



**HAL**  
open science

## Species Delimitation

Bruce Rannala, Ziheng Yang

► **To cite this version:**

Bruce Rannala, Ziheng Yang. Species Delimitation. Scornavacca, Celine; Delsuc, Frédéric; Galtier, Nicolas. Phylogenetics in the Genomic Era, No commercial publisher | Authors open access book, pp.5.5:1–5.5:18, 2020. hal-02536468

**HAL Id: hal-02536468**

**<https://hal.science/hal-02536468>**

Submitted on 10 Apr 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License


## Chapter 5.5 Species Delimitation

**Bruce Rannala**

Department of Evolution and Ecology, University of California Davis

One Shields Avenue, Davis CA USA

brannala@ucdavis.edu


 <https://orcid.org/0000-0002-8355-9955>

**Ziheng Yang<sup>1</sup>**

Department of Genetics, Evolution and Environment, University College London

London WC1E 6BT, United Kingdom

z.yang@ucl.ac.uk

 <https://orcid.org/0000-0003-3351-7981>

---

### Abstract

Species delimitation is the process of determining which groups of individual organisms constitute different populations of a single species and which constitute different species. The problem goes back to the earliest days of taxonomy and formalized processes for describing new species exist and are widely used, although the methods are time-intensive and problematic for some species. Genomic data carries extensive information about the degree of genetic isolation among species and about ancient and recent introgression. For this reason, genomic data can play an important role in species delimitation under many existing species concepts. Here we review the history of molecular species delimitation leading up to the current genomic era. We then describe the most widely used computational methods for species delimitation using single- and multi-locus genomic data. Relative strengths and weaknesses of the approaches are discussed and a new method for delimiting species based on empirical criteria is proposed.

**How to cite:** Bruce Rannala and Ziheng Yang (2020). Species Delimitation. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter No. 5.5, pp. 5.5:1–5.5:18. No commercial publisher | Authors open access book. The book is freely available at <https://hal.inria.fr/PGE>.

### 1 What is Species Delimitation?

Species play a central role in all branches of biological research and are the fundamental unit to measure biodiversity. The current rate of extinction of species on the planet due to anthropogenic activity is difficult to precisely estimate both because species boundaries are in some cases unclear and because millions of species have yet to be described. It is estimated that 80 to 90% of species on planet Earth are undiscovered, and it is likely that numerous contemporary species have already become extinct without scientists ever having documented their existence. Species delimitation is therefore an activity central to conservation of biodiversity. Several large initiatives are underway to either barcode most species (for example, by sequencing a single locus for millions of species) (International Barcode of Life <http://ibol.org/>), or sequencing whole genomes of all known species and discovering the remaining species (Earth Biogenome Project <https://www.earthbiogenome.org/>). Taxonomists are presently in a race to document the species in many groups that are

---

<sup>1</sup> Z.Y. is supported by a Biotechnological and Biological Sciences Research Council grant (BB/P006493/1).





© Bruce Rannala and Ziheng Yang.

Licensed under Creative Commons License CC-BY-NC-ND 4.0.

*Phylogenetics in the genomic era*.

Editors: Celine Scornavacca, Frédéric Delsuc and Nicolas Galtier; chapter No. 5.5; pp. 5.5:1–5.5:18

 A book completely handled by researchers.

 No publisher has been paid.

## 5.5:2 Species Delimitation

en route to extinction. All these efforts require methods for determining which individuals constitute new species (species delimitation) or should instead be assigned to an existing species (species assignment). Genomic species delimitation, the topic of this chapter, is therefore at the forefront of modern biodiversity science.

Taxonomic science has at least three distinct, but interdependent, roles in biology: the assignment of individual organisms into pre-existing species categories (species assignment), the assignment of species to higher categories (genus, family, etc), and the designation of new species categories to accommodate individual organisms that do not fit into an existing species category (species delimitation). Historically, all three roles were performed by taxonomists using morphological characters. For many species this has been effective and uncontroversial. However, some domains of life, such as bacteria, have few distinctive traits while others have distinctive morphologies that are highly plastic and associated with environmental factors leading to morphological convergence or divergence that is incompletely associated with genetic or evolutionary relatedness. For such groups, morphological delimitation of species can sometimes fail dramatically. Moreover, morphological species delimitation requires a high level of expertise and is time consuming, often making it impractical for large groups undergoing extinctions, which urgently need to be classified. A semi-automated process of delimitation in which the role of experts is to verify and refine results obtained from genomic data and computer algorithms is therefore very attractive. Such developments do not obviate the need for taxonomists but may alter their role in the process of documenting biodiversity. Automated algorithmic methods for delimiting species using genome data are the subject of this chapter. As noted by Jörger and Schrödl (2013) taxonomy includes both species discovery (delimitation) and the subsequent process of establishing a formal diagnosis and naming scheme. Diagnoses based on DNA are not considered here. As noted by Sites Jr and Marshall (2004), species concepts have received much more study than have “operational criteria” to be used for defining species in empirical studies. Operational criteria for use with molecular sequence data are the sole focus of this chapter.

## 2 A Brief History

### 2.1 Numerical taxonomy

The concept of using computer algorithms to delimit species based on morphological characters traces back to the 1960s. With the rise of computers and multivariate statistical methods the new field of Numerical Taxonomy (Sneath, 1957a) appeared poised to offer an objective solution to many of the problems of assigning and classifying species with confusing morphological variation. Bacterial taxonomists were early to adopt numerical taxonomy (Sneath, 1957b), possibly in the hope that large numbers of trait measurements could compensate for a lack of distinctive traits among bacterial strains (Goodfellow, 1971). Sneath (1976) considered a model of phenetic variation in bacteria, for example, in which the phenotype of a species is represented as a spherical multivariate normal distribution with the quantiles defining the species boundaries. Numerical taxonomy did not eliminate problems such as convergent evolution, however, which could cause unstable classifications based on morphology. Moreover, while multivariate analyses of morphology were effective at clustering individual operational taxonomic units (OTUs) into groups that shared distinctive features they did not provide *a priori* means for establishing species boundaries and could not identify the sources of morphological variation (genetic versus environmental).

## 2.2 Molecular taxonomy

During the 1970s, as molecular sequence data became available for proteins, and later for DNA and RNA, their potential utility for classifying difficult groups such as bacteria was widely recognized (Fox et al., 1977; Wayne et al., 1987; Wilson, 1995). During the 1980s, DNA-DNA hybridization (DDH) became the taxonomic gold standard for diagnosing new species of Bacteria and Archaea with a DDH similarity of less than 70% considered evidence of distinct species (Wayne et al., 1987; Meier-Kolthoff et al., 2013). However, for most groups of organisms the role that molecular versus morphological data should play in species assignment and delimitation – as well as the diagnostic criteria to be applied to each data type – remained uncertain. Traditional morphological delimitations rely on the identification of fixed differences between species. However, even for morphological characters finding convincing evidence of fixed differences requires sample sizes much larger than those found in most studies (Wiens and Servedio, 2000). Moreover, for molecular data with thousands (or millions) of bases in a sequence the probability of finding spurious fixed differences in a small sample of individuals can be very high. Fixed differences of bases at individual sites, or even of haplotypes, are commonly observed among populations within a species and are therefore not a sufficient criterion for delimiting species. Quantitative delimitation approaches that used diploid genotypic molecular data (such as allozymes) and population genetic statistics were proposed by several groups (reviewed by Sites Jr and Marshall, 2004). In particular, measures of gene flow based on Wright's  $F_{st}$  statistics were proposed as a criterion for delimiting species, with populations grouped into a single species if gene flow is detected (Porter, 1990). Criteria based on gene flow are unsatisfactory because they depend on overly simple models of population structure and a subjective threshold of gene flow for delimiting species. Furthermore, comparative genomic analyses in the past decade have highlighted the fact that gene flow between species is common in both plants and animals (Ellegren et al., 2012; Wu et al., 2018), including humans and their close relatives (Nielsen et al., 2017).

## 2.3 Patterns in gene trees

The widespread use of multilocus sequence data in the 1990s, and the development of the coalescent theory in population genetics, led several groups to develop species delimitation criteria based on observed patterns in gene trees. Avise and Ball Jr (1990) proposed a “genealogical species concept” that led to the development of several operational definitions based on patterns of shared ancestry in inferred gene trees. The most widely used criterion was “exclusivity”, also referred to as “reciprocal monophyly” (Ball et al., 1990; Baum and Donoghue, 1995; Palumbi et al., 2007), which means that all sequences from one species form a monophyletic group in the gene tree relative to the sequences from any other species. Such methods treat gene trees as observations and do not properly account for errors in inferred gene trees, which tend to reduce the frequency of reciprocal monophyly. This criterion may be too strict in some situations and not strict enough in others, depending on the population isolation time relative to the population sizes. If the populations are very large the isolation time required to reach exclusivity at all loci (or at 50% of loci, to use a weaker criterion of exclusivity) can be very long (Hudson and Coyne, 2002; Hudson and Turelli, 2003; Knowles and Carstens, 2007), and the exclusivity criterion may fail to recognise good species. In contrast, if the populations are very small (as in the case of a few founders establishing a population), the gene tree may be reciprocally monophyletic, but the population isolation time may be too short to consider them as distinct species. Templeton (2001) proposed a

## 5.5:4 Species Delimitation

genealogical “test of cohesion” for identifying species using his method of nested clade analysis (NCA). However, NCA is known to have poor statistical properties as it does not account adequately for demographic stochasticity and has extreme type I error rates (Beaumont and Panchal, 2008; Petit, 2008; Knowles, 2008).

### 2.4 DNA barcoding

Debate concerning procedures for species assignment and delimitation using molecular data was reinvigorated during the 2000s with the publication of an influential proposal for a “DNA barcoding” initiative using a single locus (the COII mitochondrial gene for animals) as a diagnostic for assigning species (Hebert et al., 2003). Advocates of DNA barcoding also proposed the use of sequence divergence thresholds (or barcode gaps) as a means for delimiting new species (Tautz et al., 2003; Blaxter, 2003). Criticisms of species delimitation using DNA barcoding included the fact that interspecific and intraspecific sequence distances may be similar in large populations, so that no fixed threshold exists, and that thresholds are subjective and often difficult to establish *a priori* (Moritz and Cicero, 2004; Will and Rubinoff, 2004; DeSalle et al., 2005). Methods based on a single locus are also expected to have low power for many recently diverged species. An advantage of using a single diagnostic locus was that it enabled high throughput analyses not possible with the multilocus sequencing approaches of the 2000s. The advance of sequencing technologies has shifted recent interests to multilocus species delimitation and assignment methods, which are more powerful than single locus methods.

### 2.5 Multispecies coalescent

The barcoding debate prompted several researchers to point out that the multispecies coalescent (MSC) (Rannala and Yang 2003; Chapter 3.3 [Rannala et al. 2020]) could provide a probability distribution for the likely gene trees given a particular species tree (in the absence of gene flow between species) and that this could provide a statistical model for assigning individuals to species based on either single-locus (Pons et al., 2006) or multilocus (Nielsen and Matz, 2006) sequence data. These approaches also offered an alternative to barcoding-inspired delimitation methods based on percent sequence divergence between species (Hebert et al., 2003) which is sensitive to the levels of polymorphism within populations. Methods based on the multispecies coalescent do not require reciprocal monophyly to delimit species (Knowles and Carstens, 2007) and can take proper account of statistical uncertainty in gene trees (Yang and Rannala, 2010). However, early methods were not able to analyze multilocus datasets and used approximations to the MSC (Pons et al., 2006). More recently, several approximate and exact multilocus methods have been developed for species delimitation under the MSC (O'Meara, 2009; Yang and Rannala, 2010) that can potentially scale to genomic datasets and provide greater power for species delimitation and species assignment. Inferences from such methods may be sensitive to model violations, however, such as gene flow between species (Leaché et al., 2018). MSC-based methods are an example of a parametric inference method.

### 2.6 Machine learning

The high performance of machine learning algorithms in solving many classification problems, and the straightforward generic nature of their application, has led to many recent applications in population genetics, phylogenetics, and other areas of evolutionary biology. Machine

learning algorithms can be broadly classified as either supervised (SML) or unsupervised (UML) machine learning, according to how training datasets are utilized. Both SML and UML approaches have been recently applied to species delimitation (Pei et al., 2018; Derkarabetian et al., 2019). Pei et al. (2018) developed an SML algorithm for species delimitation using support vector machines. Datasets generated by population genetic simulations (with or without gene flow) were used to train the algorithm. Summaries of the data were used rather than using sequence data directly to make the model and computation tractable. Five summary statistics were used: the proportion of private positions, the folded site frequency spectrum, the pairwise difference ratio, F-statistics, and the longest shared tract. For simulated data, the algorithms appeared to perform as well as the model-based species delimitation method BPP (Yang and Rannala, 2010) when species are genetically isolated and were more likely than BPP to delimit species that experienced gene flow. A similar approach was developed by Smith and Carstens (2019) but using the folded site frequency spectrum and a Random Forest (RF) classifier.

SML methods have the advantage that they can be computationally efficient and can be trained on models that are too complex to derive formal Bayesian or maximum likelihood estimators. A well-known weakness of supervised learning methods, encapsulated by the so-called “supervised learning no free lunch theorem” (Wolpert, 2002), is that they can become too specialized – they work very well for the training dataset but poorly for many other datasets. In this case, since the algorithm is trained using specific values of population genetic parameters such as  $\theta$  and  $M$  it could perform poorly outside the training range. Formal statistical methods do not have this problem since they have optimality properties that hold over the entire parameter space. For complicated models it may be impossible to train an SML algorithm over the entire state space for the parameters. Another weakness of both SML and UML methods is that they often use summary statistics rather than the full dataset. Unless the summary statistics are known to be sufficient statistics this will entail loss of information. Finally, little is known about the asymptotic statistical performance of most SML or UML methods and developers must therefore resort to simulations to evaluate them when inferring evolutionary parameters for which the true parameter values are unknown. Simulation studies can never be comprehensive.

### 3 A Survey of Species Delimitation Methods

Here we provide a concise survey of the most widely used species delimitation methods, sketching important features of the statistical and computational theory and the assumptions underlying them. We focus exclusively on methods designed for use with DNA sequence data and divide the methods into two categories: (1) heuristic methods, which use a summary statistic or algorithm for delimitation that is not derived from a formal statistical model of the population genetic structure. Heuristic methods are often computationally efficient but the results can be difficult to interpret and they may have poor statistical properties; (2) parametric methods, which are based on an explicit probabilistic model of population divergences and evolution of the genetic sequences and which select the delimitation model that maximizes the likelihood or Bayesian posterior probability. Full-likelihood methods under a parametric model are known to be asymptotically most powerful when the model is correct but are often computationally demanding. If the model is incorrect the statistical properties may become unpredictable. Most widely used parametric methods are derived based on the MSC model, which will therefore be described in some detail below. Both approximate and exact parametric methods will be described. A distinction is made between methods

## 5.5:6 Species Delimitation

designed for use with a single locus versus multilocus sequence data. We focus on methods that are applicable across the tree of life, and do not consider methods developed specifically for a certain species group, such as Genome Blast Distance Phylogeny (Meier-Kolthoff et al., 2013) for species delimitation of bacteria.

### 3.1 Heuristic methods

Most heuristic methods for species delimitation originate from the DNA barcoding initiative and are therefore designed for single locus data. Multilocus heuristic methods proposed more recently often simply concatenate genes (Zhang et al., 2013). Heuristic methods are computationally efficient and can be applied to large datasets. Their statistical performance may be good in some regions of the parameter space but poor in others. To evaluate the performance, simulation is often used because standard statistical theory (e.g., asymptotic efficiency in large datasets, etc) typically does not apply to heuristic methods.

#### 3.1.1 ABGD: Automated Barcode Gap Discovery

Early studies of pairwise sequence divergence between and within species aimed to identify a “barcode gap” which can distinguish within-population differences from differences caused by species divergences. For example, a relative difference of one order of magnitude (the  $10\times$  rule) between intra- and interspecific divergences was proposed by Hebert et al. (2004) and a maximum of 3% for intraspecific divergence (the 3% rule) was proposed by Smith et al. (2005). Such *a priori* thresholds are arbitrary and subsequent analyses suggested that the barcode gap varies among species groups based on factors such as effective population sizes and species divergence times (Hickerson et al., 2006; Rannala, 2015).

The aim of the automated barcode gap discovery (ABGD) program (Puillandre et al., 2012) is to identify barcode gap thresholds in an automated process. For a sample of  $n$  haploid sequences all  $n(n-1)/2$  pairwise distances are calculated. The distance metric may use a correction for multiple-hit substitutions. The distances are then ranked in increasing order so that  $d_i \leq d_{i+1}$  for  $i = 1, \dots, n(n-1)/2$ . For the distance of rank  $r$  the local slope is calculated as

$$s_{r,w} = \frac{d_{r+w} - d_w}{w}. \quad (1)$$

The slope is expected to be largest at the barcode gap that delimits species. Thus the method infers the barcode gap as the distance that maximizes the local slope.

Simulation under the coalescent process was used to compute a threshold value under which sequences are more likely to be intraspecific, assuming a constant population size with  $n$  and  $\theta$  specified and estimating the threshold exceeding 95% of pairwise distances. It was concluded that for  $n > 10$  the threshold was a linear function of  $\theta$  with a slope of 2.581. The rationale was that, if  $\theta$  is known, the barcode gap can be predicted from the simulation results. The parameter  $\theta$  is unknown for empirical data but for a single population the average pairwise distance provides an estimate of  $\theta$ . With more than one species in the sample the estimate of  $\theta$  will be too large. A user-specified “prior” threshold is therefore used to separate intraspecific distances from interspecific distances. Only putative intraspecific distances (those below the prior threshold) are used in estimating  $\theta$ . Once  $\theta$  is estimated a threshold distance is determined, and groups are then formed so that “the distance between sequences from different groups is always larger than the gap distance, and for each sequence of each group, there is at least one other sequence in the group at a distance smaller than

the gap distance.” This procedure is applied recursively to identify additional groups that may have different barcode gaps.

Like the PTP method discussed below, ABGD may be expected to work best when the gene tree is essentially monophyletic (i.e., there is no incomplete lineage sorting). Like other barcoding methods it only uses information from a single locus. It is computationally inexpensive and can be applied to large samples of sequences. Its weaknesses include its reliance on simple pairwise distance calculations and clustering operations, and failure to use all the information in the sequence data. The distribution of intraspecific distances may be multimodal due to factors such as population growth or selection (Rogers and Harpending, 1992; Harpending et al., 1993) potentially leading to spurious barcode gaps.

### 3.1.2 GMYC: General Mixed Yule Coalescent

The General Mixed Yule Coalescent (GMYC) (Pons et al., 2006) assumes that the waiting times between coalescence events or branch lengths in a gene tree fall into two classes: those within species with the rate determined by the coalescent process, or those between species with the rate determined by a generalization of the Yule process model of species divergences. It is assumed that a time point  $T$  exists at which the generative process for gene-tree nodes switches from a coalescent process to a Yule process. Maximum likelihood is used to estimate  $T$  and the choice of  $T$  determines the species delimitation. A likelihood ratio test is also proposed to test for the existence of multiple species ( $T > 0$ ) versus a single species ( $T = 0$ ). This method treats the inferred gene tree as known and is therefore computationally simple and fast, but has the drawback of ignoring errors in the gene tree. Another limitation is that the method can be applied to only a single locus, although an attempt was made to extend it to multiple loci (Fujisawa and Barraclough, 2013). More importantly, by using a single threshold  $T$ , the model implicitly assumes that all lineages in each population coalesce before any speciation event occurs, implying the absence of incomplete lineage sorting. The method ignores the coalescent process within ancestral species populations, in effect assuming that the gene tree is reciprocally monophyletic. The GMYC method should thus perform best for datasets with long intervals between speciation events and small population sizes, in which case incomplete lineage sorting is unlikely to occur.

### 3.1.3 PTP: Poisson Tree Process

The “Poisson Tree Process” (PTP) identifies species status based on the distribution of branch lengths in the gene tree (Zhang et al., 2013). The tree and branch lengths are inferred from a sequence alignment using maximum likelihood and then treated as known without errors. When multiple loci are available they are concatenated to infer one gene tree with branch lengths. The PTP method models the branch lengths,  $x_i$ , in a rooted non-ultrametric tree as a mixture of two exponential distributions with rates  $\lambda_S$  (between-species) and  $\lambda_C$  (within-species), respectively. A species delimitation model ( $\Lambda$ ) assigns every branch on the gene tree to one of those two classes. The log-likelihood for the delimitation, given the data ( $G$ ) of a rooted gene tree with  $n$  branches, is then

$$L(\Lambda; G) = \sum_{i=1}^k \log(\lambda_S e^{-\lambda_S x_i}) + \sum_{i=k+1}^n \log(\lambda_C e^{-\lambda_C x_i}), \quad (2)$$

where the delimitation model  $\Lambda$  partitions  $k$  branches as “between-species” and  $n - k$  branches as “within-species”. The Poisson rates ( $\lambda$ s) are estimated by using the inverse of the average



## 5.5:8 Species Delimitation

branch length in each class, so that no iteration is needed for parameter estimation. A heuristic search is used to find the delimitation that maximizes the likelihood. An extension of the method (Kapli et al., 2017) allows  $\lambda$ s to vary among populations.

A strength of the PTP approach is that it can handle large datasets with thousands of species. Unlike the GMYC method discussed above, PTP uses a rooted non-ultrametric tree so that it does not rely on the molecular clock, which may be seriously violated for distantly related species. The approach implicitly assumes reciprocal monophyly in the gene tree and a perfect match of the gene tree with the species tree. It is thus expected to work best for identifying species that are separated by long intervals between speciation events and that have small population sizes.

Some weaknesses of the method may be easily identified. The method is essentially a single-locus method, as concatenating sequences across gene loci or genomic segments does not account for the stochastic fluctuations in the coalescent process among the loci. Because the Poisson tree process describes a distribution of delimitation models, in theory there should be a prior term  $f(\Lambda)$  in equation 2, and the posterior probability for the delimitation model should be maximized  $f(\Lambda|G)$  instead of the log-likelihood:

$$f(\Lambda|G) \sim f(\Lambda) \prod_{i=b}^n \lambda_{\mathbb{I}_b} \exp\{-\lambda_{\mathbb{I}_b} x_i\} \quad (3)$$

where the indicator  $\mathbb{I}_b = \text{'S'}$  or  $\text{'C'}$  depending on whether delimitation model  $\Lambda$  partitions branch  $i$  as a between-species branch or a within-species coalescent branch. Furthermore, if more than two sequences are in one species according to the delimitation model, the waiting time until the next coalescent when there are  $k$  lineages is proportional to  $2/(k(k-1))$ . This can in theory be accommodated by redefining  $\lambda_C$  as the coalescent rate for two lineages, and applying the correction factor  $2/(k(k-1))$  if the within-species branch represents the waiting time until the next coalescent when there are  $k$  lineages in the sample in the population. Parameters  $\lambda_S$  and  $\lambda_C$  may be estimated using the method of moments, by using the average observed values so that the amount of computation remains the same. While it is unnecessary to sample multiple individuals or sequences to infer the species phylogeny, sampling multiple sequences from the same species adds much information in species delimitation (Zhang et al., 2011). Such modifications may make it possible to utilize the information in multiple samples.

### 3.2 Parametric methods

Parametric species delimitation methods are based on a probabilistic data-generating model, that is, a model of the biological processes generating gene genealogical trees and DNA sequences among individuals. This entails modeling the process of speciation, the genealogical process of coalescence within populations, and the process of DNA sequence evolution driven by genetic drift and natural selection. The canonical model relevant to species delimitation is the multispecies coalescent (MSC) with neutral evolution (Rannala and Yang, 2003). All parametric species delimitation methods considered here are based on an MSC model, although some methods (such as GMYC, Pons et al., 2006) are based on simplifications or approximations of it. The MSC model, describing the basic biological processes of reproduction and drift, is important to heuristic methods of species delimitation as well: for example, evaluation of the performance of those methods typically involves simulating genetic sequence data under the MSC model. If species delimitation can be formulated as a problem of statistical inference under the MSC model, standard theories of statistical inference can be made use of; for example, maximum likelihood and Bayesian methods are known to have

desirable asymptotic (large-sample) properties such as consistency and efficiency. However, the assumptions of the parametric approach may be violated in real data analysis, and an important question is how well the method performs when the model is misspecified.

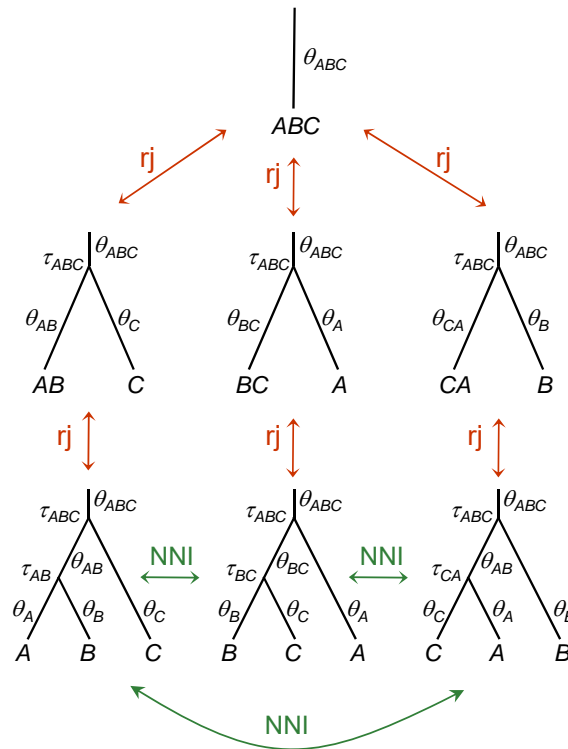
### 3.2.1 Multispecies coalescent and the distribution of gene trees

The genomes of individual organisms carry rich historical information at many levels and at different time scales. At a low level, genomes provide information about inbreeding (homozygosity), pedigree relatedness and genome sharing (identity by descent) among individuals. At a higher level, genomes are informative about population or species affiliations and the evolutionary relationships among species. Different statistical models and inference methods are often used to extract information at these different levels, yet the models used are interrelated and have many shared parameters. For example, individual inbreeding coefficients are determined by mating patterns and population structure, and gene trees of relationships among sequences are influenced by population level processes such as selection and genetic drift. Phylogenetic trees among species are, in turn, inferred from patterns observed in gene trees.

The phylogenetic relationships of species or genetically isolated populations constrain the possible relationships of genomic sequences (see Chapter 3.3 [Rannala et al. 2020]). When we trace the genealogical history of a sample of sequences backwards in time, sequences from different isolated populations cannot coalesce until we reach the common ancestor of those populations. If the sequences are in the same population the rate at which they coalesce is determined by the population size. Thus, the phylogenetic tree of populations generates a probability distribution on the possible gene trees. Conversely, the probabilities of gene trees, which underlie the genomic sequence data, can inform us about population history, such as population divergence times, population sizes, and between-population migration or introgression.

Parametric statistical approaches to species delimitation use the multispecies coalescent (MSC) model to infer the existence of genetically isolated populations that are potential species. Multilocus sequence data sampled from modern species or populations are used to calculate the posterior probabilities (Yang and Rannala, 2010) or marginal likelihoods (Grummer et al., 2013; Rannala and Yang, 2017) of different species delimitation models, where a delimitation model corresponds to a certain way of merging populations into the same species (Figure 1). In a randomly mating population the two sequences sampled from a diploid individual are equivalent to any two sequences randomly sampled from the population so only the population identity of each sequence is relevant. If strong evidence exists that two or more populations are genetically isolated, they will be assigned to distinct species. In contrast, if they constitute a single panmictic population they will be collapsed into a single species. In this approach, the level of gene flow may have a major impact, as will the amount of time since population divergence. The procedure does not derive from any particular species definition but it includes the Biological Species Concept as a particular case, so that the species status is recognized given a sufficient period of reproductive isolation.

The MSC model is parametrized by the species tree topology, as well as the species divergence times ( $\tau_s$ ) and population sizes for both modern and ancestral species ( $\theta_s$ ) (Figure 1). A (rooted) species tree for  $s$  species has  $s - 1$  internal nodes or speciation events and  $2s - 1$  nodes or populations; thus an MSC model for  $s$  species has  $s - 1$  node ages or species divergence times ( $\tau_s$ ) and  $2s - 1$  population size parameters ( $\theta_s$ ). In analysis of genomic sequence data, both  $\tau_s$  and  $\theta_s$  are measured by genetic distance or the expected number of mutations per site.



■ **Figure 1** For three populations ( $A, B, C$ ), there are five species delimitation models (one of 1 species, three of 2 species, and one of 3 species) and seven MSC models, with the delimitation of three species (bottom row) resolved into three MSC models (three species phylogenies). Each MSC model of  $s$  species has  $(s - 1)$  divergence times ( $\tau$ s) and  $(2s - 1)$  population size parameters ( $\theta$ s). Species delimitation through Bayesian model selection uses Markov chain Monte Carlo (MCMC) proposals to traverse the space of MSC models to estimate the posterior probabilities for the MSC models. The posterior for a delimitation model is the sum of the posterior probabilities for the compatible MSC models: for the example here, the probability for the existence of three species ( $A, B, C$ ) is the sum of the three MSC models on the bottom row. In BPP, subtree-pruning-and-regrafting (SPR) or nearest-neighbor-interchange (NNI) algorithms are used to move between different species phylogenies, and rjMCMC is used to move between different species delimitations (Yang and Rannala, 2010, 2014).

### 3.2.2 Posterior probabilities of species delimitation models

Given a set of  $K$  populations, different species delimitations correspond to different ways of merging populations into the same species, with the number of delimited species ranging from 1 to  $K$ . We assume that individuals are correctly assigned to populations, and a population will not be split into different species although multiple populations may be merged into the same species. As an extreme approach, each sampled individual can be assigned its own population, and the Bayesian algorithm can be used to achieve both assignment and delimitation (Olave et al., 2014). When more than two species exist in the delimitation model, there are different species phylogenies as well. The species delimitation and species phylogeny together constitutes a fully specified MSC model, which allows the definition of the parameters and the specification of the probability distribution of the gene trees (Rannala and Yang, 2003). For example, with 3 populations, there will be five delimitation models: one model of 1 species, three models of 2 species, and one model of 3 species (Figure 1),

and in the case of 3 species, there are also three species phylogenies. Thus in total there are seven MSC models.

Let  $\boldsymbol{\theta}_k$  be the parameters in the MSC model specified by a delimitation model  $\Lambda_k$ . Note that the delimitation alone (e.g., knowledge of the existence of three species without knowledge of the species phylogeny) may be insufficient to define the parameters or to specify the data-generating mechanism. Thus from now on we assume that the delimitation model  $\Lambda_k$  also specifies the species phylogeny so that the parameters can be defined. In the case of figure 1 for three populations, there are seven such models. The posterior probability of delimitation model  $\Lambda_k$  given the sequence data at  $L$  loci,  $\mathbf{X} = \{X_i\}$ , is then

$$\mathbb{P}\{\Lambda_k|\mathbf{X}\} \propto \pi_k M_k, \quad (4)$$

where  $\pi_k$  is the prior probability for model  $\Lambda_k$ , and  $M_k$  is the marginal likelihood for model  $\Lambda_k$ . The proportionality constant is to ensure that the posterior probabilities for all models sum to 1. The marginal likelihood  $M_k$  for delimitation model  $\Lambda_k$  is an integral (an average) over all possible gene trees at each locus and over the MSC parameters,

$$M_k = \iint f(\boldsymbol{\theta}_k|\Lambda_k) \prod_i^L [f(G_i|\Lambda_k, \boldsymbol{\theta}_k) f(X_i|G_i)] dG d\boldsymbol{\theta}_k, \quad (5)$$

where  $f(\boldsymbol{\theta}_k|\Lambda_k)$  is the prior on the parameters under the model,  $f(G_i|\Lambda_k, \boldsymbol{\theta}_k)$  is the MSC density for gene tree  $G_i$  at locus  $i$  (Rannala and Yang, 2003), and  $f(X_i|G_i)$  is the probability of the sequence alignment at locus  $i$ , known as the phylogenetic likelihood (Felsenstein, 1981).

The integrals of equation 5 are typically calculated numerically in the Markov chain Monte Carlo (MCMC) algorithm, as implemented in the BPP program (Yang and Rannala 2010; Rannala and Yang 2017; Chapter 5.6 [Flouri et al. 2020]). In other words, MCMC is used to traverse the model space, and the frequency at which the MCMC visits each model is the estimate of the posterior probability for that model. If a delimitation is compatible with multiple MSC models (for example, the delimitation of three species is resolved into three MSC models or three species phylogenies in Figure 1), its posterior probability is calculated as the sum of the posterior probabilities for the compatible MSC models. This is the A11 analysis by Yang (2015). In BPP, tree perturbation algorithms such as nearest-neighbor-interchange (NNI), subtree-pruning-and-regrafting (SPR), and NodeSlider are used to propose moves from one species tree to another with the species delimitation fixed (Yang and Rannala, 2010, 2014; Rannala and Yang, 2017). Moves between species delimitation models involve changes of dimension (the number of parameters), so they are implemented using a pair of reversible-jump MCMC moves ( “split” and “join” , Yang and Rannala, 2010). For example, the “split” move can be used to move from the 2-species model  $(AB, C)$ , which has 4 parameters, to the three-species model  $((AB)C)$ , which has 7 parameters, with three new parameters  $(\tau_{AB}, \theta_A, \theta_B)$  created. The reverse “join” move changes the 3-species model to the 2-species model, dropping the redundant parameters. The pair of moves constitutes one reversible-jump proposal. RjMCMC algorithms often mix poorly, but thanks to development of improved algorithms, which make coordinated changes to the species tree and the gene trees when the MSC model changes (Rannala and Yang, 2013, 2017), BPP can be used to analyze datasets with hundreds or even thousands of loci. Simulation has confirmed the overall efficiency of the method (Zhang et al., 2011, 2014).

A number of empirical studies have found that the approach to model selection implemented in BPP tends to “oversplit” , favouring delimitation models with a large number of species (Sukumaran and Knowles, 2017). In some studies, the delimited number of species is the number of populations in the dataset. The problem appears to be worse when more

## 5.5:12 Species Delimitation

data (more loci) are analyzed. Leaché et al. (2018) studied the dynamics of Bayesian model selection when the true model involves two species/populations with migration but the two compared models assume either one species or two species without gene flow. The two models considered by BPP are in this case both wrong. However, analysis suggest that the two-species model is closer (judged by Kullback-Leibler divergence) to the true model of two populations with migration and is thus less wrong than the one-species model (Leaché et al., 2019). In such a case, the two-species model will dominate, with its posterior probability approaching 100% when the amount of data (the number of loci) approaches infinity. This provides an explanation for the observation that BPP tends to favour the model of distinct species even if there is substantial gene flow between the populations. If Bayesian model selection is conducted under the MSC model with migration or introgression, the two-species model will be the correct one and will naturally dominate, and the problem of over-splitting will become even more serious. With gene flow, the distinction between populations and species in such models is arbitrary. We suggest that the approach of Bayesian model selection be used mostly in the context of delimiting sympatric species whose distinctness is maintained by a lack of gene flow rather than partial genetic isolation.

### 3.2.3 Bayes factor delimitations

Several groups (Grummer et al., 2013; Leaché et al., 2014) have suggested the use of Bayes factors to choose among a small set of species delimitation models. The Bayes factor for two delimitation models  $\Lambda_1$  and  $\Lambda_2$  is defined as the ratio of the marginal likelihoods under the two models. From equation 4,

$$BF_{12} = \frac{M_1}{M_2} = \frac{\mathbb{P}\{\Lambda_1|X\}/\mathbb{P}\{\Lambda_2|X\}}{\pi_1/\pi_2}. \quad (6)$$

In other words, the Bayes factor is the ratio of the posterior odds to the prior odds. If we assign uniform prior probabilities  $\pi_1 = \pi_2$  to the two models, the Bayes factor will simply be the posterior odds. Note that here the delimitation models  $\Lambda_1$  and  $\Lambda_2$  should be fully specified MSC models: if there are 3 or more delimited species, the species phylogeny should be considered part of the model specification as well. Otherwise the marginal likelihood is not well defined. Second, the marginal likelihood is an average over the prior distribution of parameters in the MSC model (the  $\tau$ s and  $\theta$ s). As a result, the prior on parameters may be influential on the marginal likelihood, besides the model of species divergences. The marginal likelihood is sometimes referred to as the “evidence” by Bayesians, but it should be borne in mind that it incorporates information from the prior which may represent subjective “opinions”. Bayes factors and posterior probabilities for delimitation models are equivalent if Bayes factors are applied to the complete set of all possible delimitation models. Otherwise Bayes factors provide a local comparison among the delimitations examined. Interpretation or calibration of the Bayes factor is through reference to posterior model probabilities: a Bayes factor of 99 (= 0.99/0.01 in terms of posterior odds) might be considered “strong” evidence in favour of model  $\Lambda_1$ , for example, while 9 (= 0.9/0.1) might only be considered “positive” evidence.

Grummer et al. (2013) proposed a Bayes factor test of species delimitation and used the STARBEAST program (Heled and Drummond, 2009) to calculate marginal likelihoods under different delimitation models. Leaché et al. (2014) developed an efficient Bayes factor delimitation method using the SNAPP (Bryant et al., 2012) algorithm to calculate marginal likelihoods. SNAPP is developed for single nucleotide polymorphism (SNP) data from unlinked loci. As every site (every SNP) is assumed to have its own independent history, the gene trees

including coalescent times can be integrated analytically under the MSC, eliminating the need for computationally expensive integration via MCMC. SNAPP is thus computationally efficient.

A major weakness of SNAPP, or of using unlinked SNP data, is information loss and a resulting lack of power. This occurs for two reasons. First, correction for ascertainment bias leads to loss of information. SNPs are collected because the sites are polymorphic. This fact has to be accommodated in the inference method, and correction for such “ascertainment bias” in data collection may lead to serious loss of information. As an analogy, there are two data outcomes in a binomial experiment: “success” and “failure”. If we filter out all the “failures”, it will be important to correct for such ascertainment bias and furthermore very little information remains after such data filtering. In the case of the SNP data, the information loss may not be so severe, but the correction may be expected to have a major impact on inference, in particular on branch lengths in the species phylogeny. It seems that the number of constant sites removed or the number of sites separating the SNP loci may be useful for recovering some of the information lost. Second, use of essentially independent SNP sites means that not all parameters in the MSC model are identifiable and the power for comparing different delimitation models may be affected as well. Multi-locus sequence alignments allow one to tease apart the among-site variation within the same locus given the gene tree (due to the Poisson mutation process) from the among-loci variation in the gene trees (due to the stochastic fluctuation of the multispecies coalescent process, influenced by parameters such as ancestral population sizes). As a result, all parameters in the MSC model are identifiable given the multi-locus sequence data. In contrast, a single SNP can only identify a single bipartition in a gene tree and contains very little phylogenetic information. The two sources of variation are then confounded. As a result, not all parameters in the MSC model are identifiable given SNP data.

### 3.2.4 Delimitation based on empirical criteria

Suppose we are given a detailed description of the history of two allopatric populations, including their divergence time, population sizes, and the timing, directions and intensity of migration or introgression events. Can we then decide whether the two populations belong to the same species or are two distinct species? If the two populations are sympatric, reduction or absence of gene flow revealed by genomic data will be evidence for genetic isolation and for the existence of barriers to gene flow so that distinct species status can be established. However, if the species are allopatric, genetic differentiation may be due to an absence of opportunities for interbreeding rather than the existence of isolation mechanisms or adaptation to different ecological and environmental conditions. Delineation of species boundaries in such cases will be arbitrary. The neutral genome has the power to let us infer a detailed population divergence history, but may not be informative about ecological adaptation or reproductive isolation. Delimitation in such cases may make use of empirically established criteria, based on evolutionary parameters. The MSC model, either without or with introgression, can be used to infer the MSC parameters and these then used in combination with any empirical criteria for species delimitation.

One such criterion is the genealogical divergence index (*gdi*) (Jackson et al., 2017). Suppose one samples two sequences ( $a_1$  and  $a_2$ ) from population  $A$  and one sequence ( $b$ ) from population  $B$ . If sequences  $a_1$  and  $a_2$  coalesce first, the gene tree will be  $G_a = ((a_1, a_2), b)$ . Under the MSC model without gene flow, the gene tree probability is

$$P_a = \mathbb{P}(G_a | \boldsymbol{\theta}) = 1 - \frac{2}{3}e^{-2\tau/\theta_A}, \quad (7)$$

## 5.5:14 Species Delimitation

where  $2\tau/\theta_A$  is the population divergence time in coalescent units (with one coalescent time unit to be  $2N_A$  generations) and  $e^{-2\tau/\theta_A}$  is the probability that sequences  $a_1$  and  $a_2$  do not coalesce before reaching the time of species divergence ( $\tau$ ) when we trace the genealogy backwards in time.  $P_a$  ranges from  $\frac{1}{3}$  to 1. Jackson et al. (2017) rescaled  $P_a$  to form the *gdi* index

$$gdi = (3 \times P_a - 1)/2, \quad (8)$$

so that it ranges from 0 to 1.

Based on the meta-analysis of Pinho and Hey (2010), Jackson et al. (2017) suggested the rule of thumb that *gdi* values  $< 0.2$  suggest a single species and *gdi* values  $> 0.7$  suggest distinct species, while *gdi* values within the range indicate ambiguous delimitation. Those limits correspond to 0.47 and 0.8 for  $P_a$ , and, in the case of no migration, to 0.22 and 1.20 for the population divergence time in coalescent units.

The use of the *gdi* index as a metric for delimiting species has several drawbacks. First, when populations  $A$  and  $B$  have very different sizes, it is unclear which population size should be used. For example, one can use two sequences from  $B$  ( $b_1, b_2$ ) and one sequence from  $A$  to define the probability for the gene tree  $G_b = ((b_1, b_2), a)$ , and its probability  $P_b = \mathbb{P}(G_b|\theta)$ . However, when  $\theta_A$  and  $\theta_B$  are very different,  $P_a$  and  $P_b$  may be very different, leading to the awkward situation where  $P_a$  suggests one species while  $P_b$  suggests two (Leaché et al., 2019). Second, small population sizes may cause the index to over-split. It may therefore be useful to include a minimum absolute divergence time measured in generations. Third the criterion often leads to indecision, but this may reflect the difficulty of species delimitation for allopatric populations. In light of those drawbacks, we suggest the following modifications for an operational concept for delimiting species under the MSC model. We consider two populations to be distinct species if  $P_a > 0.5$ ,  $P_b > 0.5$ ,  $M = Nm < 0.1$ , and the divergence time is more than  $10^4$  generations. Otherwise we declare one species if  $P_a < 0.4$ ,  $P_b < 0.4$ ,  $M = Nm \geq 1$ , and divergence is within  $10^3$  generations. Situations between those two extremes are undecided.

Different heuristic criteria may be designed to correspond to different species definitions. Given any empirical criterion, a hierarchical procedure can be devised to delimit species using multi-locus genetic sequence data (Leaché et al., 2019). First we estimate a population phylogeny under the MSC. This is used as a guide tree so that its topology is not changed further, while the nodes on the tree may represent either populations or species. We then attempt to merge the populations into the same species using the criterion, starting from the tips of the tree, moving towards the root, each time re-estimating the MSC parameters. An ancestral node on the guide tree is merged into one species only if its descendant nodes are already merged. The procedure ends when no populations can be unambiguously merged into one species. The procedure is applied to several simulated and real datasets in Leaché et al. (2019).

## 4 Conclusions

Genomic data provide a rich source of information concerning the evolutionary history of species and populations such as divergence times, population sizes, and migration/hybridisation intensity. For sympatric species, convincing evidence of genetic isolation may be enough for establishing species status. For populations from different geographic locations, the genetic isolation can be due to isolation by distance, and may have to be combined with other sources of evidence to justify species status. In complex cases such as a species ring, the

situation may be so complicated that delimiting species becomes an arbitrary exercise. The MSC provides the framework for estimating evolutionary parameters including migration rates or introgression intensity, which can be used to apply empirical criteria for delimiting species or to generate hypotheses of species status, to be integrated with other evidence such as morphology and ecology (Yang and Rannala, 2010). An explicit extension of BPP to allow a model including a morphological character (iBPP) has been implemented by Solís-Lemus et al. (2015).

This chapter should have made clear the fact that there exists no “magic bullet” method for species delimitation using genomic data. However, methods are gradually converging on this target and already provide many useful tools for preliminary delimitations, as well as providing a solid theoretical foundation for future developments.

## References

- Avise, J. C. and Ball Jr, R. M. (1990). Principles of genealogical concordance in species concepts and biological taxonomy. *Oxford Surveys in Evolutionary Biology*, 7:45–67.
- Ball, R. M., Neigel, J. E., and Avise, J. C. (1990). Gene genealogies within the organismal pedigrees of random-mating populations. *Evolution*, 44:360.
- Baum, D. A. and Donoghue, M. J. (1995). Choosing among alternative “phylogenetic” species concepts. *Systematic Botany*, 20:560.
- Beaumont, M. A. and Panchal, M. (2008). On the validity of nested clade phylogeographical analysis. *Molecular Ecology*, 17:2563–2565.
- Blaxter, M. (2003). Molecular systematics: counting angels with DNA. *Nature*, 421:122.
- Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. A., and RoyChoudhury, A. (2012). Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Molecular biology and evolution*, 29(8):1917–1932.
- Derkarabetian, S., Castillo, S., Koo, P. K., Ovchinnikov, S., and Hedin, M. (2019). A demonstration of unsupervised machine learning in species delimitation. *Molecular Phylogenetics and Evolution*, 139:106562.
- DeSalle, R., Egan, M. G., and Siddall, M. (2005). The unholy trinity: taxonomy, species delimitation and DNA barcoding. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360:1905–1916.
- Ellegren, H., Smeds, L., Burri, R., Olason, P. I., Backstrom, N., Kawakami, T., Kunstner, A., Makinen, H., Nadachowska-Brzyska, K., Qvarnstrom, A., Uebbing, S., and Wolf, J. B. W. (2012). The genomic landscape of species divergence in ficedula flycatchers. *Nature*, 491:756–760.
- Felsenstein, J. (1981). Evolutionary trees from dna sequences: a maximum likelihood approach. *J. Mol. Evol.*, 17:368–376.
- Flouri, T., Rannala, B., and Yang, Z. (2020). A tutorial on the use of bpp for species tree estimation and species delimitation. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 5.6, pages 5.6:1–5.6:16. No commercial publisher | Authors open access book.
- Fox, G. E., Pechman, K. R., and Woese, C. R. (1977). Comparative cataloging of 16s ribosomal ribonucleic acid: molecular approach to procaryotic systematics. *International Journal of Systematic and Evolutionary Microbiology*, 27(1):44–57.
- Fujisawa, T. and Barraclough, T. G. (2013). Delimiting species using single-locus data and the generalized mixed yule coalescent approach: a revised method and evaluation on simulated data sets. *Syst. Biol.*, 62:707–724.



- Goodfellow, M. (1971). Numerical taxonomy of some nocardioform bacteria. *Journal of General Microbiology*, 69:33–80.
- Grummer, J. A., Bryson Jr, R. W., and Reeder, T. W. (2013). Species delimitation using Bayes factors: simulations and application to the *sceloporus scalaris* species group (squamata: Phrynosomatidae). *Systematic biology*, 63(2):119–133.
- Harpending, H. C., Sherry, S. T., Rogers, A. R., and Stoneking, M. (1993). The genetic structure of ancient human populations. *Current Anthropology*, 34:483–496.
- Hebert, P. D. N., Cywinska, A., Ball, S. L., and deWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270:313–321.
- Hebert, P. D. N., Stoeckle, M. Y., Zemplak, T. S., and Francis, C. M. (2004). Identification of birds through DNA barcodes. *PLoS Biology*, 2:e312.
- Heled, J. and Drummond, A. J. (2009). Bayesian inference of species trees from multilocus data. *Molecular biology and evolution*, 27(3):570–580.
- Hickerson, M. J., Meyer, C. P., and Moritz, C. (2006). DNA barcoding will often fail to discover new animal species over broad parameter space. *Systematic Biology*, 55:729–739.
- Hudson, R. R. and Coyne, J. A. (2002). Mathematical consequences of the genealogical species concept. *Evolution*, 56:1557.
- Hudson, R. R. and Turelli, M. (2003). Stochasticity overrules the “three-times rule”: genetic drift, genetic draft, and coalescence times for nuclear loci versus mitochondrial DNA. *Evolution*, 57:182.
- Jackson, N., Carstens, B., Morales, A., and B.C., O. (2017). Species delimitation with gene flow. *Syst. Biol.*, 66:799–812.
- Jörger, K. M. and Schrödl, M. (2013). How to describe a cryptic species? practical challenges of molecular taxonomy. *Frontiers in Zoology*, 10:59.
- Kapli, P., Lutteropp, S., Zhang, J., Kobert, K., Pavlidis, P., Stamatakis, A., and Flouri, T. (2017). Multi-rate poisson tree processes for single-locus species delimitation under maximum likelihood and markov chain monte carlo. *Bioinformatics*, 33(11):1630–1638.
- Knowles, L. L. (2008). Why does a method that fails continue to be used? *Evolution*, 62:2713–2717.
- Knowles, L. L. and Carstens, B. C. (2007). Delimiting species without monophyletic gene trees. *Systematic Biology*, 56:887–895.
- Leaché, A. D., Fujita, M. K., Minin, V. N., and Bouckaert, R. R. (2014). Species delimitation using genome-wide SNP data. *Systematic biology*, 63(4):534–542.
- Leaché, A. D., Zhu, T., Rannala, B., and Yang, Z. (2018). The spectre of too many species. *Systematic Biology*, 68:168–181.
- Leaché, A. D., Zhu, T., Rannala, B., and Yang, Z. (2019). The spectre of too many species. *Syst. Biol.*, 68(1):168–181.
- Meier-Kolthoff, J. P., Auch, A. F., Klenk, H.-P., and Göker, M. (2013). Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics*, 14:60.
- Moritz, C. and Cicero, C. (2004). DNA barcoding: promise and pitfalls. *PLoS Biology*, 2:e354.
- Nielsen, R., Akey, J. M., Jakobsson, M., Pritchard, J. K., Tishkoff, S., and Willerslev, E. (2017). Tracing the peopling of the world through genomics. *Nature*, 541:302.
- Nielsen, R. and Matz, M. (2006). Statistical approaches for DNA barcoding. *Systematic Biology*, 55:162–169.

- Olave, M., Sola, E., and Knowles, L. L. (2014). Upstream analyses create problems with dna-based species delimitation. *Syst. Biol.*, 63(2):263–271.
- O'Meara, B. C. (2009). New heuristic methods for joint species delimitation and species tree inference. *Systematic Biology*, 59:59–73.
- Palumbi, S. R., Cipriano, F., and Hare, M. P. (2007). Predicting nuclear gene coalescence from mitochondrial data: the three-times rule. *Evolution*, 55:859–868.
- Pei, J., Chu, C., Li, X., Lu, B., and Wu, Y. (2018). CLADES: a classification-based machine learning method for species delimitation from population genetic data. *Molecular Ecology Resources*, 18:1144–1156.
- Petit, R. J. (2008). The coup de grâce for the nested clade phylogeographic analysis? *Molecular Ecology*, 17:516–518.
- Pinho, C. and Hey, J. (2010). Divergence with gene flow: models and data. *Ann. Rev. Ecol. Evol. Syst.*, 41:215–230.
- Pons, J., Barraclough, T. G., Gomez-Zurita, J., Cardoso, A., Duran, D. P., Hazell, S., Kamoun, S., Sumlin, W. D., and Vogler, A. P. (2006). Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *Systematic Biology*, 55:595–609.
- Porter, A. H. (1990). Testing nominal species boundaries using gene flow statistics: The taxonomy of two hybridizing admiral butterflies (Limenitis: Nymphalidae). *Systematic Zoology*, 39:131.
- Puillandre, N., Lambert, A., Brouillet, S., and Achaz, G. (2012). ABGD, Automatic Barcode Gap Discovery for primary species delimitation. *Molecular ecology*, 21:1864–1877.
- Rannala, B. (2015). The art and science of species delimitation. *Current Zoology*, 61:846–853.
- Rannala, B., Edwards, S. V., Leaché, A., and Yang, Z. (2020). The multi-species coalescent model and species tree inference. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 3.3, pages 3.3:1–3.3:21. No commercial publisher | Authors open access book.
- 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.
- Rannala, B. and Yang, Z. (2013). Improved reversible jump algorithms for Bayesian species delimitation. *Genetics*, 194:245–253.
- Rannala, B. and Yang, Z. (2017). Efficient Bayesian species tree inference under the multispecies coalescent. *Syst. Biol.*, 66:823–842.
- Rogers, A. R. and Harpending, H. (1992). Population growth makes waves in the distribution of pairwise genetic differences. *Molecular Biology and Evolution*, 9:552–569.
- Sites Jr, J. W. and Marshall, J. C. (2004). Operational criteria for delimiting species. *Annu. Rev. Ecol. Evol. Syst.*, 35:199–227.
- Smith, M. A., Fisher, B. L., and Hebert, P. D. N. (2005). Dna barcoding for effective biodiversity assessment of a hyperdiverse arthropod group: the ants of madagascar. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360:1825–1834.
- Smith, M. L. and Carstens, B. C. (2019). Process-based species delimitation leads to identification of more biologically relevant species. *Evolution*, 74(2):216–229.
- Sneath, P. H. A. (1957a). The application of computers to taxonomy. *Microbiology*, 17:201–226.
- Sneath, P. H. A. (1957b). Some thoughts on bacterial classification. *Journal of General Microbiology*, 17:184–200.
- Sneath, P. H. A. (1976). Phenetic taxonomy at the species level and above. *Taxon*, pages 437–450.

- Solís-Lemus, C., Knowles, L. L., and Ané, C. (2015). Bayesian species delimitation combining multiple genes and traits in a unified framework. *Evolution*, 69(2):492–507.
- Sukumaran, J. and Knowles, L. L. (2017). Multispecies coalescent delimits structure, not species. *Proc. Natl. Acad. Sci. U.S.A.*, 114(7):1607–1612.
- Tautz, D., Arctander, P., Minelli, A., Thomas, R. H., and Vogler, A. P. (2003). A plea for DNA taxonomy. *Trends in Ecology & Evolution*, 18:70–74.
- Templeton, A. R. (2001). Using phylogeographic analyses of gene trees to test species status and processes. *Molecular Ecology*, 10:779–791.
- Wayne, L., Brenner, D., Colwell, R., Grimont, P., Kandler, O., Krichevsky, M., Moore, L., Moore, W., Murray, R., Stackebrandt, E., et al. (1987). Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. *International Journal of Systematic and Evolutionary Microbiology*, 37:463–464.
- Wiens, J. J. and Servedio, M. R. (2000). Species delimitation in systematics: inferring diagnostic differences between species. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 267:631–636.
- Will, K. W. and Rubinoff, D. (2004). Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification. *Cladistics*, 20:47–55.
- Wilson, K. H. (1995). Molecular biology as a tool for taxonomy. *Clinical Infectious Diseases*, 20:S117–S121.
- Wolpert, D. H. (2002). The supervised learning no-free-lunch theorems. In *Soft computing and industry*, pages 25–42. Springer.
- Wu, D.-D., Ding, X.-D., Wang, S., Wojcik, J. M., Zhang, Y., Tokarska, M., Li, Y., Wang, M.-S., Faruque, O., Nielsen, R., Zhang, Q., and Zhang, Y.-P. (2018). Pervasive introgression facilitated domestication and adaptation in the bos species complex. *Nature Ecol. Evol.*, 2(7):1139–1145.
- Yang, Z. (2015). The BPP program for species tree estimation and species delimitation. *Curr. Zool.*, 61(5):854–865. <http://dx.doi.org/10.1093/czoolo/61.5.854>.
- Yang, Z. and Rannala, B. (2010). Bayesian species delimitation using multilocus sequence data. *Proceedings of the National Academy of Sciences USA*, 107:9264–9269.
- Yang, Z. and Rannala, B. (2014). Unguided species delimitation using DNA sequence data from multiple loci. *Mol. Biol. Evol.*, 31(12):3125–3135.
- Zhang, C., Rannala, B., and Yang, Z. (2014). Bayesian species delimitation can be robust to guide tree inference errors. *Syst. Biol.*, 63(6):993–1004.
- Zhang, C., Zhang, D.-X., Zhu, T., and Yang, Z. (2011). Evaluation of a Bayesian coalescent method of species delimitation. *Syst. Biol.*, 60:747–761.
- Zhang, J., Kapli, P., Pavlidis, P., and Stamatakis, A. (2013). A general species delimitation method with applications to phylogenetic placements. *Bioinformatics*, 29:2869–2876.