7 Species Delimitation Using Molecular Data

Megan L. Smith and Bryan C. Carstens

CONTENTS

7.1 INTRODUCTION

The Linnean shortfall, which describes the fact that only a small portion (1–10%) of extant species have been formally described (Brown and Lomolino 1998, but see Mora et al. 2011), is one of the most pressing challenges faced by the biological sciences. A lack of formal species description is likely to complicate conservation assessments (Beheregaray and Caccone 2007), bias evolutionary (Hortal et al. 2015), biogeographical (Whittaker et al. 2005), and ecological (e.g., Prada et al. 2014) studies, and have practical implications for disease ecology (e.g., Byrne et al. 2019), invasive species (Bickford et al. 2007), and wildlife management (Bickford et al. 2007). Amplifying this challenge is the ongoing loss of biodiversity (Costello et al. 2013), which makes addressing the Linnean shortfall a challenge with an inherent expiration date. For several decades, molecular data have been viewed as having the potential to address the Linnean shortfall (e.g., Herbert et al. 2003). However, despite their promise for this application, molecular data have a turbulent history of application in species delimitation, one that is complicated by researcher biases, a clear lack of best practices, and the varying information content of the data itself. While we do not hope to solve these problems in this chapter, we do hope that our discussion of the challenges inherent in delimiting species with genetic data will help researchers adopt useful strategies for practice.

7.2 MOLECULAR DATA AND THEIR INFLUENCE ON SPECIES DELIMITATION

Molecular data are now ubiquitous in the biological sciences. While they are easy to collect at the species level (McCormack et al. 2013) and have become central to

DOI: 10.1201/9780367855604-9

many evolutionary and ecological applications, including species delimitation, the widespread adoption and application of these data required biologists to adjust their thinking in various disciplines. For example, phylogeny inference has become far more quantitative and statistical since molecular data became common; a change prompted both by the increasing size of phylogenetic datasets and the change in the nature of the characters that form the basis of phylogenetic inference (Scornavacca et al. 2020). Similarly, taxonomists have been required to adopt both conceptual and practical changes in their approach to data analysis once massive amounts of molecular data became available to augment the trait data that were traditionally used to delimit species. Perhaps the most important of these was related to perspective. While investigations into the species level necessarily occur at the interface between phylogenetics and population genetics, initial attempts to apply molecular data to the question of species boundaries came primarily from systematists who were trained in phylogenetic biology. Influential papers encouraged researchers to apply phylogenetic thinking to intraspecific variation in a geographic context (e.g., Avise et al. 1987); a suggestion that found a receptive audience in researchers with a background in systematic biology and led to the exploding popularity of phylogeography. Phylogenetic species concepts *sensu lato*, such as genealogical species concepts (e.g., Baum and Shaw 1995) or criteria based on fixed allelic differences (e.g., population aggregation analysis; Nixon and Wheeler 1992), may have been natural outcomes of early efforts to apply phylogeographic data to detect species limits, albeit outcomes that proved difficult to apply in practice (e.g., Palumbi et al. 2001). Two developments from different disciplines, coalescent theory (Kingman 1982) and conceptual work on species concepts (e.g., Mayden 1997; de Querioz 1998), led to a remarkable shift in how phylogeographic data were applied to the question of species limits.

Once multilocus sequence data became widely available in the early 2000s, researchers began to observe substantial incongruence in the inferred gene trees across sequenced loci (e.g., Funk and Omland 2003). After researchers had been encouraged to conceptualise intraspecific variation as the end point of phylogeny, the many discordant trees that they observed prompted new ways of thinking about the phylogenies that were inferred from sequence data collected in empirical systems. For example, the concept of the species tree was introduced (i.e., gene trees in species trees; Maddison 1997) to differentiate the phylogeny that can be estimated using individual genes' sequence data from the history of organismal diversification. Ultimately, it became more useful to think about phylogeny as a property that emerges from population-level processes because this enables incongruent empirical data to be modeled using coalescent theory (Kingman 1982). This radical shift, which began when the expectations of taxonomists met the realities of phylogeographic data, has resulted in the most substantive shift in systematic biology since the introduction of cladistic analysis.

Coalescent theory describes a stochastic model of the loss of alleles in a population via genetic drift. The broader implications of coalescent theory are relevant to species delimitation, although they were underappreciated until Hudson and Coyne (2002) described in detail the mathematical consequences of using a genealogical species concept. Their argument is as follows: at neutral loci, allelic variation that is present in a lineage at the time of speciation will gradually sort into monophyletic clades in the daughter lineages, but the rate at which this occurs is a property of the effective population size (*Ne*) of the parent lineage. While the expectation of the time required for this lineage sorting to occur is 4*Ne**generations (Kingman 1982), Hudson and Coyne demonstrate that there is considerable variance around this expectation; for example, it would take 9–12 *Ne**generations for 95% of sampled loci to be reciprocally monophyletic. Even for species with modest effective population sizes (say 50,000 individuals), taxonomists would not be able to delimit species using a phylogenetic or genealogical species concept that uses monophyly as a criterion for hundreds of thousands of generations after the speciation event has occurred, even in simple cases where a single ancestor forms two new species with no further diversification. Given that many species have larger effective population sizes and complex patterns of diversification that may include introgression, the implication of coalescent theory to species delimitation is clear: genealogical and phylogenetic species concepts are difficult to apply near the species level because genealogies may not reflect the actual species phylogeny. Unless taxonomists are willing to accept that evolutionary lineages that are effectively independent of one another (and may have been for a million years!) do not obtain species status until all of this ancestral variation has sorted via genetic drift, the stochastic realities of lineage sorting require population-level thinking. Coalescent theory presently serves as the statistical foundation of modern phylogeographic inference, but another conceptual development was needed for the potential applicability of coalescent theory to the question of species boundaries to become clear.

Mayden (1997) and de Queiroz (1998) introduced a fundamental shift in how biologists thought about species concepts. They argued that while species concepts disagreed about the criteria used to recognise species (i.e., morphological distinctiveness, reproductive isolation, monophyly), all concepts fundamentally envisioned species as independent evolutionary lineages at the population or metapopulation level. The general lineage concept, proposed by de Queiroz (2005), encouraged researchers to equate species to independent evolutionary lineages regardless of the method used to identify them as such. This outlook on species fits nicely with modern coalescent-based methods for delimiting species. The conceptual unification of this concept with coalescent theory began during a symposium on species delimitation organised by the Society of Systematic Biologists at the 2006 Evolution Annual Meeting in Stoneybrook, New York. Kevin de Queiroz presented a lecture on species concepts, outlined his general lineage concept, and mentioned how coalescent theory makes it possible to extend the general lineage concept into a unified species concept, where independent lineages can be recognised as species (de Queiroz 2007). In the same symposium, Lacey Knowles presented work that described a likelihood ratio test of lineage independence that enabled researchers to delimit species without relying on monophyletic gene trees (Knowles and Carstens 2007). This test utilised data simulated under the coalescent model where two lineages were independent and compared these data with those simulated under a model where the lineages were not independent. The lasting influence of this test has been felt in the general framework of the proposed statistical comparison (i.e., modeling the statistical fit of the data given two models, one where lineages are independent and one where lineages are combined), as many newer methods are based on statistical comparisons of species trees that include different groupings of putative species (Yang and Rannala 2010; Ence and Carstens 2011; Grummer et al. 2014; Leaché et al. 2014a).

Under the unified species concept, independently evolving lineages can be delimited as distinct species. While on a superficial level, this may appear to eliminate subjective decisions from the process of species delimitation, this definition of species is likely to result in over-splitting under some models of speciation when population genetic structure is present (Sukumaran and Knowles 2017). As more genomic data are gathered, many algorithms become more effective at identifying population genetic structure, highlighting the need for sanity checks in species delimitation (e.g., Jackson et al. 2017b), where the intuition of the taxonomist is considered. Given that taxonomists generally do not wish to name all populations as species due to practical considerations (Zachos et al. 2020), additional considerations may be required. For example, Zachos et al. (2020) distinguish between the process of grouping organisms into 'species taxa' and making the subjective decision of whether these taxa should be ranked as species in the Linnaean classification system. To the extent that researchers do not view each independently evolving population as warranting species recognition, the coalescent and related models and methods cannot address this second aspect of species delineation, which requires taxonomic expertise and subjective thought. Regardless, coalescent-based approaches to species delimitation provide valuable information about the status and history of species taxa, and this information can serve as the basis for integrative taxonomic efforts.

7.3 PRACTICAL CONSIDERATIONS IN SPECIES DELIMITATION

Genomes accumulate nucleotide substitutions at a rate that is influenced by demographic processes (i.e., gene flow, population size change), natural selection, and recombination as the population evolves over time. The pattern of nucleotide variation across individuals sampled from multiple lineages within a species complex will retain information about the recent history of that complex, and any method used to delimit species with molecular data will attempt to access this information. However, decisions made by researchers can potentially influence the results of a species delimitation analysis. Perhaps the most important factor to consider at the outset of an investigation is the sample design. As with any source of inference, the strength of the signal is likely to be positively correlated with the size of the dataset, although comprehensive evaluations of this relationship have not been conducted for all methods. Note that the size of the dataset is best measured on two axes, the number of loci and the number of samples, as the former determines how many independent realizations of the coalescent process are sampled, and the latter determines how well the allelic and/or genotype frequencies of the sample match the actual values from the empirical system. An equally important consideration is to document what information exists about potential division of individuals within a nominal taxon. For example, are there described subspecies? Allopatric populations? Evident environmental gradients that could serve to divide a population? Any of these factors could serve to guide researchers as they acquire samples and choose individuals for sequencing. They can also influence the types of analyses that are chosen by researchers once the genetic data are collected. Related to each of these is the question of what type of genetic data to collect. Data can be collected on a locus-by-locus basis using polymerase chain reaction and Sanger sequencing methods, but this can be tedious work. Next-generation-sequencing technologies enable researchers to collect data from thousands of loci, either in the form of sequence capture techniques (e.g., Faircloth et al. 2012) or using restriction-digest approaches (Miller et al. 2007). Notably, the technology used for sequencing also affects downstream methodological choices, as some methods are designed for use with single nucleotide polymorphism (SNP) data while others are designed for sequence data.

Carstens et al. (2013) proposed that researchers conceptualise species delimitation as a two-step process. Since many taxa lack obvious partitions, such as described subspecies or populations that are clearly allopatric, the first analyses for many investigations should be discovery approaches that do not require samples to be partitioned prior to analysis. Discovery approaches include methods such as STRUCTURE (Pritchard et al. 2000) or ADMIXTURE (Alexander and Lange 2011), which implement algorithms that cluster samples into groups based on some criterion, such as minimizing Hardy–Weinberg disequilibrium, as well as methods based on genetic distances (e.g., Automatic Barcode Gap Discovery [ABGD]) and those based on gene tree diversification (e.g., Generalized Mixed Yule Coalescent [GMYC)]; Pons et al. 2006). The key information obtained via the use of these methods is a division of the samples into two or more groups that can serve as the basis for the next step of species delimitation. Methods that require samples to be partitioned prior to analysis, such as species-tree-based programs (e.g., BPP (Yang and Rannala 2010), BFD* (Leaché et al. 2014a) and those based on demographic models (e.g., delimitR (Smith and Carstens 2020), PHRAPL (Jackson et al. 2017a), work on some level by comparing the probability of the data given the model where a key component of the model is the assignment of samples to each putative lineage. See Box 7.1 for additional examples of species discovery and species validation approaches, and Rannala and Yang (2020) for a recent review of several approaches. Note that some investigations omit the first step (i.e., discovery) because there are *a priori* groupings of samples (e.g., Morales et al. 2018). Others conduct both steps sequentially, with sample partitions in the validation stage informed by the clustering of samples from the discovery phase (e.g., Leaché and Fujita 2010). One challenge to this approach is how to treat samples where there is evidence of admixture (i.e., genetic ancestry in an individual sample that can be traced to two or more populations). Some researchers remove these samples from the validation analysis, since we know that gene flow can interfere with species tree estimation (Eckert and Carstens 2008; Leaché et al. 2014b). However, this should not be done if divergence with gene flow models is included, because it could presumably bias the validation analysis. Notably, some discovery and most validation approaches rely on particular models of the speciation process, and the choice and application of such models can greatly impact the results of species delimitation analyses. In the following, we discuss popular models employed in species delimitation and their potential shortfalls.

7.4 THE IMPORTANCE OF MODELS

The Multispecies Coalescent Model (MSCM) addresses the difficulties of applying genealogical and phylogenetic species concepts near the species level by directly modeling the coalescent process (Knowles and Carstens 2007). By modeling the causes of incomplete lineage sorting, methods based on the MSCM allow researchers to go beyond a monophyly criterion and address whether observed genealogies are consistent with different numbers of species. Species delimitation methods based on the MSCM have proliferated since its development (e.g., Yang and Rannala 2010; Ence and Carstens 2011; Leaché et al. 2014a), and many allow researchers to use genetic data to assess the probability of different numbers of species.

While the MSCM is undoubtedly a powerful approach to delimiting species with genetic data, it is not without its limitations. As with any model-based approach, the MSCM makes certain assumptions, which if violated, may render the results of species delimitation under the model unreliable. For example, MSCM methods rely on *a priori* definitions of populations or putative species (i.e., they are validation approaches). When populations are estimated using genetic data from sparse sampling, geographic clines can be mistaken for discrete populations, and this can lead to over-splitting under the MSCM (Chambers and Hillis 2020). The MSCM also assumes that speciation is an instantaneous process, and recent results demonstrate that when this is not the case, but rather, speciation is protracted, the MSCM will over-split, delimiting population structure as distinct species (Sukumaran and Knowles 2017).

Perhaps the best-known violation of the MSCM is the presence of gene flow between populations or species. The MSCM models only genetic divergence and does not consider the possibility of post-divergence gene flow between lineages. However, gene flow is thought to be important in speciation, and is implicated in many empirical systems, including *Myotis* bats (Morales et al. 2017) and flowering plants on Lord Howe Island (Papadopulos et al. 2011). Simulation studies demonstrate that ignoring gene flow causes overestimates of population sizes and underestimates of divergence times under the MSCM (Leaché et al. 2014b), and BPP (Yang and Rannala 2010) may delimit populations as species even when levels of gene flow between populations are high (Jackson et al. 2017b; Leaché et al. 2019). However, recent attempts to use more appropriate models that consider gene flow, for example, have improved error rates and led to more meaningful species delimitation (Jackson et al. 2017b; Leaché et al. 2019; Smith and Carstens 2020).

Considering these results, it is clear that the choice of appropriate models is essential for species delimitation using genetic data. While choosing an appropriate model is not always straightforward, recent advances in simulation-based approaches provide a promising avenue for species delimitation. Software for simulating large genomic datasets under models including divergence, gene flow, and population size changes has improved vastly in speed and computational efficiency in recent years (e.g., Excoffier et al. 2013). More recently, the development of tree-sequence recording has permitted simulating tens of thousands of replicates of genomic datasets under models that include selection as well as demographic processes (Kelleher et al. 2016; Haller et al. 2019). The ability to simulate many replicates of large genomic datasets under various models permits researchers to then use either Approximate Bayesian Computation (e.g., Camargo et al. 2012) or machine learning approaches (e.g., Pei et al. 2018; da Fonseca et al. 2020; Smith and Carstens 2020) to find the model that generates data most similar to the observed data. By combining new powerful simulation approaches with machine learning, researchers are effectively limited only by their creativity and the computational resources available when designing a model set. It should be noted that larger model sets inevitably lead to increased difficulties in differentiating among models (Pelletier and Carstens 2014), even when machine learning approaches are employed (Smith and Carstens 2020), because the distance in model space between these models decreases. While choosing models to compare *a priori* requires researchers to make decisions about which processes are likely to be important in their focal system, researchers are implicitly making such decisions when they utilise tools that explore a limited number of processes, like methods based on the MSCM. By using approaches that ignore processes like gene flow, researchers assume that those processes are not important. Leaving the power to determine which models to test to researchers who are experts in their study system takes advantage of their knowledge of the taxa, similarly to traditional taxonomic investigations. We view this aspect of defining a model set as a positive aspect of species delimitation, but it could lead to biases when all models considered are a poor fit to the data, or when researchers limit their model set too strictly to match misleading *a priori* knowledge of the study system. Tools to directly assess model fit, like posterior predictive simulations (e.g., Fonseca et al. 2021) or composite likelihood ratio tests (e.g., Excoffier et al. 2013), may help researchers to diagnose such situations.

Evaluating a broader array of models not only prevents erroneous inference due to model violations but also may provide novel insights into the processes driving speciation. Different modes of speciation involve different demographic and selective processes, and by modeling these processes directly, researchers may be able to address not only how many species are present but also the processes that gave rise to these species. For example, gene flow and directional selection may play an important role when divergent ecological selection drives speciation in sympatry (Coyne and Orr 2004) or when reinforcement drives speciation between once isolated populations (e.g., in Phlox; Hopkins and Rausher 2011). On the other hand, when speciation occurs in allopatry, genetic drift, natural selection, or some combination of the two may drive divergence (Coyne and Orr 2004). By designing models based on predictions about the mode of speciation and then using machine learning or other approaches to determine which of those models best reproduces the observed data, researchers can identify the most likely mode of speciation in their system as well as the number of species present. Of course, doing so requires researchers to explicitly state an operational species concept.

While model selection itself provides insights into the number of species and the process of speciation, it also permits more accurate parameter estimation (Thomé and Carstens 2016). When parameters are accurately estimated, they provide insight into the magnitude of divergence and gene flow between populations, essential parameters for determining whether populations represent independent evolutionary units (Rannala and Yang 2020). Additionally, parameter estimates may lend insight into the correspondence of speciation events with geologic and climatic processes.

For example, more precise parameter estimates might provide resolution on how the Pleistocene glaciations impacted speciation, or the extent to which divergence can occur with gene flow.

7.5 PROSPECTS FOR THE FUTURE

As genomic data become increasingly available near the species level, opportunities to connect process to pattern in taxonomy are ripe. Already, with the rise in popularity of the Multispecies Coalescent Model in species delimitation, taxonomists have begun to embrace the link between population-level and species-level processes and to use models based explicitly on these processes to evaluate species delimitation hypotheses. With further advances in the nature of genetic data and in the models and computational tools available to taxonomists, we believe that the field of molecular-based species delimitation will rely increasingly on evolutionary genetics, linking genetic variation to the specific evolutionary processes that drove speciation. As our understanding of the importance of selective processes on structuring genetic variation increases, future developments may take advantage of this and model speciation as the complex interplay of neutral and selective processes that it is. This should shed additional light on the history of populations and prove invaluable to taxonomists when evaluating the species status of lineages.

Although molecular data held (and continue to hold) great promise for species delimitation, the importance of other data sources, including morphological and ecological data, should not be overlooked. The call for so-called integrative taxonomy (Weins and Penkrot 2002; Sites and Marshall 2004; Dayrat 2005; Winker 2009; Padial et al. 2010; Schlick-Steiner et al. 2010; Yeates et al. 2011) highlights the potential benefits of combining data types when inferring species boundaries. Phenotypic and ecological data have further power to illuminate the process of speciation and to allow researchers to distinguish among population- and species-level variation (Cadena and Zapata 2021). Further, following up molecular studies with phenotypic and ecological investigations may provide diagnostic characters, without which the recognition of distinct species in the field by conservation biologists and ecologists is impossible. In short, the availability of molecular data does not eliminate the need for phenotypic and ecological data. Rather, by combining molecular, phenotypic, and ecological data, researchers can better understand how genetic divergence, phenotypic divergence, and ecological divergence differ across putative species, which should not only inform taxonomic efforts but also shed light on the processes of speciation and diversification in a way that either data type on its own could not (Winker 2009; Cadena and Zapata 2021).

7.6 SPECIES DESCRIPTION IS A NECESSARY LAST STEP IN A DELIMITATION ANALYSIS

Although species delimitation studies have flourished in recent years, a remarkably small number of those studies follow up with the description of delimited species. Pante et al. (2015) found that ~47% of integrative taxonomy studies published between 2008 and 2013 did not describe newly delimited species. Without following up with taxonomic revisions, species delimitation studies hardly address the Linnean shortfall that we often claim as the motivation for the field. A variety of factors likely contribute to the failure of many studies to describe the species that are inferred by species delimitation investigations. First, the lack of species description might signal a lack of confidence in the results – an unwillingness to commit to the delimited species (Pante et al. 2015). As mentioned earlier, taxonomists may not view all independently evolving populations as warranting formal species recognition (Zachos et al. 2020), and thus, some lack of species description may only reflect that delimited entities do not meet a particular taxonomist's criteria for describing a new species. Second, it may be that researchers plan to follow up the delimitation results with further morphological, behavioral, or other types of taxonomic investigation (Pante et al. 2015), particularly since describing species with non-traditional characters (e.g., molecular characters) remains difficult (Satler et al. 2013). Due to the ease of collecting molecular data, it could be that these investigations are more easily completed and published than integrative work that incorporates multiple data types. Third, researchers could feel inhibited by the formal rules associated with taxonomic description in the Zoological or Botanical Codes, potentially due to a lack of training (Pante et al. 2015; Pearson et al. 2011). Finally, there are likely to be fewer professional rewards for publishing in the taxonomic literature, for while these papers have a long potential history of citation, they likely will receive less notice in the immediate future than works published in the general interest literature. The competition for space is fierce for journals in the latter category, and editors might balk at devoting several pages to species description; thus, general interest journals rarely publish taxonomic revisions (Pante et al. 2015; Agnarsson and Kunter 2007). Pressure to publish in journals with high impact factors may therefore discourage authors from including species descriptions in their work. For species delimitation to address the Linnean shortfall, these issues must be addressed, so that discovered species are subsequently described. We urge funding panels to demand that proposals which include species delimitation also include species description. Furthermore, senior scientists need to be more vocal to their administration in highlighting the importance of species description, particularly when conducted by early career researchers.

BOX 7.1 Examples of Popular Species Discovery Approaches

Population Genetic Structure. STRUCTURE uses a Bayesian clustering approach to assign individuals to populations and estimate population allele frequencies (Pritchard et al. 2000). STRUCTURE assumes that markers are unlinked and that each population is in Hardy–Weinberg equilibrium (Pritchard et al. 2000). Recent advances have improved the computational efficiency of the approach used in STRUCTURE (Raj et al. 2014). structure is often combined with ad-hoc methods (e.g., Earl 2012; Evanno et al. 2005) to estimate the number of populations. Like STRUCTURE, structurama (Huelsenbeck et al. 2011) assumes

Hardy–Weinberg equilibrium, but it also uses a Dirichlet-process prior on the number of populations and reversible jump MCMC to allow simultaneous inference of population assignments and the number of populations.

Generalized Mixed Yule Coalescent (gmyc). The GMYC (Pons et al. 2006) takes ultrametric gene trees (i.e., rooted trees where a molecular clock has been enforced) as input. It then infers the transition point between branching events, corresponding speciation events (the Yule process), and branching events corresponding to allele coalescence within species (the coalescent process). Reid and Carstens (2012) introduced a Bayesian implementation of the GMYC (**b**gmyc), which takes as input a posterior distribution of gene trees and outputs posterior distributions of the number and composition of species.

Automatic Barcode Gap Discovery (abgd)**.** ABGD (Puillandre et al. 2012) takes as input short sequences and searches for a gap in the distribution of pairwise differences between sequences. ABGD requires that the user supply a prior maximum divergence of intraspecific diversity, and this value determines how finely ABGD divides individuals into species. Like the GMYC, ABGD is limited to single-locus data.

Multivariate Methods. Multivariate methods (e.g., methods based on Principal Components Analysis [PCA]) are powerful because they are fast and do not make assumptions about evolutionary models generating population and species divergence. Adegenet is a popular software package that combines PCA and Discriminant Analysis of Principal Components to assign individuals to populations (Jombart et al. 2010).

Machine Learning. Recently, Derkarabetian et al. (2019) applied a suite of machine learning approaches to perform species discovery analysis. They applied Random Forests, Variational Autoencoders, and t-Distributed Stochastic Neighbor Embedding to assign individuals to populations (or species) and found that they have high power to identify population structure. As with the multivariate methods described earlier, these approaches do not make assumptions about the evolutionary models generating population and species divergence.

BOX 7.2 Examples of Popular Species Validation Approaches

bpp**.** BPP (Yang and Rannala 2010) is a fully Bayesian approach to multilocus species delimitation. BPP takes as input DNA sequence data and uses reversible jump MCMC to estimate the species tree topology, the number of populations, genetic diversity, and population divergence. Recently, the model underlying BPP was extended to allow estimation of introgression probabilities (Flouris et al. 2019).

Bayes Factor Delimitation. Bayes Factors can be used to compare models when marginal likelihoods of the competing models are available and were initially used in the context of species delimitation by Carstens and Dewey (2010) and Grummer et al. (2014). The popular species delimitation software BFD* (Leaché et al. 2014a) applies Bayes Factor Delimitation to genome-wide SNP data by using SNAPP to calculate marginal likelihoods for all hierarchical arrangements of individuals into predefined populations and then comparing models using Bayes Factors.

phrapl**.** PHRAPL (Jackson et al. 2017a) takes as input gene trees and population assignments and can be applied to species delimitation (Jackson et al. 2017b). To compare different delimitation hypotheses, PHRAPL approximates the likelihood of models that differ in the number of species and the species history and compares models using information theory.

clades**.** CLADES (Pei et al. 2018) simulates data belonging to either the same or different species and then trains a support vector machine (SVM) to recognise whether two samples come from the same or different species. Using this SVM, CLADES can classify DNA sequence data sampled from populations as belonging to the same or different species. Finally, if there are more than two putative species, CLADES maximises the likelihood of the species status of all populations.

delimitR. delimitR (Smith and Carstens 2020) takes as input a multidimensional Site Frequency Spectrum (mSFS) and allows users to specify models including divergence, migration (primary or upon secondary contact), and population size changes. Then, fastsimcoal2 (Excoffier et al. 2013) is used to simulate mSFS under each model. Finally, a Random Forest classifier is constructed using the R package 'abcrf' (Pudlo et al. 2016) and used to select the best model and to estimate classification error rates under each model.

soda. SODA (Rabiee and Mirarab 2021) uses the algorithm of the popular species tree inference software ASTRAL (Mirarab et al. 2014; Zhang et al. 2018) and expected patterns of quartet frequencies to infer species boundaries.

REFERENCES

Agnarsson, I., and M. Kunter. 2007. Taxonomy in a changing world: Seeking solutions for a science in crisis. *Systematic Biology* 56:531–539.

Alexander, D. H., and K. Lange. 2011. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics* 12:246.

Avise, J. C., J. Arnold, R. M. Ball, E. Bermingham, T. Lamb, J. E. Neigel, C. A. Reeb, and N. C. Saunders. 1987. Intraspecific phylogeography: The mitochondrial DNA bridge between population genetics and systematics. *Annual Review of Ecology and Systematics* 18:489–522.

Baum, D. A., and K. L. Shaw. 1995. Genealogical perspectives on the species problem. *Experimental and Molecular Approaches to Plant Biosystematics* 53:123–124.

Beheregaray, L. B., and A. Caccone. 2007. Cryptic biodiversity in a changing world. *Journal of Biology* 6:9.

Bickford, D., D. J. Lohman, N. S. Sodhi, P. K. L. Ng, R. Meier, K. Winker, K. K. Ingram, and I. Das. 2007. Cryptic species as a window on diversity and conservation. *Trends in Ecology and Evolution* 22:148–155.

Brown, J. H., and M. V. Lomolino. 1998. *Bioegeography* (2nd ed.). Sunderland: Sinauer.

Byrne, A. Q., V. T. Vredenburg, A. Martel, F. Pasmans, R. C. Bell, D. C. Blackburn, M. C. Bletz, J. Bosch, C. J. Briggs, R. M. Brown, A. Catenazzi, M. Familiar López, R. Figueroa-Valenzuela, S. L. Ghose, J. R. Jaeger, A. J. Jani, M. Jirku, R. A. Knapp, A. Muñoz, D. M. Portik, C. L. Richards-Zawacki, H. Rockney, S. M. Rovito, T. Stark, H. Sulaeman, N. T. Tao, J. Voyles, A. W. Waddle, Z. Yuan, and E. B. Rosenblum. 2019. Cryptic diversity of a widespread global pathogen reveals expanded threats to amphibian conservation. *Proceedings of the National Academy of Sciences* 116:20382–20387.

Cadena, C. D., and F. Zapata. 2021. The genomic revolution and species delimitation in birds (and other organisms): Why phenotypes should not be overlooked. *Ornithology* 138:ukaa069.

Camargo, A., M. Morando, L. J. Avila, and J. W. Sites Jr. 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: International Journal of Organic Evolution* 66:2834–2849.

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

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

Chambers, E. A., and D. M. Hillis. 2020. The multispecies coalescent over-splits species in the case of geographically widespread taxa. *Systematic Biology* 69:184–193.

Costello, M. J., R. M. May, and N. E. Stork. 2013. Can we name Earth's species before they go extinct? *Science* 339:413–416.

Coyne, J., and H. Orr. 2004. *Speciation*. Sunderland: Sinauer.

da Fonseca, E. M., G. R. Colli, F. P. Werneck, and B. C. Carstens. 2020. Phylogeographic model selection using convolutional neural networks. *bioRxiv.* https://doi.org/10.1101/2020.09.11.291856

Dayrat, B. 2005. Towards integrative taxonomy. *Biological Journal of the Linnean Society* 85:407–417.

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*, eds. D. J. Howard and S. H. Berlocher, 57–75. Oxford: Oxford University Press.

de Queiroz, K. 2005. Different species problems and their solutions. *Bioessays* 26:67–70.

de Queiroz, K. 2007. Species Concepts and Species Delimitation. *Systematic Biology* 56:879–886.

Derkarabetian, S., S. Castillo, P. K. Koo, S. Ovchinnikov, and M. Hedin. 2019. A demonstration of unsupervised machine learning in species delimitation. *Molecular Phylogenetics and Evolution* 139:106562.

Earl, D. A. 2012. STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources* 4:359–361.

Eckert, A. J., and B. C. Carstens. 2008. Does gene flow destroy phylogenetic signal? The performance of three methods for estimating species phylogenies in the presence of gene flow. *Molecular Phylogenetics and Evolution* 49:832–842.

Ence, D. D., and B. C. Carstens. 2011. SpedeSTEM: A rapid and accurate method for species delimitation. *Molecular Ecology Resources* 11:473–480.

Evanno, G., S. Regnaut, and J. Goudet. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Molecular Ecology* 14:2611–2620.

Excoffier, L., I. Dupanloup, E. Huerta-Sánchez, V. C. Sousa, and M. Foll. 2013. Robust demographic inference from genomic and SNP data. *PLoS Genetics* 9:10.

Faircloth, B. C., J. E. McCormack, N. G. Crawford, M. G. Harvey, R. T. Brumfield, and T. C. Glenn. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology* 61:717–726.

Flouris, T., X. Jiao, B. Rannala, and Z. Yang. 2019. A Bayesian implementation of the multispecies coalescent model with introgression for phylogenomic analysis. *Molecular Biology and Evolution* 37:1211–1223.

Fonseca, E. M., D. J. Duckett, and B. C. Carstens. 2021. P2C2M. GMYC: An R package for assessing the utility of the Generalized Mixed Yule Coalescent model. *Methods in Ecology and Evolution* 12:487–493.

Funk, D. J., and K. E. Omland. 2003. Species-level paraphyly and polyphyly: Frequency, causes, and consequences, with insights from animal mitochondrial DNA. *Annual Review of Ecology, Evolution, and Systematics* 34:397–423.

Grummer, J. A., R. W. Bryson, and T. W. Reeder. 2014. Species delimitation using Bayes factors: Simulations and application to the *Sceloporus scalaris* species group. *Systematic Biology* 63:119–133.

Haller, B. C., J. Galloway, J. Kelleher, P. W. Messer, and P. L. Ralph. 2019. Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes. *Molecular Ecology Resources* 19:552–566.

Hebert, P. D. N., A. Cywinska, S. L. Ball, and J. R. deWaard. 2003. Biological identifications through DNA barcodes. *Proceedings of the Royal Society B* 270:313–321.

Hopkins, R., and M. D. Rausher. 2011. Identification of two genes causing reinforcement in the Texas wildflower *Phlox drummondii*. *Nature* 469:411–414.

Hortal, J., F. de Bello, J. A. F. Diniz-Filho, T. M. Lewinsohn, J. M. Lobo, and R. J. Ladle. 2015. Seven shortfalls that beset large-scale knowledge of biodiversity. *Annual Review of Ecology, Evolution, and Systematics* 46:523–549.

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

Huelsenbeck, J. P., P. Andolfatto, and E. T. Huelsenbeck. 2011. Structurama: Bayesian inference of population structure. *Evolutionary Bioinformatics* 7:EBO-S6761.

Jackson, N. D., A. E. Morales, B. C. Carstens, and B. C. O'Meara. 2017a. PHRAPL: Phylogeographic inference using approximate likelihoods. *Systematic Biology* 66:1045–1053.

Jackson, N. D., B. C. Carstens, A. E. Morales, and B. C. O'Meara. 2017b. Species delimitation with gene flow. *Systematic Biology* 66:799–812.

Jombart, T., S. Devillard, and F. Balloux. 2010. Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. *BMC Genetics* 11:1–15.

Kelleher, J., A. M. Etheridge, and G. McVean. 2016. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLOS Computational Biology* 12:e1004842.

Kingman, J. F. C. 1982. The coalescent. *Stochastic Processes and Their Applications* 13:235–248.

Knowles, L. L., and B. C. Carstens. 2007. Delimiting species without monophyletic gene trees. *Systematic Biology* 56:887–895.

Leaché, A. D., and M. K. Fujita. 2010. Bayesian species delimitation in West African forest geckos (*Hemidactylus fasciatus*). *Proceedings of the Royal Society B* 277:3071–3077.

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

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

Leaché, A. D., T. Zhu, B. Rannala, and Z. Yang. 2019. The spectre of too many species. *Systematic Biology* 68:168–181.

Maddison, W. P. 1997. Gene trees in species trees. *Systematic Biology* 46:523–536.

Mayden, R. L. 1997. A hierarchy of species concepts: The denouement in the saga of the species problem. In *Species: The units of diversity*, eds. M. F. Claridge, H. A. Dawah, and M. R. Wilson, 381–423. London: Chapman & Hall.

McCormack, J. E., S. M. Hird, A. J. Zellmer, B. C. Carstens, and R. T. Brumfield. 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution* 66:526–538.

Miller, M. R., J. P. Dunham, A. Amores, W. A. Cresko, and E. A. Johnson. 2007. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research* 17:240–248.

Mirarab, S., R. Reaz, Md. S. Bayzid, T. Zimmermann, M. S. Swenson, and T. Warnow. 2014. ASTRAL: Genome-scale coalescent-based species tree estimation. *Bioinformatics* 30:i541–i548.

Mora, C., D. P. Tittensor, S. Adl, A. G. B. Simpson, and B. Worm. 2011. How many species are there on earth and in the ocean? *PLoS Biology* 9:e1001127.

Morales, A. E., and B. C. Carstens. 2018. Evidence that *Myotic lucifigus* "subspecies" are five nonsister species, despite gene flow. *Systematic Biology* 67:756–769.

Morales, A. E., N. D. Jackson, T. A. Dewey, B. C. O'Meara, and B. C. Carstens. 2017. Speciation with gene flow in North American Myotis bats. *Systematic Biology* 66:440–452.

Nixon, K. C., and Q. D. Wheeler. 1992. Extinction and the origin of species. In *Extinction and phylogeny*, eds. M. J. Novacek and Q. D. Wheeler, 119–143. New York: Columbia University Press.

Padial, J. M., A. Miralles, I. De la Riva, and M. Vences. 2010. The integrative future of taxonomy. *Frontiers in Zoology* 7:6.

Palumbi, S. R., F. Cipriano, and M. P. Hare. 2001. Predicting nuclear gene coalescence from mitochondrial data: The three-times rule. *Evolution* 55:859–868.

Pante, E., C. Schoelinck, and N. Puillandre. 2015. From integrative taxonomy to species description: One step beyond. *Systematic Biology* 64:152–160.

Papadopulos, A. S., W. J. Baker, D. Crayn, R. K. Butlin, R. G. Kynast, I. Hutton, and V. Savolainen. 2011. Speciation with gene flow on Lord Howe Island. *Proceedings of the National Academy of Sciences* 108:13188–13193.

Pearson, D. L., A. L. Hamilton, and T. L. Erwin. 2011. Recovery plan for the endangered taxonomy profession. *BioScience* 61:58–63.

Pei, J., C. Chu, X. Li, B. Lu, and Y. Wu. 2018. CLADES: A classification-based machine learning method for species delimitation from population genetic data. *Molecular Ecology Resources* 18:1144–1156.

Pelletier, T. A., and B. C. Carstens. 2014. Model choice for phylogeographic inference using a large set of models. *Molecular Ecology* 23:3028–3043.

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. *Systematic Biology* 55:595–609.

Prada, C., S. E. McIlroy, D. M. Beltrán, D. J. Valint, S. A. Ford, M. E. Hellberg, and M. A. Coffroth. 2014. Cryptic diversity hides host and habitat specialization in a gorgonian-algal symbiosis. *Molecular Ecology* 23:3330–3340.

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

Pudlo, P., J. -M. Marin, A. Estoup, J. -M. Cornuet, M. Gautier, and C. P. Robert. 2016. Reliable ABC model choice via random forests. *Bioinformatics* 32:859–866.

Puillandre, N., A. Lambert, S. Brouillet, and G. Achaz. 2012. ABGD, automatic barcode gap discovery for primary species delimitation. *Molecular Ecology* 21:1864–1877.

Rabiee, M., and S. Mirarab. 2021. SODA: Multi-locus species delimitation using quartet frequencies. *Bioinformatics* 36:5623–5631.

Raj, A., M. Stephens, and J. K. Pritchard. 2014. FastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics* 197:573–589.

Rannala, B., and Z. Yang. 2020. Species delimitation. In *Phylogenetics in the genomic era*. eds. C. Scornavacca, F. Delsuc, N. Galtier, 5.5:1–5.5:18. Self-published.

Reid, N. M., and B. C. Carstens. 2012. Phylogenetic estimation error can decrease the accuracy of species delimitation: A Bayesian implementation of the general mixed Yule-coalescent model. *BMC Evolutionary Biology* 12:196.

Satler, J. D., B. C. Carstens, and M. Hedin. 2013. Multilocus species delimitation in a complex of morphologically conserved trapdoor spiders (Mygalomorphae, Antrodiaetidae, *Aliatypus*). *Systematic Biology* 62:805–823.

Schlick-Steiner, B. C., F. M. Steiner, B. Seifert, C. Stauffer, E. Christian, and R. H. Crozier. 2010. Integrative taxonomy: A multisource approach to exploring biodiversity. *Annual Review of Entomology* 55:421–438.

Scornavacca, C., F. Delsuc, and N. Galtier. 2020. *Phylogenetics in the genomic era*. Open access book available from https://hal.inria.fr/PGE/.

Sites, J. W., and J. C. Marshall. 2004. Operational criteria for delimiting species. *Annual Review of Ecology, Evolution, and Systematics* 35:199–227.

Smith, M. L., and B. C. Carstens. 2020. Process-based species delimitation leads to identification of more biologically relevant species. *Evolution* 74:216–229.

Sukumaran, J., and L. L. Knowles. 2017. Multispecies coalescent delimits structure, not species. *Proceedings of the National Academy of Sciences* 114:1607–1612.

Thomé, M. T. C., and B. C. Carstens. 2016. Phylogeographic model selection leads to insight into the evolutionary history of four-eyed frogs. *Proceedings of the National Academy of Sciences* 113:8010–8017.

Weins, J. J., and T. A. Penkrot. 2002. Delimiting species using DNA and morphological variation and discordant species limits in spiny lizards (*Sceloporus*). *Systematic Biology* 51:69–91.

Whittaker, R. J., M. B. Araújo, P. Jepson, R. J. Ladle, J. E. Watson, and K. J. Willis. 2005. Conservation biogeography: Assessment and prospect. *Diversity and Distributions* 11:3–23.

Winker, K. 2009. Reuniting phenotype and genotype in biodiversity research. *BioScience* 59:657–665.

Yang, Z., and B. Rannala. 2010. Bayesian species delimitation using multilocus sequence data. *Proceedings of the National Academy of Sciences* 107:9264–9269.

Yeates, D. K., A. Seago, L. Nelson, S. L. Cameron, L. Joseph, and J. W. H. Trueman. 2011. Integrative taxonomy, or iterative taxonomy? *Systematic Entomology* 36:209–217.

Zachos, F. E., L. Christidis, and S. T. Garnett. 2020. Mammalian species and the twofold nature of taxonomy: A comment on Taylor et al. 2019. *Mammalia* 84:1–5. https://doi.org/10.1515/mammalia-2019-0009.

Zhang, C., M. Rabiee, E. Sayyari, and S. Mirarab. 2018. ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* 19:153.