Supporting Online Material

Materials and Methods

Problems with diversity estimation

Paleontological diversity curves are subject to an array of sampling biases (S1). Such biases could not be addressed decisively using older literature-based compilations because they only provided age ranges (i.e., dates of first and last appearances) for individual families or genera. The most important of these compilations, although far from the only ones, were those of Sepkoski (S2-S4). The Paleobiology Database (PaleoDB: http://paleodb.org) makes it possible to remove various sampling overprints because it ties age ranges to occurrences of taxa within individual fossil collections from specific locations and stratigraphic horizons.

Variation through time in preservation quality can be dealt with simply by excluding unusually well-preserved samples (see below). Variation in the quantity of sampling is more problematic. Most of the previous literature focuses on randomized subsampling protocols that seek to hold each time interval to a uniform quota of data items (S5-S9). The quota may be expressed in terms of the number of taxonomic occurrences (classical rarefaction: S6), fossil collections (S8), occurrences represented by the collections (S5), or sum of per-collection occurrence counts each raised to a power (S7, S9).

Item quota methods make important assumptions about uniformity through time in the evenness of abundance distributions and/or the average number of specimens found in collections (S7, S9). Relaxing such assumptions requires obtaining a large amount of high-quality specimen count data for each time interval (S9).

A more serious and entirely unavoidable problem is that item quota methods effectively sample only those taxa with a certain relative frequency (i.e., those falling above a veil line: S10). For example, if a taxon has a frequency of 1% and 200 items are drawn, there is a 1 - (1 - 0.01)^200 = 87% chance it will be sampled. With a frequency of 0.35% this chance is still 50%, but with a frequency of 0.1% it is only 18%. Therefore, the effective veil line falls roughly between 0.1% and 1%. Another useful benchmark is that a frequency of 1.5% guarantees a 95% chance of being sampled.

To put these figures in perspective let us examine one of the largest collections in the PaleoDB with abundance data (number 26003, from the Silurian of Canada), which includes 12,315 specimens of 75 genera (Fig. S1A). Of that total, 14 have frequencies above 1.5%, 44 have frequencies below 0.35%, and 17 fall in between. Effectively, then, rarefaction using a quota of 200 specimens would draw almost all of the few dominant taxa, most of the few common ones, and a few of the many rare ones, with an expected total of 33.

A simple thought experiment shows why this outcome is problematic. Suppose there is a second sample in the immediately subsequent time interval in which all of the genera have survived and their relative abundance structure has not changed at all, but the true regional genus pool has doubled in size (Fig. S1B). That is, another 75 genera
with the same distribution of abundances are available for sampling, so all of the relative abundances are halved. Eight genera previously having frequencies over 0.35% now have frequencies below this point. As a result, we now expect to sample 25 or 26 of the original genera plus the same number of new ones. The expected total is therefore between 50 and 52, instead of the 66 an accurate method would have drawn.

This logic applies to all item quota methods, regardless of what items are counted: an increase in diversity will make it harder to sample the existing pool of taxa simply because the item quota and resulting frequency cutoff (veil line) are constant. Therefore, applying any such method to a diversity curve will flatten it. These methods only yield increasing estimates when diversity rises because there is usually a favorable tradeoff between the decreasing relative abundances and the increasing pool size.

Simulation analyses could be developed to make the same point, but it is clear that something is amiss: uniformly fair sampling is not the same as uniformly accurate sampling. In other words, drawing a fixed number of items is not the same as drawing a fixed proportion of the overall species pool.

**Shareholder quorum subsampling**

The subsampling method used here does not require use of abundance data, does not employ a uniform quota of items, and does seem to produce diversity estimates that closely track the species pool size. The method stems from the mathematical concept of frequency coverage (S11), which is the basis of some equations used by ecologists to produce extrapolative diversity estimates (S12, S13). Not only are these extrapolations known to be noisy, but they are generally biased either upwards or more often downwards depending on the size of the data set (S14). Thus, developing a subsampling approach that takes advantage of coverage instead of inventing yet another extrapolation index is a justifiable strategy.

The basic idea of this algorithm is to draw fossil collections and keep track of the summed frequencies of the taxa they include (i.e., the total coverage of the frequency distribution). These frequencies are akin to shares in a company, so the taxa can be called shareholders. Shares are credited only the very first time a taxon is drawn. The desired sampling level is set in terms of coverage, and when this level is reached the subsample can be said to represent a quorum. Thus, the method is called shareholder quorum subsampling (SQ).

It must be remembered that this method does not draw a fixed number of taxa, proportion of taxa, proportion of items, or (like an item quota method) number of taxa with a minimum frequency. Coverage has no direct tie to any of these figures.

Returning to collection 26003, the 33 most common taxa have frequencies adding up to just below 0.97, with the very last one having a frequency of 0.0028 (≈ 0.28%). Thus, a quorum target of 0.97 should on average give back the same number of taxa produced by rarefying the sample down to 200 specimens. This figure does not represent 97% of the genera. It does not represent anything close to 97% of the specimens because we know that drawing 200 specimens out of the 12,315 (1.6%) will on average yield 33 genera. Finally, it does not relate to a minimum frequency.
It is easier to understand what SQ actually does do by visualizing a rank-ordered occurrence frequency distribution such as the one for collection 26003 (Fig. S2A). SQ samples a fixed proportion of this curve’s area; that is all. Most of the common taxa and a stochastic assortment of the rare ones are drawn in any given trial (so the shading in the figure is for illustrative purposes only). As previously explained, methods using simple item quotas do much the same thing, but they use an inflexible veil line instead of responding to changes in the shape of the curve (Fig. S1A).

Keeping the sampled area constant usually means sampling a constant proportion of the entire taxon pool. So, if the pool were doubled as in our hypothetical case (Fig. S2B), the quorum could be reached by drawing the original 33 genera plus another set of the exact same size having the same absolute abundances. The difference is that now the 66th most common taxon would have a frequency of 0.14% instead of 0.28%. Thus, to get that last taxon we would have to draw more items (in this case, to get 66 taxa one would have to draw 393 specimens instead of 200).

Effectively, then, SQ is a method of producing accurate relative diversity estimates with fair but uneven sampling. It succeeds because it compensates for the fact that adding more taxa will cause the relative frequency of each existing taxon to drop. Its major assumption is that taxa are added randomly to the pool, or subtracted randomly from it, going from one time interval to the next. If so it will always draw a fair fraction of the pool size, even though the true pool size is by definition unknown.

Shareholder quorum level corrections

Three modifications to the SQ algorithm are necessary. They address substantively different factors that vary among time intervals: (1) the observed total coverage of the sampling pool's true frequency distribution; (2) the evenness of this distribution; and (3) the fact that individual fossil collections in small data sets can have wildly variable sizes. The first correction is crucial and the others are minor. *Correction for overall coverage.*—Essentially, the first problem has to do with the very fact that virtually no empirical data set includes all of the taxa it could. Therefore, any full data set's coverage of its sampling pool (called $u: S11$) is less than 100%. We need to standardize coverage in the first place because coverage is almost always partial and varies amongst time intervals.

The specific computational problem with incomplete coverage is that it causes the observed taxon frequencies to be overestimated. The reason is that there should also be counts for the missing taxa, albeit fractional ones, but those counts are all treated as zero. Therefore, when we calculate $f_i = n_i/O$ where $f_i$ is the frequency of some taxon and $n_i$ is its count, we divide by a sum that is too small and end up with an $f_i$ that is too big. The correction is now obvious: we multiply each $f_i$ by the fraction $u$. An alternative, algebraically equivalent algorithm is to divide the overall quorum level $q$ by $u$ before beginning to subsample each bin's data.

One way or another, to make any correction we first need to have a robust estimate of each temporal interval's total coverage. The most common estimator is Good's equation ($S11$): $u = 1 - o_1/O$ where $o_1$ = the number of taxa each sampled
exactly once (i.e., singletons) and \( O \) = the number of occurrences. This equation is expected to be accurate given even weak assumptions about distributions (\( S15 \)), and in practice it can be shown to be highly accurate by examining random subsamples of an individual empirical data set.

Unfortunately, \( u \) makes the strong assumption that larger data sets do not systematically sample larger sampling universes. Literature-based occurrence data sets do not meet this assumption because researchers focus their work on understudied time intervals, geographic regions, environments, and taxonomic groups. As a result, more and more databasing of the literature will expand the effective sampling universe. Thus, it will flatten the occurrence frequency distribution by adding taxa with very low frequencies at an unrealistically high rate.

This problem violates every diversity estimation method's assumption that occurrence frequencies are primarily controlled by biological factors. None of the existing ones take it into account, so all of them are invalid when applied to literature-based occurrence data sets. In the current data set the bias is so large that the value of \( u \) as defined by Good actually drops in each time interval as more and more published references are added to the data set.

Fortunately, an alternative and much more conservative equation does not have this empirical property, and therefore yields consistent diversity estimates regardless of how many references are consulted. It is \( u' = 1 - p_1/O = (O - p_1)/O \) where \( p_1 \) is the number of occurrences belonging to taxa that are only documented by a single published reference. Because single-publication taxa may be present in many collections instead of only one, generally \( p_1 > o_1 \) and therefore \( u' < u \). Because it uses larger counts, the \( u' \) or "single-publication occurrences" estimator is generally more precise than Good's in addition to being more conservative than Good's. Both equations asymptote on 1 as sample sizes grow infinite, but \( u' \) may do so very slowly. Use of \( u' \) makes it unnecessary to impose a reference count quota on the data (e.g., \( S9 \)), a practice that raises serious concerns because obtaining an accurate estimate when diversity is high may simply require consulting more references.

Evenness and dominance.--The second correction addresses a potential violation of the method's key assumption. If most taxa newly added to the sampling pool are more common than the average existing taxon, the frequency distribution will become less even. Conversely, adding rare taxa will make the distribution more even. This problem has large practical consequences. For example, adding one very common taxon could cause SQ to undersample badly because drawing one occurrence of this taxon would push the running frequency coverage total close to the quorum target.

Assuming random substitution is not very problematic for occurrence distributions because they tend to be so much more flat than specimen count distributions (simply because the taxa in any given collection are each credited with a single occurrence regardless of abundances). Nonetheless, the bias is real.

A partially successful solution is to simply ignore the most common taxon during all computations involving frequencies (it is still tallied if it is subsampled, of course). The relative frequency of this dominant taxon is called the Berger-Parker dominance index \( d \) (\( S16 \)). More formally, if \( n_i \) is the dominant taxon's count, then the corrected
coverage index equals \((O - p_1 - n_1)/(O - n_1)\).

This approach has an obvious biological justification: ignoring the most common taxon makes sense because so many subsamples include it that its presence is minimally informative. The correction’s nature also means that when evenness is high, SQ will sample harder to find the rare taxa. Similar results could be obtained by modifying \(u'\) with assorted dominance or evenness indices such as Pielou’s \(J\) or Shannon’s \(H\) (S16). However, this algorithm is as simple as possible and does not require using statistics that are not frequencies and so are not commensurate with \(u'\).

It is obvious that evenness cannot be completely excluded from taxon counts. However, the same criticism applies even more strongly to all item quota methods because they sample fewer and fewer taxa as the dominant one becomes more frequent, even if nothing else changes. Indeed, the very idea of trying to remove evenness from a diversity estimate sets this method apart from the others, which in some cases have been explicitly formulated to capture an evenness signal (e.g., calibrated weights subsampling: S9). Only a few researchers have previously recognized that if subsampling is aimed at diversity estimation, then evenness is a noise factor and not a signal (S17). Given SQ's strong responsiveness to the overall pool size, any remaining sensitivity to evenness is a price very much worth paying (Fig. S2 vs. Fig. S1).

**Single large collections.**--The third and least important problem involves data sets that include very few fossil collections and therefore could be dominated by an even smaller handful of large ones. Such data sets are likely to have artificially even frequency distributions that will be treated by \(u'\) (or by Good's \(u\)) as indicating very poor coverage. For example, if there is only one large collection it will yield numerous taxa-sampled-once almost by definition, suggesting that many more remain to be found even if the sample actually has captured most of the taxa that were on the landscape.

Because the presence of an exceptionally large collection will create the appearance of poor coverage, without any further correction SQ would tend to oversample. The simple and obvious solution is to exclude occurrences of taxa only ever found in the most diverse collection \(t_{\text{max}}\) from the count of single-publication occurrences, i.e., add back their occurrences to the numerator of the equation. The index therefore changes from \((O - p_1 - n_1)/(O - n_1)\) to \((O - p_1 - n_1 + t_{\text{max}})/(O - n_1)\).

**Extrapolation methods**

One must address the possibility that extrapolation methods also involving the idea of frequency coverage might be even more accurate. More specifically, the incidence coverage estimator of Chao and Lee (ICE: S12) is known to perform relatively well with empirical abundance data sets. It is equal to \(T + t_1^2/2t_2\) where \(T\) is the total number of sampled taxa and \(t_1\) and \(t_2\) are counts of singletons and doubletons. The index is premised on the idea that right-hand part of the expression correlates with the size of the unsampled species pool, and therefore that it decreases as sampling increases.

If this estimated missing pool size really does decrease at some point, then it
cannot be true that \( t_1; T \) (the fraction of singletons) is exactly constant. However, in the real data it is nearly so. That is, \( t_1 \) is a linear function of \( T \) and ICE therefore derives almost all of its information from raw diversity \((T)\).

This fact could be made easily with the current data set simply by plotting pairs of the three variables against each other. Speaking more formally, least-squares regressions of changes in the log of ICE and in the logs of \( t_1 \) and \( T \) produce intercepts that are not significantly different from zero and slopes (0.993 and 0.934) that are also very nearly 1. In other words, ICE is nothing more than \( T \) plus \( t_1 \) times a constant. Thus, the only difference between these numbers is that on average ICE is about 2.9 times higher than \( T \) and 3.4 times higher than \( t_1 \). Any diversity curve based on ICE would therefore simply tend to track the raw data.

With regard to \( t_2 \), if we factor out the importance of \( t_1 \) simply by correlating changes in \( \log(t_2/t_1) \) and \( \log(ICE) \) we find that \( t_2 \) contains no residual information \((\rho = -0.081, p = 0.585)\). Multiple regression would show the same thing.

We can also immediately see another important problem: \( ICE \sim 2.9 \ T \), which means that if ICE is accurate, \( T \) should about one-third of true richness. However, \( T \) is typically less than this amount except in data sets documenting trivial, small species pools. Therefore, ICE will normally be too low and will climb and climb as more information is acquired, in addition to not telling us anything more about relative changes in diversity than we can tell from the raw data. This underestimation bias has already been noted empirically \((S14)\).

ICE is believed to be the most accurate of all the nonparametric methods because all of the other ones are even more downwards-biased \((S14)\), but these results suggest it is effectively uninformative. An extremely important implication follows: extrapolation methods in general appear not to work even in the case of this extraordinarily large paleontological data set. In the real world we would expect very few interesting data sets to be as well-sampled as this one, meaning that still fewer are adequate even for an ICE computation. For a hard problem such as reconstructing a global diversity curve at the genus level from fossil data, ICE and by extension all other extrapolation methods should never be used.

**Marine diversity curve analysis**

The data set consisted of 340,787 occurrences of non-tetrapod metazoan animal genera within 54,674 fossil collections. The occurrences were downloaded from the Paleobiology Database on 13 March 2010 using the almost the same sifting criteria as in two recent studies \((S9, S18)\): collections were excluded if they represented large spatial or stratigraphic scales, while occurrences were excluded if they represented form taxa, ichnofossils, or taxa not identified to down to the level of a well-defined genus. Subgenera were treated as distinct, and multiple species of the same genus in a single collection were treated as a single occurrence.

The time scale consisted of 48 temporal intervals equating to geological stages or sets of short neighboring stages \((S9, S18)\). This count excludes the earliest Cambrian (before the traditional Early Cambrian), which cannot be analyzed because...
very few fossil collections of that age have been reported. A total of 49,921 collections and 314,141 occurrences had precise enough age assignments to be fit into a single bin and thereby contribute to the diversity estimates. These collections included 20,181 genera, of which 3362 are extant.

Preservation has a modest but biologically important effect on diversity estimates in this data set because it tends to amplify other biases favoring the Cenozoic in general and the Neogene in particular (S9). Key factors include preservation of original aragonite (S19), lithification of enclosing sediments (S20, S21), and the spatial distribution of collections (S9). The aragonite and lithification biases strongly favor the Cenozoic and become severe in the Neogene. These factors are held constant by excluding the relatively few fossil collections that are marked as preserving original aragonite, coming from un lithified sediments, or coming from poorly lithified sediments that were sieved (S9, S18). Samples including compression fossils that retain soft parts also were omitted in the present study.

The target for shareholder quorum subsampling was set to 0.60 for the combined data set and to 0.50 for each of the analyses focused on particular taxonomic groups. All bins had coverage of at least 0.72, but curves with targets ranging from 0.30 to 0.80 have almost identical shapes. Thus, the threshold of 0.60 was chosen simply to make sure the resulting overall curve was comparable in height to the summation of group curves. There were 100 subsampling trials per analysis, enough to remove stochastic variation in a data set this large. Contrary to an earlier study (2008), there was no attempt to standardize the number of publications used in the data set because a strong reference count bias is not demonstrable (see below).

Taxon counts were of genera represented by at least one drawn occurrence (sampled-in-bin diversity) based on evidence that this kind of a count is the least subject to biases such as edge effects (S9). Older data sets such as Sepkoski's (S3, S4) only note first and last appearances, so they do not record which taxa are sampled in which bins and therefore must be counted with inferior, range-based methods (S8).

The sampled-in-bin counts were adjusted with the three-timer correction (S9, S18), i.e., they were rescaled using the ratio 3T/(3T + PT), where 3T = the number of genera sampled immediately before, during, and after a focal interval (three timers) and PT = the number of genera sampled immediately before and after but not during the interval (part timers). Specifically, diversity of each interval i was first divided by 3T_i/(3T_i + PT_i) and then multiplied by 3T/(3T + PT) where 3T and PT were summed over the entire data set (S18). The median interval was represented by 942 collections and 751.5 genera in the raw data and by 86.6 collections and 256.6 genera in the subsampled data.

The major taxonomic groups were defined following Sepkoski’s classification (S22, Table S1), with those represented in the database by at least 2800 occurrences being analyzed. The remaining, miscellaneous taxa were treated together. None of them are represented by more than 2000 occurrences except for Demospongea. A separate curve for this group was examined but could not be analyzed because it lacks meaningful structure: variance is extremely high, neighboring points are uncorrelated, and there is no temporal trend. Adding occurrences for the two other sponge groups
(Sclerospongia and Hexactinellida) has no visible impact on this curve. Meanwhile, Bryozoa was considered as a unit because its constituent classes Stenolaemata and Gymnolaemata did not meet the quota. Sepkoski’s "Inarticulata" was equated with Linguliformea, and his "Articulata" with the remaining Brachiopoda. Within vertebrates, Conodonta and Chondrichthyes met the minimum sampling requirement but all other non-tetrapod vertebrates ("fishes") were consigned to the miscellaneous category.

The major taxonomic groups were broken out for analysis in two different ways. First, each group was analyzed separately using the above-mentioned quorum target of 0.50. This value was high enough to result in a summed curve that closely matched the one produced from the combined data with a quorum of 0.60. Second, each group’s genera were tallied during each trial of subsampling from the combined data set.

Three-timer corrections had to be handled differently in each case. For counts extracted from the combined data set, the three-timer sampling statistic for the entire data set in each bin was applied. For the separate analyses, stochastic error in the 3T and PT counts made it unrealistic to apply the correction unless sample sizes were large. Therefore, the correction was applied only if the subsampled data for a particular bin included an average of at least five three timers. The subsampled but uncorrected data were used otherwise. If the shareholder quorum could not be met in a bin at all, the raw count of genera was substituted. Although obviously a minimum in that case, the numbers at least cannot be lower than the analogous ones produced in the overall analysis.

**Combined vs. separate analyses**

The overall shape of the diversity curve did not depend strongly on whether it was based on combined data or was a summation of curves produced by separate analyses of the taxonomic groups (Fig. S3). Both versions capture a prolonged Cambro-Ordovician radiation; a mid-Devonian plunge; a mid-Permian rebound; a Permo-Triassic mass extinction and recovery; a major Triassic-Jurassic extinction followed by a delayed and proportionately very large mid-Jurassic radiation; and high Cenozoic values that fall far short of what some literature predicts (e.g., S23) but are not out of line with what Sepkoski himself thought might be realistic (S3). With the combined analysis as a benchmark, salient features of the summed version include blunting of the Devonian peak; higher values throughout the mid-Devonian though early Permian trough; and lower late Jurassic and Paleogene values. The latter result from lower estimates for gastropods.

Surprisingly, there is no more apparent noise in the summation: it produces a wobble index of bin-by-bin variation (S18) of 0.192, as opposed to 0.241 in the combined analysis. This tiny difference is not significant according to a Wilcoxon signed rank test or a Mann-Whitney test. The summation presents a lower nonparametric serial correlation (Spearman’s ρ = 0.709 vs. 0.764) but the discrepancy is again not large. Nonetheless, the summed curve is visibly flatter because the differences outlined above, and indeed it has a substantially lower standard deviation on a log scale (0.335 vs. 0.440).
As noted, although the exact choice of targets generally has little effect on curve shapes, in this case it was necessary to use a higher quorum threshold for the combined data in order to match the magnitudes of the two curves in most places. This fact raises several related questions. Why does it matter whether the data are combined? Is the 0.50 figure commensurate amongst all data sets? And why hold it constant across the groups if combining them one by one would eventually make it necessary to raise the target?

The answer is that because the relative total diversity of any one interval is not really a function of whether the data are combined (Fig. S3), what really changes is the balance of the groups. In a combined analysis, groups with generally abundant and widespread genera are sampled first. Therefore, groups with typically rare and restricted genera will be underrepresented badly unless the threshold is set higher. So, the very fact that groups differ systematically in average abundance justifies splitting the data. If they were equally abundant, by definition the data set's overall quorum target would be commensurate for all them.

Arguably, then, the only sensible approach is to break out smaller and smaller groups until doing so does not change the shape or magnitude of their individual curves (or until it becomes impossible to meet reasonable sampling targets). It is only by doing so that we are able to untangle abundance and breadth of distribution, which we do not care about here, from taxon pool size, which is what the curves are supposed to represent. Thus, although the taxonomic group definitions employed in this study are fairly narrow and do a good job of matching those used in the original evolutionary fauna analyses (S2, S22), future analyses may benefit from using even narrower definitions.

This matter is particularly important with respect to a very important difference between the curves (Fig. S3): only the summed analysis produces something that resembles a steady Cenozoic radiation (although certainly not an exponential one). Because the discrepancy involves Paleogene and not Neogene estimates, the issue is whether the combined analysis oversamples the Paleogene or the summed curve undersamples it.

There are large differences between the analyses in the relative diversity of major groups. The separate bivalve and gastropod curves (Fig. 1) are much lower than the combined data set's versions because these taxa are very abundant, so in the combined analysis they swamp out everything else and appear to have high absolute and relative diversity. If it is true that the remaining Paleogene taxa are both rare and not very diverse, as it certainly appears, then the combined figures are more likely to be wrong.

Even if we accept this hypothesis, it is still possible that although the Paleogene figures are fair in the summed curve (Fig. 3), the Neogene figures remain distorted by preservational or geographic biases that have not yet been removed. Geographic coverage in the Neogene is truly extraordinary, with very good representation of tropical, southern, and oceanic faunas that is simply not possible in the Paleogene. The data set also quite likely still includes numerous Neogene collections that do preserve aragonite and/or come from unlithified sediments, but could not be identified as such given the available published information. Thus, it is possible that an entirely accurate curve
would be flat through the Cenozoic (as in the combined analysis) but not nearly as high (as with the Paleogene values in the summed curve).

**Residual sampling biases**

Shareholder quorum subsampling removes any statistically significant cross-correlation between changes in logged diversity (Fig. 3) and changes in either logged collection or logged occurrence counts (S9, figures S1, S14A). In other words, by any reasonable definition it does remove a gross sample size effect from the data.

A variety of item quota methods performed exactly the same way when applied to an earlier version of the data set (S9). However, they retained signals of two interrelated time series: the number of published references yielding the collections in each bin, and the number of geographic cells including collections in each bin (S9). These patterns were somewhat ambiguous because it is hard to say (for example) whether cell counts are a reflection of original marine shelf area or simply demonstrate variable geographic dispersion in sampling.

SQ does not remove these connections. For example, in the summed analysis (Fig. 3) there are is a strong cross-correlations of logged, differenced diversity with differences of logged reference counts (Spearman’s $\rho = 0.483$, $p < 0.001$). A weaker but still significant correlation also still exists in the curve produced from the combined data set (Fig. S3: $\rho = 0.350$; $p = 0.016$).

Because reference and cell count effects were considered to be a major problem in the 2008 study (S9), the number of references used in the analysis was limited. Specifically, in each subsampling trial a fixed number of references was drawn, the collections in those references were identified, and subsampling was performed only on those collections. This algorithm actually failed to remove the correlations despite pushing down the curve in most intervals, most visibly the Neogene (S9, figure S13).

However, SQ renders such measures unnecessary. Imposing a quota of 69 references per bin (the highest possible value in this case) has almost no effect on the curve produced from the combined data: the "limited" version is actually bit higher throughout much of the early Paleozoic and mid-Mesozoic; the offset is small (median log ratio = 0.063); and the logged and differenced values are very strongly correlated ($\rho = 0.968$; $p < 0.001$). Indeed, the points overlap almost completely throughout the Late Cretaceous and Cenozoic. And again, the correlation between changes in diversity and reference counts remains after this restriction ($\rho = 0.300$; $p = 0.041$).

Together, the evidence now strongly suggests that (1) the 2008 curve suffered from substantial sampling pool biases whereas the current curve does not because of the way it compensates for the scope of the raw data set (i.e., by examining the number of single-publication taxa); and (2) subsampled diversity rises when more publications have been put in databases exactly because researchers focus their efforts more strongly on legitimately diverse time intervals. It remains possible that the geographic cell count data capture information about habitat area and that shelf area has a major effect on global marine diversity (S24), although this is a topic for another paper.
Effect of turnover rates on diversity estimates

The potential distortion of diversity estimates by variable turnover rates is a persistent worry in the paleobiological literature (e.g., S1, S7, S9). Essentially, three different factors might produce biases. First, more and more species are added to an interval's overall sampling pool as diversification continues through it, so one might expect long intervals to have higher sampled-in-bin counts even if standing diversity is the same (S1). This problem by itself is unlikely to be serious because the interval definitions used here have been engineered to have roughly uniform durations (S9). Second, for a mathematically identical reason intervals with higher background turnover rates should pile up more species (S25). Finally, as turnover rates increase average durations decrease (S26), which makes it harder to sample individual taxa. This problem is obviously nonexistent for any taxon whose range entirely spans a particular time interval, but it becomes relevant for intervals capturing the first or last appearance of a taxon (particularly if both events fall in the same interval).

It has been suggested that in some data sets the third bias cancels out the others (S9), in which case sampled-in-bin counts should be robust. However, that argument only applies to subsampling methods whose goal is to make sampling intensity uniform. The reason for the cancellation is precisely the fact that item quota methods are biased: when sampling is uniform and diversity is high, subsamples will underrepresent the species pool.

In contrast, SQ seeks to draw a fixed proportion of the species pool, so it will sample harder when many taxa are each represented by a few occurrences. This situation is expected when intervals are long and/or turnover rates are high. Therefore, cancelling out should not occur when SQ is used.

None of these worries matter at all unless turnover occurs continually within time intervals, as opposed to being clustered at the boundaries between time intervals. That is, if all new taxa originate exactly at the base of an interval and all doomed taxa go out at the top, then all taxa crossing into that interval have the same number of sampling opportunities and all of them really did coexist.

There is some quantitative evidence that events do tend to cluster, at least in Phanerozoic marine data sets (S18, S27). Furthermore, extinctions must cluster in intervals such as the Maastrichtian that are terminated by catastrophes. Nonetheless, it is informative to ask just how badly the diversity estimates would be biased if all turnover was continuous going through each and every time interval. It can be shown with simple simulations that:

1. SQ is always unbiased when all turnover is at boundaries (as is intuitive).
2. When turnover is entirely continuous, the pileup of sampled taxa in intervals with rapid turnover is somewhat weaker at low sampling levels. This mild bias suggests that SQ's is unable to fairly sample rare taxa (here, short-ranging ones) when data sets are simply poor.
3. A large fraction of the rate bias can be removed by correcting the taxon counts with the equation $N' = N + 2T - 3T$, where $N =$ the number of sampled taxa, $2T =$ the
geometric mean number of two-timers crossing (a) the interval base and (b) the interval top, and $3T =$ the number of three-timers. $2T$ and $3T$ counts are intimately related to turnover rates, with both counts rising as turnover slows but $3T$ counts rising more quickly in relative terms ($S18$).

(4) This correction reduces error to few percentage points under a broad range of turnover rates and sampling targets. However, estimates tend to rise slightly when turnover rates are unusually high or low, and they can be off by more than about 10% when turnover rates are high and sampling is very poor.

Based on these observations, we can predict that if there is a rate bias problem in a real data set, counts corrected with the $2T - 3T$ equation should differ systematically from raw counts (because they are more accurate). When turnover is slow the corrected counts should be much higher, and when it is rapid they should be about the same (because hardly any taxa should fall in the $2T$ and $3T$ categories).

As it happens, there is a nearly constant offset (Fig. S4A) between a curve for the combined data set (Fig. S3) and the corrected version of that same curve. The median log ratio between the latter and former is 0.120. There is no trend through time in this ratio, as measured by its correlation with sequential interval numbers ($\rho = -0.016; p = 0.914$). The ratios in the early and mid-Paleozoic are particularly consistent despite considerable turnover rate variation in this part of the time scale ($S18$), which includes some very high Cambrian values that would be expected to create a problem. Even the large mass extinctions at the end of the Permian, Triassic, and Cretaceous seem not to change the offset. Meanwhile, apparent outliers in the Cenozoic are an illusion created by the linear scaling of the Y-axis, because the log ratio in this era is always less than 1.19.

Any large rate bias would most likely affect one group in particular: the trilobites. The reason is that they are the single most important constituent of the Cambrian and Ordovician biota (Figs. 1A, 3) and turnover rates are extraordinarily high in that exact interval ($S18$). However, a trilobite curve adjusted with its own $2T$ and $3T$ counts differs only marginally from the raw trend (Fig. S4B). Unlike the overall pattern, here the early Paleozoic values change very little, which does suggest a rate bias. The point, however, is that the trend as a whole is qualitatively the same. Furthermore the difference in absolute terms is far too small to account for the very dramatic contrast between sampling-standardized, sampled-in-bin global curves (Fig. 3, S9) and more traditional curves based on simple age range compilations ($S2, S3$), which often depict very low Cambrian diversity.

These exercises do not show that a rate bias is absent in the overall data set – only that any such bias is likely to be weak, because if there had been a real problem this correction would have visibly changed the curve’s shape. However, it is still hard to believe there is no bias because the data are so very complex. Based on the above-mentioned simulation results, there are two possible explanations.

First, it could be that sampling is still extraordinarily poor even when the quorum target is 0.60 (as in Fig. S4A) because a vast number of rare genera (albeit preservable ones) are not present anywhere in the fossil record. If so, the pattern could be
explained by the simulation results suggesting that SQ has a weak bias against finding a pileup when sampling is poor. This scenario seems implausible because SQ's rate bias is negligible in simulations using this quorum target, and because SQ yields almost the same empirical curve over a broad range of targets.

Second, because rate biases only matter when turnover is continuous, it could be that turnover in this particular data set tends to be concentrated at interval boundaries, consistent with other evidence \((S18, S27)\). Obviously, there must be some turnover in almost every interval because none of the intervals are shorter than a geological stage \((S9)\) and every stage can be broken down into (say) multiple conodont, graptolite, or ammonite zones. Even if there is, this study's use of SQ combined with sampled-in-bin counts does seem justified even in the face of a hypothetical rate bias.

**Comparison with published diversity curves**

The new diversity curve differs greatly from all published alternatives, including various treatments of Sepkoski's family- and genus-level data \((S2, S3, S23, S28)\) and a recently published curve that employs much the same data set as in this paper but a radically different sampling standardization method (calibrated weights or CW: \(S9\)). Both new curves agree on major features such as a Cambrian radiation that was so fast the curve starts out a fairly high level; a barely visible net drop after the end-Ordovician extinction; peak early Paleozoic diversity in the Early Devonian; an abrupt drop in the mid-Devonian followed by a long plateau; significant mid-Permian and mid-Jurassic radiations; no certain increase within the Cenozoic, much less a very large one; and, mostly important, a relatively modest difference between peak early Paleozoic diversity and Cenozoic diversity.

However, the SQ curve is different in important ways. First, it does not pick up a cross-correlation between changes in logged diversity and changes in running average abundance evenness in the earlier data set \((\rho = -0.050; p = 0.687)\). Second, it has a different gestalt. It suggests a steeper overall Cambro-Ordovician radiation; an even larger mid-Devonian drop; a smaller mid-Permian climb and Permo-Triassic net drop; and, most visibly, Neogene values that are about 50% higher on average than Cretaceous values. This difference is easily attributed to the extraordinary Late Cretaceous radiation of gastropods (Fig. 1A) combined with outlying Neogene values for this group and Bivalvia (Fig. 1B) that might not be real. Regardless, it does not create a large Paleozoic-Cenozoic differential exactly because the early Paleozoic radiation also goes to a higher level in this analysis.

The new curve is actually much closer to those produced by methods drawing fixed quotas of collections \((S9, \text{Fig. S8D})\) or of collections weighted by occurrences \((S9, \text{Fig. S8A})\) than to the CW curve \((S9, \text{Fig. 1})\). The first of these curves shows the same overall amplitude, and in particular some high Cenozoic figures, but differs by (for example) suggesting a dubious peak at the curve's very first point and a fairly steep downwards trend through the Carboniferous. The second curve omits the Cenozoic increase but produces many additional fine-scale features that are otherwise seen only in the new treatment (Fig. 3).
Briefly, the reason CW stands so far apart is that its estimates of available data quantity are forced to match changes in collection-level specimen abundance evenness (S9, Fig. 2). Therefore, its sampling intensity pattern covaries with this local-scale ecological factor. Like all earlier methods, CW seeks to standardize the sheer volume of data items sampled in each time interval: the only real difference is this tracking.

Thus, for example, because early Paleozoic evenness is low CW samples this interval lightly; and because there is no large change in evenness between the Cretaceous and Cenozoic, it samples both intervals with the same intensity. By contrast, the other methods seem to pay more attention to high beta diversity at both ends of the curve. CW and all other item quota methods are not well designed to pick up such variation (Fig. S1), although collection count standardization seems to do a good job (S9, figure S8D). In sum, the results are complementary: the CW curve best measures local-scale evenness, and possibly local-scale richness, whereas the SQ curve more strongly measures global species pool sizes.

Multivariate analyses of new diversity data

Factor and cluster analysis was used to confirm that the new data capture the three evolutionary faunas pattern when treated the same way as in the original study (S22). Factor analysis of the 14 individual diversity curves (Fig. 2A) was performed using the R function factanal. Varimax rotation was used in order to match Sepkoski's approach (S22). Columns were taxonomic groups, rows were time intervals, and cell values were SQ diversity estimates. In his family-level study (S22) Sepkoski treated his time intervals as variables because he had more taxonomic groups (91) than intervals (81), whereas in this study there are more intervals (48) than groups (14). Thus, the data shown in Fig. 2A are loadings instead of scores.

The first four factors explain a significant amount of variation according to a chi-square test ($p = 0.039$). However, the fourth and fifth are not biologically informative, because they mostly serve to pull out individual groups such as Cephalopoda and Trilobita. Almost all the useful information is captured in a three-factor solution, so that is discussed here in detail. The two-factor solution is illustrated because it shows most of the salient patterns (Fig. 2A)

These first three factors respectively explain 28.5%, 21.3%, and 14.3% of the variance. Factor 1 contrasts the Cambrian and Modern faunas; all four Cambrian groups fall on one side, and most Modern groups plus Anthozoa fall on the other. Here Cephalopoda is intermediate between the Paleozoic and Modern faunas. Factor 2 isolates Cephalopoda while strengthening the Modern fauna association. Factor 3 sets the Modern fauna against the three major Paleozoic fauna groups (Ostracoda, "articulates," and Crinoidea). All groups are pulled out on at least one axis.

A complete linkage cluster analysis of factor 1 loadings taken from the two-factor solution (used for comparability with Fig. 2A) confirms that its strong signal is identification of the Cambrian fauna (Fig. S5A). Removing these four groups and clustering the remaining factor 2 loadings likewise reproduces a Paleozoic vs. Modern split, although Ostracoda moves to the Modern fauna (Fig. S5B).
Factor 1's scores are extremely high in the Cambrian and Ordovician and then trail off. It is therefore most similar to the original third factor (called factor I by Sepkoski because of its temporal signal: S22). Factor 2 scores pull out the six Cenozoic intervals and also have a modest peak in the Ordovician through early Devonian. This axis does not have a clear counterpart in the original analysis. Factor 3 scores are very high through the middle and late Paleozoic and fall sharply at the Permo-Triassic boundary. Thus, it somewhat resembles Sepkoski's first and second factors (called by him factors III and II: S22), which both showed large Permo-Triassic transitions. The difference is that this factor 3 and his factor II (but not I) tend to group the Cambrian with the Mesozoic.

Diversity estimates were not log transformed, both for consistency with Sepkoski's analysis (S22) and to avoid having low counts dominate the results. Regardless, similar patterns were produced by factor analyses employing log(N+1) or sqrt(N) data where N = diversity. Similar results also were obtained with principal components analysis (PCA), principal coordinates analysis (PCoA), or correspondence analysis (CA) of the raw values. For example, PCA's axis 2 also isolated Cephalopoda, and its axis 3 also isolated "articulates" and Crinoidea. PCoA and CA pulled out the main Paleozoic groups on both axes 2 and 3. These methods set the Paleozoic fauna against Trilobita and Gastropoda on axis 2 and against Linguliformea, Echinoidea, and Cephalopoda on axis 3. In other words, all of the methods confirm the general outlines of the three fauna scheme while emphasizing the peculiar behavior of Cephalopoda.

Multivariate analyses of differenced diversity data

The goal of ordinating a matrix of differenced diversity curves (Fig. 2B) was to summarize similarities between taxonomic groups in dynamic behavior. The same first differences were explained by individual time series models (see below). As in those analyses, the curves were log transformed because differences of logged taxon counts are equivalent to net turnover rates (S7, S29). Factor analysis could not be used because zero values yield undefined ratios. Treating undefined values as (say) zeroes would severely distort the results, especially for groups such as trilobites that have been entirely extinct for hundreds of millions of years. They also capture no genuine information about dynamics. Therefore, PCoA of a Euclidean distance matrix constructed from the log ratios by ignoring the undefined values was performed instead. This method is highly similar to PCA, and because factor analysis and PCA are closely related it yields the best possible match to Sepkoski's protocol (S22).

The first two axes individually explain 20.0% and 17.6% of the variance. Axis 3 is uninformative because it simply contrasts the two brachiopod groups (Linguliformea and "articulates") with everything else. Axis 4 contrasts Echinodermata with Bryozoa and Crinoidea, which again is hard to interpret. As with the factor analysis, complete linkage cluster analysis of the axis 1 coordinates identifies the Cambrian fauna (Fig. S6A). The only difference is that Ostracoda is also included in this grouping. Clustering of the axis 2 coordinates for the remaining nine groups fails to produce anything resembling a Paleozoic vs. Modern split, with representatives of each fauna appearing in each of the
three main clusters (Fig. S6B).

There is a strong suggestion in the data that the six major groups cannot be
sorted cleanly into three clusters that equate to evolutionary faunas. The Cambrian
fauna includes only one major group to start with (Trilobita) and the Paleozoic fauna
only unambiguously includes two (Anthozoa and "articulates"). These latter two groups
simply do not hold together (Figs. 2B, S6). Furthermore, to create a Modern fauna out
of these patterns one would have to shuffle too many groups to do justice to Sepkoski's
definitions. It is more fair to say that the six groups occupy five almost evenly spaced
locations (Fig. 2B).

A bootstrap analysis emphasizes this point by showing that the lack of clustering
is not merely a product of noise in the data. Each bootstrap trial involves constructing a
matrix equal in size to the original one with exactly the same columns but with rows
drawn randomly with replacement from the real set. The matrix is subjected to PCoA
and the resulting coordinates are subjected to a Procrustes rotation with the original
ones as the target.

A set of 100 trials (Fig. S7) was sufficient to show that the six groups mostly do
sort out from each other, with only the "articulate" data points (in orange) overlapping
badly with points representing other groups (most often Gastropoda or Bivalvia). Again,
there is no evidence that Anthozoa should be associated with "articulates" because their
bootstrapped coordinates are adjacent but almost non-overlapping. Additionally, the
two traditional "Modern fauna" groups (Gastropoda and Bivalvia) almost evenly split one
large region of the plot and show little overlap, and a similar pattern is shown by the two
of the three "Paleozoic fauna" groups (Anthozoa and Cephalopoda).

**Multivariate analysis of Sepkoski’s raw data**

The most recent version of Sepkoski’s genus-level compendium (S3, S4) was
analyzed to confirm that the patterns seen in his family-level data (S2) also are robust at
lower taxonomic levels. The genus compendium’s time scale is broken into 75 stages,
and it includes enough data to separately analyze the 23 largest classes that Sepkoski
recognized. Because Sepkoski’s compendium cannot be sampling-standardized its
counts are much larger, so in this reanalysis Bryozoa was split into Stenolaemata and
Gymnolaemata and nine other, smaller groups including the three sponge clades and
"Osteichthyces" (effectively Actinopterygii) were added.

The fact that it is not possible to derive sampled-in-bin counts from the
compendium’s age ranges is again an issue here. All counting protocols involving
ranges are biased by various edge effects (S9), which are particularly severe for the
boundary-crosser counts (S5) that have been applied to Sepkoski’s data by some
authors (e.g., S23, S28). The best of these bad options is the most venerable and
commonly used method, which is simply to count all taxa first or last appearing within a
time interval or found both before and after it (i.e., range-through counts). Sepkoski
himself used such counts but discarded genera found in only one interval (singletons),
thinking that might decrease sampling biases (S3). However, it can be shown easily by
simulation that this protocol actually increases edge effects, so it is not used here.
A three-axis factor analysis of the range-through counts (Fig. S8A) explains 70.9% of the variance, as opposed to the 64.1% produced by the analogous ordination of the new data. The larger figure might relate to the larger sample sizes in the Sepkoski's compendium (S4) or, most likely, to differing counting methods. The range-through counts that had to be used because of the nature of Sepkoski's data effectively smooth the data over a window of several time intervals, which removes short-term biological signals (S9). By contrast, factor analysis of the new data was based on sampled-in-bin counts that show substantial bin-to-bin variance (Fig. 1).

The factor loadings again distinguish Sepkoski's three evolutionary faunas (Fig. S8A). Based on complete linkage cluster analysis, the first factor splits out the Modern fauna and the second distinguishes the Cambrian and Paleozoic faunas. In Sepkoski's data set the Pull of the Recent (S9, S30) imposed by range-through counts makes the "Modern" taxa seem to increase exponentially. This bias is responsible for burying the Cambrian vs. Paleozoic signal on factor 2 instead of highlighting it on factor 1, as is seen in the new data (Fig. 2A).

Most of the "misclassified" taxa are minor and combined with others in the new data set. Polychaeta, originally the second-most distinctive Cambrian group but in any case not an important one (S22), here clusters with the Modern fauna. The bryozoan subgroups Stenolaemata and Gymnolaemata are respectively split between the Paleozoic and Modern faunas, as previously shown by the family-level data (S22). Finally, the sponge groups Hexactinellida and especially Demospongea fall closer to the Paleozoic fauna than to the Modern fauna. As expected, Tergomya and Hyolitha join the Cambrian fauna, and Malacostraca, "Somasteroidea" (= Asterozoa), and "Osteichthyes" join the Modern fauna.

Sepkoski believed his first three factors to be important on the basis of a scree plot (S2). However, not only three- but four- or five-factor analyses of his genus-level data, which would respectively bring the variance explained up to 70.9, 78.3%, or 83.9%, also would have been significant according to a standard chi-square test. These additional factors would provide little biological clarity: three factors would merely place each fauna on its own axis; four factors would further distinguish the Cambrian fauna from a variety of low-diversity groups such as Hexactinellida, Demospongea, Polychaeta, and Asterozoa; and five factors would contrast the two smallest Cambrian groups (Tergomya and Hyolitha) with Cephalopoda.

**Multivariate analysis of Sepkoski's differenced data**

In a principal coordinates analyses of changes in logged diversity counts generated from Sepkoski's data set, the first two axes explain 48.6% of the variation (Fig. S8B). The analogous figure produced with the same treatment of the new data set (Fig. 2B) is 37.5%.

However, despite already explaining 37.5% of the variance by itself, axis 1 is much harder to interpret than in any of the previous ordinations. Its high end splits off four or five important Cambrian groups, but they are not far removed from a cluster of minor, mostly Modern taxa that also includes Crinoidea and Cephalopoda. However,
the other end of the axis brings together about half of the Paleozoic and Modern taxa, including some major ones. This second large cluster includes both Anthozoa and "articulate" Brachiopoda (from the Paleozoic fauna) and Gastropoda and Bivalvia (from the Modern fauna).

Axis 2 is much less informative than the last analysis' factor 2, explaining only 11.1% of the variance. It does open up a gulf between the "major taxa" cluster from the "mostly Modern, mostly minor" cluster, but its most prominent feature is a distantly placed point that represents Linguliformea (Fig. S8B). Just as with axis 1, there is no simple relationship between the curve shapes and coordinates: groups with strong Pull of the Recent patterns such as Gastropoda, Bivalvia, Gymnolaemata, Malacostraca, and "Osteichthyes" fall everywhere in the plot.

Lack of resolution in this analysis may relate to lack of precision and accuracy in Sepkoski's data. There is a strong relationship between axis 1 scores and either geometric mean or median diversity levels ($\rho = -0.565, -0.531; p = 0.006, 0.009$). The same relationship appears on axis 2, although it is weaker ($\rho = -0.469, -0.348; p = 0.025, 0.104$). Logging, differencing, and centering the data in principle should have removed any simple bias related to absolute magnitude. Therefore, the only obvious explanation is that large groups that responded jointly to evolutionary forcing factors have been sifted out from small groups whose responses are masked by noise in the counts.

It does remain possible that large and small groups have legitimately different diversity dynamics. Full resolution of this problem would require expanding the Paleobiology Database to include more information on the minor groups that for now can only be split out in Sepkoski's data set. In any case, there are so many problems with the older tabulations, such as the Pull of the Recent (S9, S30), the lack of sampling standardization, and the need to use range-based counting methods, that it would be more profitable to put them aside and focus on acquiring more collection-based data.

**Dynamic model fitting methods**

Sepkoski called his coupled logistic model "kinetic" instead of "dynamic" because "ecologists are most often interested in the outcomes of the interactions, the final equilibrium points, whereas the analysis in this paper has focused on the kinetics of the interaction, the trajectories of the two phases through time" (S34, p. 244). In other words, although the two things are intertwined and the distinction is subtle, Sepkoski was primarily concerned with patterns (curve shapes) instead of processes (model structures and parameters). Here the goal is not to explain variance in the shapes of curves, but to see if changes in these curves are better explained by some hypothesized processes than others.

Models with differing numbers of parameters are only easily compared using a maximum likelihood approach, so the usual sample-size correction version of Akaike's information criterion (AICc: S31) is employed. Models are compared by computing Akaike weights (S31), which are resemble Bayesian posterior probabilities and are very conservative compared to (say) frequentist $p$-values. It is assumed that the diversity
curves represent pure Markov processes, meaning that any change $\partial_i$ between two observed counts $n_i$ and $n_{i+1}$ is a function only of $n_i$. The fit of a model is determined by seeing whether it consistently predicts these observed $\partial_i$ values with high probabilities.

Computing likelihoods in this case is somewhat problematic because the underlying processes of speciation and extinction are fundamentally multiplicative and diversity curves are therefore best viewed on a log scale, but the diversity counts themselves have errors that are only close to normally distributed on a square root scale ($S_{32}$). Therefore, a nonparametric approach is required. The one taken here is to assume that however the changes might be distributed, we can determine the probability of observing $\partial_{i+1}$ given a predicted value $E(\partial_i)$ by seeing if $E(\partial_i)$ is closer to $\partial_i$ than $\partial_i$ is to the other $\partial$ values. In other words, we compare $E(\partial_i) - \partial_i$ to $\partial_j - \partial_i$ for all values $j$ in the time series where $i$ differs from $j$.

Absolute values of the differences are taken because it does not matter whether the predictions are specifically too high or too low. The likelihood is then the percentile score of $\text{abs}(E(\partial_i) - \partial_i)$ within the empirical distribution. For example, if there are 12 diversity counts, then there are 11 changes and we compare each predicted-observed difference of changes to the 10 observed-observed differences. Values of zero and infinity are used to bracket the rank-ordered distribution, so the differences are equated with break points between equal-sized segments (in this case there are 11). So, if the predicted-observed difference is greater than any of the observed-observed differences, the likelihood is $0.5/11 = 0.045$. If it is smaller, the likelihood is $10.5/11 = 0.955$. If five $\text{abs}(\partial_j - \partial_i)$ values are higher than $\text{abs}(E(\partial_i) - \partial_i)$ and five are lower, the likelihood is $5.5/11 = 0.5$.

This method is complicated by the fact that if all else is equal, the median change is usually the most likely prediction. For example, suppose the median is zero. If we always guess that $E(\partial_i) = 0$, then each $\text{abs}(E(\partial_i) - \partial_i)$ becomes $\text{abs}(\partial_i)$ and all of these mismatches are within the range of $\partial$. However, many of the $\partial_i - \partial_i$ comparisons will be outside of that range. A simple correction is to divide the comparisons by the relevant standard deviations. For each $\text{abs}(E(\partial_i) - \partial_i)$ this is the standard deviation of all such predicted-observed differences, and for each $\text{abs}(\partial_j - \partial_i)$ it is the standard deviation of all observed-observed differences involving that $\partial_i$. The correction usually achieves the desired result of causing a random guess drawn from $\partial$ to have a likelihood close to 0.5.

It is important to note that this method is not specific to diversity data, so it can be used to evaluate any set of competing time series models as long as the data consist of counts.

**Dynamic models**

Five fundamentally different processes were examined: random walks; exponential trends; simple density dependent (logistic) trends; complex dynamics in which equilibrium points or exponential growth coefficients changed at certain times; and coupled logistic models. More formally, for a random walk $E(n_{i+1}) = n_i$, and for an exponential trend $E(n_{i+1}) = n_i + \mu$, where $\mu$ is the average difference between neighboring points in the time series. A standard Ricker model was assumed for logistic
growth: $E(n_{i+1}) = n_i \exp(r(1 - n_i/K))$ where $r$ = an intrinsic net diversification rate and $K$ = the carrying capacity.

Sepkoski (S2, S33, S34) did not use a Ricker model. Instead, he used the more traditional logistic growth equation $E(n_{i+1}) = n_i + r n_i(1 - n_i/K) = n_i(1 + r(1 - n_i/K))$ (see S31, eqn. 11). Because the variables in this functional form are expressed on linear axes, the equation may predict a negative value of $E(n_{i+1})$ whenever $n_i$ is very high. Negative diversity does not exist. Also with respect to reality, because diversification is a multiplicative process it makes more sense to make the density dependence term a multiplier.

The "average" difference in the exponential model was based either on the arithmetic mean of the logged $n_{i+1} - n_i$ values or on the median of these values. In most cases (but far from all) the mean generated more likely predictions.

The $r$ values in the logistic models were treated as pure nuisance parameters, meaning, they were computed using a grid search and the values yielding the highest likelihoods were selected. By contrast, the $K$ (= carrying capacity) terms were fixed according to four different criteria that all had straightforward biological meanings: as the mean or median of the logged $n_i$ counts, as the maximum $n_i$, or as the very last $n_i$. The mean or median makes sense if one assumes that the equilibrium is reached very quickly and all further variation is stochastic. The maximum makes sense if one assumes that the equilibrium point is reached at some time and it is never exceeded, but there are many small perturbations that drag diversity below it. The end point is a reasonable guess if one assumes that the true equilibrium point is never reached or reached only very slowly, and that the maximum may or may not be closest to this point. The $K$ value yielding the highest likelihood was selected as most representative of the logistic model.

Models with multiple, offset exponential trends were created by sequentially imposing break points. If adding a break did not improve the fit, no further breaks were examined. Separate analyses were performed in which the $r$ values were computed using either means or medians, but only using one of these criteria at a time regardless of the number of breaks.

Sequential break points also were explored for logistic growth. Only the $K$ values were recomputed at the break points (so, for example, in the "end point" model $K$ was always set to the last $n_i$ value before a given break). The most likely $r$ value was recomputed while exploring each possible break. Allowing $r$ values to vary between segments did not make a large difference in most cases, and did not produce better AICc scores because of the penalty involved in adding a parameter.

Interaction models assumed very simple one-way density dependence, meaning, they treated the focal group's $K$ value as a joint, unweighted function of its own diversity plus that of the interacting group. In other words, it was assumed that members of the focal group competed just as strongly with each other as with members of the interacting group. The intrinsic growth rates $r$ were again allowed to vary as nuisance parameters, and the same four methods of fixing $K$ (mean, median, maximum, and end point) were examined in each case.

Finally, there is some reasonable concern about whether the appearance of an
"equilibrium" could be generated by imposing random sampling error on data having no biological signal. This model is clearly wrong in general because the diversity curves are so strongly autocorrelated (Fig. 1); if "random" in any sense, these curves are at worst random walks, not random (white) noise. The expectation given a white noise model is regression to the mean, i.e., \( E(n_{i+1}) = \text{mean}(n) \), so \( E(n_{i+1} - n_i) = \text{mean}(n) - n_i \).

This simple model was also tested in all cases, and yielded such extremely high AICc values that they never had any influence on the Akaike weight computations. Previous analyses of turnover rates in the overall data set also showed no evidence for a substantial regression to the mean effect (S18).

**Dynamic model fitting results**

Based on the Akaike weights (S31), random walk models gain virtually no support in analyses of 11 out of 14 groups (Table S1) despite the fact that they are notoriously hard to reject as null models (S35). In the three cases where random walks are weakly supported at least three alternative models cannot be ruled out. Moreover, the curves in general fail to meet the minimal description of a classic random walk because their changes (i.e., step sizes) are non-normally distributed. The logged and differenced data for "articulates," Conodonta, Chondrichthyes, and Cephalopoda are nonnormal according to either a Shapiro-Wilk test (at \( p < 0.05 \)). Additionally, at least one test statistic is significant at \( p < 0.10 \) for Gastropoda and Bivalvia, which is more than would be expected at random in an analysis involving 10 groups. The results are not as clear using a more conservative Anderson-Darling test, but seven of 11 groups have \( p \)-values below 0.2, far more than the random expectation.

Based on the Shapiro-Wilk results, the four major groups with potentially normal step size distributions are Anthozoa, Trilobita, Gastropoda, and Bivalvia. However, a random walk receives no support in except in the case of Gastropoda, where is it poor (Table S1). Pure exponential model models are almost as bad: whenever they perform well (e.g., Anthozoa, Trilobita, and Bivalvia), models including some form of density dependence are cumulatively better supported. This pattern is particularly surprising because much of the literature dwells on exponential growth (e.g., S36, S37).

Simple logistic models and one-way coupled interaction models generally perform better (Table S1). However, these cases are all ambiguous, with support for one model or the other never rising much above 0.5. The real pattern is that complex uncoupled models receive fair support in every single case, decisive support (Akaike weight > 0.95) in two cases (Linguliformea and Bryozoa), and strong support (weight > 0.75) in two other cases (Echinoidea and Graptolithina). Complex models are also overwhelmingly supported in the cases of the Modern fauna and the overall diversity curve.

Even more surprisingly, not merely a complex model in general but some version of a logistic model in particular is almost always the best. The two exceptions involve Anthozoa and Graptolithina. The Anthozoa curve is better explained by a simple exponential model but is still very flat, while the Graptolithina curve has a very steep climb followed by a shallower but still fast drop off that is best fit by two exponential
trends. Because only 13 data points are available for the latter clade, the result can hardly be termed convincing.

Visual inspection of the data (Fig. 1) makes it clear why logistic models perform so well: most curves include long plateaus with sharp offsets. This pattern is equally apparent in the data for minor groups. For example, Linguliformea persisted at subsampled diversity levels of one to four throughout almost the entire duration of the Silurian through Neogene periods; Conodonta mostly hovered between two and six genera between the Silurian and its last appearance at the end of the Triassic; and Echinoidea increased to a level of about 10 genera in the mid-Jurassic but then only episodically rose above 20.

The best estimate of the carrying capacity was most often the end point, suggesting that the initial lag phase is long for most groups (Table S2). Median values, but not mean values, also were often good estimates. The maximum only ever gave the best fit for single-equilibrium models, which never were strongly supported (Table S1). The fact that end points gave good fits is further evidence against the idea that the appearance of an equilibrium often results from a regression to the mean bias.

Random walks and simple exponential models also are essentially rejected in analyses of the three evolutionary faunas (Figs. 3, S9). The Modern fauna's trajectory is best explained by a three-phase logistic model, with the equilibrium roughly quadrupling after two increases that both occur within the Cretaceous. The Paleozoic fauna appears to track two equilibria offset by the Permo-Triassic mass extinction. A coupled logistic interaction with Trilobita also is plausible, but the joint carrying capacity in this model is almost the same as the saturation points produced by the independent logistic models. The situation for the Cambrian fauna is unclear. An exponential model cannot be excluded, but it would assume a strongly negative net diversification rate, so it would fail to explain qualitatively how this set of taxa could have increased from nothing to its maximum almost instantaneously.

Interaction models also receive little support here, but at least they do make sense in light of Sepkoski's original coupled logistic scenario (S2, S34). The Cambrian fauna interacts most strongly with Cephalopods (a major Paleozoic group) and the Paleozoic fauna with Trilobita (the major Cambrian group). The Modern fauna's strongest pairing is with all other groups combined. Thus, although independent dynamics explain the data better, it is still conceivable that much of the pattern relates to pairwise interactions: a handoff between the Cambrian and Paleozoic faunas, and a later replacement of those two assemblages by the Modern fauna (S2).

For total diversity the best model is a two-phase logistic curve, with the carrying capacity increasing by about 77% toward the end of the Cretaceous. This pattern is mostly driven by the Cretaceous gastropod radiation (Fig. 1A). A three-phase model with an additional but small increase at the Triassic-Jurassic boundary is equally plausible: adding it as a fifth alternative would result in very similar respective Akaike weights of 0.547 and 0.412 for the two- and three-phase interpretations.
Timing of radiations and extinctions

Because almost every group's curve is best explained a two-phase logistic model, we must ask whether the offsets coincide. Almost none of them do exactly, but there are several clusters.

Within the traditional Cambrian fauna, Conodonta and Linguliformea drop near or at the end of the Ordovician, but the large offset for Trilobita coincides with the mid-Devonian crash in overall marine diversity (Fig. 3).

Equilibrium points for the key Paleozoic groups Crinoidea and "articulate" brachiopods drop within the Permian. The brachiopod pattern is complicated by large, short-term peaks in the Silurian to Early Devonian and Permian (Fig. 1A) that are driven by the very different histories of two major subclades. The class Strophomenata has an essentially flat trajectory until a major radiation across the Carboniferous-Permian boundary, and then is almost erased by the Permo-Triassic mass extinction. The remaining "articulate" brachiopods exhibit something more like a Cambro-Ordovician adaptive radiation followed by a lengthy exponential decline, and within this assemblage the class Rhynchoschellata constitutes most of the Silurian-Devonian peak. The fact that brachiopod dynamics can only be understood by discussing individual classes again makes the point that seemingly complicated diversity trajectories may really be composites of simple, unrelated patterns involving major subgroups.

The only major, very long-term shift near the Permo-Triassic boundary involves echinoids, which were persistently depauperate throughout the entire Paleozoic and experienced a large radiation that seems to have accelerated in the Middle Triassic. Likewise, the Cretaceous-Paleogene boundary extinction had no large, irreversible effect on any group other than Cephalopoda, which fell to a much lower equilibrium point. It otherwise shows a remarkably aimless pattern going all the way back to its lightning-fast radiation in the earliest Ordovician.

The Mesozoic marine revolution (S38) appears to have spurred on most of the groups that were common at that time. Bryozoa and Chondrichthyes rose to higher levels as this ecological transition was taking place during the mid-Cretaceous. The data for Bivalvia suggest a slightly earlier but similarly large radiation. The Gastropoda curve shows surprisingly little change through the Cretaceous, but the end-Cretaceous point suggests a large increase before an even larger Paleocene jump. Finally, although the Anthozoa curve is not well-fit by any logistic model. However, Cretaceous counts are persistently high, and if there were any shifts in an equilibrium point one of them most likely occurred in the Early Cretaceous.

There is no strong support for the idea that any group changed its equilibrium point within the Cenozoic. Even the anomalously high Neogene counts for Gastropoda and Bivalvia are not enough to suggest secondary radiations during this period. It remains to be seen whether these Neogene spikes are real or artifacts of residual biases such as improved geographic coverage and preservation, but the latter certainly seems plausible.

Much of the preceding shows that perturbations are important but typically affect only a few groups at a time. For example, there are shifts corresponding with diversity
crashes or mass extinctions in the late Ordovician (Conodonta and Linguliformea), in the mid-Devonian (Trilobita), and at the Cretaceous-Paleogene boundary (Cephalopoda). However, a roughly equal number of cases seem to involve adaptive radiations of some kind, such as the two large brachiopod peaks, the multiple rises associated with the Mesozoic marine revolution, and the mid-Phanerozoic diversification of echinoids. The important overall pattern is, however, no pattern: different major events in Earth history seem to matter only to selected groups, even when disruptions are very large.
Table S1

Akaike weights for alternative dynamic models that account for changes in diversity in 14 major taxonomic groups, the three evolutionary faunas, and the summed diversity of all groups. Fauna = evolutionary fauna; Random = random walk; Exp = exponential; Logistic = simple logistic with an unchanging equilibrium; Complex = double logistic (2K), triple logistic (3K), or double exponential (2r); Coupled = coupled logistic involving a second group. Evolutionary fauna name abbreviations are Cm = Cambrian; Pz = Paleozoic; Md = Modern. "Articulata" = Brachiopoda minus Linguliformea. Support values < 0.05 are omitted; support values > 0.75 are highlighted in bold.

<table>
<thead>
<tr>
<th>Group</th>
<th>Fauna</th>
<th>Random</th>
<th>Exp</th>
<th>Logistic</th>
<th>Complex</th>
<th>Coupled</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anthozoa</td>
<td>Pz</td>
<td>-</td>
<td>0.329</td>
<td>0.125</td>
<td>0.336 [2r]</td>
<td>0.206</td>
</tr>
<tr>
<td>Trilobita</td>
<td>Cm</td>
<td>-</td>
<td>0.583</td>
<td>0.068</td>
<td>0.277 [2K]</td>
<td>0.069</td>
</tr>
<tr>
<td>Ostracoda</td>
<td>Pz</td>
<td>-</td>
<td>-</td>
<td>0.184</td>
<td>0.697 [3K]</td>
<td>0.119</td>
</tr>
<tr>
<td>Linguliformea</td>
<td>Cm</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.999 [2K]</td>
<td>-</td>
</tr>
<tr>
<td>&quot;Articulata&quot;</td>
<td>Pz</td>
<td>-</td>
<td>-</td>
<td>0.057</td>
<td>0.432 [2K]</td>
<td>0.504</td>
</tr>
<tr>
<td>Bryozoa</td>
<td>Md</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.994 [2K]</td>
<td>-</td>
</tr>
<tr>
<td>Crinoidea</td>
<td>Pz</td>
<td>-</td>
<td>-</td>
<td>0.085</td>
<td>0.555 [2K]</td>
<td>0.307</td>
</tr>
<tr>
<td>Echinodermia</td>
<td>Md</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.773 [2K]</td>
<td>0.105</td>
</tr>
<tr>
<td>Graptolitina</td>
<td>Cm</td>
<td>-</td>
<td>-</td>
<td>0.074</td>
<td>0.896 [2r]</td>
<td>-</td>
</tr>
<tr>
<td>Conodonta</td>
<td>Cm</td>
<td>0.148</td>
<td>0.075</td>
<td>0.063</td>
<td>0.675 [2K]</td>
<td>-</td>
</tr>
<tr>
<td>Chondrichthyes</td>
<td>Md</td>
<td>0.367</td>
<td>0.054</td>
<td>0.282</td>
<td>0.095 [2K]</td>
<td>0.205</td>
</tr>
<tr>
<td>Cephalopoda</td>
<td>Pz</td>
<td>-</td>
<td>-</td>
<td>0.317</td>
<td>0.314 [2K]</td>
<td>0.352</td>
</tr>
<tr>
<td>Gastropoda</td>
<td>Md</td>
<td>0.074</td>
<td>0.191</td>
<td>-</td>
<td>0.610 [2K]</td>
<td>0.077</td>
</tr>
<tr>
<td>Bivalvia</td>
<td>Md</td>
<td>-</td>
<td>0.342</td>
<td>0.168</td>
<td>0.191 [2K]</td>
<td>0.271</td>
</tr>
<tr>
<td>Cambrian</td>
<td>-</td>
<td>-</td>
<td>0.463</td>
<td>0.081</td>
<td>0.358 [2K]</td>
<td>0.095</td>
</tr>
<tr>
<td>Paleozoic</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.205</td>
<td>0.330 [2K]</td>
<td>0.464</td>
</tr>
<tr>
<td>Modern</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.996 [3K]</td>
<td>-</td>
</tr>
<tr>
<td>Summed</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.930 [2K]</td>
<td>NA</td>
</tr>
</tbody>
</table>
**Table S2**

Turnover rates and equilibrium diversity levels for the major groups. Median rate = median observed net diversification rate; r = most likely intrinsic diversification rate based on the most strongly supported double or triple logistic model (see Table S1); End point = diversity at the end of the diversity curve (i.e., in the late Neogene); K basis = parameter yielding equilibrium diversity level (median or end point value); Pz K = early Paleozoic equilibrium, as implied by the analyses summarized in Table S1; End K = Cenozoic equilibrium (for living groups) and terminal equilibrium (for extinct groups).

<table>
<thead>
<tr>
<th>Group</th>
<th>Median rate</th>
<th>r</th>
<th>End point</th>
<th>K basis</th>
<th>Pz K</th>
<th>End K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anthozoa</td>
<td>0.038</td>
<td>0.201</td>
<td>27.2</td>
<td>median</td>
<td>21.4</td>
<td>35.4</td>
</tr>
<tr>
<td>Trilobita</td>
<td>-0.205</td>
<td>0.258</td>
<td>NA</td>
<td>end</td>
<td>48.7</td>
<td>5.9</td>
</tr>
<tr>
<td>Ostracoda</td>
<td>0.036</td>
<td>0.591</td>
<td>26.4</td>
<td>median</td>
<td>9.8</td>
<td>29.0</td>
</tr>
<tr>
<td>Linguliformea</td>
<td>-0.021</td>
<td>0.801</td>
<td>1.7</td>
<td>median</td>
<td>11.2</td>
<td>1.9</td>
</tr>
<tr>
<td>&quot;Articulata&quot;</td>
<td>0.155</td>
<td>0.530</td>
<td>10.2</td>
<td>end</td>
<td>67.6</td>
<td>10.2</td>
</tr>
<tr>
<td>Bryozoa</td>
<td>0.048</td>
<td>0.617</td>
<td>51.4</td>
<td>median</td>
<td>11.8</td>
<td>46.2</td>
</tr>
<tr>
<td>Crinoidea</td>
<td>0.077</td>
<td>0.445</td>
<td>2.0</td>
<td>median</td>
<td>18.9</td>
<td>4.3</td>
</tr>
<tr>
<td>Echinoidea</td>
<td>0.030</td>
<td>0.540</td>
<td>20.5</td>
<td>end</td>
<td>2.0</td>
<td>20.5</td>
</tr>
<tr>
<td>Graptolithina</td>
<td>-0.098</td>
<td>0.921</td>
<td>NA</td>
<td>median</td>
<td>8.7</td>
<td>2.5</td>
</tr>
<tr>
<td>Conodonta</td>
<td>-0.110</td>
<td>0.479</td>
<td>NA</td>
<td>median</td>
<td>12.4</td>
<td>5.3</td>
</tr>
<tr>
<td>Chondrichthyes</td>
<td>0.020</td>
<td>0.280</td>
<td>13.7</td>
<td>median</td>
<td>5.9</td>
<td>20.6</td>
</tr>
<tr>
<td>Cephalopoda</td>
<td>-0.029</td>
<td>0.297</td>
<td>2.0</td>
<td>median</td>
<td>30.0</td>
<td>3.1</td>
</tr>
<tr>
<td>Gastropoda</td>
<td>0.098</td>
<td>0.331</td>
<td>248.8</td>
<td>end</td>
<td>43.9</td>
<td>248.8</td>
</tr>
<tr>
<td>Bivalvia</td>
<td>0.050</td>
<td>0.138</td>
<td>129.1</td>
<td>end</td>
<td>38.7</td>
<td>129.1</td>
</tr>
</tbody>
</table>
**Fig. S1.** Rarefaction of two rank-ordered abundance distributions. Quota of randomly drawn specimens is 200. Point identified with a number corresponds to the least abundant genus that would be drawn if the genera were sampled in order of frequency, as they more or less are (i.e., an approximate minimum). (A) Actual distribution of 75 genera in Paleobiology Database collection 26003. Quota is roughly equivalent to a minimum frequency (i.e., veil line) of 0.035 (dotted line) and on average yields a minimum draw of 33 genera. Dotted line indicates this frequency. (B) Same distribution as in (A) with a doubled genus pool but sampled with the same method and quota. Because the veil line does not change, a minimum of less than twice as many genera
are sampled (52, as opposed to 66).

**Fig. S2.** Shareholder quorum subsampling (SQ) of the same distributions shown in Fig. S1. Sampling target is a quorum (sum of proportions of represented taxa) of 0.97. Gray area under the curve adds up to this amount. Genera are again assumed to have been sampled in rank order, which in this case means that the sampled count is an absolute minimum. (A) SQ of the actual distribution of 75 genera in collection 26003, which on average yields a minimum 33 genera. (B) SQ with the same quorum target and a doubled species pool, which on average yields twice as many sampled taxa (66).
Fig. S3. Sampling standardized Phanerozoic marine diversity curves produced by SQ analyses of individual major taxa that were either performed separately for each group and then summed (thin line) or based on combined data (thick line). The respective quorum targets were 0.50 and 0.60.
Fig. S4. Effect of adjusting diversity estimates to remove the influence of variable turnover rates. Thin lines = standard SQ diversity curves; thick lines = same curves adjusted with the equation $N' = N + 2T - 3T$. (A) Curves based on the combined data set for all marine invertebrates. (B) Curves for trilobites only.
**Fig. S5.** Complete linkage cluster analyses of factor analysis data for sampling standardized but undifferenced diversity curves (Fig. 1). Coordinates are transformed into Euclidean distances before clustering. **(A)** Clusters based on axis 1 scores. **(B)** Clusters based on axis 2 scores excluding Trilobita, Linguliformea, Graptolithina, and Conodonta (i.e., the Cambrian fauna isolated on axis 1).
Fig. S6. Cluster analyses of principal coordinates data for differenced diversity curves. Methods are as in Fig. S5. (A) Clusters based on axis 1 coordinates. (B) Clusters based on axis 2 coordinates excluding Trilobita, Ostracoda, Linguliformea, Graptolithina, and Conodonta (i.e., members of the Cambrian fauna plus Ostracoda).
**Fig. S7.** Bootstrapped principal coordinates analysis scores for the six major groups whose diversity trajectories are shown in Fig. 1. Two-letter abbreviations indicate observed values; small points indicate bootstrapped values rotated by Procrustes analysis with real values as the target. Green = Anthozoa; red = Trilobita; orange = "articulate" Brachiopoda (Ar); purple = Cephalopoda; gray squares = Gastropoda; blue = Bivalvia.
Fig. S8. Multivariate ordinations of diversity counts for 23 major marine animal higher taxa across 75 time intervals. Evolutionary fauna categorizations are Sepkoski’s (S22). The Cambrian fauna (gray circles) includes Trilobita, Polychaeta, Tergomya, Linguliformea, Hyolitha, Graptolithina, and Conodonta; the Paleozoic fauna (open squares) includes Anthozoa, Ostracoda, "articulate" Brachiopoda, Stenolaemata, Cephalopoda, and Crinoidea; and the Modern fauna (black circles) includes Demospongea, Hexactinellida, Malacostraca, Gymnolaemata, Asterozoa, Echinoidea, Bivalvia, Gastropoda, Chondrichthyes, and "Osteichthyes." (A) Factor analysis of raw counts with zero values included, as in Sepkoski’s original analysis (S22). (B) Principal coordinates analysis of logged and differenced counts with zero values ignored.
Fig. S9. Separately plotted diversity curves for the Modern (black line), Paleozoic (gray line), and Cambrian (dotted line) evolutionary faunas (see also Fig. 3).
Supplemental References