The art and science of species delimitation

Bruce RANNALA[*]

Department of Evolution and Ecology, University of California Davis Davis, CA 95616, USA

Abstract DNA-based approaches to systematics have changed dramatically during the last two decades with the rise of DNA barcoding methods and newer multi-locus methods for species delimitation. During the last half-decade, partly driven by the new sequencing technologies, the focus has shifted to multi-locus sequence data and the identification of species within the framework of the multi-species coalescent (MSC). In this paper, I discuss model-based Bayesian methods for species delimitation that have been developed in recent years using the MSC. Several approximate methods for species delimitation (and their limitations) are also discussed. Explicit species delimitation models have the advantage of clarifying more precisely what is being delimited and what assumptions we are making in doing so. Moreover, the methods can be very powerful when applied to large multi-locus datasets and thus take full advantage of data generated using today's technologies [*Current Zoology* 61 (5): 846–853, 2015].

Keywords Bayesian species delimitation, Species concepts, Multi-species coalescent

DNA-based approaches to systematics have undergone a revolution during the last two decades. First, the rise of DNA barcoding initiatives during the late 1990s, aimed at using single-locus genetic sequences to identify and assign samples to species, generated large-scale taxonomy-oriented sequencing projects in several countries (Schindel and Miller, 2005; Hebert and Gregory, 2005). This was followed by extensive debate about the feasibility of single-locus DNA barcoding (Moritz and Cicero, 2004; Hickerson et al., 2006). During the last half-decade, partly driven by the new sequencing technologies, the focus has shifted to multi-locus sequence data and the identification of species within the framework of the multi-species coalescent (MSC) (Knowles and Carstens, 2007; Yang and Rannala, 2010). In this paper, I do not attempt to review the extensive literature on genetic delimitation of species but focus mainly on the model-based approaches that have been developed in recent years using the MSC. Explicit models have the advantage of clarifying more precisely what is being delimited and what assumptions we are making in doing so. Moreover, the methods can be very powerful when applied to large multi-locus datasets and take full advantage of today's technologies. By focusing on model-based approaches the aim is to provide one possible path through the maze of methods proposed for species delimitation in recent years. Other paths exist but I will not explore them here.

1　The Objective

A fundamental difficulty facing biologists interested in the genetic delimitation of species is that in order to delimit species they must first be defined. Species definitions intermingle with species concepts and the lack of concensus in this field poses a serious dilemma for the "delimiters" (Sites and Marshall, 2003). If systematists cannot agree on what defines a species how can geneticists possibly develop objective methods to identify one? Moreover, even if one were willing to adopt a particular species concept it is difficult to translate a verbal "concept" into a well-defined mathematical model. An obvious candidate for such a translation is the "Biological Species Concept" (BSC) – the requirement of reproductive incompatibility between species (Mayr, 1976). The BSC does not lead to an easily identifiable model for use in genetic analysis, however, unless the genes involved in reproductive isolation are known. For example, a pair of allopatric species might have been completely isolated for a million years, which would be readily evident from a multi-locus genetic analysis, but genetic isolation alone does not prove that the species are incapable of interbreeding.

One possible resolution to the ambiguity of genetic predictions derived from species concepts is to turn the problem on its head and instead ask whether characteristic patterns of genetic divergence emerge when we

---

∗ Corresponding author. E-mail: brannala@ucdavis.edu

compare groups that systematists have recognized as different species, comparing these with patterns for groups labeled as subspecies, as populations, and so on. Despite some success with particular taxonomic groups (Hebert et al., 2004), such approaches have most often failed to generate useful "rules of thumb." For example, the so-called 10x rule for interspecific versus intraspecific distances (Hebert et al., 2004) does not appear to be generally useful (Hickerson et al., 2006). Hey and Pinto (Hey and Pinho, 2012) applied an "empirical" delimitation approach to a range of groups and concluded that few consistent patterns emerge when comparing populations versus species (and thus that species designations may be arbitrary).

One problem with the empirical delimitation approach is that it presupposes that the traditional systematics is correct (i.e., that existing species and population boundaries provide a relevant guide); this may not be true in general. A second problem is that many empirical delimitation studies have used population-level measures of differentiation that may not provide an appropriate diagnostic. Measures of genetic differentiation based on allele (or genotype) frequencies, such as F_{ST}, can increase rapidly when severe population bottlenecks occur, founding events, etc. For example, if a population initially has allele frequency 0.5 at a locus and a new population is founded by 2 randomly chosen diploid individuals the new population is fixed for the allele with probability (1/8). Such large variances among populations lead to large F_{ST} values. Thus, even recently isolated populations can show large levels of differentiation under certain circumstances. On the other hand, consistently large populations accumulate frequency differences slowly and may show little differentiation even if completely isolated for hundreds of generations. These properties at least partially explain the frequent lack of agreement of differentiation measures such as F_{ST} with species (or population) status in empirical delimitation studies. Studies of average divergence for a single locus (Hebert et al., 2004) are prone to large variance introduced by the random processes of coalescence and mutation accumulation (Ross et al., 2008).

The focus of Bayesian species delimitation methods developed over the last decade has been on the use of multi-locus genetic data with the basic premise being that the groups we call species are genetically isolated on an evolutionary timescale. The boundaries between "evolutionary" or "mutational" versus "population genetic" timescales of isolation are fairly well def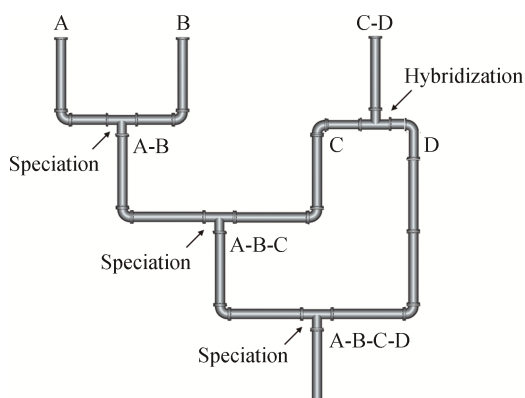ined if we consider an explicit population genetic model. A "population genetic" timescale is defined here as below the level at which mutations accumulate and would be dominated by frequency differences due to genetic drift, selection, etc. Model-based methods are developed using a multi-species coalescent (MSC) framework (Takahata et al., 1995; Rannala and Yang, 2003) in which gene trees evolve within populations that may, or may not, be connected by migration. One tentatively defines as a species any population that is significantly supported as a distinct lineage in a species-tree analysis (see below). Simulation studies (Zhang et al., 2011) suggest the migration threshold for genetic isolation is about 1 individual per 10 generations. This determines the level of gene flow below which model-based genetic delimitation identifies putative species.

One can also consider the time duration during which gene flow must be below the migration threshold in order for a species to be delimited. The rate of mutation for nuclear genes for a broad range of animals is on the order of 10^{-8} to 10^{-10} mutations per site per year. Given these rates, even with whole genome data it is difficult to obtain statistical support for species divergences less than about 10^{-6} in units of expected DNA substitutions per site per year. Thus, a bound for the divergence time of an incipient species that can be identified on a species tree is about 1,000 years ($10^{-9} \times 1,000$ years). The basic properties of a group that delimit it as a species using model-based delimitation methods are then that it has received less than about 1 migrant per 10 generations for at least 1,000 years. In practice, the threshold for divergence time based on the available loci is probably at least one or two orders of magnitude greater than this – for example, a dataset with 50 loci, each 1 kb in length, would not allow a divergence event to be detected if it occurred less than about 10,000 years ago ($50 \times 10,000 \times 10^{-9} = 5 \times 10^{-4}$). Although (Yang and Rannala, 2010.) have previously associated the model underlying Bayesian species delimitation methods with the Biological Species Concept, reproductive isolation is a sufficient condition for identification of species under the model but not a necessary one (the delimited species may be allopatric). It is more accurate to consider an operational definition of the method as presented above. However, the above description should only be interpreted as a rough guideline – more analysis is needed to determine precise bounds on the conditions for species delimitation using the MSC framework.

2 The Model

At this point we have an outline of a "species defini-

tion" corresponding to the basic genetic model described above, we now consider that model in more detail. Genetic isolation on an evolutionary time-scale is the property that allows species to be recognized using molecular genetic delimitation techniques. Under such a paradigm, one can visualize species genetic histories as an assemblage of horizontal and vertical pipes (Fig. 1). Each pipe is a conduit for gene lineages; the horizontal pipes represent an admixture of contemporaneous species and the vertical pipes represent the continuity of a species through time. If a speciation event gives rise to two descents then an upward T-fitting occurs, while if a hybridization of two species produces a single descendent species a downward T occurs (Fig. 1). One can take the analogy slightly further and imagine that the cross-sectional area of the vertical pipes represents effective population sizes and that of the horizontal pipes represents the flow of migrants. Here the analogy becomes strained because gene flow may be asymmetrical between populations, effective population sizes may vary across the genome and through time, and so on. Fortunately, the precise details of migration and population size change do not need to be known to identify species and represent their history. As noted above, there is a threshold effect to gene flow in terms of its impact on the future destiny of a species. If less that about 1 individual per 10 generations is exchanged between two species they will persist as independently evolving units while if greater numbers of individuals are exchanged they will behave as a single interbreeding

species. Thus, we can interpret a horizontal pipe as indicating sufficient gene flow to create a single species (from a genetic perspective). Of course, we could allow more than 3 pipes to connect at a speciation or hybridization event, etc, but this is unnecessary since we can allow arbitrarily short pipes connecting events. A species graph can be visualized as the outcome of this plumbing exercise (Fig. 1).

Here, we will ignore the horizontal pipes and assume that a binary tree (only vertical pipes and upward T-fittings) can be used to represent the relationships among species through time. Most parametric species delimitation methods are based on this model, which is often adequate. It is feasible to expand the model to include hybrid species but this scenario has not been implemented in current approaches. The binary tree of relationships among species is the so-called "species tree." A gene tree, on the other hand, is a binary tree representing the history of a region of a genome sampled from one or more contemporary individuals. The splitting points in the gene tree represent the coalescence of lineages to a common ancestor. Each lineage of a gene tree is associated with a particular (contemporary or ancestral) species at any point in the history. At a particular instant in time, gene tree lineages that are found in different species cannot coalesce to a common ancestor – the species tree constrains the possible coalescence events among lineages in each gene tree (Rannala and Yang, 2003).

The probability distribution of gene trees under a MSC model is completely determined once the species tree topology and branch lengths, and the ancestral population sizes, have been specified (Rannala and Yang, 2003). The contemporary and ancestral effective population sizes are normally scaled in units of expected substitutions and represented by the commonly used population genetic parameter $\theta = 4N\mu$, where N is the effective population size and μ is the mutation rate per site per generation. The expected number of substitutions per site between a randomly chosen pair of sequences from a population is equal to θ under the neutral model. The MSC model is the theoretical basis for both Bayesian and likelihood inference methods of species delimitation. However, in a Bayesian framework one can think of the MSC as a prior on gene trees; the posterior density of gene trees may be quite different from the prior – it is a mistake to assume that if the neutral coalescent model is incorrect the delimitation method will not work. The Bayesian method could be robust to the prior and perform well even in cases where

Fig. 1 Representation of the independent evolutionary units in a multi-species coalescent tree of three contemporary species (A, B, and C-D) as a series of "pipes"
There are three ancestral speciation events and one ancestral hybridization event. The ancestral species are A-B, A-B-C, C, D, and A-B-C-D. An upward facing pipe "T" fitting represents a hybridization event and a downward facing T-fitting represents a speciation event.

strong directional selection is operating on some genes (Edwards, 2009) – further simulation studies are needed to determine the robustness of Bayesian species delimitation to violations of the neutral coalescent model. With many loci, as in whole-genome datasets it appears likely that robustness can be achieved. By collapsing, or expanding nodes on the species tree one can represent different species delimitation models and by modifying branch lengths or topology one can represent different phylogenetic relationships among the delimited species.

3 The Methods

Several authors have classified species delimitation methods as either "discovery" or "validation" procedures (see O'meara, 2010). "Discovery" methods are typically "assignment" methods, aiming to identify genetic substructure in populations using genotype frequencies. One property of assignment methods (relevant to species delimitation) is that with sufficient numbers of loci they will detect very subtle differences between populations – even ones that have experienced recent or ongoing gene flow and would not normally qualify as species. This explains the need for a subsequent "validation" of the putative species using another method (Rittmeyer and Austin, 2012). It is arguably more accurate to describe such methods as population substructure detection methods rather than species discovery methods because they will often detect structure at levels well below the species. Moreover, in some cases, the population substructure is quite obvious (the potential species may be allopatric for example) and an explicit "discovery" step may not be needed. Finally, one can use alternative discovery approaches such as partitioning populations into subsets and iteratively applying "validation" methods without a prior discovery step. As the computational efficiency of validation methods increases it may become possible to increase the number of populations used in a "validation" analysis, or even treat each diploid genome as a potential species, thus eliminating the need for a discovery step. For these reasons, we focus here exclusively on the class of methods described as species validation methods – we refer to these methods simply as species delimitation methods.

Essentially all model-based species delimitation methods assume the existence of a species tree with gene trees determined by the MSC as described in the previous section. However, most methods simplify the inference procedure in one or more ways. The methods can be broadly classified into two types: (1) Data-limited methods that can only analyze particular forms of genetic data (e.g., a single non-recombining locus, or unlinked single nucleotide polymorphisms [SNPs]); and (2) Pseudo-data methods that make inferences based on inferred gene trees that are treated as if they were observations. All of these approaches reduce the computational expense and/or analytical complexity of the inference problem but at the cost of potentially increased bias and/or reduced power and accuracy. Fully Bayesian methods have been developed but require a greater computational investment. Here, we will begin by describing the general Bayesian inference framework and then outline the various alternative methods.

3.1 Bayesian species delimitation

The delimitation model, M, includes both the species delimitation and the phylogeny. If $\mathbf{G}$ is the set of gene trees for the sampled loci, $\mathbf{D}$ is the multi-locus sequence data, and Ω is the set of priors and model parameters (the dimension of Ω is variable), the posterior probability of M is

$$f\left(M|\mathbf{D},\beta\right) =$$

$$\frac{1}{f(\mathbf{D})}\int_{\Omega}\int_{\mathbf{G}} f\left(\mathbf{G}|M,\Omega\right) f\left(M|\Omega\right) f\left(\mathbf{D}|\mathbf{G},\Omega\right) f\left(\Omega|\beta\right) d\mathbf{G}d\Omega,$$

where β defines a set of fixed prior (or hyper-prior) parameters. The program BPP (Yang and Rannala, 2014) uses a reversible-jump Markov chain Monte Carlo method to numerically evaluate this probability. The *Beast module Dissect (Jones et al., 2014) uses a Dirac delta function to mimic the effect of collapsing nodes on the species tree, also with the aim of generating the posterior probabilities of delimitation models. The above formulation is entirely general, although current Bayesian implementations assume that the genetic loci are unlinked and that there is no intra-locus recombination. We now consider several alternative delimitation methods that aim to evaluate these probabilities by different means, or approximations.

3.2 Data-limited methods

The first class of alternative methods that we consider are those restricted to particular data types. Leache et al. (Leaché et al., 2014) used a Bayes Factor (BF) approach adapting the SNAPP program (Bryant et al., 2012) for phylogenetic inference using SNP data to calculate marginal likelihoods for use in comparing BFs of particular delimitation models. Strengths of this approach include the fact that SNAPP can analytically integrate over gene trees; this might offer an advantage over methods that numerically integrate over gene trees using MCMC in some circumstances. However, the cur-

rent implementation of (Leaché et al., 2014) relies on an MCMC path sampling framework to estimate marginal likelihoods which can be computationally intensive. Two limitations of the method are: (1) it may only use unlinked SNP loci and each SNP contains relatively little information about the underlying gene tree at a locus -- the branch lengths are particular poorly specified; (2) it requires that the models for comparison using BFs be specified a priori by the user as it is not practical to enumerate all possible model BFs for more than a small number of populations. This second problem applies to other recent BF-based methods (Grummer et al., 2013) as well. Another method that is data-limited is the Generalized Mixed Yule Coalescent (GMYC) model (Fujisawa and Barraclough, 2013) which uses a prior on the gene tree that assumes a mixture of a Yule branching process for the species tree and a coalescent model within populations – this cannot be easily extended to multiple loci.

3.3   Pseudo-data methods

Another recently proposed class of species delimitation methods attempt to simplify the analysis by treating point estimates of the gene trees as if they were observed data, with the MSC prior then playing the role of a likelihood function. The program spedeSTEM (Ence and Carstens, 2011) applies this approach, using Akaike Information Criterion (AIC) scores(Akaike, 1974) to choose among different delimitation models, with likelihoods calculated using maximum likelihood estimates of the gene trees. An advantage of this strategy is that it is relatively simple to estimate gene trees and maximize the likelihood when not integrating over gene trees. There are two clear disadvantages however: (1) the method does not correctly account for the uncertainty of gene tree topologies and branch lengths; (2) the AIC score is based on an asymptotic approximation and may perform poorly in practice as a model choice criterion. The AIC approach is particularly problematic in this case because the θ parameters are assumed equal for all populations in the spedeSTEM model and are user-specified, rather than estimated from the data. The AIC approximation is derived based on an assumption that the likelihoods are calculated using maximum likelihood estimates of the unknown parameters.

4   Concordance among Delimitation Methods: A Measure of Accuracy?

In a recent paper entitled "How to fail at species delimitation" Carsten et al. (2013) noted that the "poten-

tial parameter space relevant to species delimitation is larger and far more complex than that considered by even the most heavily parameterized of existing methods" and that a "naïve response to the above conundrum is to identify a single method that is demonstrably accurate in some simulation study and apply it alone to the data." The authors suggest that simulation studies designed to identify such methods have shortcomings because "no simulation study has included every potentially useful method" and "results from simulation studies are conditional on the specific attributes of the simulated data used in such studies." They conclude with the recommendation that many delimitation methods be applied and delimitations accepted that are common to all (or most) of the methods (as was done by Satler et al., 2013). This sounds like reasonable advice. However, following this advice will generally lead to the outcome that few or no species will be delimited. In fact, the more studiously a researcher applies this advice -- including additional delimitation methods and requiring agreement among them -- the greater is the likelihood that she will "fail at species delimitation."

Carsten et al's (2013) comment regarding the complexity of the parameter space is, of course, true – reality is always more complex than any particular statistical model. However, satisfactory statistical tests, or estimators, do not normally result from simply maximizing the number of parameters (the "biological realism") of the model. In virtually every situation, some aspects of the problem (additional model parameters) have an important influence on estimates of the parameters of interest, or the result of a hypothesis test, and others do not. Often the best approach is to endeavor to identify the model components that have an important effect on inferences. It is usually straightforward to examine the effect of including or excluding various model parameters, as well as the effects of violations of model assumptions, either by analytical analysis or by simulations. In the case of species delimitation, the current situation is even more straightforward since essentially all the existing parametric methods are based on the same likelihood function – the main differences among the methods are their use of different short-cuts, such as assuming that gene trees are known without error (which discards information and can introduce bias), or a reduced data structure (such as unlinked SNPs rather than sequence data). Short-cut methods are employed for analytical or computational expedience – they do not represent alternative "models". Clearly the statistical performance of such methods cannot exceed that of the

full-likelihood or Bayesian method, although the computations may be faster and issues such as MCMC convergence of less or no concern. This is very different from the situation where alternative models are employed in different methods and it is not clear which one provides better results.

Carstens at al. (2013) suggest that it is naïve to attempt to choose an optimal method for use in inference and that a cautious analyst should instead apply many of the available methods and look for agreement between them. This idea, a sort of "statistical democracy", has appeared before in the phylogenetics literature. Kim (1993), for example, argued that if UPGMA clustering, maximum parsimony, and neighbor-joining all agreed on the same tree this indicated its reliability (he proposed a reliability index for this purpose). The problem with this approach is that several methods with poor statistical properties can nonetheless make similar inferences because their underlying algorithms are similar – their agreement does not then increase the likelihood that a result is correct. Moreover, if one of the estimators is superior in some situation and generates a very different result it is discounted -- a "tyranny of the majority" arises. Statistical democracy also puts no constraints on the permissible tests, or estimators. Clearly there are many more ways to construct bad statistical tests, or estimators, than there are to construct good ones and it follows that the more tests we include in the composite test the worse the tests will be on average.

To illustrate the effect of combining tests we consider a simple example. Imagine a "good" test T_1 that has power $1 - \beta$ and type I error rate α and a "bad" test T_2 with a p-value that is uniformly distributed on $(0,1)$ irrespective of the data. Both T_1 and T_2 will have type I error rate α. However, test T_2 also has power α. Now consider a more "conservative" composite test,

$$T_3 = T_1 \cap T_2$$

Namely, we reject the null hypothesis only in the case that both T_1 and T_2 reject. The test is indeed conservative; the type I error rate is now α^2. However, the power of T_3 is $(1 - \beta)\alpha$. For example, if $\alpha = 0.01$ and $\beta = 0.2$ test T_1 has type I error rate 0.01 and power 0.8. Test T_2 has type I error rate 0.01 and power 0.01. The combined test T_3 has type I error rate 0.0001 and power 0.008. This example may seem contrived since the second test returns a random p-value. However, it illustrates two things: first, if the sampling distributions of the two test statistics are not identical the type I error rate of the composite test is less than any of the component tests (it is more conservative). Second, the power

of the composite test is less than the power of the least powerful component test – in other words there is a large penalty for adding a poor test. Furthermore, the more component tests that we add the less powerful the composite test becomes and if the component tests have very similar sampling distributions the sampling distribution of the composite test will be similar to that of every component. In that case, the composite test offers no particular benefit – one would do just as well using any one of the components.

To summarize, if one or more of the tests are bad the composite test will incur a large reduction in power with the only benefit being a decrease in type I error. If all the tests are good the composite test offers essentially no benefit. In general, a better strategy is to attempt to eliminate poor tests from consideration using simulations (or analysis) to evaluate them and if one is concerned about type I error to use a more stringent criteria for significance when applying the best test.

If we apply "statistical meritocracy" as our criterion and search for a best test we require a means for comparing the statistical performance of the available tests. Given the complex model we deal with in species delimitation this will often involve simulation studies. Carstens et al. (2013) observe that one can't compare the performance of all possible methods nor can one evaluate the performance of any given method for all possible combinations of parameters and conclude that many (or all) available methods should be used since none can be excluded apriori. As noted above, this approach carries large penalties in terms of power and offers no clear statistical advantage. The basis for this recommendation is also not well justified. First, we do not need to evaluate all the available methods using simulations because many of them, pseudo-data methods for example, or methods based on summary statistics rather than the full likelihood, cannot perform better than the full-likelihood or Bayesian method so the question comes down to how much worse they are. If we are not concerned about computational expense and just want to find the best estimator we can safely exclude them. Second, we do not need to examine the entire parameter space when comparing a given set of methods via simulation because we can often get a good indication of the relative performance of methods by simulating over a subset of the space. If the state space for simulations is reasonably large and one method outperforms another over all the simulation conditions, for example, it is unlikely that the opposite will occur in the un-sampled region of parameter space. Moreover, it is

often feasible to randomly sample the state space to obtain combinations of parameters for the simulation – this is essentially like making inferences about a population from a sample rather than a complete census. If the state-space sampling is carefully implemented this can be as informative as an exhaustive analysis would be.

In the above discussion we have advocated the statistical meritocracy approach (exclusive use of the best method) based solely on the statistical properties of the methods. However, there is also the practical concern that all computer programs contain implementation errors, numerical instabilities, and other algorithmic problems that can cause erroneous inferences (this is a failure of the implementation not the method). For this reason, we advocate the use of two or more algorithmically similar programs (i.e., based on the same or very similar statistical models) when possible. A bug in one of the programs can then be detected as a discrepancy between the results obtained using each program. Moreover, with MCMC programs one should also run each individual program multiple times with the same priors but different random number generator seeds to examine whether the results are consistent for a given program. Any discrepancies observed between the results obtained from either within or between program runs should be resolved (or otherwise explained) before publishing the results if at all possible.

5 How to Succeed at Species Delimitation

With a new version of BPP now available that jointly infers the phylogeny and delimitation (Yang and Rannala, 2014) one of the major concerns about Bayesian delimitation -- the effect of errors in the guide tree (Leaché and Fujita, 2010; Olave et al., 2013; Zhang et al., 2014) -- has been resolved. One remaining concern for the Bayesian approach is the sensitivity of delimitation outcomes to the priors on the root age τ and the population genetic parameter θ. Leache and Fujita (Leaché and Fujita, 2010) suggested that a large prior mean on θ and a small prior mean on τ are conservative (reduce the probability of splitting). However, this does not seem to be true in general; although a larger mean for the prior on θ tends to decrease the delimitation probability the effect of changing the prior on τ is less predictable. Leaché and Fujita (2010) examined the effect of changing the prior means on τ and θ by two orders of magnitude from (0.001 to 0.1). This is often too extreme – a better strategy is to set the means of both priors to reasonable values (same order of magni-

tude) based on a preliminary analysis and then to vary the prior means above and below this value within one order of magnitude of the preliminary estimates. The shape parameter α should be between 1 and 2 so that the standard deviation is roughly proportional to the mean. It is well known that by choosing a sufficiently extreme prior one can always influence the posterior distribution for finite datasets. Thus, it is important to allow only biologically reasonable values for the priors (often within an order of magnitude of the value suggested by the data).

We conclude with several recommendations for successful species delimitation: (1) use multi-locus sequence data; (2) use explicit model based approaches and be aware of the assumptions; (3) if using Bayesian methods examine the effect of the priors but do not use priors that are too extreme. We advocate the use of model-based methods for several reasons. First, many heuristic methods that are not derived based on an explicit model may make implicit assumptions; as the old saying goes "Better the Devil you know than the Devil you "t". Second, a model allows one to understand what is being inferred which is preferable to the use of a black box method with unknown properties. Finally, in a parametric framework likelihood and Bayesian based inference approaches lead to estimators and hypothesis tests with good statistical properties, especially for large samples. Model-based species delimitation is still in the early stages of development but several current programs already allow analyses of relatively large multi-locus datasets (hundreds of loci and a dozen or more populations). The availability of whole-genome sequence data offers many computational challenges that should drive new developments in the field for some time to come.

References

Akaike H, 1974. A new look at the statistical model identification. IEEE Trans Autom Contr AC 19: 716–723.

Bryant D, Bouckaert R, Felsenstein J, Rosenberg NA, Roychoudhury A, 2012. Inferring species trees directly from biallelic genetic markers: Bypassing gene trees in a full coalescent analysis. Mol. Biol. Evol. 29: 1917–1932.

Carstens BC, Pelletier TA, Reid NM, Satler JD, 2013. How to fail at species delimitation. Mol. Ecol. 22: 4369–4383.

Edwards SV, 2009. Natural selection and phylogenetic analysis. Proc. Natl. Acad. Sci. U. S. A. 106: 8799–8800.

Ence DD, Carstens BC, 2011. SpedeSTEM: A rapid and accurate method for species delimitation. Mol. Ecol. Resour. 11: 473–480.

Fujisawa T, Barraclough TG, 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.

G rummer JA, Bryson RW, Reeder TW, 2013. Species delimitation using Bayes factors: Simulations and application to the *Sceloporus scalaris* species group (Squamata: Phrynosomatidae). Syst. Biol. 63: 119–133.

Hebert PDN, Gregory TR, 2005. The promise of DNA barcoding for taxonomy. Syst. Biol. 54: 852–859.

Hebert PDN, Stoeckle MY, Zemlak TS, Francis CM, 2004. Identification of birds through DNA barcodes. PLoS Biol. 2: e312.

Hey J, Pinho C, 2012. Population genetics and objectivity in species diagnosis. Evolution 66: 1413–1429.

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

Jones G, Aydin Z, Oxelman B, 2014. DISSECT: An assignment-free Bayesian discovery method for species delimitation under the multispecies coalescent. Bioinformatics. doi: 10.1093/bioinformatics/btu770

Kim J, 1993. Improving the accuracy of phylogenetic estimation by combining different methods. Syst. Biol. 42: 331–340.

Knowles LL, Carstens BC, 2007. Delimiting species without monophyletic gene trees. Syst. Biol. 56: 887–895.

Leaché AD, Fujita MK, 2010. Bayesian species delimitation in West African forest geckos *Hemidactylus fasciatus*. Proc. Biol. Sci. 277: 3071–3077.

Leaché AD, Fujita MK, Minin VN, Bouckaert RR, 2014. Species delimitation using genome-wide SNP Data. Syst. Biol. 63: 534–542.

Mayr E, 1976. Species concepts and definitions. In: Grene M, Mendelsohn E ed. Topics in the Philosophy of Biology. Springer Netherlands, 353–371.

Moritz C, Cicero C, 2004. DNA barcoding: Promise and pitfalls. PLoS Biol. 2: e354.

Olave M, Solà E, Knowles LL, 2013. Upstream analyses create problems with DNA-based species delimitation. Syst. Biol. syt106.

O'meara BC, 2010. New heuristic methods for joint species delimitation and species tree inference. Syst. Biol. 59: 59–73.

Rannala BH, Yang Z, 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164: 1645–1656.

Rittmeyer EN, Austin CC, 2012. The effects of sampling on delimiting species from multi-locus sequence data. Mol. Phylogenet. Evol. 65: 451–463.

Ross HA, Murugan S, Li WLS, 2008. Testing the reliability of genetic methods of species identification via simulation. Syst. Biol. 57: 216–230.

Satler JD, Carstens BC, Hedin M, 2013. Multilocus species delimitation in a complex of morphologically conserved trapdoor spiders (Mygalomorphae, Antrodiaetidae, Aliatypus). Syst. Biol. 62: 805–823.

Schindel DE, Miller SE, 2005. DNA barcoding a useful tool for axonomists. Nature 435: 17.

Sites JW, Marshall JC, 2003. Delimiting species: A Renaissance issue in systematic biology. Trends Ecol. Evol. 18: 462–470.

Takahata N, Satta Y, Klein J, 1995. Divergence time and population-size in the lineage leading to modern humans. Theor. Pop. Biol. 48: 198–221.

Yang Z, Rannala B, 2014. Unguided species delimitation using DNA sequence data from multiple loci. Mol. Biol. Evol. 31: 3125–3135.

Yang Z, Rannala BH, 2010. Bayesian species delimitation using multilocus sequence data. PNAS 107: 9264–9269.

Zhang C, Zhang DX, Zhu T, Yang Z, 2011. Evaluation of a bayesian coalescent method of species delimitation. Syst. Biol. 60: 747–761.

Zhang C, Rannala B, Yang Z, 2014. Bayesian species delimitation can be robust to guide-tree inference errors. Syst. Biol. 63: 993–1004.