Supporting Online Material for

Phanerozoic Trends in the Global Diversity of Marine Invertebrates

John Alroy,* Martin Aberhan, David J. Bottjer, Michael Foote, Franz T. Fürsich, Peter J. Harries, Austin J. W. Hendy, Steven M. Holland, Linda C. Ivany, Wolfgang Kiessling, Matthew A. Kosnik, Charles R. Marshall, Alistair J. McGowan, Arnold I. Miller, Thomas D. Olszewski, Mark E. Patzkowsky, Shanan E. Peters, Loïc Villier, Peter J. Wagner, Nicole Bonuso, Philip S. Borkow, Benjamin Brenneis, Matthew E. Clapham, Leigh M. Fall, Chad A. Ferguson, Victoria L. Hanson, Andrew Z. Krug, Karen M. Layou, Erin H. Leckey, Sabine Nürnberg, Catherine M. Powers, Jocelyn A. Sessa, Carl Simpson, Adam Tomašových, Christy C. Visaggi

*To whom correspondence should be addressed. E-mail: alroy@nceas.ucsb.edu

Published 4 July 2008, *Science* **321**, 97 (2008)
DOI: 10.1126/science.1156963

This PDF file includes:

Materials and Methods
SOM Text
Figs. S1 to S21
References

Phanerozoic trends in the diversity of marine invertebrates

Supporting Online Material

Materials and methods

Data

The structure of the Paleobiology Database has been discussed previously (*S1*). In brief, geographic, stratigraphic, environmental, and taphonomic data are recorded for fossil collections that typically represent small geographic areas and narrow stratigraphic units. Presences of individual taxa in each collection (i.e., occurrences) are tracked, as are counts of individual specimens when these are available. Although species names are recorded, analyses in this paper focus on genera because that allows specifically indeterminate occurrences to be included and ensures comparability with earlier analyses.

As in an earlier study (*S1*), we have binned the collection data into a series of 49 roughly uniform time intervals, averaging 11.1 m.y. in duration. The first bin of the Cambrian, which includes the Nemakit-Daldynian (equivalent to the Manykaian) and excludes the traditional Early Cambrian, could not be included in the analyses because it was impossible to document a large number of collections and occurrences from this rather depauperate interval. The other bins are: (1) traditional Early Cambrian; (2) Middle Cambrian; (3) Late Cambrian; (4) Tremadoc; (5) Arenig; (6) Llanvirn and "Llandeilo"; (7) Caradoc; (8) Ashgill; (9) Llandovery; (10) Wenlock, Ludlow, and Pridoli; (11) Lochkovian and Pragian; (12) Emsian; (13) Eifelian and Givetian; (14) Frasnian; (15) Famennian; (16) Tournaisian; (17) early Visean; (18) late Visean and Serpukhovian; (19) Bashkirian and Moscovian; (20) Kasimovian and Gzehlian; (21) Asselian and Sakmarian; (22) Artinskian and Kungurian; (23) Roadian, Wordian, and Capitanian; (24) Wuchiapingian and Changhsingian; (25) Early Triassic; (26) Anisian and Ladinian; (27) Carnian; (28) Norian and Rhaetian; (29) Hettangian and Sinemurian; (30) Pliensbachian; (31) Toarcian and Aalenian; (32) Bajocian and Bathonian; (33) Callovian, Oxfordian, and Kimmeridgian; (34) Tithonian; (35) Berriasian and Valanginian; (36) Hauterivian and Barremian; (37) Aptian; (38) Albian; (39) Cenomanian; (40) Turonian, Coniacian, and Santonian; (41) Campanian; (42) Maastrichtian; (43) Paleocene; (44) Ypresian (Early Eocene) and Lutetian (early Middle Eocene); (45) Bartonian (late Middle Eocene) and Priabonian (Late Eocene); (46) Oligocene; (47) Early and Middle Miocene; and (48) Late Miocene, Pliocene, and Pleistocene.

An earlier study (*S1*) focused on genera also represented in Sepkoski's data

(*S2*) that belonged to the five core taxonomic groups that dominate the marine record (Anthozoa, Brachiopoda, Echinodermata, Mollusca, and Trilobita), but this restriction was only only due to perceived shortcomings in coverage of other groups that are now much less serious. The five groups constitute 71.2% of the genera in our copy of Sepkoski's invertebrate data set and 79.8% of the genera in our own data set, a difference that may be exaggerated if the additional groups are rarer and harder to sample.

We exclude tetrapods from our data set because their relative overstudy and typically very large relative body size makes them not comparable in terms of sampling methods to the molluscs, brachiopods, trilobites, and other groups that dominate our data set. In any case, tetrapods are a minor group in the marine realm, with tetrapod genera constituting only 1.7% of the Sepkoski compendium (*S2*).

Restricting the data to known metazoans less tetrapods means that only occurrences of genera with explicitly recorded taxonomic assignments were included. Thus, generically indeterminate, unnamed, or misspelled taxa were automatically excluded. The other sifting criteria were as follows. (1) For reasons explained below, most analyses excluded collections known to come from poorly lithified sediments that were sieved or from entirely unlithified sediments. (2) Collections from non-marine settings were excluded. (3) Collections of the coarsest spatial scale (basin) or stratigraphic scale (geological group) were excluded. (4) Multiple occurrences of the same genus in a single collection were treated as a single occurrence. (5) Reidentifications of individual occurrences and blanket synonymies were both employed. (6) To ensure comparability with earlier analyses (e.g., *S2, S3*), subgenera were treated as separate genera. (7) Genera whose names were qualified with the term "aff." or with quotation marks were excluded. (8) Form taxa and ichnofossils were excluded, leaving only body fossils. All of the remaining default download options given on the Paleobiology Database web site were followed, which means that no further occurrences were excluded and there were no further modifications of the included data.

Only the fields needed for analyses were downloaded. These were order name, genus name, abundance value, abundance unit, reference number for the occurrence, primary reference number for the collection, paleolatitude, paleolongitude, and time interval (bin) name.

The data were downloaded on 25 May 2008, and consisted of 44,446 collections and 284,816 occurrences of 18,702 genera drawn from 5384 references. A preliminary analysis (*S1*) employed data downloaded on 14 December 2000. By restricting a download that uses the current criteria to entries recorded before that date, we obtain 8165 collections and 71,146 occurrences. Thus, the current data set includes 5.4 times as many collections and 4.0 times as many occurrences. A

total of 39,726 collections had precise enough age assignments to be assigned to the 11 m.y.-long bins. At least 382 collections and 2398 occurrences are found in each one.

The data stem from references published by 3338 different senior authors. Assuming conservatively that each senior author's career only spanned the dates between the first and last publications under that name, a total of 16,114 career years were invested in collecting and describing these fossils. Thus, an effort to resample the record from scratch (*S4*) would either be inadequate or unrealistic.

Total counts of literature references, fossil collections, and occurrences (Figs. S1A, B) show that our sampling is particularly good by any measure in the early Paleozoic and Cenozoic, the two intervals of the most interest. Otherwise, sampling is fairly uniform because our sampling strategy succeeded in eliminating significant gaps that were intentionally left open by an initial study (*S1*). The total number of genera sampled in each bin based on all of the data (Fig. S2) covaries with the collection and occurrence counts (Fig. S1). Standardization removes two very large diversity peaks that correspond to these sampling highs and also are present in Sepkoski's data (Fig. 4), which suggests again that his compendium's data are influenced by strong sampling biases.

The unsurprising fact that standardization does remove sampling bias can be shown by cross-correlating the various sampling measures (Figs. S1) with our standard diversity curve (Fig. 1) after logging and differencing all the data sets. The resulting Spearman rank-order correlation coefficients are insignificant when using counts of collections, occurrences, and specimens (ρ = 0.072, 0.111 and -0.048; p = 0.627, 0.457, and 0.749).

Calculation of extinction percentages

We take advantage of having bin-by-bin occurrence data instead of simple age ranges by using occurrences to compute more robust extinction statistics. To do so we count the number of taxa sampled in two consecutive bins (two timers or $2T_i$), the number of those taxa also sampled in the immediately following bin (three timers or $3T_i$), the number sampled before and after the focal bin i but not within it (part timers or PT_i), and the local sampling proportion for the following bin i+1: $3T_{i+1}/(3T_{i+1} + PT_{i+1}) = S_{i+1}$. The ratio 3T/2T is a measure of cohort survivorship, but is biased because some of the two timers that continue into the third bin are not resampled at that time. The S_{i+1} term is used to counteract this problem. The resulting extinction percentage equation is $1 - 3T_i/(2T_i S_{i+1})$. This measure sidesteps edge effects, the Signor-Lipps effect, and the Pull of the Recent because of its focus on just three consecutive bins at a time.

Wobble index of short-term volatility

In some of the following discussions we use a simple index to quantify short-term variation in particular diversity curves. It is the absolute value of $\log(N_i^2/[N_{i-1}\ N_{i+1}])$ where N is the number of genera in time interval i. The index rises when an interval has relatively high or low diversity compared to both neighboring intervals. In other words, it flags single-interval spikes in the curve. It reacts strongly to perturbations such as mass extinctions only if there is an up-down pattern: if three intervals have diversity of 100, 200, and 100 genera, the index is 1.39, but if the values are 200-200-100 it is 0.69, and if there is a downwards trend such as 300-200-100 it is only 0.29.

The median of the index values across the entire time series can be used to assess the curve. For example, we obtain a value of 0.169 for the sampling-standardized curve we discuss the most often (Fig. 1), but one of 0.363 for the raw sampled-in-bin curve without any sampling standardization (Fig. S2). This difference captures the fact that the raw data include at least three very high spikes that are not nearly as prominent after standardization, two in the early Paleozoic and one late in the Jurassic, as well as two very high points in the Neogene.

We apply some caution in interpreting this index because there are different reasons why a method might remove wobbles. First, true variation could be revealed by stripping away a bias that happens to have concealed certain spikes or dips while having less of an impact elsewhere. An example would be backwards smearing before a mass extinction (the Signor-Lipps effect: *S5*). However, our use of sampled-in-bin (SIB) counts instead of range-based counts avoids this particular problem, and we suspect that similar ones also would apply less or not at all to counting methods such as SIB that do not rely on information from neighboring bins.

Second, it could be that certain methods systematically dampen real variation: simply drawing a straight line between the first and last data points would leave none at all. This scenario is more plausible and might apply (for example) to the occurrences-weighted subsampling method, but is less likely to pertain to our calibrated weights method for reasons we detail below: (1) it accounts for empirical variation in evenness; (2) it makes no strong assumptions about the uniformity of fossil collection sizes across the time series; and (3) it generates statistically independent diversity estimates in each temporal bin.

Third, short-term variation might reflect any number of sampling biases that change rapidly through time. We discuss many examples in the supporting text. On balance, then, we tend to favor analyses yielding low wobble index values, but recognize that other criteria need to be considered simultaneously.

Reference quota

Standardizing the amount of data drawn may not entirely remove sampling effects because diversity curves may be influenced by the amount of data available for subsampling in the first place. The crux of the matter is that publications do not report a random draw of all available fossils. Instead, they systematically focus on new taxa and undersampled times, places, and environments. Thus, adding more literature to a database expands the potential sampling pool faster than adding truly random reports of fossils would. If this problem is real, then a draw of (say) 100 collections from a data set of 1000 references should sample a larger pool of taxa than a draw of 100 collections from 100 references.

We later present evidence that a reference count bias has a strong influence on small-scale variation in our data set, although it is not responsible for major findings such as a relatively small the Cretaceous-Cenozoic radiation. For now, we simply note the procedure we use to remove it: limiting the number of references consulted in any one interval during any one subsampling trial. We impose this standardization by drawing references with equal probability until a fixed number is reached. If the references do not include enough estimated specimens to meet our quota of 16,200, we continue drawing more references until it is reached. This step is taken immediately before random subsampling of the available collections. We employ a quota of 65 references because all of the intervals include that many, and because during most trials 16,200 specimens can be extracted in all but a few bins without using additional references.

We frequently use our focal curve (Fig. 1) as a benchmark for evaluating different treatments of the data. To ensure consistency, we therefore impose the same quota of 65 references in all of this supplement's analyses, except in cases we note where a lower quota or no quota has to be used for particular reasons.

Calibrated weights subsampling method

Sampling standardization methods effectively seek to draw a fixed number of specimens per bin, either directly or by using a proxy based on counts of fossil collections or genus occurrences. In the latter case, the usual algorithm is to draw entire collections at random until a quota of estimated specimens is reached (*S1*). Whenever possible, we have tallied actual counts of specimens during subsampling. Otherwise the counts were estimated using methods described in this section. The data set includes an estimated 3.46 million specimens, of which 1.41 million derived from 11,078 collections were tallied directly.

Unlike conventional diversity curve construction methods that ignore

occurrences, treat ranges as continuous, and extend ranges to the Recent whenever taxa are extant, subsampling methods pay no attention to whether genera are extant or extinct, always drawing from the entire pool of fossilized taxa. Most of these methods assume that the number of specimens scales to the number of occurrences per collection by following a power law, or log-log linear curve (*S6*). All previous studies assumed as well that the pattern does not vary through time. The slope of this line is called an occurrence weight (*S6*) and can be used to compute the widely used probability of interspecific encounter (PIE) measure of evenness (*S7*).

However, this simplified approach is unrealistic when dealing with extremely long periods of time such as the entire Phanerozoic. For example, there are multiple, independent studies suggesting that evenness has in fact varied substantially (*S8-11*). The idea of varying the slope of the log-log curve in different bins instead of keeping this slope fixed was suggested in the same paper that described the occurrences-squared method, which also was the first to explicitly relate occurrence counts to estimated specimen counts (*S6*).

Previously, sufficient data were not available in enough Phanerozoic bins to calibrate slopes separately (*S1*). Our intensive and targeted data collection effort has resulted in at least six collections that include 100 or more specimens in every single one of our 48 bins. We were able to calibrate the slope employing at least five ecologically distinct collections in each bin (see below). There are at least 32 large collections and 16 ecologically distinct collections in half of them. These figures would have been far higher if we had not applied such stringent data quality standards (see below).

A simple, nonparametric method of estimating the log:log slope involves assuming that the collection curve runs from the 1 specimen/1 genus point to some *n* specimens/*g* genera point. Collections larger than this specimen count cutoff *c* are rarefied (i.e., randomly subsampled without replacement: *S12*) down to that level, and the median genus count *g* is then determined. The slope or occurrence weight *w* is then simply log *c*/log *g*. The same assumptions of linearity and rooting at the 1:1 point underlie all the preceding literature (*S1, S6*). The only difference is fixing the slope with an actual data point.

Unfortunately, all of the methods assuming linear relationships are compromised by the fact that most real rarefaction lines bend visibly within the range of typical paleoecological data, i.e., even below a sampling level of about 200 specimens. The bending is illustrated here for two of the best-sampled bins (Fig. S3). Thus, any straight line method will overestimate the number of specimens when there are few genera, and underestimate the number when there are many.

We have resolved this problem with a variant of the power law weighting function that is just as simple and conservative, namely, a blended linear and

power law function that employs a single shape parameter x. This parameter is very similar to the log ratio w. The equation is $\log g = (x + (1 - x)/n) \log n$, so $x = (n \log g - \log n)/((n - 1) \log n) - 1$. Any unknown value of n is easy to find by first estimating it with the simple power law and then using the recursive equation $n = 2n - (x (n - 1) + 1) \log n / \log g$, which stabilizes in well under 100 iterations. Like the simple power law, the blended function starts at the 1:1 point, but unlike the power law its shape at any value of n close to 1 is nearly linear with a slope of unity. There is no particular biological basis for the shape of this function because it does not relate to a particular species abundance distribution, but that problem also holds for the power law and alternatives such as the hyperbolic Michaelis-Mention equation (*S13*), and in any case the blended function fits the data well (Fig. S3).

The governing parameters of both the simple power law and the blended linear-power law function have the key property of being equivalent to PIE. The probability of an interspecific encounter is just the chance of finding a second genus upon drawing a second specimen, i.e., $g - 1$ when $n = 2$ (*S7*). This value is $2^w - 1$ or $\exp(w \log 2)$ given a power law and $\exp((x + (1 - x)/2) \log 2) - 1$ given the blended function. Although the values of w and x tend to track each other very closely, the resulting PIE values may differ substantially. For example, if $w = 0.5$ based on $n = 100$ and $g = 10$, then $x = 0.495$, but the encounter probabilities respectively derived from w and x are 0.414 and 0.679. Because the blended function consistently fits rarefaction curves much better at the base of the line (Fig. S3) and this region is so firmly nested within the data, PIE values based on x are expected to be more accurate.

A variety of alternative curvilinear fitting methods could have been used, but the obvious ones such as the Michaelis-Menten function (*S13*) all asymptote too strongly on a maximum diversity level to be of use. This behavior causes major problems because (1) asympotes must be extrapolated far beyond the range of the data and the variance on a log scale increases dramatically at high sampling levels, so estimates are highly imprecise; (2) fitting methods tend to be positively biased, meaning they project more and more taxa in the species pool as sample sizes increase (*S13*); (3) because a fitted line describes a collection of average diversity, collections with legitimately high true diversity may have sampled genus counts higher than the asymptote; and (4) if genus counts are just below the asymptote, specimen estimates will be extremely high and sensitive to small amounts of error in the fit. Poor accuracy and poor precision are very problematic when collections are large, because this error will strongly influence grand totals of specimen counts across collections. The use of a power law at high sampling levels tends to minimize these problems by producing systematically conservative specimen count estimates. Conservative values capture some of the natural variation in alpha

diversity that leads to high genus counts often being found at moderate sample sizes.

We fitted the weight parameters separately for each temporal bin and took weighted moving averages (WMAs) across a five-bin running window to reduce noise (Fig. S4A). A traditional WMA uses integer weights that peak at the focal bin and decline by one point with each step away from it. Thus, in this case the weights were 1, 2, 3, 2, and 1 for each series of five bins. To lessen the influence of poorly calibrated bins, the weights were multiplied by the number of collections used to estimate each median. For example, if the number of collections was 25, 20, 15, 10, and 5, the overall weights were 25, 40, 45, 20, and 5, and the proportional contribution to the mean of the central bin was 19, 30, 33, 15, and 4%. A five-bin window was selected because longer windows would span more than the average geological period. To reduce the effect of skewness on the data, we took geometric means of the weights. This correction had a very minor effect because the weights are closely spaced on a linear scale.

Because the weights affect the sampling level and are averaged over five bins, they could in principle dampen variation in the overall diversity curve. For example, an actual major drop in both evenness and global diversity might be obscured by low specimen estimates after the drop if high evenness beforehand influences the weighting coefficient afterwards. This bias is local and likely to be small because of the 1/2/3/2/1 weighting scheme. For example, the contribution of Permian samples to the earliest Triassic coefficient is just 32%. Raising the bar on showing a local excursion in the global diversity curve is conservative because the claim that there is no change in diversity is a null hypothesis, and it is specifically conservative for this study because we still show much higher bin-to-bin variation than in Sepkoski's compendium (*S2*).

Choice of rarefaction level

The trend in PIE produced by the weight calibration (Fig. S4A) is consistent with other studies showing that evenness increased between the Cambrian and Ordovician (*S9*) and between the early Paleozoic and Cenozoic (*S8, S10, S11*). However, these data are only as good as the assumptions underlying the calibrated weights method. The most important is that the shape of the fitted line within the range of typical data is not strongly dependent upon sample size. We have addressed this matter by estimating the value of PIE at rarefaction levels of 25, 50, 75, 100, 150, and 200 specimens, which spans the range of routine paleoecological collections. We found that the WMA curves were consistently in good agreement regardless of the cutoff (Fig. S4B). As expected, the curves did drift upwards, but slowly: the PIE values averaged across the bins are 0.711, 0.693, 0.683, 0.678,

0.669, and 0.665. These values yield estimates of 11.9, 10.3, 9.5, 9.1, 8.4, and 8.1 genera for any collection of 100 specimens.

The cutoff has hardly any effect on the shape of the evenness curve in the Paleozoic and Cenozoic, despite its influence on the offset (Fig. S4B). It therefore has nothing to do with our conclusion that evenness climbed only moderately through the Phanerozoic, which is what drives the finding that Cenozoic global diversity was relatively much lower than previously thought (Fig. 2).

However, the 200-specimen cutoff curve disagrees with the others in the Jurassic and early to middle Cretaceous. We suspect that the difference is unrelated to bending in the rarefaction curves, but instead reflects the small number of Mesozoic collections in our data set with at least 200 specimens. Typically, there are half as many useable collections per bin at this level as with a 100-specimen cutoff. This problem is demonstrated by the existence of at least two large outliers within the Cretaceous (Fig. S4A). The 100-specimen cutoff not only removes some of the noise, but also allows us to include samples that are of the typical size seen in paleoecological studies, and close to the estimated average throughout most of our time series, as discusssed later.

Selection of rarefied collections

Use of high-quality data is essential to the calibration procedure, so we applied a variety of filtering criteria. Collections were excluded if more than 95% of the specimens belonged to the most common species, which is necessary to avoid having mass mortality assemblages and samples from high-stress environments dominate the estimates. However, they were still included if the specimen count was less than 20, because in those cases it was mathematically impossible for two or more species to be present and still have such high dominance, making the hypothesis of > 95% dominance untestable. Collections also were excluded if less than 80% of the genera had abundance data of some kind, or if the publication only provided the names of selected genera.

Rarefactions excluded counts of fragments or of isolated elements such as crinoid stem ossicles, ophiuroid vertebrae, conodont apparatus elements, shark teeth, or fish scales. The marine fossil record is dominated by groups such as molluscs and anthozoans whose shells and colonies are not taphonomically comparable to these small isolated elements, because one individual may contribute dozens or hundreds of the latter. Furthermore, small isolated elements frequently are assigned to form genera that each equate to numerous whole-body genera. Percentage, grid-count, and quadrat count data also were excluded, leaving counts of distinct specimens and individuals. Such data comprise 97% of the non-percentage occurrences.

A final problem is the presence of pseudoreplicated collections from the same environments or geographic regions, or with similar taphonomic and collecting regimes. The hallmark of such collections is a similar genus list, most quickly, simply, and reliably identified by the identity of the most common (dominant) genus. We therefore avoided pseudoreplication by only employing the largest collection in each bin dominated by a given genus.

Collections excluded from the calibration analysis still were included in the subsampling analyses. If their abundances were counts of specimens and individuals, these figures were used for bookkeeping purposes during subsampling trials regardless of whether the collections themselves were used in the calibration. In all other cases specimen counts were estimated using the calibration weights method.

Inverse weighting method

Weighting by specimen counts is problematic because these counts are very unevenly distributed among collections, so a few very large collections can dominate the subsample in any bin. Because large collections represent a small geographic, environmental, and temporal window into the fossil record, they draw from a limited species pool. Furthermore, individual monographs tend to focus on one or a few major groups, so if the largest collections come from a few monographs, there may be systematic overrepresentation of those groups. Therefore, overall diversity estimates will be biased downwards whenever large collections happen to predominate.

In a perfect world, sampled specimens instead would be randomly distributed through space and environments. Collections would therefore be of uniform size. Although the actual collections without abundances cannot literally be forced to be the same size, the average number of specimens contributed by each collection across all subsampling trials can be made uniform. If, for example, two collections respectively have N and 100 N specimens, and if the probability of sampling the latter is 1/100th as great, on average they will contribute the same number of specimens across trials.

Based on this logic, we have weighted the sampling probability of each collection by the inverse of its estimated specimen total. To obtain the actual probabilities, the weights are summed and the sum is divided into the inverse value for each collection.

Inverse weighting is crucial to recovering any evidence of a Cretaceous - Cenozoic radiation. Not weighting but using exactly the same other methods results in a substantially different curve (Fig. S5), with diversity falling steeply in the latest Cretaceous (Maastrichtian) and not recovering until the early Neogene.

This striking result is no surprise, because these problem intervals include unusually large collections, as shown by the average number of specimens per collection implied by the calibrated occurrence weights (Fig. S6). Therefore, in such cases a small number of large collections drawn from only a few regions and environments can fill up a bin's quota easily, and as a result the broad species pool is not sampled.

As expected, the mean number of specimens actually drawn per collection is almost entirely governed by whether inverse weighting is used. Without this correction, the ratio is virtually the same as in the raw data (Fig. S6). With it, the variation is essentially removed (Fig. S6). Not surprisingly, without inverse weighting the greatly varying mean number of specimens correlates negatively (as predicted) with the resulting global diversity curve. The relationship is visible even after logging, differencing, and ranking the data, as is necessary for these sorts of time series (n = 47, Spearman's ρ = -0.304; p = 0.040). This clearly artifactual correlation is not present our usual curve (Fig. 1), which employs inverse weighting (n = 47, Spearman's ρ = 0.136, n. s.).

The fact that sampling must avoid extremely large collections to recover a realistic global diversity trend argues against the suggestion that global curves should be set aside in favor of a few extraordinarily large, new samples from individual field areas of several different ages (*S4*). Such an approach might say much about local or regional diversity, but likely would say little about global diversity.

Although the inverse weighting method is mathematically elegant and solves a major problem, it does not change the sampling outcomes for intervals that have barely enough collections to make the quota (Fig. S5). In such cases, all or almost all of the collections must be included in each subsampling trial, regardless of their size. Although full sampling per se is not a problem, if the distribution of collection sizes is highly uneven, there will be relative oversampling of some regions or environments. However, all of the bins have at least 15% more estimated specimens than the one setting the quota (the Late Permian), and all but five have at least 44% more. Furthermore, as shown later we obtain curves with the same shape in key intervals when we lower or raise the quota. Thus, we consider this effect not to be important.

Counting method and sampling probability correction

In addition to employing these more realistic and even-handed subsampling methods, we only counted extant and extinct genera that were actually sampled. Alternative methods treat genera as always having been sampled throughout their entire ranges. Ironically, these intuitively appealing methods create edge effects,

Signor-Lipps effects, and ultimately the Pull of the Recent (supporting online text). Our use of sampled-in-bin (SIB) counts therefore avoids major biases that call most published diversity curves into question.

We adjusted the sampled-in-bin (SIB) counts to control for short-term variation in the sampling pool that cannot be resolved simply by drawing a uniform amount of data. Such problems might result from overly focused sampling of specific geographic regions, paleoenvironments, or major taxa. The correction involves the 3T/(3T + PT) sampling completeness statistic used to compute the extinction rates.

The simplest correction would be to divide each bin's diversity value by the sampling proportion for that bin, which would favor bins that are still at a relative disadvantage after sampling standardization because of the sampling pool bias. However, doing so would amount to extrapolating diversity to estimate the sampling pool's size, and we are not really interested in extrapolation here. We simply want to know whether a particular bin has better or worse sampling than the average bin.

Therefore, after dividing the bins by their completeness estimates, we multiply the entire curve by a grand estimate for the entire data set. We do so not by averaging completeness values across all bins, because some bins have small sample sizes. Instead, we base the overall figure on grand totals of three and part timers found across all sets of three bins. This figure is computed separately in each subsampling analysis, and is 0.721 for the one used throughout most of the paper (Fig. 1). The correction has almost no effect for most bins (Fig. S7), and the only highly visible differences are removing three substantial gaps in the Cretaceous, where our sampling is poor in general (Figs. S1, S2), and moderating the latest Jurassic spike. Further details about the three timer data themselves are discussed in the supporting text.

The method does have one disadvantage that would be a problem for a short diversity curve. Because it requires information on adjacent intervals, a correction is not always possible because separate sampling probability estimates cannot be computed at the ends of time series. Thus, our curve (Fig. 1) uses the raw SIB count for the late Neogene bin. We also employ raw counts throughout the Cambrian because very high turnover rates throughout that period leave too few three timers and part timers to compute meaningful sampling proportions.

The three timer measure is different from the standard completeness proportion (*S14, S15*), which counts genera in the denominator if they are present at any time before and after a bin, not just immediately before and after. The old measure therefore includes long-ranging, poorly-sampled genera, which will be more numerous if neighboring bins are relatively well-sampled. Thus, it includes information about the quality of sampling in distant bins that is not relevant to

evaluating a particular bin. However, excluding rare genera inflates the estimates compared to the conventional proportions, so the new measure is only a relative indicator of completeness. The same is true (if less so) for the standard proportions, because they only consider genera that are common enough to be sampled at least twice in the first place. Because so many extinct taxa leave no fossils at all, the real and entirely unknown completeness proportions must always be lower than either measure, and perhaps considerably so.

The three timer correction could have a negative instead of positive impact, because it involves dividing the curve by a function of two different counts that each have random sampling error. Thus, it could decrease precision instead of increasing accuracy. The wobble index is 0.227 in the raw data and 0.169 after rescaling, and the decrease is highly significant ($p = 0.006$) according to a Wilcoxon signed rank test.

One way or another, the correction is not important for our results, because all of the key patterns still remain. Examples include the steep Jurassic and Early Cretaceous radiation and the limited difference between the early Paleozoic and Cenozoic. However, we focus throughout this paper on the corrected data because of the Wilcoxon test's straightforward implication that the correction reduces artifactual bin-to-bin variation.

Confidence interval method

Confidence intervals on diversity curves have previously been based either on rarefaction methods (e.g., *S1, 16*) or the binomial distribution's variance (e.g., *S17*). Here we use instead intervals based on Chernoff bounds (*S18*), which apparently have not been applied previously in an ecological or paleontological context. They express $\Pr(G > (1 + d)\mu)$ or $\Pr(G > (1 - d)\mu)$ where G is the observed count, d is a constant, and μ is the expected count. The former is $(\exp(d)/(1 + d)^{(1 + d)})^{\mu}$ and the latter $(\exp(d)/(1 - d)^{(1 - d)})^{\mu}$. The confidence limits are just the values of $(1 + d)\mu$ or $(1 - d)\mu$ rounded off to the nearest integer for which the probabillities are <= 0.025 or >= 0.975. They can be found by examining different values of d that are separated by small increments and span the range from zero to one or zero to negative one, as appropriate.

Confidence limits on rarefaction curves (*S12*) are problematic because they assume that specimens (or taxonomic occurrences) are drawn entirely from the set that has already been sampled. Thus, the confidence intervals start out at zero, expand, and then contract again to zero as the number of genera in a subsample approaches the total observed number. A plot of taxa against specimens showing the confidence intervals therefore resembles a banana, when one would expect a

fan shape instead.

Because of this effect, the nominal precision is zero for temporal intervals that barely attain the subsampling quota. Intuitively, the opposite should be true: the worst-sampled intervals should have very large confidence intervals at the quota. The reason for the disconnect is that the error limits only have to do with variation in subsamples from samples. Variation in samples of the entire sampling universe is the real matter of concern instead. Additionally, these expressions assume independent draws of specimens or occurrences with equal probability, and both assumptions are greatly violated when entire collections are drawn with weighted probabilities.

The variance of the binomial does pertain to draws from the entire sampling universe. It equals $Np(1 - p)$ where N is the number of genera that may be sampled, and p is the probability of sampling each one. N is unknown, but as it climbs p must fall in order to obtain the observed count G. Thus, the limit is the simplified expression Np, and the standard deviation is sqrt(Np), which is always a conservative estimate because sqrt(Np) > sqrt(Np[1-p]). In practice, N and p are unknown and the binomial distribution per se cannot be used. However, the expected value μ of Np is G, so sqrt(G) can be used to estimate the standard deviation around μ of a normal distribution, and the confidence intervals computed accordingly.

Unfortunately, computing confidence intervals using a normal distribution is categorically inappropriate for diversity data because the distribution (1) allows the mathematical possibility of obtaining a negative count; (2) is always symmetrical even when counts are small, and the true probability distribution therefore must be skewed; and (3) most importantly, assumes that all genera have equal sampling probabilities. Such an assumption is not only false but dangerously so, because observed abundance distributions are almost always dominated by a few common taxa.

Chernoff bounds avoid these problems because they are always conservative regardless of the pool size or variation in the sampling probabilities. In other words, they assume nothing whatsoever about the sampling universe, so they are robust to any departure from the ideal notion that the pool consists of a very large number of genera with equal chances of being found. The values are also never negative and always right-skewed.

Alternative subsampling methods

Alternative curves based on four established methods are shown in Fig. S8. Like calibrated weights subsampling, all of these methods succeed in the narrow sense of removing any signal of sampling intensity. Nonparametric correlations of

logged and differenced data sets fail to find any relationship between these curves and counts of collections, occurrences, or estimated specimens. However, these methods are special cases of the ones we employ, and make strict and demonstrably false assumptions about key parameters not changing through time. We discuss these methods nonetheless because they are widely used, confirm the small-scale features of our curves, and fail to show the substantial Cretaceous and Cenozoic trend seen with calibration of occurrence weights, which demonstrates the conservative nature of our conclusions and the importance of allowing these weights to vary based on rarefaction data.

The first three methods draw entire collections. Such methods are said to be "by-list" (or, by ecologists, "sample-based") because each collection is represented computationally by a list of taxa.

Occurrences-weighted by-list subsampling (OW: *S1, S6, S19*) draws collections until a quota of occurrences has been reached (Fig. S8A). It is mathematically a special case of calibrated weights because it assumes that the slope of the within-collection sampling curve is always 1.0. One way to make sense of this assumption is to suppose that if all other things are equal, evenness is always extraordinarily high, too high for the same genera to be sampled twice in average-sized collections. So, each specimen yields yet another genus.

However, this extreme interpretation is not necessary, because the same relationship would be seen if evenness were generally moderate, but evenness and sample size were positively correlated, i.e., large samples also tended to be highly even. So, for example, a collection with 300 specimens and 30 genera would imply a low but realistic log specimens:log occurrences slope of about 1.68. One with 50 specimens and 5 genera would imply a slope of 2.43, equating to much lower but still realistic evenness. In other words, if evenness and collection size really do covary, we can obtain such things as a 6:1 ratio of specimens between two collections that is matched by a 6:1 ratio of genera, justifying the notion that one occurrence buys one fixed number of specimens.

Thus, the OW method can be seen as striking a balance between assuming that collection size varies through time, and assuming that evenness does: the low assumed slope means that large collections make a higher but not very much higher contribution to the sampling quota, which is fair if they are diverse both because of biology (high evenness) and because of sampling (large size). It therefore partially accommodates greater evenness and therefore larger average occurrence counts by penalizing large collections only lightly.

The problem is that the method overpenalizes when both specimen counts and evenness increase, which appears to have been the case in the Cenozoic (Figs. S4, S6). As a result, the OW curve (Fig. S8A) has a Paleozoic peak of 697 genera, but a Cenozoic peak of only 751. It otherwise closely resembles the calibrated

weights curve (Fig. 1).

A fixed exponent of 1.4 has been advocated recently (O1.4W: *S20*) based on log-log slopes of curves produced by subsampling one pooled regional data set and one pooled global data set. These among-collection data were used to argue that subsampling methods make hidden assumptions about beta diversity (*S20*). However, among-collection patterns are methodologically irrelevant for our purposes, because the curves for individual collections used to tune specimen count estimation (Fig. S3) are not the same as the subsampling curves produced later by pooling all the collections and drawing from them (e.g., Fig. 3). Functions that describe within-collection patterns are never used by any of these methods to predict (much less constrain) the shape of the overall, among-collection diversity curve. Thus, no assumptions are made about beta diversity.

As it happens, the log specimens:log genera values in our data are always higher than 1.4, even in the Cenozoic. The median is actually 2.07. Indeed, the same paper arguing for the 1.4 exponent (*S20*) also showed with a third data set that within-collection rarefaction curves do typically have a log-log slope of about 2, within the range of most of our data.

In any case, the O1.4W method, like OW, is just another variant of calibrated weights with unrealistic assumptions of low and constant evenness (Fig. 2). Furthermore, the O1.4W curve (Fig. S8B) is generally similar to the OW curve and implies even less of a Cretaceous-Cenozoic radiation. There is little difference in short-term variability (nonparametric serial correlation $\rho = 0.721$ for OW and 0.703 for O1.4W), although the O1.4W curve seems to dampen out more variation (standard deviation of logged data 0.201 vs. 0.238). Empirically, then, there is no evidence that hidden assumptions would matter much even if there were any.

Occurrences-squared weighted, by-list subsampling (O2W: *S1, S6*) puts even more emphasis on list length by assuming an invariant slope of 2.0. The slope is so steep, and the penalty for sampling long lists therefore so severe, that the method effectively assumes there is no signal of evenness in the data. In other words, unlike OW, O2W makes no allowance for the possibility that collections in a given interval with many taxa may have more even abundances than other collections, not just more specimens. Nonetheless, the O2W curve (Fig. S8C) is still quite similar to the one produced by O1.4W, and similar on the fine scale to the OW curve, despite lacking its large excursions. Short-term variability is again about the same ($\rho = 0.674$), and overall variation is dampened further (s.d. 0.177).

We note that the similarity between the three methods would weaken if sampling probabilities were not inversely weighted by collection size, because this adjustment does have an important effect on the calibrated weights curve (Fig. S5). Methods with higher weighting coefficients by definition give more weight to large collections, so allowing those samples to be drawn frequently would cause such

methods to recover less beta diversity. Therefore, the O2W curve in particular would be too low in the Cenozoic without inverse weighting, a patttern seen as well with calibrated weights (Fig. S5).

Finally, unweighted by-list subsampling (UW: *S1, 19*) involves drawing a fixed number of lists per bin and ignoring the number of occurrences per list. This method has not been favored in recent studies (e.g., *S1, S6*). It can be seen as making any of three assumptions: specimens-vs.-occurrences curves are irrelevant because each collection has a fixed size; both things do vary, but unpredictably, because there is too much noise among rarefaction curves to generalize the relationship; or the rarefaction curve's shape changes through time, but the average number of specimens per collection does not, so weighting methods would fail but drawing a fixed number of collections would yield a fixed average number of specimens.

The first assumption is unrealistic because the size of samples with known specimen counts of course does vary greatly within time intervals. The second conflicts with the empirical data, because we see consistent patterns across real rarefaction curves placed in particular temporal bins (Fig. 3) and there is a clean temporal trend in the median shape of these curves (Fig. S4A). High variation presumably would obliterate such a signal. As for the last assumption, the average estimated number of specimens per collection actually varies by an order of magnitude through the Phanerozoic (Fig. S6), and the trend is not a methodological artifact, because it does not match what one would predict from the evenness data used to obtain the estimates (Fig. S4A). For example, the mean bottoms out during the Cretaceous, an interval of high evenness, and then skyrockets going into the Cenozoic, even though the change in evenness is small.

More importantly, even if average collection size were constant, it would be more sensible to reach the same goal of drawing a fixed number of specimens on average by explicitly tracking specimen counts, as we have done. UW is therefore just another variant of calibrated weights in which actual specimen counts are not monitored because sampling patterns are assumed to be exceptionally clean.

Because UW is a restrictive special case of calibrated weights, the two kinds of curves would match well if UW's assumptions were correct. In reality, they do not (Fig. 1 vs. Fig. S8D). UW suggests a pronounced and lengthy mid-Phanerozoic low ending with a sudden and very large increase between the mid-Cretaceous and Paleocene (Fig. S8D). Inevitably, the trend in average specimen counts (Fig. 6) mirrors the UW curve more closely than those produced by any of the other methods, suggesting that the UW curve's twofold decline throughout the Paleozoic and threefold Cretaceous-Cenozoic increase are fundamentally artifactual.

In sum, there is no biologically important differences between any of the

four curves except the last one. OW, O1.4W, and O2W depress the Cenozoic radiation because they use constant weights, even though O1.4W and especially O2W assume weights that closely match the observed data (Fig. S3). All three curves show about the same short-term volatility, with respective wobble indexes of 0.170, 0.155, 0.149, and 0.169 for OW, O1.4W, O2W, and our usual curve (Fig. 1). Meanwhile, although UW captures high Cenozoic diversity, it is still grossly inadequate because it ignores the large trend through time in average collection size (Fig. S6). It also has considerably larger spikes, with a wobble index of 0.200. Regardless of assumptions, all four curves agree on many details and fail to show anything like a three- or fourfold ratio of Cenozoic to early Paleozoic diversity. These results conflict with the suggestion (*S20*) that the choice of a weighting coefficient is a key issue.

We note that some studies (e.g., *S16*) do not draw entire collections, but instead use direct and independent draws of occurrences, which is effectively ecological rarefaction of occurrences. We do not present results based on this method because it gives the same expected diversity values as OW except at very low sampling levels, differing substantively just because it yields smaller confidence intervals as a result of assuming that the occurrences are statistically independent (*S1*).

Another caveat is that all methods in paleontology and ecology that standardize real or estimated specimen counts assume that population density does not vary between study intervals. If it does, diversity may be overestimated whenever densities actually are low, because more collections must be drawn to reach a given specimen quota. Densities could be low if body size is large on average, or the studied taxonomic group captures relatively little biomass. Overall populations of macroinvertebrates may be either high or low in the wake of mass extinctions, because body sizes may be small if the extinctions are size-selective, or biomass totals may be low if more energy is captured by micro-organisms than usual. Thus, the rate of recovery from extinctions in our analysis may be either depressed or exaggerated in (for example) the Early Triassic. It is not clear how one might correct for such an effect because population densities are very difficult to estimate with paleontological data, and in any case the issue is beyond the scope of the present study.

Supporting text

Despite the concordance between the new and old data (Fig. 4) and all of the statistical advantages of our approach, it could be argued that Sepkoski's data (*S2*) are simply better than ours, so his patterns should be preferred. Specifically, the argument could be made that Sepkoski's compilation has captured more genera (if not many more) and therefore more rare genera. It also might have more reliable individual age ranges if Sepkoski standardized taxonomy to a greater extent, strove harder to exclude spurious age range extensions, and sampled a wider variety of geographic regions in general, regardless of the tropics.

We consider the observed concordance and the statistical issues to be paramount, and therefore believe that these claims are not to the point. Complaints about data quantity and quality cut both ways, because our data do capture the main signals in Sepkoski's compendium (Fig. 4). Nonetheless, in this section we address not only overall concerns about our data, but additional methodological and taphonomic factors that support our rejection of Sepkoski's patterns at both large and fine scales.

First, we make further comparisons with Sepkoski's curve by using different counting methods to test for a Pull of the Recent effect (*S21*). At the large scale, we see a Pull of the Recent in our data regardless of whether we standardize for sampling. At the fine scale, we find that traditional counting methods smooth away interesting variation when applied to our standardized data.

Second, we discuss three additional explanations for the Cretaceous-Cenozoic radiation in traditional data: aragonite preservation, lithification, and geography. Cenozoic samples frequently preserve original aragonite and are often taken from unlithified or poorly lithified siliciclastic sediments, all of which presumably allows better preservation. Excluding either the few collections preserving original aragonite or the few collections from unlithified or sieved and poorly lithified sediments depresses the curve throughout the Cenozoic, most strongly in the Neogene. These results can be interpreted as showing preservational biases and not evolutionary signals. On the one hand, aragonitic taxa became diverse and abundant far earlier than the Neogene. On the other, unlithified samples are more often extratropical than are lithified samples, and they are more spatially concentrated. In principle, then, they should capture less diversity, not more. Thus, we would expect the opposite pattern to result if there was a true biological reason for the difference. For these reasons, all of the analyses reported elsewhere in this paper omit collections including original aragonite or coming from unlithified sediments or sieved, poorly lithified sediments.

Meanwhile, Cenozoic and early Paleozoic samples are much more broadly

distributed in space than other samples, which puts the rest of the Phanerozoic at a disadvantage with respect to capturing beta diversity. This apparent bias is not merely a function of raw sample size. Our imposition of a reference quota leaves our curve independent of spatial patterns, but if we had not used such a quota we would have seen a systematic relationship between sampled diversity and the spatial breadth of sampling.

We note that because so many biases have a strong effect on global diversity data, there is not much variation left in Sepkoski's Phanerozoic diversity curve to be explained by changes in the amount of preserved sedimentary rock (*S22-24*). If this reasoning is correct, then it is unlikely that diversity and rock quantity are largely controlled in parallel by a common factor such as sea level (*S24-26*).

Third, we show that our main diversity curve is insensitive to other manipulations of the data. We successively standardize patterns using much higher sampling quotas where this is possible, account for the effect of having inventoried more publications in some time intervals than others, exclude data not consistent with Sepkoski's compendium, lift our taxonomic vetting criteria, and lump the data into geographically and stratigraphically larger sampling units. None of these treatments has any substantial effect on our major biological conclusions.

Finally, we show that our data are robust in absolute and relative terms, regardless of how they are analyzed. Sampling of the low latitudes is strong, and age ranges are consistently well-sampled throughout the Phanerozoic. Also on this topic, we address the seeming discrepancy between our evidence of a strong Pull of the Recent and a recent analysis (*S3*) showing that Plio-Pleistocene bivalves are well-sampled, which we are not yet able to replicate and which is not necessarily apropos because of idiosyncrasies in that slice of the fossil record.

In sum, Sepkoski's three- or fourfold increase does not illustrate a real exponential radiation only visible in his data because of its supposed high quality, but instead results from unavoidable sampling and counting biases in his treatment of those data combined with biases in the fossil record such as preservation and geography. Because his raw data are ranges instead of occurrences, standardization is not possible, genera sampled in bins cannot be counted, and contextual factors cannot be addressed. Further arguments in favor of a large, seemingly exponential Cretaceous and Cenozoic increase would therefore have to rest on new fossil collections that sample the rock record very differently than the existing ones, because neither our data nor Sepkoski's legitimately document it.

The Pull of the Recent

Many of the differences between our results and earlier ones revolve around

the counting methods discussed previously. Previous Phanerozoic-scale studies (e.g., *S27-S33*) were required by the nature of simple first and last appearance data to assume continuous sampling of age ranges, instead of representing genera actually sampled within bins. Ranging data through introduces several related effects, such as an exaggerated initial climb due to censorship of the data (the edge effect) and smooth declines going into extinction events (the Signor-Lipps effect: *S5*).

The most important counting bias is the Pull of the Recent (*S21*), which is caused by the greatly superior sampling of the Recent relative to all parts of the fossil record. Indeed, the major departure between Sepkoski's genus level curve and the amount of available sedimentary rock starting in the Cenozoic (*S25*) suggests that the supposed exponential increase relates to more than just a rock area effect. Effectively, the Pull of the Recent is the reverse of an edge censorship effect, and a special case of sampling intensity biases that are removed by our subsampling protocol.

The Pull of the Recent problem is hard to quantify, but we consider it largely irrelevant to our analyses because our sampling and counting methods are statistically unbiased or nearly so, and do include extant genera when they are represented by actual fossil occurrences. Nonetheless, a simple analysis suggests that the bias may be quite large. In Fig. S9, we have counted each extant genus at each boundary it crosses only up until its last fossil appearance. Thus, extant and extinct taxa are treated identically. These boundary crosser (BC) counts are used to ensure comparability with Sepkoski's data (Fig. 4).

Truncating the ranges removes much of the Cenozoic rise: the ratio of the late Neogene to the median Ordovician through Devonian interval is now 2.16 instead of 3.74 (Fig. S9). The old and new BC curves also tilt upwards because turnover rates decline through the Phanerozoic, as discussed below.

The boundary-crosser data show the Pull of the Recent effect (*S21*) even more clearly with sampling-standardized data (Fig. S10A) than raw data (Fig. S9). The early Paleozoic plateau is removed, creating the impression that Cenozoic diversity was 4.80 times higher. The ratio of 3.1 in Sepkoski's data is actually far smaller.

As in Figs. S9 and S10A, preliminary analyses of our data using boundary crosser counts (*S1*) ended ranges of extant genera at last fossil appearances, and produced similar results. Meanwhile, Sepkoski himself attempted to deal with the Pull of the Recent problem by ending ranges at last fossil appearances (*S31*; see also *S33*). Furthermore, he counted not boundary-crossers, but all genera ranging through each bin save for those confined to it (i.e., range through genera minus singletons), thinking that this would remove sampling variance (*S31*). His indirect corrections yielded a Cenozoic increase over the early Paleozoic of just 70%,

similar to ours (Fig. 1), but the approach and the result were later put aside (e.g., *S32*, *S34*), or argued against because Sepkoski's tabulation of Plio-Pleistocene occurrences was probably incomplete (*S3*).

Unfortunately, terminating ranges at last fossil appearances is no solution because it trades the Pull of the Recent for an edge censorship effect, which should push down diversity in the last few bins. Thus, these truncation analyses may exaggerate the scale of the Pull of the Recent problem, and they only put an upper limit on it. However, SIB counts cannot suffer from censorship because they ignore bins in the past or the future. Surprisingly, the BC and SIB curves are very similar in the Cenozoic (Fig. S10B), and the former is visibly lower only in the last bin. The drop is also quite small in relative terms compared to the summed edge and Pull of the Recent effects (Fig. S10A). Therefore, any downwards exaggeration due to truncating ranges is likely to be strong only in the last bin, whereas the concern over upwards bias due to pulling ranges forward to the Recent is much more serious because it has to do with the entire late Cretaceous and Cenozoic.

In sum, all methods that assume sampling of genera everywhere within their ranges, including both traditional range through and BC counts, are biased by Signor-Lipps, edge, and Pull of the Recent effects, so they are unacceptable for global diversity studies. In our data, the largest of these biases appears to be the last one.

We note that a recent study (*S3*) found almost no Pull of the Recent with respect to Plio-Pleistocene bivalves, but the result is generally consistent with ours because that group and interval are exceptionally well sampled (see below).

Sampling biases in Sepkoski's data

Sampling bias apparently is not just relevant to comparisons of the early Paleozoic and Cenozoic. There also appears to be pronounced undersampling throughout large parts of the late Paleozoic and early Mesozoic, as suggested by three lines of evidence. First, we were unable to fill the long mid-Phanerozoic sampling low despite years of concentrated effort aimed at raising our overall quota (Figs. S1 and S2). Second, our raw BC curve shows lower relative diversity in the late Paleozoic and early Mesozoic than our sampling standardized BC curve (Fig. S9 vs. Fig. 10A). Finally, even though our data set is weakest in the mid-Phanerozoic (Figs. S1 and S2), we still match or even exceed Sepkoski's curve throughout this interval (Fig. 4). At the same time, we match his curve in the early Paleozoic only with a relatively much larger overall pool of data.

The clear implication that Sepkoski's curve is low in the mid-Phanerozoic because the overall literature is poor again supports the legitimacy of long-standing

concerns about sampling bias (*S1, S21-26, S35, S36*). Sepkoski's curve is not only too high at the end, but too low in the early Mesozoic. The combination of effects is biologically important because it again suggests that the Meso-Cenozoic diversification was constrained, instead of being an exponential radiation that picked up in the late Cretaceous and Cenozoic. Indeed, much of the post-Paleozoic radiation may already have been over by the mid-Jurassic (Fig. 1).

Ordinal level patterns

Our data also already capture Sepkoski's pattern at the ordinal level (Fig. S11) in addition to the genus level (Fig. 4). We are not able to replicate his curve with exactly the same methods for a technical reason: Sepkoski's genus-level data pull ranges to the Recent, but we can use those data only to infer ends of ordinal ranges that go up to the ends of the fossilized genus ranges these orders include. Thus, we are only able to infer that an order is extant if at least one of its extant genera happens to have a fossil record.

We can, however, bracket his curve by counting orders in two different ways (Fig. S11): first, by simply pulling all the ranges of our extant orders to the Recent, because we have marked all the extant orders in our somewhat differing taxonomy as such; and second, by only pulling their ranges up to the last fossil occurrences of the genera they include. One way or another, the magnitudes are always very similar, and the bracketing suggests that we have not overlooked a large fraction of Cenozoic orders. This convergence is an important indicator of our compilation's thoroughness, but not entirely unexpected because recording an order only requires sampling one of its genera.

Additional effects of counting methods

There are additional problems with methods that count taxa as present throughout their entire age ranges, not just the Pull of the Recent, edge effects, and the Signor-Lipps effect.

First, BC counts are skewed upwards by low turnover rates, because longer-lived genera are more likely to be sampled both before and after a boundary (*S1, S33*). It is well-known that turnover rates declined through the Phanerozoic (*S33, S37, S38*), and the BC curve indeed rises through this interval (Fig. S10B). Thus, the very steep Cenozoic increase seen in many earlier curves (e.g., *S2, S33*) may well be a joint artifact of two counting biases: a rate effect and the Pull of the Recent.

A similar but smaller artifact could decrease SIB curves as well: more taxa within a bin have short ranges when turnover rates are high, and therefore

individual sampling probabilities are lower on average. However, it can be shown by simulation that this bias is almost exactly cancelled out by the very fact that having more taxa appear within a bin inflates sampled diversity. As a result, SIB turns out to be an excellent estimator of average standing diversity at any one point within a bin despite nominally being a measure of cumulative diversity across an entire bin. This ameliorating property does not apply to boundary crosser counts precisely because they apply to a fixed pool of taxa existing at one moment in time, so they ignore the large number of taxa that appear above the base of a bin and therefore augment the sampling pool when turnover rates are high.

Secondly, although standardization should remove variance, both of the BC curves (Figs. S9 and S10A) show about as much volatility as our SIB curve (Fig. 1). These raw and standardized BC curves respectively have wobble indices of 0.166 and 0.144, in the latter case substantially lower than the value of 0.169 for the SIB curve. The lack of a difference results from ranging taxa across bins, which smooths out curves by effectively averaging diversity over a window. Unsampled genera are counted in a bin based on presences before and after it. These taxa are most likely to have been sampled in nearby intervals, because most genera have short ranges.

As a result of this smoothing, the raw genus-level curves based on both the Sepkoski and new data sets (Fig. 4) show high serial correlations (Spearman's ρ for values in each bin *b* vs. values in *b-1*: 0.844 and 0.893). After standardizing and switching to SIB counts (Figs. 1, S10B), the correlation is about the same (ρ = 0.902) even though the values were computed independently. Note that the raw time series are autocorrelated, and therefore the correlations can only be used to compute similarity coefficients, so *p*-values are not reported for them.

The BC and SIB curves (Fig. S10B) thus differ in more than just major trends. Specifically, all of the BC curves group together. First differences of logged counts in the Sepkoski BC curve are strongly cross-correlated with differences in both the raw and standardized Paleobiology Database BC curves, even after removing the Pull of the Recent from them (ρ = 0.823, 0.663; $p <$ 0.001). However, the midpoint values of the Sepkoski, raw, and standardized BC curves (thick lines in Figs. 4, S9, and S10A) are much less cross-correlated with the differenced SIB curve (ρ = 0.307, 0.404, 0.405; p = 0.041, 0.006, 0.006). Removing the Cretaceous and Cenozoic data barely strengthens the correlation between Sepkoski's curve and the SIB curve (ρ = 0.347, p = 0.052), so the less than overwhelming match is not just due to the Pull of the Recent.

This evidence demonstrates that fine-scale features are much more dependent on counting methods than on either data sets or sampling methods. Arguably, ranging methods smooth away or distort most of the biological signal at

all temporal scales, although the Sepkoski diversity curve's underlying turnover rates may still be informative. Meanwhile, the one most debated large-scale feature of Sepkoski's curve, the Cretaceous and Cenozoic climb, also seems to be sensitive to counting methods (i.e., pulling ranges forward to the Recent or not: Figs. S9, S10, and S11). Thus, Sepkoski's curve may be compromised at all scales by the limited counting methods available with that data set, although we believe that sampling is just as important.

The aragonitic preservation bias

A possible explanation for the high and rising level of diversity in the Cenozoic (Fig. 1) is a preservational bias favoring that interval. Two obvious features of Cenozoic collections are frequent preservation of fossils with original aragonite and reduced lithification of many source rocks.

The aragonite bias has been argued to have no strong quantitative effect on sampled diversity within collections (*S9-11, S39*). On the other hand, it has been suggested that the taphonomic loss of aragonitic molluscs was quite large in at least some Paleozoic and early Mesozoic environments (*S40, S41*).

Original aragonite is very rarely recorded as being present before the Cretaceous. Although most publications do not report preservational details, it is striking that the Cenozoic encompasses 12.6% of the binned collections in our standard data set (4984/39,708), but 49.3% (269/546) of our binned collections known to preserve aragonite. Before the Maastrichtian we have a significant with-aragonite data set (55 collections) only in our third (Carnian) Triassic bin. Most of those samples represent the exceptionally well-preserved biota of the Cassian Formation. All of these figures exclude collections from unlithified sediments and sieved, poorly lithified sediments.

Our data suggest that the increasingly frequent preservation of aragonite does have an impact on diversity. Although just 6.7% (192/2880) of the Neogene samples in our with-original aragonite data set actually are marked as preserving original aragonite, evenness is markedly higher when we include these collections: PIE rises from 0.666 to 0.718 in the early Neogene, and from 0.732 to 0.737 in the late Neogene. In the Carnian, PIE again rises substantially (0.627 vs. 0.656).

Given the increase in evenness, it is no surprise that an otherwise identically prepared diversity curve exceeds our standard one specifically in the Neogene (Fig. S12A). For example, our late Neogene count rises by 15.1% from 642 genera to 739.

We acknowledge but cast doubt on five arguments that we are mistaken about the frequency and impact of aragonite preservation. First, given the difficulty of scoring preservational information from many publications, it is

possible that this information happens to be commonly available only in the Neogene. This scenario seems not only coincidence-driven but implausible in light of the striking statistical disparities we see. For example, between the last Triassic bin (Norian-Rhaetian) and seventh Cretaceous bin (Campanian) we record just 70 collections preserving original aragonite, but 12,137 others.

Second, it is imaginable that few of the researchers entering pre-Neogene data have made an effort to score preservational data. However, for example, one of us (Hendy) has entered 5978 collections, and original aragonite is preserved in 464 of the 2851 Neogene collections (14.8%) but only 84 of the 3127 older collections (2.7%). Another (Fürsich) has entered 610 Jurassic collections that mostly represent personal field data. He has noted the preservation of original calcite in 411 of them, but that of original aragonite in none of them. Thus, this second ad hoc argument also seems poorly motivated. Importantly, both arguments only have to do with whether our categorizations are complete, so neither one explains why we would see a difference in evenness and diversity simply based on which collections we include.

Third, aragonitic preservation might for whatever reason be better in environments that also happen to have higher true diversity. A correlation with latitude would be of the most concern because we do see a steep latitudinal diversity gradient in the Neogene (Fig. 3). However, excluding collections that preserve aragonite has no effect on the proportion of Neogene collections having paleolatitudes below 30º. This figure is exactly 24.4% in the early Neogene with or without these data, and falls only from 40.4% to 37.9% in the late Neogene if they are excluded.

Fourth, we may have greatly decreased the available sampling pool by creating fewer opportunities to recover aragonitic taxa. At the extreme, if original aragonite had been preserved in 100% of Neogene collections with aragonitic taxa but never in older collections, we would have ended up with a sampling pool composed only of calcitic taxa. However, so few Neogene collections preserve original aragonite that the restriction hardly changes the proportion that includes aragonitic taxa: in the early Neogene, it is 83.1% (1119/1347) without the restriction and 82.3% (1045/1270) with it, whereas in the late Neogene it is 86.6% (1376/1589) without the restriction and 85.3% (1210/1418) with it. Thus, we have not forbidden sampling of particular aragonitic taxa but avoided sampling of particular well-preserved fossils.

Finally, it could be that aragonite preservation simply tracks the evolutionary radiation of aragonitic forms. However, as noted above aragonite preservation is only common in any sense during the Neogene, but the proportion of genera thought to have aragonitic shells rose steadily and steeply during the Mesozoic as aragonitic groups like neogastropods diversified (*S42*). This rise to dominance was

effectively over by the Cretaceous-Tertiary boundary, at which point the great majority of molluscan occurrences were already of aragonitic forms.

In summary, we fully acknowledge that more research needs to be done on this difficult issue, and that our case for the importance of an aragonite bias is not certain. Nonetheless, because we are not able to explain why an effect on evenness and global diversity would be restricted to our Neogene data in the absence of a bias, we consider it prudent at this time to exclude collections preserving original aragonite from our analyses. This manipulation should be harmless at worst because of the very fact that the large majority of our Cenozoic molluscan collections include aragonitic taxa, regardless of data restrictions.

The lithification bias

Lithification is demonstrably even more important. Collections from unlithified sediments are unknown before the Jurassic, and both unlithified and poorly lithified samples become common only in the Cenozoic. Reporting in this case is better. We have 22,891 collections of any known lithology that do not preserve original aragonite and can be placed in a bin, of which 19,627 (85.7%) are known to come from lithified sediments, 2365 (10.3%) from poorly lithified sediments, and another 899 (3.9%) from entirely unlithified sediments.

Mean genus counts for the 779 Cenozoic collections from unlithified sediments are at least 11.3 genera in four of six bins, but in our standard Cenozoic data set the mean is 8 - 10 genera per bin except in the late Neogene, where it is 11.4 (Fig. S12B). Ratios between the means for unlithified and other collections are 1.5, 2.2, and 2.9 for the Paleocene, early Eocene, and early Neogene bins.

A possible explanation for the difference is that unlithified samples include a much larger average number of specimens per collection, undoubtedly due to the ease of extracting specimens in general. Indeed, even if we exclude unlithified samples but include poorly lithified samples, this ratio is dramatically higher in the Cenozoic (Fig. S6).

If specimen counts per se are the only issue, then it should make no difference whether we include unlithified or sieved and poorly lithified collections in our diversity curve analyses. Calibrated weights subsampling controls for this factor by making sure that each collection contributes the same number of specimens on average, even in the Cenozoic (Fig. 6). As we show later, even extensive lumping of small collections does not defeat the method's ability to accurately estimate evenness and therefore specimen counts.

However, when we add back in this small number of unlithified and sieved, poorly lithified samples, the standardized curve rises visibly in the Cenozoic, making it seem that there is an important radiation (Fig. S12C). The two versions

are overwhelmingly cross-correlated regardless of whether the data are differenced (Spearman's $\rho = 0.990$ for raw data; $\rho = 0.967$, and $p < 0.001$ for logged and differenced data). However, median Paleozoic and late Neogene diversity are now off by 116% (820 vs. 380 genera) instead of 74% (642 vs. 369).

There are two viable explanations for the difference. First, for coincidental reasons there could be real biological differences between samples that happen to be lithified during diagenesis and samples that don't. Second, unlithified samples could be better preserved in addition to larger on average.

It is unclear why postdepositional factors should relate to ecology and biogeography, except that unlithified samples are almost always siliciclastic. However, carbonate samples should be more and not less diverse because they are more often tropical and more often represent reefal environments. Indeed, the median absolute paleolatitude of the unlithified late Neogene collections is 37.8º (n = 447), but the median in our overall data set is 26.9º (n = 1418). Putting lithology and latitude aside, it also is not true that unlithified samples in our particular data set are broadly distributed and therefore likely to pick up significant geographic beta diversity: 241 (53.9%) of the late Neogene unlithified collections are from the Gulf Coast and Atlantic Coastal Plain, between 23.5º - 50º N and 70 - 100º W, but only 176 (12.4%) of the other late Neogene collections are. Both of these strong differences suggest that excluding unlithified samples should visibly raise instead of lower the curve by enhancing the number of tropical samples and improving geographic coverage.

Because the biological factors work in the wrong direction, the sampling differences related to lithification (Fig. S12B, S12C) must be fundamentally preservational and therefore a bias. The most likely mechanism is that lithification makes some taxa difficult or impossible to preserve and/or collect. One reason might be differential taphonomic loss and degradation. A second mechanism may be the difficulty of observing and extracting fragile or very small specimens in lithified rocks. Sampled diversity and evenness within collections drops precipitously in unlithified bulk samples if a mesh size of about 10 mm or more is used (*S43*), and certain groups such as asterozoans are primarily described from elements that would only be sampled with much finer meshes than that. Although this factor does not relate to preservation biases per se, perhaps specimens smaller than 10 mm are harder to capture or identify by visual inspection of lithified samples.

The importance of the lithification effect has been independently confirmed by explicit rarefaction analyses of large individual collections (*S44*). In that case, unlithified samples recovered more taxa even when sample size was standardized. These results and ours make it clear that the higher diversity of unlithified samples is a bias and not a biologically meaningful factor.

Given these revelations, previous studies of evenness and richness at the collection level that compare the Paleozoic to the Cenozoic (*S8, S10, S11*) are likely to be compromised. A hypothesized tripling or quadrupling of sampled diversity within collections between the early Paleozoic and Cenozoic has been used to explain and justify the similar increase in Sepkoski's global diversity curve (*S11*), but that study did not account for the lithification bias. Furthermore, its limited raw data actually only demonstrate a within-collection increase of about 90%, our data imply an approximate 71% increase between the late Ordovician and late Neogene at the 200 specimen sampling level (Figs. 2, S3, S4), and a recent comprehensive analysis gave estimates that fall close to our figure (*S10*). The observed increase in evenness matches our global diversity curve (Fig. 1) and therefore is not nearly great enough to explain the steep trend in Sepkoski's data (*S2*).

Reference count bias

Strong cross-correlations between reference counts (Fig. S1) and different versions of our diversity curve explain why we have had to impose a reference quota. After logging and differencing, we find rank-order correlations that are significant at $p < 0.001$ for multiple treatments that do not use a reference count quota. These include a curve that is otherwise treated exactly the same way as our main one (Fig. S13A: $\rho = 0.631$), a curve produced without inverse weighting of sampling probabilities ($\rho = 0.332$, $p = 0.023$), and curves produced with inverse weighting and using the OW, O1.4W, O2W, and UW subsampling methods ($\rho =$ 0.460, 0.556, 0.635, and 0.426). Unsurprisingly, the greatest correlation of all ($\rho =$ 0.773) is produced by using unstandardized sampled-in-bin counts (Fig. S2).

Imposing a reference quota in our standard analysis (Fig. 1) has substantially weakened the cross-correlation with reference counts after logging and differencing ($\rho = 0.408$, $p = 0.005$ vs. $\rho = 0.631$, $p < 0.001$). In some ways differences are small: for example, imposing the reference quota does little to reduce brief excursions (wobble index 0.162 vs. 0.169). However, several other patterns suggest that there has been improvement. (1) While neither version cross-correlates with logged and differenced collection or estimated specimen counts (Fig. S1), the reference quota removes a nearly significant cross-correlation with differenced occurrence counts ($\rho = 0.111$, $p = 0.457$ vs. $\rho = 0.251$, $p = 0.88$). (2) We are only able to obtain a significant (albeit still weak) cross-correlation with changes in evenness (Fig. 2) after imposing this quota ($\rho = 0.332$, $p = 0.023$ vs. $\rho =$ 0.282, $p = 0.055$). We would expect such a pattern simply because we use abundance distributions to constrain our randomized subsampling. Thus, limiting

the data counter-intuitively strengthens an important biological signal whose absence would be a strong warning sign. (3) Relaxing the quota increases the curve everywhere, with the median rise being 13.6% (Fig. S13A), which again shows that changes in the extent of the universe being sampled can curtail the effect of sampling standardization.

The curve is still likely to be biased in some ways because inverse weighting has a smaller effect when fewer references are employed. For example, it reduces the median number of collections sampled in each bin from 645 to 407, so it sampled fewer environments and locations. Nonetheless, the curve suggests that even after standardization, the great abundance of available data in the early Paleozoic, Late Jurassic, and Cenozoic (Figs. S1, S2) would otherwise cause us to overestimate diversity in those parts of the curve (Fig. S13A).

The possibility that reference counts bear a biological signal is cause for concern. For example, researchers may systematically publish more about intervals with high true diversity. If so, then either beta diversity or local abundance patterns should correlate with reference counts. We do not yet have an independent beta diversity measure, but our evenness data (Fig. 2) do not correlate with reference counts after logging and differencing each variable, regardless of whether we use raw values ($\rho = 0.173$, $p = 0.245$) or running weighted averages ($\rho = 0.103$, $p = 0.491$). We note also that even if high true diversity does spur more intense study, this fact alone would not give cause for allowing reference counts or any other measure of sampling intensity to vary in an analysis.

Nonetheless, we would still have a problem if diverse intervals were represented by publications that detailed fewer specimens because a publishable unit of taxonomy could be based on a small but rich collection. As a result, it might be harder to hold the number of collections per interval relatively constant and independent of estimated diversity. However, changes in logged collection/reference, collection/reference, or specimen/reference ratios do not cross-correlate with changes in logged reference counts ($\rho = -0.180$, -0.086, and 0.042; $p = 0.226$, 0.564, and 0.779). In other words, references provide the same amount of data on average regardless of underlying biological diversity.

Regression-based adjustment for reference counts

Another way to get at the reference count problem is to perform a post hoc correction of the diversity curve (Fig. S13B). We start by regressing the logged and differenced genus and reference counts: $\log N_i - \log N_{i+1} \sim m(\log R - \log R_{i+1}) + b$. By definition, this relationship predicts the change in log N given any difference whatsoever between logged R values. We can therefore estimate the

difference between log N_i and the mean value of log N across the entire time series as m(log R_i - mean(log R)) + b. Subtracting the estimate from the real, observed difference log N_i - mean(log N) yields a residual that accounts for the reference count bias and does not change the mean of zero. Because we have done nothing more than correct observed departures from a constant, namely, mean(log N), adding the residual values to this observed mean gives us corrected estimates of log N_i per se. Algebraically, the equation reduces to exp(log N - m(log R + mean(log R)) + b.

As with the reference quota analysis (Fig. S13A), with this correction we find a less visible late Jurassic peak and much lower Neogene values (Fig. S13B). Indeed, the residual curve's changes are more highly correlated with changes in the reference quota curve ($\rho = 0.838$, $p < 0.001$) than with changes in the version that does not use a reference quota ($\rho = 0.728$, $p < 0.001$). The residual curve otherwise does differ in a few other details and is generally flatter, but these three points of similarity are quite important.

We infer that extraordinarily even Neogene coverage may have created all of the remaining upwards trend that we would see if we did not account for the reference count bias (Fig. S13). Without the early and late Neogene outliers there is no Cenozoic radiation (Fig. 1). Instead, diversity may have changed essentially not at all from the mid-Cretaceous on, as opposed to increasing a modest 42% since the Albian, 106 m.y. ago.

Putting these results aside, even the high Neogene values seen without reference count standardization (Fig. S13A) cannot be construed to demonstrate that a large Cenozoic radiation occurred. An increase is present, but far smaller than in Sepkoski's data (*S2;* Fig. 4), and the two standardized curves (Fig. S13A) are far more similar to each other than to the traditional one. Thus, our decision to limit the references used in our main analysis (Fig. 1) has not had a major impact.

Large-scale geographic bias in sampling

The fact that adding references expands the effective size of the sampling universe raises the question of what is being expanded. Do publications systematically target understudied taxonomic groups, environments, geographic regions, or something else?

A comparison of environments (or lithologies) would not be meaningful because the early Paleozoic rock record is dominated by carbonates, whereas the Cenozoic record is dominated by siliciclastics (*S45*). Likewise, a straightforward comparison of taxonomic group coverage is not possible because the dominant groups have changed dramatically. Furthermore, there is no a priori reason to

think that the sampling is uneven across any of these categories.

However, we are able to demonstrate such a strong link betweeen publication rates and geographic coverage that we believe this latter factor to be the key one governing the extent of our theoretical sampling universe. There is almost a 1:1 match of reference counts and counts of approximately equal area paleogeographic grid cells that have been sampled (Fig. S14A: the cells are 10º in height and in width from 10º at the equator to 120º next to the poles). The effect is so pronounced that the rank order correlation of the curves ($\rho = 0.844$) is about as strong as the correlation of the curves' logged differences ($\rho = 0.880$, $p < 0.001$).

High Neogene peaks in both data sets are of particular concern (Fig. S14A). The early and late Neogene data are respectively drawn from 276 and 312 grid cells. Indeed, we have 922 late Neogene collections from south of 20º N paleolatitude or east of 20º E paleolongitude, areas that exclude the United States, Canada, and Western Europe. These figures exceed most of the global totals for individual bins throughout the rest of the fossil record (Fig. S14A).

Effect of geographic bias on sampled diversity

Given the match between reference and cell counts (Fig. S14A) and the fact that reference counts bias diversity curves, it is not surprising to find that changes in geographic cell counts predict changes in the diversity curve that was generated without using a reference quota (Fig. S13A: Spearman's $\rho = 0.579$, $p < 0.001$). We also arrive at essentially the same diversity curve when we use the regression-based adjustment method but substitute cell counts for reference counts (Fig. S13B vs. Fig. S14B).

For example, all three corrections push down our first two Cambrian values, compress the late Jurassic outlier, and push the curve's two last points down so far that there a slight increase at best between the mid-Cretaceous and late Neogene instead of a large radiation. Indeed, the two adjusted-after-the-fact curves are very strongly correlated after logging and differencing ($\rho = 0.923$, $p < 0.001$), and there are substantially weaker correlations between either time series and the no-reference-quota curve's first differences (residual curve based on cell counts: $\rho = 0.811$, $p < 0.001$; residual curve based on reference counts: $\rho = 0.728$, $p < 0.001$).

Meanwhile, our reference quota curve (Fig. 1) shares all of these major patterns (vs. cell count residual curve: $\rho = 0.877$, $p < 0.001$; vs. reference count residual curve: $\rho = 0.838$, $p < 0.001$). Switching to this curve also removes much of the correlation between differenced logs of cell and genus counts ($\rho = 0.579$, $p < 0.001$ vs. $\rho = 0.375$, $p = 0.010$). We therefore believe that the reference quota has mostly removed the one most dramatic bias in our data related to the sampling

universe issue.

Nonetheless, it might be tempting to use a restriction algorithm that directly focuses on geography instead of merely limiting reference counts. We have not done so because we view the reference bias as objectively important, and we believe the reference quota method potentially solves a number of related problems in a simple and elegant manner. Furthermore, the alternatives are limited. One approach would be to sample from areas of a fixed size in each interval, ignoring collections that fall outside of the standardized area. However, doing so would unfairly penalize intervals with broad longitudinal dispersion of the continents, as opposed to the tight clustering seen during the mid-Phanerozoic existence of Pangea. The early Paleozoic would be affected the most, so our conclusions about the Cretaceous and Cenozoic would not be altered. An even more serious problem is that the longitudinal and/or latitudinal breadth of sampling is not the real issue: spatial aggregation per se is (Fig. S14). Restricting the area of sampling would not necessarily remove this bias.

Rock quantity

It has long been recognized that the availability of fossiliferous rock may indirectly control sampled diversity by influencing sample size (*S22, S23*), regardless of whether the data are spatially clumped per se. Recent studies (*S25, S26*) have gone beyond this argument to posit that variation in rock quantity also signifies fluctuations in sea level that directly modulate true marine diversity by changing the area of benthic habitat. In other words, sea level may be the "common cause" of variation in sample size and true diversity.

Nothing in this paper addresses the issue of habitat area, which may indeed explain certain biological patterns in our diversity curve, as well as patterns in turnover rates and genus durations (*S25, S26*). However, several lines of evidence suggest that sea level change does not strongly bias our particular data set.

(1) Apart from sample size, most of the factors demonstrated here to have a major effect on diversity curves do not relate to the extent and duration of sedimentary packages. These include edge effects, the Pull of the Recent, the aragonite bias, and the lithification bias.

(2) Any direct sample size effect has been removed by our standardization of estimated specimen counts, as measured by cross-correlations between changes in diversity and various measures of sample size we discuss elsewhere.

(3) Almost as importantly, we have removed most of the correlation between the geographic reach of sampling (Fig. S14) and diversity (Fig. 1) by imposing a reference count quota. We would expect to see a strong correlation if sea level controlled rock quantity in a way that controlled our cell counts.

(4) Because almost all of the variation in our cell counts is explained by their relationship to reference counts (Fig. S14A), there is little left that could be explained by sea level or other factors.

(5) The number of sampled geographic cells (Fig. S14A) only varies about half as much as our specimen counts (Fig. S1B), despite our success in documenting a large number of fossils in every single bin. Specifically, the respective standard deviations of the logged tallies are 0.410 and 0.727 If the production of fossils per unit area were on average constant through time, a doubling in area would produce a doubling in specimen counts, so the ratio between the two would be unity.

(6) Finally, there are direct reasons to doubt that the geographic extent of sampling per se controls our specimen counts. The clearest evidence is that after logging and differencing the cell and specimen counts, we find a rank-order correlation of 0.395, explaining only 15.6% of the variance. The connection may actually be indirect because changes in reference counts more strongly predict both changes in cell counts ($\rho = 0.880$, $p < 0.001$) and changes in specimen counts ($\rho = 0.479$, $p < 0.001$).

Because rock volume effects apparently do not explain the large amount of variation in our data set's sample sizes, we might ask what does. First, much of the variation was produced by the particular focuses of our data collection effort, such as our intentionally strong emphasis on the Neogene. Second, sampling relates strongly to intrinsic, socioeconomic factors that include proximity of researchers to outcrops of certain ages and researcher interest in selected time intervals or taxonomic groups (*S22, S25, S46, S47*). Third, sampling is controlled by other extrinsic, physical factors such as facies distribution, diagenesis, and differential preservation potential. One of the key physical factors may be the proportion of rock that is exposed at the surface, a quantity expected to rise with decreasing burial probabilities as one moves towards the Recent, and especially going into the Cenozoic. Some recent studies on the rock availability factor focus on exposed rock area (*S24*), but the common cause hypothesis is grounded on rock quantity estimates that do not distinguish between surface and subsurface sediment (*S25, S26*).

Quantity of data

Admittedly, Sepkoski's curve is somewhat higher than ours in many intervals when treated the same way (Fig. 4), and given that our data are still further subsampled, many other described but relatively rare genera have been omitted in our analyses. It is easy to show that increasing our sampling quotas to capture such genera would not affect our conclusions, because we can do exactly

that in selected intervals.

Subsampling quotas can be extended up to 80,000 specimens in the late Ordovician and Cenozoic, where individual sampling curves maintain a nearly constant ratio even at very high quotas (Fig. S15). For example, the genus counts for the late Neogene and latest Ordovician bins have a ratio of 1.95 at the 16,000 specimen level, 1.91 at the 32,000 specimen level, and 1.89 at the 96,000 specimen level.

Quotas can be doubled throughout most of the time series, including almost all of the Cenozoic and early Paleozoic (Fig. S16). Increasing the quota from 16,200 to 32,400 estimated specimens naturally does increase the magnitude of the curve, but there is very little change in the relative order of the points ($\rho = 0.953$) or of their differences after logging ($\rho = 0.824$, $p < 0.001$; median Ordovician - Devonian and late Neogene diversity 582 and 997 genera, 71% higher). So, methodological issues such as counting procedures (Figs. S9 - S11) and sampling algorithms (Fig. S8) are far more important than the size of the data set per se.

Our raw data already approach or exceed Sepkoski's throughout much of the mid-Paleozoic (Fig. 4), so there may not be enough published data left to be added to our compilation to even double the data set in that interval. Because we could easily double, triple, or even quadruple the sampling quota in the early Paleozoic and Cenozoic data (Figs. S2, S15), our job may be done with respect to the early Paleozoic-Cenozoic comparison.

In any case, the question of whether our database is big enough relative to Sepkoski's seemingly impressive sample of 4399 extant genera is moot, because even that data set is extremely incomplete in absolute terms. About 180,000 extant species of marine invertebrates had already been described a decade ago (*S48*), and there are 1514 extant genera and 6497 extant species just of molluscs on the western margin of the Atlantic (*S49*). The extant species total and 4.3:1 molluscan species:genus ratio together suggest about 42,000 invertebrate genera globally, many of which are soft-bodied or otherwise not easily fossilized. Thus, the difference between our curve and Sepkoski's (Fig. 4) is just the difference between a large sample of common, easily preserved genera and a somewhat larger one.

Quality of data

Taxonomic error in our data set may be a concern. Earlier studies have shown that it is present in both our data set and Sepkoski's (*S2*) but has no consistent effect on either of them (*S3, S50-53*). There are other reasons to believe our data are not particularly suspect, such as the fact that our restriction of the data to occurrences of formally classified genera removes most misspelled names. We also have updated occurrences with all synonomies known to us. However, our list

of synonyms is not yet comprehensive, and the data set may include many synonyms known to Sepkoski.

We addressed this issue with two radically different treatments of the data that alternatively brought them closer to and farther apart from Sepkoski's. We then performed calibrated weights subsampling, and in each case we recovered the same major trends, and strong cross-correlations between the baseline curve (Fig. 1) and the modified curves (Fig. S17, S18).

In the first treatment, we included genera only if they were also listed in Sepkoski's compendium, and furthermore excluded occurrences that fell out of the age ranges defined by Sepkoski (Fig. S17A). The curve is still strongly correlated with our usual one (Spearman's $\rho = 0.874$, $p < 0.001$ after logging and differencing; median early Paleozoic and late Neogene diversity 270 and 391 genera, 45% higher).

Differences in shape are small and inconsistent. The largest is that when scaled up, the compendium-vetted curve is relatively higher, not lower, in the early Paleozoic. This fact belies the argument that our finding of a small Cenozoic/early Paleozoic ratio relates to a quirk in our data.

On the other hand, the compendium-based curve is relatively lower in the Triassic. The Triassic gap is due to the fact that more than half of the Triassic genera in each of our 11 m.y. bins either do not range into them in the compendium, or are not listed anywhere by Sepkoski (Fig. S17B). It is not clear why spurious range extensions would increase diversity substantially in this interval and none other.

The gap is small but important because it creates the appearance of a weaker Triassic rebound followed by a steady Meso-Cenozoic radiation (Fig. 17A). Thus, some of the supposed exponential pattern may be due not just to the Pull of the Recent and overall trends in sampling intensity, but specifically to undersampling of the literature by Sepkoski (*S2*) in the Triassic. On the other hand, both Sepkoski's family level data (*S2*, *S29*) and data from the Fossil Record 2 compendium (*S30*) depict an even slower early Mesozoic rebound than do his genus level data (*S2*). Thus, it does remain possible that our data are distorted by exceptionally well-preserved and diverse parts of the record that Sepkoski did not inventory carefully, but on the other hand are not representative of the Triassic in general. A potential example is the mid-Triassic Cassian Formation of Italy.

Also of note is the consistently higher percentage overlap of genus counts with Sepkoski's in the Paleozoic than post-Paleozoic (Fig. S17B). It remains to be seen whether this prolonged drop implies systematic undersampling by Sepkoski or systematic inaccuracy in our data. However, the former hypothesis is buttressed by the fact that the better Paleozoic match is reflected in the Sepkoski-restricted data set yielding a relatively higher Paleozoic curve (Fig. 17A). As for a

mechanism, we note that Sepkoski presented a family level Phanerozoic diversity curve (*S29*) only after beginning his work in the Paleozoic (*S54*), and published repeatedly on Paleozoic but not Meso-Cenozoic faunal distributions (e.g., *S55*).

In the second treatment (Fig. S18), we intentionally degraded the data set by including all occurrences of any kind marked as belonging to the Database's marine invertebrate working group, thereby including not just marine microfossils and vertebrates, but even plants that happen to be preserved in marine deposits. We also reversed almost all our standard downloading options by counting multiple occurrences of the same genus in the same collection separately; not applying synonymies or reidentifications of individual occurrences; not treating subgenera as separate genera; and including "aff." genus names, taxonomically questioned genus names, informal non-Linnean names, form taxa, ichnofossils, and occurrences of higher order taxa indeterminate to the genus level, all treated as if they actually were valid genera. The data set could hardly be less taxonomically reliable.

Nonetheless, we found almost the same pattern (Spearman's $\rho = 0.920$ and $p < 0.001$ after logging and differencing; peak early Paleozoic and late Neogene diversity 466 and 692 genera, 48% higher). The wobble index is actually lower in this treatment (0.169 for the usual data vs. 0.132 in this treatment). These similarities combined with those seen in the preceding analysis suggest that data quality has no bearing on the general shape of the curve. Because degrading our data has almost no consistent effect, but forcing the data to be consistent with Sepkoski's compendium (*S2*) does, the gap we see in the Triassic (Fig. S17B) may have more to do with missing data in the compendium than spurious range extensions in the Paleobiology Database (for better or worse).

Geographic and stratigraphic scale

To make the spatiotemporal scales of the collections uniform, all of our analyses exclude collections representing entire geographic basins or stratigraphic groups. However, collections ranging from hand samples to local areas, and individual beds to entire formations, all are included. To see if reducing the variation in scale might matter, we lumped together small collections of the same geological age from the same stratigraphic formation and member, and from the same exact geographic coordinate (Fig. S19). The curves have nearly identical magnitudes even though the lumped data set draws from fewer collections (fine-scale data: geometric mean = 417 genera and 436 collections; large-scale data: geometric mean = 405 genera, but 365 collections). The reason is that the calibrated weights method correctly identifies the lumped collections as having more specimens, and the sampling quota in terms of specimens is identical, so

fewer collections are drawn. As a result, the cross-correlation again is high, and early Paleozoic and Cenozoic diversity do not diverge considerably in the large-scale data (Spearman's $\rho = 0.980$ and $p < 0.001$ after logging and differencing; median early Paleozoic and late Neogene diversity 364 and 641 genera, 76% higher).

Latitudinal distribution of data

One concern about a preliminary version of this data set (*S1*) was whether it sampled the tropics fairly in the Cenozoic (*S56-58*). Our new data confirm that tropical regions are more diverse in the Cenozoic (Fig. 3). Thus, if tropical collections are underrepresented in that interval, the young end of the curve (Fig. 1) may be depressed.

There are two separate issues here: whether our sampling compares well to the literature and earlier compilations, and whether the sampling is good in general. We consider it unlikely that both problems exist at once because there is a good match between our raw data curve and Sepkoski's throughout most of the Cretaceous and Paleogene (Fig. 4). Sepkoski's supposed radiation supposedly focused in the tropics (*S56-58*) would not also be visible in a comparable treatment of our data if our sampling was poor in both relative and absolute terms and his was not.

Nonethelesss, we obtained comparative data about the nature of sampling to address these two questions more directly. We used latitudinal distributions of countries named in references to quantify relative differences between four databases: ours; Sepkoski's genus-level compendium (*S2*); GeoRef, a proxy for the literature at large; and the systematics portion of The Treatise on Invertebrate Paleontology, Part N (Bivalvia). We note that the Treatise data were used heavily in the compilation of Sepkoski's compendium. The benchmark for assessing the absolute quality of tropical sampling in these databases is the relative amount of modern shelf area within 30º of the equator, which is 37% (*S45*).

For the Georef search, we worked from a list of all systematic references on Cenozoic molluscs. For Sepkoski's compendium and the Treatise, the entire reference list was checked. We scanned the titles to determine whether each paper focused on the Cenozoic and where the study was done. Based on the latter information, we assigned each study a latitude. Centroids were used for countries and states, where possible. We defined high and low latitudes as being above or below 30º N or S, respectively.

The data (Fig. S20) illustrate that the Paleobiology Database contains about 28 - 32% references on low latitude regions for the Cenozoic in general, whereas the other sources contain 19 - 27% low latitude data. Because the modern shelf

area figure of 37% (*S45*) is not much higher, none of the data sources dramatically undersamples this region. Our data actually oversample it relative to the literature in the late Neogene: respectively 24 and 38% (310/1270 and 538/1418) of our early and late Neogene collections are from below 30º.

Thus, our sampling is good in both relative and absolute terms. In relative terms, Sepkoski's low latitude sampling is not better than ours, and the raw data sets yield very similar patterns (Fig. 4), so the entire issue has nothing to do with any difference between our results and his. In absolute terms, we have substantial to excessive tropical data (Fig. S20), document a strong gradient even in the early Paleozoic (Fig. 3), and see an increase in diversity in the temperate zone, not just the low latitudes (Fig. 3A vs. Fig. 3B). Thus, putting other compilations aside, it seems unlikely that we have overlooked a large radiation confined to the tropics that did not accelerate until the Cretaceous and Cenozoic. To the contrary, we suspect that the anomalous uptick at the very end of our curve relates to our intentional oversampling of the tropics in the Neogene.

Completeness of age ranges

Because the Cenozoic is sampled heavily (Figs. S1, S2) and because adding more data to analyses does not change critical parts of the curve (Figs. S15, S16), it seems quite plausible that a similar proportion of late Cenozoic genera has been captured relative to earlier intervals. Indeed, even within the poorly preserved collections that make up our standard data set, the late Neogene includes actual occurrences of 1936 genera, 28% more than appear in any Paleozoic bin. Nonetheless, this argument may be hard to believe if one assumes from the start that the Cenozoic is dramatically more diverse, making it intuitive that our data might fail to capture a fair proportion of late Cenozoic diversity because not all the literature is included.

That there is no such failure is suggested by an analysis of the completeness data for genus-level sampling, as measured by the ratio of three timers to three timers plus part timers (see Counting method). In our understandarized data set, completeness mostly varies only from around 70 - 85%, with no clear temporal trend (Fig. S21A). Thus, there is no indication that we might have sampled the Cenozoic more poorly than the Paleozoic. In fact, the Cenozoic values are entirely within the range of the rest of the data.

The three timer statistic ignores the Recent, but the standard completeness proportion (*S14, S15*) for any one interval takes all other time intervals into account. Thus, if we include Recent as well as fossil occurrences this method infers that many more genera existed during the last few bins even though they have no fossil occurrences afterwards (*S21, S33*). With the Pull of the Recent

added, it is 31% lower in the early Neogene, at least 20% lower throughout the rest of the Cenozoic, and at least 11% lower throughout the Mesozoic. We mention the early Neogene because the late Neogene completeness value cannot be computed in the regular data set, exactly because it ignores the list of Recent taxa. These changes show that even in the well-sampled early Neogene there was a distinct statistical population of genera that failed to be preserved in the fossil record, but can be inferred to exist because they happen to have survived to the present day.

A key benefit of completeness statistics is their ability to measure the impact of sampling standardization (Fig. S21B). Our data collection protocol was designed to make sampling uniform even before sampling standardization (Figs. S1, S2), but without standardization there is a correlation between changes in estimated specimen counts and changes in the logit of the three timer completeness proportions ($\rho = 0.510$, $p < 0.001$: Fig. S21A). In other words, completeness goes up when more specimens are available. However, this relationship disappears when the completeness data are based on subsampled ranges ($\rho = 0.052$, $p = 0.734$: Fig. S21B), showing that standardization does what it is claimed to do. The reason for this large difference may be that standardization greatly decreases the variability of the proportions (standard deviation of logit-transformed values 0.574 vs. 0.291).

Bivalve age ranges

In contrast to our results, a recent study (*S3*) argued that the Pull of the Recent (*S21*) is unimportant because the diversity of Plio-Pleistocene bivalves is only slightly inflated by ranging Sepkoski's genera through to the Recent: 95% of genera and subgenera ranging through actually are sampled in that interval. By comparison, we recover only 53% of ranged-through invertebrate genera and subgenera in the late Neogene, which is twice as long as the Plio-Pleistocene by itself.

A basic possible reason for the discrepancy is that we have attempted unbiased sampling of the published fossil record because concentrating on unusual occurrences of rare genera would have compromised our sampling standardization protocol. This fact partially explains why our curve is still short of Sepkoski's in certain intervals when treated the same way (Fig. 4), despite the large size of our data set (Figs. S1, S2). However, Jablonski et al. (*S3*) tried to locate all published genera thought to range into the Plio-Pleistocene, including those with the poorest fossil records. Thus, while we might have found higher completeness in the Cenozoic by systematically documenting at least one occurrence of each genus instead of pursuing an unbiased sampling strategy throughout, the same would likely have been true of the rest of the Phanerozoic.

Another possibility is that bivalves are not representative of the overall fauna because of their high preservability, as suggested by three lines of evidence. First, Sepkoski records 4399 extant marine invertebrate genera with a fossil record, of which 839 (19%) are bivalves. However, there are about 180,000 extant marine invertebrate species, of which only 20,000 (11%) are bivalves (*S48*). Second, a recent, exhaustive study has found fossil records for 76% of all living marine bivalve genera (*S59*). Counts of extant marine invertebrate genera are rarely given, but as mentioned there are very approximately 42,000 extant invertebrate genera, which based on Sepkoski's figure suggests that on the order of 10% have been described as fossils. Even if this percentage is off by a factor of three or four, it is still far lower than 76%. Finally, putting aside genera, direct tabulations show that bivalves have an unusually high percentage of families with a fossil record compared to other large groups of extant skeletonized marine animals (*S60*).

Sampling of bivalves in the Paleobiology Database, which includes 69,039 occurrences of 2246 genera and subgenera, is also more complete than for marine invertebrates as a whole. For example, the 53% figure for the overall completeness of late Neogene invertebrates compares with 63% for late Neogene bivalves after pulling ranges to the Recent (which is much lower than Jablonski et al.'s 95% figure [*S3*]). Based on three timer analysis, which is robust to biases in overall completeness, bivalves are almost invariably better sampled after the Paleozoic (circles in Fig. S21A). This consistent and often large difference persists in the sampling standardized data (Fig. S21B). Nonetheless, pulling stratigraphic ranges to the present day leads to many range extensions of sparsely fossilized genera that are not sampled in the early Neogene. Thus, if we do not pull ranges to the Recent, early Neogene bivalve completeness in the raw data grows from 70% to 84%, and in the standardized data from 48% to 69%.

A different explanation would be that the Cenozoic and particularly Plio-Pleistocene fossil record in general, not just for bivalves, is exceptionally well sampled in the overall literature. One hint is that ranging genera forward only to their last fossil appearance in Sepkoski's data, which must remove any trace of the Pull of the Recent, still leaves an exponential, if not nearly as large, Meso-Cenozoic radiation (*S31, S33*). We see the same pattern in our data (Fig. S9).

The key evidence, however, relates to the fact that we have intentionally sampled the Neogene very intensively (Figs. S1, S2), with more collections (1418) falling in our final bin than in any other. Despite this effort, we still cannot fully match Sepkoski's enormous Cenozoic figures, even though long intervals in our curve are just as diverse as in his (Fig. 4). If the gap were due to poor Cenozoic sampling in the Paleobiology Database, sampling standardization would narrow it in relative terms by bringing the other intervals down farther. However, this pattern is not seen in curves that truncate ranges of extant taxa at their last fossil

appearances (thin lines in Figs. 5 and S8A). Standardizing decreases the relative height of the Cenozoic, essentially flattening the curve starting in the Cretaceous, which widens the gap in relative terms. By contrast, leaving in the Pull of the Recent (thick lines in Figs. S9 and S10A) permits an exponential radiation to appear one way or another. Thus, although the latter effect seems to govern whether a truly massive radiation is seen, the data suggest that there is a large Pull of the Cenozoic (or Plio-Pleistocene).

Finally, it is possible that Jablonski et al.'s sampling protocol (*S3*) produced a maximum completeness estimate instead of an exact figure, because corrections were only said to have been made to Plio-Pleistocene data. Thus, the number of genera sampled in that one interval was increased with better identifications, taxonomic standardization, and more examination of collections and published data, but it is not clear whether similarly thorough efforts were made to find even older records of other extant genera. Such records would decrease the apparent completeness percentage unless they were matched immediately by new Plio-Pleistocene occurrences. In other words, existing holes in the 5.3 m.y.-long Plio-Pleistocene record were filled, but we cannot be sure whether new holes were opened up in older intervals. Similar reasoning would apply to Jablonski et al.'s estimate of Maastrichtian completeness (*S3*).

It does seem likely that many Recent genera still have misidentified or unpublished fossil records before but not within the Plio-Pleistocene. Only 124 (14.8%) of Sepkoski's Recent genera first appear in the Plio-Pleistocene, but 121 (14.4%) go back into the Mesozoic or even Paleozoic, so truly old appearances are routine. Thus, much may remain to be ferreted out of the literature and museum collections. Indeed, Jablonski et al. (*S3*) added 144 additional genera to Sepkoski's tally of 814 Recent genera with fossil records (our version of Sepkoski's compendium includes 839 fossilized Recent genera, but the discrepancy is minor). Their extension of the fossilized Recent bivalve roster by 18% with a study of just 8% of the Cenozoic record (*S3*) shows that Sepkoski's compilation is quite incomplete (Fig. S17B). Thus, regardless of our effort, undersampling in Sepkoski's own data may be sufficient to allow a strong Pull of the Recent, even though a greatly improved data set for one group in particular does not show it.

Regardless of how the results of Fig. S21 and those of Jablonski et al. (*S3*) are ultimately reconciled, it is clear that the Plio-Pleistocene fossil record is not representative of the deeper Cenozoic record, bivalves are not representative of marine invertebrates, and Sepkoski's compilation is not a highly complete sample of the paleontological literature. Of more importance is the fact that applying Sepkoski's methods to our data set simultaneously brings it into accord with his (Fig. 4) and creates a large Pull of the Recent (Fig. S9). Thus, it is hard to see how this bias could have no visible effect whatsoever on his data.

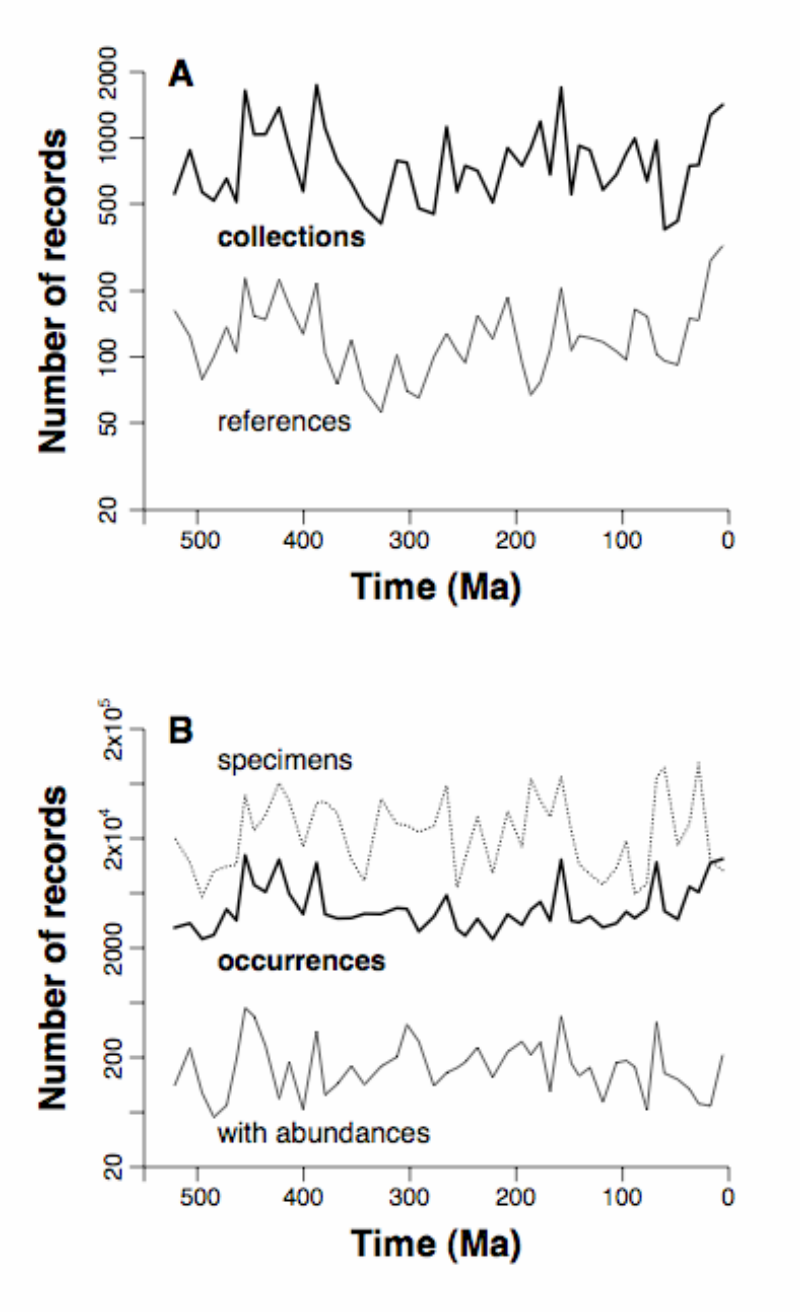

Fig. S1. (A) Number of fossil collections (thick line) and literature references (thin line) recorded in each temporal bin. **(B)** Total of directly counted and estimated specimens or individuals (dotted line), number of taxonomic occurrences (thick lines), and number of collections including direct specimen or individual count data (thin line).

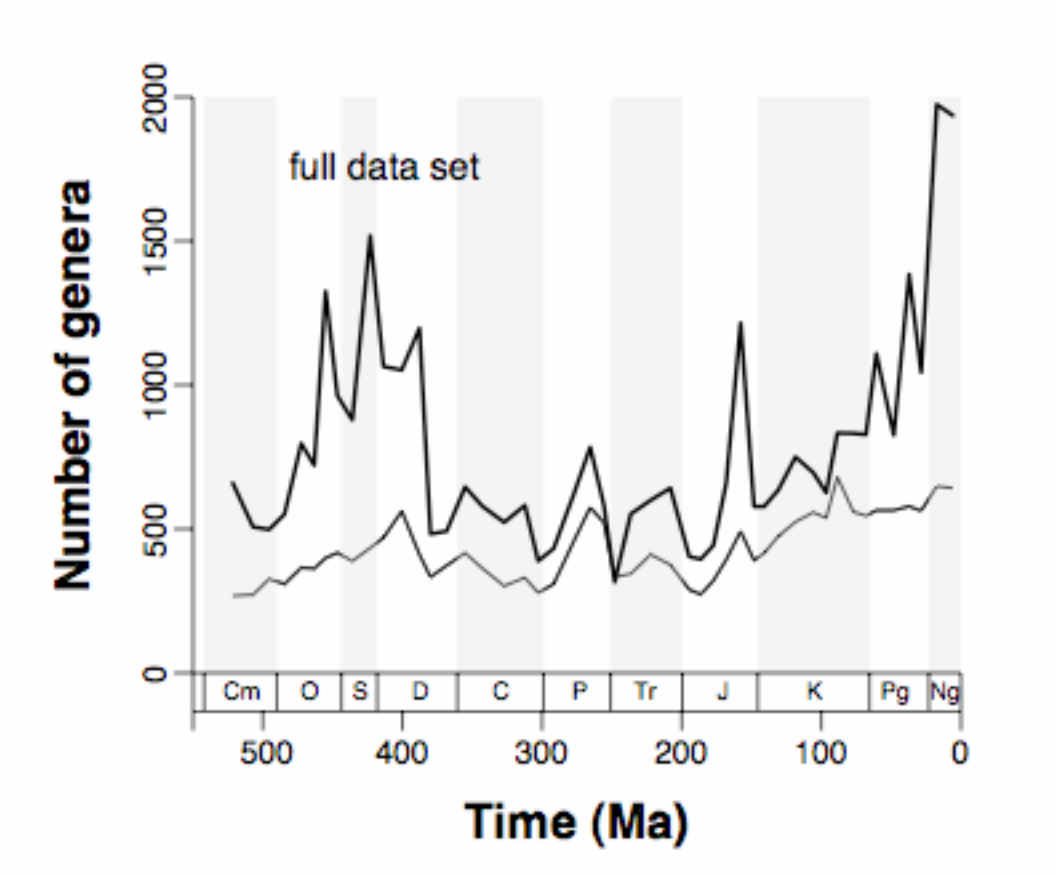

Fig. S2. Counts of all genera in the full data set that fall in each bin (thick line), and estimated counts using the subsampled data (thin line; see also Fig. 1).

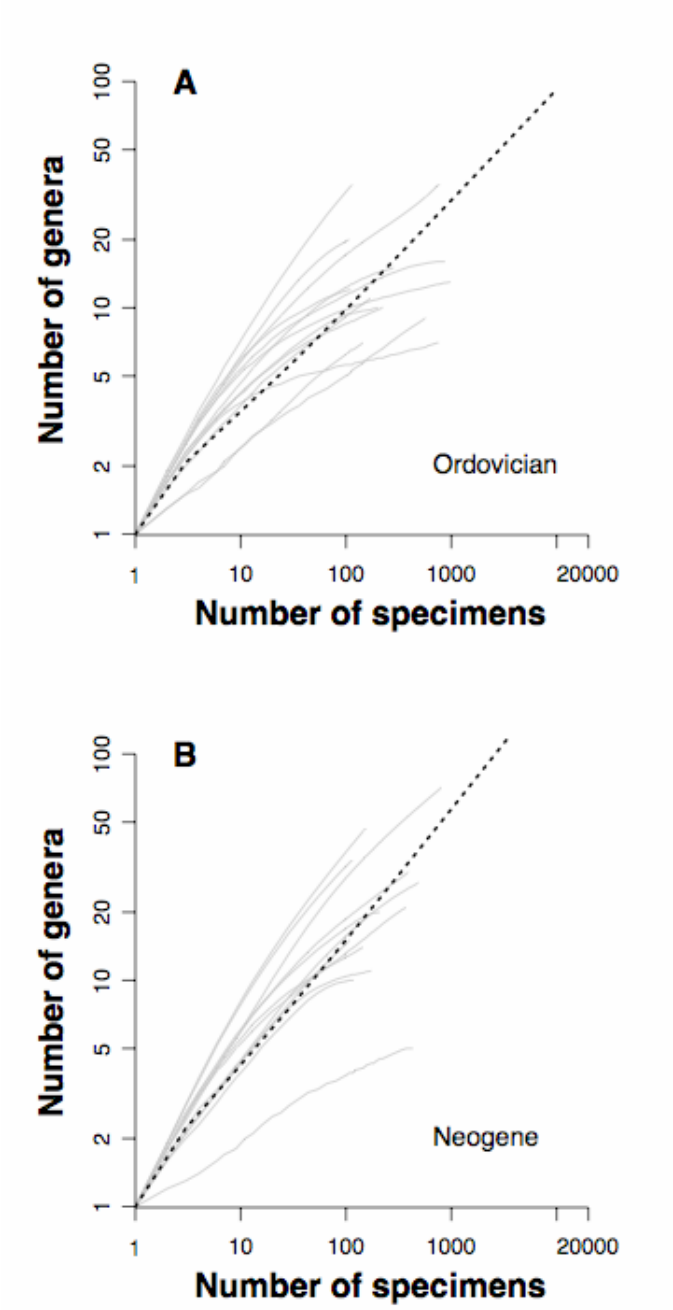

Fig. S3. Rarefaction curves for individual collections belonging to two bins. In each panel, the collection with the median rarefied diversity at a sample size of 100 is used to calibrate a blended linear-power law function (thick black line). **(A)** Data for the Caradoc bin, the best sampled in the Ordovician. **(B)** Data for the late Neogene bin, the best sampled in the Cenozoic.

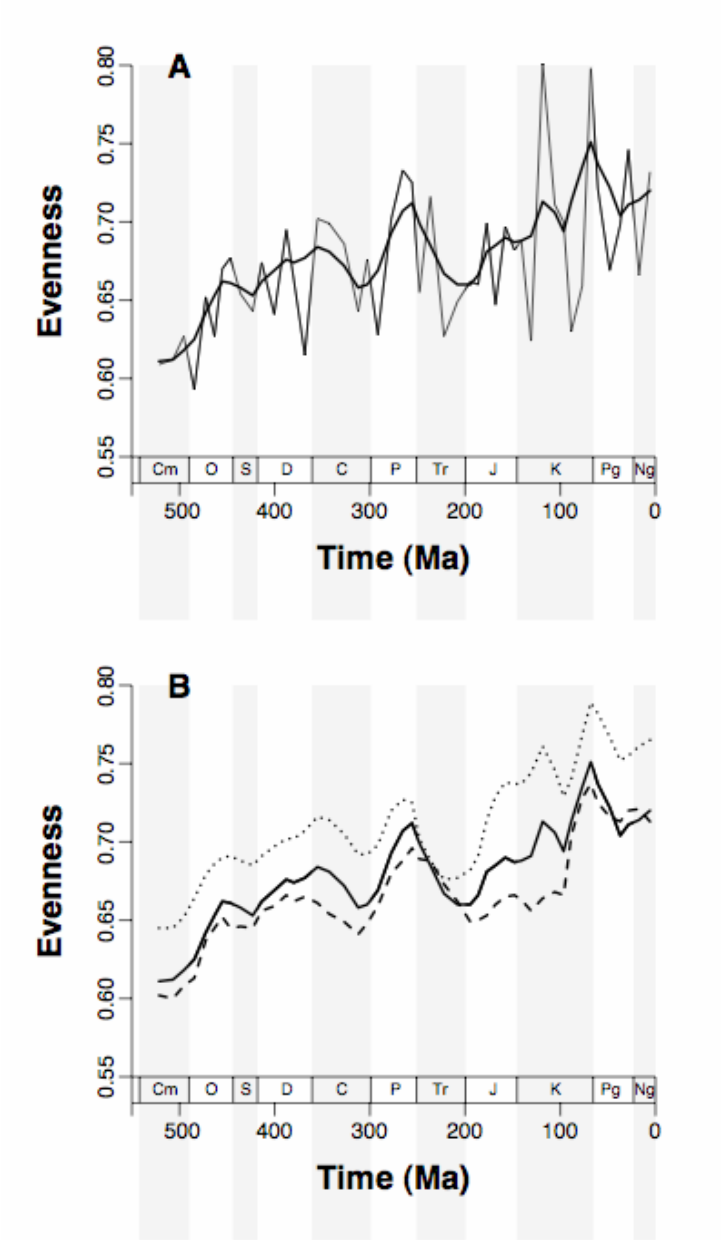

Fig. S4. Evenness estimates for 11 m.y.-long bins. **(A)** Raw data based on separate analyses in each bin (thin line) and weighted moving averages (WMAs) across five-bin windows (thick line). The values derive from lines fitted to pass through the median genus count at a 100 specimen sampling level (Fig. S3). **(B)** WMAs based on analyses of median genus counts at the 25, 100, and 200 specimen levels (dotted, solid, and dashed lines, respectively).

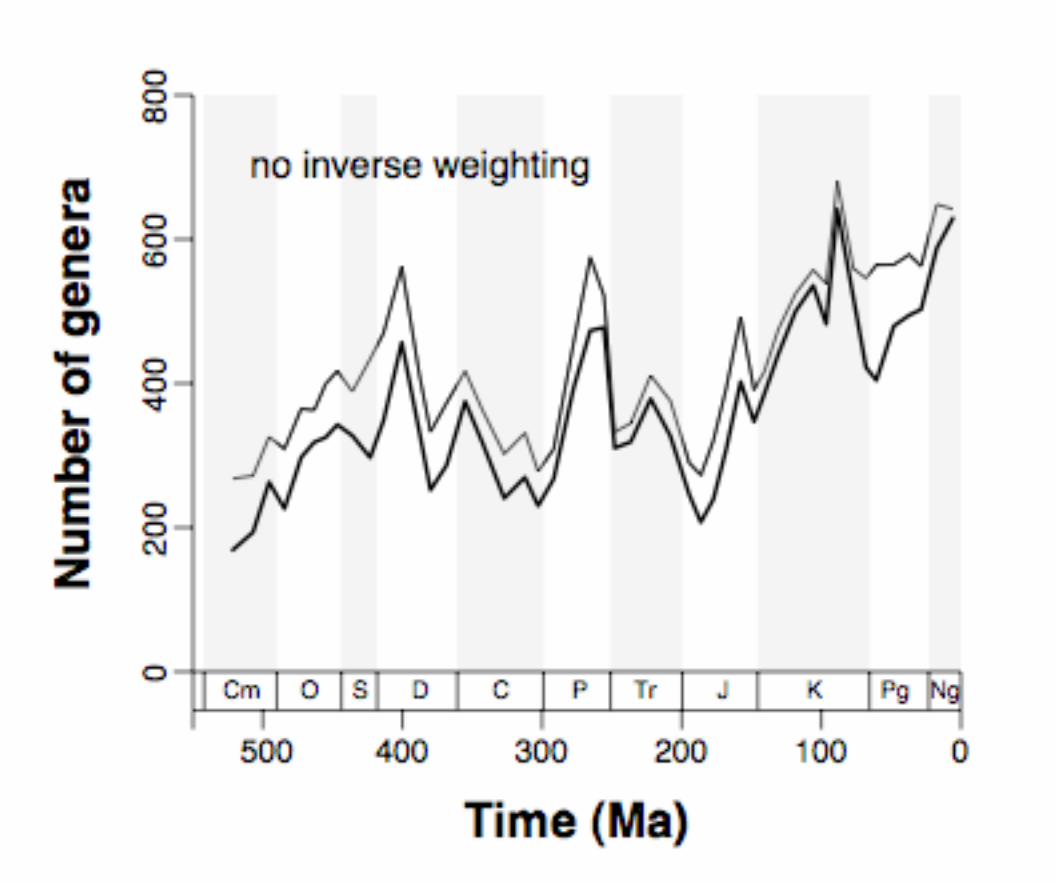

Fig. S5. Effect of not weighting sampling probabilities by the inverse of estimated collection size. Thin and thick lines respectively show calibrated occurrence weights curves with and without inverse weighting.

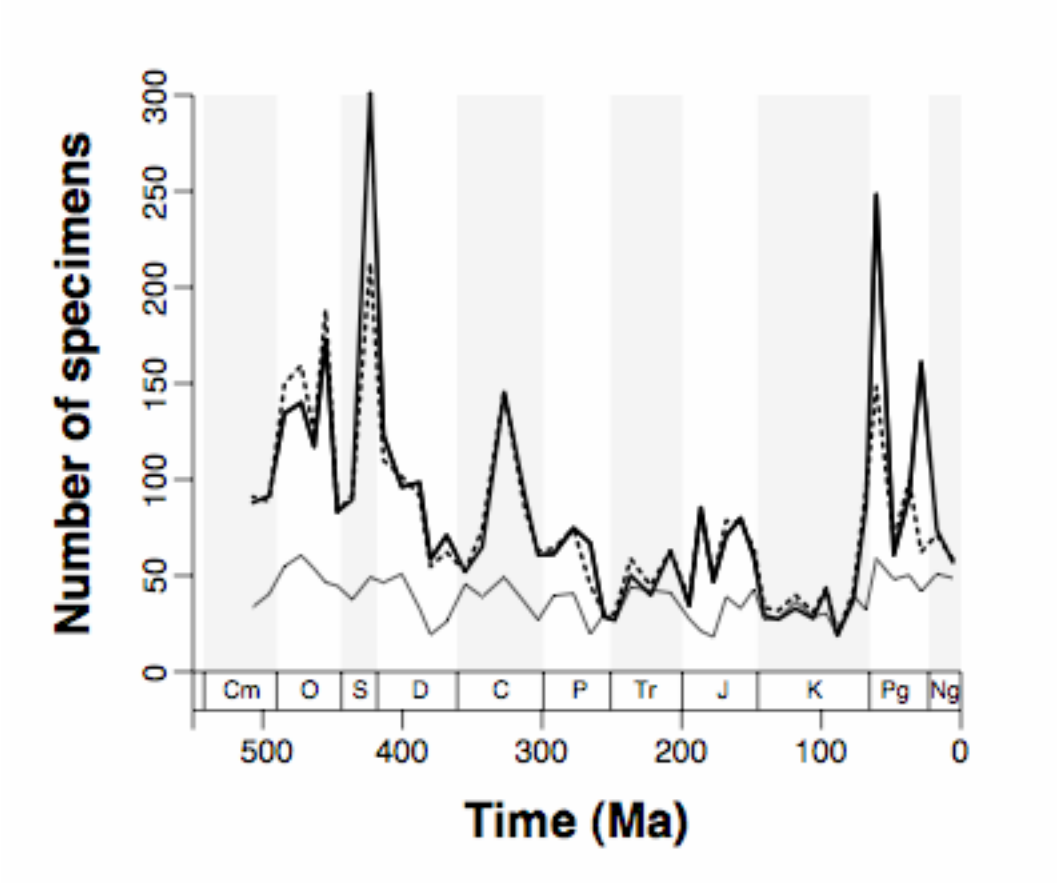

Fig. S6. The average number of specimens per fossil collection estimated using the calibrated weights method (Figs. S3, S4). Counts in the overall data (thick line) mirror those in subsampled data produced without inverse weighting (dotted line) but not those in our usual subsampled data produced with inverse weighting (thin line).

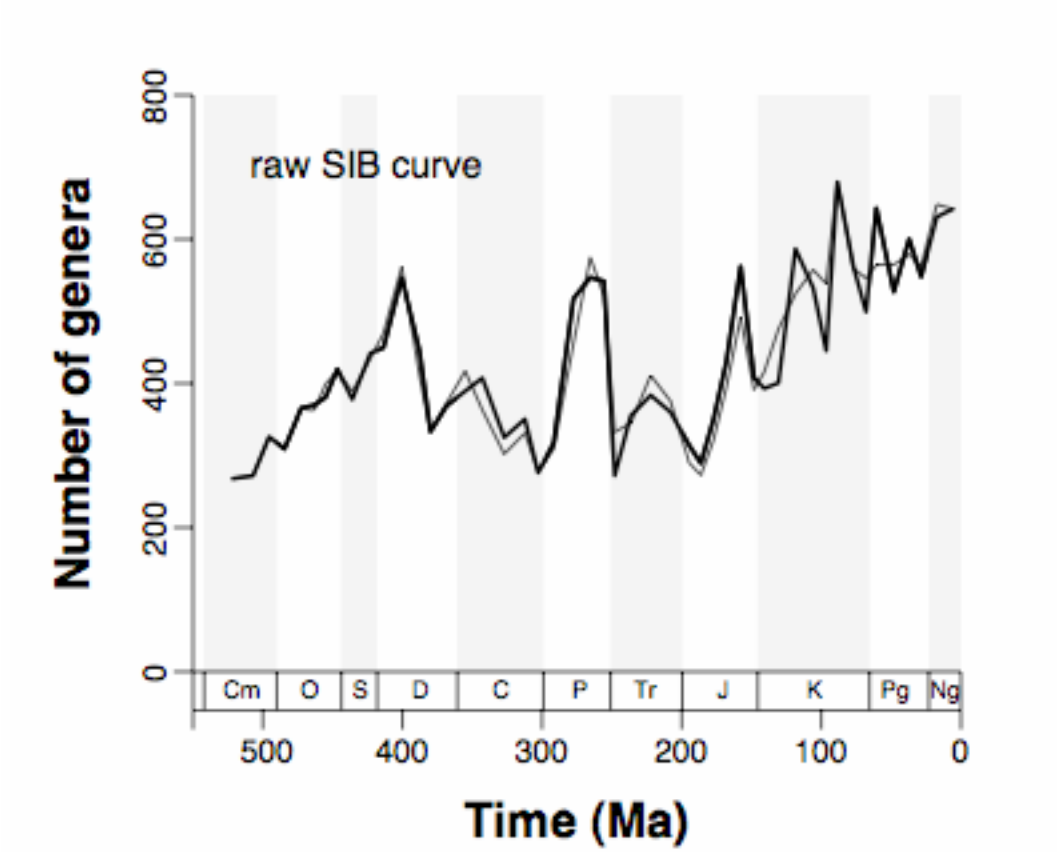

Fig. S7. Alternative SIB diversity curves based on different counting methods. The raw SIB curve (thick line) shows less variance once it is corrected for residual sampling error that is indicated by part timer counts (thin line).

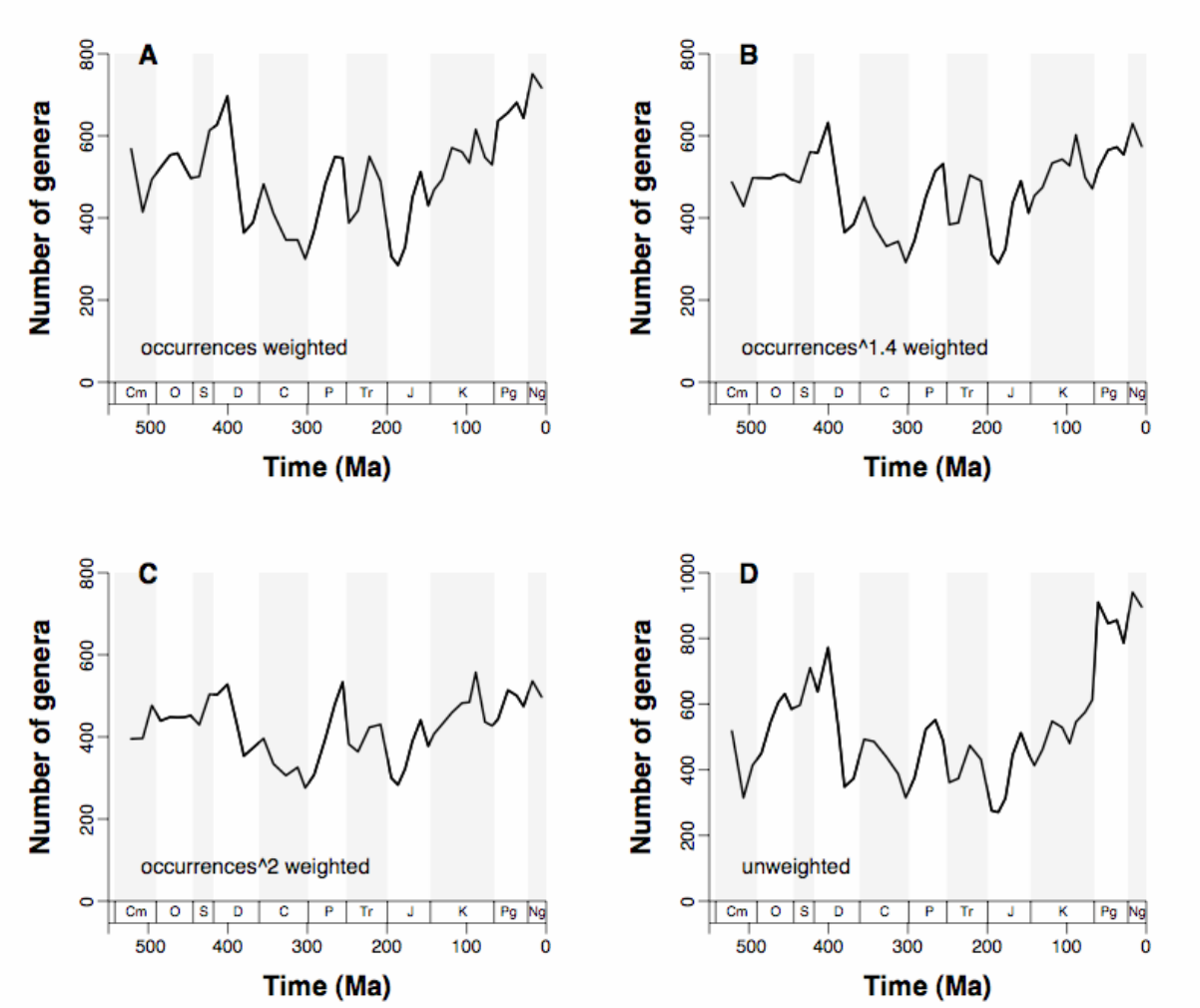

Fig. S8. Additional SIB diversity curves based on alternative methods of sampling standardization. As usual, sampling probabilities for each collection are subjected to inverse weighting. **(A)** Counts of genera based on drawing collections until a quota of 2390 taxonomic occurrences has been reached (OW). **(B)** Counts based on drawing collections up to a quota of 5220 specimens, which is estimated by raising occurrence counts to the power of 1.4 and summing them across collections (O1.4W). **(C)** Counts based on a quota of 17,700 specimens estimated by squaring and then summing the number of occurrences (O2W). **(D)** Counts based on drawing 380 taxonomic collections per bin regardless of specimen counts (UW).

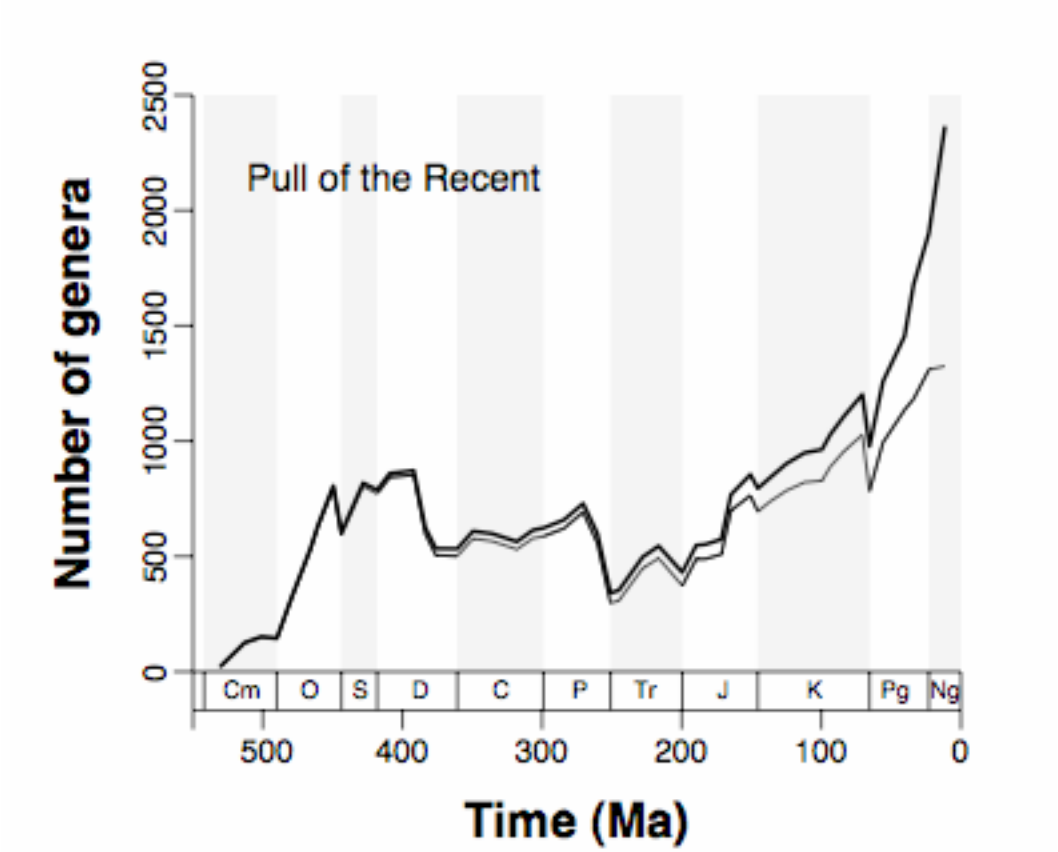

Fig. S9. Paleobiology Database boundary-crosser curves with no sampling standardization in which extant genera are pulled forward to the Recent (thick line) or only to their last fossil occurrences (thin line).

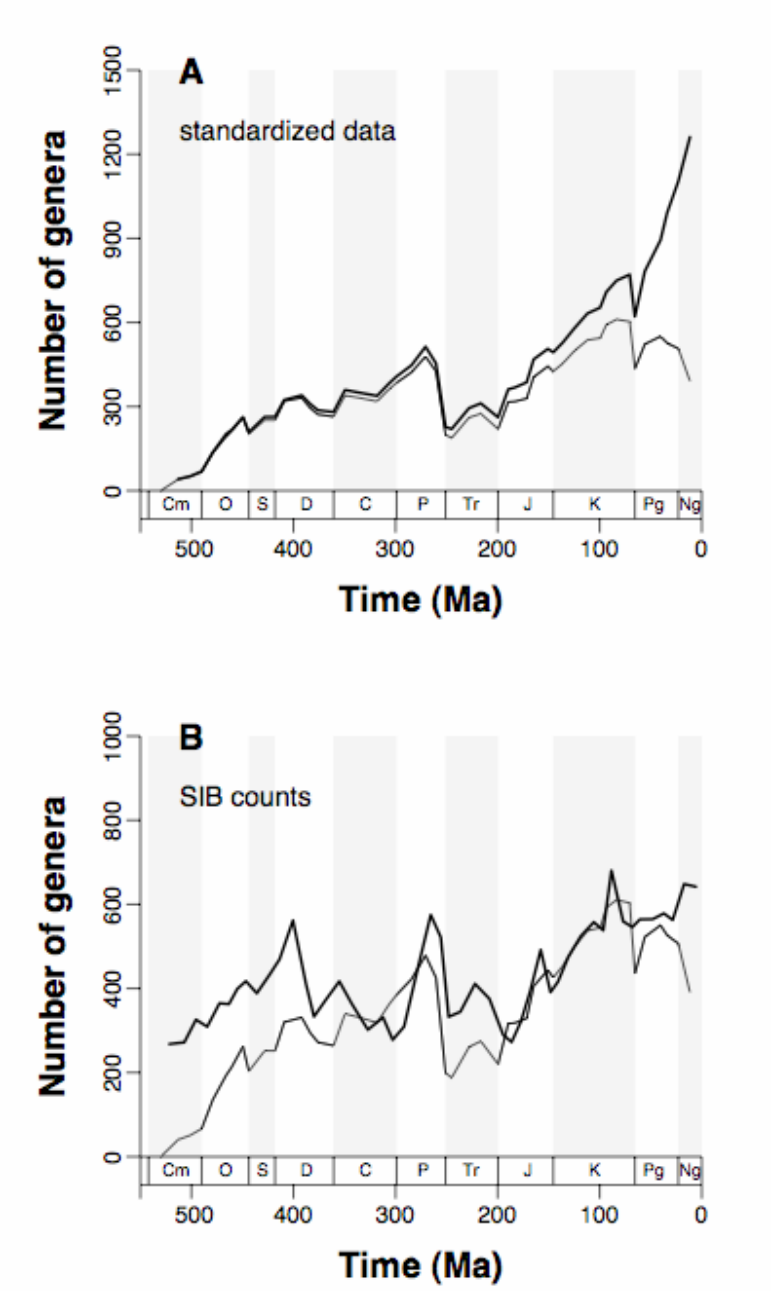

Fig. S10. Diversity curves generated by different counting methods, but using the same sampling-standardized data produced by calibrated weights subsampling. **(A)** Boundary-crosser curves with extant genera pulled forward to the Recent (thick line) or not (thin line). **(B)** Boundary-crosser (thin line) or SIB (thick line) curves, with genera not pulled forward. The SIB curve (same as Fig. 1) is sometimes higher because more genera must have existed within the entirety of a bin than at its boundaries, but often equal and sometimes lower because BC includes many genera with long age ranges that are not actually sampled in specific bins. Extant genera are also included in SIB analyses, but treated the same as any

others.

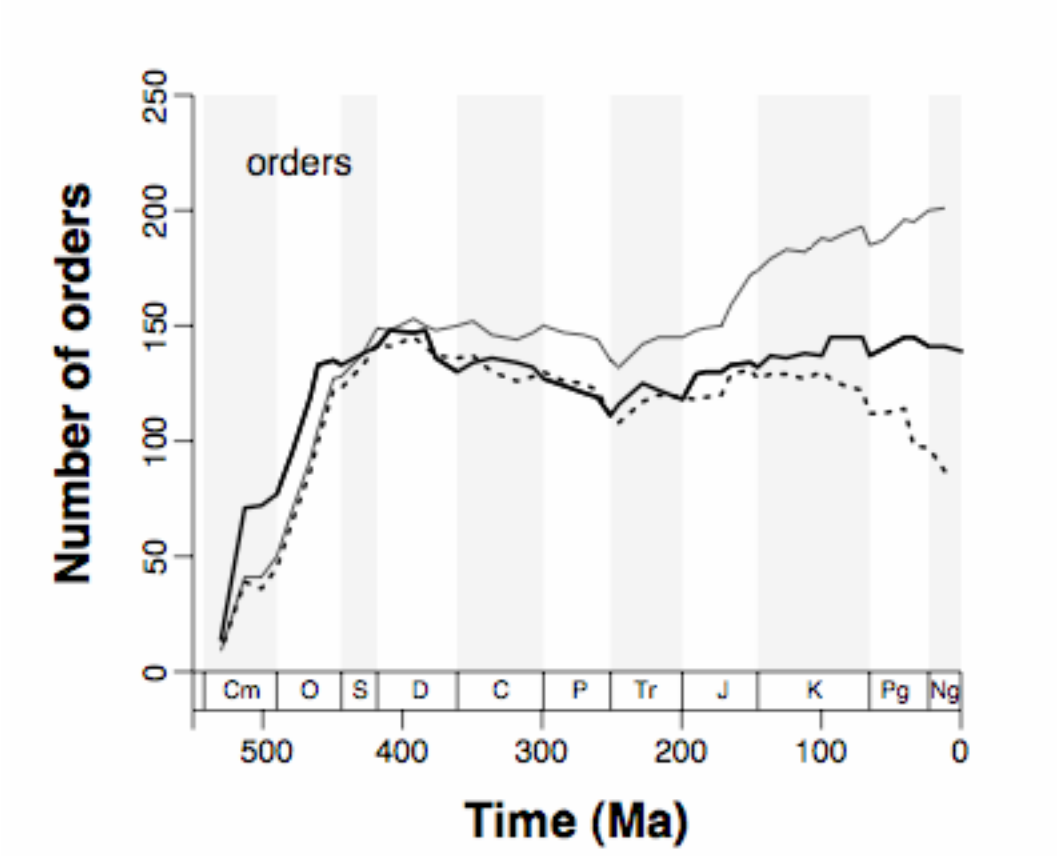

Fig. S11. Ordinal diversity curves based on Sepkoski's compendium (*S2*: thick line) and the new data (thin and dashed lines). Counts are of orders crossing boundaries between temporal bins (boundary crossers). Ranges of extant orders are consistently pulled forward either to the Recent (thin line) or to last fossil occurrences (dashed line) in our data, but in Sepkoski's data they can only be pulled forward to the last nominal occurrences of included genera, so some extant orders are not pulled forward to the Recent.

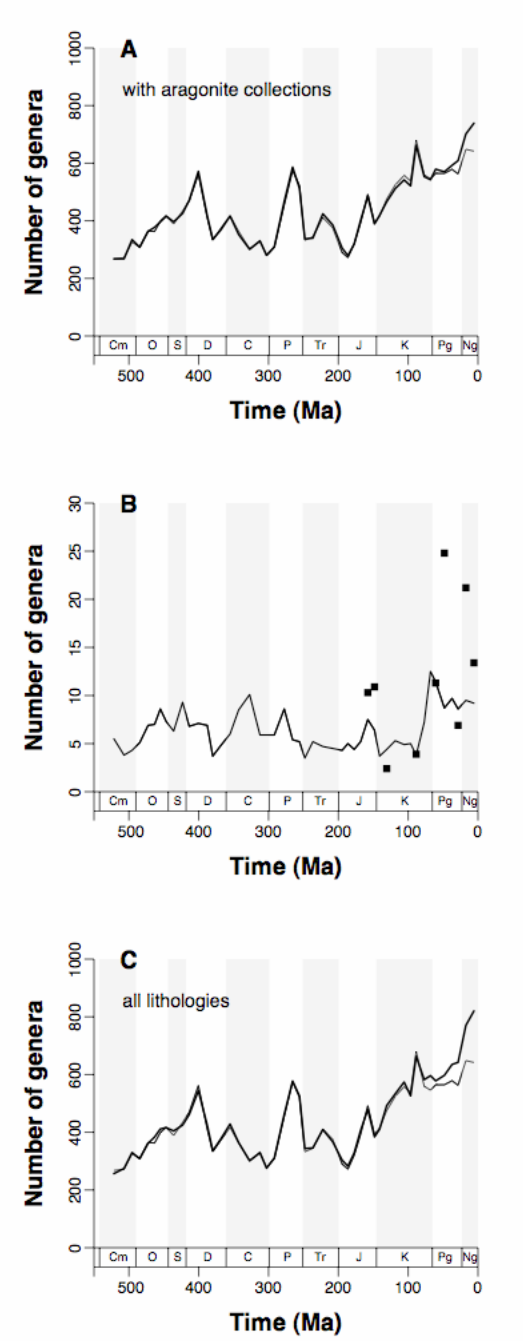

Fig. S12. Effects of aragonite preservation and lithification on diversity. **(A)** Standardized curves including and excluding collections that preserve original aragonite (thick and thin lines). Quota is 16,200 estimated specimens. **(B)** Average number of genera per collection in each temporal bin. Unlithified samples (squares) show higher richness than other samples (line). Data points including at least 10 collections are shown. **(C)** Standardized curve including and excluding collections from unlithified or sieved, poorly lithified sediments (thick and thin lines). Quota is 16,200 estimated specimens.

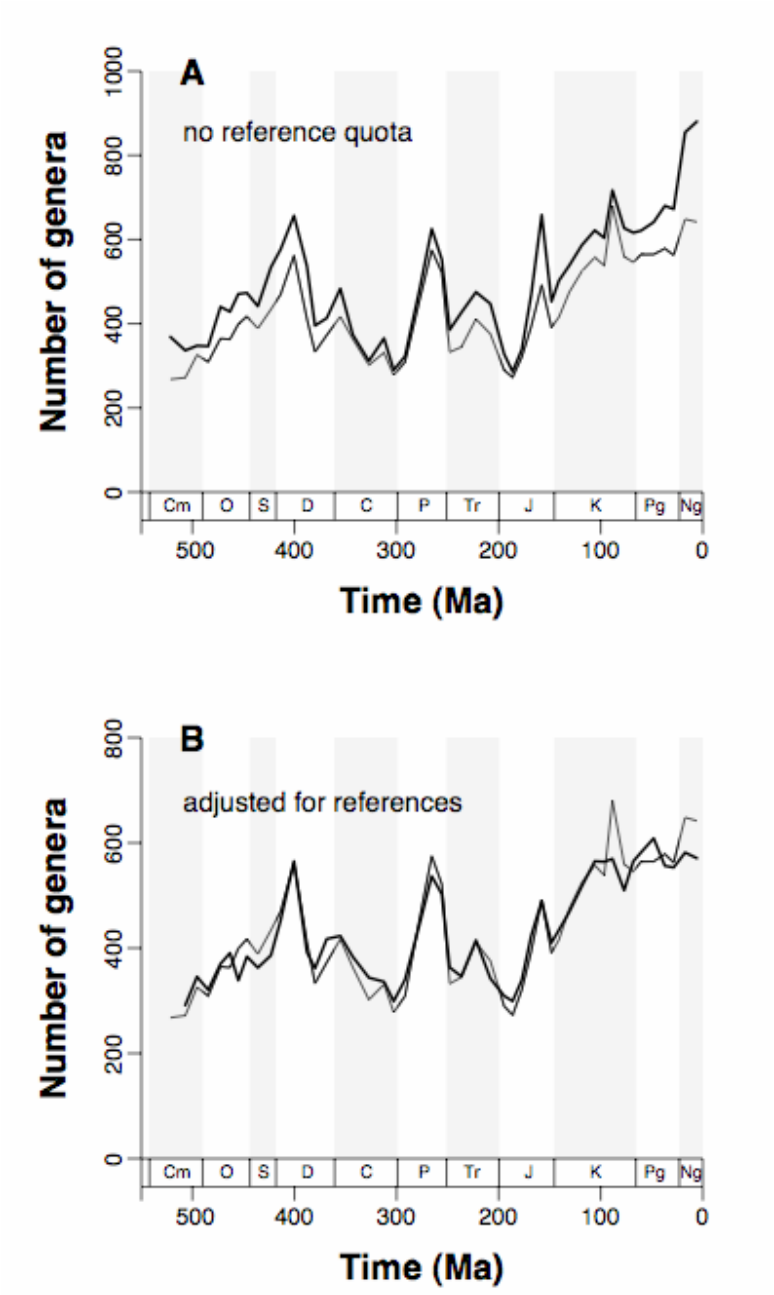

Fig. S13. Effect of reference counts on diversity. **(A)** Sampling standardized curve produced by drawing from all available references during each trial (thick line) contrasted with our otherwise identical diversity curve produced by using a quota of 65 references drawn randomly during each trial (thin line; see also Fig. 1). **(B)** Adjusted version of our standard diversity curve (thick line) produced by regressing changes in log genus counts in the curve with no reference quota (panel A) on changes in log cell counts, and then using the slope to rescale the counts.

Thin line is the standard curve (Fig. 1). Adjusted curve is scaled down by a factor of 8/9 for ease of comparison.

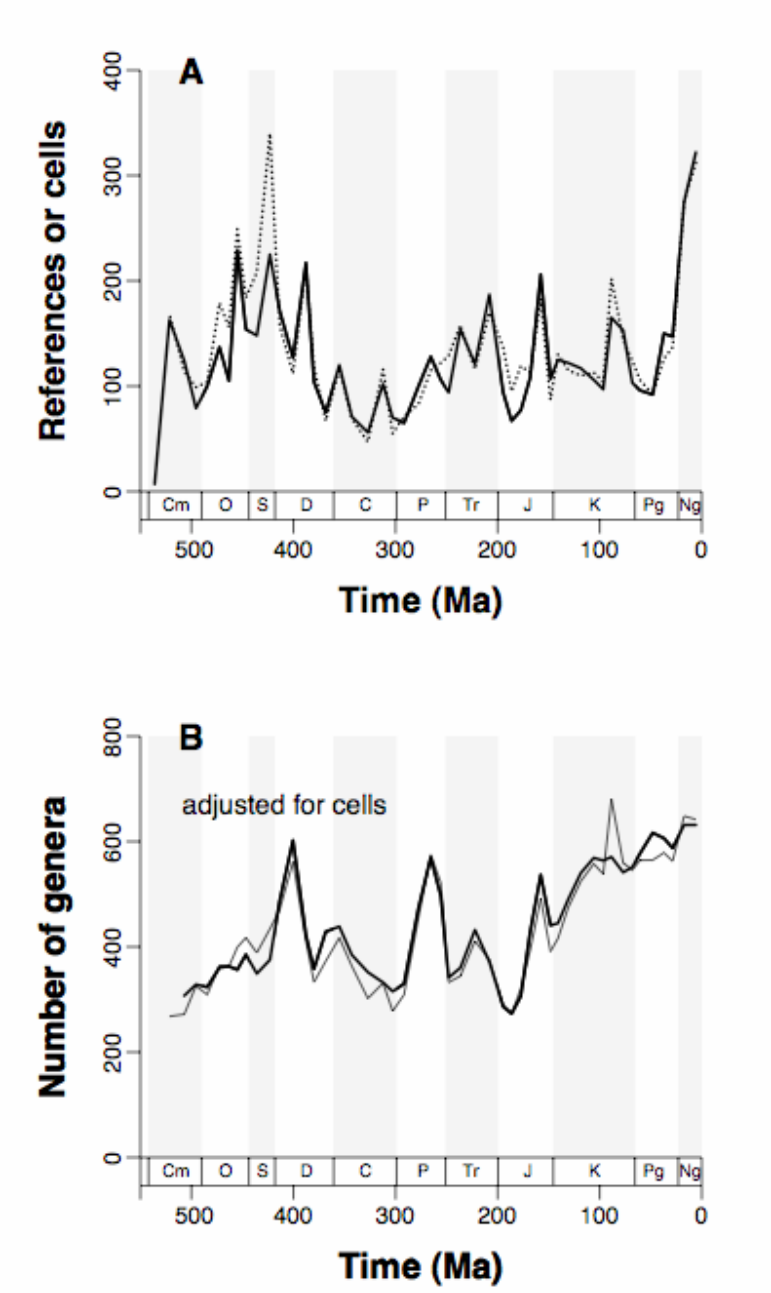

Fig. S14. Paleogeographic sampling through the Phanerozoic and its effect on diversity. **(A)** Counts of sampled paleolatitude/paleolongitude grid cells that are approximately equal in size (thick line) contrasted with reference counts (dotted line). Each longitudinal band of cells is 10º in height, and there are respectively 36, 35, 33, 29, 25, 21, 15, 9, and 3 in each band going from the equator to the

poles. **(B)** Adjusted diversity curve based on geographic cell counts, produced using the same method as in Fig. S13B and scaled down by a factor of 8/9. Thin line is the standard curve (Fig. 1).

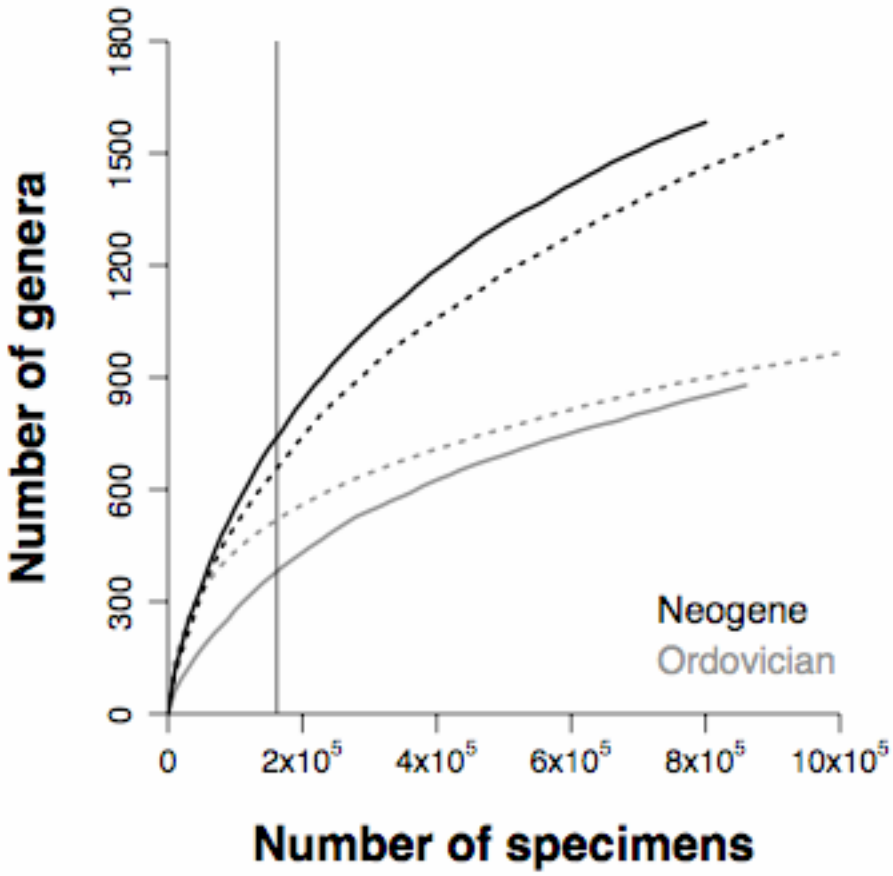

Fig. S15. Subsampling curves for four individual 11 m.y.-long temporal bins comparing sampled genera to estimated specimen counts. Regular sampled-in-bin counts are given instead of corrected counts because the three timer correction requires having data in multiple consecutive bins, which is not possible at very high sampling levels. Vertical line marks the quota of 16,200 specimens used to construct the calibrated weights curve (Fig. 1). Bins are Caradoc (dashed gray line), Ashgill (solid gray line), Early/Middle Miocene (dashed black line), and Late Miocene/Pliocene/Pleistocene (solid black line). These bins respectively include about 284,200, 86,300, 93,800, and 80,900 estimated specimens.

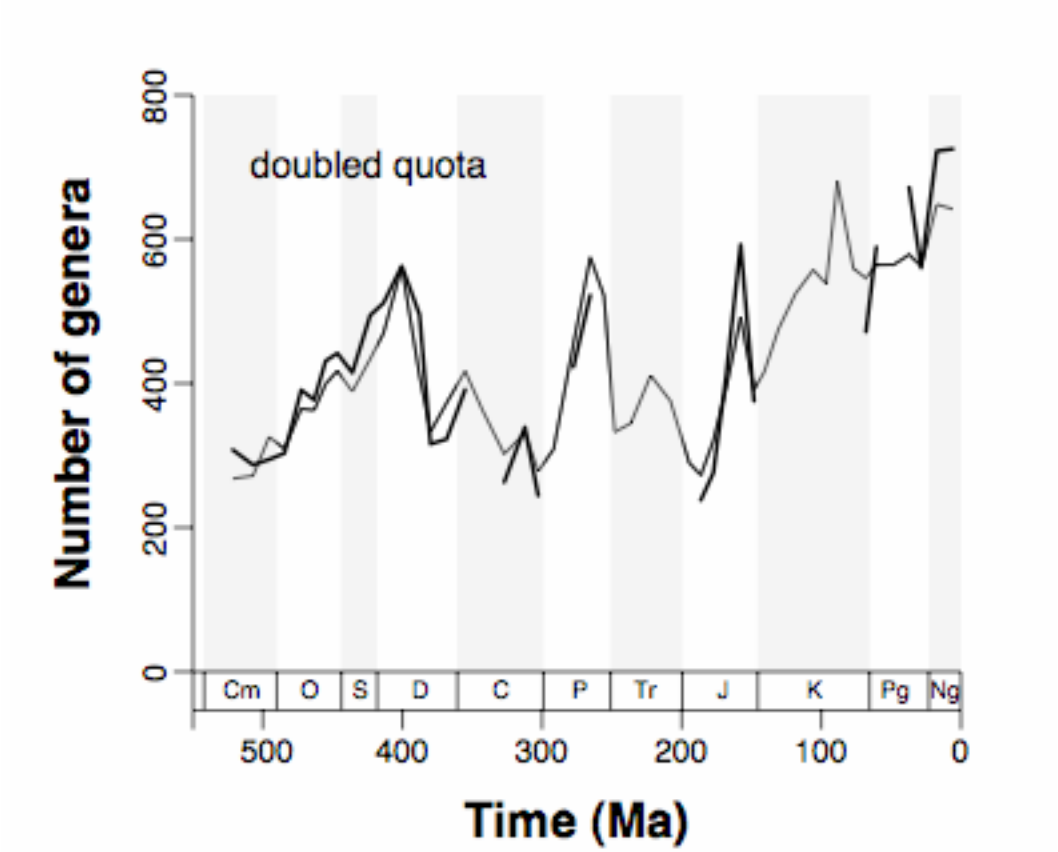

Fig. S16. Effect of doubling the sampling quota for each bin from 16,200 estimated specimens (thin line, see also Fig. 1) to 32,400 specimens (thick line). In the latter case, no correction is made for local variation in sampling completeness because there are too many gaps in the data to compute three timer and part timer counts consistently. Gaps in the curve span bins that do not meet the doubled quota. Curve based on the higher quota is scaled down by a factor of 8/11 to show the similarity in shape.

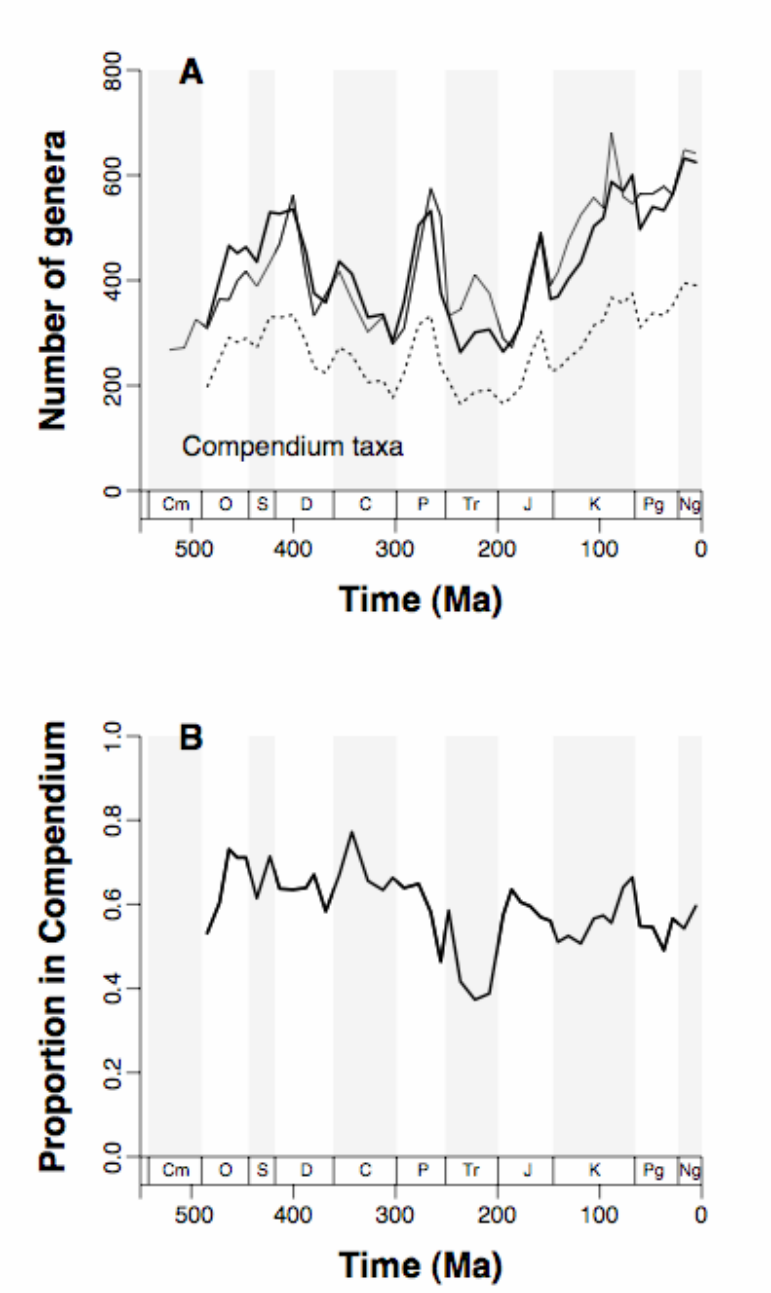

Fig. S17. Effects of restricting data to be in accord with Sepkoski's compendium (*5*). **(A)** Curve including only the genera found in the compendium, and excluding occurrences of these genera falling outside of the age range defined in the compendium (thick and dotted lines), contrasted with the curve for the standard data set (Fig. 1, thin line). Sampling quota is 9220 estimated specimens per bin. Thin and dotted lines are scaled to a maximum of 800 genera, whereas thick line is the same curve scaled down by a factor of 5/8 to allow easy comparisons with the standard curve. **(B)** Proportion of genera sampled in bins that also range into the bins in Sepkoski's compendium.

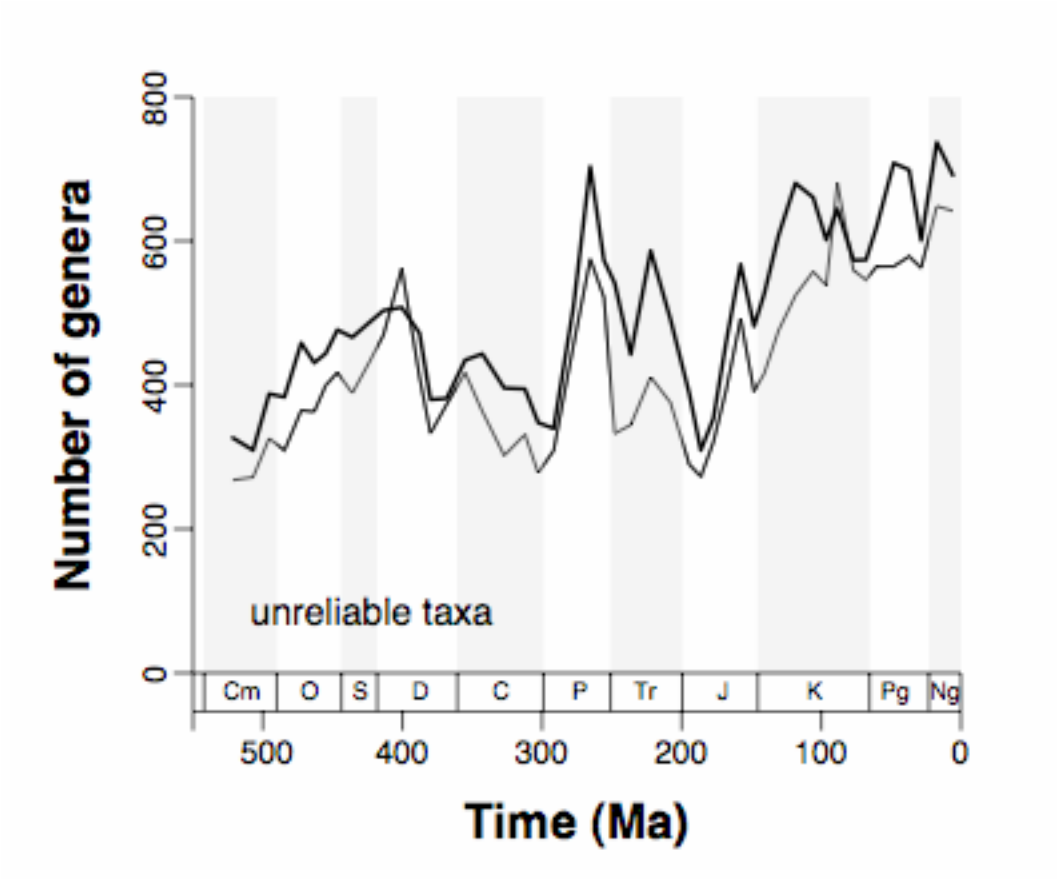

Fig. S18. Effects of loosening download criteria to include as much unreliable taxonomic information as possible. The unreliable curve (thick line) includes all occurrences of all taxa at any rank in marine collections, not just marine invertebrate genera, and is contrasted with the standard curve (thin line). Sampling quota is 16,200 estimated specimens per bin.

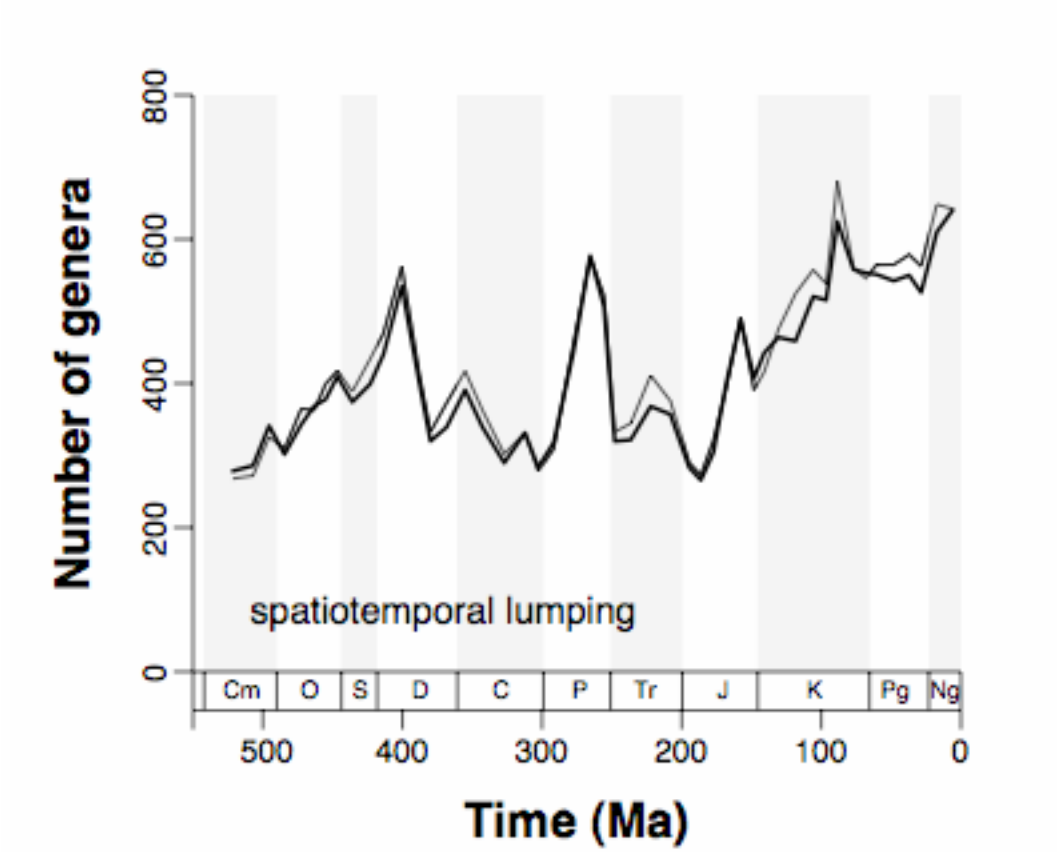

Fig. S19. Effect of coarsening the spatiotemporal scale of collections and thereby making them more uniform. Sampling quota is 13,100 estimated specimens per bin, below that used in the main analysis (Fig. 1). The large-scale data (thick line) are contrasted with the fine-scale data from the standard analysis (thin line).

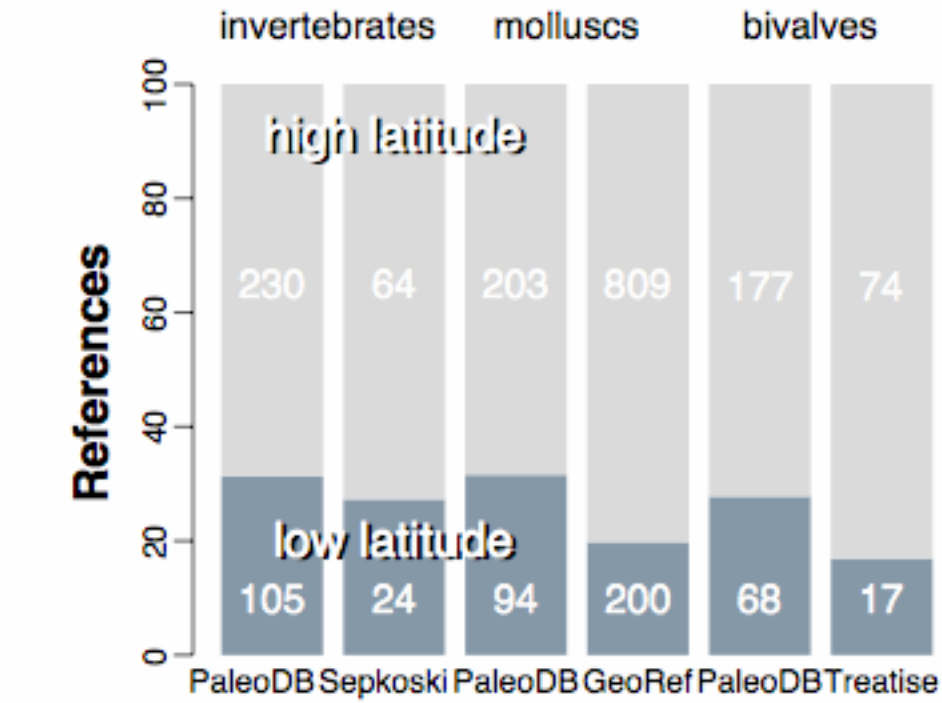

Fig. S20. Latitudinal distribution of Cenozoic data in the Paleobiology Database and other sources. Data are binned into low current latitude (< 30º) and high current latitude (> 30º) categories. Counts within the bars are of references pertaining to each group (labels at top) in each source (labels at bottom). "Invertebrates" includes Anthozoa, Brachiopoda, Bryozoa, Cirripedia, and Mollusca. PaleoDB = Paleobiology Database; Sepkoski = Sepkoski's compendium of fossil marine genera (*S2*); GeoRef = GeoRef bibliographic database; Treatise = Bivalvia volumes of the Treatise on Invertebrate Paleontology.

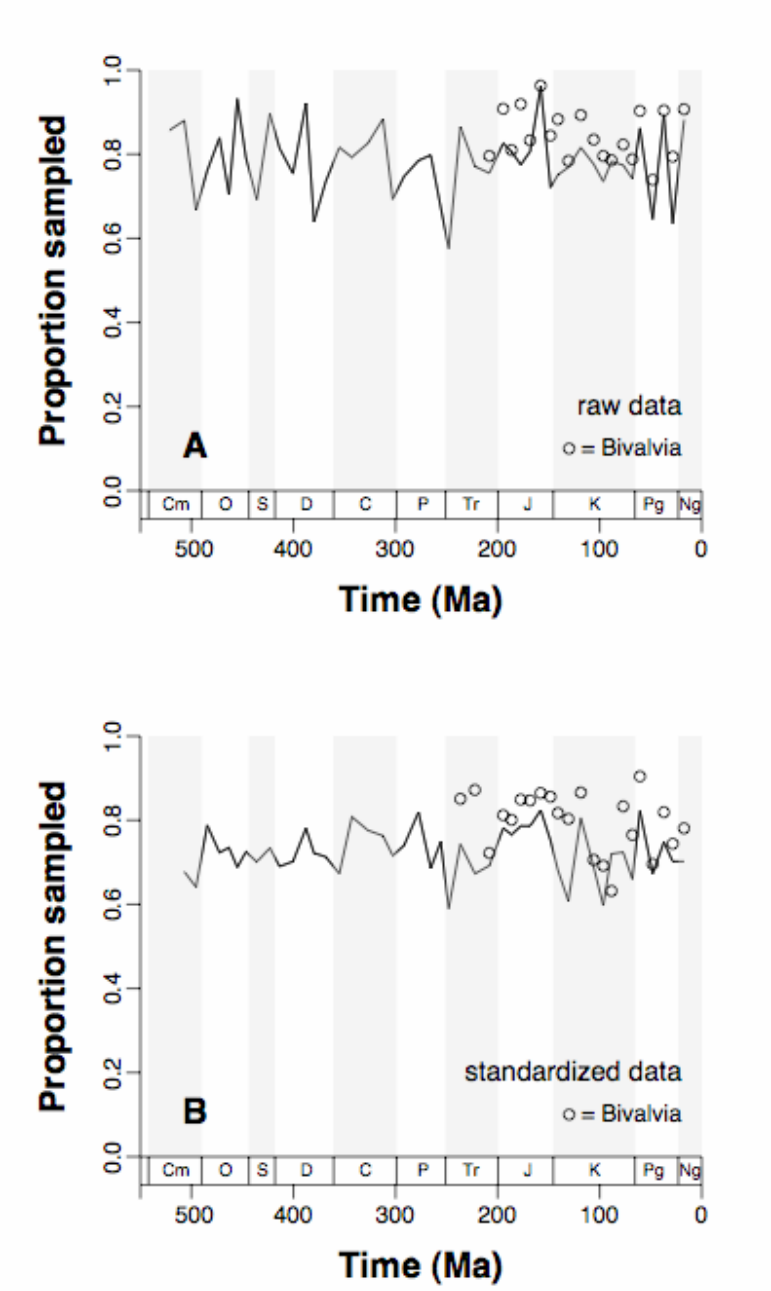

Fig. S21. Completeness of sampling through the Phanerozoic. Proportions are of genera occuring immediately before and after each bin that are sampled within it. Circles are data for bivalves, separately sampling-standardized to the level of 190 collections without any weighting. Bins falling under this quota are not represented. **(A)** Proportions for data without standardization. **(B)** Proportions for sampling-standardized data only.

References

S1. J. Alroy et al., *Proc. Natl. Acad. Sci. U.S.A.* **98**, 6261 (2001).
S2. J. J. Sepkoski, Jr., in *Global Events and Event Stratigraphy*, O. H. Walliser, Ed. (Springer, Berlin, 1996), pp. 35-52.
S3. D. Jablonski, K. Roy, J. W. Valentine, R. M. Price, P. S. Anderson, *Science* **300**, 1133 (2003).
S4. J. B. C. Jackson, K. G. Johnson, *Science* **293**, 2401 (2001).
S5. J. Alroy, *Paleobiology* **26**, 707 (2000).
S6. T. D. Olszewski, *Oikos* **104**, 377 (2004).
S7. M. G. Powell, M. Kowalewski, *Geology* **30**, 331 (2002).
S8. S. E. Peters, *Paleobiology* **30**, 325 (2004).
S9. M. Kowalewski, W. Kiessling, M. Aberhan, F. T. Fürsich, D. Scarponi, S. L. Barbour Wood, A. P. Hoffmeister. *Paleobiology* **32**, 509 (2006).
S10. A. M. Bush, R. K Bambach, *J. Geol.* **112**, 625 (2004).
S11. J. C. Tipper, *Paleobiology* **5**, 423 (1979).
S12. R. K. Colwell, J. A. Coddington, *Phil. Trans. R. Soc. Lond. B* **345**, 101 (1994).
S13. C. R. C. Paul, in *Problems of Phylogenetic Reconstruction*, K. A. Joysey, A. E. Friday, Eds. (Academic, London, 1982), pp. 75-117.
S14. M. Foote, D. M. Raup, *Paleobiology* **22**, 121 (1996).
S15. A. I. Miller, M. Foote, *Paleobiology* **22**, 304 (1996).
S16. M. Foote, *Paleobiology* **20**, 320 (1994).
S17. H. Chernoff, *Ann. Math. Stat.* **23**, 493 (1952).
S18. J. Alroy, *Palaeogeog. Palaeoclimat. Palaeoecol.* **127**, 285 (1996).
S19. A. M. Bush, M. J. Markey, C. R. Marshall, *Paleobiology* **30**, 666 (2004).
S20. J. W. Valentine, *Palaeontology* **12**, 684 (1969).
S21. J. J. Sepkoski, Jr., R. K. Bambach, D. M. Raup, J. W. Valentine, *Nature* **293**, 435 (1981).
S22. J. J. Sepkoski, Jr., *Paleobiology* **10**, 246 (1984).
S23. M. J. Benton, *Science* **268**, 52 (1995).
S24. J. J. Sepkoski, Jr., *J. Paleont.* **71**, 533 (1997).
S25. R. K. Bambach, *Geobios* **32**, 131 (1999).
S26. M. Foote, *Paleobiology*, **26** suppl., 74 (2000).
S27. D. M. Raup, *Bull. Carnegie Mus. Nat. Hist.* **13**, 85 (1979).
S28. D. M. Raup, *Science* **177**, 1065 (1972).
S29. D. M. Raup, *Paleobiology* **2**, 289 (1976).
S30. A. B. Smith, (2001). *Philos. Trans. Roy. Soc. Lond. B, Biol. Sci* **356**, 351 (2001).
S31. S. E. Peters, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 12326 (2005).
S32. S. E. Peters, *Paleobiology* **32**, 387 (2006).

S33. P. W. Signor, J. H. Lipps, *Geol. Soc. Am. Spec. Pap.* **190**, 291 (1982).
S34. R. K. Bambach, A. H. Knoll, J. J. Sepkoski, Jr., *Proc. Natl. Acad. Sci. U.S.A.* **99**, 6854 (2002).
S35. S. E. Peters, M. Foote, *Paleobiology* **27**, 583 (2001).
S36. J. S. Crampton, M. Foote, A. G. Beu, P. A. Maxwell, R. A. Cooper, I. Matcham, B. A. Marshall, C. M. Jones, *Paleobiology* **32**, 509 (2006).
S37. D. M. Raup, J. J. Sepkoski, Jr., *Science* **215**, 1501 (1982).
S38. M. Foote, *Paleobiology* **27**, 602 (2001).
S39. L. Cherns, V. P. Wright, *Geology* **28,** 791 (2000).
S40. P. Wright, L. Cherns, P. Hodges, *Geology* **31,** 211 (2003).
S41. A. K. Behrensmeyer, F. T. Fürsich, R. A. Gastaldo, S. M. Kidwell, M. A. Kosnik, M. Kowalewski, R. E. Plotnick, R. R. Rogers, J. Alroy, *Paleobiology* **31**, 607 (2005).
S42. J. S. Madin, J. Alroy, M. Aberhan, F. T. Fürsich, W. Kiessling, M. A. Kosnik, P. J. Wagner, *Science* **312**, 897 (2006).
S43. M. Kowalewski, A. P. Hoffmeister, *Palaios* **18**, 460 (2003).
S44. A. J. W. Hendy, Geol. Soc. Am. Abs. Prog. **37**, 117 (2005).
S45. L. J. Walker, B. H. Wilkinson, L. C. Ivany, *J. Geol.* **110**, 75 (2002).
S46. G. G. Simpson, in *Evolution after Darwin, Vol. I, The Evolution of Life*, S. Tax, Ed. (University of Chicago Press, Chicago, 1960), pp. 117-180.
S47. W. Kiessling, *Facies* **51,** 27 (2005).
S48. M. L. Reaka-Kudla, in *Biodiversity II: Understanding and Protecting our Biological Resources*, M. L. Reaka-Kudla, D. E. Wilson, E. O. Wilson, Eds. (Joseph Henri, New York, 1997), pp. 83-108.
S49. G. Rosenberg, *Malacolog 4.0: A Database of Western Atlantic Marine Mollusca [WWW Database (Version 4.0.2)]* URL http://data/acnatsci.org/wasp (2005).
S50. C. Patterson, A. B. Smith, *Nature* **330**, 248 (1987).
S51. P. J. Wagner, *Paleobiology* **21**, 410 (1995).
S52. J. M. Adrain, S. R. Westrop, *Science* **289**, 110 (2000).
S53. P. J. Wagner, M. Aberhan, A. Hendy, W. Kiessling, *Proc. Roy. Soc. Lond. B, Biol. Sci* **274**, 1471(2007).
S54. J. J. Sepkoski, Jr, *Paleobiology* **5**, 222 (1979).
S55. J. J. Sepkoski, Jr., A. I. Miller, in *Phanerozoic Diversity Patterns*, J. W. Valentine, Ed. (Princeton University Press, Princeton, 1985), pp. 153-190.
S56. M. A. Buzas, L. S. Collins, S. J. Culver, *Proc. Natl.. Acad. Sci. U.S.A.* **99**, 7841 (2002).
S57. K. Roy, D. Jablonski, J. W. Valentine, G. Rosenberg, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 3699 (1998).
S58. K. Roy, D. Jablonski, J. W. Valentine, *Proc. Royal Soc. Lond. Ser. B* **267**,

293 (2000).
S59. J. W. Valentine, D. Jablonski, S. Kidwell, K. Roy, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 6599 (2006).
S60. M. Foote, J. J. Sepkoski, Jr., *Nature* **398**, 415 (1999).