

Endless tests: guidelines for analysing non-nested sister-group comparisons

Steven M. Vamosi* and Jana C. Vamosi

*Department of Biological Sciences, University of Calgary,
2500 University Drive NW, Calgary, Alberta T2N 1N4, Canada*

ABSTRACT

Question: How can we best use phylogenies and sister-group comparisons to understand the impact of ecological and life-history traits on diversification rates?

Analysis: Brief review of the basic structure of sister-group comparisons. Description of the sign, Slowinski-Guyer and species-diversity contrast tests. Elucidation of the potential shortcomings of these tests.

Conclusions: The Slowinski-Guyer test has statistical flaws and should no longer be used to analyse hypotheses about the effects of traits on diversification rates. Species-diversity contrast tests are the most conservative, yet powerful, methods for analysing sister-group comparisons.

Keywords: diversification, key innovation, sister-group comparison, Slowinski-Guyer test.

INTRODUCTION

Sister-group comparisons are usually invoked when attempting to address whether a particular trait or syndrome is associated with variation in diversification (i.e. speciation minus extinction) rates (Slowinski & Guyer, 1993; Nee *et al.*, 1996; Barraclough *et al.*, 1998; de Queiroz, 1998). Several traits have been shown to be correlated with either decreased or increased diversification rates, including body size (Gittleman & Purvis, 1998) and phytophagy (Mitter *et al.*, 1988) in animals, and floral symmetry (Sargent, 2004) and sexual system (Heilbuth, 2000; Vamosi and Vamosi, 2004) in plants. Heilbuth (2000), for example, used sister-group comparisons to demonstrate that the low representation of dioecy in angiosperms was associated with lower speciation and/or higher extinction rates of dioecious lineages. This conclusion was based on the observation that dioecious lineages tended to have fewer species than their non-dioecious sister groups. If a single well-resolved phylogeny is available for a group that possesses variation in a continuous trait, one can analyse whether differences in species richness correlate with differences in that trait using MacroCAIC (Agapow and Isaac, 2002). However, in the case of categorical traits that are relatively rare (e.g. dioecy in angiosperms), it is necessary to use several different phylogenies to address whether a relationship is present between diversification and the trait of interest. In these cases, the preferred methodology to address

* Author to whom all correspondence should be addressed. e-mail: smvamosi@ucalgary.ca
Consult the copyright statement on the inside front cover for non-commercial copying policies.

the effect of a trait on diversification is to consult phylogenies of the groups of organisms of interest, and identify sister taxa that differ in the trait of interest (see Fig. 1). In Fig. 1, the lineage with the trait of interest has more species than its sister group, which is associated with the other state of the trait. This sister-group pair would constitute one independent replicate, and multiple other sister-group pairs composed of different groups of genera (or other higher-order taxa) would be required to allow the application of recommended statistical tests of the hypothesis. Because sister taxa are the same age by definition (Vrba, 1980; Felsenstein, 1985), this approach accounts for the effects of shared ancestry and focuses on the differences in diversification that have accrued during the time since the taxa last shared a common ancestor. Rather than elaborating on our summary, we point readers to more detailed descriptions of the method that can be found elsewhere (e.g. Nee *et al.*, 1996; Barraclough *et al.*, 1998; Heilbut, 2000).

ANALYSING SISTER-GROUP COMPARISONS

Throughout this paper, we assume that the reader is interested in the practical aspects of determining the evolutionary effects of the presence or absence of a trait on diversification. For some perspective on the philosophical debate that has surrounded the practice of correcting for phylogenetic relatedness of lineages, we point interested readers to Rosenzweig's (1996) discussion about the potential loss of statistical power associated with considering only sister taxa with contrasting values of the trait of interest, and Barraclough and colleagues' (1998) detailed reply that elaborated on the justification for the use of sister-group comparisons in addressing the correlates of diversification. We also reiterate the caveat that traits may open up novel ecological opportunities without affecting speciation or extinction rates (Schluter, 2000). A potentially interesting application of the use of sister-group comparisons to explore ecological differentiation of related taxa, illustrated with a consideration of seed size evolution, was recently developed by Ackerly and Nyffeler (2004). Finally, a sister-group comparison framework may also be applied when attempting to understand the effects of human activities on taxa possessing different ecological and life-history traits (Vamosi and Vamosi, 2005).

The two main decisions faced when undertaking a sister-group comparison are how to: (1) assign a value to the contrast in species richness for each sister-group pair, and (2) combine the information from a number of sister-group pairs to produce a meaningful test (especially when the total number of sister-group pairs is low). Several methods have been devised for the analysis of sister-groups comparisons, with various approaches to the two steps outlined above (e.g. Mitter *et al.*, 1988; Slowinski and Guyer, 1993; Wiegmann *et al.*, 1993; Barraclough *et al.*,

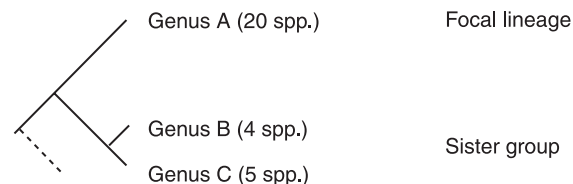


Fig. 1. An example of a sister-group pair that would constitute a single replicate in a sister-group comparison. In this example, the clade with the trait of interest (i.e. the focal lineage) has more species than its sister group, which is lacking the trait of interest. The focal lineage has 20 species, whereas the sister group has 9 species.

1995, 1996). Here, we will not consider a further alternative, the maximum likelihood approach recently advocated by McConway and Sims (2004). The main impetus for developing their approach was spurred by a concern over the shortcomings of the Slowinski-Guyer method (Slowinski and Guyer, 1993). Their analyses, accordingly, are focused primarily on comparisons between the two methods. Interestingly, they state that neither method is appropriate when one 'labels' *a priori* a particular member of a sister-group pair, which is exactly what one would do before determining whether a particular trait is associated with higher or lower diversification rates. Thus, their analysis is most appropriate for determining whether heterogeneity exists, not if that heterogeneity is associated with a particular trait.

To help illustrate our discussion of the various methods, we consider the relationship between species diversity and fruit characteristics (fleshy or dry) for 19 sister-group pairs of angiosperms (Table 1). In each sister-group pair, one clade has fleshy fruits and the other has dry fruits. These sister-group pairs were derived from a larger sample of non-dioecious (i.e. monoecious and hermaphroditic) families considered by Vamosi and Vamosi (2004). We chose this data set for two reasons. First, the sample size is within the range of several published studies [for example, 10 (Slowinski and Guyer, 1993); 19 (Sargent, 2004); 31 (Barracough *et al.*, 1995)]. Second, fruit characteristics have been hypothesized to influence diversification rates, although a consensus about whether fleshy or dry fruits are correlated with increased diversification rates has not been reached (e.g. Donoghue, 1989; Tiffney and Mazer, 1995; Smith, 2001; Vamosi and Vamosi, 2004). Our goal here is to contrast the various methods, not to demonstrate conclusively the impact of fruit characteristics on angiosperm diversification.

SIGN TEST

The sign test was the first test applied to the analysis of sister-group comparisons (e.g. Mitter *et al.*, 1988; Farrell *et al.*, 1991). Considering our example, we can ask: 'Do families with fleshy fruits *typically* have more species than their sister groups (that lack fleshy fruits)?' Here we are not concerned with effect size, rather with whether the trait of interest is associated with the larger clade more often than expected based on the binomial distribution. The clade with fleshy fruits had more species in only 7 of the 19 comparisons, whereas the reverse was true in the remaining 12 comparisons (Table 1). This pattern is judged to be not significantly different from the random expectation with a one-tailed sign test ($P=0.36$). For 19 replicates, the larger clade would have to be associated with the trait of interest in 14 or more sister-group pairs to reject the null hypothesis. Because the sign test ignores the magnitude of differences in species richness, it has considerably less power than the Wilcoxon signed-rank test (Zar, 1984), which is typically applied when using the various species diversity contrast methods. Use of the sign test in sister-group comparisons should be limited to cases where only approximate numbers of species or the order of the comparison is known with confidence.

SLOWINSKI-GUYER METHOD

A common criticism of studies that utilize sister-group comparisons is that the results *need* to be analysed with the method first proposed by Slowinski and Guyer (1993). The perception that the Slowinski-Guyer method is the *de facto* test for sister-group comparisons is perhaps surprising, given that its various flaws have been discussed by a number of authors (e.g. Nee *et al.*, 1996; de Queiroz, 1998; Goudet, 1999; Schluter, 2000; McConway and Sims, 2004). The very fact that studies that

Table 1. Non-dioecious angiosperm sister-group pairs in which one clade has fleshy fruits and the other has dry fruits

Fleshy clade	No. of species	Dry clade	No. of species	Sign	SloGuy <i>p</i>		deQu <i>p</i>		Contrasts		
					(F > D)	(D > F)	(F > D)	(D > F)	WMF (1993)	BHN (1995)	BHN (1996)
<i>Pangium</i>	1	<i>Acharia</i> + <i>Kigellaria</i>	2	-	1.000	0.500	1.000	0.560	-1	0.333	-1.585
* <i>Cyrilla</i>	1	<i>Clethra</i>	64	-	1.000	0.016	1.000	0.172	-63	0.015	-6.022
<i>Roussea</i>	1	<i>Lobelia</i>	300	-	1.000	0.003	1.000	0.054	-299	0.003	-8.234
<i>Myrtophyllum</i> + <i>Haloragis</i> + <i>Penthorum</i>	1	<i>Tetracarpaea</i>	89	-	1.000	0.011	1.000	0.344	-88	0.011	-6.492
<i>Austrobaileya</i>	1	<i>Illicium</i> + <i>Schisandra</i>	67	-	1.000	0.015	1.000	0.172	-66	0.015	-6.087
<i>Davidsonia</i>	3	<i>Bauera</i>	4	-	0.667	0.500	0.279	0.560	-1	0.429	-1.161
<i>Mitchella</i>	3	<i>Pentas</i>	34	-	0.944	0.083	0.817	0.323	-31	0.081	-2.565
<i>Milligania</i>	5	<i>Borya</i>	10	-	0.714	0.357	0.667	0.516	-5	0.333	-1.339
<i>Sambucus</i>	9	<i>Viburnum</i>	150	-	0.949	0.057	0.817	0.226	-141	0.057	-2.179
<i>Pereskia</i>	16	<i>Mollugo</i>	35	-	0.700	0.320	0.656	0.495	-19	0.314	-1.265
* <i>Decaisnea</i> + <i>Sargentodoxa</i> + <i>Tinospora</i> + <i>Menispermum</i> + <i>Nandina</i> + <i>Caulophyllum</i> + <i>Hydrastis</i> + <i>Glaucidium</i>	33	<i>Euptelea</i>	2	+	0.059	0.971	0.108	0.830	31	0.943	3.210

<i>Tetracera</i>	40	<i>Dillenia</i>	60	-	0.606	0.404	0.237	0.527	-20	0.400	-1.107
<i>Osbeckia</i>	50	<i>Mouriri</i>	81	-	0.623	0.385	0.634	0.516	-31	0.382	-1.121
<i>Hippocratea</i>	100	<i>Plagiopteron</i>	1	+	0.010	1.000	0.108	1.000	99	0.990	6.658
* <i>Cyclanthus</i> +	216	<i>Petrosavia</i> + <i>Japonlirion</i>	3	+	0.014	0.991	0.140	0.892	213	0.986	3.881
<i>Sphaeradenia</i> +											
<i>Freycinetia</i>											
<i>Bixa</i>	393	<i>Theobroma</i> + <i>Grewia</i> + <i>Tilia</i> + <i>Sterculia</i> + <i>Durio</i>	1	+	0.003	1.000	0.043	1.000	392	0.997	8.622
<i>Impatiens</i>	850	<i>Idria</i>	11	+	0.013	0.988	0.129	0.871	839	0.987	2.715
<i>Lamium</i> + <i>Clerodendrum</i> + <i>Callicarpa</i> + <i>Phyla</i> + <i>Pedicularis</i> + <i>Paulownia</i>	947	<i>Euthystachys</i>	1	+	0.001	1.000	0.022	1.000	946	0.999	9.889
<i>Solanum</i>	1700	<i>Nolana</i>	18	+	0.010	0.990	0.108	0.892	1682	0.990	2.526

Note: Information on phylogenetic relationships and species richness obtained from the data set used by Vamosi and Vamosi (2004). Test statistics presented for: sign test (+ = fleshy clade with more species; - = dry clade with more species); Slowinski-Guyer (SloGuy) test ($F > D$ = testing hypothesis that fleshy clades are larger than dry clades; $D > F$ = testing hypothesis that dry clades are larger than fleshy clades); de Queiroz's (deQu) modification to the Slowinski-Guyer test (hypotheses tested same as for Slowinski-Guyer test); Wiegmann and colleagues' (1993) test for differences in species richness (positive values assigned to contrasts in which the fleshy clade is larger than the dry clade); Barracough and colleagues' (1995) test (proportion of species in clade with trait of interest/total number of species in sister-group pair); Barracough and colleagues' (1996) test [$\log$ (proportion of species in larger clade + 1)/ $\log$ (proportion of species in smaller clade + 1)]. Asterisks indicate the sister-group pairs that are used for the analysis in Table 2.

utilize this method continue to be published (e.g. Smith, 2001; Holliday and Steppan, 2004) suggests that readers have not appreciated the descriptions of the underlying flaws and/or that suitable alternatives have not been clearly identified. Because of this, we illuminate the limitations of the Slowinski-Guyer method using plain language and a real data set, referring readers to the results of simulation studies where appropriate. Later, we use the same approach when discussing the preferred alternative methods.

First, we review the basic structure of this method. For each sister-group pair, given a species in the lineage with the trait of interest and b species in the sister group, one calculates:

$$P = \frac{b}{(a + b - 1)} \quad (1)$$

Each of these proportions is log transformed and the sum of these values is multiplied by -2 and compared with a χ^2 distribution with $2k$ degrees of freedom, where k is the number of sister-group pairs. Applying the Slowinski-Guyer method to our data in Table 1, we find highly significant support ($\chi^2 = 71.28$, $P = 0.0009$) for the hypothesis that fleshy-fruited clades experience enhanced diversification rates.

Such striking rejection of the null hypothesis using an analysis based on only 19 replicates warrants closer inspection. Recall that only 7 of the 19 sister-group pairs were in the predicted direction (i.e. clade with fleshy fruits larger than sister group with dry fruits; see Sign column in Table 1). In other words, the clade with fleshy fruits is actually smaller than the clade with dry fruits in more than half of the sister-group pairs. Furthermore, the clade with dry fruits was strikingly larger than the clade with fleshy fruits in several of the 12 sister-group pairs that were not in the predicted direction, yet this did not appear to diminish the support for the hypothesis. Finally, if we had instead used the same data set to test whether dry fruits lead to increased diversification (i.e. the opposite hypothesis), we would have rejected the null hypothesis in this case as well ($\chi^2 = 58.8$, $P = 0.017$).

How can a single data set provide support for opposing hypotheses? The root of the problem lies with the use of Fisher's combined probability test, and is perhaps best stated thus: 'For any given set of k comparisons, if $k - n$ comparisons produce a test statistic greater than the critical value, then the direction or magnitude of the remaining n comparisons has no consequence for the outcome'. Nee *et al.* (1996, p. 246) first pointed out that 'using this procedure, it would be possible to establish significance . . . with just one sister-group comparison, if there is a large difference in size between clades'. Curiously, this statement, and the accompanying discussion of significance regions for combined probability tests, has been largely ignored or dismissed. This issue crops up in our analyses of the data in Table 1 and is best illustrated with a subset of the sister-groups (Table 2). Heterogeneity in species richness for these six sister-group pairs is random with respect to the trait being investigated. However, the Slowinski-Guyer method would lead one to reject with confidence the null hypothesis ($P = 0.03$). Furthermore, this problem becomes progressively more serious with increased sample sizes. For example, doubling the cumulative probability observed in Table 2 (corresponding to 12 sister-group pairs, with two instances of each sister-group pair that support the prediction and two instances of each sister-group pair that oppose the prediction) results in a highly significant (false) rejection of the null hypothesis ($P = 0.006$). Because they do not rely on combined probabilities, neither the sign test nor any of the species diversity contrast methods reject the null hypothesis in either case.

Table 2. Heterogeneity in species richness among clades in sister-group pairs and its effects on the Slowinski-Guyer method

Sister-group pair	Clade with trait	Sister group without trait	Slowinski-Guyer P	Probability	Cumulative probability
1	216	3	0.014	8.54	8.54
2	64	1	0.016	8.27	16.81
3	33	2	0.059	5.66	22.47*
4	2	33	0.971	0.06	22.53
5	3	216	0.991	0.02	22.55
6	1	64	1.000	0	22.55

Note: Each of three sister-group pairs from Table 1 each appear twice, once in the predicted direction (i.e. clade with trait is larger than sister group lacking the trait) and once in the opposite direction. For example, the contrast of 216 species vs. 3 species appears in sister-group pairs 1 and 5. Although the heterogeneity in species richness is random with respect to the trait being considered, the null hypothesis of no association between the trait and diversification is rejected ($P = 0.03$). The asterisk indicates that a consideration of only the first three sister-group comparisons was required to reject the null hypothesis (i.e. the cumulative probability from sister-group pairs 1 to 3 was greater than the critical value of 21.026 with $k = 12$ degrees of freedom).

The Slowinski-Guyer method is thus prone to elevated Type I errors, and this bias is especially pronounced when the data possess a U-shaped distribution (Goudet, 1999). U-shaped distributions, which arise when there are large differences in species diversity of sister lineages regardless of their traits, have been frequently reported (e.g. de Queiroz, 1998; Mooers and Heard, 1997; Goudet, 1999; Sims and McConway, 2003). Indeed, Goudet (1999) found that 7 of 11 data sets used to test for associations between traits and diversification rates, including the one used by Slowinski and Guyer (1993) to illustrate their method, possessed U-shaped distributions of P -values. We can demonstrate the effect by plotting the proportion of species within each sister-group pair that have fleshy fruits (Fig. 2A). Species with fleshy or dry fruits comprised 80% or more of the total number of species in the sister-group pair in six (32%) or seven (37%), respectively, of the sister-group pairs. Furthermore, only one of the 19 sister-group pairs [*Davidsonia* (fleshy clade) vs. *Bauera* (dry clade)] had a roughly equal number of species in each lineage. Thus, both states of the categorical trait being considered were associated with a few sister-group pairs that had many species (and whose sister group had few species).

Recognizing that large differences in species richness of members of sister-group pairs were quite common, de Queiroz (1998) advocated a modification: test the proportions generated by the Slowinski-Guyer test against the 'true' null distribution of species diversity differences derived from real phylogenies. We generated a null distribution for the data in Table 1 by sampling a set of 94 non-dioecious sister-group pairs from Vamosi and Vamosi (2004). The focal clades and their sister groups in this sample are random with respect to their fruit characteristics. Not surprisingly, the expectation in this null distribution was not 0.5, with the median proportion being 0.31. Furthermore, over half of the comparisons produced extreme values: about one-third (32 of 94) of the pairs had proportions ≤ 0.10 and another fifth (20 of 94) had proportions ≥ 0.90 . We then determined what proportion of the sister-group pairs in this larger data set generate Slowinski-Guyer probabilities at least as extreme as observed for each of the 19 sister-group pairs. We found no support for either

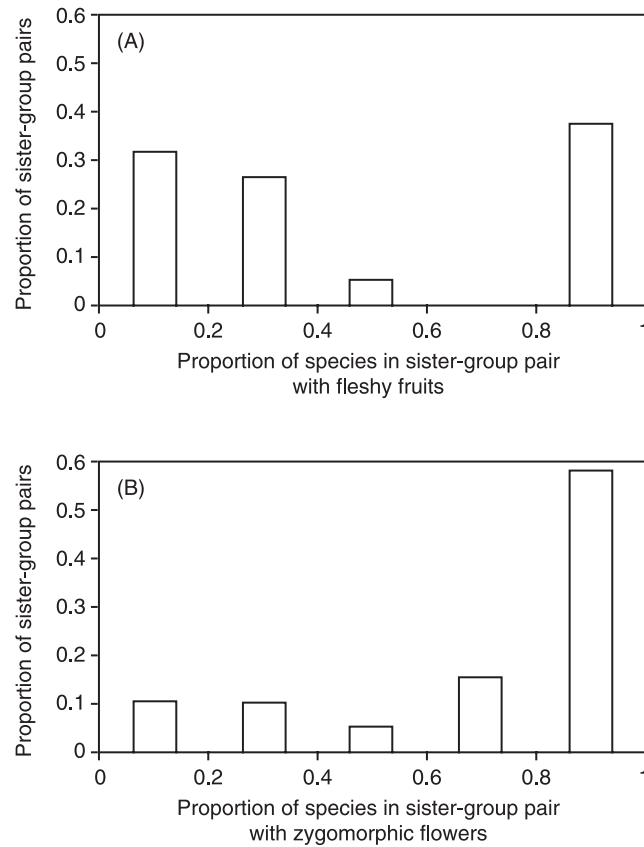


Fig. 2. The distribution of the proportion of species (categorized as 0–20%, >20–40%, . . . >80–100%) with the trait of interest [A: fleshy fruits, data from Table 1; B: zygomorphic flowers, data from Sargent (2004)] in each sister-group pair. In panel A, there are a roughly equal number of sister-group pairs that are composed almost exclusively of species with dry (i.e. 0–20%) or fleshy (>80–100%) fruits. Conversely, in panel B, there are few sister-group pairs that are composed almost exclusively of species with actinomorphic (i.e. 0–20%) flowers. Species diversity contrast methods confirm that zygomorphic flowers, but not fleshy fruits, are associated with increased diversification. $N = 19$ sister-group pairs in both panels.

hypothesis – that is, that fleshy ($P = 0.23$) or dry ($P = 0.85$) fruits are associated with higher species richness.

The use of random null distributions has not caught on, probably for two good reasons. First, there is no way to know with any confidence the appropriate null distribution, as admitted by de Queiroz (1998). Certainly our null distribution captured the inherent asymmetry in species richness among the lineages that comprise sister-group pairs. However, if we were to base our conclusions about the impact of fruit characteristics on diversification on the results of this analysis, we should not be surprised to encounter a number of criticisms, not the least of which would be: why did we use this particular sample of sister-group pairs? Second, this modified test continues to rely on a combined probability framework, which is fraught with a number of shortcomings.

SPECIES DIVERSITY CONTRAST METHODS

Several methods exist that consider both the sign and magnitude of species richness differences among members of sister-group pairs, yet do not use combined probabilities (e.g. Wiegmann *et al.*, 1993; Barraclough *et al.*, 1995, 1996). The diversity of alternatives may have inadvertently contributed to the continued use of the Slowinski-Guyer method; indeed, Barraclough and colleagues (1995, 1996) used two different variants in subsequent papers with no explanation for the change or, for that matter, the choice of method in either study. The latter contrast measure, which relies on log transformations of species numbers, likely derives from the expectation that the log of species number may be proportional to net speciation rate under certain models of diversification (e.g. Stanley, 1979).

The general approach when applying species diversity contrast methods involves a series of simple steps. First, contrasts in species diversity are tabulated for every species pair. Second, these contrasts are ranked from largest to smallest irrespective of whether they support or oppose the hypothesis (i.e. by their absolute value). Third, contrasts that are in agreement with the hypothesis are assigned a '+' and those that oppose the hypothesis are assigned a '-'. Finally, the values are tested against the null expectation with a non-parametric test, typically either the randomization test for matched pairs (if sample sizes are low) or the Wilcoxon signed-rank test (Siegel, 1956).

The major difference between the various methods is the way in which the species diversity contrast values are calculated. Wiegmann *et al.* (1993) (hereafter the WMF method) simply calculated the difference between the species richness of the focal clade and its sister group for each sister-group pair. For example, the WMF method would produce a species diversity contrast value of $20 - 9 = 11$ for the sister-group pair in Fig. 1. Barraclough *et al.* (1995) (hereafter the BHN95 method) calculated the proportion of species in the focal lineage [i.e. $20/(20 + 9) = 0.69$ for the example in Fig. 1]. In contrast, Barraclough *et al.* (1996) (hereafter the BHN96 method) calculated:

$$\frac{\log(x)}{\log(y)} \quad (2)$$

for each sister-group pair, where x is the number of species in the larger clade and y is the number of species in the smaller clade. Using our sister-group pair in Fig. 1, the BHN96 method would produce a value of $\log(20)/\log(9) = 1.36$. The differences in calculating the contrasts also change the null expectation. With multiple sister-group pairs, the null expectations for the three methods are 0, 0.5 and 0, respectively.

Although we advocate the general approach, there are some difficulties with the calculation of the individual contrasts that first need to be dealt with. Because it relies on raw differences, the WMF method would produce the same test statistic (i.e. 10) for the following sister-group pairs: (1020 vs. 1010 species) and (20 vs. 10 species). Intuitively, the latter contrast should be assigned greater support for the hypothesis than the former contrast. The BHN96 method solves this problem but introduces a related one. Consider the following sister-group pairs: (500 vs. 50 species) and (50 vs. 5 species). There are 10 times as many species in the focal group as in its sister group in both pairs, yet their analysis would produce a score of 1.59 and 2.43 for the first and second comparison, respectively. This property of the test would lead to the inflation of Type I error if young lineages were disproportionately represented among the groups in favour of the hypothesis. Also, if the data set includes even a single clade that is monotypic (as in Table 1), one needs to add

the value of one to all signed-ranks values to prevent division by zero. Finally, sister-group pairs in which the two clades have equal species richness need to have their contrast value manually changed to zero.

The power and error rates of the different metrics have recently been examined using simulations of nested sister-group comparisons (Isaac *et al.*, 2003). Although they were interested in continuous variables, the findings of their study are applicable to our discussion (A. Mooers, personal communication). The use of raw differences in species richness (i.e. the WMF method) was observed to have high and variable Type I error rates and low power, while using log-transformed ratios (BHN96 method) or proportional differences (BHN95 method) was observed to have acceptably low Type I error rates and high power (Isaac *et al.*, 2003). Their recommendation was to use proportional differences if branch lengths are unknown and sample sizes are low, and log ratios in all other cases. Because a simple sister-group comparison does not typically incorporate branch length information, it may follow that the BHN95 method would be the most reliable way to perform a sister-group comparison. However, the log transformation employed by the BHN96 method has the advantage of controlling for the multiplicative nature of diversification (i.e. that initial differences in species diversity become ever larger over time). The randomization test for matched pairs as used by Barraclough *et al.* (1995) is needed only if sample sizes are low [Siegel (1956) suggests $N = 12$ or fewer]; for larger sample sizes, the Wilcoxon signed-rank test may be applied. Because sister-group pairs used in a study are frequently of different ages, and hence are drawn from different distributions, parametric tests should be avoided. Monte Carlo simulations suggest that the Wilcoxon signed-rank test is as powerful as the *t*-test when the assumption of normality is not upheld and, indeed, may be more powerful with small sample sizes (Tanizaki, 1997).

Because these methods are not influenced by data with U-shaped distributions (e.g. Goudet, 1999), we do not find support for the hypothesis that possessing fleshy fruits enhances diversification (using the sister-group pairs in Table 1). Applying the Wilcoxon signed-rank test, we obtain the following *P*-values: 0.38, 0.41 and 0.51 for the WMF, BHN95 and BHN96 methods, respectively.

SLOWINSKI-GUYER VERSUS SPECIES DIVERSITY CONTRAST METHODS

Although we largely agree with assessments of the Slowinski-Guyer test made previously (Nee *et al.*, 1996; de Queiroz, 1998; McConway and Sims, 2004), we wish to address a comment made by de Queiroz (1998) regarding alternative methods. In his consideration of the statement made by Nee *et al.* (1996) that the Slowinski-Guyer test can reject the null hypothesis with a single sister-group pair, de Queiroz (1998, p. 711) argued, 'Nee *et al.*'s criticism should be generalized to all tests that consider the magnitude of differences in clade size'. This statement appears to have been misinterpreted as indicating that only the sign test avoids this criticism of the Slowinski-Guyer test. The null hypothesis of the species diversity contrast methods is conceptually identical to the sign test, rather than to the Slowinski-Guyer test. In the latter, the null hypothesis is that species richness of sister groups should be roughly equivalent. In the sign test and species diversity contrast methods, however, the null hypothesis is that the directions of species diversity contrasts are random with respect to any given trait. Therefore, unlike in the Slowinski-Guyer test, if no clear pattern exists, extreme species diversity contrasts in one direction will tend to be balanced by extreme contrasts in the opposite direction. Furthermore, one cannot conduct a randomization

or Wilcoxon signed-rank test with only a single replicate as one can for the Slowinski-Guyer test.

Because we are trying to settle an issue rather than fuel a debate, we would like to point out that both de Queiroz (1998) and McConway and Sims (2004) admitted, near the end of their papers, that species diversity contrast methods (e.g. Barraclough *et al.*, 1996) were reasonable choices. In the former paper, however, this statement appears to have been overlooked by readers and, in the latter, it was mentioned as an afterthought. Thus, we feel the need to unequivocally state that these species diversity contrast methods are the most conservative yet powerful methods available for ascertaining whether a particular trait, or trait complex, is associated with variation in diversification rates.

A PROSPECTUS

Data exploration has long been considered a good first step in analysing the results of an experiment. In an analogous fashion, we propose that the distribution of the proportion of species with the trait of interest for all sister-group pairs be considered before formal analyses *and* presented in articles. Applying this approach to the 19 sister-group pairs of angiosperms studied by Sargent (2004), for example, we find a right-skewed distribution (Fig. 2B), lending support to her conclusion that angiosperm lineages with zygomorphic flowers are typically larger than sister groups with actinomorphic flowers. Combined with a proper analysis, a data set possessing a U-shaped distribution (e.g. Fig. 2A) should not be considered as providing support for an association between a particular trait and diversification.

We advocate with some vigour that further use of the Slowinski-Guyer method be abandoned. Although it spurred interest in analysing the correlates of diversification, its methodological shortcomings are legion (Nee *et al.*, 1996; de Queiroz, 1998; Goudet, 1999; Schluter, 2000; McConway and Sims, 2004; this paper). To quote Nee *et al.* (1996, p. 246), the Slowinski-Guyer method simply ‘cannot be used to test hypotheses about the correlates of diversification’. The sign test, although it avoids the pitfalls of the combined probability framework, is rather limited in its applicability. For the vast majority of cases, we recommend the use of species diversity contrast methods, with the possible exception of the differences method first used by Wiegmann *et al.* (1993). The log transformation method of Barraclough *et al.* (1996) may be most appropriate, given the multiplicative nature of diversification that is typically assumed. Finally, contrasts should be analysed with non-parametric tests, typically the Wilcoxon signed-rank test or randomization procedures when few sister-group pairs are available.

Understanding the role of phenotypic traits and ecological interactions in macroevolution is a fundamental goal of evolutionary biology and should not be hampered by confusion over the assumptions and limitations of statistical methods. We hope that our treatment of sister-group comparisons will help remove these impediments and encourage further studies of the processes that generate biological diversity.

ACKNOWLEDGEMENTS

We wish to thank A. de Queiroz, L. Harder, R. Laird, S. Mazer, A. Mooers, M. Routley, D. Schluter, D. Sikes and J. Tindall for discussions and comments on previous versions of the manuscript and NSERC (Canada) for financial support.

REFERENCES

- Ackerly, D.D. and Nyffeler, R. 2004. Evolutionary diversification of continuous traits: phylogenetic tests and application to seed size in the California flora. *Evol. Ecol.*, **18**: 249–272.
- Agapow, P.M. and Isaac, N.J.B. 2002. MacroCAIC: revealing correlates of species richness by comparative analysis. *Divers. Distrib.*, **8**: 41–43.
- Barracough, T.G., Harvey, P.H. and Nee, S. 1995. Sexual selection and taxonomic diversity in passerine birds. *Proc. R. Soc. Lond. B*, **259**: 211–215.
- Barracough, T.G., Harvey, P.H. and Nee, S. 1996. Rate of *rbcL* gene sequence evolution and species diversification in flowering plants (angiosperms). *Proc. R. Soc. Lond. B*, **263**: 589–591.
- Barracough, T.G., Nee, S. and Harvey, P.H. 1998. Sister-group analysis in identifying correlates of diversification. *Evol. Ecol.*, **12**: 751–754.
- de Queiroz, A. 1998. Interpreting sister-group tests of key innovation hypotheses. *Syst. Biol.*, **47**: 710–718.
- Donoghue, M.J. 1989. Phylogenies and the analysis of evolutionary sequences, with examples from seed plants. *Evolution*, **43**: 1137–1156.
- Farrell, B., Dussourd, D.E. and Mitter, C. 1991. Escalation of plant defense: Do latex and resin canals spur plant diversification? *Am. Nat.*, **138**: 881–900.
- Felsenstein, J. 1985. Phylogenies and the comparative method. *Am. Nat.*, **125**: 1–15.
- Gittleman, J.L. and Purvis, A. 1998. Body size and species-richness in carnivores and primates. *Proc. R. Soc. Lond. B*, **265**: 113–119.
- Goudet, J. 1999. An improved procedure for testing the effects of key innovations on rate of speciation. *Am. Nat.*, **153**: 549–555.
- Heilbuth, J.C. 2000. Lower species richness in dioecious clades. *Am. Nat.*, **156**: 221–241.
- Holliday, J.A. and Steppan, S.J. 2004. Evolution of hypercarnivory: the effect of specialization on morphological and taxonomic diversity. *Paleobiology*, **30**: 108–128.
- Isaac, N.J.B., Agapow, P.M., Harvey, P.H. and Purvis, A. 2003. Phylogenetically nested comparisons for testing correlates of species-richness: a simulation study of continuous variables. *Evolution*, **57**: 18–26.
- McConway, K.J. and Sims, H.J. 2004. A likelihood-based method for testing for nonstochastic variation of diversification rates in phylogenies. *Evolution*, **58**: 12–23.
- Mitter, C., Farrell, B. and Wiegmann, B.J. 1988. The phylogenetic study of adaptive zones: has phytophagy promoted insect diversification? *Am. Nat.*, **132**: 107–128.
- Mooers, A.Ø. and Heard, S.B. 1997. Inferring evolutionary process from phylogenetic tree shape. *Quart. Rev. Biol.*, **72**: 31–54.
- Nee, S., Barracough, T.G. and Harvey, P.H. 1996. Temporal changes in biodiversity: detecting patterns and identifying causes. In *Biodiversity: A Biology of Numbers and Difference* (K.J. Gaston, ed.), pp. 230–252. Oxford: Oxford University Press.
- Rosenzweig, M.L. 1996. Colonial birds probably do speciate faster. *Evol. Ecol.*, **10**: 681–683.
- Sargent, R.D. 2004. Floral symmetry affects speciation rates in angiosperms. *Proc. R. Soc. Lond. B*, **271**: 603–608.
- Schluter, D. 2000. *The Ecology of Adaptive Radiation*. Oxford: Oxford University Press.
- Siegel, S. 1956. *Nonparametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill.
- Sims, H.J. and McConway, K.J. 2003. Nonstochastic variation of species-level diversification rates within angiosperms. *Evolution*, **57**: 460–479.
- Slowinski, J.B. and Guyer, C. 1993. Testing whether certain traits have caused amplified diversification: an improved method based on a model of random speciation and extinction. *Am. Nat.*, **142**: 1019–1024.
- Smith, J.F. 2001. High species diversity in fleshy-fruited tropical understory plants. *Am. Nat.*, **157**: 646–653.
- Stanley, S.M. 1979. *Macroevolution: Pattern and Process*. San Francisco, CA: Freeman.

- Tanizaki, H. 1997. Power comparison of non-parametric tests: small-sample properties from Monte Carlo experiments. *J. Appl. Stat.*, **24**: 603–632.
- Tiffney, B.H. and Mazer, S.J. 1995. Angiosperm growth habit, dispersal and diversification reconsidered. *Evol. Ecol.*, **9**: 93–117.
- Vamosi, J.C. and Vamosi, S.M. 2004. The role of diversification in causing the correlates of dioecy. *Evolution*, **58**: 723–731.
- Vamosi, J.C. and Vamosi, S.M. 2005. Present day risk of extinction may exacerbate the lower species richness of dioecious clades. *Divers. Distrib.*, **11**: 25–32.
- Vrba, E.S. 1980. Evolution, species and fossils: how does life evolve? *S. Afr. J. Sci.*, **76**: 61–84.
- Wiegmann, B., Mitter, C. and Farrell, B. 1993. Diversification of carnivorous parasitic insects: extraordinary radiation or specialized dead end? *Am. Nat.*, **142**: 737–754.
- Zar, J.H. 1984. *Biostatistical Analysis*, 2nd edn. Englewood Cliffs, NJ: Prentice-Hall.

