Opinion

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.

http://dx.doi.org/10.1016/j.tree.2012.04.012

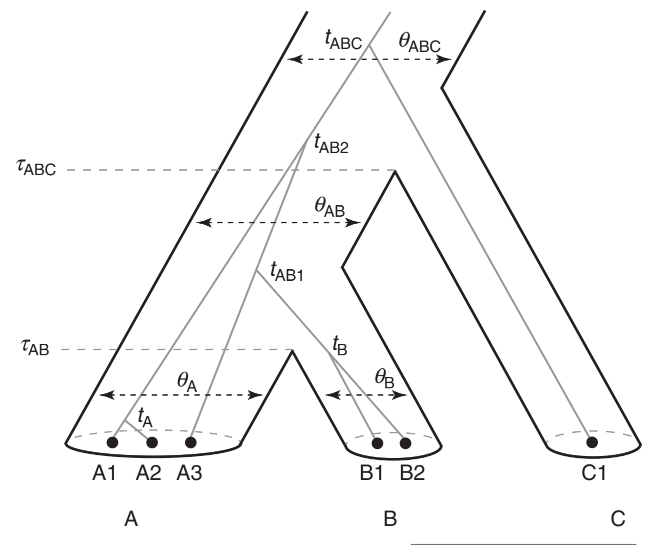
characters regardless of their significance or type under the phylogenetic species concept (morphological or molecular).

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 [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.



TRENDS in Ecology & Evolution

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.

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].