modeling evolutionary relationships probabilistically (e.g., Rannala and Yang, 2002; Rosenberg, 2002; Liu and Pearl, 2006; Carstens and Knowles, 2007a; Edwards et al., 2007). It is a mistake to think that combining the loci for an analysis of the concatenated data (e.g., Rokas et al., 2003) will necessarily yield "better" representations of evolutionary history. Studies have shown that when the gene trees of loci are discordant (as expected with recent species divergence) concatenation of the nucleotide data across loci can result in positively misleading inferences about the history of divergence (Kubatko and Degnan, 2007). Moreover, because the loci have different evolutionary histories, there is no way to tell whether the estimated tree is (or is not) a reliable representation of the species' histories (Maddison and Knowles, 2006; Carstens and Knowles, 2007a).

Regardless of which approach is used for making interpretations about species status, it is important to recognize that recommendations about species delimitation based on genetic data are inferences based on the process of neutral divergence among species. If speciation involves selectively driven divergence, then decisions based on neutral DNA divergence will tend to be too conservative (i.e., will fail to recognize species) if the taxa have recently originated. Differences in the temporal dynamics between selected versus neutral divergence (Gavrilets, 2003; Turelli et al., 2001) result in a lag time where differentiation will not be observed in neutral markers, and this period might be rather large (Hudson and Coyne, 2002). The potential to be misled by processes that result in discord between the actual history of lineage splitting and the data used to infer species boundaries highlights the inherent limitations of relying on a single locus (e.g., Fig. 4) or character. Consideration of multiple data types provides a context for identifying differences in historical signal (e.g., Wiens and Penkrot, 2002; Dettmann et al., 2003a, 2003b; Sites and Marshall, 2004b; Ross and Shoemaker, 2005; Starrett and Hedin, 2006; Marshall et al., 2006; Payne and Sorenson, 2007). Comparative analyses are especially important for avoiding biases in species delimitation when a mismatch between the species boundary and a specific type of data reflects the way in which genetic data are interpreted (e.g., Hickerson et al., 2006). For example, recently evolved species (e.g., Meier et al., 2006; Carstens and Knowles, 2007b), such as those arising via divergent selection (e.g., Turner, 1999; Schluter, 2000; Sorensen et al., 2003; Mendelson and Shaw, 2005; Zigler et al., 2005), would not be recognized under methods that rely on the reciprocal monophyly of neutrally evolving gene trees. This false-negative error (i.e., failing to discover new species) clearly arises from how the genetic data are interpreted. Given enough time, the species would be recognized using the genetic exclusivity criteria of reciprocal monophyly at each of the sampled loci (and most of the genome).

CONCLUSIONS

What is recognized as a species boundary is very much influenced by the method used to delimit species (Sites and Marshall, 2004b), which has important ramifications that extend beyond the issue of species delimitation. Species are the basic unit of biodiversity and, as such, are inextricably linked to the study of the processes involved in speciation (Moritz et al., 1992; Agapow et al., 2004; de Queiroz, 2005a, 2005b), as well as the conservation of diversity (Hey et al., 2003; Gompert et al., 2006). Consequently, discrepancies in species identification have focused attention on the methods used to delimit species, an issue of pressing concern with the increase in popularity of genetic data (e.g., Moritz et al., 1992; Sites and Crandall, 1997; DeSalle et al., 2005; Hickerson et al., 2006; Meier et al., 2006). The approach presented here addresses two primary problems with species delimitation based on genetic data—biases in the detection of species reflecting when and how speciation occurred and failure to account for the high variance of genetic processes when inferring species boundaries (Hudson and Coyne, 2002; Matz and Nielson, 2005; Hickerson et al., 2006; Pons et al., 2006). Preliminary results from the probabilistic-modeling approach indicate that accurate species delimitation is possible, despite widespread incomplete lineage sorting and discordance among loci. These findings confirm that it is not necessary to rely on exclusivity criteria (such as genetic thresholds and general rules), and, therefore, that the common problems associated with species delimitation—misleading conclusions arising from how genetic data are interpreted—can be avoided (Hudson and Coyne, 2002; Hickerson et al., 2006; Meier et al., 2006). Genetic data can then provide important corroboration of species boundaries suggested by other sources of information (e.g., morphology) when the temporal dimension influencing the degree of congruence between the genetic data and the species boundaries (and other sources of data) is taken into account (de Queiroz, 2005a, 2005b). However, the study indicates the importance of sampling, both with regards to loci and individuals (see also Matz and Nielson, 2005; Pons et al., 2006), reinforcing the danger of using single-locus data for species delimitation. Sampling design is also key to successful implementation of the approach described here, because it will be critical to parameterizing the model, which ultimately determines whether species boundaries will be inferred correctly.

ACKNOWLEDGMENTS

We thank members of the Knowles lab, John Wiens, Tim Barraclough, and an anonymous reviewer for their input, and a special thanks to Dick Hudson for sharing his program for calculating the probability of reciprocal monophyly. We also appreciate the opportunity to participate in the symposium and thank John Wiens for inviting us. The research was funded by a National Science Foundation grants (DEB-04-47224 and DEB-07-15487) to LLK.

REFERENCES

Agapow, P. M., O. R. P. Bininda-Edmonds, K. A. Crandall, J. L. Gittleman, G. M. Mace, J. C. Marshall, and A. Purvis. 2004. The impact of species concept on biodiversity studies. Q. Rev. Biol. 79:161–179.

Avise, J. C., and M. R. Ball, Jr. 1990. Principles of genealogical concordance in species concepts and biological taxonomy. Pages 45–67 *in* Oxford survey evolutionary biology (D. J. Futuyma and J. Antonovics, eds.). Oxford University Press, New York.

Balakrishnan, R. 2005. Species concepts, species boundaries and species identification: A view from the tropics. Syst. Biol. 54:689–693.

Baum, D. A., and K. L. Shaw. 1995. Genealogical perspectives on the species problem. Pages 289–303 *in* Experimental and molecular approaches to plant biosystematics (P. C. Hoch and A. G. Stephenson, eds.). Missouri Botanical Gardens, St. Louis, Missouri.

Carstens, B. C., and L. L. Knowles. 2007a. Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: An example from *Melanoplus* grasshoppers. Syst. Biol., 56:400–411.

Carstens, B. C., and L. L. Knowles. 2007b. Shifting distributions and speciation: species divergence during rapid climate change. Mol. Ecol. 16:619–627.

DeChaine, E. G., and A. P. Martin. 2004. Historic cycles of fragmentation and expansion in *Parnassius smintheus* (Papilionidae) inferred using mitochondrial DNA. Evolution 58:113–127.

Degnan, J. H., and L. A. Salter. 2005. Gene tree distributions under the coalescent process. Evolution 59:24–37.

de Queiroz, K. 2005a. Different species problems and their solution. BioEssays 26:67–70.

de Queiroz, K. 2005b. Ernst Mayr and the modern concept of species. Proc. Natl. Acad. Sci. USA 102:6600–6607.

DeSalle, R., M. G. Egan, and M. Siddal. 2005. The unholy trinity: Taxonomy, species delimitation, and DNA barcoding. Phil. Trans. R. Soc. B. 360:1905–1916.

Dettman, J. R., D. J. Jacobson, and J. W. Taylor. 2003a. A multilocus genealogical approach to phylogenetic species recognition in the model eukaryote *Neurospora*. Evolution 57:2703–2720.

Dettman, J. R, D. J. Jacobson, E. Turner, A. Pringle, and J. W. Taylor. 2003b. Reproductive isolation and phylogenetic divergence in *Neurospora*: Comparing methods of species recognition in a model eukaryote. Evolution 57:2721–2741.

Donnely, P., and S. Tavaré. 1995. Coalescents and genealogical structure under neutrality. Annu. Rev. Genet. 29:401–421.

Edwards, S. V., L. Liu, and D. K. Pearl. 2007. High-resolution species trees without concatenatioin. Proc. Natl. Acad. Sci. USA 104:5936–5941.

Felsenstein J. 2006. Accuracy of coalescent likelihood estimates: Do we need more sites, more sequences, or more loci? Mol. Biol. Evol. 23:691–700.

Gavrilets, S. 2003. Models of speciation: What have we learned in 40 years? Evolution 57:2197–2215.

Gompert, Z., C. C. Nice, J. A. Fordyce, M. L. Forister, and A. M. Shapiros. 2006. Identifying units for conservation using molecular systematics: The cautionary tale of the Karner blue butterfly. Mol. Ecol. 15:1759–1768.

Hebert, P. D. N., A. Cywinska, S. L. Ball, and J. R. DeWaard. 2003. Biological identifications through DNA barcodes. Proc. R. Soc. Lond. B 270:313–321.

Hey, J., R. S. Waples, M. L. Arnold, R. K. Butlin, and R. G. Harrison. 2003. Understanding and confronting species uncertainty in biology and conservation. Trends Ecol. Evol. 18:597–603.

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

Hudson, R. R. 1990. Gene genealogies and the coalescent process. Pages 1–44 *in* Oxford survey evolutionary biology (D. J. Futuyma and J. Antonovics, eds.). Oxford University Press, New York.

Hudson, R. R. 1992. Gene trees, species trees and the segregation of ancestral alleles. Genetics 131:509–512.

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

Hudson, R. R., and M. Turelli. 2003. Stochasticity overrules the "three-times" rule: Genetic drift, genetic draft, and coalescence times for nuclear loci versus mitochondrial DNA. Evolution 57:182–190.

Janzen, D. H. 2004. Now is the time. Phil Trans. Royal. Soc. Lond. B 359:731–732.

Jennings, W. B., and S. V. Edwards. 2005. Speciational history of Australian grass finches (*Poephila*) inferred from thirty gene trees. Evolution 59:2033–2047.

Kingman, J. F. C. 1982. The coalescent. Stochastic Process. Appl. 13: 235–248.
Knowles, L. L. 2004. The burgeoning field of statistical phylogeography, J. Evol. Biol., 17:1–10.
Knowles, L. L., and B. C. Carstens. 2007. Inferring a population-divergence model for statistical-phylogeographic tests in montane grasshoppers. Evolution 61:477–493.
Knowles, L. L., and W. P. Maddison. 2002. Statistical phylogeography. Mol. Ecol. 11:2623–2635.
Kubatko, L. S., and J. H. Degnan. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst. Biol. 56:17–24.
Liu, L., and D. K. Pearl. 2006. Species trees from gene trees: Reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Technical report #53, Ohio State University.
Maddison, W. P. 1997. Gene trees in species trees. Syst. Biol. 46:523–536.
Maddison, W. P., and L. L. Knowles. 2006. Inferring phylogeny despite incomplete lineage sorting, Syst. Biol. 55:21–30.
Maddison, W. P., and D. R. Maddison. 2004. MESQUITE: A modular system for evolutionary analysis. Version 1.01. http://mesquiteproject.org.
Marshall, J. C., E. Arevalo, E. Benavides, J. L. Sites, and J.W. Sites. 2006. Delimiting species: Comparing methods for Mendelian characters using lizards of the *Sceloporus grammicus* (Squamata: Phrynosomatidae) complex. Evolution 60:1050–1065.
Masta, S. E., and W. P. Maddison. 2002. Sexual selection driving diversification in jumping spiders. Proc. Natl. Acad. Sci. USA 99:4442–4447.
Matz, M. V., and R. Nielsen. 2005. A likelihood ratio test for species membership based on DNA sequence data. Phil. Trans. R. Soc. Lond. B 360:1969–1974.
Mayden, R. L. 1997. A heirarchy of species concepts: The denouement in the saga of the species problem. Pages 381–422 *In* The units of biodiversity (M. F. Claridge, H. A. Dawah, and M. R. Wilson, eds). Chapman and Hall, New York.
Mayr, E. 1963. Animal species and evolution. Harvard University Press, Cambridge, Massachusetts.
Meier, R., K. Shiyang, G. Vaidya, and P. K. L. Ng. 2006. DNA barcoding and taxonomy in Diptera: A tale of high intraspecific variability and low identification success. Syst. Biol. 55:715–728.
Mendelson, T. C., B. D. Inouye, and M. D. Rausher. 2004. Quantifying patterns in the evolution of reproductive isolation. Evolution 58:1424–1433.
Mendelson, T. C., and K. L. Shaw. 2005. Rapid speciation in an arthropod: The likely force behind an explosion of new Hawaiian cricket species revealed. Nature 433:375–376.
Milot, E., Gibbs, H. L. and K. A. Hobson. 2000. Phylogeography and genetic structure of northern populations of the yellow warbler (*Dendroica petechia*). Mol. Ecol. 9:677–681.
Moritz, C., and C. Cicero. 2004. DNA barcoding: Promise and pitfalls. PLoS Biol. 2:1529–1531.
Moritz, C., C. J. Schneider, and D. B. Wake. 1992. Evolutionary relationships within the *Ensatina eschscholtzii* comples confirm the ring species interpretation. Syst. Biol. 41:273–291.
Moyle, L. C., M. S. Olson, and P. Tiffin. 2004. Patterns of reproductive isolation in three angiosperm genera. Evolution 58:1195–1208.
Omland, K. E., J. M. Baker, and J. L. Peters. 2006. Genetic signatures of intermediate divergence: Population history of Old and New World holarctic ravens (*Corvus corax*). Mol. Ecol. 15:795–805.
Orr, H. A., and L. H. Orr. 1996. Waiting for speciation: The effect of population subdivision on the time to speciation. Evolution 50:1742–1749.
Pamilo, P., and M. Nei. 1988. Relationships between gene trees and species trees. Mol. Biol. Evol. 5:568–583.
Panchal, M., and M. A. Beaumont. 2007. The automation and evaluation of nested clade phylogeographic analysis. Evolution 61:1466–1480.
Payne, R. B., and M. D. Sorenson. 2007. Integrative systematics at the species level: Plumage, songs, and molecular phylogeny of quail-finches *Ortygospiza*. Bull. Brit. Orn. Club 217:4–26.
Pons, J., T. G. Barraclough, J. Gomez-Zurita, A. Cardoso, D. P. Duran, S. Hazell, S. Kamoun, W. D. Sumlin, and A. P. Vogler. 2006. Sequence-based species delimitation for the DNA taxonomy of undescribed insects. Syst. Biol. 55:595–609.
Rannala, B., and Z. Yang. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164:1645–1656.
Rieseberg, L. H., T. E. Wood, and E. J. Baack. 2006. The nature of plant species. Nature 440:524–527.
Rokas, A., B. L. Williams, N. King, and S. B. Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798–804.
Rosenberg, N. A. 2002. The probability of topological concordance of gene trees and species trees. Theor. Popul. Biol. 61:225–247.
Rosenberg, N. A. 2003. The shapes of neutral gene genealogies in two species: Probabilities of monophyly, paraphyly, and polyphyly in a coalescent model. Evolution 57:1465–1477.
Ross, K. G., and D. D. Shoemaker. 2005. Species delimitation in native South American fire ants. Mol. Ecol. 14:3419–3438.
Schluter, D. 2000. The ecology of adaptive radiation. Oxford University Press, Oxford, UK.
Sites, J. W. and K. A. Crandall. 1997. Testing species boundaries in biodiversity studies. Conserv. Biol. 11:1289–1297.
Sites, J. W., and J. C. Marshall. 2004a. Empirical criteria for delimiting species. Ann. Rev. Ecol. Evol. Syst. 35:199–227.
Sites, J. W., and J. C. Marshall. 2004b. Delimiting species: A renaissance issue in systematic biology. Trends Ecol. Evol. 18:462–470.
Sorenson, M. D., K. M. Sefc, and R. B. Payne. 2003. Speciation by host switch in brood parasitic indigobirds. Nature 424:928–931.
Starrett, J., and M. Hedin. 2006. Multilocus genealogies reveal multiple cryptic species and biogeographic complexity in the California turret spider *Antrodiaetus riversi* (Mygalomorphae, Antrodieatidae). Mol. Ecol. 16:583–604.
Takahata, N. 1989. Gene genealogy in three related populations: Consistency probability between gene and population trees. Genetics 122:957–966.
Takahata, N., and M. Nei. 1985. Gene genealogy and variance of interpopulational nucleotide differences. Genetics 110:325–344.
Turelli, M., N. H. Barton, and J. A. Coyne. 2001. Theory and speciation. Trends Ecol. Evol. 16:330–343.
Turner, G. F. 1999. Explosive speciation of African cichlid fishes. Pages 113–129 *in* Evolution of biological diversity (A. E. Magurran and R. M. May, eds.). Oxford University Press, Oxford, UK.
Wakeley, J. 2006. Coalescent theory: An introduction. Roberts & Co., Greenwood Village, Colorado.
Wiens, J. J. 2004. What is speciation and how should we study it? Am. Nat. 163:914–923.
Wiens, J. J., and T. A. Penkrot. 2002. Delimiting species using DNA and morphological variation and discordant species limits in spiny lizards (*Sceloporus*). Syst. Biol. 51:69– 91.
Won, Y.-J., A. Sivasunder, Y. Wang, and J. Hey. 2005. On the origin of Lake Malawi cichlid species: A population genetic analysis of divergence. Proc. Natl. Acad. Sci. USA 102:6581–6586.
Zigler, K. S., M. A. McCartney, D. R. Levitan, and H. A. Lessios. 2005. Sea urchin bindin divergence predicts gamete compatibility. Evolution 59:2399–2404.

First submitted 19 November 2006; reviews returned 16 January 2007; final acceptance 21 August 2007
Guest Associate Editor: John Wiens