Opinion

Cell PRESS

Coalescent-based species delimitation in an integrative taxonomy

Matthew K. Fujita[1*], Adam D. Leaché[2], Frank T. Burbrink[3], Jimmy A. McGuire[4] and Craig Moritz[4]

[1] Museum of Comparative Zoology and Department of Organismic and Evolutionary Biology, Harvard University, 26 Oxford Street, Cambridge, MA 02138, USA
[2] Department of Biology and Burke Museum of Natural History and Culture, University of Washington, Box 351800, Seattle, WA 98195, USA
[3] Biology Department, College of Staten Island, City University of New York, Staten Island, 2800 Victory Blvd, New York, NY 10314, USA
[4] Museum of Vertebrate Zoology and Department of Integrative Biology, University of California, 3101 Valley Life Sciences Building, Berkeley, CA 94720, USA

The statistical rigor of species delimitation has increased dramatically over the past decade. Coalescent theory provides powerful models for population genetic inference, and is now increasingly important in phylogenetics and speciation research. By applying probabilistic models, coalescent-based species delimitation provides clear and objective testing of alternative hypotheses of evolutionary independence. As acquisition of multilocus data becomes increasingly automated, coalescent-based species delimitation will improve the discovery, resolution, consistency, and stability of the taxonomy of species. Along with other tools and data types, coalescent-based species delimitation will play an important role in an integrative taxonomy that emphasizes the identification of species limits and the processes that have promoted lineage diversification.

Coalescent theory takes its place in species delimitation

Systematics is a vital discipline in biology that focuses on investigating the origins and causes of biological diversity. The species category is a fundamental unit in biology, and developing robust and highly replicable measures for identifying distinct evolutionary lineages is a central goal of species delimitation (see Glossary). Morphological data and approaches have necessarily dominated species delimitation for centuries, and the emergence of molecular and genomic data sets, together with contemporary species concepts, has brought species delimitation to an interesting crossroads where diverse methodological and philosophical approaches meet. Species delimitation is an integrative field that depends on increasingly diverse data types, yet it remains rife with arguments and opposing opinions regarding the relative utility of alternative approaches. Although this dynamism reflects a vibrant field, it can also impede the stabilization of alpha taxonomy, which can differ significantly depending on alternative applications of the 30+ criteria for delimiting species [1]. Establishing a stable taxonomy is particularly important

Glossary

Akaike information criterion: a measure used to quantify the improvement of fit of a complex model over a less-complex model, thereby justifying the inclusion of additional parameters.
Allopatric speciation: speciation resulting from divergence via geographic isolation.
Alpha taxonomy: the branch of taxonomy focused on discovering, describing, and naming species.
Biological species concept (BSC): a species concept that defines a species as a group of interbreeding populations that is reproductively isolated from other such groups [52].
Coalescent theory: the mathematical and probabilistic theory underlying the evolutionary history of alleles.
Divergence time: the time since two organismal lineages diverged.
Effective population size: the number of breeding individuals in a population that will contribute to the gene pool in the next generation. This is a fundamental quantity in population genetics, often represented as the parameter theta (θ).
Evolutionary species concept (*sensu* [53]): a species concept that defines a species as '...a lineage of ancestral descendant populations which maintains its identity from other such lineages and which has its own evolutionary tendencies and historical fate.'
Gene flow: the movement of genes among populations as a result of migration.
Gene tree: the genealogical relationships among alleles of a gene.
General lineage concept: a species concept that defines a species as an independently evolving lineage [54]. This concept reconciles other species concepts, which differ according to their criteria for identifying the point of lineage divergence.
Genetic drift: the stochastic changes of allele frequencies in a population.
Incomplete lineage sorting: the process by which ancestral alleles are inherited and lost by diverging lineages, resulting in non-monophyly of alleles relative to species trees.
Integrative taxonomy: an approach to taxonomic research that aims to incorporate the diverse data types and methods used in systematic biology to document biodiversity and the evolutionary processes that promote divergence.
Multilocus data: data collected from many unlinked, orthologous segments of nucleic acids (or amino acids). Many applications of population genetics and phylogenetics require these data from multiple individuals per population and/or species.
Parapatric speciation: speciation that has resulted from divergence despite some levels of gene flow between incipient species.
Phylogenetic species concept (*sensu* Cracraft [55]): a species concept that defines a species as '...an irreducible (basal) cluster of organisms, diagnosably distinct from other such clusters, and within which there is a parental pattern of ancestry and descent.' The diagnostic character can be from any trait (morphological or molecular) and of any significance (e.g., a single base pair).
Species delimitation: the process of determining the boundaries and numbers of species from empirical data.
Species trees: a phylogenetic tree showing branching relationships among lineages (species), rather than relationships among alleles (gene trees).
Taxonomic inflation: the artificial increase in the number of species in a group resulting from elevation of geographical variants (often recognized taxonomically as subspecies) to species status. This typically arises when using diagnostic

Corresponding author: Fujita, M.K. (mkfujita@uta.edu)
[*] *Current address*: Department of Biology, University of Texas, 501 S. Nedderman Drive, Box 19498, Arlington, TX 76019-0498, USA.

 0169-5347/$ – see front matter http://dx.doi.org/10.1016/j.tree.2012.04.012 Trends in Ecology and Evolution, September 2012, Vol. 27, No. 9

characters regardless of their significance or type under the phylogenetic species concept (morphological or molecular).
Theta (θ): a fundamental population parameter that is the product of effective population size and mutation rate. It can be interpreted in several ways: as population size comparisons between populations with similar mutation rates, as levels of genetic diversity within populations, and as the capacity for populations to maintain genetic variability [56].

for any field that relies on accurate measures of biodiversity, including ecology and conservation, as well as for research programs dedicated to understanding the evolution of organismal traits, including developmental biology, comparative biology, and genomics [2–6]. An unstable taxonomy also has immense practical ramifications: continuously splitting and lumping taxa based on subjective criteria generates confusion regarding alpha taxonomy, potentially wasting tens of millions of dollars in conservation effort for species listed under the US Endangered Species Act [4].

Delimiting species among sympatric forms is generally non-controversial because reproductive isolation (and thus *de facto* species status) is often readily inferable on the basis of morphological, behavioral, or ecological evidence; rather, the primary challenge usually regards delimiting allopatric species. For most cases of allopatry, the various criteria, which generally serve as proxies for reproductive potential or lineage status, are difficult to measure objectively. For example, proponents of the Biological Species Concept (BSC) are often forced to decide whether some degree of morphological divergence is sufficient to reflect intrinsic reproductive isolation. In an effort to reduce the inherent subjectivity required by the application of such proxies, there has recently been a push in the literature to pursue an 'integrative taxonomy', which attempts to make use of many different sources of data (e.g., molecular, morphological, behavioral, and/or ecological data) to delimit species in a stable and transparent manner [3,7–11] (the term 'iterative taxonomy' has also been suggested [12]). Although we are strong advocates for the application of diverse data types to species delimitation problems, we note that cryptic and allopatric species present great challenges for the field. Here, data sources are usually limited to a combination of geography, ecology, genetics, and morphology; therefore, it is imperative to apply methods that provide objective measures for identifying distinct evolutionary lineages. In our view, and as we discuss below, coalescent theory provides a fundamentally different and stronger framework for objectively identifying cryptic and/or allopatric species using genetic data than is possible using the subjective assessment of morphological proxies for reproductive potential or gene flow.

Applying coalescent theory to species delimitation can infer the dynamics of divergence, the interplay of evolutionary processes, and the relationships among taxa [13–16]. Analytical methods that merge the properties of population genetic processes with phylogenetics have resulted in an important paradigm shift in systematics, where the point of inference is now species trees rather than gene trees [14,17–19]. In turn, these coalescent-based models have provided methods that help researchers identify speciation events, understand processes of speciation, and quantify the probability of evolutionary independence

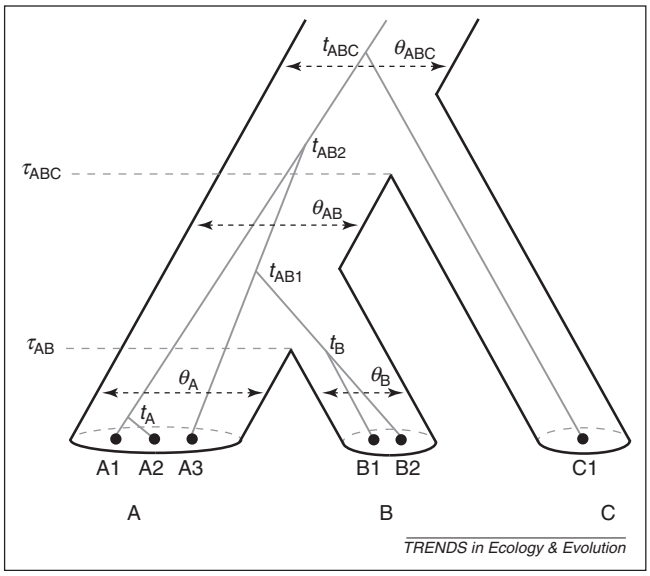

Figure 1. The multispecies coalescent and the associated parameters used in coalescent-based species delimitation models. The bold branches represent organismal lineages (with species A, B, and outgroup C), with their widths corresponding to effective population size (measured as θ) and the nodes correspond to the time of speciation (τ). The solid gray tree within the species tree is a single gene tree, the nodes of which correspond to coalescence times of alleles in the population (t). Note that the gene tree is discordant with respect to the species tree.

[15,20–24]. Successful application of these models hinges upon the availability of now-common multilocus data, where individual gene trees contribute to understanding the depths (divergence times) and widths (effective population sizes) of species trees [17] (Figure 1). In this opinion, we argue for the use of coalescent-based species delimitation as a method to test species delimitation hypotheses. Importantly, coalescent methods should play an important role in stabilizing taxonomy because they have the potential to reduce investigator-driven biases in species delimitation. We first provide an overview of the theory behind coalescent-based species delimitation and then describe how these methods can play an important role in integrative taxonomy by stabilizing taxonomic inferences.

Coalescent-based species delimitation

Coalescent theory provides an opportunity to calculate the probability of speciation

The central aim of coalescent-based approaches is to identify independently evolving lineages, each representing a species. Until recently, species delimitation using molecular data relied on reciprocal monophyly or diagnostic states (e.g., fixed differences) as important criteria for identifying species [3]. Although a single locus can support these criteria, this is often not the case across multiple loci. Alternatively, coalescent-based species delimitation methods use probabilistic approaches that do not require reciprocal monophyly of alleles or fixed differences. This is an important distinction because most alleles are not expected to be reciprocal monophyletic among lineages across most of the genome, particularly at the timescale of recent speciation [25]. Instead, coalescent-based species delimitation uses multilocus data to test alternative hypotheses of lineage divergence that allow for gene tree discordance under genetic drift (Figure 1) [17,20,23].

Understanding the relationship between species delimitation and speciation processes using genetic data relies on the basic fundamentals of coalescent theory. Coalescent theory provides a framework for determining the shapes and patterns of species trees, contingent on demographic parameters such as population size (often designated $\theta = 4N_e\mu$, where N_e is the effective population size and μ is the DNA mutation rate), lineage divergence times (τ), and species tree topology (Figure 1) [26]. Rannala and Yang [17] developed a framework for estimating the likelihood of multilocus data given a species tree, $f(D|S)$, by integrating over gene trees; this forms the foundation for many of the species tree reconstruction methods for estimating the probability of a species tree, $f(S)$ (which includes the parameters θ and τ) [18]. An extension of this framework also allows calculation of the probability of a particular species delimitation given multilocus data (Equation 1):

$$f(S, \Lambda|D) = \frac{1}{f(D)} f(D|S)\, f(S|\Lambda)\, f(\Lambda) \quad (1)$$

where $f(S|\Lambda)$ is the prior distribution of species phylogenies and $f(\Lambda)$ denotes the prior distributions of delimitation models [23]. A delimitation model is a representation of the species tree where each node of the tree represents a speciation event; collapsing nodes merges putative species and represents a model with fewer species (Box 1). This is the framework for Bayesian species delimitation as implemented in the program Bayesian Phylogenetics & Phylogeography (BP&P) (Box 1) [17], which identifies independent evolutionary lineages given multilocus data and a starting guide species tree (Box 2). Studies have already successfully applied BP&P in species delimitation. For example, Leaché and Fujita [27] used a multilocus data set and BP&P to infer four species of West African forest geckos (*Hemidactylus*) that share similar morphologies and ecologies. Under an integrative taxonomic framework, Burbrink *et al.* [28] used BP&P to test whether two divergent lineages of mountain kingsnake represented two species, as hypothesized based on geographic distributions, phylogeny, and ecological niche modeling. Setiadi *et al.* [29] used BP&P to identify cryptic species in an assemblage of Indonesian *Limnonectes* fanged frogs.

Other methods for coalescent-based species delimitation

Several other approaches are available for delimiting species using coalescent techniques (Table 1). Knowles and Carstens [20] developed a method that uses likelihood ratio tests or the Akaike Information Criterion to determine whether a collection of gene trees better fit a single-species model or a two-species model, relying on coalescent estimators (e.g., divergence times and population sizes) to obtain likelihood values for each hypothesis [30]. O'Meara [21] extended the Knowles and Carstens [20] approach by developing methods that do not require pre-specifying the species tree, therefore simultaneously delimiting species and inferring the species tree that maximizes the probability of the gene trees. The program SpedeSTEM [31] estimates the likelihood of a species tree given a collection of independent gene trees and uses information theory to

Box 1. The coalescent model and species delimitation

Phylogenetic inference and population genetic models are becoming more integrated than ever before [14], and species delimitation methods are also benefitting from this union. Coalescent theory [57–59] provides a tractable theoretical framework for modeling population history, and the multispecies coalescent model [17] is now used widely in phylogenetic and species delimitation methods [18,23,31]. The multispecies coalescent model tracks the genealogical history of samples back to a common ancestor for samples representing multiple species.

Bayesian and maximum likelihood (ML) methods are available for coalescent-based species delimitation. The methods use the same underlying multispecies coalescent model [17], but species delimitation models are evaluated differently. The ML method uses point estimates for the genealogies at each locus (obtained from ML phylogenetic estimation) and for the effective population size parameter theta (θ) [20,31]. Alternative hierarchical species delimitation models that differ with respect to the numbers of species are evaluated using a hierarchical likelihood ratio test or an information-theoretic approach [31]. The Bayesian method incorporates genealogical uncertainty by estimating gene trees directly from the sequence data for each locus, and prior distributions are used for θ and the depth of the species tree [23]. Reversible-jump Markov chain Monte Carlo (rjMCMC) is used to obtain the posterior probability distribution of species delimitation models that differ in species numbers (Box 2) [23]. Although these approaches are quite different both analytically and philosophically, the fundamental goal of identifying distinct evolutionary lineages using a coalescent model remains the same.

Coalescent methods for species delimitation [23,31] can detect lineages that are evolutionarily distinct at very shallow timescales. It might seem somewhat paradoxical that these methods interpret genealogical discordance as deep coalescence, given that gene exchange among populations and/or species is more probable among recently diverged species [15]. Fortunately, although gene exchange is not modeled, simulations show that the methods are conservative in that they will lump species together that are exchanging genes at a population frequency exceeding as few as one migrant per generation [24], the classic inflection point between homogenization via gene flow and divergence under genetic drift [60].

Table 1. Implementation of coalescent-based species delimitation

Program	Analytical framework	Input	Output	Refs
GMYC	Best-fit tree branching models (coalescent vs Yule)	Ultrametric gene tree	Transition point from species to populations, and estimate of species number	[32,61]
Brownie	Maximum likelihood or gene tree parsimony	Gene trees	Species tree of delimited species and group membership	[62]
SpedeSTEM	Maximum likelihood and/or information theory	Sequence alignments and group membership	Species tree of delimited species	[31]
BP&P	Bayesian and/or reversible-jump MCMC	Sequence alignments, group membership, and guide tree	Posterior probability distribution of species delimitation models, coalescent times, and population sizes	[23]

Box 2. Coalescent species delimitation in practice

We illustrate how Bayesian species delimitation is used to test alternative species delimitation models. We first propose that the focal clade contains four species, and that we have sampled *n* specimens per species (Figure I). The goal of the method is to obtain a posterior probability distribution for species delimitation models that consist of the four-species model and the alternative models that contain as few as a single species. The fact that each model contains different numbers of species means that they are also described by different numbers of model parameters (e.g., species divergence times and effective population size parameters; Figure I), and this necessitates the use of an algorithm that can estimate a posterior probability distribution for multidimensional problems, namely reversible-jump Markov chain Monte Carlo (rjMCMC) [23]. The method evaluates alternative models derived from all possible subtrees that are generated by collapsing or splitting nodes on the guide tree (Figure I). Only trees that are less resolved than the guide tree are evaluated. The rjMCMC algorithm performs two types of moves to traverse the alternative species delimitation models: moves that propose splitting species, and moves that propose joining species. The proportion of time spent on each model is proportional to the posterior probability of the model.

General guidelines for effective sampling strategies for Bayesian species delimitation are emerging from simulation studies and analyses of empirical data. Although it is possible to implement the method with only a single locus, sampling multiple sequences per species increases accuracy [23,24], although computational limitations could prohibit the method from working efficiently with data sets containing hundreds of sequences (e.g., human population data [23]). Zhang *et al.* [24] used simulations to demonstrate that species delimitations using BP&P was feasible with 5–10 sequences for 1–2 loci when the tree depth was large (e.g., τ_0 = 0.01) and population size was small (e.g., θ = 0.001); shorter divergence times and larger populations sizes will require increased sampling. Ence and Carstens [31] found that evolutionary lineages as young as 0.5*n* generations can be validated as distinct using as few as five loci. The guide tree plays a crucial role in the BP&P method, and misspecification of the guide tree can result in strong support for models containing more species [29]. Supplying a guide tree with resolution finer than the species level is important if the purpose is to explore models that contain diversity among phylogeographic groups.

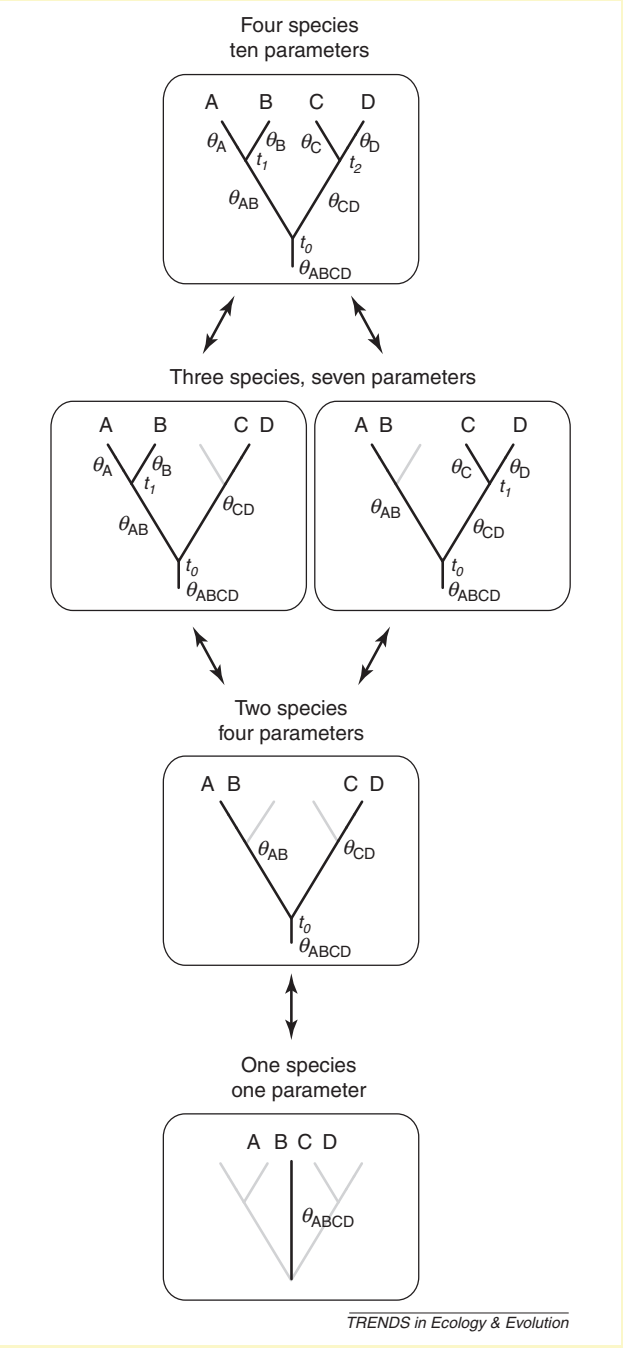

Figure I. An example of Bayesian species delimitation on a four-species symmetric guide tree using BP&P [23]. The fully resolved guide tree contains ten population demographic parameters: seven effective population sizes (θ) for the four species and the three ancestral populations, and three species coalescence times (τ). The model evaluates all subtrees that are fully compatible with the guide tree, and this produces the alternative models containing fewer numbers of species (and parameters), all the way down to the one-species model. The rjMCMC algorithm produces a posterior probability distribution for the species delimitation models.

generate metrics of comparison [22]. The General mixed Yule coalescent (GMYC) method, introduced by Pons *et al.* [32], uses the distinct branching patterns between divergence (Yule model) and intraspecific diversification (coalescent model) to distinguish between species processes (e.g., speciation and extinction) and population processes (coalescence of alleles). The transition between these distinct branching patterns is the threshold used to delimit species using GMYC. Table 2 lists recent examples of species delimitations that have employed these coalescent-based methods.

Coalescent-based species delimitation as a component of integrative taxonomy

Coalescent-based species delimitation alone cannot fully illuminate all features of the lineage diversification process, which is a major goal of integrative taxonomy. In many instances, researchers first suspect that they have discovered new species by identifying divergences among individuals or groups using morphological or single-locus barcode data. In an integrative framework, further evidence is collected and analyzed to determine whether patterns of divergence are seen in different data types,

Table 2. Recent empirical examples of coalescent-based species delimitation

Species complex	Method	Significance	Refs
Hemidactylus geckos	BP&P	One of the first instances of using Bayesian species delimitation in a cryptic complex, using population structure to inform the guide tree	[27]
Myotis bats	Hierarchical likelihood-ratio test, Bayes Factor, and information theoretic	Consistently delimited species using three different approaches	[22]
Corallorhiza orchids	BP&P	Highlights the difficulties, and often conflicts, of applying different operational criteria to species delimitation	[41]
Lampropeltis snakes	BP&P	Uses BP&P to delimit two species, then gene flow analysis to infer process	[28]
Heliconius butterflies	BP&P	Confirmed the distinctness of two species under a variety of analytical scenarios and algorithms	[24]
Drosophila flies, *Manacus* manakins, *Lactarius* fungi, and *Melanoplus* grasshoppers	Kingman's coalescent and nonparametric delimitation, and simultaneous species tree inference	Simultaneously delimited species, assigned individuals to species, and inferred species phylogeny	[21]
Rivacindela beetles	GMYC model	Introduces the GMYC model to identify speciation events where branching rates switch from intraspecific ('coalescent') to interspecific ('diversification') patterns	[32]
Madagascar insects	GMYC model	Develops a multiple-threshold approach for the GMYC model	[61]
Sceloporus (fence lizards)	BP&P	BP&P delimits five species using 29 nuclear loci; consistent with estimates based on a mitochondrial DNA genealogy	[23]
Homo sapiens	BP&P	Human ethnic populations are collapsed to a single species using the t-threshold approach	[23]
Rotifers	BP&P	Four species of asexual bdelloid rotifers were supported with cytochrome c oxidase I and 28S sequences	[23]
Tiny greenbul (*Phyllastrephus debilis*)	BP&P	When testing for speciation between three subspecies, BP&P did not support the elevation of some subspecies to species status	[66]
Etheostoma darters	Information theoretic	Demonstrated that coalescent-based species delimitation can infer fewer species than morphological-based inferences	[67]
Rana chinensis species group (East Asian brown frogs)	BP&P	Inferred four species of frogs that appear associated with ecological divergence	[68]
Notiospathius wasps	GMYC	Compared species delimitation using alternative models and morphology, finding some discordance between methods and markers, but ultimately identifying cryptic species	[69]
Limnonectes (fanged frogs)	BP&P	Although identifying multiple species using BP&P, the authors consider the assumptions and limitations that may lead to false positives	[29]
Xanthoparmelia fungi	BP&P	Significant cryptic diversity and species delimitation based on BP&P highlight incongruences with taxonomy based on morphology	[70]
Typhlichthys (southern cavefish)	BP&P and Brownie	Cryptic diversity was found using both methods of species delimitation, although results differed based on sampling of both individuals and loci	[71]
Liolaemus (Lizards)	BP&P, SpeDeSTEM, and Approximate Bayesian Computation (ABC)	This study introduces an ABC method for species delimitation that incorporates gene flow. While BP&P was the most accurate, the ABC method performed well under scenarios with gene flow	[72]

further supporting species status (e.g., [7]). However, there is still substantial subjectivity in determining how to interpret morphological data, which can be influenced by non-heritable factors, such as environmental or maternal conditions. As discussed above, single-locus data represent the history of a single gene that might not be representative of organismal history. Thus, both subjective interpretation of morphological data and the idiosyncratic history of a single-locus can confound species delimitation.

By contrast, a multilocus investigation using coalescent-based species delimitation tools (as described above) embraces the stochastic nature of the coalescent to calculate the posterior probabilities of a speciation event. These methods are highly replicable and largely free from

investigator bias so their results should be robust and stable; application of the method using the same or similar molecular data by different investigators should produce the same species delimitation. Other coalescent methods can be used to further investigate speciation processes, a major goal of integrative taxonomy. For example, quantifying migration rates between species to determine whether lineage divergence occurred in the absence of gene flow (allopatric speciation) or with some levels of gene flow (e.g., parapatric speciation) is possible using programs such as IMa2 [33,34]. Indeed, demonstrating the cessation of gene flow corroborates the inference of evolutionary independence of lineages (although inferring patterns of gene flow from complex evolutionary histories can be difficult [35,36]). Identifying morphological and/or ecological differences between lineages can also help understand processes of speciation. For example, Hoskin *et al.* [37] found that response to past climate change had important influences on both genetic divergence and ecologically relevant morphological diversification in the rainforest frog, *Cophixalus ornatus*.

We fully appreciate that it will be a challenge to convince traditional taxonomists that coalescent-based species delimitations are valid, particularly when such lineages are not found in post-hoc analysis to have been morphologically diagnosable. However, we ask skeptics to consider the nature of species. If they agree that species are lineages (and they might not), then it should be clear that the real goal of species delimitation is to test hypotheses of evolutionary independence regardless of whether putative lineages differ in phenotypic character systems that are readily apparent to human observers. If morphological character differences have served as proxies for reproductive isolation and lineage independence, then scientists should be willing to consider direct genetic evidence of lineage status as particularly appropriate for the question at hand, particularly when analytical results are presented in a rigorous statistical framework.

Increasing taxonomic stability

Species concepts have different criteria for delimiting species

A species is a hypothesis based on data that supports its evolutionary independence from other lineages. Coalescent-based species delimitation identifies independent evolutionary lineages and, as such, satisfies the criteria of several species concepts under the umbrella of the General Lineage Concept, such as the Evolutionary Species Concept (which equates to the BSC assuming no gene flow between species [38]). Subsequent to identification, systematists must name new species according to nomenclatural codes whose primary purpose is to stabilize taxonomy, a highly desirable goal given that biological disciplines rely on accurate species designations. The reality is that the current system has allowed a plurality of incomparable nomenclatural practices that has led to taxonomic instability in some taxa, for example the apparent 'taxonomic inflation' in primates [39,40]. This instability arises because of differences in species concepts and their criteria; for instance, a complex of species delimited under the criterion of the Phylogenetic Species Concept, which requires only character diagnosability regardless of the biological significance of that character (and regardless of reproductive isolation), might be a single species under the criterion for the BSC, but the reverse might not necessarily be so [39]. For example, Barrett and Freudenstein [41] conducted a study in which the application of different criteria (diagnosability and identification of allele pools *sensu* Doyle [42], or lineage independence *sensu* BP&P [23]) and data types (DNA and morphology) resulted in nearly opposing species delimitation of mycoheterotrophic orchids (*Corallorhiza*), an example of data discordance that confounds many practicing systematists.

Others have argued that this instability reflects hypothesis-driven science and represents progress, and that taxonomic stability is unrealistic [43–45]. In many cases, morphology would have difficulty detecting cryptic species. Therefore, in the case of cryptic species, it is likely that only multilocus data could be used to properly delimit taxa and so morphological data and associated criteria cannot be used to undermine decisions rendered from coalescent methods. The result of these arguments is that substantial subjectivity remains and investigator bias can undermine the comparability across taxonomic groups and treatments.

Coalescent-based species delimitation is a step towards objectivity and comparability

Insofar as the goal of taxonomy is to arrive at robust estimates of species identity, then repeatability in species delimitation is essential, while understanding that species are hypotheses subject to change with new data and discoveries. It is important to note that coalescent-based species delimitation does not fall under the family of phylogenetic species concepts and does not require character diagnosability, a criterion often associated with investigator bias. Importantly, as these methods treat data equally among all living things, the inferred species are immediately comparable; that is, there is essentially a *de facto* standardization in coalescent-based species delimitation methods. The results should also be robust and the data are highly recyclable; adding additional samples (perhaps from new populations, species, or sequences) will build upon the species inferences produced from previous analyses.

Limitations of coalescent-based species delimitation: feasibility, assumptions, selection, and sampling

From a practical standpoint, coalescent-based species delimitation can be difficult for some researchers for whom collection of multilocus data is impractical. First, the method cannot be used for fossil taxa or taxa that lack suitable genetic resources. Second, collecting multilocus data can still be prohibitively expensive for some researchers, thereby preventing their adoption of coalescent methods. Unfortunately, this can be the case in regions that urgently need these methods, including many tropical regions that harbor exceptional cryptic diversity and that have urgent conservation needs (e.g., [46]). That said, it is important to note that we view coalescent-based species delimitation as an additional tool that can supplement or fine-tune understanding of species diversity; we expect

that traditional morphological approaches will, by necessity, continue to be the basis for the lion's share of new species descriptions.

Methodological limitations and considerations

Coalescent techniques could miss instances of very recent speciation caused by selection at a few loci. Indeed, neutrality is an assumption for most methods that use genealogies to investigate species histories. Speciation caused by selection at a few loci would be difficult to detect because the neutral markers that are typically targeted for study might not carry any record of species divergence [47,48]. Similarly, due to contrasting effects of selection and drift, rates of phenotypic and (neutral) molecular divergence are often discordant: on the one hand, morphological conservatism can mask deep molecular divergence and, on the other, rapid and repeated evolution of adaptive phenotypic traits can result in incorrect species delimitation if these same traits are used in alpha taxonomy. Second, gene flow is still unaccounted for in coalescent models of species delimitation, although Zhang *et al.* [24] found that some gene flow (0.1 migrants per generation) does not significantly affect the accuracy of BP&P (Box 2). Nevertheless, an important avenue for future research will be to incorporate gene flow thresholds into coalescent models of species delimitation. We note that the conditions under which coalescent-based species delimitation is expected to perform poorly (e.g., speciation caused by selection at a few loci) should result in underestimation of species diversity rather than oversplitting.

Sampling and guide trees

Sampling of both individuals and markers is an immediate consideration for any species delimitation method [10,49], including coalescent-based species delimitation. Sufficient intraspecific sampling is necessary for each species to represent geographic distributions and genetic diversity. In particular, it is important to avoid false positives from limited geographic sampling; this might occur, for instance, with sampling at the ends of strong spatial gradients of genetic diversity (isolation-by-distance). Furthermore, all coalescent-based species delimitation methods should use a multilocus data set to avoid the idiosyncratic histories of gene trees, which can differ because of several processes, including incomplete lineage sorting [50]. Simulations suggest that over a reasonably broad set of divergence histories, a modest number of individuals (5–10) and independent loci (5–10) provide considerable power for correct species delimitation [24] (Box 2). Further studies using empirical data will help elucidate the necessary individual and marker sampling, as well as parameter sensitivity, to conduct robust coalescent-based species delimitation.

Several coalescent-based species delimitation methods require a 'guide tree' that limits the tree space required to find the optimal species delimitation model (Box 2). Despite reducing computational burden, supplying a misspecified guide tree can overestimate the number of species (Box 2; [29]). Ideally, coalescent-based species delimitation methods will simultaneously delimit species and infer species trees [21]. Finally, comparisons between different methodologies (e.g., maximum likelihood and Bayesian) are necessary to measure the consistency between different coalescent approaches for species delimitation.

Box 3. The role of statistical species delimitation in taxonomy

Although some taxonomists might express some trepidation that coalescent-based species delimitation will result in radical changes to traditional taxonomic research and species description, we suggest that these fears are unfounded. For example, according to the International Code of Zoological Nomenclature [63], a valid species description requires only a character description that is adequate to identify the holotype specimen; no formal diagnosis is required. In practice, most modern species descriptions include a character-based diagnosis [64], and we are entirely supportive of this convention. However, we suggest that species descriptions would benefit from the inclusion of coalescent-based definitions when they are available, and we suggest further that they should serve the same purpose as a standard diagnosis when the species in question is not diagnosable on the basis of morphology alone [27,65]. The logic behind this suggestion is that a morphological diagnosis is taken as evidence of independent lineage status or lack of gene flow; a coalescent-based diagnosis is based on a more direct assessment of gene flow through genetic analysis. We envision coalescent-based diagnoses taking the following form: Species A is composed of all populations in Region X, which cluster as a lineage distinct from all other populations according to coalescent-based genetic analysis of multiple genetic loci.

Concluding remarks

There have been tremendous advances in analytical methods to study speciation, species delimitation, and species relationships, many of which are founded on coalescent theory. Along with the ability to collect large multilocus data sets, researchers can implement coalescent models to identify the evolutionary processes that contribute to speciation. It is perhaps this framework that has fueled a sharpened focus of speciation studies that aim to understand the processes of lineage divergence [51]. Nevertheless, there is a need (perhaps more than ever during the contemporary biodiversity crisis) to keep pace with the naming of new species and so formally recognize their existence (Box 3). Along with other tools and data types, coalescent-based species delimitation should play an important role in an integrative taxonomy that emphasizes the identification of species and the processes that have promoted lineage diversification.

Acknowledgement

The authors would like to thank the editor and two anonymous reviewers for their insightful comments that have improved the quality of this manuscript.

References

1 de Queiroz, K. (2007) Species concepts and species delimitation. *Syst. Biol.* 56, 879–886
2 Mace, G.M. (2004) The role of taxonomy in species conservation. *Philos. Trans. R. Soc. B* 359, 711–719
3 Sites, J.W. and Marshall, J.C. (2004) Operational criteria for delimiting species. *Annu. Rev. Ecol. Evol. Syst.* 35, 199–227
4 Agapow, P.M. *et al.* (2004) The impact of species concept on biodiversity studies. *Q. Rev. Biol.* 79, 161–179
5 Blackburn, T.M. and Gaston, K.J. (1998) Some methodological issues in macroecology. *Am. Nat.* 151, 68–83
6 Cracraft, J. (2002) The seven great questions of systematic biology: an essential foundation for conservation and the sustainable use of biodiversity. *Ann. Mo. Bot. Gard.* 89, 127–144

7 Padial, J.M. *et al.* (2010) The integrative future of taxonomy. *Front. Zool.* 7, 16
8 Schlick-Steiner, B.C. *et al.* (2010) Integrative taxonomy: a multisource approach to exploring biodiversity. *Annu. Rev. Entomol.* 55, 421–438
9 Leaché, A.D. *et al.* (2009) Quantifying ecological, morphological, and genetic variation to delimit species in the coast horned lizard species complex (*Phrynosoma*). *Proc. Natl. Acad. Sci. U.S.A.* 106, 12418–12423
10 Dayrat, B. (2005) Towards integrative taxonomy. *Biol. J. Linn. Soc.* 85, 407–415
11 Wiens, J.J. and Penkrot, T.A. (2002) Delimiting species using DNA and morphological variation and discordant species limits in spiny lizards (*Sceloporus*). *Syst. Biol.* 51, 69–91
12 Yeates, D.K. *et al.* (2011) Integrative taxonomy, or iterative taxonomy? *Syst. Entomol.* 36, 209–217
13 Hey, J. (2010) The divergence of chimpanzee species and subspecies as revealed in multipopulation isolation-with-migration analyses. *Mol. Biol. Evol.* 27, 921–933
14 Edwards, S. (2009) Is a new and general theory of molecular systematics emerging? *Evolution* 63, 1–19
15 Pinho, C. and Hey, J. (2010) Divergence with gene flow: models and data. *Annu. Rev. Ecol. Evol. Syst.* 41, 215–230
16 McCormack, J. *et al.* (2009) Maximum likelihood estimates of species trees: how accuracy of phylogenetic inference depends upon the divergence history and sampling design. *Syst. Biol.* 58, 501–508
17 Rannala, B. and Yang, Z. (2003) Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* 164, 1645–1656
18 Liu, L. *et al.* (2009) Coalescent methods for estimating phylogenetic trees. *Mol. Phylogenet. Evol.* 53, 320–328
19 Knowles, L.L. and Kubatko, L.S., eds (2010) *Estimating Species Trees: Practical and Theoretical Aspects*, Wiley Blackwell
20 Knowles, L.L. and Carstens, B.C. (2007) Delimiting species without monophyletic gene trees. *Syst. Biol.* 56, 887–895
21 O'Meara, B.C. (2010) New heuristic methods for joint species delimitation and species tree inference. *Syst. Biol.* 59, 59–73
22 Carstens, B.C. and Dewey, T.A. (2010) Species delimitation using a combined coalescent and information-theoretic approach: an example from North American *Myotis* bats. *Syst. Biol.* 59, 400–414
23 Yang, Z. and Rannala, B. (2010) Bayesian species delimitation using multilocus sequence data. *Proc. Natl. Acad. Sci. U.S.A.* 107, 9264–9269
24 Zhang, C. *et al.* (2011) Evaluation of a Bayesian coalescent method of species delimitation. *Syst. Biol.* 60, 747–761
25 Hudson, R.R. and Coyne, J.A. (2002) Mathematical consequences of the genealogical species concept. *Evolution* 56, 1557–1565
26 Hudson, R.R. (1990) Gene genealogies and the coalescent process. *Oxford Surv. Evol. Biol.* 7, 1–44
27 Leaché, A.D. and Fujita, M.K. (2010) Bayesian species delimitation in West African forest geckos (*Hemidactylus fasciatus*). *Proc. Biol. Sci. B* 277, 3071–3077
28 Burbrink, F.T. *et al.* (2011) Speciation at the Mogollon Rim in the Arizona Mountain Kingsnake (*Lampropeltis pyromelana*). *Mol. Phylogenet. Evol.* 60, 445–454
29 Setiadi, M.I. *et al.* (2011) Adaptive radiation and ecological opportunity in Sulawesi and Philippine fanged frog (*Limnonectes*) communities. *Am. Nat.* 178, 221–240
30 Degnan, J.H. and Salter, L.A. (2005) Gene tree distributions under the coalescent process. *Evolution* 59, 24–37
31 Ence, D. and Carstens, B. (2010) SpedeSTEM: a rapid and accurate method for species delimitation. *Mol. Ecol. Resour.* 24, 473–480
32 Pons, J. *et al.* (2006) Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *Syst. Biol.* 55, 595–609
33 Hey, J. (2006) Recent advances in assessing gene flow between diverging populations and species. *Curr. Opin. Genet. Dev.* 16, 592–596
34 Hey, J. (2010) Isolation with migration models for more than two populations. *Mol. Biol. Evol.* 27, 905–920
35 Gaggiotti, O.E. (2011) Making inferences about speciation using sophisticated statistical genetics methods: look before you leap. *Mol. Ecol.* 20, 2229–2232
36 Strasburg, J.L. and Rieseberg, L.H. (2011) Interpreting the estimated timing of migration events between hybridizing species. *Mol. Ecol.* 20, 2353–2366
37 Hoskin, C.J. *et al.* (2011) Persistence in peripheral refugia promotes phenotypic divergence and speciation in a rainforest frog. *Am. Nat.* 178, 561–578
38 Coyne, J.A. and Orr, H.A. (2004) *Speciation*, Sinauer Associates
39 Isaac, N.J. *et al.* (2004) Taxonomic inflation: its influence on macroecology and conservation. *Trends Ecol. Evol.* 19, 464–469
40 Tattersall, I. (2007) Madagascar's lemurs: cryptic diversity or taxonomic inflation? *Evol. Anthropol.* 16, 12–23
41 Barrett, C.F. and Freudenstein, J.V. (2011) An integrative approach to delimiting species in a rare but widespread mycoheterotrophic orchid. *Mol. Ecol.* 20, 2771–2786
42 Doyle, J.J. (1995) The irrelevance of allele tree topologies for species delimitation, and a non-topological alternative. *Syst. Biol.* 20, 574–588
43 Sangster, G. (2009) Increasing numbers of bird species result from taxonomic progress, not taxonomic inflation. *Proc. Biol. Sci. B* 276, 3185–3191
44 Padial, J.M. and De la Riva, I. (2006) Taxonomic inflation and the stability of species lists: the perils of ostrich's behavior. *Syst. Biol.* 55, 859–867
45 Sangster, G. (2000) Taxonomic stability and avian extinctions. *Conserv. Biol.* 14, 579–581
46 Funk, W.C. *et al.* (2011) High levels of cryptic species diversity uncovered in Amazonian frogs. *Proc. Biol. Sci. B.* http://dx.doi.org/10.1098/rspb.2011.1653
47 Wu, C-I. and Ting, C-T. (2004) Genes and speciation. *Nat. Rev. Genet.* 5, 114–122
48 Hickerson, M. *et al.* (2006) DNA barcoding will often fail to discover new animal species over broad parameter space. *Syst. Biol.* 55, 729–739
49 Valdecasas, A.G. *et al.* (2008) 'Integrative taxonomy' then and now: a response to Dayrat (2005). *Biol. J. Linn. Soc.* 93, 211–216
50 Maddison, W.P. (1997) Gene trees in species trees. *Syst. Biol.* 46, 523–536
51 Hart, M.W. (2010) The species concept as an emergent property of population biology. *Evolution* 65, 613–616
52 Mayr, E. (1995) Species, classification, and evolution. In *Biodiversity and Evolution* (Arai, R. *et al.*, eds), pp. 3–12, National Science Museum Foundation
53 Wiley, E.O. (1978) The Evolutionary Species Concept reconsidered. *Syst. Zool.* 21, 17–26
54 de Queiroz, K. (1998) The general lineage concept of species, species criteria, and the process of speciation: a conceptual unification and terminological recommendations. In *Endless Forms: Species and Speciation* (Howard, D.J. and Berlocher, S.H., eds), pp. 57–75, Oxford University Press
55 Cracraft, J. (1989) Speciation and its ontology: The empirical consequences of alternative species concepts for understanding patterns and processes of differentiation. In *Speciation and Its Consequences* (Otte, D. and Endler, J.A., eds), pp. 28–59, Sinauer Associates
56 Kuhner, M.K. (2009) Coalescent genealogy samplers: windows into population history. *Trends Ecol. Evol.* 24, 86–93
57 Kingman, J.F.C. (1982) On the genealogy of large populations. *J. Appl. Probab.* 19, 27–43
58 Kingman, J.F.C. (1982) The coalescent. *Stoch. Proc. Appl.* 13, 235–248
59 Kingman, J.F.C. (2000) Origins of the coalescent. 1974–1982. *Genetics* 156, 1461–1463
60 Wright, S. (1943) Isolation by distance. *Genetics* 28, 114–138
61 Monaghan, M.T. *et al.* (2009) Accelerated species inventory on Madagascar using coalescent-based models of species delineation. *Syst. Biol.* 58, 298–311
62 O'Meara, B.C. *et al.* (2006) Testing for different rates of continuous trait evolution using likelihood. *Evolution* 60, 922–933
63 International Commission of Zoological Nomenclature (1999) *International Code of Zoological Nomenclature*, (4th edn), International Trust for Zoological Nomenclature
64 Bauer, A. *et al.* (2010) Availability of new Bayesian-delimited gecko names and the importance of character-based species descriptions. *Proc. Biol. Sci. B* 278, 490–492
65 Fujita, M. and Leaché, A. *et al.* (2011) A coalescent perspective on delimiting and naming species: a reply to Bauer. *Proc. Biol. Sci. B* 278, 493–495
66 Fuchs, J. *et al.* (2011) Diversification across an altitudinal gradient in the Tiny Greenbul (*Phyllastrephus debilis*) from the Eastern Arc Mountains of Africa. *BMC Evol. Biol.* 11, 117

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

68 Zhou, W-W. *et al.* (2012) Speciation in the *Rana chensinensis* species complex and its relationship to the uplift of the Qinghai-Tibetan Plateau. *Mol. Ecol.* 21, 960–973

69 Ceccarelli, F.S. *et al.* (2012) Species identification in the taxonomically neglected, highly diverse, neotropical parasitoid wasp genus *Notiospathius* (Braconidae: Doryctinae) based on an integrative molecular and morphological approach. *Mol. Phylogenet. Evol.* 62, 485–495

70 Leavitt, S.D. *et al.* (2011) Species delimitation in taxonomically difficult lichen-forming fungi: an example from morphologically and chemically diverse *Xanthoparmelia* (Parmeliaceae) in North America. *Mol. Phylogenet. Evol* 60, 317–332

71 Niemiller, M.L. *et al.* (2012) Delimiting species using multilocus data: diagnosing cryptic diversity in the Southern Cavefish, *Typhlichthys subterraneus* (Teleostei: Amblyopsidae). *Evolution* 66, 846–866

72 Camargo, A. *et al.* (2012) Species delimitation with ABC and other coalescent-based methods: a test of accuracy with simulations and an empirical example with lizards of the *Liolaemus darwinii* complex (Squamata: Liolaemidae). *Evolution* http://dx.doi.org/10.1111/j.1558-5646.2012.01640.x