Palaeontology
The Palaeontological Association
www.palass.org

[Palaeontology, Vol. 53, Part 6, 2010, pp. 1211–1235]

GEOGRAPHICAL, ENVIRONMENTAL AND INTRINSIC BIOTIC CONTROLS ON PHANEROZOIC MARINE DIVERSIFICATION

by JOHN ALROY*

Paleobiology Database, National Center for Ecological Analysis and Synthesis, 735 State Street, University of California, Santa Barbara, CA 93101, USA; e-mail alroy@nceas.ucsb.edu
*Present address: Department of Biological Sciences, Macquarie University, Sydney, NSW 2109, Australia

Typescript received 21 January 2010; accepted in revised form 12 September 2010

Abstract: The Paleobiology Database now includes enough data on fossil collections to produce useful time series of geographical and environmental variables in addition to a robust global Phanerozoic marine diversity curve. The curve is produced by a new 'shareholder quorum' method of sampling standardization that removes biases but avoids overcompensating for them by imposing entirely uniform data quotas. It involves drawing fossil collections until the taxa that have been sampled at least once (the 'shareholders') have a summed total of frequencies (i.e. coverage) that meets a target (the 'quorum'). Coverage of each interval's entire data set is estimated prior to subsampling using a variant of a standard index, Good's *u*. This variant employs counts of occurrences of taxa described in only one publication instead of taxa found in only one collection. Each taxon's frequency within an interval is multiplied by the interval's index value, which limits the maximum possible sampling level and thereby creates the need for subsampling. Analyses focus on a global diversity curve and curves for northern, southern and 'tropical' (30°N to 30°S) palaeolatitudinal belts. Tropical genus richness is remarkably static, so most large shifts in the curve reflect trends at higher latitudes. Changes in diversity are analysed as a function of standing diversity; the number, spacing and palaeolatitudinal position of sampled geographical cells; the mean onshore–offshore position of cells; and proportions of cells from carbonate, onshore and reefal environments. Redundancy among the variables is eliminated by performing a principal components analysis of each data set and using the axis scores in multiple regressions. The key factors are standing diversity and the dominance of onshore environments such as reefs. These factors combine to produce logistic growth patterns with slowly changing equilibrium values. There is no evidence of unregulated exponential growth across any long stretch of the Phanerozoic, and in particular there was no large Cenozoic radiation beyond the Eocene. The end-Ordovician, Permo–Triassic and Cretaceous–Palaeogene mass extinctions had relatively short-term albeit severe effects. However, reef collapse was involved in these events and also may have caused large, longer term global diversity decreases in the mid-Devonian and across the Triassic/Jurassic boundary. Conversely, the expansion of reef ecosystems may explain newly recognized major radiations in the mid-Permian and mid-Jurassic. Reef ecosystems are particularly vulnerable to current environmental disturbances such as ocean acidification, and their decimation might prolong the recovery from today's mass extinction by millions or even tens of millions of years.

Key words: adaptive radiation, diversity estimation, global change, macroevolution, mass extinction, rarefaction.

MUCH of the palaeobiological literature adheres to the notion that environmental factors determine levels of local, regional and global diversity. For example, claims that diversity in the deep fossil record tracks a latitudinal gradient or reflects biogeographical provincialism go back at least four decades (Fischer 1960; Stehli *et al.* 1969; Valentine 1970). However, even after so many years of discussion we are still at an early stage of testing such ideas with rigorous statistical analyses, thanks in no small part to the extreme difficulty of removing sampling overprints.

Solving this problem requires assembling large amounts of data resolved to the level of individual fossil localities (e.g. Knoll *et al.* 1979) instead of working with simple lists of geological first and last appearances that lack any contextual information (e.g. Sepkoski 1984, 1997). It also requires separating the concepts of accurate, comprehensive and uniformly sampled. I will show that these properties are not only different but impossible to attribute to a single diversity curve. Finally, it requires rethinking what it would mean to demonstrate causal relationships between diversity and either biotic or abiotic factors (McKinney and Oyen 1989).

Here, I employ these new strategies to investigate controls on Phanerozoic marine invertebrate diversity at the

doi: 10.1111/j.1475-4983.2010.01011.x

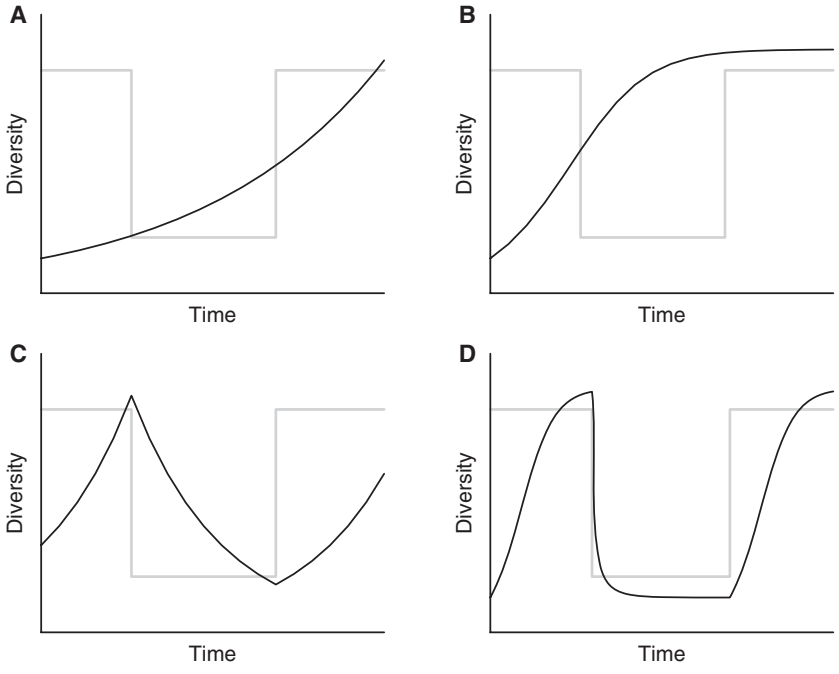
global scale and within three major palaeolatitudinal belts. The analysis focuses primarily on comparisons between global, southern and tropical trends because there is little northern hemisphere shelf area throughout much of the early Palaeozoic. It employs new statistical methods (Alroy 2010) that explicitly seek to reconstruct each taxonomic sampling pool's relative size, i.e. its richness in the strict ecological sense (cf. Hurlbert 1971; Gotelli and Colwell 2001), as opposed to drawing a fixed amount of data for no other reason than that it seems fair (Alroy 1996; Miller and Foote 1996; Alroy *et al.* 2008). It works within a framework of rigorous time series analysis and instead of taking a piecemeal approach, it considers biotic, environmental and geographical factors simultaneously. Gaps in coverage of the literature are still evident and additional variables may need to be considered, but strong patterns in the data suggest that diversification is neither random nor governed by a single set of factors.

CONCEPTUAL FRAMEWORK

There are many technical difficulties with analysing diversity data, but proper categorization of rival hypotheses is a more fundamental and in fact overriding complication. Four strong themes in the palaeontological literature are relevant. First, any diversity pattern may result from speciation and extinction outcomes that appear to be random when nothing is known about population-level processes, and therefore seem to be governed by underlying turnover probabilities that are effectively constant (Raup *et al.* 1973; Van Valen 1973). Second, diversity may increase randomly (and therefore exponentially) but appear to be limited because large mass extinctions keep it in check (Hoffman and Fenster 1986; Benton 1995, 2009). Third, diversity may have hard limits imposed by ecological interactions such as competition. If all other things are equal, diversification should follow a logistic trend as the equilibrium carrying capacity is approached and turnover rates are brought into equality (Sepkoski 1978). Finally, diversity may track some physical factor such as latitude (Fischer 1960; Stehli *et al.* 1969) or habitable geographical area (Schopf 1974; Sepkoski 1976).

Some workers have dichotomised these themes by emphasizing 'biotic' and 'abiotic' factors (e.g. Miller 1998; Benton 2009). However, the best models of underlying dynamics may need to combine the two things. For example, diversity might track area (theme 4) exactly because competition over a resource creates an equilibrium (theme 3) and the resource's amount is correlated with area. In fact, the seemingly universal relationship between species richness and geographical area was the very pattern that motivated the early ecological literature on logistic diversity dynamics (MacArthur and Wilson 1967; Schopf 1974; Simberloff 1974). Resource-based equilibrium models also could explain latitudinal diversity gradients. Nonetheless, resources might govern intrinsic diversification rates (theme 2) instead of carrying capacities, with the rates not changing until the resources base does.

A simple sketch (Text-fig. 1) makes it much easier to understand the alternative models and their alternative predictions:



TEXT-FIG. 1. Alternative views of diversity dynamics. Black lines indicate diversity on a log scale; grey lines indicate a key extrinsic forcing factor such as a palaeogeographic or environmental variable or the diversity of a different taxonomic group. A, pure exponential growth with no extrinsic or intrinsic controls (Red Queen model). B, logistic growth with no extrinsic controls but with a limit to diversity imposed by intrinsic factors such as ecological competition. C, exponential growth with a rate determined by a changing extrinsic factor. D, logistic growth with a diversity limit imposed by a changing extrinsic factor.

1. If rates are truly constant and therefore independent of changes in the environment, diversity should grow exponentially (Text-fig. 1A). This prediction follows from the 'Red Queen' theory that was formulated to explain why extinction probabilities seem to be independent of taxon ages in many data sets ('Van Valen's law', Van Valen 1973).
2. If biotic interactions cause extinction rates to rise or origination rates to fall as diversity increases (i.e. growth is density dependent), diversity should increase logistically (Text-fig. 1B). This is not to say that abiotic factors are irrelevant; they must be, or else there could not be an equilibrium. It is also not to say that organisms cannot change the equilibrium point through evolutionary innovation. It merely supposes that the resources limiting diversity do not wax or wane during the interval of study, and that these resources continue to be partitioned in much the same way.
3. Diversity may never approach an equilibrium, but if the underlying rates correlate with some changing factor it still may not grow constantly (Text-fig. 1C). This factor may or may not be environmental. For example, evolutionary innovations within the group under study or in a clade of competitors, predators or prey also may change the trajectory.
4. Finally, diversity may track an equilibrium that in turn tracks a varying resource such as marine shelf area (e.g. Simberloff 1974), or perhaps the varying richness of an interacting taxonomic group (Sepkoski 1979). The resulting pattern would most likely be complex (Text-fig. 1D), and declines could be rapid enough to suggest an instantaneous catastrophe despite being driven by medium- or long-term shifts in the carrying capacity (Sepkoski 1984).

We can now see that Van Valen's law (Text-fig. 1A) is actually inconsistent with and an alternative to equilibrial diversity dynamics (Text-fig. 1B); that complex changes in diversity can occur even when relevant environmental factors are constant (Text-fig. 1B; Sepkoski 1984); that environmental changes can be important regardless of whether there is saturation (Text-fig. 1C–D); and that a strongly equilibrial dynamic can be present even if a diversity curve does not seem logistic (Text-fig. 1D). Indeed, the one scenario most likely to generate a highly visible latitudinal diversity gradient in the Recent is the one scenario combining biotic and abiotic controls (Text-fig. 1D). This fact has long been recognized by researchers trying to explain why latitudinal gradients seem to exist in stasis over such large periods of time (e.g. Fischer 1960; Stehli *et al.* 1969).

The inferences drawn from Text-figure 1 are not intuitive, and it is no coincidence that so much of the palaeobiological literature adheres to a different conceptual framework. For example, a recent review (Benton 2009) emphasized commonalities between the Red Queen and equilibrial accounts; interpreted the latter as being at odds with explanations involving physical factors, although an equilibrium can only exist if limitations are imposed by physical factors; and emphasized theoretical claims that competition and predation cannot control global diversification when this idea is not a deduction from low-level patterns but an empirical hypothesis to be tested directly against data, as done here.

These distinctions are more than philosophical; they point the way to resolving the debate through concrete statistical analyses. To begin with, the variable to be explained is the rate of growth in taxonomic richness, not its absolute magnitude. The fact that dynamics concern rates and not simple curves is overlooked in much of the literature (e.g. Benton 1995). If we assume that speciation and extinction are random and continuous probabilistic processes, we can define the net diversification rate as the difference between neighbouring diversity levels on a log scale, i.e. the log ratio (Raup 1985; Alroy 2000, 2008, 2010).

We expect nothing to correlate with net rates in the Red Queen scenario (Text-fig. 1A). Rates should be correlated negatively with standing diversity if resources are constant and enough time elapses to approach the equilibrium point (Text-fig. 1B), and they should correlate positively with some physical factor if that factor changes through time and biotic interactions are still weak at the highest diversity levels that are reached (Text-fig. 1C).

Finally, diversity should change in lockstep with physical factors if (1) competition is strong enough to keep diversity close to its saturation point at all times, (2) the right physical factors are measured, and (3) these factors change greatly through time (Text-fig. 1D). This deduction may seem confusing; one might think instead that the strongest evidence in favour of 'abiotic' controls would be a direct relationship between changes in diversity and changes in the environment. To the contrary, such a correlation would be the best possible evidence for 'biotic' logistic dynamics. Alternatively, if diversity is equilibrial but often not at its equilibrium, both changes in the environment and in standing diversity should be statistically important.

In sum, we can think of the problem in terms of a multiple regression of diversification rates on (1) standing diversity levels, (2) absolute magnitudes of physical variables, and (3) changes in the same physical variables. The combination of variables that proves to be significant will allow us to identify the most important model categories (Text-fig. 1). With very few exceptions (e.g. Sepkoski 1976; McKinney and Oyen 1989; Peters and Foote 2001; Kiessling and Aberhan 2007), macroevolutionary studies have instead tended to focus on a single factor at a time.

Additionally, older literature used crude tabulations that could not be sampling standardized (e.g. Sepkoski 1976), and more recent and sophisticated analyses have tended to address relatively short time intervals (McKinney and Oyen 1989; Connolly and Miller 2002; Kiessling and Aberhan 2007).

MATERIALS AND METHODS

Taxonomic occurrence data

Diversity curves were obtained through sampling standardization of genus-level taxonomic occurrence data downloaded on 2 September 2010 from the Paleobiology Database (PaleoDB: http://paleodb.org). An occurrence is defined as the presence of one genus in one fossil collection that typically originates from a single narrow stratigraphic horizon in a small geographical area. Almost all palaeontological literature on global diversity works at the rank of genus or above (e.g. Sepkoski 1984, 1997), so analyses were kept at this level even though species-level patterns are highly similar. Likewise, there is an almost linear relationship between species- and genus-level diversity in extant benthic molluscs, the Recent group that dominates the Cenozoic fossil record (Roy *et al.* 1996).

Occurrences were binned into 50 time intervals equivalent to geological stages or sets of neighbouring stages. The interval definitions (Alroy 2010) are much the same as those used in several recent papers (e.g. Alroy *et al.* 2008), which are built in to the PaleoDB website. However, the Cenozoic is split into eight bins instead of six. Respectively, these are Palaeocene; Early, Middle and Late Eocene; Oligocene: Early Miocene; Middle Miocene plus Late Miocene; and Pliocene plus Pleistocene.

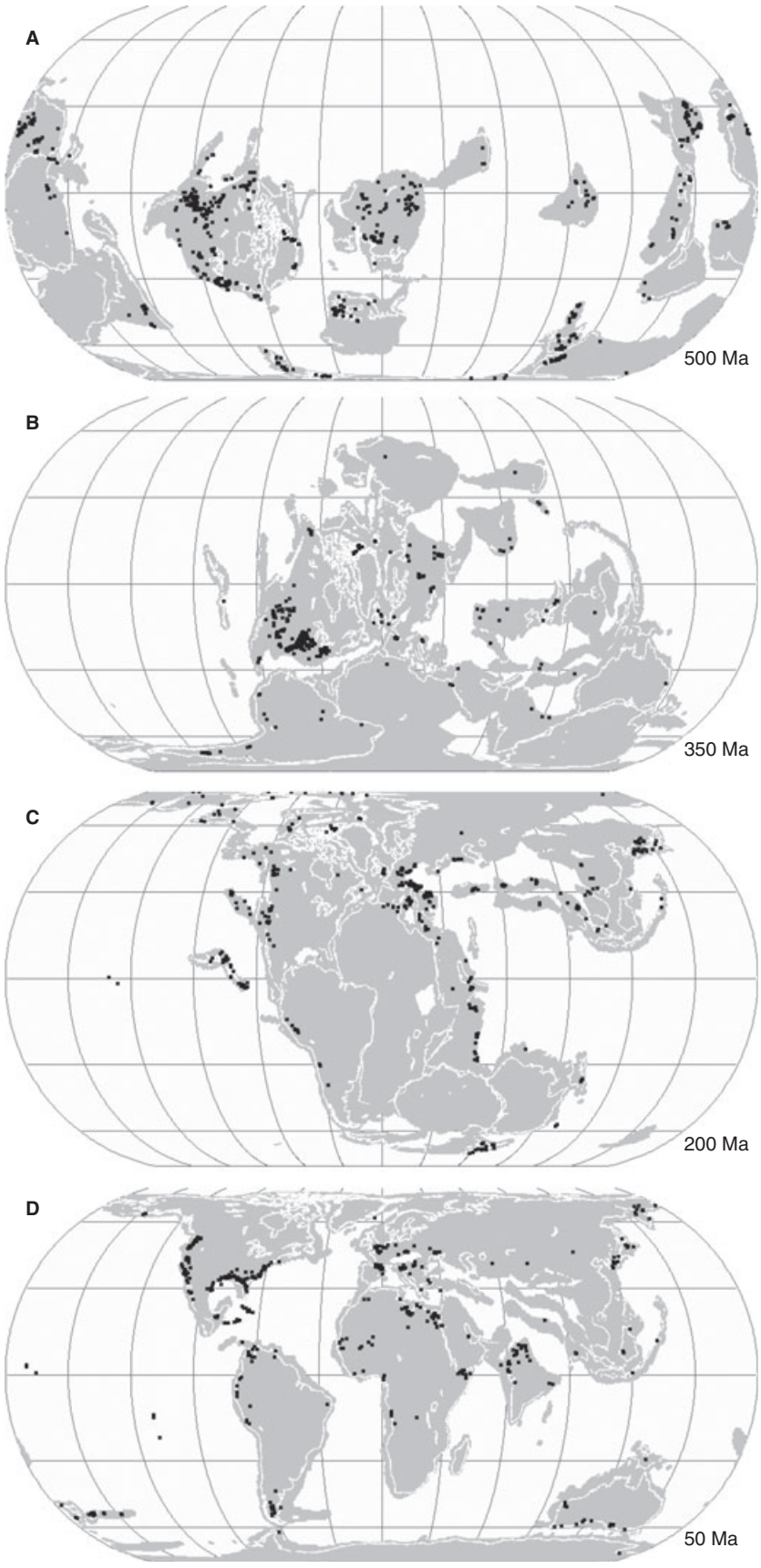
Sifting criteria generally followed those used by Alroy *et al.* (2008). All metazoan groups except Tetrapoda were included. Collections coming from terrestrial environments or representing basin-scale areas or entire stratigraphic groups were excluded. To guarantee that comparisons of global and regional diversity were consistent, only those collections with unequivocal tectonic plate assignments were included. The remainder fall in gaps between the plates having known rotations. Preservation quality was held constant by excluding the relatively few fossil collections that preserved original aragonite or soft tissue, came from unlithified sediments or came from poorly lithified sediments that were sieved but not treated chemically. Synonymies and reidentifications were employed whenever possible. Occurrences of generically indeterminate taxa or genus names qualified with quotation marks, aff., *ex gr.* or *sensu lato* were excluded. Subgenera were treated as separate taxonomic units for consistency with earlier literature (e.g. Sepkoski 1997; Alroy *et al.* 2008). Finally, form taxa and ichnofossils were excluded.

Palaeogeographic reconstructions were based on rotation and plate identification data supplied by C. Scotese and rotation algorithms programmed by the author for use on the PaleoDB website (Text-fig. 2). The 'tropics' were defined as the region within 30 degrees of the equator instead of 23.4 degrees. Northern Pacific bivalve diversity does show a large break at the latter point (Roy *et al.* 2000). However, gastropod diversity begins to rise sharply around 35°N and reaches its peak at 25°N in both the north-western Atlantic and the north-eastern Pacific (Roy *et al.* 1998), so the broader definition seems fair.

Excluding data records that could not be assigned to a unique temporal bin, the global data set included 317 450 occurrences of 20 366 genera distributed among 50 728 collections and drawn from 5994 references. A total of 3355 genera (16.5 per cent) are extant. The tropical data set included 166 507 occurrences of 13 967 genera, 27 030 collections and 3254 references, with 1950 of the genera (14.0 per cent) being extant. There was almost no shelf area above 30°N in the early Palaeozoic and still very little in the mid-Palaeozoic (Text-fig. 1C–D), so the northern data set is only close to complete for the post-Palaeozoic. It includes 89 854 occurrences of 7679 genera, 13 690 collections and 1707 references. Its post-Palaeozoic focus allows it to capture relatively more extant genera (2139 = 27.9 per cent). Finally, for complementary palaeogeographic reasons, the southern data set skews back again towards the Palaeozoic, with 53 507 occurrences of 6793 genera, 9177 collections, 1179 references and 832 extant genera (12.2 per cent).

Time series analysis

Most analyses are based on first differences of the variables, which are indicated with a Δ preceding the variable name. Differences are taken to remove problems with time series autocorrelation that make it impossible to infer causal relationships (McKinney and Oyen 1989). Variables that are counts of items such as genera, references, collections or geographical cells are log transformed prior to any analysis. To extract more information about changes between neighbouring values that are close to one limit or the other, variables that are proportions are rescaled to range between -1 and 1 (i.e. by subtracting one-half and multiplying by two) and then arcsine transformed. Correlations are quantified using Spearman's ρ, a nonparametric estimator that weakens the effects of outliers. When there are none, it provides much the same result as a standard, parametric Pearson's product–moment correlation. Kendall's τ is in some ways a better nonparametric measure of association than ρ. However,



TEXT-FIG. 2. Palaeogeographic reconstructions for selected time intervals using Eckert IV projections. Some small plates lacking any fossils older than the reconstruction date are omitted from the Palaeozoic reconstructions. A, Cambrian data points on a 500-Ma map. B, Early Carboniferous (Pennsylvanian) data points on a 350-Ma map. Collections are much more widespread in most intervals (Text-fig. 8). C, Late Triassic data points on a 200-Ma map. Points in the Pacific fall on small terranes that are now in Alaska and British Columbia. D, Eocene data points on a 50-Ma map. Points in the Pacific are from the Marshall Islands and ODP cores.

the two tests yield almost identical p-values and τ is not commensurate with ρ, so τ results are not reported. Multiple regressions and ordinations were respectively performed with the R statistical package's glm and princomp functions.

Concerns over sampling biases

Questions about whether sampling biases obscure palaeontological diversity patterns go back all the way to the earliest global studies (e.g. Newell 1952; Simpson 1960; Valentine 1970; Raup 1972; Knoll *et al.* 1979). Now that large relational databases can be used to eliminate these biases, we should be past the point where a diversity curve is only ever seen as tabulation (e.g. Sepkoski 1984; Benton 1995) instead of a statistical estimate (e.g. Raup 1976). Nonetheless, some debate continues about whether to remove overprints with sampling standardization (Alroy *et al.* 2008) or consider such concerns to be 'navel gazing' and leave them in (Benton and Emerson 2007).

These arguments hearken back to earlier concerns about interpreting curves taken straight from the literature (e.g. Raup 1976; Sepkoski *et al.* 1981). Some researchers recognized that sampling biases are pervasive but were willing to take the data at face value (e.g. Sepkoski 1978). At the other extreme, some believed sampling problems were so severe and insurmountable that taxon counts were categorically meaningless (Sheehan 1977).

This original debate led nowhere. One major reason was that 'corrected' curves could only be produced by assuming that some isolated variable such as worker interest (Sheehan 1977) or rock amount (e.g. Raup 1976; Sepkoski 1976) strictly reflected sampling bias and was the only important measure of such bias. It was never clear whether these measures were truly relevant. After all, the proximal cause of bias is variation in the amount and topical focus of published data, and more distal factors such as rock volume might or might not compel more publication (Newell 1959; Sheehan 1977).

Nonetheless, it would make sense to use rock amount as a proxy for sample size (e.g. Peters and Foote 2001; Smith 2001; McGowan and Smith 2008) if the rock record was highly incomplete but the parts that were preserved were very well sampled (Smith 2001). It is certainly true that the extent of the rock record varies greatly, but the idea that sampling of available rocks is very good is justified by older statistical analyses (e.g. Paul 1982; Foote and Sepkoski 1999) that are certainly too sanguine (Alroy *et al.* 2008).

This hypothesis still must be considered, and fortunately it makes two very simple predictions that can be tested with the available data set. First, if the species pool available for sampling is almost fully known, or at least known to a uniform extent, standardization should make little difference. Second, sampled geographical area should correlate strongly with diversity levels regardless of standardization. These predictions are not borne out by the analyses presented below; raw taxon counts – but not standardized ones – are primarily controlled by counts of data items such as published references, and area variables are only very weakly correlated with standardized diversity. Therefore, it may be time to table the side-debate over rock amount and focus strictly on palaeontological sampling intensity *per se*.

Shareholder quorum subsampling

Almost all parties now agree that using standardized subsamples is a better idea than using raw counts (Benton 2009). Thus, the real problem is how to standardize; the relevant methods are still in a state of flux (Alroy *et al.* 2008; Alroy 2010). Numerous methods have been proposed, and until recently all of them have aimed to represent each time interval with a randomly drawn and entirely uniform quota of data items, such as estimated counts of specimens (Alroy 1996, 2000; Alroy *et al.* 2008), counts of occurrences (Miller and Foote 1996) or counts of fossil collections (Alroy *et al.* 2001). These methods implicitly equate uniform sampling with accurate sampling.

However, it is not true that uniform sampling is fair. The reason is that if the size of a species pool increases through speciation or immigration, the chance of discovering any one species with a draw of any one data item will drop. Therefore, drawing a fixed amount of data must yield a smaller fraction of the existing species, and for no good reason. If, say, given some item quota and given a mirror-image doubling of the species pool the fraction that is drawn drops by 20 per cent, subsampling will yield $2 \times 0.8 = 1.6$ times as many species instead of twice as many.

Based on this one simple argument, it should be clear that an accurate method must sample harder – and therefore nonuniformly – when richness increases. A simple solution is to track not the number of items that are drawn but the 'coverage' of the data set represented by the species that have been drawn (Alroy 2010). The coverage of any one species is its relative frequency, i.e. the proportion of occurrences that belong to it (Good 1953). If, say, species A, B and C have respective frequencies of 0.2, 0.1 and 0.05, drawing a single fossil collection that includes A, B and C will immediately produce 35 per cent coverage. Standardization can be achieved by drawing enough collections in each time interval to generate the same coverage level. By analogy, the taxa can be thought of as 'shareholders' with their frequencies being shares;

the target coverage level is then a 'shareholder quorum' (Alroy 2010).

The shareholder quorum subsampling (SQS) method does not reduce to drawing a fixed fraction of all the data items because the target is based on frequencies, not counts. In the preceding example, we might achieve a quorum by drawing one collection or two or ten. However many we must draw, that number will be the same regardless of whether we start with (say) 10, 1000 or a million collections.

In real data sets, many taxa are entirely unknown, so coverage of the entire frequency distribution is often highly incomplete. Unknown taxa effectively have frequencies rounded down to zero, so when there are many the known taxa have inflated frequencies. Ignoring this fact and using the empirical frequencies to track the progress of sampling would therefore bias the results against intervals with poor initial coverage.

To fix the problem, we must estimate the coverage of each interval's entire data set and then oversample when coverage is low (Alroy 2010). Good's u, the most commonly used estimator, compares the number of single-occurrence taxa to the total number of occurrences using the equation $1 - n_1/O$, where n_1 = the number of taxa seen only once and O = the total number of observations (Good 1953). If there are no such taxa, coverage is 100 per cent. In theory, dividing the sampling quorum q in each particular interval by u should remove all of the bias against small data sets. An algebraically identical algorithm would be to draw collections until the sum of $u(n_i/\mathrm{O})$ equals q.

The use of a coverage estimator is the crux of the SQS method; it is the reason the method is algorithmically simple but not conceptually trivial. After all, without a correction there would be no reason to subsample; regardless of whether the underlying frequency distribution is log series, geometric series, log normal, etc., no taxa are missing when the full data set in each interval produces 100 per cent coverage (Good 1953; Colwell and Coddington 1994).

Corrections to the subsampling method

Any simple simulation analysis will show that Good's u is extremely accurate and quite precise when overall data sets are produced by entirely random sampling. However, the point of publishing is to describe new phenomena, not to list further random samples of what might already be well-known times, places, environments and taxonomic groups. As a result, literature compilations almost never meet this requirement. The bias is so great that u may actually decrease as literature data sets expand. Fortunately, it so happens that u can be made to work well even in such cases simply by counting occurrences of single-publication taxa (say, p_1) instead of single-occurrence taxa (n_1), which produces a revised equation $u' = 1 - p_1/O$ (Alroy 2010).

Two minor modifications to the method (Alroy 2010) are made necessary by the vagaries of the published literature. Both corrections only become relevant in small and idiosyncratic data sets, but they do seem to decrease residual sampling error even in the current global analysis (as defined below).

First, the most common taxon is ignored in frequency calculations (but still tallied) to avoid conflating apparent ecological dominance with species pool size. Superabundant genera are drawn anyway in almost any subsample, so this correction is at worst harmless. Such genera may be form taxa, determinable but badly undersplit groups of species, organisms with unusually favourable taphonomic characteristics, or ecosystem engineers such as oysters. One way or another, their abundance is unlikely to say very much about overall ecological structure.

Second, taxa only ever appearing in the very most diverse collection are excluded from the count of single-publication taxa. Such collections may distort the results if they include extremely large numbers of specimens or benefit from exceptional preservation, which according to Good's equation would make coverage seem low and therefore cause sampling to be excessive. Obviously, one would think instead that the presence of such collections indicates good coverage.

Quorum levels are set to be moderately high but not entirely maximal in each analysis, the highest possible sampling level being the point past which certain intervals cannot meet sampling targets. A quorum of 0.40 is used for the three latitudinally restricted data sets, and the target varies for global analyses depending on the context. In stark contrast to the alternatives, the exact choice of a sampling level is not important when using this method (see below).

Counts are corrected with a simple 'three timer' equation that removes residual bin-to-bin sampling error (Alroy 2008; Alroy *et al.* 2008). It involves comparing genera sampled in three consecutive bins (three timers or *3T*) and genera sampled immediately before and after but not within focal bins (part timers or *PT*). The chance of sampling is then $3T/(3T + PT)$. This fraction is similar to the simple completeness metric of Paul (1982) that is still often used (e.g. Mander and Twitchett 2008). However, by considering only three intervals at a time it entirely avoids time series censorship biases such as the Signor–Lipps effect or the Pull of the Recent (Raup 1979).

The correction is to divide each bin's genus count by the sampling probability for that one bin. To make sure the original and rescaled curves are comparably centred, the entire curve is then multiplied by the average

probability for the entire data set that is derived from the grand totals for *3T* and *PT*. To avoid problems introduced by binomial error in the counts, a three-timer correction is only made when *3T* is at least two for a given bin.

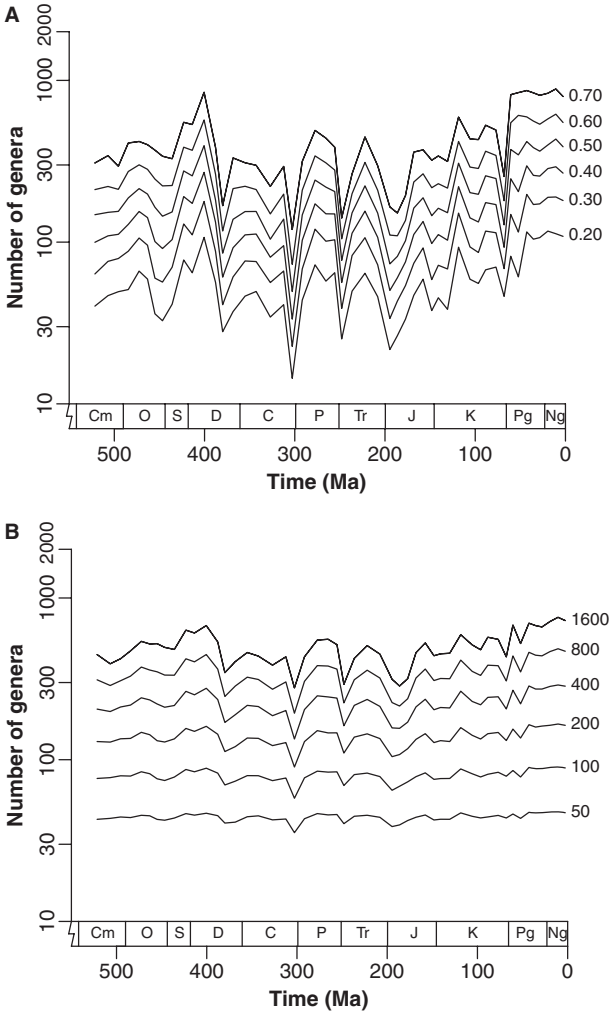
Effect of standardized sampling levels

It is easily shown by simulation that unlike any other method, SQS tracks the size of the species pool with complete accuracy when additions and subtractions are random. That said, like any other method it will tend to underestimate when additions are mostly of very rare species. Nonetheless, its biases are quite weak, and it also has two distinctive and very important properties. First, it produces almost the same relative diversity estimates regardless of the subsampling level. Second, it also produces much the same result regardless of whether a data set is analysed in full or after being subdivided into major subcomponents.

The first point is easily demonstrated by comparing curves produced by SQS and by the subsampling method that has the longest history in palaeontology and is still the most popular, classical rarefaction (CR; Hurlbert 1971; Raup 1975; Tipper 1979; Miller and Foote 1996). SQS curves are almost identical across a broad range of sampling quorums, i.e. from 0.20 to 0.70 (Text-fig. 3A). For that reason, the standard deviations are similar (respectively, 0.459 and 0.504) even though the geometric means differ by a factor of seven (53 and 384 genera).

However, rarefaction produces flatter and flatter diversity curves as sampling quotas decrease (Text-fig. 3B). These curves are punctuated only by a few small excursions, and they depict a steady but weak post-Palaeozoic trend. A single occurrence of course includes only one taxon, so CR must produce a flat line as its quota falls to this level. The problem is that changes in the shape of the diversity curve are easily visible at each step going down from what is clearly a high sampling level (1600 occurrences) to what is normally considered a small but still acceptable level (50 occurrences). Even the logged 1600 occurrence curve has a modest standard deviation of 0.246, and the 50 occurrence curve reduces this value to a mere 0.057.

Nonetheless, the two methods do produce curves with strong underlying similarities that can be demonstrated by cross-correlating the highest ones in each plot ($\rho = 0.840$, $p < 0.001$). Indeed, the lowest curves are even more strongly cross-correlated ($\rho = 0.956$, $p < 0.001$). These correspondences result from the fact that any evening of the occurrence frequency distribution will most likely cause both methods to draw more genera. The real differences therefore have to do with amplitude, not rank ordering.



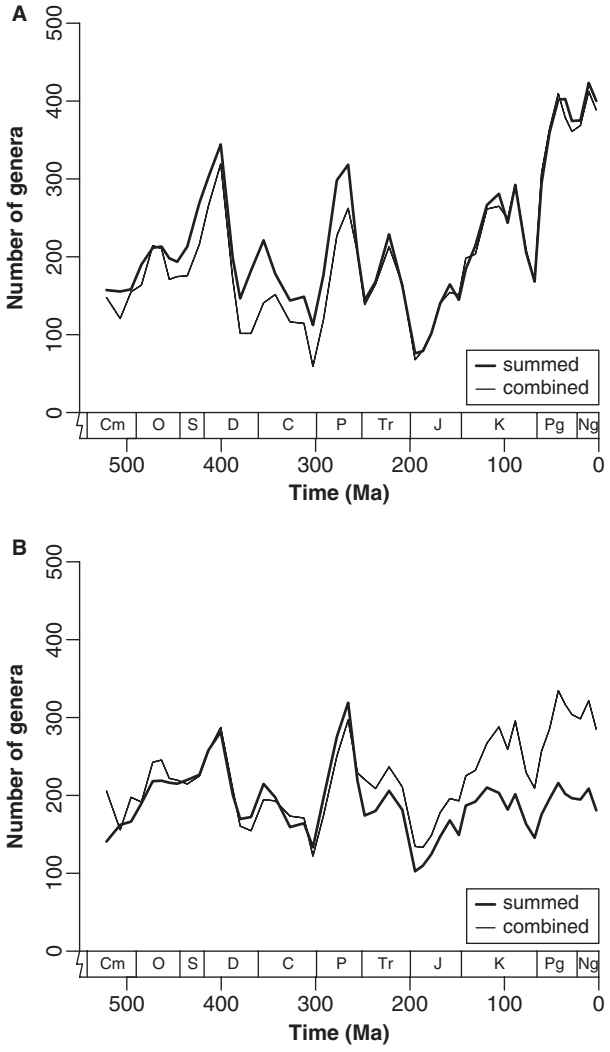
TEXT-FIG. 3. Diversity curves produced by two different subsampling methods at different sampling levels. Curves are not subjected to three-timer correction because this step would restore much of the variance eliminated by the methods, and the point is to show whether such variance scales to the sampling level. A, shareholder quorum subsampling (SQS) with quorums ranging from 0.70 (top line) to 0.20 (bottom line). Early Eocene value in the 0.70 curve is interpolated from neighbouring ones because coverage falls short of that target in that interval. B, classical rarefaction (CR) at levels ranging from 800 occurrences (top line) to 50 occurrences (bottom line).

Thanks to so greatly underestimating the amount of variance, the CR curves recover little of the Cambro–Ordovician radiation, the mid-Devonian drop and the offset between the mid-Cretaceous and Cenozoic. By contrast, the SQS curves imply that there were at least a half-dozen large and rapid increases or decreases during the Phanerozoic. Thus, although all the CR and SQS curves agree with earlier ones in depicting a minor net change between the early Palaeozoic and Cenozoic (Alroy *et al.* 2001, 2008), they have profoundly different biological implications.

Effect of data partitioning

SQS's second important property, robustness to partitioning, can be shown by (1) analysing a complete global Phanerozoic marine data set and then (2) producing separate curves for brachiopods and for other marine animals and adding them together (Text-fig. 4A). Brachiopods are rare in the post-Palaeozoic, so a relatively low quorum of 0.50 was used in this case.

In the SQS analysis, Palaeozoic estimates – and particularly mid- to late Palaeozoic estimates – are consistently higher when the data set is subdivided. The reason is that the high per-taxon abundance of brachiopods conceals the diversity of other groups. However, there is no qualitatively important disagreement with respect to long-term trends or the magnitude of large jumps and crashes in the curve (Text-fig. 4A).



TEXT-FIG. 4. Diversity curves computed either from the combined global data set (thin line) or by summing a brachiopod-only curve and a brachiopods-excluded curve (thick line). A, SQS with a quorum of 0.50 in all cases. B, CR with a quota of 400 occurrences for the full data set and quotas of 200 occurrences each for the brachiopod-only and brachiopods-excluded data sets.

By contrast, with rarefaction the choice of whether and how to subdivide the data effectively determines the outcome (Text-fig. 4B). Specifically, the summation produces a curve that is almost flat after the Permian, so it grows farther and farther apart from the other CR curve. This is true despite having used the three-timer correction to eliminate residual sampling biases and despite having employed seemingly reasonable quotas of 200 brachiopod occurrences and 200 non-brachiopod occurrences.

The reason for the difference is CR's complete inflexibility. Once brachiopods become a minor component of the biota, 'spending' 200 occurrences on them keeps CR from sampling the other taxa as intensely as it should. By contrast, SQS recognizes that the dominant post-Palaeozoic groups must be sampled heavily regardless of whether this by-then rare and depauperate clade is included.

Previous literature has assumed that rarefaction has no such bias. Indeed, some have specifically argued that each major taxonomic group should be analysed separately to avoid 'taxonomic dilution' of its diversity levels when it has low abundance (Westrop and Adrain 1998). This argument turns out to be prescient (see below), but it does not work for CR. Nonetheless, further discussion has focused on whether and how changes in absolute abundance can be demonstrated while accepting the premise that diversity estimates should be made on a group-by-group basis (Finnegan and Droser 2005). CR causes such estimates to be meaningless, while SQS makes them meaningful by disentangling changes in abundance and richness.

The bottom line is that the two diversity curves produced by CR (Text-fig. 4B) cannot be right: one of them is certainly very wrong, and perhaps both of them are. The dilemma is insurmountable even if we only want to create a curve for a single group such as brachiopods. If combined data sets are unreliable (Westrop and Adrain 1998), then we must ask whether the brachiopod subset itself should be subdivided, and if so when we should stop subdividing. If subset curves are unreliable, we still must ask whether the combined diversity curve is accurate because it too is a subset: only marine animal groups with high preservation potential figure in diversity curves, and preservation varies widely even among shelly taxa (Foote and Sepkoski 1999).

Thus, rarefaction's problems are conceptual and not just technical. All other existing subsampling methods also conflate uniform sampling with fair sampling and thereby underestimate amplitude, so all of them create

and destroy patterns depending on how a data set is circumscribed. The fossil record itself is circumscribed by many different factors, not the least being geography; most fossil collections are from certain parts of Europe and North America. Therefore, this problem is lethal. Any method that is highly sensitive to partitioning cannot be expected to produce regional diversity curves that are truly comparable, although rarefied curves (e.g. Miller 1997) are still more likely to be accurate than raw counts.

Effect of data set size

One key problem with earlier standardization methods is that their estimates track the size of the raw data set. Specifically, these methods produce a positive relationship between supposedly standardized taxonomic richness and the number of published references pertaining to each time interval (Alroy *et al.* 2008). This problem was seen not only with raw data but with subsampled diversity curves generated by several earlier standardization methods (Alroy *et al.* 2008).

However, the current study's Δ SQS and Δ reference counts do not correlate at $p = 0.05$ in the global, northern, southern or tropical data ($\rho = 0.245$, 0.106, 0.147, 0.090). Why, then, do the results change? There is a deep connection between this question and the problem of untangling spatial signals because both of the obvious explanations are likely to involve geography.

First, sample size might reflect palaeontological interest that in turn reflects true historical diversity. If so, then high counts in some time intervals are not necessarily problematic, a good method should show a correlation, and SQS may have actually destroyed a signal.

Second, a correlation might only indicate failure to remove a researcher effort bias. Such failure could involve variable geographical concentration of sampling (or database entry), although uneven coverage of taxonomic groups or environments cannot be discounted. A bad method would show a diversity–reference count relationship and a good method would not.

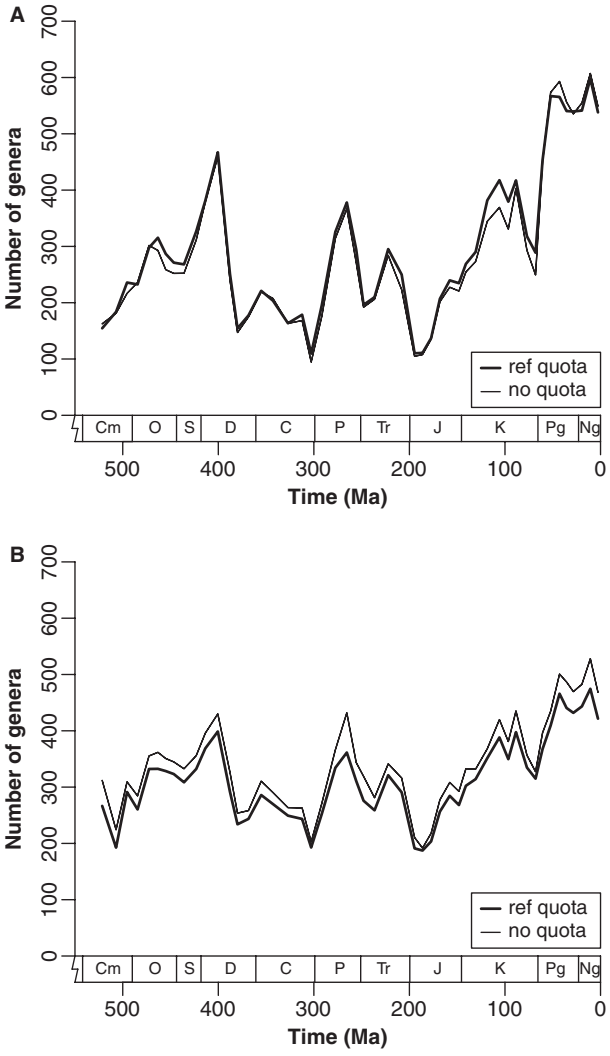
Because SQS's removal of the correlation is consistent with either scenario, a better test of these hypotheses is needed. Such a test should exploit the fact that regardless of what governs reference counts, a good method simply should not care about them. The same result should be obtained even if a database does not reflect worker interest, but instead reflects (say) a decision not to waste time documenting overstudied intervals. Indeed, the marine invertebrate data set was built by seeking to bring every interval up to a minimum standard of coverage regardless of what was in the overall literature (Alroy *et al.* 2008). Therefore, forcing effective researcher effort to be constant should make no difference to the shape or magnitude of a curve.

This prediction can be tested by comparing a curve based on the entire data set to a curve produced by holding the number of available references constant during each subsampling trial (i.e. by imposing a reference quota prior to further standardization: Alroy *et al.* 2008). If accurate, this constrained curve should not be substantially lower. Nonetheless, earlier analyses found that a reference quota did in fact push down the curve regardless of what method was used. At the time, the best of two bad options was to use a quota consistently (Alroy *et al.* 2008).

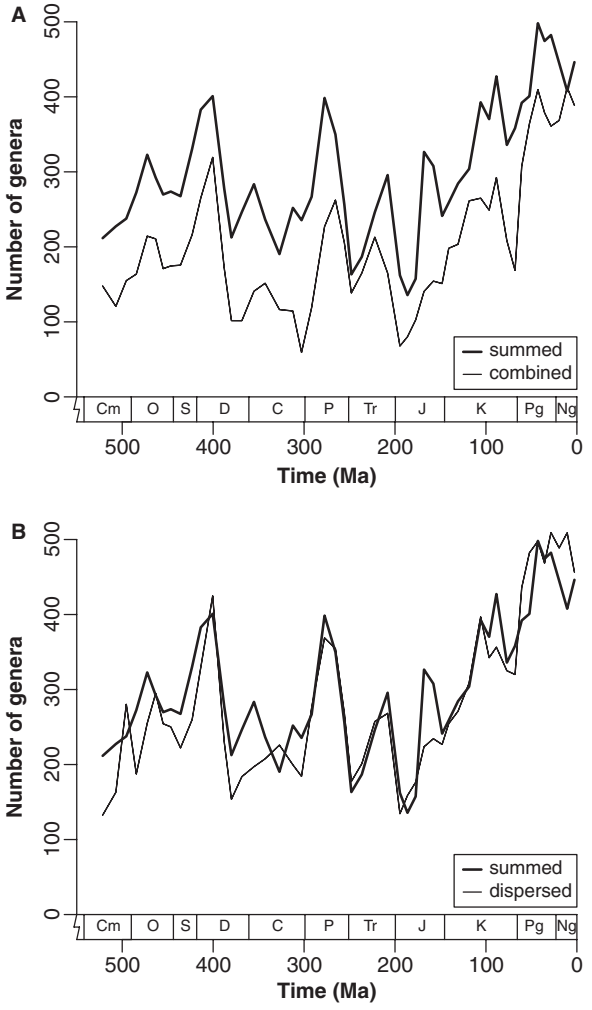
In the current data set, the minimum number of references in each time interval, which sets the highest possible reference quota, is 70. This figure is half the median of 148.5 and far below the maximum of 319. SQS curves generated by using this reference quota and a quorum of 0.60 are offset by a small and inconsistent amount (Text-fig. 5A). The restricted counts are actually higher in 40 of 50 temporal bins, with the difference averaging about 4 per cent (mean ratio = 1.041, median = 1.036). Even more remarkably, there is no large discrepancy in the Neogene even though sampling outside of Europe and North America is exceptionally strong at this time.

By contrast, CR produces a reference quota curve that is unambiguously biased (Text-fig. 5B) despite the fact that it systematically removes variance of all kinds (Text-fig. 3). The average genus count in the restricted analysis is about 8 per cent lower (mean ratio = 0.921, median = 0.924), and not one is higher. The uncorrected, underlying subsampled counts now show a strong cross-correlation between changes in this log ratio and changes in the reference count ($\rho = -0.873$, $p < 0.001$). That is, the more references there are, the more the quota matters. This bias is absent in the analogous SQS data ($\rho = -0.151$, n.s.).

Because the unrestricted data set captures a much greater amount of researcher effort but SQS still produces the same curve, it cannot be the case that its estimates are biased by the volume of raw data. Instead, the lack of a correlation between the global SQS curve and reference counts suggests either that researchers do not study diverse intervals more intensely or that the protocol used to construct the database obscured any such pattern. One way or another, because reference quotas seem to make no difference to SQS they are not used in the remainder of this paper's analyses. It is nonetheless important to point out that changes in raw genus counts are overwhelmingly controlled by changes in reference counts ($\rho = 0.880$, $p < 0.001$), which demonstrates that sheer sample size has a profound influence on unstandardized data (Raup 1976).



TEXT-FIG. 5. Effect of restricting the data set in each subsampling trial to 70 randomly drawn references. Diversity curves generated with a quota (thick line) are contrasted with standard curves (thin line). A, effect on SQS with a quorum of 0.60. B, effect on CR with a quota of 800 occurrences.



TEXT-FIG. 6. Alternative methods of correcting for uneven taxonomic coverage. Quorum is 0.50 in all cases. A, effect of summing 15 separately computed diversity curves for major groups and a 'miscellaneous' category (Alroy 2008). B, effect of throwing back fossil collections during subsampling with a probability inversely proportional to the number of collections yielded by the same published reference.

Correction for taxonomically concentrated data

Alroy (2010) found that summing separate standardized curves for 14 major groups and a 'miscellaneous' category produced consistently higher total diversity estimates throughout the Palaeozoic and Mesozoic (Text-fig. 6A), although the difference amounts to little more than scaling by a ratio that changes in only place (near the Cretaceous/Palaeogene boundary). The lower values for the regular data set resulted from swamping of rare but sometimes diverse groups by more frequent ones such as brachiopods and molluscs.

It would not be feasible to split the data both by taxonomic group and by latitudinal belt. Fortunately, much the same pattern can be generated in a heuristic way by making sure that subsampling draws are more evenly dispersed among published references (Text-fig. 6B). The reason is that references listing many fossil collections tend to focus on dominant groups, whereas references concerning rare groups tend to describe a single fossil collection. The trick is to throw back the randomly drawn collections with a probability of $(N - 1)/N$ where N is the number of collections pertaining to the same reference as the collection in hand. That is, collections are always used if no other collection was described in the same paper, but otherwise might be discarded temporarily. A somewhat related algorithm was used by Alroy *et al.* (2008) to address the similar problem of individual

fossil samples having unequal sizes, but this one more directly addresses the issue of uneven taxonomic coverage.

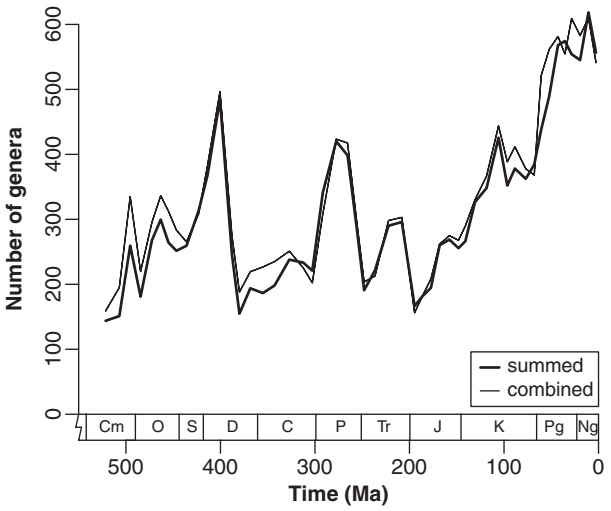
If the quorum level is low, each reference will on average yield the same number of collections in a given analysis (as long as that average is less than one). If the target is very high, throwbacks will make little difference because most collections will have to be drawn eventually anyway.

Regardless of what the additional throwback step does to the sampled distribution of collections across references, it should not affect the outcome if there is no reporting bias to start with. Likewise, taxonomic partitioning should have no effect unless abundant but not very diverse groups are over-studied. Because dispersed sampling and taxonomically split sampling do produce essentially the same pattern by raising the curve (Text-fig. 6B), use of one protocol or the other would seem advisable. The former approach is the only practicable one given the need to subdivide the data spatially, so it used throughout the rest of this paper.

It should be noted that the curves based on taxonomic splitting (Alroy 2010) and dispersed sampling (Text-fig. 6B) do disagree on minor features such as the steepness of the Cambrian rise and mid-Devonian crash and the existence of a late Jurassic peak. Too much attention might be paid to the fact that the dispersed curve suggests consistently high Cenozoic diversity but the summed curve is more erratic. This fact is irrelevant because all sampling standardized curves agree on the key points: (1) Cenozoic diversity is exaggerated in the earlier literature; (2) there was no strong increase within the Cenozoic; (3) high Cenozoic values are driven almost entirely by the Cretaceous radiation of gastropods; and (4) there is ample evidence of logistic dynamics at the global scale (Alroy 2008, 2010; Alroy *et al.* 2008).

Combination of latitudinally restricted data sets

Finally, it is also important to show that SQS generates much the same pattern regardless of whether it is applied to the entire global data set or used to produce separate curves for the northern, tropical and southern regions that are then combined. Genera may be found in multiple regions, so they are tallied not by simple addition but by examining the proportion p of subsampling trials that recover each of them in each belt (respectively p_N, p_T and p_S). The weighted count for a genus is $1 - (1 - p_N)(1 - p_T)(1 - p_S)$, i.e. the chance that it is not overlooked in all three areas. Weighted counts for three timers and part timers are computed with similar equations. Raw genus counts are used when a bin's overall coverage falls short of the 0.40 quorum.

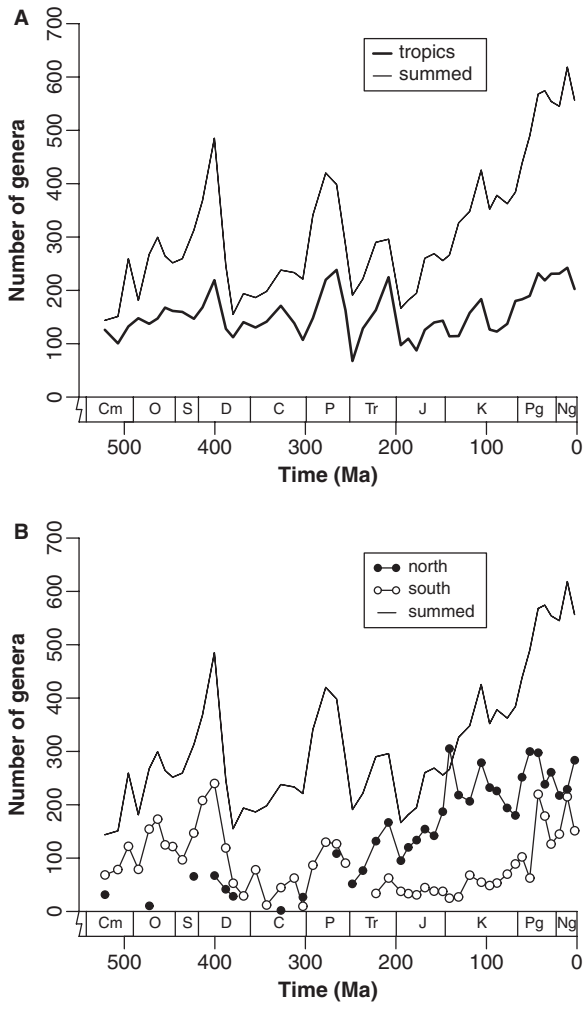


TEXT-FIG. 7. Alternative Phanerozoic marine diversity curves employing global data (thin line) and data for the northern, tropical and southern palaeolatitudinal belts combined after subsampling (thick line). Subsampling quorum is 0.55 for the global data set and 0.40 for the split data sets. The target of 0.55 is selected to make comparisons easier, but the exact figure does not influence the curve's shape (see Text-fig. 3).

The overall and combined curves do resemble each other very closely when they are scaled for comparison (Text-fig. 7). None of the few small differences are really noteworthy. For example, combination yields lower estimates through the Cambrian and Ordovician and again in the Late Devonian and Carboniferous, but trends within these intervals are very similar. Differences between these two curves and the one based on separate analyses of major groups (Text-fig. 6B) are larger but still not very meaningful.

All three regions make important contributions to the global pattern (Text-fig. 8). The tropical curve is relatively flat apart from early Devonian, Permian, Triassic and Cenozoic peaks that also register in the global curve (Text-fig. 8A), which means that the Cambro–Ordovician rise and mid-Devonian crash are almost entirely driven by trends in the south (Text-fig. 8B). The northern curve has large gaps in the Palaeozoic (Text-fig. 8B), but it is likely that nothing important has been missed because little habitat area existed there at the time (Text-fig. 2) and the available Palaeozoic estimates are all low. Post-Palaeozoic northern diversity climbs slowly and is static after the mid-Cretaceous (Text-fig. 8B). Therefore, the Cenozoic peak, which is achieved by the late Eocene, only reflects tropical and southern patterns.

The relative size of this climb may be exaggerated in the southern hemisphere because Cretaceous sampling is not particularly good in key southern areas such as New Zealand and Argentina. Regardless, there is no evidence



TEXT-FIG. 8. Diversity curves for latitudinal belts contrasted with the global pattern (thin line). Global curve is a summation of the regional curves, as in Text-figure 7. A, Tropical diversity (thick line). B, Northern higher latitude diversity (filled circles) and southern higher latitude diversity (open circles).

from any treatment of the data set that a strong radiation occurred anywhere within the Cenozoic, and certainly not in the tropics. This fact is contrary to much speculation in the older literature (e.g. Valentine 1970; Benton 1995; Bambach 1999).

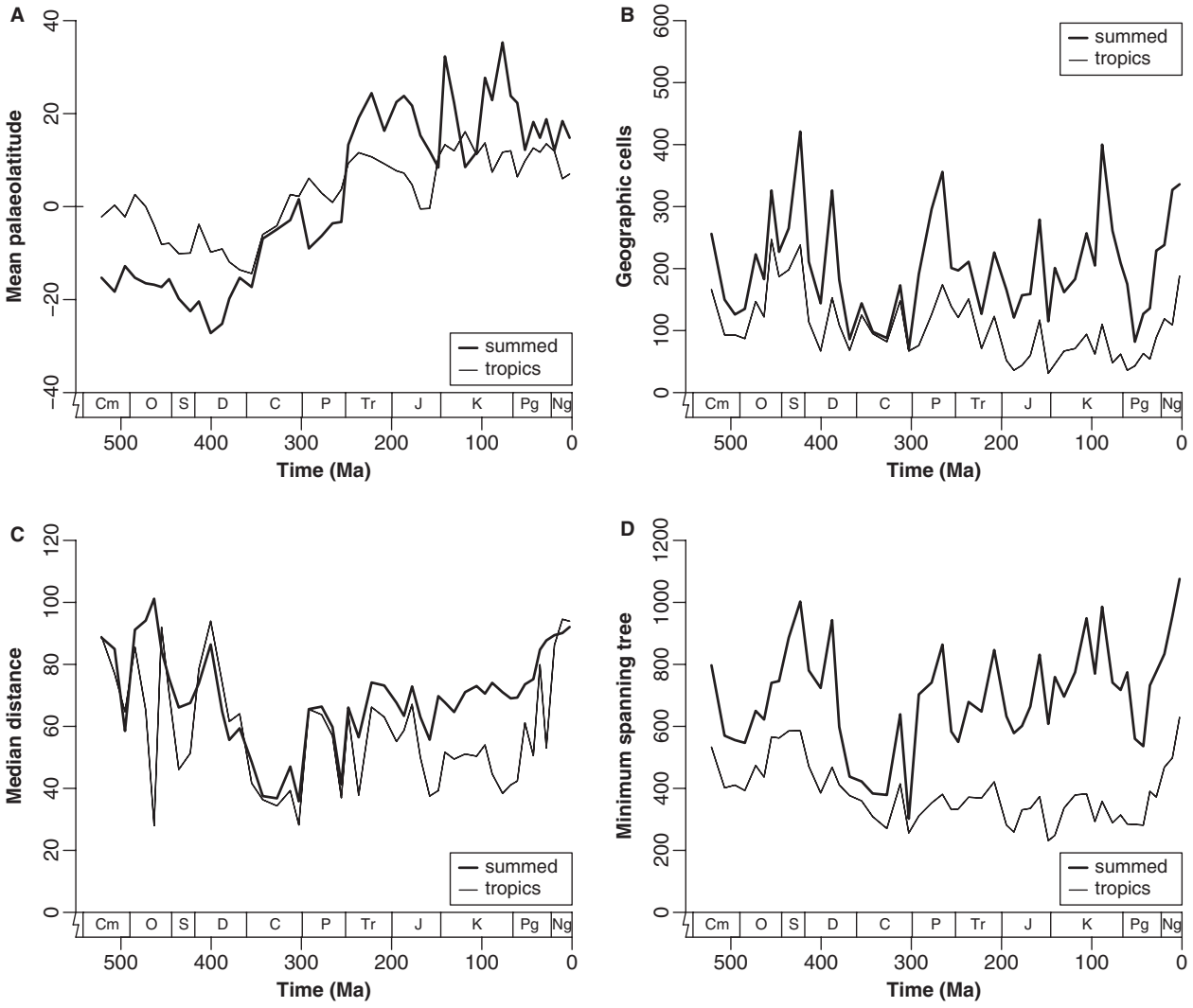
Geographical variables

Several palaeogeographic variables were used to capture the effects of changing latitude, area and continental dispersion (Text-fig. 9). To avoid monographic effects, statistics were computed for palaeogeographic cells instead of individual fossil collections. Cells were defined to be 1 degree on a side and counted only if they were occupied by collections. Based on the logic illustrated in Text-figure 1, both raw and differenced (Δ) values were used as independent variables. The four main categories were as follows:

1. Palaeolatitude and the absolute value of palaeolatitude, which capture changes in palaeocontinental positions (Text-fig. 9A). Absolute values were investigated because more of the literature dwells on temperate–tropical zone comparisons (e.g. Stehli *et al.* 1969; Roy *et al.* 1996, 1998, 2000) than on north-south gradients per se. The nominally tropical collections generally do fall close to the equator, with the mean cell position never straying more than 16 degrees from it. The global trend very roughly tracks the average palaeolatitude of North American and European marine sedimentary rocks (Allison and Briggs 1993), although Cretaceous means in the new data are not much higher than Cenozoic means.
2. The raw number of distinct occupied cells, which measures geographical area (Text-fig. 9B). The grid size is so coarse that this proxy should be relatively robust to sampling intensity. Furthermore, it would be unwise to adjust the counts for the number of available collections because it is possible that collection counts track original habitat area (Peters 2005) instead of, say, variation in preserved rock amount that is unrelated to palaeogeography (Raup 1976).
3. The median great circle distance (GCD) between all pairs of occupied cells, which reflects a combination of area and dispersion (Text-fig. 9C). The main reason for using the median instead of the mean is that it is consistently lower, indicating a skewed distribution.
4. The summed length of a minimum spanning tree (MST) connecting all the cells, which measures a combination of the last two factors (Text-fig. 9D). This statistic should be more resistant than the median GCD to heavy spatial skewing of the data or large geographical gaps in the data. For example, if there are 10 very closely spaced cells plus one other cell very distant from the others, the median GCD will be very small but the summed MST will be large. If there are ten cells in each of these two regions, the median GCD will connect a pair in different areas because there are 90 intraregional pairs but 100 interregional pairs. However, the MST will change very little. Likewise, adding points close to the line that connects the regions will have little effect on the summed MST but will lower the median GCD.

Environmental variables

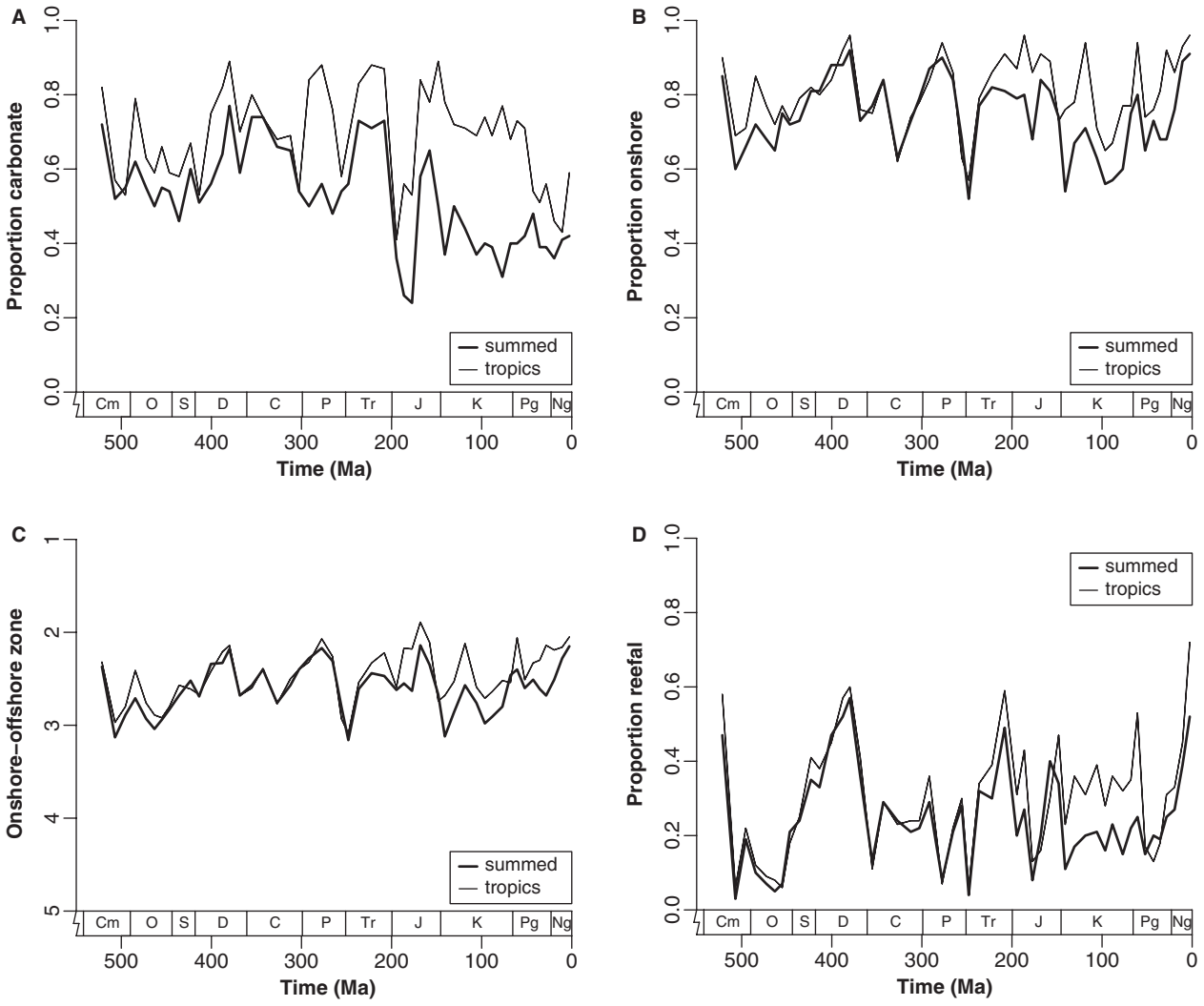
The lithology and environmental context of sediments yielding fossil collections are routinely recorded in the



TEXT-FIG. 9. Phanerozoic trends in geographical measures. Data for the tropics (thin lines) are contrasted with data for the entire globe (thick lines). A, mean palaeolatitude of geographical cells occupied by the fossil collections used in this study. Each cell measures one palaeolatitudinal and palaeolongitudinal degree on a side. B, number of cells sampled at least once. C, median great circle distance between all pairs of occupied cells. D, length of minimum spanning tree connecting all occupied cells.

PaleoDB, making it possible to quantify at least four useful variables (Text-fig. 10) very similar to ones used in a recent study of the Triassic and Jurassic (Kiessling and Aberhan 2007):

1. The proportion of cells including collections that come from carbonates, as opposed to siliciclastic rocks (Text-fig. 10A). Common examples in the former category include limestone, lime mudstone and dolomite; examples in the latter include sandstone and shale. Two major lithologies can be recorded for each collection, so a collection is put aside if both carbonate and siliciclastic lithologies were present and it is not clear which one yielded the fossils.
2. The proportion coming from onshore environments, defined as marginal marine, shallow subtidal or deep subtidal (Text-fig. 10B), as opposed to offshore environments (including continental slopes and basins).
3. The mean onshore–offshore zone on a five-point scale (Text-fig. 10C). In order, the categories are marginal marine, including peritidal areas, lagoons, estuaries, bays and delta plains; shallow subtidal, including foreshore, shoreface, sand shoal, delta front and all reefal environments; deep subtidal, including the shoreface–offshore transition zone; offshore, including prodeltas; and slopes or basins. This classification broadly resembles those used in other studies (e.g. Sepkoski 1988).
4. The proportion coming from reefal environments (Text-fig. 10D). These data are congruent with both qualitative (Copper 1988) and quantitative (Kiessling



TEXT-FIG. 10. Phanerozoic trends in environmental measures. Line thicknesses are as in Text-figure 9. A, proportion of cells sampling carbonate lithologies. B, proportion sampling onshore environments. C, mean onshore–offshore depth zone of cells. D, proportion sampling reefal environments.

2002, 2009) assessments of reef disappearance and growth through the Phanerozoic.

All calculations ignored cells with missing data where appropriate. There was no weighting by collection count for the purposes of computing proportions and averages. In other words, if a cell included one collection, it contributed no more or less to a given proportion or average than a cell with 100 collections.

RESULTS

Equilibrial dynamics

Density dependence in time series leads to logistic growth and the maintenance of equilibria (Sepkoski 1978). Such behaviour is hard to show when time series have high sampling error (Alroy 2008), but the current data seem not to, as demonstrated by visibly strong correlations of neighbouring data points in the diversity curves (i.e. serial correlations). Fitting an exponential curve to the global data set and taking residuals does not remove this serial correlation ($\rho = 0.695$, $p < 0.001$). Another way to measure noise in the data set is the wobble index (Alroy 2008), which quantifies the size of short-term departures from trends. The median wobble of the data points is 0.161, substantially lower than the figure of 0.308 for the raw data (Text-fig. 6B). This value is also not much higher than the one computed for a previously published Phanerozoic curve that had lower amplitude thanks to the subsampling method that created it (Alroy *et al.* 2008).

Given that the data are well behaved, the most simple test for density dependence is to correlate changes in logged diversity with standing diversity in immediately preceding intervals (Alroy 2008). In this case, even a two-tailed nonparametric correlation is marginally significant ($\rho = 0.332$, $p = 0.020$) and the relationship has the predicted sign, so the global data do suggest saturation.

Unfortunately, there is more noise in the tropical and particularly southern data sets, and the northern data set is too incomplete in the Palaeozoic to perform dynamic analyses (Text-fig. 8). The first two time series have median wobbles of 0.229 and 0.463, as opposed to 0.433 and 0.572 in the raw data. On the other hand, after detrending they still have nonrandom fine-scale patterns, as measured by serial correlations ($\rho = 0.549$ and 0.684, $p < 0.001$ and 0.001).

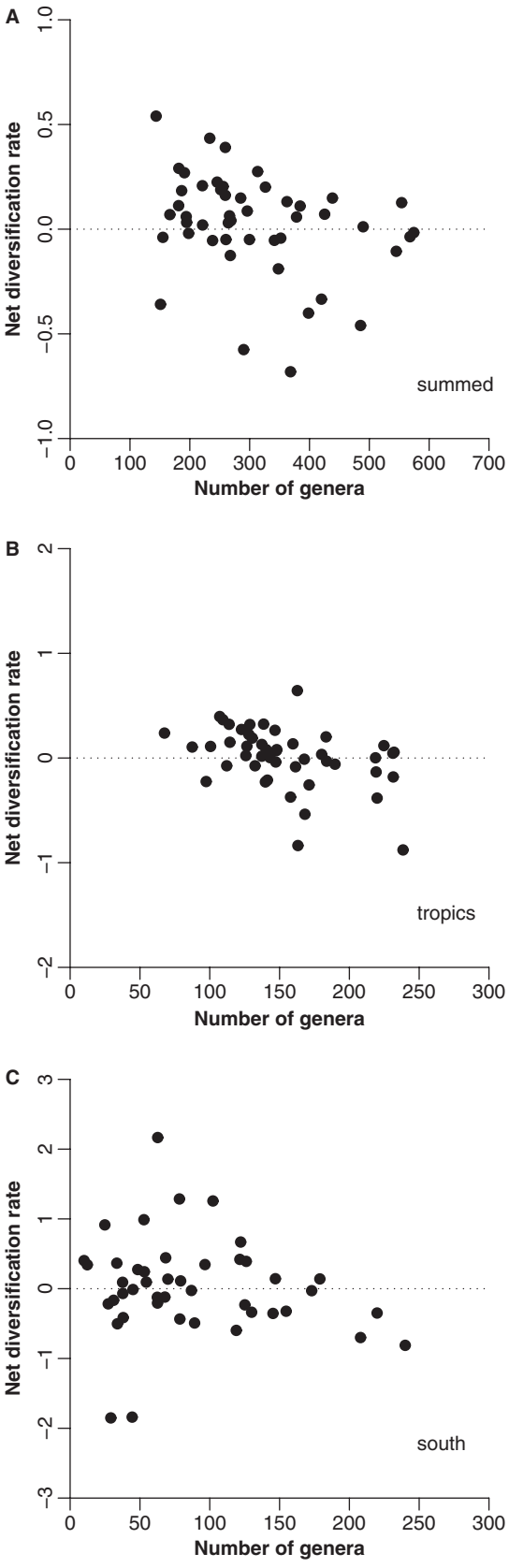
The well-known regression to mean effect nonetheless becomes serious when there is so much sampling error (Alroy 2008), so a simple correlation-based density dependence test cannot be justified. Fortunately, there is no such problem when it comes to correlations between standing diversity levels and changes two intervals in the future (instead of one). The reason is that entirely independent counts are involved.

As it happens, the global data produce a slightly stronger correlation with the additional lag ($\rho = 0.354$, $p = 0.014$: Text-fig. 11A). A similar value is seen in the tropical data ($\rho = 0.471$, $p < 0.001$: Text-fig. 11B), but there is no relationship in the southern data ($\rho = 0.145$, n.s.: Text-fig. 11C). All three values are essentially unchanged when the residuals produced by fitting exponential lines are substituted for the raw diversity values ($\rho = 0.330$, 0.452, 0.171; $p = 0.022$, 0.001, n.s.).

These results are not completely surprising given the lack of strong upwards trends in the curves (Text-fig. 8). Nonetheless, there is no *a priori* reason to think any saturation effect would be seen so far into the future, so the global and tropical patterns are particularly suggestive of genuine equilibrial dynamics.

Combined models of diversity dynamics

Simplistic multiple regressions involving both the plethora of explanatory variables (Text-figs 9, 10) and their first differences would be too cumbersome to be interpretable. Furthermore, these variables are often very strongly correlated. So, for example, trimming down the full set through stepwise regression would not work well



TEXT-FIG. 11. Correlations between absolute diversity levels within each time interval $t - 1$ and rates of change in diversity between t and $t + 1$. A, summed global data. B, tropical data. C, southern data.

because it would put an undue focus on a few selected variables when others could just as easily have been picked.

Instead, redundancy is addressed here by performing a principal components analysis on all the geographical and environmental variables and then using the axis scores in multiple regressions. Prior to ordination each variable is centred and scaled to have the same variance. For reasons stated previously, standing diversity is also used as an explanatory variable in these regressions. It is not log transformed because a standard Ricker model of density dependence implies that the net diversification rate should be a function of diversity on a linear scale. Standard *F*-ratio tests are used to identify significant predictors. The ordination axes are by definition independent, so it makes no real difference how many are examined at once for the purpose of conducting these tests. Finally, loadings on axes contributing significantly to the regressions are examined to determine which variables best explain changes in diversity.

To address the regression to the mean problem, analyses are performed a second time after lagging the diversity data back one interval. In other words, the regression model is reconfigured to explain the change in diversity from time t to time $t + 1$ as a function of standing diversity within $t - 1$ plus axis scores that summarize physical variables within t and differences of those variables between t and $t + 1$.

In the global data set and without lagging diversity backwards, only axis 1 and standing diversity are significant predictors ($p = 0.014$, $p = 0.020$; 19.4 per cent variance explained (VE)). Exactly the same result is seen when standing diversity is lagged back another interval ($p = 0.019$, $p = 0.018$; VE 20.2 per cent). Axis 1 in the physical factor ordination contrasts the four flavours of onshore–offshore variables with the four carbonate and reef variables, so it is dominantly environmental. Axis 2, which again explains no added variance, pulls out the four variables related to geographical cell counts and minimum spanning tree lengths. It is therefore a clear-cut proxy for habitat area.

In the tropical analysis, only axis 2 and diversity at time t are important and the regression model is more powerful ($p = 0.003$ and $p < 0.001$; VE 36.0 per cent). Lagging the taxon counts flags the same variables but weakens the result ($p = 0.025$, $p = 0.005$; VE 25.5 per cent). Axis 2 again primarily captures variation in environmental variables, contrasting depth measures with reef and carbonate measures. Axis 1 relates to latitude, and axis 3 is again geographical and again irrelevant. That is, because shelf area in the tropics changes only very slowly (Text-fig. 9D), tropical richness is primarily controlled by the frequency of onshore carbonate environments (Text-fig. 10).

Finally, in the southern data set standing diversity is irrelevant regardless of whether it is lagged, but the first three axes are all significant predictors or nearly so ($p < 0.001$, $p < 0.001$, $p = 0.029$; VE 48.6 per cent). Lagging also makes little difference ($p < 0.001$, $p < 0.001$, $p = 0.069$; VE 47.1 per cent using axes 1 and 2). Axis 1 is primarily a function of pole-to-pole latitude and distance to the equator, axis 2 captures the six variables related to reef proportion and onshore–offshore position, and axis 3 also mostly has to do with latitude and depth, although it mixes in some of the geographical area variables.

Alternate models of diversity dynamics

The remaining problem is how to relate these empirical patterns to the general models outlined in Text-figure 1. If physical factors control the average rate of diversification but do not set limits (Text-fig. 1C), we should expect raw values to be better predictors of rates than changes. The opposite should hold if these factors set hard limits to diversity (Text-fig. 1D). The obvious test is to exclude the differenced values in one set of analyses and then exclude the raw values in a second set. To simplify the discussion, only the regression analyses including lagged standing diversity as a predictor are detailed; analyses not lagging diversity yield similar results.

The global data set's axis 1 drops out when raw values are used, so standing diversity becomes the only significant predictor of diversification rates – but not a good one ($p = 0.031$; VE 9.7 per cent). The tropical model again involves its axis 2 (onshore–offshore position and reef/carbonate area) and standing diversity ($p = 0.021$, $p < 0.011$), and likewise is no worse (VE 26.2 per cent). In the southern analysis, the newly computed axis 2 stands alone ($p = 0.001$) and explains much less (VE 21.2 per cent). It also now has a mixed signal, with the most heavily loading variables being reef proportion, cell counts, mean onshore–offshore zone and summed MST.

Using differenced values, the results for global diversity are the same as in the full model; axis 1 and standing diversity are both significant ($p = 0.045$, $p = 0.026$; VE 17.5 per cent). In the tropical analysis, the environment drops out, leaving standing diversity and taking much of the explanatory power with it ($p = 0.004$; VE 16.7 per cent). Finally, in the southern analysis the model changes a second time but the fit is as good as in the full model, with only axes 1 (latitude and area) and 2 (environment) being significant ($p < 0.001$ for both; VE 51.2 per cent).

DISCUSSION

Dynamical models

Several key inferences can be drawn from these results. The most obvious is that because different diversity trends (Text-fig. 8) and causal factors (Text-fig. 11) are seen globally and within each latitudinal belt, there is clear support for the general hypothesis that trends at different scales are governed by partially independent factors (Miller 1997).

The results also yield yet more evidence for the importance of density dependence, which has been shown to operate in this data set previously (Alroy 2008) and has an effect on most major groups (Alroy 2010). However, standing diversity terms are only significant when habitat area is roughly constant – that is, in the global and tropical data but not the southern data.

Environmental variables usually relating to depth are important in all three cases. Their importance suggests that onshore environments have higher carrying capacities. In the global and tropical data, diversification is particularly influenced by the dominance of reefs and carbonate environments, which is consistent with recent work on the interaction between environmental factors and origination rates that used a largely overlapping data set (Kiessling *et al.* 2010). By contrast, the most rigorous previous study to have employed sampling standardization and multiple regression found evidence for equilibrial dynamics but not for environmental controls during the Ordovician radiation (Connolly and Miller 2002).

However, proxies for shelf area have no definite influence on any of the data sets. This pattern is not consistent with the idea that logistic dynamics are driven by simple species–area relationships (MacArthur and Wilson 1967; Sepkoski 1976; Peters 2005). Likewise, the median distance between samples is not a strong predictor in any analysis, so there is no clear evidence from this study that diversity is a strong function of dissimilarity among regional biotas (i.e. geographical beta diversity) created simply by geographical distance.

Nonetheless, the apparent importance of latitude in the southern data might actually be an echo of a species–area effect because Δ latitude and Δ summed MST have an inverse relationship in the south ($\rho = -0.314$, $p = 0.023$). For example, in the mid-Devonian this proxy for area decreases (Text-fig. 9D) while collections move to the northern edge of the southern belt; by the Permian, area increases again and points move southwards.

There is mixed evidence with respect to whether changes in physical factors are more important than static values. Changes are definitely the key drivers in the south, where diversification is most predictable but simple density dependence is absent. However, the evidence is ambiguous for the global and tropical data. The southern pattern indicates that equilibria vary greatly but do exist, and that they are tracked fairly quickly on a geological time scale. That is, the evidence favours the causal model illustrated in Text-figure 1D, which cannot be shoehorned into a simple narrative contrasting biotic and abiotic factors because both matter. The direct correlation between diversity and net diversification in the global and tropical data (Text-fig. 11) in combination with environmental effects is also consistent with this more synthetic paradigm.

In sum, the most parsimonious conclusion is that diversity is in fact equilibrial, but limits increase whenever shelf habitats are extensive or shallow-water carbonate environments, and reefs in particular, are common. Latitude and geographical area may or may not be important outside of the tropics, but neither thing drives legitimately exponential growth rates.

These general conclusions are clear-cut, so the only outstanding question is whether the physical variables reflect preservational bias or historical reality. Of most concern, there would be a problem if (1) the geographical variables were strongly influenced by the geographical breadth and concentration of sampling instead of historical trends, and (2) the diversity curves were biased by these same factors. However, patterns in the curves are not qualitatively consistent with this interpretation. For example, richness estimates do not rise anywhere in the Neogene (Text-fig. 8) despite unusually strong geographical coverage during this interval (Text-fig. 9B). Furthermore, independent regressions would show that summed MST is a stronger predictor than simple cell counts, and MST designed to be less vulnerable to sampling effects.

These concerns cannot be dismissed entirely, so individual cases where shifts may correspond to gaps in sampling are discussed below. Such problems may pertain both to geography (e.g. several large swings in cell counts: Text-fig. 9B) and to environment (e.g. two Jurassic outliers in carbonate proportion: Text-fig. 10A).

Methods and data

It has been suggested previously that global Phanerozoic diversity patterns cannot be understood unless they are dissected both geographically and environmentally (Miller 1997, 1998; Connolly and Miller 2002), and in fact much is learned through such an exercise (Text-figs 8–11). However, it must be emphasized that the accuracy of the global diversity curve (Text-fig. 7; Alroy 2010) seems not to be compromised in any way by differences among regions (Text-fig. 8).

The real problems are therefore not a matter of scale or data but a matter of concepts and methods: what do we consider to be accurate diversity estimation, and which methods meet this criterion of accuracy? It is now clear that all sampling standardization methods heretofore used in the palaeobiological literature (e.g. Alroy 1996, 2000; Alroy *et al.* 2008) are premised on a conflation of uniformity with accuracy that leads to underestimation of variance (Text-fig. 3B) and internally inconsistent estimates (Text-figs 4B, 5B). Imposing uniformity amounts to assuming that all categories are equally diverse until proven otherwise, which might seem persuasive as a null hypothesis but also guarantees spurious results.

These problems are not trivial because most palaeontological studies employ a single method, classical rarefaction, that demonstrably flattens and otherwise distorts the global curve (Text-fig. 3). A few examples out of many include Miller and Foote (1996), Kammer *et al.* (1998), Westrop and Adrain (1998), Olszewski and Patzkowsky (2001), Krug and Patzkowsky (2004), Finnegan and Droser (2005), Kiessling *et al.* (2007, 2008, 2010), Clapham *et al.* (2009) and Hendy (2009).

However, rarefaction is much more than a palaeontological method applied to taxonomic occurrences. It is arguably the gold standard of diversity estimation in the ecological literature (Gotelli and Colwell 2001), where it is far more often applied to within-collection specimen counts than among-collection occurrence counts of the kind treated here. According to ISI Web of Knowledge data surveyed in September, 2010, the keystone paper on rarefaction in the palaeobiological literature (Tipper 1979) has been cited 127 times, whereas the analogous paper in ecology (Hurlbert 1971) has been cited well over 1400 times.

Even worse, ecologists usually seek to construct equal-sized samples to start with. They turn to rarefaction only when, say, looking at fixed geographical areas or for fixed durations of time is not practical or yields highly variable numbers of individuals. The logic of SQS challenges the very notion that experimental designs should predefine sample sizes. Instead, it suggests that sample sizes should be allowed to increase until quorums are attained, which means that sampling effort should not be planned ahead of time.

Despite everything, it must be admitted that rarefaction has been preferred for a very sound reason: previously, the only clear alternative was to extrapolate out from the data to an asymptotic pool size estimate. The many alternative methods for doing so (1) are assumption-laden, (2) provide noisy estimates, and (3) usually provide at best a lower bound on the true number of species (Colwell and Coddington 1994). Thus, when most taxa remain undescribed we truly have no good alternative to subsampling of some kind. Nonetheless, rarefaction's deadly embrace of the null hypothesis has now rendered a large body of literature entirely suspect.

Previously published alternative subsampling methods may do a better job of imposing uniformity than does rarefaction (e.g. Alroy 1996, 2000; Alroy *et al.* 2001, 2008), but that does not mean they are any more accurate. We instead need methods that capture high-amplitude trends and yield consistent patterns regardless of how the data are subsetted or limited (Text-figs 3–8). Although the particular shareholder quorum algorithm used here may eventually be discarded in favour of a better one, it is therefore likely that future methods will also seek to impose uniform frequency coverage instead of uniform sample sizes.

Even granting that a methodological advance has been made, some limitations to the current data set need to be remedied by additional compilation. Gaps in the northern temperate zone need to be filled (Text-fig. 8B). There is an overall deficit throughout the Carboniferous and too much concentration of Carboniferous data in the then-equatorial continent of Laurentia (Text-figs 2, 9). An early Cretaceous excursion in sampling to the south (Text-fig. 9A) is also most likely artefactual. It is not possible to improve the geographical range of sampling everywhere, so a satisfactory solution to these problems may also need to involve more sophisticated methods of dispersing sampling (Text-fig. 6) or balancing the weight of geographical regions (Text-fig. 7) than the ones employed here.

Mass extinctions

The statistical results provide a useful summary, but apparent cases where a singular event had a major effect on taxonomic richness do merit attention. These are discussed here in chronological order, with the summation of three separately computed curves for latitudinal belts (Text-fig. 7) serving as the main reference point.

All of the traditionally identified Big Five mass extinctions (Raup and Sepkoski) deserve some comment even though it makes more sense to talk about a 'Big Three' because the first two do not stand out strongly against revised extinction rates also derived from PaleoDB occurrences (Alroy 2008). The first four correspond with crises in reefal carbonate production, although richness per se was not always affected and reef development was only briefly interrupted at most of these times (Kiessling 2009).

The end-Ordovician mass extinction. The first widely discussed mass extinction came at the end of the Ordovician. The net loss of global biodiversity within a few Myr of the event was nil (Krug and Patzkowsky 2004; Text-fig. 7), tropical diversity did not change (Text-fig. 8A),

and the extinction rate did not diverge remarkably from the average for the early Palaeozoic (Alroy 2008). Nothing in the environmental data at this scale of resolution suggests a crisis at this time (Text-fig. 10) although reef development per se was briefly constricted immediately after the extinction (Kiessling 2009).

That said, there are important changes in geography even though the number of occupied cells and the median distance between cells remain about the same (Text-fig. 9B–C). First, the minimum spanning tree increases in length by 19 per cent and continues to rise through the Silurian. Second, the average latitude of the sampled cells jumps 4 degrees away from the equator as sampling moves southwards. The first pattern cannot account for the extinction because greater habitat area should mean greater beta diversity (Schopf 1974; Sepkoski 1976, 1988). The second is also not a plausible biological mechanism because it most likely reflects overconcentration of sampling in eastern Laurentia, Baltica and Avalonia, all of which were still well south of the equator at this time.

In sum, the end-Ordovician extinction was relatively small and cannot be tied convincingly to any of this study's historical variables, so it is not possible to show here that it reflected global physical perturbations. It may well be best to interpret this episode as a summation of more local events (Miller 1998) and exclude it from the Big Five.

The mid-Devonian crash. Despite the incontrovertible fact that the Permo-Triassic extinction was much more severe than any other during the Phanerozoic (Alroy 2008, 2010), net declines in diversity that were just as large in proportional terms took place in the mid-Devonian and across the Triassic-Jurassic boundary. Extinction rates were not particularly elevated going into the Late Devonian, so this drop relates more strongly to depressed origination rates (Alroy 2008) and like the end-Ordovician episode does not merit inclusion in a Big Five list.

The Late Devonian and end-Triassic both corresponded with a great reduction in the number of reef sites (Kiessling 2002, 2009), and there indeed were very sharp decreases in the proportion of reefal cells at both times (Text-fig. 10D). These two shifts are unlikely to be artefacts because the among-cell reef proportion curve also closely matches the Phanerozoic trend in the relative abundance of corals within fossil collections (Kiessling *et al.* 2008). Furthermore, the mid-Devonian decline is present in the older data of Sepkoski (1997) and cannot be accounted for as an artefact of rock outcrop area (McGowan and Smith 2008).

More worrisome concerns about the quality of the Devonian data are raised by the fact that the number of sampled cells and their dispersion both fall through this period (Text-fig. 9B–C). However, the reality of this event's great apparent magnitude (Text-fig. 7) is affirmed not only by the presence of an independently demonstrated environmental shift with biological underpinnings at this time, but by the long-term nature of the shift – there was no clear recovery until the mid-Permian, a period of more than 100 myr that witnessed large changes in palaeogeography and the environment.

The Permo–Triassic mass extinction. It comes as no surprise that diversity declined precipitously across the Permo/Triassic boundary both globally and at low latitudes (Text-fig. 8A). There is good evidence that this extinction was caused by environmental perturbations triggered by Siberian Trap volcanism, such as anoxia, global warming and the direct effects of elevated CO_2 levels (Knoll *et al.* 2007). None of these factors are likely to have persisted on a time scale long enough to figure in this study's environmental time series, which in any case do not directly relate to them (Text-fig. 10). Nonetheless, large changes in ocean chemistry do explain why this greatest of all extinctions was strongly selective against reefal organisms such as sponges, corals and bryozoans (Stanley 1988; Knoll *et al.* 2007; Alroy 2010), and therefore why the proportion of reefal cells is extremely low in the Early Triassic (Text-fig. 10D).

The Triassic–Jurassic mass extinction. Although it has been known for many years that the end-Triassic witnessed a very large extinction (Raup and Sepkoski 1982; Kiessling and Aberhan 2007; Alroy 2008), the new data suggest it had a much larger medium-term effect than even the Permo–Triassic crisis (Text-fig. 7). The immediate cause of the extinction is not very well understood, but was likely to have involved physical perturbations such as global warming that may have been catastrophically rapid (McElwain *et al.* 2009). Nonetheless, it has already been shown that a combination of extinction and reduced origination around the Triassic-Jurassic boundary targeted infaunal taxa and organisms common in the tropical/carbonate/onshore/reefal environmental complex (Kiessling and Aberhan 2007; Kiessling *et al.* 2007; Mander and Twitchett 2008).

The distribution of reefal environments was quite broad in the latest Triassic, and the drop was very sharp (Stanley 1988; Kiessling 2002, 2009; Text-fig. 10D). Although reef habitats were present in the Early Triassic, they were spatially restricted and uncommon (Kiessling 2002). There was also a massive, medium-term decline in the frequency of carbonate environments (Text-fig. 10A). By contrast, there is no excursion in the onshore–offshore zone data (Text-fig. 10B–C; Kiessling and Aberhan 2007), which is not surprising given that the sea level lowstand near the boundary was brief (Mander and Twitchett 2008). Furthermore, none of the geographical variables suggest a large change (Text-fig. 9), which rules out geographical

sampling bias as a cause of the pattern. Thus, all evidence points to reef collapse, possibly related to global warming, as a major factor in this mass extinction (McElwain *et al.* 2007; Kiessling 2009).

The Cretaceous–Palaeogene mass extinction. No event in the deep fossil record is better studied than the global extinction that ended the Mesozoic and is now agreed by all reasonable parties to have been caused by a large extraterrestrial impact (Schulte *et al.* 2010). Marine invertebrate survivorship patterns are consistent with a short-term decline in primary productivity that favoured small, infaunal, deposit-feeding and nonplanktotrophic taxa (Hansen *et al.* 1993; Aberhan *et al.* 2007). By contrast, if there was any global climate shift at this exact time, it had to have been short-lived and without any long-term effect (Wilf *et al.* 2003).

This study's geographical variables are clearly not relevant to an impact scenario, and indeed they show no interesting shifts across the boundary (Text-fig. 9). Of slightly more interest, the environmental variables also are static (Text-fig. 10). This fact comes as no great surprise because both the current data set (Text-fig. 10D) and more detailed analyses (Kiessling 2009) suggest there was no major loss of reefal habitats.

However, the new sampling-standardized curves do yield a striking insight about the Cretaceous–Palaeogene extinction. The Palaeocene values seen in all three latitudinal belts (Text-fig. 8) are as high as, or even higher than, the Maastrichtian values. Thus, marine invertebrates recovered everywhere from the event within just a few million years, entirely unlike the situation in the Early Triassic and Early Jurassic.

Evolutionary radiations

The Cambro-Ordovician radiation. Efforts to reconstruct diversity through the Cambrian and earliest Ordovician are complicated by extremely high turnover rates during this interval (Raup and Sepkoski 1982; Alroy 2008). When turnover is rapid and mostly occurs within an interval instead of at its boundaries, more and more taxa join the sampling pool despite never having coexisted. This problem can make estimates of overall pool size – the desideratum of shareholder quorum subsampling – essentially meaningless. It is strictly analogous to the long-understood bias introduced by variation in time interval durations (Raup 1972), which is also only relevant when turnover is continuous.

The rate bias cannot be handled simply by dividing counts by interval lengths because the relationship between rates and overall pool sizes is expected to be nonlinear (Raup 1985). More and more taxa become confined to single intervals when rates are high, so another approach would be to simply discard those taxa. Sepkoski (1997) did just that for a different reason (i.e. hoping to remove sampling biases). However, doing so also makes matters worse because it overcorrects (Alroy *et al.* 2008), creating the appearance of very low Cambrian diversity in Sepkoski's compendium (Sepkoski 1997) and some treatments of the PaleoDB data (Alroy *et al.* 2001). Finally, it could be assumed that the rate bias will be cancelled out by the decreasing probability of sampling individual taxa when their ranges within intervals are very short (Alroy *et al.* 2008), but if the current sampling method is as accurate as is claimed, this contrary bias should not exist.

One way or another, in the current data set there is probably no large bias, because several hypothesis tests suggest that most marine turnover is concentrated at boundaries between intervals (Foote 1994; Alroy 2008, 2010). An unrelated problem is low evenness of local abundance distributions in the early Paleozoic (Peters 2004; Alroy *et al.* 2008), which simply makes it harder to sample rare taxa. Standardization methods drawing fixed numbers of items are influenced strongly by evenness, sometimes by design (Alroy *et al.* 2008), and it is possible that SQS is still vulnerable to it. However, instead of low Cambrian values both the current analysis (Text-fig. 7) and a recent one using different standardization methods (Alroy *et al.* 2008) generated surprisingly high counts followed only by small increases.

These trends suggest that rapid Cambrian diversification at low taxonomic levels allowed an equilibrium to be approached almost immediately, consistent with the extremely rapid filling of morphospace at this time (Briggs *et al.* 1992). Thus, the most interesting patterns throughout the Cambro–Ordovician interval, as opposed to those seen near its base, are relatively rapid shifts in the dominance of major taxonomic groups (Sepkoski 1984; Alroy 2010).

The Permian radiation. Most older diversity curves suggested no major changes throughout the mid- and late-Palaeozoic (e.g. Sepkoski 1997), but curves based on the PaleoDB (Alroy *et al.* 2008) depict a sharp mid-Permian diversification that is confirmed by this study's improved standardization methods (Text-fig. 7). The jump is both relatively and absolutely larger in the tropics than elsewhere (Text-fig. 8), and it mostly reflects an evolutionary radiation of strophomenate brachiopods (Alroy 2010).

Modern studies of Permian diversity and turnover have mostly focused on extinctions at the end of the Guadalupian and Lopingian (e.g. Clapham *et al.* 2009), and lack of interest in immediately preceding diversification patterns may simply reflect the traditional lack of evidence for any large, global transition between the Carboniferous and Early Permian (e.g. Sepkoski 1984, 1997). Thus, it

remains to be shown what exactly drove this radiation. There is approximate stasis in the geographical and general environmental trend lines (Text-figs 9, 10) and fine-scale trends in reef data are erratic (Kiessling 2002, 2009; Text-fig. 10D). Nonetheless, there appears to have been a net expansion of shallow-water environments between the mid-Devonian crisis and the end of the Permian (Text-fig. 10B–C).

The mid-Jurassic radiation. Traditional compilations uniformly suggested that while there was a strong increase in taxonomic richness through the Jurassic, it was merely a phase of a steady, somewhat linear trend across the entire Mesozoic that was barely interrupted by the Triassic–Jurassic extinction (e.g. Valentine 1969; Sepkoski 1997). In proportional terms, however, the new diversity curve's mid-Jurassic radiation is quite steep, and unlike other recoveries it is preceded by a lag of several temporal bins. For example, rebounds from the Permo–Triassic and Cretaceous–Palaeogene events were largely complete in the space of one or two bins (Text-fig. 7). Again speaking in proportional terms, the Jurassic rise is greater than any potential Cretaceous rise, and there may well have been none (Text-fig. 7). Thus, the traditional view of a steady Mesozoic radiation is an illusion created by undersampling of the Triassic, oversampling of the Late Cretaceous and a legitimate but in fact underestimated Jurassic increase (Alroy *et al.* 2008) that might have been followed by a proportionately smaller Cretaceous rise seen mostly at high northern latitudes (Text-fig. 8B).

Published turnover rate data are ambivalent, suggesting that if anything there was an increase in extinction and decrease in origination through the Jurassic (Kiessling and Aberhan 2007). These data are slightly suspect now that better turnover rate equations (Alroy 2008) and subsampling protocols (Text-figs 3–6) are available. However, they are at least consistent with the current results in suggesting that net diversification rates peaked in the mid-Jurassic (Kiessling and Aberhan 2007). Importantly, this recent study suggested that organisms common in reefs diversified much more quickly than others, and additional comparisons such as carbonate vs. siliciclastic, onshore vs. offshore, and tropical vs. extratropical showed no large differences (Kiessling and Aberhan 2007). It also has been observed qualitatively that reefs were largely absent throughout the Early Jurassic but expanded rapidly afterwards (Stanley 1988; Kiessling 2002, 2009).

The Cenozoic radiation. There is deep division in the literature over the claim that diversity increased exponentially throughout the Late Cretaceous and Cenozoic, which if true would enshrine the Recent as the pinnacle of evolution. Diversity curves at the family, order or class level show no such increase (Valentine 1969; Raup 1972, 1975; Sepkoski 1978, 1984), although some of them depict a weak upwards trend following a recovery from the Cretaceous–Palaeogene mass extinction (e.g. Sepkoski 1984). However, early literature speculated that genus- or species-level curves would depict an extremely steep trend, either for biological reasons (Valentine 1970) or as a sampling artefact (Raup 1972).

Unfortunately, Sepkoski's genus-level Phanerozoic data set (Sepkoski 1997), the only one available until recently, can be made to show anything. For example, one can generate the appearance of a very strong exponential increase by exclusively counting genera that range across individual stage boundaries (Bambach 1999). Conversely, by counting genera that range across or into each stage but only using fossil occurrences to define last appearances and excluding genera only found in a single stage, one can create an almost linear Mesozoic trend followed by a near-plateau (Sepkoski 1997).

This dilemma has been resolved by clear evidence that at least three strong and unrelated biases favour the Cenozoic: a simple sample size effect (Raup 1976; Sheehan 1977; Alroy *et al.* 2001, 2008); preservational factors such as lithification and preservation of aragonite that decrease apparent local-scale richness in the deeper record (Cherns and Wright 2000; Alroy *et al.* 2008; Hendy 2009; Sessa *et al.* 2009); and the 'Pull of the Recent' (Raup 1979; Peters and Foote 2001), a reverse edge effect that results from relatively complete sampling in the Recent combined with the common practice in older literature of counting each extant genus as present in every interval between its first fossil appearance and the Recent. The confluence of these factors explains why Sepkoski *et al.* (1981), in an analysis that stifled almost all discussion of bias for two decades, found that both local and global diversity were much higher in the Cenozoic than Palaeozoic.

The current analysis suggests no trend at all within the Cenozoic, regardless of data set partitioning (Text-fig. 7; Alroy 2010). This fact is even more remarkable because changes in two geographical variables indicate improved coverage: a shift in sampling towards the equator right after the Cretaceous (Text-fig. 9A) and an increase in the median distance between sampled cells that is particularly rapid during the Palaeogene but continues into the Neogene (Text-fig. 9C). Spanning trees in the Neogene are also exceptionally large (Text-fig. 9D), possibly reflecting improved tropical southern hemisphere sampling that is demonstrated by the median distance data. Finally, coverage of reefs is also exceptional in the Neogene.

None of these trends have a clear connection to palaeogeographic movements or other obvious physical factors. For example, the jump in sampled area is unlikely to be real because sea level fell through the Cenozoic, with much of the decline occurring during the Oligocene (Kominz *et al.* 1998). Instead, there seems to be a combination of

geographical sampling biases that might be responsible in part for the Cenozoic 'radiation' seen in some older analyses. SQS seems to have resolved such problems within this part of the geological time scale, but more should be done to improve geographical coverage in such intervals as the Palaeogene.

Future implications

Two major conclusions can be drawn from this paper's empirical results. First, diversity limits exist and are related to environmental variables such as onshore–offshore depth and to some extent species–area effects, as one might expect. Second, the Big Three major mass extinctions and radiations varied spatially and involved a wide variety of causes and consequences instead of being governed by a single key factor.

Nonetheless, three of the five traditionally recognized Phanerozoic mass extinctions (mid-Devonian, Permo–Triassic and Jurassic–Cretaceous) correspond with a major disruption of either reef ecosystem productivity or taxonomic composition. Large rises in the mid-Jurassic and possibly mid-Permian also related to the fortunes of shallow-water ecosystems. That said, reef declines unaccompanied by mass extinctions occurred at the end of the Silurian and Jurassic (Kiessling 2009).

Even though these large biotic shifts resulted from a variety of proximal causes (combinations of turnover rates) and more distal causes (environmental perturbations such as volcanism or bolide impacts), there is good quantitative and qualitative evidence to support the idea that healthy reef ecosystems in particular promote diversification (Kiessling *et al.* 2010). Contemporary increases in atmospheric CO_2 and attendant ocean acidification appear to be the single largest threat faced by reefs, and the tipping point that will result in their near-complete loss may be surpassed within just a few decades (Hoegh-Guldberg *et al.* 2007). If reef extent is indeed the most important environmental factor governing long-term departures from diversity equilibria, then reef collapse is likely to delay recovery from the current mass extinction for the duration of an entire geological period or more.

Acknowledgements. I thank P. Donoghue, G. Harrington and the Palaeontological Association for providing the opportunity to contribute to this issue and its accompanying symposium; M. Benton, A. McGowan and A. Miller for critiquing the manuscript; M. Foote for testing and debating the subsampling methods; M. Kosnik for technological support; and C. Scotese for providing crucial palaeogeographic rotation data. My interest in this topic was spurred by A. Hendy, W. Kiessling, A. Miller, S. Peters and other members of the PaleoDB's working groups. I am grateful to M. Clapham, A. Hendy, W. Kiessling and many others for recent contributions to the PaleoDB that made the analyses possible. This research was independently funded, and this paper is Paleobiology Database official publication 112.

Editor. Philip Donoghue

REFERENCES

ABERHAN, M., WEIDEMEYER, S., KIESSLING, W., SCASSO, R. A. and MEDINA, F. A. 2007. Faunal evidence for reduced productivity and uncoordinated recovery in Southern Hemisphere Cretaceous-Paleogene boundary sections. *Geology*, **35**, 227–230.

ALLISON, P. A. and BRIGGS, D. E. G. 1993. Paleolatitudinal sampling bias, Phanerozoic species diversity, and the end-Permian extinction. *Geology*, **21**, 65–68.

ALROY, J. 1996. Constant extinction, constrained diversification, and uncoordinated stasis in North American mammals. *Palaeogeography, Palaeoclimatology, Palaeoecology*, **127**, 285–311.

—— 2000. New methods for quantifying macroevolutionary patterns and processes. *Paleobiology*, **26**, 707–733.

—— 2008. Dynamics of origination and extinction in the fossil record. *Proceedings of the National Academy of Sciences, USA*, **105**, 11536–11542.

—— 2010. The shifting balance of diversity among major marine animal groups. *Science*, **329**, 1191–1194.

—— MARSHALL, C. R., BAMBACH, R. K., BEZUSKO, K., *et al.* (21 others) 2001. Effects of sampling standardization on estimates of Phanerozoic marine diversification. *Proceedings of the National Academy of Sciences, USA*, **98**, 6261–6266.

—— ABERHAN, M., BOTTJER, D. J., FOOTE, M., *et al.* (31 others) 2008. Phanerozoic trends in the diversity of marine invertebrates. *Science*, **321**, 97–100.

BAMBACH, R. K. 1999. Energetics in the global marine fauna: a connection between terrestrial diversification and change in the marine biosphere. *Geobios*, **32**, 131–144.

BENTON, M. J. 1995. Diversification and extinction in the history of life. *Science*, **268**, 52–58.

——2009. The Red Queen and the Court Jester: species diversity and the role of biotic and abiotic factors through time. *Science*, **323**, 728–732.

—— and EMERSON, B. C. 2007. How did life become so diverse? The dynamics of diversification according to the fossil record and molecular phylogenetics. *Palaeontology*, **50**, 23–40.

BRIGGS, D. E. G., FORTEY, R. A. and WILLS, M. A. 1992. Morphological disparity in the Cambrian. *Science*, **256**, 1670–1673.

CHERNS, L. and WRIGHT, V. P. 2000. Missing molluscs as evidence of large-scale, early skeletal aragonite dissolution in a Silurian sea. *Geology*, **28**, 791–794.

CLAPHAM, M. E., SHEN, S. and BOTTJER, D. J. 2009. The double mass extinction revisited: reassessing the severity, selectivity, and causes of the end-Guadalupian biotic crisis (Late Permian). *Paleobiology*, **35**, 32–50.

COLWELL, R. K. and CODDINGTON, J. A. 1994. Estimating terrestrial biodiversity through extrapolation. *Philosophical*

Transactions of the Royal Society of London, Series B, **345**, 101–118.

CONNOLLY, S. R. and MILLER, A. I. 2002. Global Ordovician faunal transitions in the marine benthos: ultimate causes. *Paleobiology*, **28**, 26–40.

COPPER, P. 1988. Ecological succession in Phanerozoic reef ecosystems: is it real? *Palaios*, **3**, 136–151.

ERWIN, D. H. 2009. Climate as a driver of evolutionary change. *Current Biology*, **19**, R575–R583.

FINNEGAN, S. and DROSER, M. L. 2005. Relative and absolute abundance of trilobites and rhynchonelliform brachiopods across the Lower/Middle Ordovician boundary, eastern Basin and Range. *Paleobiology*, **31**, 480–502.

FISCHER, A. G. 1960. Latitudinal variations in organic diversity. *Evolution*, **14**, 64–81.

FOOTE, M. 1994. Temporal variation in extinction risk and temporal scaling of extinction metrics. *Paleobiology*, **20**, 424–444.

—— and SEPKOSKI, J. J. Jr 1999. Absolute measures of the completeness of the fossil record. *Nature*, **398**, 415–417.

GOOD, I. J. 1953. The population frequencies of species and the estimation of population parameters. *Biometrika*, **40**, 237–264.

GOTELLI, N. J. and COLWELL, R. K. 2001. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters*, **4**, 379–391.

HANSEN, T. A., FARRELL, B. R. and UPSHAW, B. III 1993. The first 2 million years after the Cretaceous-Tertiary boundary in east Texas: rate and paleoecology of the molluscan recovery. *Paleobiology*, **19**, 251–265.

HENDY, A. J. W. 2009. The influence of lithification on Cenozoic marine biodiversity trends. *Paleobiology*, **35**, 51–62.

HOEGH-GULDBERG, O., MUMBY, P. J., HOOTEN, A. J., STENECK, R. S., *et al.* (13 others) 2007. Coral reefs under rapid climate change and ocean acidification. *Science*, **318**, 1737–1742.

HOFFMAN, A. and FENSTER, E. J. 1986. Randomness and diversification in the Phanerozoic: a simulation. *Palaeontology*, **29**, 655–663.

HURLBERT, S. H. 1971. The nonconcept of species diversity: a critique and alternative parameters. *Ecology*, **52**, 577–586.

KAMMER, T. W., BAUMILLER, T. K. and AUSICH, W. I. 1998. Evolutionary significance of differential species longevity in Osagean-Meramecian (Mississippian) crinoid clades. *Paleobiology*, **24**, 155–176.

KIESSLING, W. 2002. Secular variations in the Phanerozoic reef ecosystem. *Phanerozoic Reef Patterns, SEPM Special Publication*, **72**, 625–690.

—— 2009. Geologic and biologic controls on the evolution of reefs. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 173–192.

—— and ABERHAN, M. 2007. Environmental determinants of marine benthic biodiversity dynamics through Triassic-Jurassic time. *Paleobiology*, **33**, 414–434.

————BRENNEIS, B. and WAGNER, P. J. 2007. Extinction trajectories of benthic organisms across the Triassic-Jurassic boundary. *Palaeogeography, Palaeoclimatology, Palaeoecology*, **244**, 201–222.

———— and VILLIER, L. 2008. Phanerozoic trends in skeletal mineralogy driven by mass extinctions. *Nature Geoscience*, **1**, 527–580.

——SIMPSON, C. and FOOTE, M. 2010. Reefs as cradles of evolution and sources of biodiversity in the Phanerozoic. *Science*, **327**, 196–198.

KNOLL, A. H., NIKLAS, K. J. and TIFFNEY, B. H. 1979. Phanerozoic land-plant diversity in North America. *Science*, **206**, 1400–1402.

—— BAMBACH, R. K., PAYNE, J. L., PRUSS, S. and FISCHER, W. W. 2007. Paleophysiology and end-Permian mass extinction. *Earth and Planetary Science Letters*, **256**, 295–313.

KOMINZ, M. A., MILLER, K. G. and BROWNING, J. V. 1998. Long-term and short-term global Cenozoic sea-level estimates. *Geology*, **26**, 311–314.

KRUG, A. Z. and PATZKOWSKY, M. E. 2004. Rapid recovery from the Late Ordovician mass extinction. *Proceedings of the National Academy of Sciences, USA*, **101**, 17605–17610.

MACARTHUR, R. H. and WILSON, E. O. 1967. *The theory of island biogeography*. Princeton University Press, Princeton, NJ, 203 pp.

MANDER, L. and TWITCHETT, R. J. 2008. Quality of the Triassic–Jurassic bivalve record in northwest Europe. *Palaeontology*, **51**, 1213–1223.

MCELWAIN, J. C., BEERLING, D. J. and WOODWARD, F. I. 2007. Fossil plants and global warming at the Triassic–Jurassic boundary. *Science*, **285**, 1386–1390.

—— WAGNER, P. J. and HESSELBO, S. P. 2009. Fossil plant relative abundances indicate sudden loss of Late Triassic biodiversity in East Greenland. *Science*, **324**, 1554–1556.

MCGOWAN, A. J. and SMITH, A. B. 2008. Are global Phanerozoic marine diversity curves truly global? A study of the relationship between regional rock records and global Phanerozoic marine diversity. *Paleobiology*, **34**, 80–103.

MCKINNEY, M. L. and OYEN, C. W. 1989. Causation and nonrandomness in biological and geological time series: temperature as a proximal control of extinction and diversity. *Palaios*, **4**, 3–15.

MILLER, A. I. 1997. Comparative diversification dynamics among palaeocontinents during the Ordovician Radiation. *Geobios*, **30**(suppl. 1), 397–406.

—— 1998. Biotic transitions in global marine diversity. *Science*, **281**, 1157–1160.

—— and FOOTE, M. 1996. Calibrating the Ordovician radiation of marine life: implications for Phanerozoic diversity trends. *Paleobiology*, **22**, 304–309.

NEWELL, N. D. 1952. Periodicity in invertebrate evolution. *Journal of Paleontology*, **26**, 371–385.

——1959. Adequacy of the fossil record. *Journal of Paleontology*, **33**, 488–499.

OLSZEWSKI, T. D. and PATZKOWSKY, M. E. 2001. Evaluating taxonomic turnover: Pennsylvanian–Permian brachiopods and bivalves of the North American Midcontinent. *Paleobiology*, **27**, 646–668.

PAUL, C. R. C. 1982. The adequacy of the fossil record. 75–117. *In* JOYSEY, K. A. and FRIDAY, A. E. (eds). *Problems of phylogenetic reconstruction*. Academic Press, London, 442 pp.

PETERS, S. E. 2004. Evenness of Cambrian–Ordovician benthic marine communities in North America. *Paleobiology*, **30**, 325–346.

—— 2005. Geologic constraints on the macroevolutionary history of marine animals. *Proceedings of the National Academy of Sciences, USA*, **102**, 12326–12331.

—— and FOOTE, M. 2001. Biodiversity in the Phanerozoic: a reinterpretation. *Paleobiology*, **27**, 583–601.

RAUP, D. M. 1972. Taxonomic diversity during the Phanerozoic. *Science*, **177**, 1065–1071.

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

—— 1976. Species diversity in the Phanerozoic: an interpretation. *Paleobiology*, **2**, 289–297.

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

—— 1985. Mathematical models of cladogenesis. *Paleobiology*, **11**, 42–52.

—— and SEPKOSKI, J. J. JR 1982. Mass extinctions in the marine fossil record. *Science*, **215**, 1501–1503.

—— GOULD, S. J., SCHOPF, T. M. and SIMBERLOFF, D. S. 1973. Stochastic models of phylogeny and the evolution of diversity. *Journal of Geology*, **81**, 525–542.

ROY, K., JABLONSKI, D. and VALENTINE, J. W. 1996. Higher taxa in biodiversity studies: patterns from eastern Pacific marine molluscs. *Philosophical Transactions of the Royal Society of London, Series B*, **351**, 1605–1613.

———— and ROSENBERG, G. 1998. Marine latitudinal gradients: tests of causal hypotheses. *Proceedings of the National Academy of Sciences, USA*, **95**, 3699–3702.

—— —— —— 2000. Dissecting latitudinal diversity gradients: functional groups and clades of marine bivalves. *Proceedings of the Royal Society of London, Series B*, **267**, 293–299.

SCHOPF, T. M. 1974. Permo-Triassic extinctions: relation to sea-floor spreading. *Journal of Geology*, **82**, 129–143.

SCHULTE, P., ALEGRET, L., ARENILLAS, I., ARZ, J. A., *et al.* (37 others) 2010. The Chicxulub asteroid impact and mass extinction at the Cretaceous-Paleogene boundary. *Science*, **327**, 1214–1218.

SEPKOSKI, J. J. JR 1976. Species diversity in the Phanerozoic: species-area effects. *Paleobiology*, **2**, 298–303.

—— 1978. A kinetic model of Phanerozoic taxonomic diversity. I. Analysis of marine orders. *Paleobiology*, **4**, 223–251.

—— 1979. A kinetic model of Phanerozoic taxonomic diversity. II. Early Phanerozoic families and multiple equilibria. *Paleobiology*, **5**, 222–251.

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

—— 1988. Alpha, beta, or gamma: where does all the diversity go? *Paleobiology*, **14**, 221–234.

—— 1997. Biodiversity: past, present, and future. *Journal of Paleontology*, **71**, 533–539.

—— BAMBACH, R. K., RAUP, D. M. and VALENTINE, J. W. 1981. Phanerozoic marine diversity: a strong signal from the fossil record. *Nature*, **293**, 435–437.

SESSA, J. A., PATZKOWKSY, M. E. and BRALOWER, T. J. 2009. The impact of lithification on the diversity, size distribution, and recovery dynamics of marine invertebrate assemblages. *Geology*, **37**, 115–118.

SHEEHAN, P. M. 1977. Species diversity in the Phanerozoic: a reflection of labor by systematists? *Paleobiology*, **3**, 325–328.

SIMBERLOFF, D. S. 1974. Permo-Triassic extinctions: effects of area on biotic equilibrium. *Journal of Geology*, **81**, 267–274.

SIMPSON, G. G. 1960. The history of life. 117–180. *In* TAX, S. (ed.). *Evolution after Darwin. Volume I. The evolution of life*. University of Chicago Press, Chicago, 629 pp.

SMITH, A. B. 2001. Large-scale heterogeneity of the fossil record: implications for Phanerozoic biodiversity studies. *Philosophical Transactions of the Royal Society of London, Series B*, **356**, 351–367.

STANLEY, G. D. JR 1988. The history of early Mesozoic reef communities: a three-step process. *Palaios*, **3**, 170–183.

STEHLI, F. G., DOUGLAS, R. D. and NEWELL, N. D. 1969. Generation and maintenance of gradients in taxonomic diversity. *Science*, **164**, 947–949.

TIPPER, J. C. 1979. Rarefaction and rarefiction: the use and abuse of a method in paleoecology. *Paleobiology*, **5**, 423–434.

VALENTINE, J. W. 1969. Patterns and taxonomic and ecological structure of the shelf benthos during Phanerozoic time. *Palaeontology*, **12**, 684–709.

—— 1970. How many marine invertebrate fossil species? A new approximation. *Journal of Paleontology*, **44**, 410–415.

VAN VALEN, L. 1973. A new evolutionary law. *Evolutionary Theory*, **1**, 1–30.

WESTROP, S. R. and ADRAIN, J. M. 1998. Trilobite alpha diversity and the reorganization of Ordovician benthic marine communities. *Paleobiology*, **24**, 1–16.

WILF, P., JOHNSON, K. R. and HUBER, B. T. 2003. Correlated terrestrial and marine evidence for global climate changes before mass extinction at the Cretaceous-Paleogene boundary. *Proceedings of the National Academy of Sciences, USA*, **100**, 599–604.