CHAPTER TWO

On the arbitrary identification of real species

JODY HEY

Introduction

The detection of a new species is the result of a decision-making process, one that has traditionally and primarily been based upon the discovery of distinguishing characters (Cronquist 1978; Mayr 1982; Winston 1999; Sites & Marshall 2004). This process, called diagnosis, is acutely important as it necessarily lies at the crux of the discovery of biodiversity, including the identification of conservation units. With the rise of quantitative phylogenetic methods, and the increasing accessibility of molecular data, numerous methods for diagnosis have been proposed in recent years (Sites & Marshall 2004).

For many situations the identification of a new species does not particularly require quantitative methodology. These are the cases where the organisms of the putative new species are conspicuously divergent from all known species and, thus, where the new species is identified as a sister taxon to previously identified groups. But increasingly, as more new species are described and as more species are the subject of additional investigation, the questions of diagnosis arise within previously described species, wherein patterns of differentiation among populations of the same species must be interpreted in taxonomic terms. In short: how do we decide when a closer look at one taxonomic species actually reveals the presence of more than one species?

A similar question arises in conservation contexts: do the data from one species (or one conservation unit) actually reveal the presence of multiple units, each of which merit recognition and possibly protection? This question may not be cast in terms of the taxonomic rank of species, and so the criteria used for diagnosis of a new species may differ from those used for diagnosing a population in terms of meriting conservation status. For example, in the United States the language of the amended Endangered Species Act (16 USC §§1531–1544) refers to a 'distinct population segment of any species of vertebrate fish or wildlife which interbreeds when mature'. In effect this conservation policy has identified a taxonomic category of 'distinct population segment' or 'DPS' as it is called in discussions on biodiversity conservation, one for which criteria and protocols for diagnosis have been much discussed (Waples 1991a; Moritz 1994;

Speciation and Patterns of Diversity, ed. Roger K. Butlin, Jon R. Bridle and Dolph Schluter. Published by Cambridge University Press.

Vogler & DeSalle 1994; Waples 1995; Pennock & Dimmick 1997; Dimmick *et al.* 1999; Goldstein *et al.* 2000).

This chapter addresses basic questions about quantitative approaches to the diagnosis of closely related taxa, such as those that might be applied when a single variable species is studied, and where investigators are faced with the question of whether or not different populations of one species each merit some taxonomic status. It is shown how most quantitative methods for determining the taxonomic status of populations base the taxonomic decision on a finding of statistically detectable divergence, and thus base the decision inherently upon the sample size of the study.

On the nature of natural populations

Given that most described species consist of multiple partly differentiated populations (Hughes *et al.* 1997), how might we consider the reality or the objectivity of local populations? Consider as a starting point a single species taxon which possibly includes multiple populations that merit their own taxonomic status. One possible observation is that the different populations are found to be clearly differentiated from one another on the basis of multiple readily observable characters. In such cases, the diagnosis of additional taxa is likely to be supported by any diagnostic method. At the other extreme it is possible that no geographic, morphological or genetic differentiation is observed. Such a case would suggest the presence of a single species and would not meet the diagnostic criteria for additional taxa under any method. It is in between these two extremes, of strong differentiation on the one hand and of random mating with an absence of differentiation on the other hand, where lies the wilderness of taxon diagnosis. This difficult territory includes most of the millions of species that occur as structured populations. In this context 'population structure' is a catch-all that includes a wide array of phenomena, including: hierarchical patterns of population structure, multiple separate populations, isolation by distance and metapopulation dynamics.

Importantly for conspecific populations, all of the processes that give rise to population structure are expected to vary continuously – they are not 'all or none' phenomena. Even locally fixed alleles, which are restricted to one or a subset of populations, will occur effectively as a matter of degree in many situations. Such fixations can occur by chance, or by selection, and they can occur anywhere in the genome. For populations that have been separated for some time with limited gene flow, private alleles are likely to be present to some degree somewhere in the genome. In short, population differentiation is expected to vary continuously in complex ways.

The key corollary of these points is that we do not have a basic expectation that evolution will generate a tangible threshold or trigger point of differentiation, beyond which new taxa are unambiguously diagnosable. Finally, it follows

necessarily that there can be no objective criterion for diagnosis over a broad range of levels and patterns of population differentiation.

These points are commonplace observations and are really just a roundabout way of affirming Darwin's point of 'how entirely vague and arbitrary is the distinction between species and varieties' (Darwin 1859, p. 48). If we are going to give some taxonomic status to varieties, whether they are species or subspecies or even some smaller conservation-based taxonomic rank, then this continuum of differentiation will be a dominating fact of the matter for investigators involved in diagnosis. The general and inherent ambiguity is well summarized by Sites and Marshall (2004):

> there is no objective criterion for how much morphological divergence is enough to delimit a species, what threshold frequency of intermediates is needed to delimit species by genotypic clusters (Mallet 1995), what proportion of unlinked loci are needed to delimit coalescent species (Hudson & Coyne 2002), or what frequency cut-off most appropriately indicates that no significant gene flow is occurring between populations.
>
> (Wiens & Servedio 2000)

The challenges of using hypothesis tests for species diagnosis

It is common to propose that species be diagnosed in a hypothesis-testing framework (Mallet 1995; Sites & Crandall 1997; Wiens & Servedio 2000; Templeton 2001; DeSalle *et al.* 2005). Of the many methods proposed for diagnosing species in recent years, nearly all are cast in either the general language of hypothesis testing or explicitly as a mathematical and statistical method of hypothesis testing (Sites & Marshall 2004). Consider the following conventional hypotheses:

Null model = one single species = no significant differentiation
Alternative model = two (or more) species = a statistical finding of differentiation between populations

Take note that the alternative hypothesis explicitly equates a statistically significant finding of differentiation with the presence of more than one species. Now recall the reality that many species actually do consist of multiple partly differentiated populations. Suppose that these two hypotheses are to be considered in light of typical real world data from two related but partly differentiated populations. It is possible that small samples drawn from multiple populations of the same species might not reveal differentiation. However, as the sample size grows, some level of differentiation between the samples will be found to be greater than what could be expected by chance. In other words, we might expect to reject the null hypothesis for virtually any situation given enough data. In effect, the decision (one species or more than one species) is expected to be a function of the sample size. Figure 2.1 demonstrates this general relationship between sample size and a conclusion of multiple species.

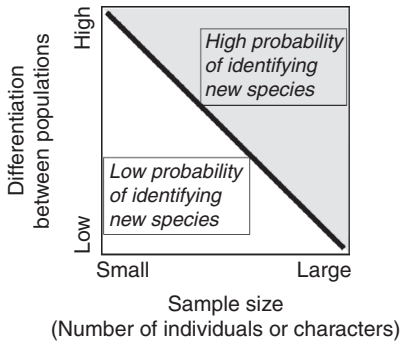

Figure 2.1 The likelihood of species diagnosis is described as a function of both sample size and divergence between populations. Studies with small sample sizes are unlikely to reveal a statistically significant finding, regardless of the degree of divergence, whereas large studies will find divergence even when it is slight. Areas above and to the right of the line will be associated with high rates of species diagnosis, whereas areas below and to the left will have low rates of species diagnosis.

The source of this difficulty lies in the specific choice of the alternative hypothesis which states that additional species are recognized with a statistical finding of *any* differentiation, no matter how slight it is. Since virtually all local populations will be somewhat differentiated from others, this particular hypothesis framework is a recipe for finding as many species as one might have resources for sampling.

Sites and Marshall (2004) reviewed a dozen methods, most of which exhibit the property described here. For example, Good and Wake (1992) proposed a test for assessing whether divergence is due to more than isolation by distance, using a regression of genetic distances among populations against geographic distances. If a significant non-zero intercept is observed, then isolation by distance is rejected. Clearly the more data there is the more likely a finding of a non-zero intercept. In this context 'more data' means more pairs of populations.

Under Population Aggregation Analysis (PAA) (Davis & Nixon 1992), populations are grouped together if they share character states. Successive aggregation of populations is done until only those populations that are separated by fixed character state differences remain. Clearly the more characters or genes that are studied, the greater the chance that one of them will show a fixed difference distinguishing more populations (Wiens 1999; Yoder *et al.* 2000).

Templeton proposed a variation of Nested Clade Analysis (NCA) (Templeton 2001) for identifying species. Like other applications of NCA (Templeton *et al.* 1995; Templeton 1998) the method begins with a statistical test of association of geographic distribution with phylogenetic patterns. Once again, larger sample sizes will mean that a significant association is more likely to be observed.

Methods having a bias in the reverse direction

Another set of methods also exhibit a direct dependency on sample sizes, but exhibit that effect in the opposite direction such that smaller samples are more likely to support the alternative hypothesis and lead to a rejection of the null hypothesis of a single species. These methods employ the same general null and alternative hypotheses as described above; however, they treat the variation that is observed in a sample with what is observed in an entire population. It is

Population 1 individuals	Characters A	B	C	D	E	F	G	H
#1	1	0	1	0	1	1	0	1
#2	1	1	1	0	0	0	1	0
#3	0	1	1	0	0	1	0	1
#4	1	1	1	0	1	0	1	0
#5	1	0	1	0	1	0	0	1
#6	1	0	1	0	1	0	1	1
#7	0	0	1	0	1	0	1	1
#8	1	0	1	0	1	0	1	1
Population 2 individuals								
#1	1	0	1	0	0	0	1	1
#2	1	0	1	0	0	1	1	0
#3	1	0	1	0	0	1	1	0
#4	1	0	1	0	0	1	1	1
#5	1	0	1	0	0	1	0	0
#6	1	1	1	0	0	0	1	1
#7	1	0	1	0	0	1	1	1
#8	1	1	1	0	0	1	0	1

Figure 2.2 An example of the effect of sample size on the discovery of fixed differences between populations. Eight characters are shown for samples from two populations, with no character showing a fixed difference in the full sample. However, smaller samples (grey areas) reveal fixed differences in characters E and F.

possible then to observe a pattern of variation in a sample – such as a fixed difference between populations – that would not be observed in a larger sample. Paradoxically then, such methods are more likely to lead to conclusions of multiple species when sample sizes are small.

The method of PAA (Davis & Nixon 1992) suffers from this problem precisely because it implicitly assumes that character values, which are observed as fixed within samples, are also fixed in species (Wiens 1999; DeSalle & Amato 2004). The problem is demonstrated in Fig. 2.2 which shows a hypothetical data set of eight binary characters for samples of size eight for each of two populations. As constructed the sample from two populations now show fixed differences and would be combined in a PAA analysis. But also highlighted in Fig. 2.2 are smaller samples which do reveal fixed differences (at characters E and F).

This kind of sample size effect will also occur under the method sketched out by Baum and Shaw (1995) to identify genealogically exclusive groups. This approach is based on gene tree estimates obtained for multiple loci, in which multiple sequences have been obtained from each population under consideration and the trees are combined into a strict consensus tree. Any clusters of individuals that are present on the multi-locus strict consensus tree are identified as exclusive groups and meet the species criterion. But clearly the more individuals sampled per population, and the more loci studied, the less likely that a strict consensus tree will reveal exclusive groups.

Estimator bias and consistency

By considering the decision process in a slightly different light we can perceive the dependency of species diagnosis on sample size as a kind of estimator bias. By definition, estimator bias is the difference between the true value of a

parameter and the expected value of an estimate of that parameter. If in the present case the parameter is the true number of species that actually exist (under some criterion), then we would hope that the expected value of an estimate of that value would be close to the true value. But if the expected value changes with the sample size, then clearly there must be a bias, at least over some ranges of sample size.

The property of being biased is not necessarily a large problem for an estimator, particularly if the bias is small. Certainly many estimators in many different contexts have some degree of bias. Also, it is common for estimators who are biased to be *consistent*, meaning that the bias becomes less as the sample size grows, so that with larger sample sizes the expected value of the estimate converges on the true value of the parameter. However, the bias that is described here arises because the estimation procedure is directly equated with hypothesis testing and thus with statistical power. This necessarily creates a strong and direct link between the finding of additional species and the sample size. For those species that consist of populations with complex and varying degrees of differentiation, we might expect that the number of detected species will continue to increase with sample size almost indefinitely as larger and larger samples reveal finer and finer, but still detectable, patterns of differentiation.

Finally, it bears noting that this kind of bias is not one that is associated with any particular species concept, except insofar that some species concepts do not, or have not yet, lent themselves to being cast in terms of a hypothesis-testing framework. Any species diagnostic protocol in which the presence of additional species is equated simply with a p value (i.e. probability of a Type 1 error), or the equivalent thereof, can be expected to suffer this difficulty.

Overcoming sample size effects using cut-off criteria

One way to partly overcome a direct dependence of species diagnosis on sample size is to require that the degree of differentiation between populations (by whatever method is being used) pass some particular threshold, or cut-off, value. Methods like this can also be, and have been, cast in explicit hypothesis-testing frameworks. In such cases the general hypotheses are

Null model = one single species = differentiation does not significantly exceed the previously specified threshold value

Alternative model = two (or more) species = a finding of differentiation between populations that does significantly exceed the threshold value

Cut-off values have been proposed in several different ways. For example, Porter (1990) proposed that an estimated population migration rate (Nm, the product of the effective population size and the migration rate per generation) which is less than 0.5 was likely to be an indicator of populations that are in fact diverging from one another and that may merit some taxonomic status. Similarly Highton

proposed that a genetic distance value greater than 0.15, for Nei's *D* measure of allozyme distance (Nei 1972), was a likely indication of species status, based on the observed distribution of *D* for known species (Highton *et al.* 1989; Highton 1990). Wiens and Servedio (2000) suggested that rather than requiring species to have fixed differences, which is difficult to assess without complete sampling, a cut-off frequency of 0.05 be used as part of a statistical test of species status. Hebert *et al.* (2004) suggested that a difference between DNA sequences of the *Cytochrome Oxidase 1* mitochondrial gene, which is greater than ten times that found within species, would be a useful indication of species status.

However, cut-off values raise a new set of issues. One is that the index or indices of differentiation that are used, whatever they may be, must be well motivated by our understanding of the process of divergence. Such a motivation could possibly come from any of several sources, including evolutionary models of divergence, population genetic theory or particular species concepts. Another general concern that arises for any particular method based on cut-off values is that it may not be suitable for different taxonomic groups that necessarily vary in the ways that populations tend to be structured and in the ways that speciation occur (DeSalle *et al.* 2005). It would be better if the cut-off methods developed and used actually make sense for a wide range of taxonomic groups.

Another possible concern is that cut-off values must be partly arbitrary. If a protocol is designed with a divergence criterion x, then it follows that more new species are going to be identified than would be the case if $2x$ were the criterion. Given the inherent continuous nature of divergence among populations within species, this is unavoidable. However this component of arbitrariness may be difficult to accept in contexts where species are thought of as inherently fundamental and unitary.

The insufficiency of overall summaries of differentiation

In the continuum of degrees of differentiation among populations, divergence is a complex process that is not likely to be well encapsulated by any single index or parameter. Consider the issues that arise in the use of F_{ST}, Wright's (1951) fixation index. F_{ST} is readily estimated for most kinds of genetic variation and it is a ubiquitous feature of studies on population structure. Under a model in which population structure is at an equilibrium of genetic drift and gene flow, F_{ST} can be used to estimate a population gene flow level. However, F_{ST} is calculated just as easily for any pair of populations, regardless of the role of gene flow in their divergence and regardless of whether divergence is at equilibrium (Whitlock & McCauley 1999).

It is useful to consider F_{ST} with respect to two complementary population genetic models. If two populations are exchanging genes at some level for a long period of time, then they will approach an equilibrium level of F_{ST} and a corresponding equilibrium level of divergence. Under such situations, F_{ST} can

be used to estimate the population migration rate Nm (Wright 1951; Slatkin & Voelm 1991). But now consider a radically different model in which two populations separate out of a single population and exchange zero migrants thereafter. At any point in time after the split there will be some level of divergence, and F_{ST} can be calculated and will reflect that divergence. Similarly it is possible to estimate the time of splitting using F_{ST}, assuming zero gene flow (Takahata & Nei 1985). From these two examples we see that F_{ST} can be used to estimate entirely different quantities, migration rate and splitting time, under two radically different models – one of which predicts that divergence will increase and the other which has divergence at an unchanging equilibrium. Clearly F_{ST} by itself cannot be a suitable measure of differentiation upon which to base a cut-off value for species diagnosis, nor is the problem limited to F_{ST}. Because of the very different nature of population splitting time and gene flow, it is likely that no single measure of divergence can capture both of these key aspects of the divergence process.

Considering model-based approaches

Notwithstanding the complexity of the divergence process, it is possible to capture many of the dynamics of divergence in a quantitative model. Figure 2.3 shows the Isolation with Migration model (Wakeley & Hey 1998; Nielsen & Wakeley 2001; Hey & Machado 2003), which is intended to represent the divergence history of a pair of sister populations or species. The model includes six parameters, including: the time when the populations separated from one another; the effective population sizes N, for the two populations; the effective population size for the ancestral population (before the splitting time); and two unidirectional migration rates. Together these parameters capture many of the demographic components of the divergence process. For example, divergence will be reduced if genetic drift is slow due to large effective population sizes; if the time of population splitting was recent; or if migration rates are high.

Clearly, the Isolation with Migration model cannot capture many things. In particular it does not contain any parameters that correspond to population-specific adaptation. However, the model can indirectly inform on adaptation in

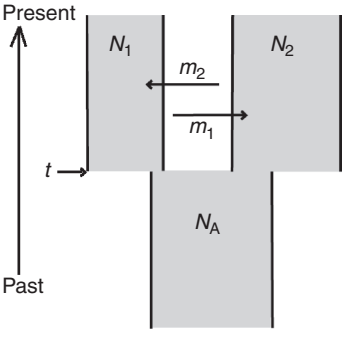

Figure 2.3 The Isolation with Migration model.

some circumstances. For example, if two populations show evidence of gene flow, then this elevates the relevance of any additional information on differential adaptation. Since only a small amount of gene flow is needed to retard divergence, the presence of phenotypic differentiation *together* with evidence of gene flow can indicate that the phenotypic differences are indeed under divergent selection (Machado *et al.* 2002; Bull *et al.* 2006).

Suggestions for criteria for species diagnosis

Any discussion of numerical cut-off values for taxon diagnosis requires a motivating argument that justifies some particular index of differentiation and some particular cut-off value. We can use as a motivating claim the idea that putative new taxa should be evolutionarily independent to a sufficient degree for them to be expected to continue to diverge. This is a claim similar to that for 'independent evolutionary trajectories', which was proposed for a well-used (and much discussed) concept of an 'evolutionary significant unit' or ESU (Waples 1991b), but one which is also consistent with or implicit in many species concepts (Mayden 1997).

Given the roles that population size, population splitting time and migration all play in the divergence process, it seems unlikely that a single numerical criterion could be used as an indication of whether or not divergence is likely to continue. However, some basic population genetic theory does tell us that we can expect gene flow to be low for diverging populations, because a population migration rate of Nm greater than 1 is sufficient to strongly limit divergence (Wright 1931). Therefore, one choice of a cut-off value would be that Nm be less than 1. However, estimating gene flow also requires that confounding factors, such as the time of population splitting, be accounted for (Whitlock & McCauley 1999).

An estimate of the time since two populations split can also be a useful indicator of their evolutionary independence, as the greater the time the more opportunity there has been for fixation of alleles, including selected alleles, in individual populations. Also populations that split many generations ago have necessarily stayed at least partly separate throughout whatever other demographic vagaries have occurred in the history of the populations since that time. The trouble with splitting time, as an indicator of divergence, is that it cannot be used directly if it is expressed in units of years. The genetics of the divergence process plays out on a timescale of generations, so that it would make sense for a splitting time cut-off to be expressed in units of generations. A related point is that divergence time, even when cast in terms of generations, will matter more for small populations than for large populations. This is because small populations experience rapid genetic drift and will fix alleles more rapidly than will large populations. Thus it would be useful to express a splitting time criterion on a scale of generations, and in a way that reflects effective population size. In

fact this kind of timescale is regularly used in population genetic models of divergence, where the time parameter is literally given on a scale of 'effective population size generations'. If we let τ be the time since separation, scaled by effective population size and generation, then a scaled time of 1 (i.e. $\tau = 1$) would mean that N generations had passed. For example in a *Drosophila* species with 10 generations per year and an effective size of 1 million, a value of $\tau = 1$ corresponds to $0.1 \times 1\,000\,000 = 100\,000$ years.

An example using specific migration and splitting time cut-offs

From the arguments given above we can consider a species diagnostic protocol with two criteria:

1. The time since splitting, in units of N_e generations, τ, should be significantly greater than 1 for each of the populations.
2. The population migration rate in each direction with the most closely related population should be significantly less than 1.

Together these criteria could be the basis of a statistical approach to species diagnosis. Both criteria 1 and 2 are required because it is not sufficient to simply reject a null hypothesis of a single species on the basis of either divergence time or a low level of gene flow.

Statistical cut-off criteria like these will still necessarily retain a dependency upon sample size. In general small samples are not expected to meet either criterion 1 or 2, regardless of the true values of τ and Nm. However, unlike methods that equate *any* significant non-zero differentiation with species diagnosis, the bias that arises with the use of joint cut-off values should diminish as sample size grows. This diminishing effect of sample size means that the protocol should thereby be statistically consistent. This is because as the sample size grows the estimates of τ and Nm should become more and more precise, so that it will become clearer whether or not criteria 1 and 2 are met.

For example, consider the case of two subspecies of the common chimpanzee *Pan troglodytes troglodytes* (the central African chimpanzee) and *Pan troglodytes verus* (the western African chimpanzee), the divergence of which were studied using a data set of 48 loci (Won & Hey 2005). The estimated effective sizes were 27 900 for *P. t. troglodytes*, 7 600 for *P. t. verus* and 5 300 for their ancestral population; their splitting time was estimated to be 422 000 years; and a clear signal of gene flow was observed from *P. t. verus* into *P. t. troglodytes*, but not in the reverse direction. Figure 2.4 shows posterior probability density estimates for τ and Nm in both directions. By calculating the area under the curves with respect to the two criteria listed above, we find that the probability that $Nm \geq 1$ for gene flow into *P. t. troglodytes* is 0.004, while the probability that $Nm \geq 1$ for gene flow into *P. t. verus* is 0.0. Clearly the gene flow criterion is met in this case. Turning to τ, we find that the density curve for *P. t. verus* is far to the right and clearly exceeds 1. However, in

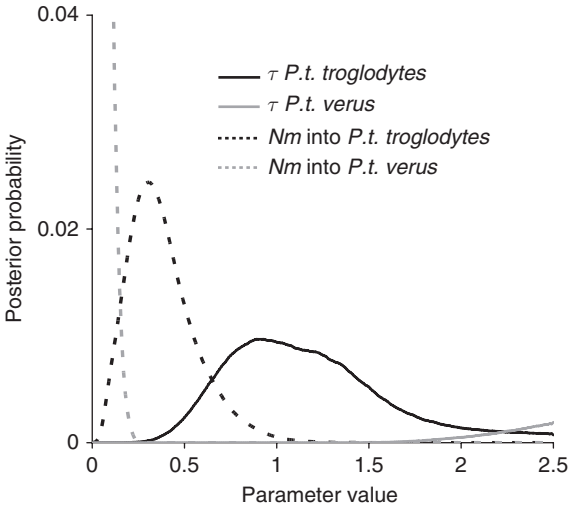

Figure 2.4 Posterior probability density estimates of τ and Nm. The IM program (Hey & Nielsen 2004) was run using the same data and protocol as in Won and Hey (2005). In that program, the parameters were as follows: tu, the product of number of generations since splitting and the mutation rate; $4Nu$, the product of four times the effective population size and the mutation rate; and m/u, the ratio of the migration rate per generation and the mutation rate. Since $\tau = 4 \times tu/4\,Nu$ and $Nm = 4\,Nu \times m/u/4$, it was possible to record the necessary quantities over the course of the analysis. For Nm into *P. t. verus* the peak of the curve is at 0, with a probability value that is above the axis limit shown. In the case of τ for *P. t. verus* the majority of the probability density was far to the right of the portion of parameter value axis that is shown.

the case of τ for *P. t. troglodytes* we find that the probability that $\tau \leq 1$ is 0.3878. In this case, we cannot reject the null hypothesis of a single species on the basis of criterion 1. Figure 2.4 also shows clearly how the decision-making process can depend strongly upon the selected cut-off values. If instead we use a value of 0.5 for criterion 1, then we find that for *P. t. troglodytes* the probability that $\tau \leq 0.5$ is only 0.0168, which would lead to the rejection of the null hypothesis of a single species and a conclusion that the two subspecies be elevated to species status.

Conclusion

If the diagnosis of new species is to be based on hypothesis testing then it will be difficult if not impossible to completely remove the statistical bias, such that large samples reveal more new species than do small samples. However, if non-zero cut-off values for well-motivated indicators of divergence are used, then tests using large samples should converge on a robust result of either acceptance or rejection of the null hypothesis.

The indicators of divergence that are presented here, τ and Nm, were selected based on a population genetic understanding of the divergence process. However, they are not the only possible indicators, and some points should be emphasized when indicators are considered:

- No single measure of divergence can be expected to capture all of the key dynamics of the divergence process, including: separation time, gene flow and adaptation.
- Even a large set of divergence measures or estimated parameters is unlikely to fully capture the divergence history in any particular case. Whatever method is used, there will necessarily be a trade-off between the accessibility and simplicity of the diagnostic method and the completeness of the divergence estimates.
- Divergence measures should be selected to be applicable to as wide an array of kinds of organisms as possible. It would be nice if a finding of a new species within one group of organisms conveyed a similar degree of distinction as a finding of a new species in a distantly related group of organisms.

Finally, the use of specific cut-off values does highlight an apparent arbitrariness to species diagnosis. This is simply because the cut-off values must be selected by human investigators and because shifting them will lead to different rates of species diagnosis. This element of arbitrariness is inherent to the diagnosis process whenever closely related populations are being considered, and no quantitative protocol can remove it. However, this does not mean that divergence is not real and objective, and it does not mean that species that are identified do not reflect real divergence and are not, in a similar sense, real and objective.

Acknowledgements

Thanks to Jack Sites, John Bridle, Jim Mallet and Roger Butlin for comments on the manuscript.

References

Baum, D. A. and Shaw, K. L. (1995) Genealogical perspectives on the species problem. In: *Experimental and Molecular Approaches to Plant Biosystematics* (ed. P. C. Hock and A. G. Stevenson), pp. 289–303. Missouri Botanical Garden, St. Louis.

Bull, V., Beltran, M., Jiggins, C. D., *et al.* (2006) Polyphyly and gene flow between non-sibling Heliconius species. *BMC Biology* **4**, 11.

Cronquist, A. (1978) Once again, what is a species? In: *Biosystematics in Agriculture* (ed. J. A. Romberger), pp. 3–20. Allanheld & Osmun, Montclair, NJ.

Darwin, C. (1859) *On the Origin of Species by Means of Natural Selection*. Murray, London.

Davis, J. I. and Nixon, K. C. (1992) Populations, genetic variation, and the delimitation of phylogenetic species. *Systematic Biology* **41**, 421–435.

DeSalle, R. and Amato, G. (2004) The expansion of conservation genetics. *Nature Reviews Genetics* **5**, 702–712.

DeSalle, R., Egan, M. 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.

Dimmick, W. W., Ghedotti, M. J., Grose, M. J., *et al*. (1999) The importance of systematic biology in defining units of conservation units. *Conservation Biology* **13**, 653–660.

Goldstein, P. Z., Desalle, R., Amato, G. and Vogler, A. P. (2000) Conservation genetics at the species boundary. *Conservation Biology* **14**, 120–131.

Good, D .A. and Wake, D. B. (1992) Geographic variation and speciation in the torrent salamanders of the genus Rhyacotriton (Caudata: Rhyacotritonidae). *University of California Publications in Zoology* **126**, 1–91.

Hebert, P. D. N., Stoeckle, M. Y., Zemlak, T. S. and Francis, C. M. (2004) Identification of birds through DNA barcodes. *PLoS Biology* **2**, e312.

Hey, J. and Machado, C. A. (2003) The study of structured populations – new hope for a difficult and divided science. *Nature Reviews Genetics* **4**, 535–543.

Hey, J. and Nielsen, R. (2004) Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics* **167**, 747–760.

Highton, R. (1990) Taxonomic treatment of genetically differentiated populations. *Herpetologica* **46**, 114–121.

Highton, R., Maha, G. C. and Maxson L. R. (1989) Biochemical evolution in the slimy salamanders of the *Plethodon glutinosus* complex in the eastern United States. *Illinois Biological Monographs* **57**, 1–78.

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

Hughes, J. B., Daily, G. C. and Ehrlich, P. R. (1997) Population diversity: its extent and extinction. *Science* **278**, 689–692.

Machado, C. A., Kliman, R. M., Markert, J. M. and Hey, J. (2002) Inferring the history of speciation from multilocus DNA sequence data: the case of *Drosophila pseudoobscura* and its close relatives. *Molecular Biology and Evolution* **19**, 472–488.

Mallet, J. (1995) A species definition for the modern synthesis. *Trends in Ecology and Evolution* **10**, 294–299.

Mayden, R. L. (1997) A hierarchy of species concepts: the denouement in the saga of the species problem. In: *Species: The Units of Biodiversity* (ed. M. F. Claridge, H. A. Dawah and M. R. Wilson), pp. 381–424. Chapman and Hall, London.

Mayr, E. (1982) *The Growth of Biological Thought*. Harvard University Press, Cambridge, MA.

Moritz, C. (1994) Defining 'evolutionary significant units' for conservation. *Trends in Ecology and Evolution* **9**, 373–375.

Nei, M. (1972) Genetic distance between populations. *American Naturalist* **106**, 283–292.

Nielsen, R. and Wakeley, J. (2001) Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics* **158**, 885–896.

Pennock, D. S. and Dimmick, W. W. (1997) Critique of the evolutionary significant unit as a defnition for 'distinct population segments' under the U.S. Endangered Species Act. *Conservation Biology* **11**, 611–619.

Porter, A. H. (1990) Testing nominal species boundaries using gene flow statistics: the taxonomy of two hybridizing admiral butterflies (Limenitis: Nymphalidae). *Systematic Zoology* **39**, 131–147.

Sites, J. W. and Crandall, K. A. (1997) Testing species boundaries in biodiversity studies. *Conservation Biology* **11**, 1289–1297.

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

Slatkin, M. and Voelm, L. (1991) F_{ST} in a hierarchical island model. *Genetics* **127**, 627–629.

Takahata, N. and Nei, M. (1985) Gene genealogy and variance of interpopulational nucleotide differences. *Genetics* **110**, 325–344.

Templeton, A. R. (1998) Nested clade analyses of phylogeographic data: testing hypotheses about gene flow and population history. *Molecular Ecology* **7**, 381–397.

Templeton, A. R. (2001) Using phylogeographic analyses of gene trees to test species status and processes. *Molecular Ecology* **10**, 779–791.

Templeton, A. R., Routman, E. and Phillips, C. A. (1995) Separating population structure from population history: a cladistic analysis of the geographical distribution of mitochondrial DNA haplotypes in the tiger salamander, Ambystoma tigrinum. *Genetics* **140**, 767–782.

Vogler, A. P. and DeSalle, R. (1994) Diagnosing units of conservation management. *Conservation Biology* **8**, 354–363.

Wakeley, J. and Hey, J. (1998) Testing speciation models with DNA sequence data. In: *Molecular Approaches to Ecology and Evolution* (ed. R. DeSalle and B. Schierwater), pp. 157–175. Birkhäuser Verlag, Basel.

Waples, R. S. (1991a) *Definition of 'Species' Under the Endangered Species Act: Application to Pacific Salmon*, p. 29. National Marine Fisheries Service, Seattle, WA.

Waples, R. S. (1991b) Pacific salmon, *Oncorhynchus* spp., and the definition of 'species' under the Endangered Species Act. *Marine Fisheries Review* **53**, 11–22.

Waples, R. S. (1995) Evolutionarily significant units and the conservation of biological diversity under the Endangered Species Act. In: *Evolution and the Aquatic Ecosystem: Defining Unique Units in Population Conservation* (ed. J. L. Nielsen), pp. 8–27. American Fisheries Society, Bethesda, MD.

Whitlock, M. C. and McCauley, D. E. (1999) Indirect measures of gene flow and migration: FST not equal to 1/(4 Nm + 1). *Heredity* **82** (Pt 2), 117–125.

Wiens, J. J. (1999) Polymorphism in systematics and comparative biology. *Annual Review of Ecology and Systematics* **30**, 327–362.

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.

Winston, J. (1999) *Describing Species*. Columbia University Press, New York, NY.

Won, Y. J. and Hey, J. (2005) Divergence population genetics of chimpanzees. *Molecular Biology and Evolution* **22**, 297–307.

Wright, S. (1931) Evolution in Mendelian populations. *Genetics* **16**, 97–159.

Wright, S. (1951) The genetical structure of populations. *Annals of Eugenics* **15**, 323–354.

Yoder, A. D., Irwin, J. A., Goodman, S. M. and Rakotoarisoa, S. V. (2000) Genetic tests of the taxonomic status of the ring-tailed lemur (Lemur catta) from the high mountain zone of the Andringitra Massif, Madagascar. *Journal of Zoology* **252**, 1–9.