Paleobiology, 31(1), 2005, pp. 21–34

Estimating paleodiversities: a test of the taxic and phylogenetic methods

Abigail Lane, Christine M. Janis, and J. John Sepkoski Jr.

Abstract.—The traditional ''taxon counting'' method of estimating ancient biodiversity is open to many criticisms, not least of which are the problem of inconsistency in the preservation of fossil organisms and the associated error on first and last appearance times of taxa. Construction of phylogenetic trees provides a way of correcting the first appearance of a taxon based on the origination time of its sister group. Workers have suggested that biodiversity studies include such phylogenetically implied range extensions. Potential problems with this method, in particular the bias inherent in altering origination—but not extinction—times, and the potential for incorrect addition of ghost ranges if the ancestor of a taxon is defined as its sister, are investigated by using a new computer simulation. The program creates a phylogeny, samples it and then adds ghost lineages, with diversity counts being made at all three stages. Results show that under certain conditions, such as in the case of a taxonomic group with many extant representatives, the phylogenetic method is superior to the taxic at capturing diversity pattern. However, there are also important conditions where the taxic approach provides an equal or superior estimate of diversity, such as if the group is extinct or has few extant lineages. Use of the phylogenetic method has the effect of magnifying the Signor-Lipps sampling effect seen before mass extinction events, and if ancestral species within a phylogeny are misdiagnosed as the sister species of their descendants, the phylogenetic method also consistently overestimates diversity magnitudes.

Abigail Lane. Department of Earth Sciences, Wills Memorial Building, Queens Road, Bristol, BS8 1RJ, United Kingdom
Christine M. Janis. Department of Ecology and Evolutionary Biology, Brown University, Providence, Rhode Island 02912. E-mail: Christine_Janis@Brown.edu*
J. John Sepkoski, Jr.†
*Corresponding author
† Deceased

Accepted: 21 April 2004

Introduction

Investigations into the patterns of ancient biodiversity, or taxonomic richness, are concerned with uncovering the diversification history of a particular group through a particular time period. The group can range in size from diversity within a single family in a specific location, up to the clade of all life found globally, and time periods can range from tens of thousands of years up to hundreds of millions of years (e.g., Valentine 1969; Raup 1972; Sepkoski 1984). However, the method of uncovering such diversification patterns has traditionally been the same. Some level in the taxonomic hierarchy is selected, and the earliest and latest records of any member species are used to define the total geological range of this taxon. Hence there is a phylogenetic aspect to this method—if two individual species belonging to the same genus are discovered on either side of an interval, a generic range extension is assumed, spanning the gap. The numbers of taxa present within a sequence of time intervals are summed to produce a series of diversity counts, which together make up the overall pattern. This reliance on the observed stratigraphic occurrence of taxa has been termed the ''taxic'' (Levinton 1988) or ''taxon counting'' approach.

More recently this direct reading of the history of life from stratigraphy has come under increasing criticism for its reliance on what is perceived as an incomplete and biased sample as represented by the fossil record (e.g., Novacek and Norell 1982; Norell and Novacek 1992a,b; Smith 1988). A second method thus uses various solutions to the problems associated with the incompleteness of the fossil record have been used to attempt to enhance taxonomic richness and diversity estimates:

 0094-8373/05/3101-0003/$1.00

these include idealized sampling theory estimates of unsampled ranges (e.g., Signor and Lipps 1982; Strauss and Sadler 1989; Marshall 1990, 1991); extrapolation of taxonomic richness based on empirical distributions (Anderson et al. 1996); and standardization techniques to correct for differences in sampling rate between time intervals and taxonomic groups (e.g., Raup 1975; Miller and Foote 1996; Alroy 2000; Alroy et al. 2001).

A third method, which has grown out of the use of cladistics to determine phylogeny, is termed the ''phylogenetic'' approach (Smith 1994). This uses the relationships between taxa, as recovered by cladistic analysis and calibrated by stratigraphy, to predict the maximum age of divergence of sister groups, and hence also calculates a minimum estimate of unsampled range prior to first appearance of taxa in the fossil record. This unsampled range has been termed the ''ghost lineage'' (Norell 1993) and is assumed to be a result of gaps in preservation, although Wagner (2000a) suggested such gaps reflect a number of additional parameters including speciation/extinction rates and taxonomic philosophy. A detailed method, formulated by Norell (1992, 1993), proposes that the application of cladistics to patterns of biodiversity offers an alternative and superior estimation of the history of life to that obtained by direct reading of the rock record. The result of applying this method is a taxic history predicted by phylogeny and not always in accordance with the stratigraphic occurrence of fossils (e.g., Smith 1988). Both the taxic and the phylogenetic approaches are attempts to reconstruct the richness component of biodiversity, i.e., numbers of taxa. These are the approaches that are tested here.

The Phylogenetic Method

The basic premise underpinning the phylogenetic method is an extension of the theory of speciation by bifurcation—that is a taxon splitting into two new daughter species, and going extinct in the process (Hennig 1965; Doyle and Donoghue 1993). This view of speciation is at odds with the budding speciation hypothesis (e.g., Mayr 1963; Eldredge and Gould 1972) though it has received support elsewhere (Vrba 1993). The bifurcation model is consistent with Hennig's (1965) rule that sister taxa must have equal first appearance times, and it is this assumption upon which the phylogenetic method is based (Norell 1993).

A logical outcome of the concept of bifurcation is that sister species and sister groups must have originated at the same point in time and so any given taxon must be as old as its sister. The practical application of this logic as embodied by the phylogenetic method causes the first appearance times of taxonomic groups to be adjusted to that of the oldest known occurrence of their sisters. As a consequence, in most situations the range in time of many taxa or groups will be extended backward, regardless of their actual fossil occurrences (Fig. 1).

The unsampled portion of a taxon or group's range prior to the first stratigraphic appearance is known as a ghost lineage. The recovery of some or all of a ghost lineage will extend the range of a taxon backward in time. A specific kind of ghost lineage that extends the range of a group (rather than the range of a single taxon) is known as a ghost taxon. These correspond to the internal, or ancestral, segments of a cladogram and as such are an estimation of the extent of unsampled ancestral lineages within a phylogeny (Fig. 1). The temporal extension of fossil ranges has huge consequences for estimates of biodiversity (= biotic richness) through time. The phylogenetic method for estimating diversity sums not only the known fossil ranges occurring in any given time interval but also the ghost ranges added by cladistic implication. These ghost ranges produce an increase in diversity and often their inclusion will severely modify temporal diversity patterns (Smith 1988; Norell 1993).

There are several criticisms of the phylogenetic method of estimating diversity. First, it has been predicted that there is an inherent bias involved with correcting only the first appearance times of taxa and taxonomic groups (e.g., Wagner 1995, 2000b; Foote 1996a). If a ghost lineage is the unsampled initial portion of a taxon's range, then the corresponding unsampled terminal portion can also be defined.

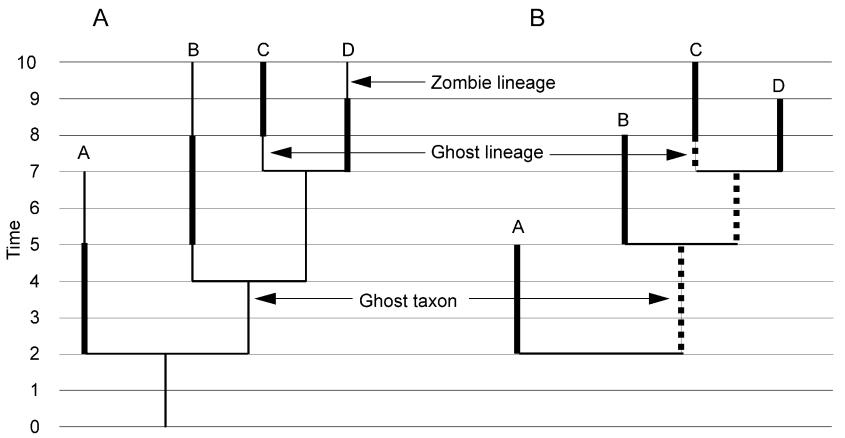

FIGURE 1. The phylogenetic method of recovering unsampled taxon ranges. A, A phylogeny consisting of four terminal taxa plus ancestral lineages, with their complete temporal ranges indicated by the fine lines. Bold lines show just the sampled part of the range from first to last appearance in the fossil record. Examples of a ghost lineage prior to first appearance, and a zombie lineage subsequent to last appearance, are indicated. Also highlighted is an unsampled ancestral ghost taxon. B, The relationships between the taxa are used by the phylogenetic method to recover some portion of the ghost lineages, indicated by the dashed lines. In contrast the zombie lineages of terminal taxa A, B and D are unable to be recovered. Ghost taxa are created to fill in the unsampled ancestral lineages of the crown groups. This reconstructed phylogeny represents an estimation of the real phylogeny shown in A.

The terms artificial range truncation (Signor and Lipps 1982) and Signor-Lipps range (Wagner 2000b) have been suggested. Here we define a new term, zombie lineage, as that unsampled portion of a taxon's range occurring after the final appearance of the taxon in the fossil record prior to its actual extinction, from the notion of a ''zombie'' representing the living dead—in this case the extension of a taxon range past its last apparent living appearance. The extent of a taxon's zombie lineage can be inferred at some level of probability by using the methods of Strauss and Sadler (1989) and Marshall (1990) but it cannot be inferred by phylogeny. If we assume that there is no tendency for taxa to be preferentially sampled at either the early or late end of their ranges, it must be assumed that the amount of zombie lineage in a sampled phylogeny is likely to be equal to the amount of ghost lineage. Therefore, the addition of ghost lineages potentially produces a bias toward the early part of time ranges, and hence a skew backward in time in diversity counts. Wagner (2000b) compared diversity counts produced by the phylogenetic method as applied to both the true tree and to the most parsimonious tree of a sampled clade. He did indeed find a heightening of diversity counts at the start of a clade's history, but he did not specifically compare results from the phylogenetic method with true diversity levels as obtained from the unsampled clade.

A second potential problem of the phylogenetic method is the assumption that ancestral taxa are rarely or never found in the fossil record; indeed, the creation of a ghost taxon to connect a sister group to its next nearest relative is in effect creating a ghost ancestor, a taxon represented by an internal segment of a cladogram (Norell 1993). However, the cladistic method does not state that ancestors are never found, only that they are non-diagnosable; i.e., that they can only be defined by a lack of characters. It may be the case that many ancestral taxa are included in cladistic analyses but mistakenly defined as terminal taxa, especially if character loss is involved in bifurcation. Using empirically derived models of species origination, extinction and preservation, Foote (1996b) predicted that 1–10% of marine invertebrates in the fossil record are directly ancestral to other known fossil species. Many population biology models imply that species properties encouraging speciation (e.g., wide geographic range and numerous

populations within a species) also encourage preservation in the fossil record (Wagner and Erwin 1995). This is supported by molecular studies (e.g., Omland 1997) that show that geographically widespread taxa tend to be paraphyletic compared with more restricted taxa. Hence there is evidence to suggest that taxa with many ancestral lineages also have an increased preservation potential. The unnecessary addition of ghost taxa in situations where sampled ancestors are misdiagnosed may seriously overinflate estimates of diversity.

Finally, a major drawback of the phylogenetic method is that it assumes a ''true'' cladogram for the group is available and that any errors in the inferred taxon relationships are minor when estimating diversity. In reality several well-supported cladograms are often available for any particular group, and the use of different trees will produce very different reconstructions of biodiversity (Wagner 2000b).

Two of the above criticisms of the phylogenetic method are here investigated: the first is the issue of potential bias in correcting only the first appearance times of taxa, and the second is of the possible error associated with ancestral lineages being included in the analysis. The investigations were conducted by using a computer simulation of phylogeny growth and subsequent sampling.

GHOSTRANGE Computer Simulation

The unavoidable problem when trying to assess the usefulness of the various methods of enhancing diversity counts is that of not knowing the actual complete diversity count. One answer is to produce an artificial phylogeny by using computer programs coded with algorithms to simulate the origination and extinction of taxa. An evolutionary tree can be ''grown'' by such a program, its final topology dependent upon initial parameters input by the user; i.e., origination and extinction rates, and other options such as the simulation of mass extinction events and diversity equilibrium levels. Such phylogeny modeling has been instrumental in answering questions concerning the randomness of clade shape (Gould et al. 1977), and speciation (Bookstein 1987), and the inclusion of paraphyletic taxa in diversity counts (Sepkoski and Kendrick 1993; Robeck et al. 2000) among others.

A computer-generated phylogeny has a perfectly known diversity history, against which any incomplete diversity estimates can be compared. It is true that such simulations produce idealized and greatly simplified pictures of evolutionary history, in which many of the complexities of taxon origination and extinction are either disregarded or averaged together. However, the advantage of computer models is that they allow us to run repeated experiments in which we can examine the effects of particular parameters while controlling the effects of others. They provide a framework in which to test evolutionary hypotheses, and diversity-summing methods, in a manner not possible in the real world.

A new computer program, GHOSTRANGE, has been designed to test the use of the phylogenetic method of enhancing diversity counts.

Phylogeny Generation.—GHOSTRANGE simulates the growth of an evolutionary tree, starting from one initial taxon at time step 1, diversifying to a total of x number of taxa by time step n, both small (about 100 taxa) and large (up to 1000 taxa) phylogenies can be generated. The program follows convention in using a time step interval representing 1 million years, which should encompass the time necessary for speciation, local adaptation, and biogeographic expansion (Sepkoski and Kendrick 1993). The simulated taxa can be seen as analogous to genera or families, the usual focus of diversity studies. When the program samples the taxonomic ranges (see below) it simulates the discovery in the fossil record of one or more member species of the higher taxa.

Tree growth rate is controlled by origination and extinction rates input to the program combined with random number generation. If an origination event occurs, it does so by bifurcation: i.e., one taxon splits to become two descendant taxa, the ancestor going extinct in the process. This mechanism of speciation is by no means universally accepted among evolutionary biologists and the alternative budding model has been used in other computer

simulations (e.g., Sepkoski and Kendrick 1993; Robeck et al. 2000; Wagner 2000b), but the bifurcation model is used here because of the nature of the phylogenetic method being assessed.

The phylogeny can grow either exponentially or logistically, and there is also an option to simulate mass extinction events. For a logistic pattern the origination rate is diversity dependent, decreasing as the diversity level approaches the equilibrium level. Once this equilibrium is reached, the origination rate is constantly adjusted as the diversity level changes, in order to maintain stasis.

The equation for calculating the diversity-dependent origination rate is

$$r_o = k_o - D\,((k_o - k_e)/D_{eq}) \qquad (1)$$

where

r_o = diversity-dependent per-taxon rate of origination
k_o = initial per-taxon rate of origination
k_e = initial per-taxon rate of extinction
D = standing diversity
D_{eq} = equilibrium diversity.

This equation is a combination of Sepkoski's models for diversity-dependent origination and extinction rates, and the equilibrium diversity constant (Sepkoski 1978: p. 231, eq. 6,7, p. 232, eq. 10).

Mass extinctions are simulated by GHOSTRANGE using the procedures of Sepkoski and Kendrick (1993) and Robeck et al. (2000). At intervals through the program run (arbitrarily set to every 20 time steps) the background extinction rate is increased to 0.9 for one time step only. Extinction rate is then returned to its original level, and diversity once more increases exponentially, or logistically to equilibrium, depending on the diversification model in use. In this way the magnitudes of the simulated mass extinctions vary; comparison of the ''real'' and estimated diversity counts tests how well the phylogenetic and taxic approaches depict these events.

Phylogeny Sampling.—The phylogeny is randomly sampled according to a sampling rate input by the user. This gives each taxon a chance to be ''found'' in any one time step, and simulates the discovery of one or more member species of the taxon in the fossil record. A taxon is assumed to be present in all time steps between the first and last sampled occurrences. An option is also available to simulate the ''Pull of the Recent'' (Raup 1979). This assumes that all lineages surviving to the final time interval are extant and will always be found; i.e., a perfect sample. In reality the probability of all extant taxa in a clade being found in the Recent varies greatly from group to group, and as such should only be interpreted as an end-member state. Excepting the final time interval under the Pull of the Recent option, this simulation assumes a constant sampling probability throughout the entire time period. Although this does not reflect the true nature of preservation rates, it does allow for the testing of each diversity-summing method under ''ideal'' homogeneous sampling conditions, without the masking of any skew by rate variation. It is important that the absence of any such bias is established before any claim is made as to the use of diversity enhancement methods to correct for sampling heterogeneity.

Insertion of Ghost Lineages.—GHOSTRANGE reconstructs the relationships between the sampled taxa to form a new phylogeny based on the incomplete data. The gaps in the ranges and relationships of the sampled taxa are filled up with ghost lineages and taxa according to Norell's (1992, 1993) method. The mode of dealing with ancestral taxa when using the phylogenetic method is a difficult issue. Norell (1993) assumes that no ancestors are present in his theoretical examples, and he saw the addition of ghost taxa as the phylogenetic recovery of ancestral taxa that have not been discovered in the fossil record. Smith (1994), however, acknowledged that ancestral ''metataxa'' may be discovered in the fossil record and included in cladograms and evolutionary trees, and suggests that ghost range extensions should be taken from the first appearance of the descendant group only as far as the last appearance of the putative ancestor. This is the method used in Wagner's (2000b) phylogeny reconstruction models.

These conflicting viewpoints are incorporated into the GHOSTRANGE program by using three options for dealing with ancestral

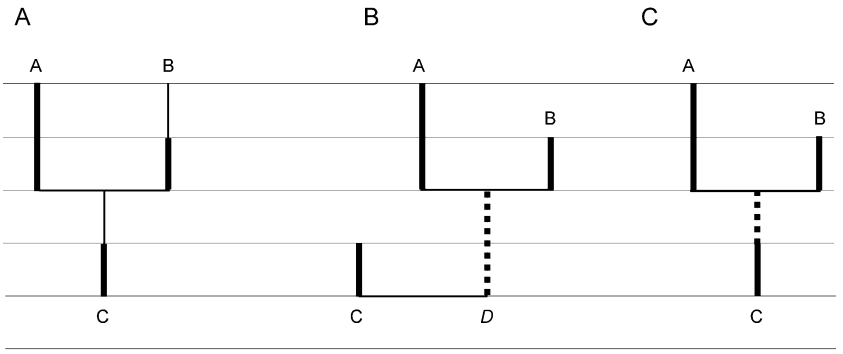

FIGURE 2. Dealing with ancestral taxa. A, A simple phylogeny consisting of crown group AB and ancestral taxon C. Bold lines show an example of the sampled ranges produced by the GHOSTRANGE program to simulate the sampling of the fossil record. B, A simulated recovery of the unsampled range where ancestors are mistaken as the sister taxon of their descendant groups. The GHOSTRANGE program adds a ghost taxon D (dashed line) down to the first appearance of C. C, A simulated recovery of the unsampled range where ancestors are correctly diagnosed. The GHOSTRANGE program adds a ghost lineage down to the last appearance of C. In this way the zombie lineage of ancestral taxon C is recovered by the phylogenetic method. However, the zombie lineage of terminal taxon B remains un-recovered.

taxa. The first assumes that ancestors are never recovered from the fossil record; the program does not allow them to be sampled, and they are not included in estimated diversity counts or ghost lineage insertion.

The second, perhaps more plausible, option allows ancestors to be sampled with as much likelihood as terminal taxa. However, when it comes to ghost range insertion an ancestral species is misdiagnosed as the sister species of its descendants. GHOSTRANGE inserts a ghost taxon between the first occurrence of a sampled sister group and the first appearance of its most recent sampled ancestor (Fig. 2B).

Finally, a third option assumes that ancestral taxa are found and correctly diagnosed as such, and ghost lineages are therefore only inserted from the first appearance of a descendant group down to the last appearance of their nearest sampled ancestor (Fig. 2C). This final method allows the phylogenetic estimate to recover the zombie as well as the ghost lineages of ancestral taxa, and it does not overestimate diversity in earlier intervals, as does the previous method (Fig. 2B). However, the manner of dealing with ancestors does not change the phylogenetic method's inability to recover the zombie lineages of terminal taxa (Fig. 2B,C).

The process of adding ghost lineages and ghost taxa continues until the sampled phylogeny is completely reconstructed. At this point there will no longer be any time range gaps separating taxa and taxonomic groups.

Diversity Estimates.—Three diversity counts are produced. The first is the actual taxon count per time step compiled from the original unsampled phylogeny; the remaining two are the diversity estimates based on the two methods under test. The taxic method simply sums the number of sampled taxa in each time step; the phylogenetic method also includes ghost lineages. These two estimates are compared with the complete data to assess how well each has performed in capturing both the magnitude (percentage of) and the pattern of the real diversity count. Pattern comparisons are made using the squared product-moment correlation coefficient (r^2) and also the squared partial correlation on time removed. The simple r^2 value can give misleading results when applied to time series data such as diversity counts (see Connor 1986; Harvey and Pagel 1991) and so the second metric is used on time-detrended data. It is essentially the r^2 value between residuals calculated from linear regressions of the data on time, and has been used in previous computer simulations of diversity patterns (Sepkoski and Kendrick 1993; Robeck et al. 2000).

Hence the simulations produced by the program compare a diversity count made at a particular taxonomic level (e.g., generic or familial) with an estimate based on a phylogeny

TABLE 1. Summary of results. Data shown are mean values for all runs incorporating the stated parameter on either small or large phylogenies; *n* in each case is 128 (excepting the two ways of diagnosing ancestors where *n* is 64). MEx = mass extinctions, SI = sampling intensity, PR = ''Pull of the Recent.'' All r^2 values are significant at the 0.05 probability level.

	Parameter	Taxic estimate mean partial r^2	Phylogenetic estimate mean partial r^2
Small clades	Exponential diversification	0.50	0.55
	Logistic diversification	0.37	0.65
	MEx not included	0.47	0.65
	MEx included	0.38	0.54
	SI = 0.1	0.23	0.45
	SI = 0.5	0.63	0.74
	PR not included	0.44	0.48
	PR included	0.42	0.71
	Ancestors not included	0.42	0.65
	Ancestors included	0.52	Misdiagnosed 0.62
			Correctly diagnosed 0.56
Large clades	Exponential diversification	0.67	0.68
	Logistic diversification	0.59	0.83
	MEx not included	0.68	0.80
	MEx included	0.57	0.71
	SI = 0.1	0.40	0.63
	SI = 0.5	0.85	0.88
	PR not included	0.70	0.65
	PR included	0.56	0.86
	Ancestors not included	0.55	0.78
	Ancestors included	0.64	Misdiagnosed 0.69
			Correctly diagnosed 0.77

constructed at the same taxonomic level. In this way it is similar to empirical comparisons of estimated species richness with phylogenetic diversity derived from a species-level phylogeny (Smith 1988).

Parameters

The following parameters were used when conducting program runs:

Number of Taxa.—For ''small'' phylogeny generation a figure of 100 was input, and for ''large'' phylogenies this was increased to 500. This covers the range from average-size cladistic studies to very large investigations.

Initial Diversification Rates.—Per-taxon origination rate varied between 0.25 and 0.35, and extinction rates were kept steady at 0.05, values similar to those used in previous phylogeny simulations (e.g., Gould et al. 1977; Sepkoski and Kendrick 1993).

Sampling Rates.—Set at either 0.1 or 0.5. A probability of 0.5 finds per million years is very high for most fossil taxa; however, a high sampling rate was included to investigate subtle differences between the two estimates of diversity. A rate of 0.1 finds per million years is plausible for genera within a well-preserved clade.

All combinations of program options were run for each sampling rate, encompassing the three methods of dealing with ancestral taxa, and inclusion or not of Pull of the Recent. This produced a total of 512 program runs.

Results

Results of the analysis are shown in Table 1, divided into small and large clades.

Diversity Pattern Capture.—In the majority of simulations the phylogenetic estimate captured the pattern of real diversity better than the taxic estimate. The partial r^2 values for the phylogenetic estimate were considerably (greater than 5%) higher than those of the taxic estimate in 346 out of the 512 simulations (68%). An example is shown in Figure 3A. There are two exceptions to this general rule, where the taxic count is on a par with, or exceeds, the phylogenetic estimate. An exponen-

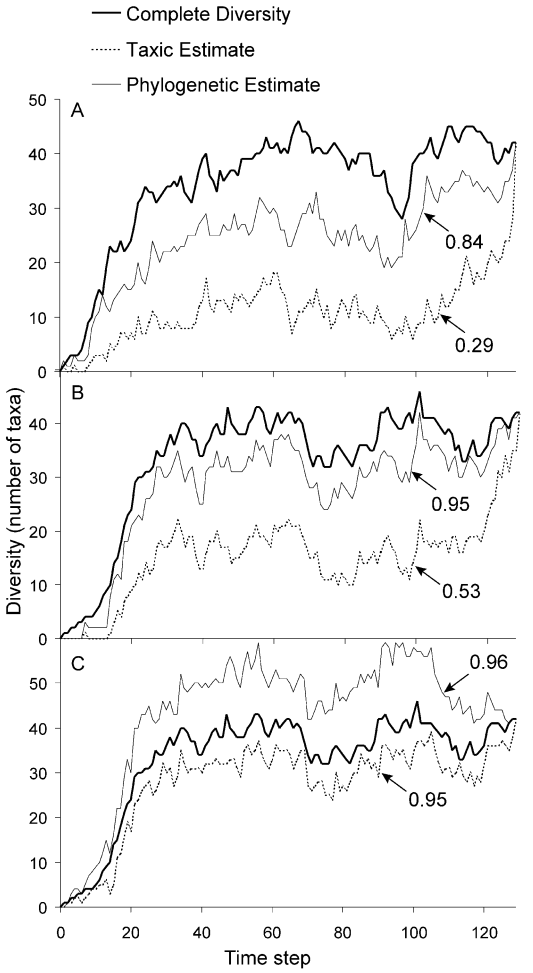

FIGURE 3. Examples of typical logistic diversification simulations. Figures given are partial r^2 values for the two diversity estimates. The ''Pull of the Recent'' is simulated in all examples. A, Sampling rate 0.1. Ancestors correctly diagnosed. The low sampling rate degrades the taxic estimate; the phylogenetic estimate is able to recover many of the lost taxa. B, Sampling rate 0.5. Ancestors not sampled. Again the unsampled ancestral taxa lower the performance of the taxic estimate, but they are recovered with the phylogenetic method. C, Sampling rate 0.5. Ancestors misdiagnosed as sister taxa. Both estimates capture diversity pattern accurately; however, the phylogenetic estimate consistently overestimates diversity magnitude owing to the needless addition of ghost lineages.

tial diversification pattern brings the taxic method up to equal the performance of the phylogenetic method, in contrast to logistic patterns where the latter invariably outperforms the former. The same is true for inclusion of the Pull of the Recent. If all taxa are permitted to be ''found'' in the final time interval of the program run, the phylogenetic estimate performs better than the taxic. However, if the Pull of the Recent is not simulated, the phylogenetic performance significantly drops, whereas the opposite is true of the taxic method, which increases to equal or outperforms it. We will return to this important point below.

Increasing sampling intensity obviously increases the performance of both methods of estimating diversity, though the increase is greater in the taxic estimate. Similarly, the analyses on larger clades tend to produce a better performance from both estimates than those on smaller clades.

Diversity Magnitude Capture.—The phylogenetic method significantly overestimates diversity magnitudes in the many of program runs. In 208 of the simulations (41%) the phylogenetic estimate has a mean maximum diversity magnitude that exceeds that of the real data. Stripping of ancestors can discard as many as 50% of a phylogeny's taxa, severely reducing the amount of diversity magnitude that the taxic method can capture (Fig. 3B). In contrast, when ancestors are included in the analysis but are misdiagnosed, needless addition of ghost lineages can increase the diversity magnitude estimates of the phylogenetic method by anything up to double the actual count (Fig. 3C).

Mass Extinctions.—The inclusion of mass extinction events within a clade's history does not effect the pattern capturing performance of the two estimates. However, what of the magnitude of mass extinction events? How well do the two estimates capture this important measure? Ten further phylogenies were created and analyzed to calculate the magnitude of their mass extinction events in terms of percentage loss of taxa during the event time step, in comparison with the standing diversity of the time step before. The mean results are shown in Table 2.

These results indicate that during the time interval of the mass extinction the phylogenetic estimate considerably dampens the event, exaggerating the Signor-Lipps effect (see later discussion), whereas the taxic count captures the magnitude of loss fairly faithfully although with more variation around the

TABLE 2. Magnitude of mass extinction events in the complete diversity curves and as captured by the two diversity estimates. Measurements given as mean and standard deviation of percentage taxonomic loss; $n = 62$.

	Actual diversity	Taxic estimate	Phylogenetic estimate
Mean % loss of taxa	46.4	44.8	25.6
Standard deviation % loss of taxa	17.5	30.4	30.1

mean than in the actual data. However, examination of the diversity curves containing mass extinction events reveals that the phylogenetic estimate does capture the magnitude of taxonomic loss, but prolonged over a greater number of time steps (Fig. 4A). The simple taxic count smears the mass extinction events to a certain degree, but the phylogenetic estimate exaggerates this, causing the diversity falloff prior to a mass extinction to start earlier in the time sequence. As an example, in Figure 4B the actual extinction event involves a 54% loss of diversity over one time step, whereas the phylogenetic estimate records a loss of 52% of taxa over nine time steps.

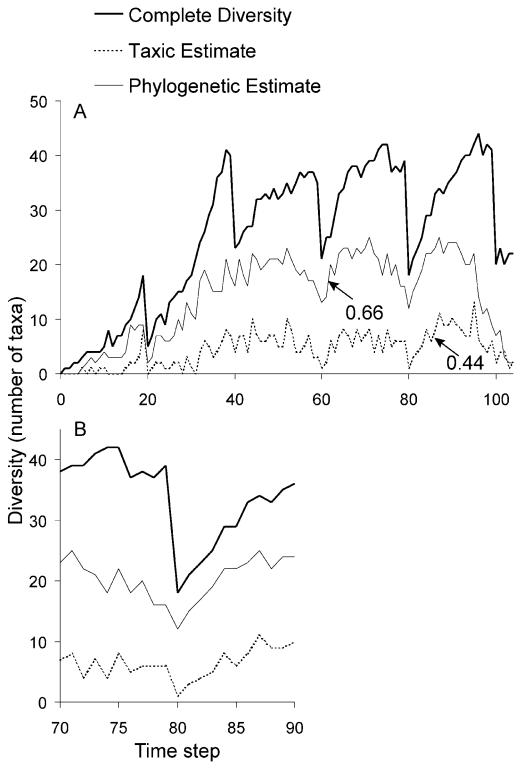

FIGURE 4. Logistic diversification simulations including mass extinctions. Figures given are partial r^2 values for the two diversity estimates. The ''Pull of the Recent'' is simulated in all examples. A, Sampling rate 0.1. Ancestors correctly diagnosed. Both estimates smear the extinction events backward in time, although the effect is more pronounced when using the phylogenetic method. B, Detail of the extinction event at time step 80. A 52% drop in taxa over one time step results in a 54% drop over nine time steps when using the phylogenetic estimate. The taxic estimate produces a 88% loss over five time steps; however, this is an unreliable estimate because of the low diversity level of the taxic count.

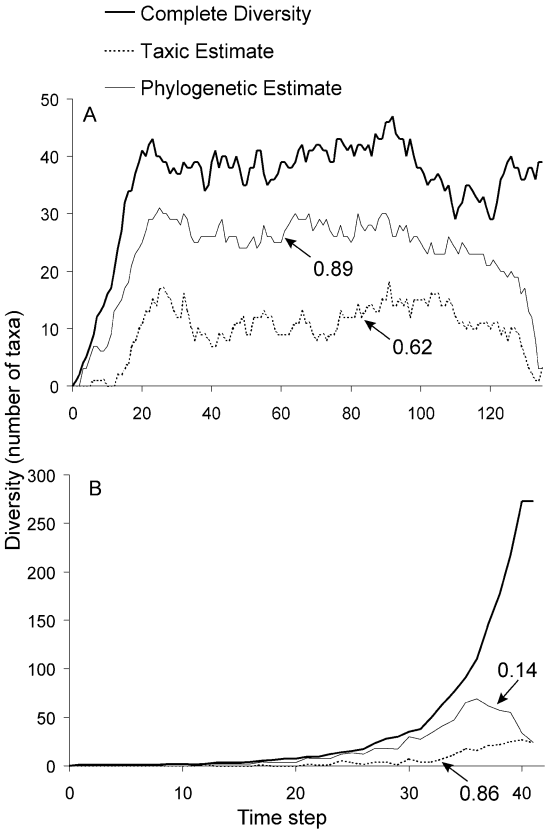

FIGURE 5. The Signor-Lipps effect at the end of diversity periods, seen if no ''Pull of the Recent'' is simulated. This is due to the coordinated burst of last appearances at the end of the program run and will also be apparent in diversity studies that do not take into account taxon appearances after the end of the study period. A, Sampling rate 0.1. Ancestors correctly diagnosed. The pronounced skew at the end of the phylogenetic estimate is not enough to lower its performance below that of the taxic estimate (B). However, when an exponential diversification pattern is simulated, the skew is enough to make the phylogenetic estimate inferior. The taxic estimate is high, despite the low sampling rate, as an exponential pattern is simple to capture.

Pull of the Recent.—The most striking ex-

amples of the taxic estimate outperforming the phylogenetic occur when a combination of an exponential diversification pattern and no Pull of the Recent are simulated. In 60% of such program runs the taxic count performs best. This difference in performance is heightened by a high sampling rate. Figure 5A illustrates an exponential diversification pattern sampled without Pull of the Recent.

The falloff in diversity seen in both estimates at the end of the curve is caused by the Signor-Lipps effect (Signor and Lipps 1982). This is the drop in numbers of recorded taxon ranges seen in the lead-up to a co-coordinated set of extinctions, such as a mass extinction event or in this case the end of the program run. In this respect the program is simulating studies that sum diversity over a limited time period, but where the sampled occurrences of taxa extending beyond the end of the period are not taken into account. Not accounting for these occurrences reduces range-through data and so artificially reduces the number of recorded ranges in the final time intervals of the study period. Conversely, this drop in diversity is not seen when 100% sampling in the final time interval is imposed, effectively eliminating zombie lineages from the crown taxa.

The dropoff in diversity levels toward the end of the time period is also evident in logistic examples (Fig. 5B) in both estimates, although again it is more pronounced in the phylogenetic estimation. This skew is not enough to alter the phylogenetic pattern to the same extent as in exponential diversification, and in most cases the corrected count performs better than the uncorrected count under conditions of logistic growth. However, the results indicate that the phylogenetic method exaggerates the Signor-Lipps sampling effect both at the termination of diversity curves and prior to mass extinction events.

Discussion

It may be true that in some instances the addition of ghost lineages to taxon counts captures more of a clade's diversity history, as asserted by Norell (1992, 1993) and Smith (1994). Although we show here that in the majority of the diversification scenarios simulated the phylogenetic method of estimating diversity is superior to the taxic, this is true only under certain conditions, for example, if the clade under analysis is extant. However, the expected backward skew in diversity counts predicted by the biased nature of only correcting the first appearance times of taxa (e.g., Wagner 2000b) is apparent in many other circumstances. These include time intervals leading up to an ''event horizon'' such as a mass extinction event, the termination of a clade, or the end of an analysis time period. In these situations the predicted bias in the phylogenetic method does have an impact, causing the exaggeration the Signor-Lipps effect. This exaggeration is only enough to reduce the performance of the phylogenetic estimate to below that of the taxic in situations of exponential diversification.

Why, then, if zombie lineages are occurring throughout the sampled phylogeny, is the predicted bias of not accounting for them apparent only under certain conditions and not throughout the clade's diversity history?

Diversity Skew.—The problem of not accounting for zombie lineages is relevant only in situations where there is an increased proportion of terminal taxa zombie lineages as compared with ghost lineage and sampled ranges. If the amount of ghost and zombie range is uniformly spread throughout the phylogeny, the addition of ghost lineages will raise the diversity count in each time step but will not skew the diversity pattern. This is the situation during periods of diversity stasis.

During times of increasing diversity, when origination rate is significantly higher than extinction rate, the majority of taxa are ancestral and there are few terminal taxa lineages. As Figure 2C shows, the phylogenetic method is able to recover the zombie lineages of ancestral taxa as well as the ghost lineages. Because the proportion of terminal taxa is low, the problem of not accounting for terminal taxon zombie lineages does not arise, and so the predicted diversity skew is not seen. A previous study of the use of the phylogenetic method (Wagner 2000b) reported heightened diversity counts in early time intervals of a clade's history when ghost lineages are included. However, this study only looked at diversity counts of a small number of taxa over a limited num-

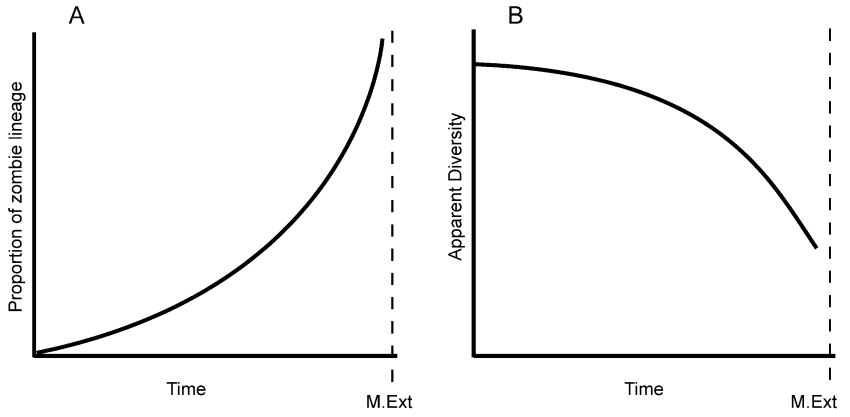

FIGURE 6. The Signor-Lipps effect. Any single taxon's chance of being resampled after its first appearance decreases with decreasing time to extinction. A, Prior to a coordinated burst of extinctions, e.g., a mass extinction event, the proportion of zombie lineage within a phylogeny increases relative to ghost lineage and sampled range. B, This has the effect of steadily lowering sampled diversity and hence causes a sudden mass extinction to look gradual. Adapted from Signor-Lipps 1982: Fig. 2.

ber of time steps. The actual bias is not toward heightened diversity levels in the early part of a clade's history, but rather toward depressed diversity levels at the terminal end.

Therefore, we should expect to encounter a diversity skew in the phylogenetic estimate only during times of an increased proportion of terminal taxon zombie lineages relative to ghost lineage and sampled range. Under what conditions will this occur?

The Exaggeration of the Signor-Lipps Effect.—From the time of a taxon's origination, the probability of any one time interval being recorded as part of its range (i.e., of becoming its first appearance in the fossil record) is equal to the sampling rate. However, after its first appearance, the probability of any individual time interval being included in the taxon's range becomes a function of the sampling rate and the number of time intervals remaining until extinction. This is because a taxon only needs to be resampled once in order to include all the preceding intervals back to the first appearance within the recorded range. The more time there is between first appearance and extinction, the more chance there is of such a resampling taking place. After a taxon's first appearance, the probability that any time interval (x) is included in the final recorded range is simply the inverse of the probability that the taxon will be missed in that interval and in all subsequent intervals, up to and including that of its extinction (t):

$$p_{range} = 1 - (1 - r_s)^{1+(t-x)} \quad (2)$$

where

p_{range} = probability of time interval being included in the final taxonomic range
r_s = sampling rate.

A logical outcome of this relationship is that as time to extinction decreases, so the probability of any individual time interval being recorded in the final range also decreases, and hence the probability of the interval being part of a zombie lineage increases. During most of a clade's diversification history these increasing probabilities are evenly distributed through time as taxa originate and go extinct randomly. However, at coordinated extinction times, such as mass extinctions, many taxa reach the ends of their life spans simultaneously, and the result is an overall increase in the proportion of unrecorded zombie lineages. This is the sampling artifact, first outlined by Signor and Lipps (1982), that can make a sudden mass extinction look prolonged (Fig. 6), and an important characteristic of this artifact is that the increasing proportion of zombie lineages belong to terminal taxa that the phylogenetic method cannot account for. These are the conditions under

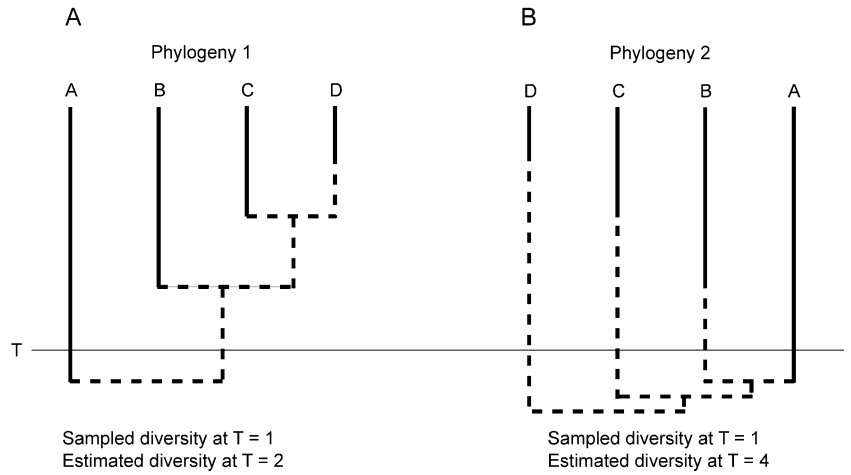

FIGURE 7. The use of different phylogenetic theories affects the diversity estimate produced by the phylogenetic method. A, In phylogeny 1, taxa C and D are sisters, with B and A progressively less closely related. Bold lines show sampled range, dotted lines show ghost range extensions. B, Phylogeny 2 gives a different view of the taxonomic relationships, with A and B as sisters. Hence different range extensions are required and an estimated diversity at time T is produced, which is double that of phylogeny 1.

which the phylogenetic estimate should be expected to distort diversity patterns.

The results of the computer simulations show that distortion does occur prior to mass extinctions (which includes the extinction event at the end of a clade's life span), and also prior to the end of diversity study periods if sampled occurrences beyond the final time step are not taken into consideration. This distortion of the diversity pattern takes the form of a magnification of the Signor-Lipps sampling effect, i.e., an artificial lowering of diversity counts.

Other Problems with the Phylogenetic Method.– The results of this analysis confirm that although the phylogenetic method of estimating diversity does improve diversity patterns in the majority of situations for clades of 100 to 500 taxa, caution needs to be applied in other circumstances, especially as many phylogenetic studies are performed on much smaller clades diversifying over few time intervals. It is likely that the diversity distortion of the phylogenetic method prior to clusters of extinctions as identified here will become of greater consequence for smaller-scale studies. At the opposite end of the spectrum, the phylogenetic method is impractical for large-scale investigations involving many different groups of organisms, such as the analysis of global diversification, as it can be used only on clades that have completely known phylogenetic relationships.

Finally, the results and conclusions of this analysis are based on a theory of speciation by bifurcation as is consistent with the phylogenetic method of estimating diversity (Norell 1992, 1993). Further research into the effect of alternative speciation theories (e.g., budding cladogenesis or anagenesis), and the necessary changes to the phylogenetic method, is required to ascertain if this technique is applicable to all evolutionary scenarios. The phylogenetic method is only as reliable as the cladogram it is based on. The relationships within many groups are in constant review, and not only might they change with each new discovery, but several conflicting trees may be published at any one time. The use of different cladograms will produce different diversity estimates when the phylogenetic method is used (Fig. 7).

Conclusions

These analyses present a cautionary tale concerning the use of the phylogenetic method in estimations of past diversity. Although phylogenetic estimate captures more of the real pattern of diversity than the taxic estimate in the majority of simulations, there are

numerous and important exceptions, especially in situations of exponential diversification where there is imperfect sampling in the final time interval. We note that the phylogenetic methods add diversity to the base of a lineage's diversification, in the form of ghost lineages and ghost taxa, representing diversity probably present but unpreserved or unrecognized. However, phylogenetic methods fail to account for potential real diversity at the other end of the diversification: that is, survival of lineages at the tips of clades beyond their final appearance in the fossil record, which we term here ''zombie lineages.'' Thus there is the potential for phylogenetic methods to bias the estimations of diversity in one direction (i.e., at the base of the clade), especially in conditions where the top of the clade is not delimited by a natural barrier such as an extinction event or the Recent time line.

Additionally, the phylogenetic approach overestimates the magnitude of diversity in 41% of simulations. This overestimation is significant and consistent in conditions where ancestors are sampled and misdiagnosed as sister taxa to their descendants, or where sampling rate is high.

Taxic and phylogenetic events also perform differently in conditions where there is imperfect sampling in the final time interval. The taxic estimate reduces diversity at the end of a clade's history owing to the Signor-Lipps sampling effect, whereas the phylogenetic estimate magnifies and prolongs this diversity falloff. For this reason the phylogenetic estimate will give a misleading pattern for clades that diversify over a small number of time steps. The phylogenetic estimate captures the magnitude of mass extinction events, but prolonged over a greater number of time steps. Hence it exaggerates the Signor-Lipps sampling effect seen before mass extinctions. If the phylogenetic estimate is used, caution should be taken when interpreting the duration of any mass extinction event.

The addition of ghost lineages to diversity counts will enhance the estimated diversity patterns in cases where a clade diversifies logistically and is considered to be both poorly sampled and lacking any sampled ancestral lineages. If it is suspected that ancestral lineages are sampled, any ghost lineage extensions of the descendants should be taken only to the last appearance of the suspected ancestor, not the first. If a clade is considered to be well sampled and includes many putative ancestral lineages, the taxic approach is adequate for estimating diversity patterns.

Acknowledgments

The notion of needing to consider zombie lineages as well as ghost lineages had its inception a decade ago in collaborative discussions between Christine Janis and Jack Sepkoski, amid considerable scheming and hilarity. (Christine regrets being dissuaded from subtitling this contribution ''zombie lineages and ghost authors.'') The GHOSTRANGE simulation is based on an original computer program written by Jack Sepkoski before his death in 1999, and the research contained herein is an extension of the conclusions and ideas he gained from his results. We would like to thank K. Harcourt-Brown for discussion during the conception of this project, M. J. Benton for critical suggestions on drafts of the paper, and P. Wagner for meticulous comments on near-final versions. We also thank the University of Bristol for providing funding via the Benjamin Meaker visiting Professor Fellowship for Christine Janis to visit Bristol in the spring of 2001, during which time Abby Lane took on this project.

Literature Cited

Alroy, J. 2000. New methods for quantifying macroevolutionary patterns and processes. Paleobiology 26:707–733.

Alroy, J. (and 25 others). 2001. Effects of sampling standardization on estimates of Phanerozoic marine diversification. Proceedings of the National Academy of Sciences USA 98: 6261–6266.

Anderson, J., H. Anderson, P. Fatti, and H. Sichel. 1996. The Triassic Explosion(?): a statistical model for extrapolating biodiversity based on the terrestrial Molteno Formation. Paleobiology 22:318–328.

Bookstein, F. L. 1987. Random walk and the existence of evolutionary rates. Paleobiology 13:446–464.

Connor, E. F. 1986. Time series analysis in the fossil record. Pp. 119–147 *in* D. M. Raup and D. Jablonski, eds. Patterns and processes in the history of life. Springer, Berlin.

Doyle, J. A., and M. J. Donoghue. 1993. Phylogenies and angiosperm diversification. Paleobiology 19:141–167.

Eldredge, N., and S. J. Gould. 1972. Punctuated equilibria: an alternative to phyletic gradualism. Pp. 82–115 *in* T. J. M. Schopf, ed. Models in paleobiology. Freeman Cooper, San Francisco.

Foote, M. 1996a. Perspective: evolutionary patterns in the fossil record. Evolution 50:1–11.

———. 1996b. On the probability of ancestors in the fossil record. Paleobiology 22:141–151.

Gould, S. J., D. M. Raup, J. J. Sepkoski Jr., T. J. M. Schopf, and D. S. Simberloff. 1977. The shape of evolution: a comparison of real and random clades. Paleobiology 3:23–40.

Harvey, P. H., and M. D. Pagel. 1991. The comparative method in evolutionary biology. Oxford University Press, Oxford.

Hennig, W. 1965. Phylogenetic systematics. University of Illinois Press, Urbana.

Levinton, J. 1988. Genetics, paleontology, and macroevolution. Cambridge University Press, Cambridge.

Marshall, C. R. 1990. Confidence intervals on stratigraphic ranges. Paleobiology 16:1–10.

———. 1991. Estimation of taxonomic ranges from the fossil record. *In* N. L. Gilinsky and P. W. Signor, eds. Analytical paleobiology. Short Courses in Paleontology 4:19–38. Paleontological Society, Knoxville, Tenn.

Mayr, E. 1963. Animal species and evolution. Harvard University Press, Cambridge.

Miller, A. I., and M. Foote. 1996. Calibrating the Ordovician radiation of marine life: implications for Phanerozoic diversity trends. Paleobiology 22:304–309.

Norell, M. A. 1992. Taxic origin and temporal diversity: the effect of phylogeny. Pp. 89–118 *in* M. J. Novacek and Q. D. Wheeler, eds. Extinction and phylogeny. Columbia University Press, New York.

———. 1993. Tree-based approaches to understanding history: comments on ranks, rules, and the quality of the fossil record. American Journal of Science 293-A:407–417.

Norell, M. A., and M. J. Novacek. 1992a. Congruence between superpositional and phylogenetic patterns: comparing cladistic patterns with fossil evidence. Cladistics 8:319–337.

———. 1992b. The fossil record and evolution: comparing cladistic and paleontologic evidence for vertebrate history. Science 255:1690–1693.

Novacek, M. J., and M. A. Norell. 1982. Fossils, phylogeny, and taxonomic rates of evolution. Systematic Zoology 31:266–275.

Omland, K. E. 1997. Examining two standard assumptions of ancestral state reconstruction: repeated loss of dichromatism in dabbling ducks (Anatini). Evolution 51:1636–1646.

Raup, D. M. 1972. Taxonomic diversity during the Phanerozoic. Science 231:177:1065–1071.

———. 1975. Taxonomic diversity estimation using rarefaction. Paleobiology 1:333–342.

———. 1979. Biases in the fossil record of species and genera. Bulletin of the Carnegie Museum of Natural History 13:85–91.

Robeck, H. E., C. C. Maley, and M. J. Donoghue. 2000. Taxonomy and temporal diversity patterns. Paleobiology 26:171–187.

Sepkoski, J. J. Jr. 1978. A kinetic model of Phanerozoic taxonomic diversity. I. Analysis of marine orders. Paleobiology 4:223–251.

———. 1984. A kinetic model of Phanerozoic taxonomic diversity. III. Post-Paleozoic families and mass extinctions. Paleobiology 10:246–267.

Sepkoski, J. J., Jr., and D. C. Kendrick. 1993. Numerical experiments with model monophyletic and paraphyletic taxa. Paleobiology 19:168–184.

Signor, P. W., and J. H. Lipps. 1982. Sampling bias, gradual extinction patterns and catastrophes in the fossil record. Geological Society of America Special Paper 190:291–296.

Smith, A. B. 1988. Patterns of diversification and extinction in early Palaeozoic echinoderms. Palaeontology 31:799–828.

———. 1994. Systematics and the fossil record. Blackwell Science, Oxford.

Strauss, D., and P. M. Sadler. 1989. Confidence intervals and Bayesian probability estimates for ends of local taxon ranges. Mathematical Geology 21:411–427.

Valentine, J. W. 1969. Patterns of taxonomic and ecological structure of the shelf benthos during Phanerozoic times. Palaeontology 12:684–709.

Vrba, E. S. 1993. Turnover-pulses, the Red Queen, and related topics. American Journal of Science 293-A:418–452.

Wagner, P. J. 1995. Diversification among early Paleozoic gastropods—contrasting taxonomic and phylogenetic descriptions. Paleobiology 21:410–439.

———. 2000a. Phylogenetic analyses and the fossil record: tests and inferences, hypotheses and models. *In* D. H. Erwin and S. L. Wing, eds. Deep time: *Paleobiology*'s perspective. Paleobiology Memoir 26(Suppl. to No. 4):341–371.

———. 2000b. The quality of the fossil record and the accuracy of phylogenetic inferences about sampling and diversity. Systematic Biology 49:65–86.

Wagner, P. J., and D. H. Erwin. 1995. Phylogenetic tests of speciation hypotheses. Pp. 87–122 *in* D. H. Erwin and R. L. Anstey, eds. New approaches for studying speciation in the fossil record. Columbia University Press, New York.