THE ROYAL SOCIETY

Species delimitation in systematics: inferring diagnostic differences between species

John J. Wiens[1]* **and Maria R. Servedio**[2]

[1]*Section of Amphibians and Reptiles, Carnegie Museum of Natural History, Pittsburgh, PA 15213-4080, USA*
[2]*Center for Population Biology, University of California, Davis, CA 95616, USA*

Species are fundamental units in studies of systematics, biodiversity and ecology, but their delimitation has been relatively neglected methodologically. Species are typically circumscribed based on the presence of fixed (intraspecifically invariant or non-overlapping) diagnostic morphological characters which distinguish them from other species. In this paper, we argue that determining whether diagnostic characters are truly fixed with certainty is generally impossible with finite sample sizes and we show that sample sizes of hundreds or thousands of individuals may be necessary to have a reasonable probability of detecting polymorphisms in diagnostic characters at frequencies approaching zero. Instead, we suggest that using a non-zero frequency cut-off may be a more realistic and practical criterion for character-based species delimitation (for example, allowing polymorphisms in the diagnostic characters at frequencies of 5% or less). Given this argument, we then present a simple statistical method to evaluate whether at least one of a set of apparently diagnostic characters is below the frequency cut-off. This method allows testing of the strength of the evidence for species distinctness and is readily applicable to empirical studies.

Keywords: species; species delimitation; systematics; statistics; sample sizes

1. INTRODUCTION

Species are fundamental units of systematic, ecological and evolutionary studies and the accurate documentation and delimitation of species is increasingly important as the species diversity of the world's biota is increasingly reduced and threatened. The discovery and description of species is a major endeavour of the field of systematics. However, in stark contrast to phylogeny reconstruction, the other major endeavour of systematics, there has been relatively little progress in the statistical methodology of species delimitation, at least as practiced by most alpha taxonomists. Although there has been considerable interest recently in the use of haplotype phylogenies from DNA sequence data to infer species boundaries (e.g. Avise & Ball 1990; Baum & Donoghue 1995; Graybeal 1995; Olmstead 1995; Templeton 1998; Brower 1999), most species continue to be circumscribed based on morphological comparisons of museum specimens. For example, published descriptions of new plant and animal species almost always include a museum specimen designated as a holotype and a list of diagnostic morphological features.

The basic procedure of most morphological alpha-level systematic studies is to compare character distributions between geographical samples and determine which sets of populations are delimited by seemingly fixed diagnostic differences (that is, differences which are inferred to be invariant within the putative species or are at least non-overlapping). These fixed differences may indicate an absence of gene flow between putative taxa and the presence of two or more distinct species. Advocates of the phylogenetic species concept have claimed that fixed diagnostic differences are a necessary criterion for species delimitation (e.g. Eldredge & Cracraft 1980; Nixon & Wheeler 1990; Davis & Nixon 1992) and proponents of the evolutionary species concept have claimed that this concept is operationally equivalent to the phylogenetic species concept in terms of the evidence required (fixed differences) (Frost & Kluge 1994). It seems that the common criterion for species recognition in most of the empirical systematics literature over the past 100 years is the presence of one or more apparently fixed or non-overlapping differences between putative species (regardless of the underlying species concept), even if this criterion is rarely made explicit or theoretically justified (Nixon & Wheeler 1990).

The traditional approach of delimiting species based on one or more apparently fixed differences was codified by Davis & Nixon (1992) in a methodology called population aggregation analysis (PAA). PAA involves systematically comparing character state distributions among populations, aggregating sets of populations which differ only in polymorphic traits and considering sets of populations which differ by at least one seemingly fixed difference (or which share no states for that character) to be different species. However, as pointed out by Davis & Nixon (1992), PAA is problematic in that (i) unless many characters are sampled, the number of species present may be underestimated because characters with fixed states may not be observed and (ii) unless many individuals are sampled, the number of species may be overestimated by considering traits which are actually polymorphic to be fixed. PAA has never been modified to account for these problems or to detect when the data are inadequate for making a clear decision. However, these are not problems unique to PAA. In fact, most empirical systematic studies and more conceptual discussions of species delimitation (e.g. Nixon & Wheeler 1990; Davis & Nixon 1992; Frost & Kluge 1994) either do not consider the issue of sampling error or else do not address it in a rigorous fashion.

*Author for correspondence (wiensj@clpgh.org).

Received 6 October 1999 *Accepted* 25 November 1999

In this paper, we present a statistical method of assessing confidence in species-level decisions. We first argue that determining whether traits are truly fixed with certainty requires sampling every single individual in the species and we show that having a reasonable probability of detecting polymorphisms at frequencies approaching zero may require sampling hundreds or thousands of individuals per species. We then suggest that it may be reasonable to infer that species are distinct even if there is a possibility of polymorphisms in the diagnostic characters occurring at low frequencies. Finally, we present a hypothesis test to make such an inference statistically using a frequency cut-off.

2. SPECIES CONCEPTS

Before we discuss species delimitation, we briefly digress to define what we mean by 'species'. In a recent review, de Queiroz (1998) suggested that the plethora of proposed species concepts (e.g. biological, phylogenetic, evolutionary and cohesion) agree fundamentally on what species are: for sexual organisms, a species is a lineage which is unified primarily by sexual reproduction or gene flow among its constituent parts. We follow this general lineage concept of species (de Queiroz 1998). Furthermore, we consider species to be real entities which exist regardless of whether there is sufficient evidence to recognize them (Frost & Kluge 1994). Thus, we distinguish between a species concept (an idea of what kind of entity species are) and a species criterion (a methodological approach to recognizing species in a particular case) (de Queiroz 1998). In this paper, we use the presence of one or more diagnostic characters which distinguish a given species from all others as a species criterion (following common practice in empirical studies); by 'diagnostic' we mean characters which have the alternate state below a given frequency cut-off, including fixation (see §4). However, we acknowledge that in certain cases some real species may fail to pass this test (i.e. all operational species criteria will fail in some cases) (Frost & Kluge 1994).

3. DETECTING POLYMORPHISM IN FIXED DIAGNOSTIC CHARACTERS

Character fixation is the common criterion for species delimitation and we are interested in determining the confidence levels associated with this criterion. However, claiming that a trait is truly fixed (frequency = 100%) assumes that there is not a single individual in the species which possesses the alternate trait. To be certain that this is the case would require sampling every single individual in the species, which is clearly an impossibility in most empirical studies. However, given a large enough sample size, a polymorphism would have to occur at a very low frequency to avoid being detected. We will explore the sample sizes needed to detect polymorphism confidently in a putatively diagnostic character at a given non-zero frequency.

Swofford & Berlocher (1987) discussed the problem of distinguishing polymorphic and fixed traits with a finite sample size in the context of phylogenetic analysis of polymorphic data. To find the probability that a rare trait will go undetected within a sample of n individuals when it occurs at frequency p they provided the equation from the binomial distribution

$$\text{probability(rare trait undetected)} = (1 - p)^n. \quad (1)$$

We have eliminated a factor of 2 from the exponent of the original Swofford & Berlocher (1987) equation (for alleles in diploids) because we assume that we are looking at morphological characters in which heterozygotes are unlikely to be detectable as such.

The approach of Swofford & Berlocher (1987) can be used to develop estimates of the sample sizes (number of individuals) needed to have a reasonable probability that a putatively fixed diagnostic character used in an alpha-level systematic comparison is not actually polymorphic (i.e. has the trait of the other species at a given frequency). All tests in this paper are designed to be applied to one species at a time and the evidence for a diagnostic difference between two species must be evaluated in each species separately.

Focusing on one of the species in a pairwise comparison, we would like to know how many individuals must be sampled for an apparently fixed character to reduce the probability that we are failing to detect the state from the other species to 5% (following the standard acceptable error rate in statistics), under the assumption that this rare trait is present at frequency p (we assume that all the individuals sampled for this character are invariant). Setting equation (1) equal to 0.05 and solving for n, we obtain

$$n = \frac{\log(0.05)}{\log(1 - p)}. \quad (2)$$

Figure 1 shows the sample size (n) needed from each putative species to have only a 5% probability of failing to detect a polymorphism occurring at various frequencies (p) in an apparently fixed character. The more individuals sampled, the more likely it is that polymorphisms have not gone undetected.

An important conclusion from figure 1 is that, in order to have only a 5% probability that a rare trait at a relatively low frequency (i.e. 0.01) has been missed in a seemingly fixed character, the sampling effort required is probably unattainable for most empirical studies (e.g. several hundred specimens). Lower frequencies of the rare trait would require even greater sampling. Given these results and our argument that determining fixation with certainty requires sampling every single individual of a species, claiming that differences between species are truly fixed or even very close to fixed represents little more than a guess, at least from a statistical perspective. This point is largely absent from empirical and theoretical studies of species delimitation (e.g. Nixon & Wheeler 1990; Davis & Nixon 1992; Davis 1996). While some may argue that the problem of undetected polymorphism will only be an issue if traits often occur at very low frequencies, the work of Wright (1937) suggested that polymorphisms in neutral traits are more likely to occur at very low frequencies than at higher frequencies in natural populations.

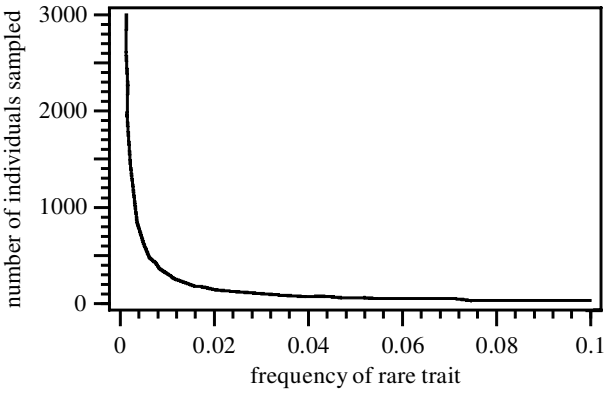

Figure 1. The data (number of individuals sampled) which must be obtained to have only a 5% probability of failing to detect a rare trait when it occurs within a putative species at various frequencies.

4. EVALUATING A FREQUENCY CUT-OFF

Rather than clinging to the almost impossible ideal of distinguishing fixation and polymorphism, a more reasonable approach may be to allow some level of polymorphism in the diagnostic characters (e.g. assume that a trait present at a frequency of 95% or higher in one population and 5% or lower in the other population is 'close enough' to being a fixed difference in that it indicates very low levels of gene exchange between the putative species). Some previous authors have suggested the use of a non-fixed frequency cut-off in making species decisions (e.g. McKitrick & Zink 1988).

Requiring fixed differences between two putative species is probably not necessary to demonstrate that there may be negligible amounts of gene flow between them. Because gene flow can homogenize trait frequencies over a short time-period, even relatively small differences in trait frequencies might be strong evidence of reduced or absent gene flow. To give a simple numerical example, say we have two populations (putative species), each of which has a neutral diagnostic character state from the other population present at a frequency of 5%. If these populations were to start exchanging only one migrant per 100 individuals per generation, the frequencies of the foreign trait would reach as high as 20% in each population in roughly 20 generations (as determined through iterations of the appropriate recursion equations).

We argue that if there are large differences in trait frequencies between two putative species, this may be good evidence that there is little or no gene flow between them. A reasonable cut-off for trait frequencies to indicate low or absent gene flow may depend upon several factors, including effective population sizes, population substructuring and whether the traits considered are under selection. Porter (1990) suggested incorporating these parameters in species delimitation using statistical measures of gene flow (e.g. Wright 1931, 1978; Slatkin & Barton 1989). Although this approach seems promising, the necessary parameters may be difficult to estimate, particularly for morphological characters. Given that the genetic data necessary to quantify gene flow more precisely are generally unavailable and may be impractical to collect for most studies of species delimitation, we believe that the frequency cut-off approach described here provides a practical proxy for assessing the amount of genetic exchange.

Once a frequency cut-off for rare traits is decided upon, a modified version of a binomial test (presented below) can be used to evaluate whether there is sufficient evidence to conclude that, in a sample of apparently fixed characters, at least one of these characters has the foreign trait at a frequency below the predetermined cut-off. Following the convention that there need be only one fixed difference between a pair of putative species to indicate that gene flow is absent, we assume that having at least one character in which the foreign trait is absent or below the frequency cut-off is an indication of negligible gene flow. The test may be modified by researchers who wish to alter this convention (e.g. requiring two diagnostic characters instead of one).

We would like to be able to reject the null hypothesis that any rare states in the diagnostic characters are actually present at a frequency greater than p (the frequency cut-off). The data necessary to make this determination are the number of individuals sampled (n), the total number of characters surveyed for potential diagnostic differences (c) and the number of those characters which were found to be fixed for the diagnostic character state (k), at least among the individuals sampled. From equation (1), we know that the probability of finding only the common trait for a specific character (we will call this probability F) if the rare trait is actually present at a frequency greater than or equal to p is at most $(1-p)^n$. This is, in effect, the probability of finding that a given character lies within the set of diagnostic characters k. Applying the binomial distribution, the probability of obtaining at least k such characters out of c total characters, assuming the frequency of each rare trait is greater than or equal to p, is at most

$$P = \sum_{r=k}^{c} (\binom{c}{r})(F^r)(1-F)^{(c-r)}). \quad (3)$$

These P-values are calculated for selected ranges of p, n, c and k in figure 2 and electronic Appendices A–C (available at The Royal Society Web site). If the value of P obtained for a given set of data is less than the alpha level, the null hypothesis can be rejected and the researcher can conclude that at least one of the putative diagnostic characters is fixed or has the rare trait at a frequency less than the cut-off p. For example, assume a researcher has sampled 20 individuals of a given species, has found five apparently fixed diagnostic characters (k) out of ten characters sampled (c) and is willing to accept a frequency cut-off of 10% ($p = 0.10$) using a 5% confidence interval ($a = 0.05$). Looking at figure 2, the intersection of $P = 0.05$ and $k = 5$ falls between lines b ($n = 20$) and c ($n = 10$), showing that 20 individuals are sufficient evidence to reject the null hypothesis that the actual frequency of all of the apparently absent traits is above 10% (and in electronic Appendix A, the P-value lies below the predetermined alpha for $n = 20$ and $c = 10$ with only four diagnostic characters). The researcher can therefore accept the hypothesis that in at least one of the apparently fixed characters the alternate (foreign) state is below a frequency of 10%.

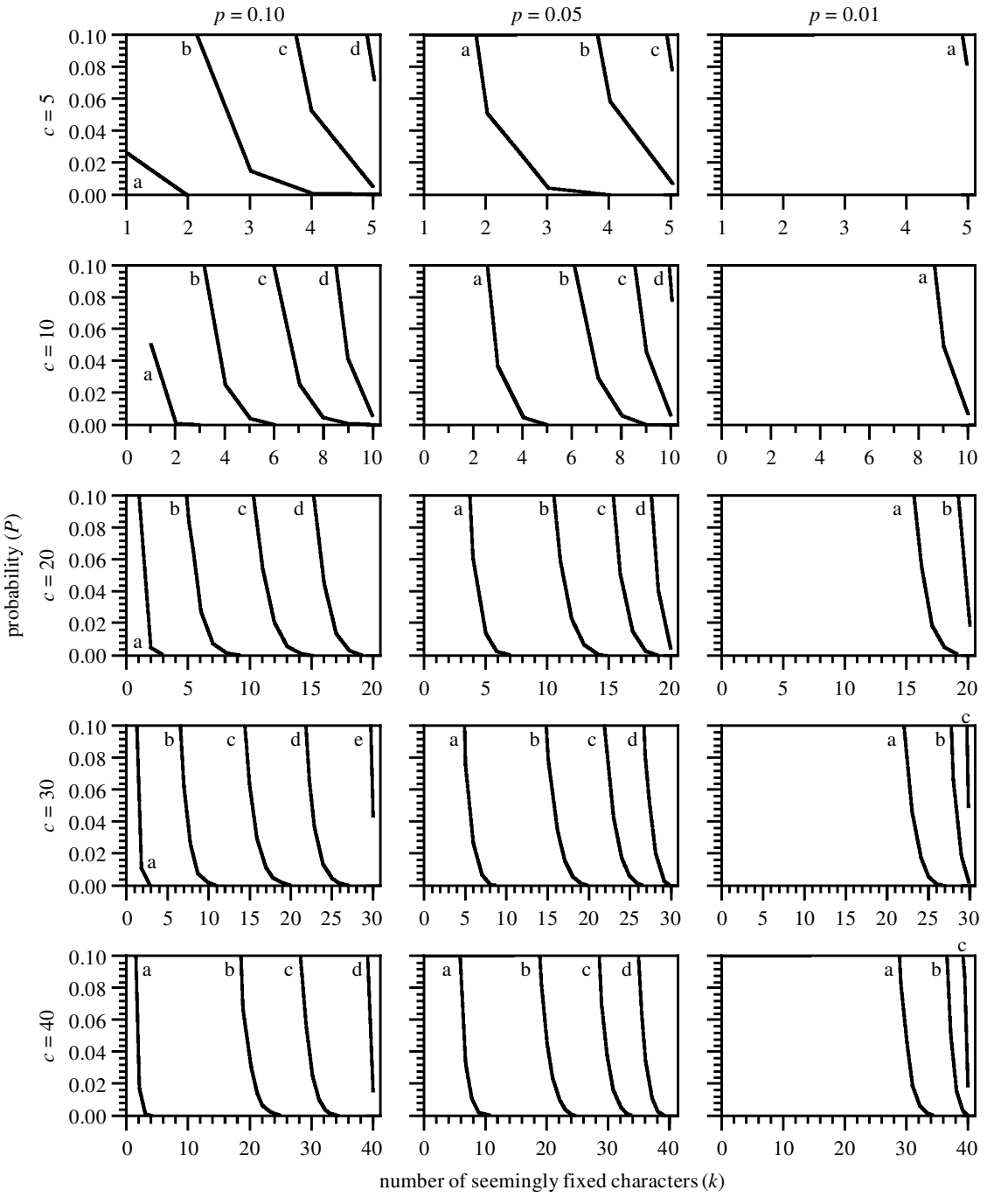

Figure 2. Significance levels for a binomial test evaluating whether there is sufficient data (characters and individuals sampled) to conclude that at least one of the apparently fixed traits has a frequency below the predetermined cut-off p (i.e. is diagnostic). Different lines represent different numbers of individuals sampled for the putative species (a, 50 individuals; b, 20 individuals; c, ten individuals; d, five individuals; and e, one individual). For example, line a in the upper left hand corner box ($c = 5$ and $p = 0.10$) indicates the probability of failing to detect polymorphism (at a frequency of 10% or higher) in all of the seemingly invariant characters, given that 50 individuals were sampled for all five characters. c is the total number of characters sampled, including polymorphic and seemingly fixed characters.

There is obviously a trade-off between the number of individuals sampled per species (n) and the number of characters which appear to be fixed for a diagnostically different character state (k). If many seemingly fixed characters are sampled (assuming a constant sample size) we can be more confident that at least one of these characters does not have the foreign trait above a given frequency. In essence, the more seemingly fixed characters, the greater the chances that two species may be demonstrably distinct. Likewise, the more individuals sampled, the more likely it is that polymorphisms have not gone undetected.

An important caveat which should be made about this test is that it does not say anything about which of the apparently fixed characters actually has the foreign trait below the cut-off. Therefore, if two species are compared

for a number of apparently fixed character differences, the fact that both species pass the test does not mean that the same character is below the cut-off in each (unless, of course, there is only one diagnostic character). A possible way around the problem might be to first determine with confidence what set of specific characters had rare states below the frequency cut-off in the more well sampled of the two species (for a specific character, a simple binomial test based on equation (1) can be used; corresponding tables can be found in standard books on non-parametric statistics, e.g. Conover (1971)). Once this set of characters has been determined, they can be used as the total set of characters (c) in the test proposed above to determine whether any one of them were below the frequency cut-off in the more poorly sampled species (i.e. one would be looking to see whether the common state in the well-sampled species was below the frequency cut-off in the poorly sampled species).

5. DISCUSSION

In this paper, we have developed a statistical method which can be applied to traditional character-based species delimitation. This method allows systematists to estimate whether or not they have sampled sufficient characters and individuals to infer that a species is delimited by one or more diagnostic traits statistically (i.e. traits which are fixed or variable below a given frequency cut-off). Although (seemingly) fixed character differences have been the basic evidence used in most species-level decisions by empirical systematists, we argue that, strictly speaking, character fixation can almost never be claimed with certainty. Furthermore, having a high probability of detecting polymorphisms at frequencies approaching zero requires sampling hundreds or thousands of individuals per species. These observations render the 'fixed difference' criterion largely meaningless. We argue that accepting some non-zero frequency cut-off (e.g. 5 or 10%) may be more realistic and practical than assuming or requiring fixed differences between putative species. Although we do not know what exactly the preferred frequency cut-off should be, we provide a methodology allowing researchers to evaluate whether sufficient data are present to statistically support whichever cut-off is chosen. One could also use our approach to determine the smallest frequency cut-off that would be statistically supported by the data at hand.

Our approach is intrinsically probabilistic and, therefore, assumes a model. We have used a very simple model for the distribution of trait frequencies within a species (binomial distribution), which assumes no population subdivision. Rannala (1995) recently addressed the effects of population subdivision and non-random mating on errors in estimating allele frequencies. Based on his work, we note that the presence of population subdivision within putative species would probably require sampling more individuals and characters in order to detect rare polymorphisms in putatively fixed traits. If population subdivision is considered likely, the estimates obtained from our approach should be treated as minimum estimates of the number of individuals and characters that should be sampled. Furthermore, we assume that each of the seemingly fixed characters has an equal evidential value or weight for species delimitation. Equal weighting is a common assumption in systematics, but there may be some characters which are far more important (and less likely to be intraspecifically variable) than others in species delimitation, such as characters involved in reproductive isolation (e.g. genital morphology) (Eberhard 1985) and complex characters controlled by many different genes. We also assume that all characters are independent and uncorrelated. If characters are non-independent or correlated, then the method may overestimate the confidence in the distinctness of putative species. Finally, our approach assumes that traits are qualitative or discrete rather than being continuous, although it should be possible to use our approach on continuous characters which are treated qualitatively (for example, treat a range of meristic trait values as a single qualitative character, e.g. three or fewer scales versus four or more). Quantitative traits can also be evaluated using standard univariate and multivariate statistical methods.

Our method requires keeping track of the total number of characters examined for potential diagnostic traits, not merely those which exhibit seemingly fixed differences. This may seem unusual to some systematists, but is extremely important. It makes intuitive sense that there would be much stronger evidence for two species being distinct if five out of six characters surveyed exhibited fixed diagnostic differences than if only five out of 100 differed. At least in a crude way, our method makes use of all of the characters scored in an alpha taxonomic study, rather than merely those that are found to be different. In most closely related groups of organisms, there is a finite set of characters which is routinely used by systematists diagnosing and distinguishing species in the group and these may often be the most appropriate sets of characters to consider for our method. Including large numbers of invariant characters which are not relevant to the taxonomic level at hand (e.g. presence of a head or DNA) will artificially decrease the chances of finding a species to be statistically distinct. In addition to incorporating the overall number of characters used in species-level comparisons, our approach also underscores the need to report sample sizes explicitly (the number of individuals sampled per species) as well.

We suggest that our method is most appropriate for morphological and allozyme data. DNA sequence or restriction site data might be better applied to species delimitation using a tree-based approach (*sensu* Baum & Donoghue 1995) to evaluate whether haplotypes of each putative species cluster together as monophyletic groups (e.g. Avise & Ball 1990; Templeton 1998; Brower 1999). A tree-based approach can be applied to morphological and/or allozyme data as well (e.g. Hollingsworth 1998). However, using morphological or allozyme data, populations can be united by shared trait frequencies or similarities in the means of their quantitative traits and, thus, the exclusive clustering of a set of populations need not indicate a cessation (or extreme reduction) of gene flow with other populations, as it potentially does for DNA sequence data. The application of tree-based species delimitation to morphological data (and other sets of unlinked characters) remains an area in need of further study.

A clear implication of our study is that having statistical confidence in species-level decisions based on

diagnostic character differences is difficult and requires extensive sampling of individuals and characters. In many empirical cases, species may be distinct but the supporting character data may be weak, particularly when only a few specimens are available and/or the species diverged too recently for any or enough diagnostic differences to evolve. In such cases, other lines of evidence might be considered, such as (i) obvious reproductive isolation due to distant allopatry or impassable barriers to gene flow, and (ii) the phylogenetic relationships of the taxa compared (e.g. the putative species are not closely related). Differences in the frequencies of more variable traits might also provide important evidence for species delimitation and it may be useful for future studies to develop ways of incorporating frequency differences from all characters. The use of diagnostic character differences (with a frequency cut-off) requires thorough sampling of individuals and characters to allow for statistical confidence, but even with extensive sampling it may be unclear which characters are truly diagnostic in each species. We have presented a methodology for evaluating some of the uncertainty associated with species delimitation using this widespread criterion, but our results also suggest that other criteria, particularly ones which may be more powerful with limited data, should be explored.

We are grateful to Kevin de Queiroz, Brad Livezey, Charles McCulloch, John Rawlins and Rocco Servedio for useful discussion and/or comments on the manuscript.

REFERENCES

Avise, J. C. & Ball, R. M. 1990 Principles of genealogical concordance in species concepts and biological taxonomy. *Oxford Surv. Evol. Biol.* **7**, 45–67.

Baum, D. A. & Donoghue, M. J. 1995 Choosing among alternative 'phylogenetic' species concepts. *Syst. Bot.* **20**, 560–573.

Brower, A. V. Z. 1999 Delimitation of phylogenetic species with DNA sequences: a critique of Davis and Nixon's population aggregation analysis. *Syst. Biol.* **48**, 199–213.

Conover, W. J. 1971 *Practical nonparametric statistics*. New York: Wiley.

Davis, J. I. 1996 Phylogenetics, molecular variation, and species concepts. *Bioscience* **46**, 502–511.

Davis, J. I. & Nixon, K. C. 1992 Populations, genetic variation, and the delimitation of phylogenetic species. *Syst. Biol.* **41**, 421–435.

de Queiroz, K. 1998 The general lineage concept of species, species criteria, and the process of speciation: a conceptual unification and terminological recommendations. In *Endless forms: species and speciation* (ed. D. J. Howard & S. H. Berlocher), pp. 57–75. Oxford University Press.

Eberhard, W. G. 1985 *Sexual selection and animal genitalia*. Cambridge, MA: Harvard University Press.

Eldredge, N. & Cracraft, J. 1980 *Phylogenetic patterns and the evolutionary process*. New York: Columbia University Press.

Frost, D. R. & Kluge, A. G. 1994 A consideration of epistemology in systematic biology, with special reference to species. *Cladistics* **10**, 259–294.

Graybeal, A. 1995 Naming species. *Syst. Biol.* **44**, 237–250.

Hollingsworth, B. D. 1998 The systematics of chuckwallas (*Sauromalus*) with a phylogenetic analysis of other iguanid lizards. *Herpetol. Monogr.* **12**, 38–191.

McKitrick, M. C. & Zink, R. M. 1988 Species concepts in ornithology. *Condor* **90**, 1–14.

Nixon, K. C. & Wheeler, Q. D. 1990 An amplification of the phylogenetic species concept. *Cladistics* **6**, 211–223.

Olmstead, R. G. 1995 Species concepts and plesiomorphic species. *Syst. Bot.* **20**, 623–630.

Porter, A. H. 1990 Testing nominal species boundaries using gene flow statistics: the taxonomy of two hybridizing admiral butterflies (*Limenitis*: Nymphalidae). *Syst. Zool.* **39**, 148–161.

Rannala, B. 1995 Polymorphic characters and phylogenetic analysis: a statistical perspective. *Syst. Biol.* **44**, 421–429.

Slatkin, M. & Barton, N. H. 1989 A comparison of three indirect methods for estimating average levels of gene flow. *Evolution* **43**, 1349–1368.

Swofford, D. L. & Berlocher, S. H. 1987 Inferring evolutionary trees from gene frequency data under the principle of maximum parsimony. *Syst. Zool.* **36**, 293–325.

Templeton, A. R. 1998 Species and speciation: geography, population structure, ecology, and gene trees. In *Endless forms: species and speciation* (ed. D. J. Howard & S. H. Berlocher), pp. 32–43. Oxford University Press.

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

Wright, S. 1937 The distributions of gene frequencies in populations. *Proc. Natl Acad. Sci. USA* **23**, 307–320.

Wright, S. 1978 *Evolution and the genetics of populations. IV. Variation within and among natural populations*. University of Chicago Press.

Electronic appendices to this paper can be found at (http://www.pubs.royalsoc.ac.uk/publish/pro_bs/rpbl444.htm).