# Assessing the quality of the fossil record: insights from vertebrates

MICHAEL J. BENTON[1]*, ALEXANDER M. DUNHILL[1], GRAEME T. LLOYD[2] & FELIX G. MARX[1,3]

[1]*School of Earth Sciences, University of Bristol, Wills Memorial Building, Queen's Road, Bristol, BS8 1RJ, UK*

[2]*Department of Palaeontology, The Natural History Museum, Cromwell Road, London, SW7 5BD, UK*

[3]*Department of Geology, University of Otago, PO Box 56, Dunedin 9054, New Zealand*

**Corresponding author (e-mail: mike.benton@bristol.ac.uk)*

**Abstract:** Assessing the quality of the fossil record is notoriously hard, and many recent attempts have used sampling proxies that can be questioned. For example, counts of geological formations and estimated outcrop areas might not be defensible as reliable sampling proxies: geological formations are units of enormously variable dimensions that depend on rock heterogeneity and fossil content (and so are not independent of the fossil record), and outcrop areas are not always proportional to rock exposure, probably a closer indicator of rock availability. It is shown that in many cases formation counts will always correlate with fossil counts, whatever the degree of sampling. It is not clear, in any case, that these proxies provide a good estimate of what is missing in the gap between the known fossil record and reality; rather they largely explore the gap between known and potential fossil records. Further, using simple, single numerical metrics to correct global-scale raw data, or to model sampling-driven patterns may be premature. There are perhaps four approaches to exploring the incompleteness of the fossil record, (1) regional-scale studies of geological completeness; (2) regional- or clade-scale studies of sampling completeness using comprehensive measures of sampling, such as numbers of localities or specimens or fossil quality; (3) phylogenetic and gap-counting methods; and (4) model-based approaches that compare sampling as one of several explanatory variables with measures of environmental change, singly and in combination. We suggest that palaeontologists, like other scientists, should accept that their data are patchy and incomplete, and use appropriate methods to deal with this issue in each analysis. All that matters is whether the data are *adequate* for a designated study or not. A single answer to the question of whether the fossil record is driven by macroevolution or megabias is unlikely ever to emerge because of temporal, geographical, and taxonomic variance in the data.

The fossil record is far from perfect, and palaeontologists must be concerned about inadequacy and bias (Raup 1972; Benton 1998; Smith 2001, 2007*a*). Fundamental issues concerning the quality and completeness of the fossil record were enunciated clearly by Charles Darwin (1859, pp. 287–288), who wrote:

> That our palæontological collections are very imperfect, is admitted by every one. The remark of that admirable palæontologist, the late Edward Forbes, should not be forgotten, namely, that numbers of our fossil species are known and named from single and often broken specimens, or from a few specimens collected on some one spot. Only a small portion of the surface of the earth has been geologically explored, and no part with sufficient care, as the important discoveries made every year in Europe prove. No organism wholly soft can be preserved. Shells and bones will decay and disappear when left on the bottom of the sea, where sediment is not accumulating… With respect to the terrestrial productions which lived during the Secondary and Palæozoic periods, it is superfluous to state that our evidence from fossil remains is fragmentary in an extreme degree.

Raup (1972) clarified the situation when he compared the 'empirical' model of Valentine (1969), a literal reading of the fossil record, with his 'bias simulation model' that explained the bulk of the apparent low diversity levels of marine invertebrates in the Palaeozoic as a sampling error. Two opposite viewpoints have been argued, either that the fossil record is good enough (e.g. Sepkoski *et al.* 1981; Benton 1995; Benton *et al.* 2000; Stanley 2007) or not good enough (e.g. Raup 1972; Alroy *et al.* 2001, 2008; Peters & Foote 2002; Alroy 2010) to show the main patterns of global diversification through time. A resolution between these opposite viewpoints does not appear close (Benton 2009; Erwin 2009; Marshall 2010).

*From*: McGowan, A. J. & Smith, A. B. (eds) *Comparing the Geological and Fossil Records: Implications for Biodiversity Studies*. Geological Society, London, Special Publications, **358**, 63–94.
DOI: 10.1144/SP358.6 0305-8719/11/$15.00 

Key objective evidence for bias in the fossil record could be the extraordinary and ubiquitous correlation of sampling proxies and diversity curves: why is there such close tracking of measures of rock volume by palaeodiversity? There are three possible explanations: (1) rock volume/sampling drives the diversity signal (Peters & Foote 2001, 2002; Smith 2001, 2007*a*; Butler *et al.* 2011); (2) both signals reflect a third, or 'common', cause such as sea-level fluctuation (Peters 2005; Peters & Heim 2010); or (3) both signals are entirely or partially redundant (= identical) with each other. In reality, the close correlation probably reflects a combination of all three factors in different proportions in any test case, and so it is probably fruitless to prolong the debate about which of the three models is correct, and which incorrect.

Much of the literature on the quality of the rock and fossil records has focused on marine settings. This reflects the interests of palaeontologists who engage with these questions, and the fact that many marine rock records are more complete than most terrestrial (continental) rock records. However, the terrestrial fossil record is worth considering for several reasons: terrestrial life today is much more diverse than marine life, perhaps representing 85% of modern biodiversity (May 1990; Vermeij & Grosberg 2010), terrestrial life includes many major taxa that are sensitive to atmospheric, temperature, and topographic change and so are key indicator species in studies of global change, and for many terrestrial groups (e.g. angiosperms, insects, vertebrates) there are mature morphological and molecular phylogenies that enable cross-comparison between stratigraphic and cladistic data.

In this paper, we explore the use of sampling proxies, and suggest that some commonly used measures, notably formation counts and outcrop areas, may not be useful or accurate measures of sampling. Indeed, we suggest that there is probably no single numerical metric that captures all aspects of sampling (= rock volume, accessibility, effort), and recent attempts to correct the raw data, or to model sampling-driven patterns, may be premature. We then look at some case studies of patchy fossil records in taxa with good phylogenetic data, and suggest that in some cases at least the rock volume and fossil occurrence measures are identical, and so correlate almost perfectly. Finally, we suggest that such global-scale confrontations of sampling proxies and fossil data are not adequate at present, and recommend instead study-scale approaches to detect and correct sampling, involving direct evidence for missing data (e.g. Lazarus taxa; ghost ranges), direct evidence for sampling (e.g. number of localities or samples per time bin; fossil specimen completeness), and an integrated, model-based approach to incorporating sampling and explanatory models into explaining particular diversity curves.

## The fossil record, reality and sampling

### *The fossil record, collector curves and assessing reality*

The *known fossil record*, meaning our present understanding as represented by the literature and museum collections, is a subsample of the *potential fossil record*, all the fossils in the rocks, including undocumented materials (Fig. 1). The potential fossil record is itself a subsample of *all life that ever existed*, or reality, and this includes many soft-bodied and microscopic organisms that have never been fossilized and so can never be known.

The difference between the potential fossil record and reality may be very large (Paul 1988; Forey *et al.* 2004). An estimate of this difference has been made based on the proportion of fossilizable to non-fossilizable modern animals: of 1.2 million living species named at the time, Nichol (1977) estimated that fewer than 0.1 million (8%)

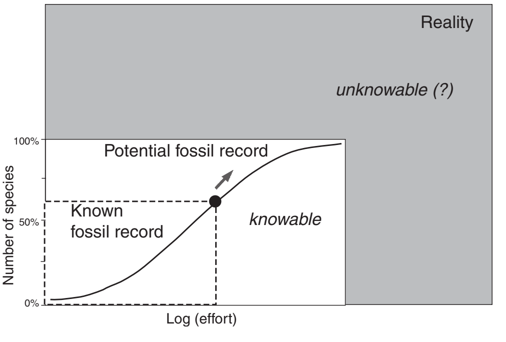


**Fig. 1.** The known and potential fossil record, and reality. Many sampling methods assess our position on the collector curve, which tracks the accumulation of knowledge, assessed against a measure of effort (e.g. number of specimens, area sampled, number of person-days work), from no knowledge to complete knowledge of the potential fossil record. The moving point shows where current knowledge stands on the collector curve trajectory. Other methods are required to assess the difference between the known fossil record and reality; the proportion of known (or knowable) fossil record (white) to the unknowable portion of reality (grey) is entirely hypothetical, and could range from 100% (where everything is known, or could be found) to infinitesimal (where nothing, or almost nothing, of a particular record at a particular level of focus, taxonomically and stratigraphically, is known); in fact the potential fossil record could represent about 10% of reality (Paul 1988; Forey *et al.* 2004). The collector curve is traced from its original presentation in Preston (1948).

were skeletized, and so potentially preservable as fossils. This kind of figure has also emerged from comparisons of organisms preserved in fossil Lagerstätten, such as the Burgess Shale, of which 10–15% of species are skeletized, which led Paul (1988) to suggest that, in normal conditions of preservation, perhaps 10% of Phanerozoic animal species might be preservable, leaving 90% unknowable, except for hints here and there from soft-bodied organisms glimpsed in sites of exceptional preservation. If the proportions of soft-bodied to hard-bodied organisms have remained constant through geological time, such an estimate may be helpful; if, however, the proportion has varied substantially, then even the best-sampled fossil record may say little about the true global pattern of palaeodiversity through time.

It is assumed that the documented fossil record is improving, as more fossils are found and as more researchers investigate and revise earlier work; doubtless some day all fossils in the rocks could be known scientifically (Maxwell & Benton 1990; Smith 2001; Forey *et al.* 2004). Certainly, the rate of naming of new species in some fossil groups is approaching saturation, as determined from *collector curves* (Preston 1948), where the trajectory has apparently reached the asymptote region (Fig. 1; e.g. Maxwell & Benton 1990; Benton & Storrs 1994; Benton 1998, 2008*a*; Paul 1998; Smith 2007*b*; Tarver *et al.* 2007, 2011; Bernard *et al.* 2010). Further, comparisons of knowledge through research time show that new finds often do not much affect overall macroevolutionary patterns in broad-scale studies (Maxwell & Benton 1990; Sepkoski 1993), although more focused analyses of patchily occurring groups such as dinosaurs may show changing phylogenetic trees as new fossils are found (Tarver *et al.* 2011). In addition, predicted gaps (e.g. ghost ranges, Lazarus gaps) are often filled by new fossils, and so congruence of cladograms with the fossil record improves (Benton & Storrs 1994) or remains static (Benton 2001) at higher taxonomic levels, although new fossil finds in finer-scale studies can still fill gaps or create new gaps (Weishampel 1996). Finally, comparisons (Foote & Raup 1996; Foote & Sepkoski 1999) of the probability of sampling fossil genera in one, two or three stratigraphic units (the FreqRat method) with proportions of living families with a fossil record show that, as expected, some taxa are apparently well sampled (e.g. ammonoids, conodonts, brachiopods, bryozoans, echinoids, ostracods, bivalves) while others are not (e.g. polychaetes, malacostracans). For well-sampled clades, such as trilobites, bivalves, and mammals, Foote & Raup (1996) estimated that 60% of actual species palaeodiversity had been recorded.

Any metric that measures sampling then is a measure of our trajectory to complete knowledge of the potential fossil record, not of reality, and this is well understood (e.g. Valentine 1969; Paul 1988; Maxwell & Benton 1990; Peters & Foote 2001; Smith 2001, 2007*a*; Forey *et al.* 2004). However, it might be easy to forget this, and to assume that comparisons of recorded fossil record data with sampling proxies provide the true pattern of palaeodiversity. This could only be the case if the proportions of soft-bodied to hard-bodied taxa within a clade have remained constant through time, or if the analyst is referring specifically to a single well-skeletonized group, such as dinosaurs or brachiopods. Even for statements about single clades, there are unassessable variables that might make the potential fossil record depart substantially, or unpredictably, from reality, especially if certain key habitats are never, or rarely, sampled. Patchy sampling by geographical region or by stratigraphic unit can be detected by comparative analyses (e.g. Smith 2007*b*), but the true proportions of different habitats occupied by members of a clade may never be known.

Sampling proxies that measure human effort, such as 'palaeontological interest units' (e.g. Raup 1977; Sheehan 1977) only refer to the unknown part of the potential fossil record. Further, as we argue below, counts of 'number of formations' are so intimately linked with historical aspects of species discovery that they too address only the unknown part of the potential fossil record, and could hardly be used to predict true palaeodiversities. Indeed, this was suggested by the observation (Benton 2008*a*) that the best way to guarantee to find new genera and species of dinosaurs is to find new basins/formations. Historical studies of the pattern of naming of new species of dinosaurs show that the best determinant of new taxa is the discovery of new formations, not intensified collecting in known sedimentary basins. Whereas the discovery curve for dinosaurs showed evidence of an asymptote ten years ago (Benton 1998), the subsequent, and continuing, exploration of new formations in China and South America has generated a rapid rise in naming of new species in the past decade (Benton 2008*a*). Nevertheless, the collector curve can never extend beyond the ceiling of the potential fossil record to explore all the dinosaurian species that once lived but were never fossilized (Fig. 1).

A number of attempts have been made to estimate the total species richness of dinosaurs, using a model for estimating future discoveries and generic longevity (Dodson 1990), using assumptions about species-area relationships (Russell 1995), and using relative abundance plots of rare to common species within local faunas (Wang & Dodson 2006). All these methods focus on the potential fossil record, and do not necessarily identify taxa that have never been preserved.

Is the 'unknowable' part of palaeodiversity (Fig. 1, grey) forever unknowable? It might be possible to seek approaches or metrics that at least point to some of the palaeodiversity of soft-bodied organisms or of unpreserved habitats by a variety of means. Some approaches might be (1) *phylogenetic*, using say evidence from a complete molecular phylogeny of living forms to identify likely missing fossils, (2) *stratigraphic*, using information about the known stratigraphic distributions of fossils to predict those that may have been there but can never be sampled, (3) *geographical*, attempting to identify gaps in fossil distributions, or (4) *ecological*, perhaps looking at trophic webs or pyramids to try to pinpoint missing taxa. Some efforts in this area include comparisons of cladograms and stratigraphic order of fossils (e.g. Norell & Novacek 1992; Benton & Storrs 1994; Huelsenbeck 1994; Wills 1999, 2007; Benton *et al.* 2000), comparison of fossil and molecular estimates of clade origins to identify durations and distributions of gaps (e.g. Smith 2007*b*), and calculation of confidence intervals on stratigraphic ranges of fossils using known collecting patterns (e.g. Marshall 1990, 1997). None of these approaches, however, gives an estimate of the difference between the potential fossil record and reality, merely a proportional measure on comparing two time bins or geographical regions, or a measure of whether the difference is likely to be large, or at least likely to bias the data.

Measures of rock volume might provide approaches to bridging the gap between the potential fossil record and reality (Fig. 1), but only if used in models to estimate original areas of habitats and missing taxa predicted to have occupied such habitats, but that have not been, or cannot be, sampled. For example, conservation biologists identify the niche of a modern species, and then plot potential geographical distributions according to the wider distribution of the precise habitats suitable for that species (e.g. Guisan & Zimmermann 2000; Guisan & Thuiller 2005). The assumption is that the species could occupy all of its potential habitat if human and other historical pressures did not prevent it. Such techniques might be used by palaeontologists to identify 'missing species' by combining rock volume data with palaeogeographical maps, and such hypothetical missing species could fall in the potentially knowable or unknowable zones (Fig. 1).

### Sampling

*Sampling* is a set of statistical procedures to explore how subsets of individual observations within a population of individuals may yield wider knowledge about the population of concern, especially for the purpose of making predictions based on statistical inference. In the context of the fossil record, sampling can refer to two research themes, (1) the choice of subsamples of a greater whole as a practical means to determine aspects of the wider sample, and (2) the degree to which the known fossil record is itself a suitable subsample of the greater whole. That 'greater whole' could be simply the potential fossil record (Fig. 1), but it is usually assumed to be reality.

Reasons that the known fossil record falls short of the potential fossil record or reality are a mix of geological and human factors (Raup 1972) that fall into three categories: (1) rock volume, the progressive geological bias against preservation and discovery of ever-older fossils (diagenesis, metamorphism, erosion, covering by younger rocks); (2) accessibility, the currently available rock area or volume; and (3) effort, human factors, such as geographical location, ease of access, and subject interest (by age, location, or fossil group).

### Adequacy and applications

An important caveat is that most of the previous discussions of fossil record completeness have concentrated on bias and sampling as they relate to a single use of the fossil record, namely to represent global diversity patterns. Such discussions might, or might not, have an impact on other uses of the fossil record, for example in studies of individual lineages or clades, in local- or regional-scale studies, phylogenetic analyses, ecological studies of communities, or anatomical and functional studies. So, an entirely biased fossil record that provides spurious data on the scaling of a global-scale radiation or mass extinction, might nonetheless provide numerous near-perfect fossil Lagerstätten that represent entire fossil communities and individual fossils that provide remarkable anatomical detail for functional, ecological, and evolutionary studies. As Donovan & Paul (1998) stated, 'the fossil record may be incomplete, but it is entirely adequate for many and most requirements of palaeontology'. Benton *et al.* (2000) made this point in a specific case, when they showed that age v. clade congruence is good at the scale of stratigraphic stages and taxonomic families, and shows no time dependence or bias back through the 550 Ma of the Phanerozoic. At finer temporal scales, Wills (2007) also showed constant levels of completeness through most of the Phanerozoic, but a substantial drop in the Cambrian and, surprisingly, in the Neogene. These may be partly 'edge' effects, but the Cambrian drop likely mixes sampling failure combined with obscure taxa and many soft-bodied forms.

Some observations based on analyses Darwin could not have predicted might be said to suggest that the fossil record, error-ridden and incomplete as it is, is adequate for many purposes, although none of these provides evidence that error in the fossil record is negligible: (1) the order of fossils in the rocks generally matches closely the order of nodes in morphological and molecular trees (Norell & Novacek 1992; Benton & Hitchin 1997); (2) at coarse scales of observation (families and stratigraphic stages), there is no evidence that this matching becomes worse deeper in time (Benton *et al.* 2000; Wills 2007); (3) macroevolutionary patterns, including posited mass extinctions and diversifications, are largely immune to changes in palaeontological knowledge, even over 100 years of research time (Maxwell & Benton 1990; Sepkoski 1993; Adrain & Westrop 2000); (4) congruence between stratigraphy and phylogeny has also been largely stable through the 20th century, despite an order-of-magnitude increase in the number of fossils (Benton 2001); (5) new fossil finds, even of reputedly poorly sampled groups such as primates and humans, do not always alter perceptions of evolutionary patterns (Tarver *et al.* 2011); and (6) new post-Cambrian Lagerstätten rarely add new families to existing knowledge, just new species and genera (Benton *et al.* 2008).

## Sampling proxies

### *Definition*

A *sampling proxy* is a metric that represents 'collecting effort' in some way, the *x*-axis of the species v. effort plot (Fig. 1). Such a sampling proxy should represent some or all of the geological and human factors that can introduce error into interpretations of data from the fossil record, typically time series of diversity. The sampling proxy ought to be a measure of biasing factors such as rock volume, accessibility, or human effort, and it should be *independent of the signal it seeks to correct*, namely the documented fossil record. This might seem evident, but it has rarely been demonstrated, or even discussed, especially in the case of the commonly used universal sampling proxy of formation numbers. Further, sampling proxies generally assess only the difference between the known and potential fossil records, and probably have little to say about the unknowable part of reality (see above).

Sampling proxies can be used in various ways, (1) to assess whether a fossil record is biased, (2) to assess the nature of the bias, and (3) to correct the bias and produce a true evolutionary signal. In the first two cases, the sampling proxy may be partial, in that it documents some aspect of rock volume, accessibility, or human effort. In the third case, however, if the sampling proxy is to be used as a correction factor or the basis of a sampling-free model, it should be *comprehensive*, and so represent the three key factors of rock volume, accessibility, and effort. In cases where supposedly error-free records have been generated and then used to make statements about evolution (e.g. Peters & Foote 2001, 2002; Smith 2001; Smith & McGowan 2007; Barrett *et al.* 2009; Wall *et al.* 2009; Benson *et al.* 2010; Butler *et al.* 2011), the sampling proxy, whether formation counts or outcrop areas, was not demonstrated to encapsulate all, or even most, of the error signal – this was an assumption. At best, these papers could be said to have provided examples of how to identify parts of the fossil record that cannot be explained by sampling, and so *might* be real. At worst, they show very little because the universal sampling proxies used may have been partially redundant with the fossil record they sought to correct (formation counts) or were uncertain measures of rock volume/accessibility (outcrop areas). This of course is a criticism of the assumption that the sampling proxies were universal and comprehensive, not a claim that the fossil record is complete and unbiased.

A sampling proxy may be compared for matching with the raw diversity signal in various ways: (1) visual inspection; (2) correlation or rank order correlation (e.g. Fröbisch 2008; Barrett *et al.* 2009; Butler *et al.* 2009; Benson *et al.* 2010); (3) rarefaction to equalize sample sizes in each time bin (e.g. Benton *et al.* 2004; Lloyd *et al.* 2008); or (4) modelling a sampling-predicted diversity pattern based on numbers of formations or relative map areas, and comparing this with observed diversity to explore correlations and residuals (e.g. Peters & Foote 2001, 2002; Smith 2001; Smith & McGowan 2007; Barrett *et al.* 2009; Benson *et al.* 2010; Butler *et al.* 2011; Mannion *et al.* 2011; Lloyd *et al.* 2011).

There are two commonly used sampling proxies, outcrop area and numbers of formations, and a third might be human effort. These three are discussed here, before assessing the merits of the first two from some recent studies.

### *Sampling proxies 1: outcrop (map) area*

It has been argued (e.g. Smith 2001, 2007*a*; Crampton *et al.* 2003; Smith & McGowan 2008; Wall *et al.* 2009) that map area (= area of outcrop) is a good proxy for sampling. Rocks that cover large areas of the landscape are likely to have been much more sampled than those that do not outcrop widely, so this metric incorporates aspects of rock volume, rock availability at the surface, and human factors.

Practitioners of the map area approach have used a variety of methods. Some (e.g. Wall *et al.* 2009) have extracted information from global-scale summary geological maps (e.g. Ronov 1978, 1994), although these suffer from enormous generalizations and simplifications both in terms of the ages of the rocks and their areas. Time divisions in such studies have necessarily been broad to accommodate the difficulties of welding together information from disparate and uncoordinated national geological surveys; the OneGeology programme (http://www.onegeology.org/) which aims to produce a single, comprehensive world geological map, may help to rectify this. Others have used maps and memoirs from single countries or groups of countries (e.g. NW Europe; Smith 2007*a*; McGowan & Smith 2008; Smith & McGowan 2007), consulting sheet memoirs to record those stratigraphic units as present that have yielded fossils belonging to particular zones. A problem with the latter approach has been that it was used for comparisons of like with unlike (local or regional map data v. global fossil record). These authors argued that their approach was valid as errors become negligible once the map sample is as large as 1300 – Smith & McGowan (2007) showed extremely high congruence between sampled rock areas through geological time obtained from France and Spain. This presumably indicates a considerable amount of fundamental geology shared between the two countries, relating to bedrock, topography, and soil cover, and does not necessarily say much about accessibility (exposure) or about world geology. Furthermore, we see no evidence that scaling up small-scale data that may contain errors and mismatches will necessarily gloss and over-ride those errors. Nonetheless, when comparing like with like (the fossil record and geological map areas for New Zealand), Crampton *et al.* (2003) found that outcrop area was a good proxy for sampling because it correlated closely with number of collections.

Conclusions from such map area v. palaeodiversity studies have been very different, with some (e.g. Peters 2005, 2008) arguing that the covariation of fossil diversity and outcrop areas indicates a common driver such as sea-level change, others (e.g. Wall *et al.* 2009) arguing for a strong bias on estimates of fossil diversity from outcrop area, and yet others (e.g. Marx 2009) finding only modest evidence for an influence of outcrop area on fossil diversity.

A key issue, perhaps not appreciated fully before, is that outcrop (= map area) is not always correlated with rock exposure (= area of rock exposed at the surface), but is heavily modified by overlying deposits and depends on factors such as topographic elevation and coastal intersection (Dunhill in press). As a substantial proportion of outcropping rock is not exposed at the surface, and fossil specimens can only be recovered easily from exposed localities, it can be argued that rock exposure area represents a better proxy for the amount of sedimentary rock available for study than outcrop (i.e. map) area. Rock outcrop area can only be regarded as a good sampling proxy if it proves to correlate well with current rock exposure area, and, importantly, to be a good representation of the total historical accessibility of geological units, through a combination of current exposure and past collecting opportunities in old quarries, mines, railway cuttings, and landslips. Previous difficulties in quantifying the effects of exposure failure on sampling (Peters & Heim 2010) have been overcome by using remote sensing, in the form of Google Earth™ imagery, and a Geographic Information System (GIS) to map and quantify rock exposure accurately on a local scale (Dunhill 2011).

Dunhill (2011) showed that rock outcrop and exposure area are not positively correlated across 50 sample localities in England and Wales (Fig. 2a, b), and a further investigation showed these results to be consistent with results obtained, using the same methodology, in New York State (Dunhill in press). However, after data manipulation, rock outcrop and exposure area did display positive correlations in data sets from California and Australia (Dunhill in press), suggesting that variables such as climate and population density may have an influence on the amount of rock exposed at the surface. Further tests showed that proportional rock exposure is dependent on a number of variables that are themselves independent of outcrop area, in particular proximity to the coast, elevation, and bedrock age (Dunhill 2011, in press). Coastal areas exhibit consistently higher proportional exposure than inland localities because of constant high erosion regimes. A common pattern in the data was greater proportional exposure of older bedrock compared to younger bedrock (Fig. 2d), which might be explained by the presence of a greater proportion of harder lithologies of Palaeozoic age and more softer, unlithified sediments of Cenozoic age (Hendy 2009). Areas of higher elevation, both in inland and coastal areas, consistently exhibit higher proportional exposure (Fig. 2c), a result of increased erosion at altitude and greater exposure in areas of high coastal cliffs (Dunhill in press). It is apparent that many of the variables influencing the amount of rock exposed at the surface are not independent of one another, with bedrock age and elevation correlating positively in all four sampled regions (Dunhill in press). This can probably be explained by the simple fact that older rocks have been around for longer and have thus had more time to become uplifted.

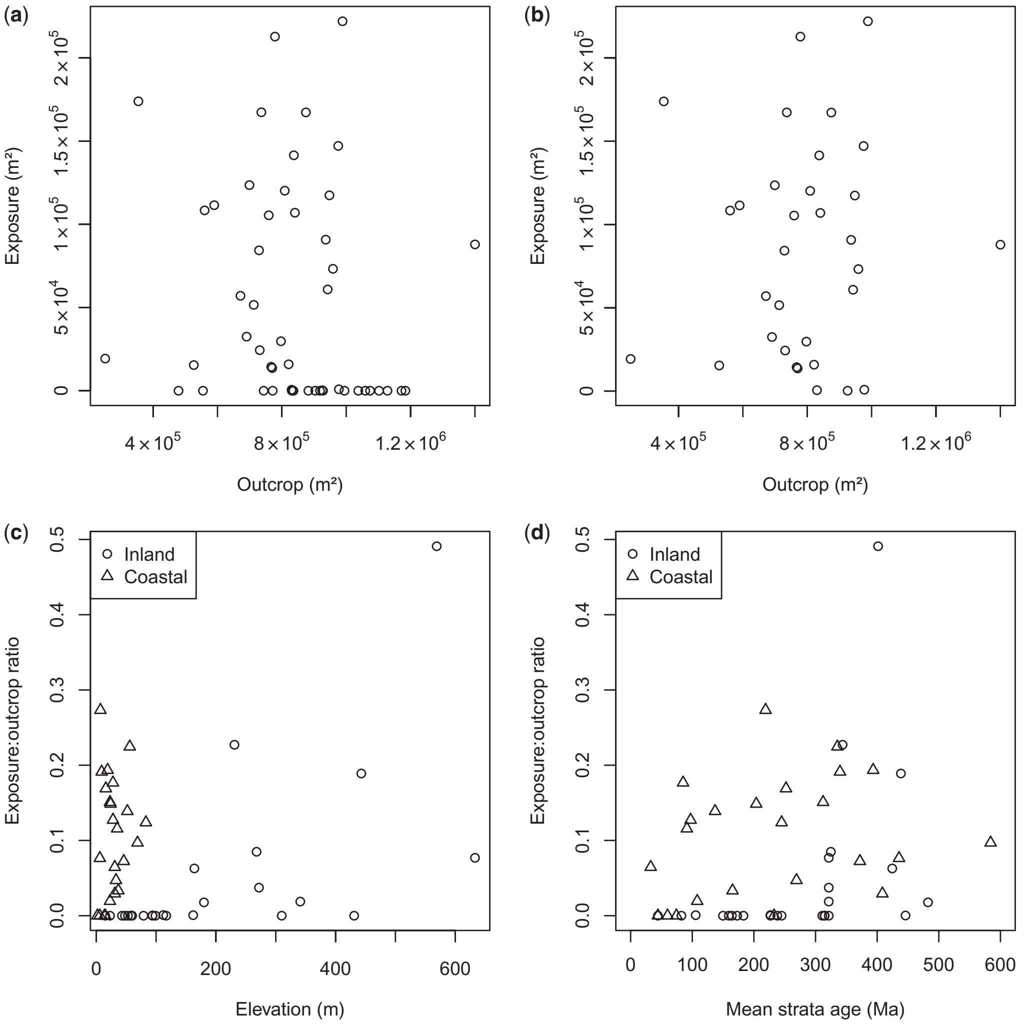


**Fig. 2.** Comparisons of outcrop (= map area) and exposure, and implications for sampling proxies. (**a**) Negative correlation between outcrop and exposure area for all of the sample localities in the England and Wales data set (Spearman: $r_s = -0.28$, $P = 0.05$). (**b**) Lack of correlation between outcrop and exposure area for the England and Wales data set after areas with zero rock exposure have been removed (Spearman: $r_s = 0.07$, $P = 0.68$). (**c**) Correlation between elevation and exposure:outcrop ratio for all the sample localities in England and Wales (Spearman: $r_s = -0.03$, $P = 0.85$), inland localities only (Spearman: $r_s = 0.72$, $P < 0.001$) and coastal localities only (Spearman: $r_s = 0.15$, $P = 0.47$). (**d**) Correlation between mean strata age and exposure:outcrop ratio for all the sample localities in England and Wales (Spearman: $r_s = 0.37$, $P = 0.009$), inland localities only (Spearman: $r_s = 0.7$, $P < 0.001$) and coastal localities only (Spearman: $r_s = 0.11$, $P = 0.34$) (modified from Dunhill 2011).

The fact that outcrop and exposure area do not consistently correlate across continents, and that rock exposure appears to be controlled by a number of variables that are independent of outcrop area, non-independent of each other, and inconsistent in their effects across different regions of the world, suggests the need for further investigation of the value of outcrop (i.e. map) area as a proxy for sampling. Further, if rock volume is linked with processes that regulate biodiversity, through for example sea-level change on continental shelves (see below, common cause hypothesis), then it might be advisable to exercise caution in interpreting correlations of rock volume sampling metrics and fossil diversity, and especially in using such metrics as the basis for generating a supposedly sampling-free diversity curve (e.g. Smith 2001, 2007*a*; Smith & McGowan 2007; Wall *et al.*

2009). If, as seems likely (see above), these rock volume and area measures (outcrop area, exposure area) assess the potential for future discoveries (i.e. potential minus known fossil record), rather than predict into the unknown part of reality (Fig. 1), then the claim that 'sampling-free' palaeodiversity curves can be generated from them is further put in doubt.

### *Sampling proxies 2: number of formations*

Counts of numbers of geological formations have been used widely as a sampling proxy (e.g. Peters & Foote 2001, 2002; Fröbisch 2008; Barrett *et al.* 2009; Butler *et al.* 2009; Benson *et al.* 2010; Mannion *et al.* 2011). These authors found close or very close covariation between numbers of formations and palaeodiversity, and this has generally been interpreted as evidence for a control by rock volume on apparent biodiversity. For example, Barrett *et al.* (2009), Butler *et al.* (2009), Benson *et al.* (2010), and Mannion *et al.* (2011) used their counts of numbers of formations as a measure of the reality of sampling, and so they modelled 'sampling-corrected' global diversity curves, and based revised interpretations of the history of life on the residuals.

Counts of numbers of formations may reflect a combination of some aspects of exposure area, total thickness, lithological heterogeneity, and intensity of study, and so they might seem to be ideal sampling proxies that encapsulate elements of rock volume, accessibility, and human effort (Peters & Foote 2001). However, formation counts have been widely criticized as sampling proxies because:

(1) Their definitions are arbitrary, being human inventions. The thickness, duration, area covered, and heterogeneity may depend on the customs of geologists, whether by nationality or by main research subject-area, as well as on the recency of revision and definition of formal formation names (Upchurch & Barrett 2005; Peters & Heim 2010).
(2) They do not generally correlate with rock exposure area measurements (Dunhill 2011, in press), which at least offer the simplest guide to accessibility, even though they do not offer a single universal sampling proxy.
(3) They do not necessarily correlate with collection effort (Crampton *et al.* 2003), but see Upchurch *et al.* (2011).
(4) They may largely reflect rock heterogeneity (Crampton *et al.* 2003; Smith 2007*a*), such that highly varied sediments lead to more and thinner formations, whereas formations are much thicker when the rock types are unchanging over long time spans, for example, continental red beds or Chalk.
(5) Importantly, they may also depend on fossil abundance and diversity (Wignall & Benton 1999): if fossils are highly abundant and diverse, formations may be subdivided more finely than if fossils are sparse or absent. This means that formation counts are not necessarily independent of the signal they seek to test or correct.
(6) Formations vary enormously in scale: formations vary over several orders of magnitude in thickness and geographical area (e.g. Williams 1901; Peters 2006; Peters & Heim 2010). At the smaller end of the scale are formations within well-studied and highly fossiliferous divisions of the Jurassic or Carboniferous of England that are only metres thick and cover only a few square kilometres. For example, the Beacon Limestone Formation (Lower Jurassic) of the southern UK, reaches a maximum thickness of 5 m (Cox *et al.* 1999), but is typically less than 1 m thick (Simms *et al.* 2004) with an outcrop coverage of around 20 $km^2$ (BGS digital bedrock geology DiGMapGB-50 of the UK), so representing a maximum volume of 0.073 $km^3$. At the other end of the scale are units such as the Late Jurassic Morrison Formation of the Midwestern United States that is up to 150 m thick and covers an area of 1.5 million square kilometres (Carpenter 1997), equivalent to a maximum volume of 225 000 $km^3$. These two examples show eight orders of magnitude difference in rock volume of named formations that might otherwise be treated as equivalent sampling units.

A further issue with the use of formation counts as a measure of sampling is that analysts usually count only *fossiliferous* formations, not all possible formations. Therefore, anything that means organisms are rare and so not fossilized will reduce the fossiliferous formation count, while this figure will rise when fossils are common and abundant. A particular example might be the time bin following a mass extinction, where palaeontologists might recover low numbers of fossils, and indeed those low sample sizes correspond to low numbers of fossiliferous formations. But this does not necessarily indicate sampling bias (Wignall & Benton 1999; *contra* Smith 2001, 2007*a*, *b*). Perhaps organisms were truly rare following the mass extinction; therefore fossils are not often found, and so rare fossils mean low numbers of fossiliferous formations. But this is a reversal of the normal sampling assumption. Here fossils determine formation numbers, not the other way round. Smith (2001, p. 355) argued that such a phenomenon could not occur: 'Taxon absences arise because of lack of habitat continuity and/or changing preservational opportunity in the fossil record, not as a result of fluctuating abundance within a uniformly sampled habitat.' Peters & Heim (2010) note, however, that absence of fossils in a stratigraphic unit can reflect sampling failure, human error in reporting, or true absence. Preservation failure is doubtless often the case, but there is no fundamental reason why abundance and diversity of organisms in their communities cannot be reflected in the fossil record.

There are three ways to document the fossiliferous formation count (FFC): a *strict FFC*, consisting of only those formations that have produced named fossils included in the diversity measure; a *wider FFC* consisting of all formations that have ever produced any kind of fossil of the group in question, whether named, unidentified elements, or trace fossils; and a *comprehensive FFC* that includes all formations of the correct facies that have produced, or might produce fossils of the group in question. This allows for future finds, but also documents formations that could contain fossils of the group in question, but do not, and so includes failure of sampling (Wignall & Benton 1999). The need for a wider definition of FFCs was noted by Upchurch & Barrett (2005), when they suggested that a count of dinosaur-bearing formations (DBFs), as used in Barrett *et al.* (2009), and other papers, ought to subsequently sample as widely as possible and include all possible opportunities to observe.

The strict FFC fails as a sampling proxy because it can only assess the difference between the potential and known fossil records (Fig. 1), and nearly always shows strong correlation with palaeodiversity (see below). A wider or comprehensive FFC allows some view into the unknown portion of reality (Fig. 1): for example, in assessing sampling of bird fossils, a comprehensive FFC of all vertebrate-bearing formations will show many that could well have sampled bird fossils, but that entirely lack them, and so can give a proportion of what is missing. However, what is not assessed are the times and places where birds lived in the past, but where rock was not accumulating or has not survived (and so there are no geological formations).

### *Sampling proxies 3: human effort*

Human factors are doubtless hugely important in determining our knowledge of the fossil record, the difference between potential and known (Fig. 1). These human factors were explored especially following Raup's (1972) paper. For example, Sheehan (1977) defined the Paleontologic Interest Unit (PIU) as a measure of the effort devoted to acquiring knowledge, counted in numbers of people, years, or publications on a particular subject, such as fossils of a specified geological interval or taxonomic group. These showed considerable mismatches in terms of effort per million years with, for example, eight times as many palaeontologists (per million years) working on Cenozoic fossils as on Cambrian fossils (Sheehan 1977). Further, it is well understood that the fossils of certain continents have been more intensively collected and studied than those from other continents (e.g. Kiessling 2005; McGowan & Smith 2008) as is confirmed by comparisons of collector curves, where numbers of new taxa identified from Europe and North America have reached an asymptote for many taxa, such as dinosaurs (Benton 1998, 2008*a*), conodonts (Wickström & Donoghue 2005), echinoids (Smith 2007*b*), trilobites (Tarver *et al.* 2007), and basal tetrapods (Bernard *et al.* 2010), whereas new taxa from other continents are still on the rising phase of the curve.

The idea of using a measure of human effort as a sampling proxy is reflected in the fossil collections data of the Paleobiology Database (PaleoDB; Alroy *et al.* 2001, 2008; Butler *et al.* 2011), where each collection is an assemblage of fossils collected from a particular formation often by a single palaeontologist or team. Formations with many reported collections then have been very actively sampled. Other measures of effort, as indicated by Sheehan (1977), such as numbers of palaeontologists or numbers of published papers, have proved hard to turn into an acceptable correction algorithm. Further, it would be hard to distinguish whether human factors drive or follow the data: perhaps concentrations of workers reflect abundant and diverse fossils (Raup 1977; Purnell & Donoghue 2005). Until a way can be found to disentangle the direction of causation between human effort and fossil diversity, PIUs may fail as a correction for error in the fossil record because of the probable intimate interconnection of both signals. Further, of course, these measures only assess our current position on the collector curve trajectory to complete knowledge of the potential fossil record (Fig. 1). Other measures of 'effort' though, such as numbers of collections, numbers of sampled localities, numbers of individual fossils (for rarer groups), or specimen quality (e.g. Benton *et al.* 2004; Smith 2007*b*) may provide useful insights into sampling of reality (see below).

## Bias, common cause or redundancy

### *Models*

Many measures of palaeodiversity and of rock volume show long-term covariation through geological time. Rock volume per million years increases towards the present day (Raup 1976; Smith 2001, 2007*a*), as does geological completeness of palaeontological sampling (Peters & Heim 2010). Further, there are substantial rises and falls in rock volume per million years within any selected time span, resulting from sea-level change (Smith 2001, 2007*a*; Peters 2005). Generally, the apparent diversity of marine animals follows both of these signals (Smith 2001), rising from the Cambrian onwards and rising and falling in line with sea-level rises and falls, except in the past 100 Ma, when sea-levels have generally fallen while marine diversity has risen. Terrestrial diversity likewise correlates with rock volume (Kalmar & Currie 2010), although causes of variations in rock volume for continental habitats are complex.

The close covariation of rock volume and palaeodiversity (Smith 2001, 2007*a*; Smith & McGowan 2007, 2008), and counts of geological formations (Peters & Foote 2001, 2002; Fröbisch 2008; Barrett *et al.* 2009; Butler *et al.* 2009; Wall *et al.* 2009; Benson *et al.* 2010) or sequence stratigraphic rock packets (Peters 2005, 2008; Peters & Heim 2010) can be interpreted in various ways, namely (1) rock bias; (2) common cause; and (3) redundancy.

The *rock bias hypothesis* is that the fossil record is wholly or partly a result of sampling bias, primarily determined by rock volume and accessibility (e.g. Raup 1972; Peters & Foote 2001, 2002; Smith 2001, 2007*a*; Smith & McGowan 2007; Fröbisch 2008; Barrett *et al.* 2009). In simple terms, the more rock, the more fossils (combining aspects of all three biasing factors – rock volume, accessibility, and human effort). Low volumes of rock (low outcrop area or low numbers of formations) means fossils cannot be found and so low fossil diversity is interpreted as sampling failure. The implication is that more intensive search of these formations, or ideally the discovery of new formations or areas of a poorly sampled time interval then ought to yield fossils at a faster rate than those from a well-sampled time bin.

The *common cause hypothesis* (the ‘biologically driven’ model of Smith 2001; Peters 2005, 2008) is that measures of rock volume and fossil diversity or abundance are correlated because they are driven by a third, common cause, such as sea-level change. It could be that biodiversity is driven by rock volume, so that marine life is much more diverse at times of rapid sedimentation (wide continental shelves; high habitat heterogeneity; high productivity, with much organic matter swept in from the land) than at times of low sedimentation (narrow continental shelves; low habitat heterogeneity; low productivity). The expanding and contracting marine shelf then drives expansions and contractions in marine biodiversity, which is dominated by the portion that occupies continental shelves, a kind of species-area effect (Smith 2001). Sea-level has also been posited as a driver of terrestrial diversity (e.g. Smith 2001; Mannion & Upchurch 2010), with high sea-levels corresponding either to low terrestrial diversity because of smaller land areas and fewer opportunities for preservation of habitats (Markwick 1998; Smith 2001) or to high terrestrial diversity because of more islands, leading to more endemism, as well as more chances for skeletons to fall into aquatic habitats and for coastal habitats to be swamped by the sea (Haubold 1990; Mannion *et al.* 2011). However, these suggestions are not supported by evidence from the Cretaceous terrestrial record (Fara 2002), and, although terrestrial diversity correlates with rock volume (Kalmar & Currie 2010), it is not clear how the volume of terrestrial clastic sediment could relate to sea-level (Butler *et al.* 2011).

The *redundancy hypothesis*, proposed here, is that in many cases the supposed sampling proxy and the fossil record are to a greater or lesser extent redundant with each other. This is especially true when the fossil record is patchy, perhaps for birds or pterosaurs – each fossil find adds a new fossiliferous formation to the roster. In such cases (e.g. Butler *et al.* 2009), the correlation between fossil count and formation count will be nearly perfect. The simplest ‘correction’, where number of fossil taxa is divided by number of formations will yield a flat line, as found by Lloyd *et al.* (2008) when they subsampled numbers of dinosaur species by dinosaur localities through 12 Mesozoic time bins. But such a confrontation of two redundant signals says nothing about sampling or the true unbiased pattern.

We now present re-analyses of two recently published studies, one on anomodonts and one on pterosaurs, to explore situations where bias (or ‘megabias’) was claimed, but where the supposed sampling proxy is really redundant with the palaeontological diversity signal.

## *Anomodonts: redundancy of data and sampling proxy*

As noted earlier, it is important to consider the kind of formation count employed, whether the strict, wider, or comprehensive fossiliferous formation count (FFC). We explore the effects of using these three variants in a recent study of anomodont diversity through the Permian and Triassic (Fröbisch 2008). Fröbisch (2008, 2009) divided the fossil record of anomodonts into a variety of time bins: land vertebrate faunachrons (LVFs), standard marine stages, and million-year divisions. The LVFs were his first determination of ages, and the others were derived from them, so the test here focuses on the LVFs. All 128 anomodont species were assigned to the 13 LVFs according to data from Fröbisch's (2008) paper, and the numbers of anomodont-bearing formations for each time bin were listed (Table 1). In addition, the numbers of named stratigraphic units with *any* tetrapod fossils were also listed, based on unpublished data from Benton *et al.* (2004), Sahney & Benton (2008), Sahney *et al.* (2010), and other sources. As Fröbisch (2008) found, the time series of anomodont species diversity correlates significantly with the time series of anomodont-bearing formations, but not with the time series of all tetrapod-bearing formations, with which anomodont diversity shows a negative, but not significant, correlation (Fig. 3; Table 2).

In order to test further whether the strict FFC could ever be an independent sampling proxy for 'diversity of X', a series of randomized trials was carried out to see whether it would be possible to break the strong correlation of time series of anomodont species diversity with number of anomodont-bearing formations. The numbers of recorded formations per time bin (Table 1) were used as a starting point. For each of the 84 anomodont-bearing formations reported in Fröbisch (2008), a total species diversity from 0–5 was generated using random numbers – the minimum of 0 allows for no finds in a formation, and 5 is the maximum reported, for the Gamkan 1 time interval, in Fröbisch (2008). These were summed for each time bin according to the reported numbers of formations, and forty simulations were performed, sufficient to assess the statistical significance of the results at the 95% level for a two-tailed test (95% probability = 1 in 20; × 2 for a two-tailed test). Remarkably, the simulated anomodont totals and the actual numbers of anomodont-bearing formations were significantly correlated in all but three cases (Table 2). In fact these randomized trials gave rank correlation values in a tight range around a mean and mode of 0.60, identical to the actual value obtained by Fröbisch (2008), and with significance values equal to or slightly better than in the real example (Table 2). This suggests not only that random data *can* produce results similar to those found with real data, but that this will occur most of the time. Hence, it is probably impossible ever to find a non-significant relationship between time series of sparsely occurring fossils and their strict FFC, and so the discovery of such a correlation in such cases says nothing about sampling, quality of the fossil record, or megabiases.

In order to avoid any spurious correlations arising from autocorrelation within the time series, the data were detrended by taking first differences of anomodont diversity and formation numbers, and these confirmed the significant correlation of changes in anomodont diversity and changes in strict FFC, but an absence of correlation of changes in anomodont diversity and changes in the comprehensive FFC (Table 2). The same is true also for generalized differencing (GD), a more comprehensive technique that incorporates detrending and differencing but modulates the differences by the strength of the correlation between successive data points (McKinney 1990). The GD results show a highly significant correlation between numbers of

**Table 1.** *Comparison of diversity of anomodont fossils through 13 time bins, spanning from the Middle Permian (bottom) to Late Permian*

| Stage | Lucas LVF | Karoo LVF | Myr | Anomodont species | Anomodont-bearing fms | All fmns |
|---|---|---|---|---|---|---|
| Nor(l) | Revueltian | | 2 | 2 | 2 | 16 |
| Crn(u) | Adamanian | | 3.5 | 2 | 2 | 17 |
| Crn(m) | Otischalkian | | 2.5 | 2 | 3 | 17 |
| Lad(u)-Crn(l) | Berdvankian | | 5 | 3 | 6 | 31 |
| Ans(u)-Lad(l) | Perovkan | *Cynognathus C* | 4 | 23 | 8 | 40 |
| Ole(u)-Ans(l) | Nonesian | *Cynognathus A, B* | 6.5 | 13 | 13 | 59 |
| Ind-Ole(l) | Lootsbergian | *Lystrosaurus* | 2.5 | 9 | 8 | 49 |
| Chx(m-u) | Platbergian | *Dicynodon* | 3 | 34 | 12 | 8 |
| Wuc(u)-Chx(l) | Steilkransian | *Cistecephalus* | 1.5 | 27 | 6 | 16 |
| Wuc(m) | Hoedemakeran | *Tropidostoma* | 2.5 | 16 | 8 | 15 |
| Cap(u)-Wuc(l) | Gamkan II | *Pristerognathus* | 2 | 7 | 4 | 11 |
| Cap(l-m) | Gamkan I | *Tapinocephalus* | 4.5 | 13 | 3 | 11 |
| Wor | Kapteinskraalian | *Eodicynodon* | 2 | 5 | 2 | 15 |

Standard stratigraphic stages, and two systems of Land Vertebrate Faunachrons (LVF) are given, as well as durations of these intervals, in Myr (millions of years). The Dicynodont-bearing formations (fmns) are those from Fröbisch (2008) that yielded anomodont fossils, whereas 'All formations' are all named stratigraphic units that have yielded any kind of fossil tetrapod remains. Stage-name abbreviations: Ans, Anisian; Cap, Capitanian; Chx, Changhsingian; Crn, Carnian; Ind, Induan; Lad, Ladinian; Nor, Norian; Ole, Olenekian; Wor, Wordian; Wuc, Wuchiapingian.

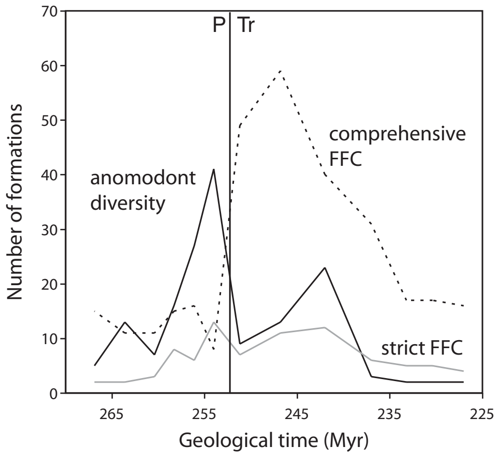


**Fig. 3.** Diversity of anomodont species through the Permo-Triassic, showing the close matching of the time series for anomodont diversity (black line) and anomodont-bearing formations (grey line; strict FFC). The time series for *all* tetrapod-bearing formations (dashed line; comprehensive FFC) follows a rather different course. The Permo-Triassic (P-Tr) boundary is indicated by a vertical line. Based on data in Fröbisch (2008, 2009), and unpublished.

anomodont species and the strict FFC, but a highly insignificant negative correlation with the comprehensive FFC.

As a non-independent metric, these results suggest that anomodont FFC cannot be used as a sampling proxy or correcting factor for the time series of anomodont genera because the two signals are essentially identical. Much more fruitful would be to compare the diversities of anomodonts with total numbers of tetrapod-bearing formations (Table 2), and to seek to understand why, for example, anomodonts were apparently highly diverse during *Tapinocephalus*, *Cistecephalus*, *Dicynodon*, and *Cynognathus* C Zone times when the comprehensive FFC was relatively low (Fig. 3).

### *Pterosaurs: redundancy and the Lagerstätten effect*

The fossil record of pterosaurs is an extreme example of episodicity (Barrett *et al.* 2008; Butler *et al.* 2009), where a dozen or so Lagerstätten, such as the Late Triassic Zorzino Limestone of North Italy, the Late Jurassic Solnhofen Limestone of Germany, the Early Cretaceous Santana Formation of Brazil and the Jehol Group of China, and the Late Cretaceous Niobrara Chalk Formation of North America account for more than half the species and genera of pterosaurs ever recorded. This gives the time series of pterosaur species occurrences a spiky appearance (Fig. 4a), each peak representing one or more Lagerstätten.

When the data are detrended and normalized, the strong correlation between pterosaur-bearing formations and pterosaur diversity remains with first differences (FD), but becomes a highly non-significant negative correlation with generalized differencing (GD; Table 3). This mixed finding suggests that much of the correlation between pterosaur palaeodiversity and pterosaur-bearing formations may relate to the overall trend (GD), but the surviving correlation with FD may be a trivial result because the count of pterosaur-bearing formations (PBF) is not a comprehensive FFC (it includes formations from which named pterosaurs were found as well as others that yielded pterosaur fragments), and so both curves are intimately linked, each documenting the same episodic preservation of fossils, and it cannot be said that one metric explains the other. The key question is whether a patchy fossil record such as this is simply tied to the Lagerstätten, or might also reflect some wider dependence on rock volume or accessibility.

**Table 2.** *Correlations of anomodont species diversity through time with different proxies for formation numbers, showing rank-order correlations for the raw data and for first differences (FD) and generalized differences (GD)*

| | Spearman's $\rho$ | $P$ |
|---|---|---|
| Anomodont-bearing formations | 0.60 | 0.034* |
| Randomized species numbers ($n = 40$) | 0.53–0.66 | 0.015* − 0.062 |
| Mode | 0.60 | 0.029* |
| Mean | 0.60 | 0.033* |
| FD anomodont-bearing formations | 0.58 | 0.041* |
| GD anomodont-bearing formations | 0.76 | 0.006** |
| All tetrapod-bearing formations | −0.22 | 0.472 |
| FD all tetrapod-bearing formations | −0.13 | 0.646 |
| GD all tetrapod-bearing formations | −0.12 | 0.716 |

Probabilities ($P$) for each correlation measure are given, and these are marked as significant ($P < 0.5$*) and highly significant ($P < 0.005$**).

**Table 3.** *Correlation of phylogenetically corrected diversity estimate (PDE) for species of pterosaurs (actual records plus ghost ranges), from Butler* et al. *(2009) with number of pterosaur-bearing formations (PBFs) from Butler* et al. *(2009), counted as raw data, $log_{10}$ of raw values, to create a normal distribution of the data, and first differences (FD) to detrend the data, and divided by substage duration (∂t) to standardize for time. Comparisons are also made of raw pterosaur counts (TDE) and PDE with dinosaur-bearing collections (DBC) and dinosaur-bearing formations (DBF), as raw data and $log_{10}$-tranformed data, both from Butler* et al. *(2011), to approximate a comprehensive FFC*

| | Pearson's $r$ | $P$ | Spearman's $\rho$ | $P$ | Kendall's $\tau$ | $P$ |
|---|---|---|---|---|---|---|
| PDE v. PBF | 0.61 | 0** | 0.56 | 0** | 0.46 | 0** |
| $\log_{10}$ (PDE v. PBF) | 0.49 | 0** | 0.56 | 0** | 0.46 | 0** |
| FD (PDE v. PBF) | 0.40 | 0** | 0.34 | 0** | 0.30 | 0.002** |
| GD (PDE v. PBF) | −0.05 | 0.645 | −0.09 | 0.429 | −0.07 | 0.406 |
| $FD/\partial t$ (PDE v. PBF) | 0.97 | 0** | 0.35 | 0** | 0.29 | 0.002** |
| TDE v. DBC | 0.32 | 0.113 | 0.64 | 0** | 0.51 | 0** |
| TDE v. DBF | 0.49 | 0.012* | 0.54 | 0.004** | 0.41 | 0.005* |
| $\log_{10}$ (TDE v. DBC) | 0.60 | 0.001** | 0.65 | 0** | 0.48 | 0.001** |
| $\log_{10}$ (TDE v. DBF) | 0.59 | 0.002** | 0.55 | 0.004** | 0.42 | 0.004** |
| PDE v. DBC | 0.08 | 0.699 | 0.37 | 0.061 | 0.31 | 0.032* |
| PDE v. DBF | 0.20 | 0.319 | 0.21 | 0.314 | 0.17 | 0.238 |
| $\log_{10}$ (PDE v. DBC) | 0.21 | 0.313 | 0.36 | 0.068 | 0.31 | 0.034* |
| $\log_{10}$ (PDE v. DBF) | 0.14 | 0.499 | 0.20 | 0.333 | 0.17 | 0.246 |

Probabilities ($P$) for each correlation measure are given, and these are marked as significant ($P < 0.5$*) and highly significant ($P < 0.005$**).

If the measure of rock volume had been a comprehensive one, such as 'all fossiliferous Mesozoic rock units' or 'all Mesozoic rock units with vertebrate fossils', then the sampling measure would be less evidently redundant with the pterosaur fossil record. In this study, we chose counts of all dinosaur-bearing formations and all dinosaur collections from the PaleoDB, as given by Butler *et al.* (2011), as proxies for comprehensive FFCs (Fig. 5). Correlations are strong with the raw diversity measure (TDE), but limited with the phylogenetically corrected measure (PDE) (Table 3; Fig. 5). For the raw data, all three correlation measures yielded largely significant results, although the strict correlation (Pearson) provides generally less significant results than the rank-based measures (Spearman, Kendall). This suggests, perhaps surprisingly, that there is a sampling signal linked to the wider availability of suitable rocks through the Mesozoic, lying behind the dominant sampling signal from the small number of crucial Lagerstätten. Interestingly, linear correlations were very poor (Pearson's $r$), rank-order correlations were poor (Spearman's $\rho$), and yet the phylogenetically corrected pterosaur numbers and dinosaur-bearing collections showed some evidence that rises and falls were in phase with each other (Kendall's $\tau$). This study confirms that the pterosaur fossil record is dominated by ten or twelve Lagerstätten, as already shown, but that the Mesozoic record of fossiliferous units (whether DBF or DBC) apparently follows the pattern of occurrence of those Lagerstätten, and so covaries with the pterosaur palaeodiversity curve to some extent as well (Fig. 5); in other words, and unexpectedly, the concentration of Lagerstätten in the Late Triassic, late Jurassic, and mid Cretaceous matches times of high numbers of dinosaur collections in the PaleoDB.

To return to the Butler *et al.* (2009) paper, it is not, however, clear what the modelled pterosaur diversity curve, with wider FFC removed, actually documents (Fig. 4a): it is hardly a 'true' or corrected global signal of pterosaurian palaeodiversity because Lagerstätten and rock volume have nothing to do with each other. In other words, if a particular fossil record is dominated by Lagerstätten, rock volume and diversity need not correlate. If there had been a clear correlation between number of formations and palaeodiversity, so that each spike in diversity really averaged out across several formations or localities, then this would not be a record dominated by Lagerstätten. Both could occur at the same time, with rock volume and palaeodiversity rising and falling together and, on top of that, times of particularly high diversity might be made even more pronounced by the presence of one or more Lagerstätten. However, these are two separate things – Lagerstätten are rich in fossils, not rich in rock.

If this is the case, then a fossil record dominated by Lagerstätten, such as that of pterosaurs, is largely determined by intimate details of how each extraordinarily rich deposit is exploited – the 'Jehol peak' (Aptian–Albian), for example, was zero a

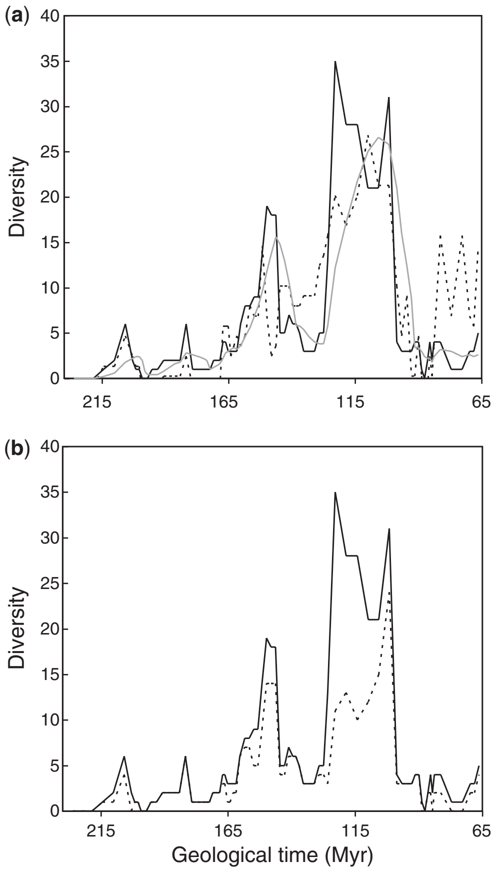


**Fig. 4.** Diversity of pterosaur species through the Mesozoic. (**a**) Raw species count (solid black line), modelled species diversity, with the effect of number of formations removed (dashed line), and five-point moving average (grey line). (**b**) Raw species diversity according to a 2008 data base (solid black line), and based on the number of pterosaurian taxa named before 1990 (dashed line). Data from Barrett *et al.* (2008) and Butler *et al.* (2009).

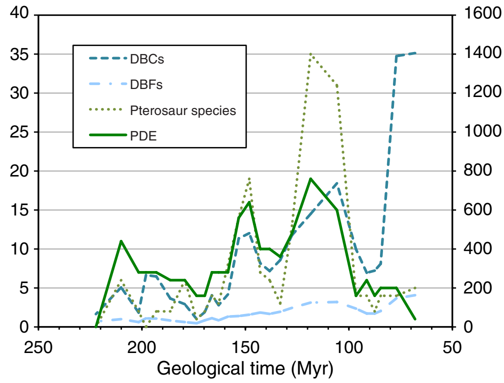


**Fig. 5.** Covariance of pterosaur diversity and wider sampling measures, the number of dinosaur collections in the PaleoDB (DBC) and the number of dinosaur-bearing formations (DBF). The DBC measure in particular shows a spiky pattern, with highs in the Late Jurassic (Kimmeridgian, Tithonian), mid Cretaceous (Aptian, Albian), and latest Cretaceous (Campanian, Maastrichtian). The first two of these peaks correspond to times of pteroaur-bearing Lagerstätten, and this is highlighted especially in the raw pterosaur palaeodiversity measure (TDE), but much less so in the phylogenetically corrected measure, including ghost ranges (PDE). Correlation measures are shown in Table 3.

few years ago (Fig. 4b), and now consists of 20 pterosaur species, and soon might reach 25 or 30, or it might fall if [sp.] taxonomic revisions reveal some synonymy of species names. The Jehol peak is founded on the fossil contents of two formations, the Yixian and Jiufotang, in NE China. The raw data, and the modelled sampling time series, are equally dependent on the current state of research, and so neither can 'correct' the other. This can be illustrated by stripping the raw data back to the position at the end of 1989 – all post-1990 finds are excluded from the Barrett *et al.* (2008) data base, and the high peaks in particular are substantially reduced (Fig. 4b). This removal of the past 20 years of research effort effectively halves the total number of pterosaur species known (from 118 today to 66 at the end of 1989), but 20 of the 52 species removed come from the Yixian and Jiufotang formations of China, eight from the Crato and Santana formations of Brazil, and the remainder scattered throughout other less productive localities. The great majority of new finds reported since 1990 are from previously known formations, and most are from a small number of Lagerstätten. The shape of the diversity time series, whether based on raw or modelled data, in a Lagerstätten-driven signal such as the pterosaur fossil record, is dependent more on intensity of collecting in known Lagerstätten rather than the number and distribution of those fossil-bearing formations. Further, at any time a new Lagerstätte may be found or exploited, as the Jehol Group formations were in the 1990s, and the addition of numerous fossil taxa corresponds to only a comparatively trivial addition to the FFC, the basis of the modelling approach employed by Butler *et al.* (2009).

Is there any meaningful way to turn such a Lagerstätten-driven fossil record into a sampling-free distribution? One might apply various techniques to reduce the spikiness of the plot, such as reading only the residuals after formation numbers have been considered, or rarefying, or shareholder quorum sampling (Alroy 2010), although the latter two are likely to return a flat line. The modelling approach (Smith & McGowan 2007) is to rank the species and number of formations time series, calculate their straight-line relationship ($y = 1.1076x - 4.1948$, in this case; Butler *et al.* 2009), then apply a correction by subtracting this modelled diversity estimate (MDE) from the observed diversity estimate (TDE) for the time bin, that is assuming that true diversity is constant and observed diversity is driven entirely by sampling. The modelled diversity estimates track the raw data closely for much of the signal (Fig. 4a), but lie below or above in places, so suggesting the influence of factors other than sampling on the signal (Butler *et al.* 2009). However, as noted, this method has not accounted for sampling, and it has probably removed much real diversity signal.

A second approach might be to seek to smooth the spikiness of the species diversity time series. This has already been done to some extent in the lumping approach taken to the stratigraphic substages, similar to the method employed in Barrett *et al.*'s (2009) dinosaur study. For example, the Yixian Formation is dated by Barrett *et al.* (2008) as 'late Barremian to Aptian', the Jiufotang Formation as 'Aptian', the Crato Formation as 'late Aptian to early Albian', and the Santana Formation as 'Aptian–Albian'. These age designations correctly reflect current uncertainties, and they vary from 2 to 6 substages in duration. Each valid pterosaurian species from those formations was then scored from 2 to 6 times, depending on the age uncertainty, but this bears no relation to the actual age, which might eventually turn out to fall entirely within a single substage for each of the formations. A similar 'smearing effect' could be achieved by adopting a 5-point moving average, for example (Fig. 4a), but there is no justification for either approach. Alternatively, a 'tightrope' could be drawn, linking the high peaks of pterosaurian diversity, based on the assumption that the Lagerstätten reveal something about the true diversity of pterosaurs at any time. This approach at least avoids the problem of all sampling standardization techniques in that they penalize the best fossil records in favour of the poorest fossil records in a time series. However, any such corrections are transient, dependent on minute details of the study of a small number of geological formations, and impossible to interpret, representing as they do sporadic and geographically restricted samples. Probably both approaches are best avoided with such sporadic fossil records as that of pterosaurs in that any 'corrections' add levels of uncertainty and hypothesis to an already uncertain and patchy fossil record. In conclusion, the pterosaur record is patchy – we know that – and for phylogenetic interpretation we can identify weaker and stronger episodes, but statistical manipulations probably add little information.

## *Collector curves and age v. clade metrics*

In both studies, the authors (Fröbisch 2008; Butler *et al.* 2009) argued that they had demonstrated that their fossil records were biased. In doing so, they rejected the common cause hypothesis, and did not consider the redundancy hypothesis advanced here.

Certainly, we would argue that Fröbisch (2008) did not demonstrate 'an obvious rock record bias affecting the diversity curve of anomodonts during at least parts of the Permian and Triassic', nor that Butler *et al.* (2009) showed that the pterosaur fossil record is 'controlled by geological and taphonomic megabiases rather than macroevolutionary processes'. Although they are almost certainly right, their method did not demonstrate what was claimed.

The key point of the redundancy hypothesis is that it rejects the possibility of using the rock volume measure as a sampling proxy. It does not address whether the record is biased or not. Our point is that the studies by Fröbisch (2008) and Butler *et al.* (2009) tell us nothing about whether the fossil records of anomodonts or pterosaurs are good or bad – other investigations are needed to assess that. Nor are we arguing that these two papers are uniquely uninformative – such assumptions have been made in many other papers, all of which require careful reconsideration along the lines we suggest.

Our key concern is that, in cases such as these, the authors show a 'corrected' diversity curve, as if the error has been removed (e.g. Fröbisch 2008; Barrett *et al.* 2009; Butler *et al.* 2009). And yet, having failed to distinguish the empirical fossil record signal from the sampling signal (strict FFC data for anomodonts; wider FFC data for pterosaurs), the 'corrected' curves might represent something closer to the true diversity pattern than the uncorrected curves, but equally they might not. In other words, modifying the raw data with information from any measure of FFC, gives a different pattern of diversity through time, but it is unclear what is represented.

Something more can be said about the quality of the fossil record of anomodonts and pterosaurs. Even though the formations proxy approach has said nothing about the quality of these respective fossil records, there are at least two established methods that provide some insights, namely collector curves and age v. clade metrics. Collector curves (= species discovery curves) for anomodonts and pterosaurs show rather different trajectories for each group (Fig. 6). The anomodonts show a steady accretion of new species from 1850 to the present day, close to the pattern of species discovery detected for North American fossil mammals and for trilobites. The pterosaurs, on the other hand, show a rapid accumulation of valid species in the nineteenth century, relatively faster than for any of

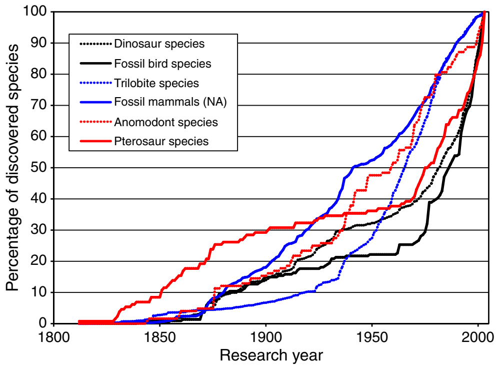


**Fig. 6.** Collector curves, or species discovery curves, for several fossil groups, including anomodonts and pterosaurs. All discovery curves are shown as percentages, even though final totals, in 2003, are very different: trilobites ($n = 4126$), early tetrapods ($n = 515$), dinosaurs ($n = 694$), fossil birds ($n = 221$), fossil mammals of North America ($n = 3340$), anomodonts ($n = 124$), and pterosaurs ($n = 130$). The 50% line marks the 'half life' of the discovery curve, the date by which half the currently valid taxa had accumulated. Data from these sources: trilobites (Tarver *et al.* 2007), dinosaurs (Benton 2008*a*), fossil birds (Fountaine *et al.* 2005), fossil mammals (Alroy 2002), anomodonts (Fröbisch 2008, 2009), and pterosaurs (Butler *et al.* 2009; Dyke *et al.* 2009).

the other test groups (although note the relatively small sample size for pterosaurs), and somewhat akin to the species discovery record of fossil birds, although beginning much earlier. Fossil birds, dinosaurs, and pterosaurs share a pattern of steeply rising rates of species discovery since 1970, a pattern not seen in the other groups. None of the species discovery curves (Fig. 6) shows a convincing asymptote, although North American fossil mammals come closest, followed by anomodonts. These observations suggest that the potential anomodont fossil record is probably better explored than the potential pterosaur fossil record, confirming the expectation of most palaeontologists. The collector curve approach cannot shed any light on how well the fossil records of these taxa reflect reality (Fig. 1).

An alternative method of exploring quality and sampling, and one that does have the potential to compare the fossil record with reality, is to confront age (stratigraphic) and clade (phylogenetic) data (Norell & Novacek 1992; Benton & Storrs 1994; Benton *et al.* 2000): good congruence between the two indicates that the phylogeny is reasonably accurate and that the fossil record is good enough to document fossils in the right order, whereas low congruence could mean that either the phylogeny or the fossil record, or both, are at fault. Age v. clade congruence metrics for anomodonts and pterosaurs are good, but not exceptional. For example, 13 cladograms of synapsids and therapsids, the larger clades including anomodonts, have Stratigraphic Consistency Indices (SCI; Huelsenbeck 1994) of 0.60–0.86 (mean, 0.74), Relative Completeness Indices (RCI; Benton & Storrs 1994) of 66.7–97.9 (mean, 80.4), and Gap Excess Ratios (GER; Wills 1999) of 0.50–0.96 (mean, 0.82), all well above the mean values for a sample of 1000 cladograms of all taxa (Benton *et al.* 2000), including plants, invertebrates, and vertebrates (SCI, 0.55; RCI, 31.13; GER, 0.56). In all cases, values of 1.00 (SCI, GER) or 100 (RCI) indicate perfect congruence. In his comparison of anomodont cladograms, Angielczyk (2002) found a range of RCI values from 28.2–63.2 (mean, 41.5), and GER values from 0.68–0.86 (mean, 076), again well above the global means, and so suggesting that, at the scale of genera and stratigraphic stages or substages, the anomodont fossil record shows better-than-average congruence with phylogeny. In the case of pterosaurs, Dyke *et al.* (2009) found RCI of 39.4 and −102.1, SCI of 0.58 and 0.62, and GER of 0.85 and 0.82 for two cladograms of pterosaurs; apart from the devastatingly low RCI value for the second cladogram, which reflects an enormous amount of ghost range, the values are well above average, suggesting generally excellent congruence between the fossil record of pterosaurian genera and phylogeny.

## Exploring fossil record incompleteness

The position reached so far is not that the fossil record is good or bad, but rather that many of the global-scale methods used recently to explore the bias and incompleteness of the fossil record fail in their core aim. The key question in the minds of palaeontologists is whether the fossil record is adequate to make a particular macroevolutionary or palaeobiological study, or not. Two major subsidiary issues are (1) testing the validity of the bias, common cause, and redundancy hypotheses, and (2) seeking to correct the empirical fossil record time series to generate a truer signal.

We consider these two issues first, and then outline four approaches for exploring error and bias in the fossil record, (1) regional exploration of geological completeness; (2) regional and local exploration of sampling completeness; (3) phylogenetic and gap-counting methods; and (4) model-based comparison of sampling bias and other explanatory variables.

### *Common cause or bias?*

Peters (2005, 2008) advanced the common cause hypothesis as the best explanation for the pervasive correlation of fossil and rock volume signals in the marine realm, and Smith (2007*a*) argued against it and for the bias hypothesis. So far, a decisive test has not been attempted on the global-scale data.

In a more focused study, Butler *et al.* (2011) attempted to test between the bias and common cause hypotheses in the evolution of dinosaurs; they showed how the various signals are mixed and indeed how difficult it is to devise a conclusive test. They found strong correlation among all metrics, namely between the fossil record signal of dinosaurian species diversity through time with measures of sampling (dinosaur-bearing collections from the PaleoDB, and dinosaur localities), and with measures of sea-level (two sea-level curves, estimated non-marine surface areas). The results became clearer after detrending, when the strong linkage between diversity and the sampling measures was confirmed, but not with the sea-level measures. They rejected the 'terrestrial common cause hypothesis' and they considered 'variation in sampling to be the preferred null hypothesis for short-term diversity variation in the Mesozoic terrestrial realm.' This is still suggestive, however, as Butler *et al.* (2011) acknowledge, when they add that 'The long-term trend towards increased sampling and dinosaur taxic diversity through the Mesozoic may result from a genuine increase in dinosaur diversity through this time period, increased opportunities to sample dinosaurs in younger rocks, or a combination of these two factors.' When numerous metrics covary in different ways and to varying degrees, it is hard to reject the influence of one or other factor.

A further issue with this study is that one 'terrestrial common cause' model has been rejected, but others almost certainly exist: is the terrestrial rock record and biodiversity driven by sea-level, continental area, mountain building, rifting, or climate change? All could affect rock volume and biodiversity. Most likely, habitable areas on land are determined by a combination of these factors, and terrestrial biodiversity may be influenced by a combination of such physical environmental drivers as these, as well as opportunism and the evolution of novel adaptations (e.g. the evolution of the ability to fly in pterosaurs and birds), and so it would be hard to capture such complexity in a comparison of diversity and physical signals.

### *Correcting the fossil record for sampling*

Attempts have been made to correct the raw fossil record data with evidence from sampling proxies, namely outcrop areas (e.g. Smith 2001; Smith & McGowan 2007; Wall *et al.* 2009) or formation counts (e.g. Barrett *et al.* 2009; Benson *et al.* 2010; Butler *et al.* 2010). In these cases, the method calculates a modelled diversity estimate that represents the diversity expected if observed diversity variations result solely from the correcting factor (outcrop area or formation count). Diversity residuals (i.e. the differences between modelled diversity values and actual diversity values) following correction for sampling then provide the supposedly sampling-free signal that may represent genuine biological signal, or may be explained in other ways. Note, however, that the 'sampling-free' residuals left after sampling standardization by outcrop areas differ in the studies by Smith & McGowan (2007) and Wall *et al.* (2009), as a result of the different palaeodiversity estimates and the different metrics of outcrop area used by both teams. It is not clear then which of these two 'corrected' curves is likely to be more informative about the true palaeodiversity signal, or whether in fact either of them is closer to the truth than the empirical data.

We have presented evidence above that number of formations is a poor global sampling proxy because of huge variations in the scale and definition of formations, and because formations document sediment and fossil heterogeneity and so are not independent of the signal they seek to correct. Further, map (= outcrop) area may also be suspect as a global sampling proxy because it does not always correspond to exposure area, perhaps a closer measure of rock availability (Dunhill 2011, in press). Therefore, these proxies on their own may be inadequate as simple correction metrics, and yet they might be elaborated to assess rock volume, accessibility, and human effort by the use of alternative metrics such as counts of collections or localities, considerations of fossil quality, and Lagerstätten. In all these cases, however, the risk of circularity (two-way causation), in other words the partial to complete redundancy of empirical and sampling signals, as noted above for formation counts and palaeontological interest units (Raup 1977; Smith 2007*a*), must be considered. It is not wise to term any of these sampling proxies simply 'sampling', as if the complex interdependence of data and sampling signal does not exist.

As argued below, the use of more subtle sampling measures, such as the number of localities or fossil collections, used as a basis for sampling standardization by several authors (e.g. Alroy *et al.* 2001, 2008; Crampton *et al.* 2003; Benton *et al.* 2004; Alroy 2010; Butler *et al.* 2011), may offer a better approach. Other sampling measures might include number of specimens (whether raw numbers or relative numbers), dispersion of sampling sites (Barnosky *et al.* 2005), and quality of specimens (completeness of preservation), but all of these are unlikely to be practical for global-scale studies incorporating diverse taxa. They are perhaps best employed at regional or clade scale, where rarefaction and other sample standardization methods may be used (e.g. Raup 1975; Benton *et al.* 2004; Barnosky *et al.* 2005).

It should be noted that rarefaction, although commonly used to standardize sample sizes, makes the assumption that 'even sampling is fair sampling', which requires that organismal abundance is constant through time, which is unlikely. In addition, there may also be a problem with community composition – if a diverse community is dominated by one or more particularly abundant taxa, it will be undersampled in terms of diversity. Overall, subsampling global data by rarefaction risks seriously 'dampening' the results, reducing peaks in palaeodiversity to a flat line (Bush *et al.* 2004; Marshall 2010). Alroy's (2010) shareholder quorum sampling method seeks to reduce this problem, ensuring that uncommon taxa are more fairly represented, but the method can only retrieve 'most of the common taxa and a stochastic assortment of the rare ones', and so some of the global signal-damping effects of rarefaction are retained. Finally, it may be that rarefaction, and equivalent techniques, really lead to a 'relative' diversity estimate, rather than an absolute one, being equivalent to, for example, 'diversity per X samples'. It is not clear how this relates to actual palaeodiversity.

This then casts doubt on the usefulness of global palaeodiversity curves 'corrected' by the use of sampling standardization (e.g. Alroy *et al.* 2001, 2008; Alroy 2010). The correction techniques themselves have been questioned (Bush *et al.* 2004), and these approaches correct only the collections included in the study, and do not consider missing collections (Smith 2007*a*). Further, the empirical curves, and the 'corrected' versions have evolved substantially over the ten-year span of these studies, as more data have been added to the PaleoDB. The 'corrected' curves differ from the empirical curve (Sepkoski 1993; Benton 1995) not only in suggesting that diversity in the sea reached modern levels in the Palaeozoic, but also in highlighting elevated diversity spikes in the Devonian, Permian, Late Cretaceous, and Palaeogene. These could be novel discoveries that require explanation, or they could reflect uneven data entry into the PaleoDB, or they may have been generated in part from the data manipulations.

## *Testing and correcting for bias*

*Regional exploration of geological completeness.* Peters and colleagues (Peters 2005, 2008; Peters & Heim 2010) have pioneered a new approach to investigating the completeness of the rock record by focusing on gap-bound sediment packages. Some 19 000 such units spanning the Phanerozoic, and encompassing all recorded surface and subsurface rock sections from the United States and Canada, are compiled in their macrostratigraphic database 'Macrostrat' (http://macrostrat.geology.wisc.edu). These sediment packages are not subject to human whim or dependent on habitat or fossil heterogeneity, as are geological formations. In their analysis of these data, Peters & Heim (2010) identify a long-term increase in rock record completeness through the Phanerozoic, with many rises and falls, an especially high peak in the latest Cretaceous, and a dip to early Mesozoic completeness levels in the Neogene. The Cretaceous peak and Neogene dip correspond to a similar phenomenon reported by Wills (2007) in assessing congruence indices through geological time, perhaps indicating a real pattern of rock record completeness.

These stratigraphic data allow detailed estimates of rock volume through time, as well as estimates of completeness of representation of fossiliferous units taken from the PaleoDB. The Mesozoic and Cenozoic are better sampled than the Palaeozoic; on average, Cenozoic time intervals have a geological completeness that is approximately 40% greater than mean Palaeozoic completeness (Peters & Heim 2010).

This approach to assessing geological completeness, limited to North America at present, has the benefit of representing sedimentary rock volume in a more comprehensive and accurate manner than counts of geological formations or map areas. The implications for assessing fossil record bias through time may also be important. For example, sampling standardization of PaleoDB data using rarefaction and equivalent techniques, omission of taxa with extant members, and other data processing approaches (Alroy *et al.* 2001, 2008; Alroy 2010) produce corrected curves for marine diversification through time that confirm Raup's (1972) bias simulation model, namely that most of the apparently low diversity levels in the Palaeozoic and Mesozoic represent sampling failure, and that the rise in diversity over the past 100 Ma is not real. The Macrostrat data, on the other hand, seem to imply that sampling cannot be solely or even largely responsible for the observed increase in marine generic diversity in the past 100 Ma (Sepkoski 1997), especially in view of the steep dip in sampling proportions in the Neogene according to the three criteria assessed by Peters & Heim (2010).

*Regional and local exploration of sampling completeness.* Areas of outcropping sedimentary rock may not yield any fossils at all, and large expanses of homogeneous outcrop might very well yield the same fossil assemblage throughout, or might show increasing diversity with area. It may, therefore, be preferable to use a measure of sampling directly related to the diversity data under scrutiny. Direct measures of sampling might be the volume of material collected (e.g. number of specimens), the number of individual find spots within a location, or an assessment of the effort involved. Although relatively straightforward to apply to small-scale field studies, it is difficult to apply such direct measures to large-scale palaeodiversity studies, and at best it may be possible to compare global or continental datasets to the number (Fara 2002; Lloyd *et al.* 2008) or area (Barnosky *et al.* 2005) of recorded fossil-bearing localities.

Direct sampling measures are most applicable to small-scale studies, where palaeontologists might sample a fixed volume (Mander *et al.* 2008) or area (Barras & Twitchett 2007) of sediment at defined sampling horizons (Little 1996), or spend a fixed amount of time sampling at each locality to ensure parity between samples. Mander *et al.* (2008) provide an example of controlling for sampling bias in their palaeoecological study of the Late Triassic mass extinction event in the SW UK, where fixed samples of 1.6 kg of sediment were collected for diversity and abundance analysis at intervals of 1 m. However, it has been noted that bulk-sampling methods are not always effective at recording rare species, and it may be necessary to integrate field samples with data from museum collections and the literature to gain a more reliable picture of palaeodiversity (Harnik 2009).

It could be contended that poorly sampled time intervals might sometimes correspond to times when fossils are of poor quality. Especially among complex organisms that are rarely preserved, such as vertebrates, it could be worth assessing whether some time bins have yielded more complete skeletons than others, and whether mean specimen completeness correlates with apparent diversity. If this were the case, then specimen quality might provide a guide to sampling.

Fossil quality has been considered in previous studies of dinosaurs (e.g. Benton 2008*a*, *b*; Mannion & Upchurch 2010) where the quality of type specimens was found to have improved through research time. In their detailed study of sauropodomorph fossils, Mannion & Upchurch (2010) found that mean skeletal completeness and mean character completeness varied between time bins, but roughly halved from the Triassic and Early Jurassic to the Late Cretaceous. Mannion & Upchurch (2010, p. 291) note that 'The ... results suggest that sea-level has, in some fashion, controlled the quality of the sauropodomorph fossil record, but only through part of the group's evolutionary history, with high sea-level correlated with low average completeness scores, and low sea-level with high completeness scores in the Jurassic–Early Cretaceous.' It is equally likely that the apparent sporadic covariation of sauropodomorph specimen quality with sea-level does not indicate a causal link at all: note that many of the completeness measures are based on relatively small sample sizes ($n = 4$–$24$ taxa), so a single locality can dominate the findings within a time bin. Perhaps such studies of fossil specimen quality based on modest numbers of specimens and localities cannot address sampling at the global scale.

In a basinwide study of the Late Permian and Early to Middle Triassic red beds of the South Urals basins (Benton *et al.* 2004), some 289 localities, assigned to 13 stratigraphic divisions in succession, have yielded 675 identified tetrapod fossils. These were assigned to four 'quality' categories, namely (1) single isolated fragments, (2) several individual elements of a taxon, (3) one or more skulls, and (4) one or more skeletons. All the noted materials, even the fragments, could be identified at least to family level, and so bone scrap is excluded. Across the whole study, the numbers of specimens in each category were 313, 288, 63, and 11 respectively. As reported before (Benton *et al.* 2004), the sampling measures of number of localities and number of specimens per time bin covary (Fig. 7a), but these do not covary with either diversity of genera or families. The 'quality' measure (number of good specimens/total number of specimens), where 'good specimens' are the complete skulls or skeletons, shows a different pattern (Fig. 7b) from either locality or specimen numbers. Ignoring the first value, which is based on a very low sample size, fossil quality in the Permian is out of synchrony with generic and familial diversity (Fig. 7b), but seems to be in line with generic diversity in particular in the Triassic. However, the measure of fossil specimen quality does not appear to covary with number of localities or specimens (Fig. 7a, b). The change in behaviour of the specimen quality measure across the Permo-Triassic boundary is probably not a sample-size artefact: Permian samples range from 11–63 specimens per time bin, excluding the first time bin (mean, 38.6) and Triassic sample sizes range from 17–147 (mean, 68). Certainly, in the late Early and Middle Triassic (time divisions 11–13), numbers of localities and specimens appear to peak in time bin 11 with number of families, but not number of genera (Fig. 7a). Further, the specimen quality index peaks in time bin 12 with number of genera (Fig. 7b). None of the three sampling measures, including specimen quality, could be said, however, to show a convincing covariation with apparent diversity, so suggesting that much of the palaeodiversity signal is probably real.

In a similar study of echinoids, Smith (2007*b*) showed that the Triassic fossil record was much poorer than that of the Jurassic. He noted that 27%

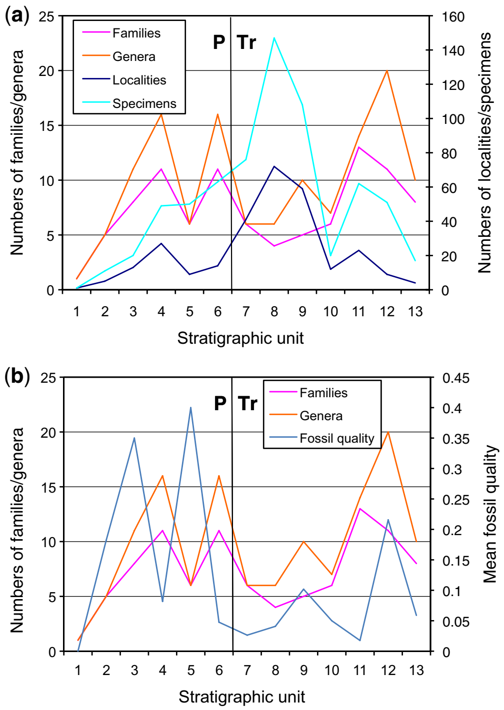


**Fig. 7.** Diversity and sampling through the Permo-Triassic continental redbeds of the South Urals basin, Russia. (**a**) Diversity of families and genera covary, as do the sampling measures of numbers of localities and specimens, but the diversity measures do not covary with the sampling measures. (**b**) Mean fossil quality (number of 'good' specimens: all specimens) does not covary with diversity of genera or families in the Permian, but appears to do so in the Triassic. Data from Benton *et al.* (2004).

Triassic species were based on relatively complete material (whole tests or whole tests plus spines), compared to 69% from the Lower Jurassic. Further, among Triassic species, 60% had been established on isolated spines or dissociated interambulacral plates, whereas only 21% of Jurassic species are named on such incomplete material. The relatively poorer Triassic record was confirmed also by slower accumulation of Triassic taxa in a comparison of collector curves, longer implied gaps from a study of molecular trees, and more ghost lineages. Smith (2007*b*) explained these differences by a combination of more limited marine rocks in the Triassic when compared to the Jurassic, as well as to evolutionary changes among the echinoids, which acquired more robust tests in the Jurassic. Such a clear demonstration of relative differences in sampling quality of members of a single clade between two time units then points to the possibility of exact numerical correction when comparing the palaeodiversity signals of Triassic and Jurassic echinoids.

*Phylogenetic and gap-counting methods.* The debate about whether number of formations is a covariate or a determinant of palaeodiversity could continue on its circular way unless additional data can be introduced. A possible source of such information might be ghost ranges and Lazarus taxa. A ghost range, or ghost lineage, is the minimum gap implied by a cladogram where the oldest fossils of two sister lineages differ in age (Norell 1992), and a Lazarus taxon (Flessa & Jablonski 1983) is a break in the record of a lineage that exists below and above a particular sampling horizon. In both cases, providing the cladogram is correct in the first case, and providing the taxa below and above the gap are the same in the second case, these can both provide independent evidence for a failure in sampling.

This has been noted before (Paul 1998). For example, Smith (2001, p. 364) pointed out, 'The only realistic way to distinguish between sampling and biologically driven patterns is to gather phylogenetic information. The key here is the recognition of ghost lineages and the stratigraphical distribution of pseudoextinctions.' Pseudoextinctions are false extinctions marking artificial truncations of lineages, which, when corrected indicate Lazarus gaps. The correct identification of such gaps can change perceptions of evolutionary pattern: for example, when Modesto *et al.* (2001) revised the cladogram of procolophonids, a clade of small reptiles from the Permian and Triassic, they found that several ghost lineages spanned the Permo-Triassic boundary, so showing that the group was not so severely affected by the Permo-Triassic mass extinction as had been thought previously.

Here, we concentrate on ghost ranges. A number of gap analyses have been carried out (e.g. Paul 1982, 1998; Flessa & Jablonski 1983; Benton 1987; Fara & Benton 2000; Smith 2001, 2007*a*, *b*; Fara 2002), and we cannot add to what was said in those papers, other than to urge caution. It has so far been generally assumed that gaps in ranges occur when fossils are not found as a result of missing rocks and missing sampling. This is doubtless commonly the case, but there is a risk that Lazarus gap analysis might still involve a measure of circularity in that the method cannot distinguish poor sampling from low abundance and diversity. A lineage that showed genuine rises and falls in abundance or breadth of geographical distribution might show Lazarus gaps even if sampling is constant throughout, but this would be an evolutionary, not a sampling, signal. This was the contention by Wignall & Benton (1999) for times of low diversity and high gap counts following mass extinctions.

A further weakness of gap analysis is that the detection of Lazarus taxa becomes harder as the gap duration increases (Benton 1987), and of course the method cannot detect gaps before and after the currently known stratigraphic range. These criticisms are true also of ghost ranges, and one might very well expect that many ghost ranges do indeed arise from low diversity and abundance of lineages and clades soon after they became established and before they had diversified fully.

Nevertheless, it is worth perhaps exploring ghost range distribution in time as an independent guide to sampling (Paul 1998). The assumptions would be (1) that ghost ranges might be distributed in time in negative proportion to putative sampling proxies such as comprehensive counts of formations or localities, and (2) that raw counts of taxa should correlate better with those putative sampling proxies than phylogenetially corrected counts of taxa. The rationale of this last suggestion emerges from a comparison of three counts of palaeodiversity, the taxon diversity estimate (TDE), a raw count of numbers of taxa reported per time bin, the ghost range diversity estimate (GDE), based on a cladogram plotted against time, and the phylogenetic diversity estimate (PDE), the sum of TDE and GDE (Barrett *et al.* 2009; Mannion & Upchurch 2010; Mannion *et al.* 2011).

We present four examples, the first two based on relatively small sample sizes, the second two on large examples, and these show broadly that phylogenetic gaps can indicate sampling failure.

*Triassic archosaurian diversity.* A recent comprehensive cladistic analysis of the relationships of Triassic archosaurs (Brusatte *et al.* 2010, fig. 8) offers a summary phylogeny plotted against time, and highlighting ghost ranges. Lazarus taxa are not shown. From this (Fig. 8; Table 4), ghost ranges were summed for substages, and compared with numbers of archosaur-bearing and tetrapod-bearing formations for those same substages (data from Benton *et al.* 2004; Sahney & Benton 2008; and unpublished). Note that formations were assigned to substages based on independent stratigraphic evidence in each case, and with no interpolation. Two counts of archosaur-bearing formations were considered, first the ‘strict’ count, taken only from the taxa included in the Brusatte *et al.* (2010) cladogram, and then the ‘all archosaur’ count, based on all archosaurs known from the Triassic. The sums of these three counts vary substantially: 37 actual archosaur (strict) FFC, 94 all-archosaur (wider) FFC, and 292 all-tetrapod (comprehensive) FFC.

To compare gaps in the stratigraphic record with ghost ranges, a measure of the ‘absence of formations’ is required. As a visual approximation (Fig. 8), the inverse of the number of formations per time bin was taken, by subtracting the actual number from the maximum possible number of formations (maximum number of formations per time bin were: actual archosaur-bearing formations, 7 in the upper Carnian and in the lower Norian; all archosaur-bearing formations, 14 in the middle Norian; all tetrapod-bearing formations, 26 in the upper Olenekian). There is no obvious visual matching (Fig. 8) of times of significant ghost range, such as the Anisian to Carnian interval, with times of lower sampling (the Ladinian and lower Carnian coincide, but the later Triassic epochs do not.

In comparisons of GDE (Table 5), the strict and comprehensive FFC gave non-significant negative associations between ghost ranges and formation counts, whereas the wider FFC correlated negatively highly significantly with ghost ranges. However, this strong correlation disappears when first differences are considered (Table 5), so the strong correlation might be an artefact of parallel trends of increasing numbers through the Triassic combined with small data sets. Comparison of TDE and PDE (Table 5) shows that TDE correlates with the strict FFC, but only at $P < 0.1$ with the wider FFC, and not at all with the comprehensive FFC. These relationships disappear with the PDE, which shows both negative and positive non-significant correlations with the sampling counts.

These results are equivocal, confirming the proposition that PDE correlates much less well with formation counts than TDE, but highlighting the odd result that ghost range counts (GDE) also correlate strongly with the wider FFC. The Triassic archosaur fossil record is not simply dependent on rock volume (no correlation with the comprehensive FFC), and it is unresolved how well ghost ranges predict sampling.

*Mesozoic bird diversity.* When the simple cladogram of Mesozoic birds from Chiappe & Witmer (2002) is compared with minimum and maximum estimates of bird-bearing formations, there is no correlation with ghost ranges, whether using the raw data or detrended data (Table 6). The minimum estimate of locality numbers (= strict FFC) consists of just the localities that yielded the bird taxa included in Chiappe & Witmer’s (2002) cladogram, whereas the comprehensive FFC comes from all records of Mesozoic birds, as documented by Fountaine *et al.* (2005). The fossil record of Mesozoic birds certainly includes very many ghost ranges (55 stage-level ghost ranges and only 29 stage-level records), and formation numbers, whether as a strict or wider FFC, might be thought to have been a suitable predictor of ghost ranges, but this is not the case whether for total or detrended data (Table 6).

When the raw palaeodiversity (TDE) is compared, however, it shows a remarkably strong

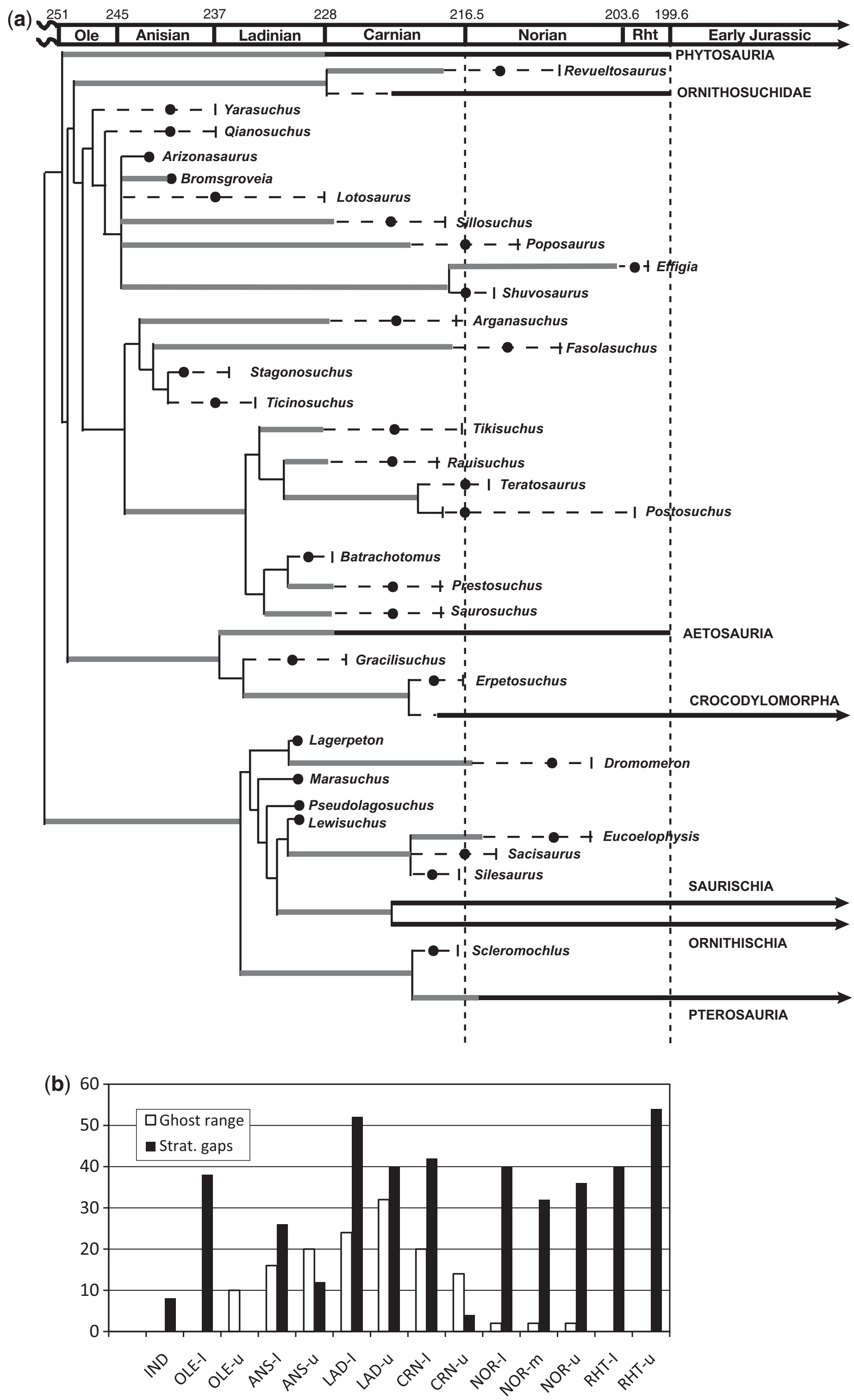
(a)
251
245
237
228
216.5
203.6
199.6
Ole
Anisian
Ladinian
Carnian
Norian
Rht
Early Jurassic
PHYTOSAURIA
Revueltosaurus
ORNITHOSUCHIDAE
Yarasuchus
Qianosuchus
Arizonasaurus
Bromsgroveia
Lotosaurus
Sillosuchus
Poposaurus
Effigia
Shuvosaurus
Arganasuchus
Fasolasuchus
Stagonosuchus
Ticinosuchus
Tikisuchus
Rauisuchus
Teratosaurus
Postosuchus
Batrachotomus
Prestosuchus
Saurosuchus
AETOSAURIA
Gracilisuchus
Erpetosuchus
CROCODYLOMORPHA
Lagerpeton
Dromomeron
Marasuchus
Pseudolagosuchus
Lewisuchus
Eucoelophysis
Sacisaurus
Silesaurus
SAURISCHIA
ORNITHISCHIA
Scleromochlus
PTEROSAURIA
(b)
Ghost range
Strat. gaps
0
10
20
30
40
50
60
IND
OLE-l
OLE-u
ANS-l
ANS-u
LAD-l
LAD-u
CRN-l
CRN-u
NOR-l
NOR-m
NOR-u
RHT-l
RHT-u

**Fig. 8.**

**Table 4.** *Comparison of phylogenetically implied gaps and formation numbers for Triassic archosaurs*

| Substage | Duration (My) | Ghost range | Archosaur formations | Stratigraphic gaps | All-tetrapod formations | Stratigraphic gaps |
|---|---|---|---|---|---|---|
| Ind | 1.5 | 0 | 5 | 9 | 32 | 4 |
| Ole(l) | 1.5 | 0 | 3 | 11 | 17 | 19 |
| Ole(u) | 1.5 | 5 | 7 | 7 | 36 | 0 |
| Ans(l) | 3.5 | 8 | 6 | 8 | 23 | 13 |
| Ans(u) | 3.5 | 10 | 6 | 8 | 30 | 6 |
| Lad(l) | 1.5 | 12 | 3 | 11 | 10 | 26 |
| Lad(u) | 1.5 | 16 | 2 | 12 | 16 | 20 |
| Crn(l) | 3.35 | 10 | 1 | 13 | 15 | 21 |
| Crn(u) | 3.5 | 7 | 13 | 1 | 34 | 2 |
| Nor(l) | 6.5 | 1 | 10 | 4 | 16 | 20 |
| Nor(m) | 6.5 | 1 | 14 | 0 | 20 | 16 |
| Nor(u) | 6.5 | 1 | 13 | 1 | 18 | 18 |
| Rht(l) | 3 | 0 | 8 | 6 | 16 | 20 |
| Rht(u) | 3 | 0 | 3 | 11 | 9 | 27 |
| Totals | 46.85 | 71 | 94 | 102 | 292 | 212 |

Data are tabulated from a recent cladistic analysis in Brusatte *et al.* (2010, fig. 8), from which phylogenetically implied gaps (‘Ghost range’) are drawn, and then compared with the inverse of the number of formations, as a measure of absence of information (‘Stratigraphic gaps’ = maximum number of formations in a time bin [14] minus actual number). Comparisons are made with the strict FFC (‘Archosaur formations’) and the comprehensive FFC (‘All-tetrapod formations’).
Abbreviations of stratigraphic stage names as in Table 1, plus Rht, Rhaetian.

**Table 5.** *Correlations of archosaurian ghost ranges (GDE, Ghost range diversity estimate) with counts of restricted and all archosaur-bearing and all tetrapod-bearing formations (strict, wider, and comprehensive FFCs) for the Brusatte* et al. *(2010) study of Triassic archosaur phylogeny, showing rank-order correlations for the raw data and for first differences (FD). Taxon diversity estimates (TDE) and phylogenetic diversity estimates (PDE = TDE + GDE) are also compared with the three formation counts*

| | Spearman’s $\rho$ | $P$ |
|---|---|---|
| GDE v. Strict FFC | −0.41 | 0.237 |
| GDE v. Wider FFC | −0.86 | 0.001** |
| GDE v. Comprehensive FFC | −0.34 | 0.337 |
| GDE v. FD strict FFC | 0.20 | 0.577 |
| GDE v. FD wider FFC | 0.15 | 0.681 |
| GDE v. FD comprehensive FFC | 0.27 | 0.451 |
| TDE v. Strict FFC | 0.79 | 0.006 |
| TDE v. Wider FFC | 0.57 | 0.088 |
| TDE v. Comprehensive FFC | −0.02 | 0.947 |
| PDE v. Strict FFC | 0.40 | 0.249 |
| PDE v. Wider FFC | −0.27 | 0.443 |
| PDE v. Comprehensive FFC | −0.24 | 0.498 |

**Fig. 8.** Phylogeny of basal archosaurs (**a**), showing dates of the major lineages and ghost ranges. (**b**) Histograms across the bottom show number of ghost ranges and a measure of the ‘absence of formations’ (= maximum number minus actual number; Strat. gaps, stratigraphic gaps) for each time bin. Abbreviations: ANS, Anisian; CRN, Carnian; IND, Induan; l, lower; LAD, Ladinian; NOR, Norian; OLE, Olenekian; RHT, Rhaetian; u, upper. A, based on data in Brusatte *et al.* (2011).

**Table 6.** *Correlations of Cretaceous bird ghost ranges (GDE) with bird-bearing FFCs, read as a strict FFC, representing only those formations with the named bird taxa (from Chiappe & Witmer 2002) and wider FFC figures (from Fountaine* et al. *2005), showing rank-order correlations for the raw data and for first differences (FD). Correlations between these measures and TDE and PDE are also given. Data are calculated from Berriasian to Campanian only, to avoid the edge effects of wide variation in number of ghost ranges in the first time bin (Tithonian), and necessary absence of ghost ranges in the last (Maastrichtian)*

| | Spearman's $\rho$ | $P$ |
|---|---|---|
| GDE v. Strict FFC | −0.44 | 0.180 |
| GDE v. Wider FFC | −0.22 | 0.518 |
| GDE v. FD strict FFC | −0.43 | 0.191 |
| GDE v. FD wider FFC | −0.17 | 0.619 |
| TDE v. Strict FFC | 0.90 | 0.000** |
| TDE v. Comprehensive FFC | 0.65 | 0.032* |
| PDE v. Strict FFC | 0.38 | 0.248 |
| PDE v. Comprehensive FFC | 0.47 | 0.147 |

correlation with a strict count of formations with fossil birds and a weaker correlation with the wider formation count (Table 6), but these two signals are doubtless essentially redundant with each other, as in the pterosaur case above. These correlations disappear for the phylogenetically corrected diversity estimate (PDE; Table 6).

In this rather extreme case, with high proportions of ghost ranges (relative completeness index $= -0.527$), these minimum estimates of phylogenetically determined gap may provide a guide to sampling that is not achievable through the various strict and wider FFCs.

*Dinosaurs.* In an attempt to go beyond such small-scale studies, an analysis of the dinosaurian fossil record was conducted. This consists of the 420 species included in the formal dinosaur supertree of Lloyd *et al.* (2008), plotted against time, using stratigraphic data to establish stage-level divisions of the Mesozoic. Dinosaurian distribution data comes from the Paleobiology database (http://paleodb.org/; download of all non-avian body fossil data on 29th June, 2010). We compared the GDE:PDE ratio, diversity (GDE) and phylogenetically corrected diversity (PDE = TDE + GDE) to three sampling proxies: (1) number of dinosaur-bearing formations (DBFs), (2) number of dinosaur-bearing localities (DBLs), and (3) the palaeoarea of a spherical polygon described by drawing a convex hull around the DBLs (Fig. 9).

As with previous dinosaur studies (e.g. Lloyd *et al.* 2008; Barrett *et al.* 2009; Butler *et al.* 2011) we find strong correlation between all of our sampling proxies and species diversity (Fig. 9). However, we note a consistent weakening of this relationship when PDE is used instead of TDE (Table 7), despite a strong correlation between GDE and sampling. The sampling proxies, dinosaur-bearing formations and dinosaur-bearing localities, doubtless mix some redundancy (many formations/localities yield a single species) with genuine sampling signal, as discussed above, and so the strong correlation between sampling proxy and palaeodiversity could reflect a mix of true sampling signal and redundancy. A better sampling proxy in these cases might be the total number of formations that have yielded any kind of vertebrate fossil, or that are of the correct facies to do so: this would allow inclusion of localities and formations that have been searched, but failed to produce dinosaur specimens.

In seeking to understand whether the relative proportion of ghost ranges might provide a more reliable guide to sampling than the traditional geological measures, the weak negative relationship between the GDE:PDE ratio and the formation/locality counts (only barely significant at $P < 0.05$ in the DBF case) is suggestive and indeed is strengthened when generalized differencing is used (McKinney 1990), where all three proxies show a strong negative correlation (Table 7). Consequently, despite being only a minimal correction it does appear that for dinosaurs at least the proportion of phylogenetically-inferred to sampled lineages is a good predictor of sampling.

Data and R code for all analyses are available from GTL.

*Baleen whales.* In a further large-scale study, taxonomic and phylogenetic diversity estimates of mysticete whales were considered. The phylogeny is based on Marx (2010, fig. 2), with *Cetotherium megalophysum* excluded owing to a lack of precise stratigraphic information. The formations that produced the taxa included in the tree of Marx (2010) (strict FFC), as well as the total number of cetacean-bearing formations (wider FFC), and all marine fossiliferous formations (comprehensive FFC) assignable to stage level, which were downloaded from the PaleoDB on 1st June, 2010, were all compared to the two diversity estimates using Spearman rank correlation (Table 7). While both the strict and wider raw FFCs showed a significant positive correlation with the raw taxonomic diversity estimate, this correlation disappeared when phylogenetic diversity was considered instead. Furthermore, the comprehensive FFC did not correlate with either taxonomic or phylogenetically adjusted diversity. When generalized differences were used instead of the raw data, the correlation of the taxonomic diversity estimate with all three formation counts was

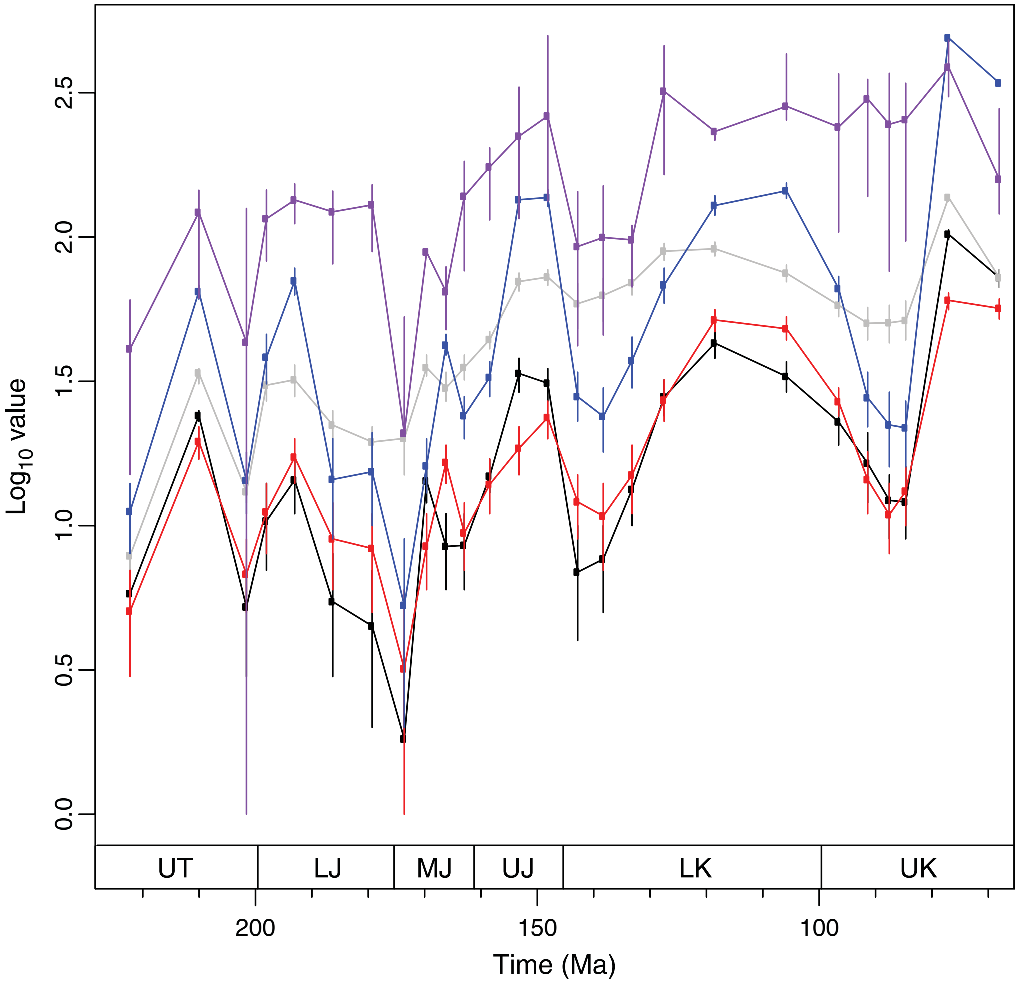


**Fig. 9.** Dinosaur diversity and sampling. Taxonomic Diversity Estimate (TDE; black), Phylogenetic Diversity Estimate (PDE; grey, based on Lloyd *et al.* 2008), Dinosaur-bearing Formations (DBFs; red), Dinosaur-bearing Localities (DBLs; blue) and palaeoarea of a spherical polygon encompassing the DBLs (purple). Values are logged to allow plotting on same scale. NB: Palaeoarea measure is further modified to allow plotting on same scale as values are orders of magnitude larger than for other variables. Vertical lines indicate 95% confidence interval reflecting 1000 randomizations of dating uncertainty. Stratigraphic divisions: UT, Upper Triassic; LJ, Lower Jurassic; MJ, Middle Jurassic; UJ, Upper Jurassic; LK, Lower Cretaceous; UK, Upper Cretaceous.

weakened, and indeed obliterated in the case of the strict FFC. By contrast, the correlation of the phylogenetic diversity estimate with the rock record was strengthened, but also turned negative in all cases. However, none of the correlations were statistically significant following differencing.

*Perspective.* In light of the need to distinguish between the bias and common cause models (e.g. Peters 2005, 2008; Smith & McGowan 2007), the observation that phylogenetic diversity estimates seem to decrease or remove existing correlations between taxonomic estimates and a range of different measures of sampling gives rise to two possible interpretations. First, if it were assumed that the phylogenetic trees used are a reasonable representation of biological reality, the weakening of the diversity-sampling correlation might indicate that observed taxonomic diversity is largely driven by bias, with the phylogenetically adjusted estimate offering an improved and, presumably, truer picture of biological reality. However, if the cladograms we used to perform this correction were in some way flawed, 'correcting' diversity based on

**Table 7.** *Correlations of dinosaur and mysticete taxonomic and phylogenetic diversities with different measures of sampling*

| Taxon | Correlation | Spearman's $\rho$ | $P$ |
|---|---|---|---|
| Dinosauria | GDE:PDE v. DBF | −0.39 | 0.047* |
| | GDE:PDE v. DBL | −0.37 | 0.060 |
| | GDE:PDE v. Palaeoarea | −0.30 | 0.138 |
| | TDE v. DBF | 0.91 | 0.000** |
| | TDE v. DBL | 0.88 | 0.000** |
| | TDE v. Palaeoarea | 0.73 | 0.000** |
| | PDE v. DBF | 0.80 | 0.000** |
| | PDE v. DBL | 0.76 | 0.000** |
| | PDE v. Palaeoarea | 0.70 | 0.000** |
| | GDE v. DBF | 0.71 | 0.000** |
| | GDE v. DBL | 0.66 | 0.000** |
| | GDE v. Palaeoarea | 0.66 | 0.000** |
| | GD GDE:PDE v. DBF | −0.64 | 0.000** |
| | GD GDE:PDE v. DBL | −0.65 | 0.000** |
| | GD GDE:PDE v. Palaeoarea | −0.51 | 0.010* |
| | GD TDE v. DBF | 0.80 | 0.000** |
| | GD TDE v. DBL | 0.82 | 0.000** |
| | GD TDE v. Palaeoarea | 0.73 | 0.000** |
| | GD PDE v. DBF | 0.67 | 0.000** |
| | GD PDE v. DBL | 0.78 | 0.000** |
| | GD PDE v. Palaeoarea | 0.63 | 0.001** |
| | GD GDE v. DBF | 0.55 | 0.006* |
| | GD GDE v. DBL | 0.64 | 0.000** |
| | GD GDE v. Palaeoarea | 0.48 | 0.016* |
| Mysticeti | TDE v. strict FFC | 0.64 | 0.035* |
| | PDE v. strict FFC | 0.30 | 0.366 |
| | TDE v. wider FFC | 0.71 | 0.015* |
| | PDE v. wider FFC | 0.17 | 0.610 |
| | TDE v. comprehensive FFC | 0.40 | 0.227 |
| | PDE v. comprehensive FFC | −0.23 | 0.503 |
| | GD TDE v. strict FFC | −0.01 | 0.973 |
| | GD PDE v. strict FFC | −0.60 | 0.067 |
| | GD TDE v. wider FFC | 0.50 | 0.144 |
| | GD PDE v. wider FFC | −0.53 | 0.105 |
| | GD TDE v. comprehensive FFC | 0.39 | 0.248 |
| | GD PDE v. comprehensive FFC | −0.33 | 0.330 |

*Abbreviations*: DBF, dinosaur-bearing formations; DBL, dinosaur-bearing localities; FFC, fossiliferous formation count; GD, generalized differences (McKinney 1990); PDE, phylogenetic diversity estimate; TDE, taxonomic diversity estimate; strict FFC, number of formations from which the taxa in the tree were recovered; wider FFC, total number of cetacean-bearing formations, as downloaded from the PaleoDB; comprehensive FFC, total number of marine fossiliferous formations as downloaded from the PaleoDB.
Probabilities ($P$) for each correlation measure are given, and these are marked as significant ($P < 0.05$*) and highly significant ($P < 0.005$**).

their topology might result in the addition of more noise than signal. In this case, the observed correlation between the sampling proxies and diversity might either be the result of an actual bias, or of a common cause – in either case, the addition of a large number of spurious ghost ranges could obliterate any statistically significant relationship. In addition, it is also worth noting that any phylogenetic correction fundamentally relies on the assumption that cladograms ignoring the potential presence of ancestral taxa in the fossil record are an adequate model of evolution. However, treating genuine ancestor-descendant pairs as sister taxa may lead to the inference of ghost lineages where none exists, and hence the over-inflation of taxon estimates per time bin, which could bias phylogenetic diversity corrections even if the topology of the cladogram itself were accurate.

Finally, cladograms may also suffer from other problems, including the one-sidedness of the correction they provide (for obvious reasons, no ranges leading upwards in time can be inferred from them), and non-random taxon sampling, particularly when the cladogram was constructed to analyse the relationships of a particular subgroup of the taxon in question (Lane *et al.* 2005). These factors certainly have the potential to bias diversity estimates, and, if the taxa included in the tree present a non-random or very small sample of the taxon of interest, could even lead to completely spurious diversity patterns.

*Model-based comparison of bias and other explanatory variables.* While a large number of studies have investigated the impact of geological or human bias on measures of palaeodiversity, relatively few have tried to contrast the latter with the explanatory power of potentially biologically relevant variables that might account for some, or most of the observed diversity signal (e.g. Smith & McGowan 2007; Barrett *et al.* 2009; Benson *et al.* 2010). While it may well be possible that genuine biological signals in the fossil record are overwhelmed by geological biases, this assumption needs to be tested explicitly. Correcting palaeodiversity data using some form of sampling proxy (e.g. Smith & McGowan 2007; Barrett *et al.* 2009) and then attempting to interpret the residuals in a biologically meaningful way may be counterproductive in this sense, as it runs the risk of throwing the baby out with the bathwater: if the presumed sampling proxy is either redundant with diversity or the result of a common cause, removing it from the data may obliterate the actual (biological) signal of interest, leaving little more than a flat line or random noise to be interpreted by the researcher.

One way to address this issue may be to consider both potential sampling proxies and evolutionarily meaningful variables at the same time, giving them an equal chance to explain the data of interest (e.g. Mayhew *et al.* 2007; Marx & Uhen 2010). If in such an analysis the explanatory power of sampling proxies outperforms that of the proposed biological model, the case for a large-scale bias in the data is corroborated. If, on the other hand, the biological model outperforms the bias hypothesis, a common cause or sampling proxy/diversity data redundancy explanation may be considered. Finally, the best model might also include aspects of both sampling bias and a biological signal. One example of this approach was recently implemented by Mayhew *et al.* (2007), who tested for, and found, a significant association of Phanerozoic diversity with temperature, while simultaneously assessing the effect of sampling probability on their results. Similarly, Marx & Uhen (2010) applied a series of models including food abundance, climate change, and a sampling proxy (number of fossiliferous marine formations) to late Oligocene to Pleistocene neocete whale diversity (Fig. 10; Table 8), and assessed their respective goodness-of-fit using the second-order Akaike's Information Criterion (AICc) and Akaike weights (Sugiura 1978; Burnham & Anderson 2002).

It is clear that in both cases the models chosen were far from exhaustive in their exploration of

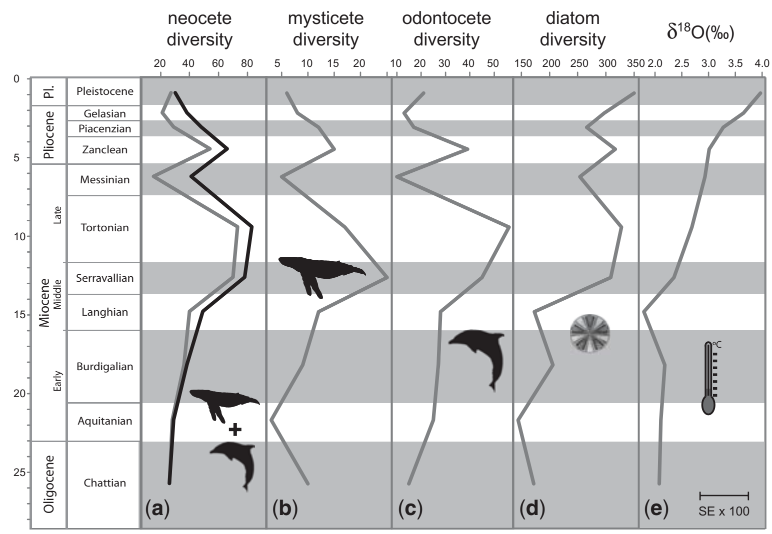


**Fig. 10.** Comparison of (**a**) neocete; (**b**) mysticete; (**c**) odontocete palaeodiversity with (**d**) diatom paleodiversity and (**e**) global $\delta^{18}O$ values (E), from Marx & Uhen (2010). Cetacean diversity is shown as sampled in bin data as downloaded from the Paleobiology Database (grey) and as a ranged through estimate (black). Based on data in Marx & Uhen (2010).

**Table 8.** *Estimated best-fit model parameters for the neocete, mysticete and odontocete datasets, as reported by Marx & Uhen (2010)*

| | Neoceti sampled in bin | | Neoceti ranged through | | Mysticeti sampled in bin | | Odontoceti sampled in bin | |
|---|---|---|---|---|---|---|---|---|
| | Estimate | S.E. | Estimate | S.E. | Estimate | S.E. | Estimate | S.E. |
| Intercept | 6.694 | 1.465 | 6.649 | 1.111 | 2.214 | 0.179 | 5.566 | 1.179 |
| st. dur. | −0.047 | 0.189 | −0.068 | 0.167 | 0.096 | 0.023 | 0.049 | 0.166 |
| Diatom | 0.029 | 0.005 | 0.028 | 0.003 | 0.015 | <0.001 | 0.020 | 0.003 |
| $\delta^{18}O$ | −2.881 | 0.577 | −2.253 | 0.422 | −1.077 | 0.081 | −2.147 | 0.351 |
| Rock | – | – | −0.013 | 0.006 | – | – | – | – |

Explanation of terms: $\delta^{18}O$, oxygen isotope records used as proxy for climate change; diatom, diatom species diversity (Neptune database); rock, total number of fossiliferous marine formations as downloaded from the Paleobiology Database; st. dur., geological stage duration; the latter was included as a non-optional predictor in all models on order to account for the potentially biasing effects of unequal Caenozoic stage durations. Based on data from Marx & Uhen (2010).

potential predictors, and other methods of simultaneously assessing the relative impact of bias and biology may be envisaged. Nevertheless, they make the point that combining sampling proxies and potential evolutionary drivers represents a more inclusive and, most likely, fairer way of assessing palaeodiversity than analysing either of these variables in isolation.

## Conclusions

While it is evident that the fossil record is incomplete, some recent approaches to identifying bias, or ‘megabias’, have been flawed. The sampling proxies, such as number of formations containing particular fossils, or map areas from particular parts of the world, may not suffice as independent evidence for sampling failure. The two signals, rock volume and palaeodiversity, often covary closely, but this need not indicate that the former drives the latter. In fact, as already suggested (Peters 2005), both may be driven by an external ‘common cause’ such as sea-level change or, in the case of terrestrial organisms, by rates of uplift and by the weather, and consequent variation in volumes of sedimentary rock accumulation.

Further, as argued here, much of the covariance of rock volume metrics and palaeodiversity is likely a result of redundancy of the signals – the number of formations containing dinosaurs is tightly linked to the number of dinosaur species because finds are sporadic and interdependent (Benton 2008*a*). Removing the formation count from the species count produces a flat line because all signals, both geological and biological, have been removed. This observation of redundancy is a criticism of the assumption that measures of rock volume are independent proxies for sampling, and it says nothing about the quality of the fossil record of dinosaurs (and other similar terrestrial taxa), which is undoubtedly patchy and incomplete.

We suggest four reasonable approaches to exploring sampling of the fossil record that avoid the problems of recent global-scale numerical explorations of covariance. First, regional-scale explorations of sampling may work because sampling metrics can be more detailed and can explore aspects of both rock volume and human effort. Further, explorations of rock volume that avoid the confusions of imprecise measurements of map areas that may not relate to rock availability (Dunhill 2011, in press) and the human quirks of geological formations (that scale over at least eight orders of magnitude), may provide independent estimators of sampling potential. A third approach may be to explore gaps (Lazarus gaps) and ghost ranges, which are both independently determined measures of known fossil absences. Our initial studies here are only moderately promising, however. A fourth approach, and perhaps the best of all because it does not assume primacy of either the fossil record or the sampling metrics, is to compare multiple models with a palaeodiversity curve, some models reflecting changes in the environment and others reflecting sampling, and yet others combining environmental change and sampling. The benefit of this approach is that there are no prior assumptions, and it assesses a variety of models for goodness of fit; the weakness is that the real drivers of palaeodiversity in any particular case may elude measurement and so may be missed.

Our proposal is that palaeontologists should be less obsessed about the poor quality of the fossil record, and that global-scale, single-hit analyses may never address the issue of whether the fossil record is good or bad, whether it is driven primarily by macroevolution or megabiases. Each time bin, each geographical region, and each clade is sampled differently, and so a global answer can probably never be found. Paul (1998) noted how palaeontologists seem to over-react and make a special issue out of fossil record incompleteness when compared to other biologists and earth scientists, who are comfortable that their data are not perfect, and who use standard methods to explore quality and confidence issues appropriate to each study.

We thank A. Smith and A. J. McGowan for organizing the meeting, and for their invitation to participate. This work was funded in part by NERC grant NE/C518973/1 to MJB, NERC doctoral training grant NE/H525111/1 to AMD, GTL was supported by NERC grant NE/F016905/1 to A. Smith, J. Young and P. Pearson, and FGM by a University of Otago Doctoral Scholarship. We thank A. McGowan, A. Smith, M. Wills, and an anonymous reviewer for reading the manuscript at various stages, and for helpful comments.

## References


ADRAIN, J. M. & WESTROP, S. R. 2000. An empirical assessment of taxic paleobiology. *Science*, **289**, 110–112.

ALROY, J. 2002. How many named species are valid? *Proceedings of the National Academy of Sciences, USA*, **99**, 3706–3711.

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

ALROY, J., MARSHALL, C. R. *ET AL.* 2001. Effects of sampling standardization on estimates of Phanerozoic marine diversification. *Proceedings of the National Academy of Sciences, USA*, **98**, 6261–6266.

ALROY, J., ABERHAN, M. *ET AL.* 2008. Phanerozoic trends in the global diversity of marine invertebrates. *Science*, **321**, 97–100.

ANGIELCZYK, K. D. 2002. A character-based method for measuring the fit of a cladogram to the fossil record. *Systematic Biology*, **51**, 176–191.

BARNOSKY, A. D., CARRASCO, M. A. & DAVIS, E. B. 2005. The impact of the species-area relationship on estimates of paleodiversity. *PLoS Biology*, **3**, 1356–1361.

BARRAS, C. G. & TWITCHETT, R. J. 2007. Response of the marine infauna to Triassic–Jurassic environmental change: ichnological data from Southern England. *Palaeogeography, Palaeoecology, Palaeoclimatology*, **244**, 223–241.

BARRETT, P. M., BUTLER, R. J., EDWARDS, N. P. & MILNER, A. R. 2008. Pterosaur distribution in time and space: an atlas. *Zitteliana B*, **26**, 61–107.

BARRETT, P. M., MCGOWAN, A. J. & PAGE, V. 2009. Dinosaur diversity and the rock record. *Proceedings of the Royal Society, B*, **276**, 2667–2674.

BENSON, R. J., BUTLER, R. J., LINDGREN, J. & SMITH, A. S. 2010. Palaeodiversity of Mesozoic marine reptiles: mass extinctions and temporal heterogeneity in geologic megabiases affecting vertebrates. *Proceedings of the Royal Society, B*, **277**, 829–834.

BENTON, M. J. 1987. Mass extinctions among families of non-marine tetrapods: the data. *Mémoires de la Société Géologique, France*, **150**, 21–32.

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

BENTON, M. J. 1998. The quality of the fossil record of the vertebrates. *In*: DONOVAN, S. K. & PAUL, C. R. C. (eds) *The Adequacy of the Fossil Record.* Wiley, New York, 269–303.

BENTON, M. J. 2001. Finding the tree of life: matching phylogenetic trees to the fossil record through the 20th century. *Proceedings of the Royal Society, B*, **268**, 2123–2130.

BENTON, M. J. 2008*a*. Fossil quality and naming dinosaurs. *Biology Letters*, **4**, 729–732.

BENTON, M. J. 2008*b*. How to find a dinosaur, and the role of synonymy in biodiversity studies. *Paleobiology*, **34**, 516–533.

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

BENTON, M. J. & HITCHIN, R. 1997. Congruence between phylogenetic and stratigraphic data on the history of life. *Proceedings of the Royal Society, B*, **264**, 885–890.

BENTON, M. J. & STORRS, G. W. 1994. Testing the quality of the fossil record: paleontological knowledge is improving. *Geology*, **22**, 111–114.

BENTON, M. J., WILLS, M. & HITCHIN, R. 2000. Quality of the fossil record through time. *Nature*, **403**, 534–537.

BENTON, M. J., TVERDOKHLEBOV, V. P. & SURKOV, M. V. 2004. Ecosystem remodelling among vertebrates at the Permian–Triassic boundary in Russia. *Nature*, **432**, 97–100.

BENTON, M. J., ZHOU, Z., ORR, P., ZHANG, F. & KEARNS, S. 2008. The remarkable fossils from the Early Cretaceous Jehol Biota of China and how they have changed our knowledge of Mesozoic life. *Proceedings of the Geologists' Association*, **119**, 209–228.

BERNARD, E. L., RUTA, M., TARVER, J. E. & BENTON, M. J. 2010. The fossil record of early tetrapods: worker effort and the end-Permian mass extinction. *Acta Palaeontologica Polonica*, **55**, 213–228.

BRUSATTE, S. L., BENTON, M. J., DESOJO, J. B. & LANGER, M. C. 2010. The higher-level phylogeny of Archosauria (Tetrapoda: Diapsida). *Journal of Systematic Palaeontology*, **8**, 3–47.

BRUSATTE, S. L., BENTON, M. J., LLOYD, G. T., RUTA, M. & WANG, S. C. 2011. Macroevolutionary patterns in the evolutionary radiation of archosaurs (Tetrapoda: Diapsida). *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, **101**, 285–299.

BURNHAM, K. P. & ANDERSON, D. R. 2002. *Model Selection and Multimodel Inference: A Practical Information-theoretic Approach.* Springer-Verlag, New York.

BUSH, A. M., MARKEY, M. J. & MARSHALL, C. R. 2004. Removing bias from diversity curves: the effects of spatially organized biodiversity on sampling-standardization. *Paleobiology*, **30**, 666–686.

BUTLER, R. J., BARRETT, P. M., NOWBATH, S. & UPCHURCH, P. 2009. Estimating the effects of the rock record on pterosaur diversity patterns: implications for hypotheses of bird/pterosaur competitive replacement. *Paleobiology*, **35**, 432–446.

BUTLER, R. J., BENSON, R. J., CARRANO, W. T., MANNION, P. D. & UPCHURCH, P. 2011. Sea level, dinosaur diversity and sampling biases: investigating the 'common cause' hypothesis in the terrestrial realm. *Proceedings of the Royal Society, B*, **278**, 1165–1170.

CARPENTER, K. 1997. Morrison formation. *In*: CURRIE, P. J. & PADIAN, K. (eds) *Encyclopedia of Dinosaurs*. University of California Press, Berkeley, California, 5.

CHIAPPE, L. M. & WITMER, L. M. (eds) 2002. *Mesozoic Birds: Above the Heads of Dinosaurs*. University of California Press, Berkeley, California.

COX, B. M., SUMBLER, M. G. & IVIMEY-COOK, H. C. 1999. *A formational framework for the Lower Jurassic of England and Wales (onshore area)*. British Geological Survey Research Report **RR/99/01**, 1–28.

CRAMPTON, J. S., BEU, A. G., COOPER, R. A., JONES, C. M., MARSHALL, B. & MAXWELL, P. A. 2003. Estimating the rock volume bias in palaeodiversity studies. *Science*, **301**, 358–360.

DARWIN, C. 1859. *On the Origin of Species by Means of Natural Selection*. John Murray, London.

DODSON, P. 1990. Counting dinosaurs, how many kinds were there? *Proceedings of the National Academy of Sciences, USA*, **87**, 7608–7612.

DONOVAN, S. & PAUL, C. R. C. (eds) 1998. *The Adequacy of the Fossil Record*. Wiley, New York.

DUNHILL, A. M. 2011. Using remote sensing and a GIS to quantify rock exposure in England and Wales: implications for paleodiversity studies. *Geology*, **39**, 111–114.

DUNHILL, A. M. in press. Problems with using rock outcrop area as a paleontological sampling proxy: comparing rock outcrop and exposure area in California, New York State, Australia, and England and Wales. *Paleobiology*, **38**, in press.

DYKE, G. J., MCGOWAN, A. J., NUDDS, R. L. & SMITH, D. 2009. The shape of pterosaur evolution: evidence from the fossil record. *Journal of Evolutionary Biology*, **22**, 890–898.

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

FARA, E. 2002. Sea-level variations and the quality of the continental fossil record. *Journal of the Geological Society, London*, **159**, 489–491.

FARA, E. & BENTON, M. J. 2000. The fossil record of Cretaceous tetrapods. *Palaios*, **15**, 161–165.

FLESSA, K. W. & JABLONSKI, D. 1983. Extinction is here to stay. *Paleobiology*, **9**, 315–321.

FOOTE, M. & RAUP, D. M. 1996. Fossil preservation and the stratigraphic ranges of taxa. *Paleobiology*, **22**, 121–140.

FOOTE, M. & SEPKOSKI, J. J. 1999. Absolute measures of the completeness of the fossil record. *Nature*, **398**, 415–417.

FOREY, P., FORTEY, R. A., KENRICK, P. & SMITH, A. B. 2004. Taxonomy and fossils: a critical appraisal. *Philosophical Transactions of the Royal Society, B*, **359**, 639–653.

FOUNTAINE, T., BENTON, M. J., DYKE, G. J. & NUDDS, R. L. 2005. The quality of the fossil record of Mesozoic birds. *Proceedings of the Royal Society, B*, **272**, 289–294.

FRÖBISCH, J. 2008. Global taxonomic diversity of anomodonts (Tetrapoda, Therapsida) and the terrestrial rock record across the Permo-Triassic boundary. *PLoS One*, **3**, e3733.

FRÖBISCH, J. 2009. Composition and similarity of global anomodont-bearing faunas. *Earth-Science Reviews*, **95**, 119–157.

GUISAN, A. & THUILLER, W. 2005. Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, **8**, 993–1009.

GUISAN, A. & ZIMMERMANN, N. E. 2000. Predictive habitat distribution models in ecology. *Ecological Modelling*, **135**, 147–186.

HARNIK, P. G. 2009. Unveiling rare diversity by integrating museum, literature, and field data. *Paleobiology*, **35**, 190–208.

HAUBOLD, H. 1990. Dinosaurs and fluctuating sea levels during the Mesozoic. *Historical Biology*, **4**, 75–106.

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

HUELSENBECK, J. P. 1994. Comparing the stratigraphic record to estimates of phylogeny. *Paleobiology*, **20**, 470–483.

KALMAR, A. & CURRIE, D. J. 2010. The completeness of the continental fossil record and its impact on patterns of diversification. *Paleobiology*, **36**, 51–60.

KIESSLING, W. 2005. Habitat effects and sampling bias on Phanerozoic reef distribution. *Facies*, **51**, 24–32.

LANE, A., JANIS, C. M. & SEPKOSKI, J. J. JR. 2005. Estimating paleodiversities: a test of the taxic and phylogenetic methods. *Paleobiology*, **31**, 21–34.

LITTLE, C. T. S. 1996. The Pliensbachian–Toarcian (Lower Jurassic) extinction event. *Geological Society of America, Special Paper*, **307**, 505–512.

LLOYD, G. T., DAVIS, K. E. ET AL. 2008. Dinosaurs and the Cretaceous Terrestrial Revolution. *Proceedings of the Royal Society, B*, **275**, 2483–2490.

LLOYD, G. T., SMITH, A. B. & YOUNG, J. R. 2011. Quantifying the deep-sea rock and fossil record bias using coccolithophores. *In*: MCGOWAN, A. J. & SMITH, A. B. (eds) *Comparing the Geological and Fossil Records: Implications for Biodiversity Studies*. Geological Society, London, Special Publications, **358**, 167–177.

MANDER, L., TWITCHETT, R. J. & BENTON, M. J. 2008. Palaeoecology of the Late Triassic extinction event in the SW UK. *Journal of the Geological Society*, **165**, 319–332.

MANNION, P. D. & UPCHURCH, P. 2010. Completeness metrics and the quality of the sauropodomorph fossil record through geological and historical time. *Paleobiology*, **36**, 283–302.

MANNION, P. D., UPCHURCH, P., CARRANO, W. T. & BARRETT, P. M. 2011. Testing the effect of the rock record on diversity: a multidisciplinary approach to elucidating the generic richness of sauropodomorph dinosaurs through time. *Biological Reviews*, **86**, 157–181.

MARKWICK, P. J. 1998. Fossil crocodilians as indicators of Late Cretaceous and Cenozoic climates: implications for using palaeontological data in reconstructing palaeoclimate. *Palaeogeography, Palaeoecology, Palaeoclimatology*, **137**, 205–271.

MARSHALL, C. R. 1990. Confidence intervals on stratigraphic ranges. *Paleobiology*, **16**, 1–10.

MARSHALL, C. R. 1997. Confidence intervals on stratigraphic ranges with non-random distributions of fossil horizons. *Paleobiology*, **23**, 165–173.

MARSHALL, C. R. 2010. Marine biodiversity dynamics over deep time. *Science*, **329**, 1156–1157.

MARX, F. G. 2009. Marine mammals through time: when less is more in studying palaeodiversity. *Proceedings of the Royal Society, B*, **276**, 887–892.

MARX, F. G. 2010. The more the merrier? A large cladistic analysis of mysticetes, and comments on the transition from teeth to baleen. *Journal of Mammalian Evolution*, **28**, 77–200.

MARX, F. G. & UHEN, M. D. 2010. Climate, critters, and cetaceans: Cenozoic drivers of the evolution of modern whales. *Science*, **327**, 993–996.

MAXWELL, W. D. & BENTON, M. J. 1990. Historical tests of the absolute completeness of the fossil record of tetrapods. *Paleobiology*, **16**, 322–335.

MAY, R. M. 1990. How many species? *Philosophical Transactions of the Royal Society of London, Series B*, **330**, 293–304.

MAYHEW, P. J., JENKINS, G. B. & BENTON, T. G. 2007. A long-term association between global temperature and biodiversity, origination and extinction in the fossil record. *Proceedings of the Royal Society B*, **275**, 47–53.

MCGOWAN, A. & SMITH, A. 2008. Are global Phanerozoic 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. 1990. Classifying and analysing evolutionary trends. *In*: MCNAMARA, K. J. (ed.) *Evolutionary Trends*. Belhaven Press, London, 28–58.

MODESTO, S., SUES, H.-D. & DAMIANI, R. 2001. A new Triassic procolophonoid reptile and its implications for procolophonoid survivorship during the Permo-Triassic extinction event. *Proceedings of the Royal Society, B*, **268**, 2047–2052.

NICHOL, D. 1977. The number of living animals likely to be fossilized. *Florida Scientist*, **40**, 135–139.

NORELL, M. 1992. Taxic origin and temporal diversity: the effect of phylogeny. *In*: NOVACEK, M. & WHEELER, Q. (eds) *Extinction and Phylogeny*. Columbia University Press, New York, 89–118.

NORELL, M. A. & NOVACEK, M. J. 1992. The fossil record and evolution: comparing cladistic and paleontologic evidence for vertebrate history. *Science*, **255**, 1690–1693.

PAUL, C. R. C. 1982. The adequacy of the fossil record. *In*: JOYSEY, K. A. & FRIDAY, A. E. (eds) *Problems of Phylogenetic Reconstruction*. Academic Press, London, 75–117.

PAUL, C. R. C. 1998. Adequacy, completeness and the fossil record. *In*: DONOVAN, S. K. & PAUL, C. R. C. (eds) *The Adequacy of the Fossil Record*. Wiley, New York, 1–22.

PETERS, S. E. 2005. Geologic constraints on the macroevolutionary history of marine animals. *Proceedings of the National Academy of Sciences, USA*, **102**, 12 326–12 331.

PETERS, S. E. 2006. Macrostratigraphy of North America. *Journal of Geology*, **114**, 391–412.

PETERS, S. E. 2008. Environmental determinants of extinction selectivity. *Nature*, **454**, 626–629.

PETERS, S. E. & FOOTE, M. 2001. Biodiversity in the Phanerozoic: a reinterpretation. *Paleobiology*, **27**, 583–601.

PETERS, S. E. & FOOTE, M. 2002. Determinants of extinction in the fossil record. *Nature*, **416**, 420–424.

PETERS, S. E. & HEIM, N. A. 2010. The geological completeness of paleontological sampling in North America. *Paleobiology*, **36**, 61–79.

PRESTON, F. W. 1948. The commonness, and rarity, of species. *Ecology*, **29**, 254–283.

PURNELL, M. A. & DONOGHUE, P. C. J. 2005. Between death and data: biases in interpretation of the fossil record of conodonts. *Special Papers in Palaeontology*, **73**, 7–25.

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

RAUP, D. M. 1975. Taxonomic diversity estimates under rarefaction. *Paleobiology*, **1**, 333–342.

RAUP, D. M. 1976. Species diversity in the Phanerozoic: a tabulation. *Paleobiology*, **2**, 279–288.

RAUP, D. M. 1977. Systematists follow the fossils. *Paleobiology*, **3**, 328–329.

RONOV, A. B. 1978. The earth's sedimentary shell. *International Geology Review*, **24**, 1313–1363.

RONOV, A. B. 1994. Phanerozoic transgressions and regressions on the continents: a quantitative approach based on areas flooded by the sea and areas of marine and continental deposition. *American Journal of Science*, **294**, 777–801.

RUSSELL, D. A. 1995. China and the lost worlds of the dinosaurian era. *Historical Biology*, **10**, 3–12.

SAHNEY, S. & BENTON, M. J. 2008. Recovery from the most profound mass extinction of all time. *Proceedings of the Royal Society, B*, **275**, 759–765.

SAHNEY, S., BENTON, M. J. & FERRY, P. A. 2010. Links between global taxonomic diversity, ecological diversity, and the expansion of vertebrates on land. *Biology Letters*, **6**, 544–547.

SEPKOSKI, J. J. JR. 1993. Ten years in the library: how changes in taxonomic data bases affect perception of macroevolutionary pattern. *Paleobiology*, **19**, 43–51.

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

SEPKOSKI, J. J. JR, BAMBACH, R. K., RAUP, D. M. & VALENTINE, J. W. 1981. Phanerozoic marine diversity and the fossil record. *Nature*, **293**, 435–437.

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

SIMMS, M. J., CHIDLAW, N., MORTON, N. & PAGE, K. N. 2004. *British Lower Jurassic Stratigraphy*. Geological Conservation Review Series, **30**. JNCC, London.

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

SMITH, A. B. 2007*a*. Marine diversity through the Phanerozoic: problems and prospects. *Journal of the Geological Society, London*, **164**, 731–745.

SMITH, A. B. 2007*b*. Intrinsic versus extrinsic biases in the fossil record: contrasting the fossil record of echinoids in the Triassic and early Jurassic using sampling data, phylogenetic analysis, and molecular clocks. *Paleobiology*, **33**, 310–323.

SMITH, A. B. & MCGOWAN, A. J. 2007. The shape of the Phanerozoic marine palaeodiversity curve: how much can be predicted from the sedimentary rock record of Western Europe? *Palaeontology*, **50**, 1–10.

SMITH, A. B. & MCGOWAN, A. J. 2008. Temporal patterns of barren intervals in the Phanerozoic. *Paleobiology*, **34**, 155–161.

STANLEY, S. M. 2007. An analysis of the history of marine animal diversity. *Paleobiology*, **33** (Suppl. 4), 1–55.

SUGIURA, N. 1978. Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics, Series A*, **7**, 13–26.

TARVER, J. E., BRADDY, S. J. & BENTON, M. J. 2007. The effects of sampling bias on Palaeozoic faunas and implications for macroevolutionary studies. *Palaeontology*, **50**, 177–184.

TARVER, J. E., DONOGHUE, P. C. J. & BENTON, M. J. 2011. Is evolutionary history repeatedly rewritten in light of new fossil discoveries? *Proceedings of the Royal Society, B*, **278**, 599–604.

UPCHURCH, P. & BARRETT, P. M. 2005. Phylogenetic and taxic perspectives on sauropod diversity. *In*: CURRY-ROGERS, K. A. & WILSON, J. A. (eds) *The Sauropods: Evolution and Paleobiology*. University of California Press, Berkeley, California, 104–124.

UPCHURCH, P., MANNION, P. D., BENSON, R. B. J., BUTLER, R. J. & CARRANO, M. T. 2011. Geological and anthropogenic controls on the sampling of the terrestrial fossil record: a case study from the Dinosauria. *In*: SMITH, A. B. & MCGOWAN, A. J. (eds) *Comparing the Geological and Fossil Records: Implications for Biodiversity Studies.* Geological Society, London, Special Publications, **358**, 209–240.

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

VERMEIJ, G. J. & GROSBERG, R. K. 2010. The great divergence: when did diversity on land exceed that in the sea? *Integrative and Comparative Biology*, **50**, 675–682.

WALL, P., IVANY, L. & WILKINSON, B. 2009. Revisiting Raup: exploring the influence of outcrop area on diversity in light of modern sample-standardization techniques. *Paleobiology*, **35**, 146–167.

WANG, S. C. & DODSON, P. 2006. Estimating the diversity of dinosaurs. *Proceedings of the National Academy of Sciences, USA*, **103**, 13 601–13 605.

WEISHAMPEL, D. B. 1996. Fossils, phylogeny, and discovery: a cladistic study of the history of tree topologies and ghost lineage durations. *Journal of Vertebrate Paleontology*, **16**, 191–197.

WICKSTRÖM, L. M. & DONOGHUE, P. C. J. 2005. Cladograms, phylogenies and the veracity of the conodont fossil record. *Special Papers in Palaeontology*, **73**, 185–218.

WIGNALL, P. B. & BENTON, M. J. 1999. Lazarus taxa and fossil abundance at times of biotic crisis. *Journal of the Geological Society, London*, **156**, 453–456.

WILLIAMS, H. S. 1901. The discrimination of time-values in geology. *Journal of Geology*, **9**, 570–585.

WILLS, M. A. 1999. Congruence between phylogeny and stratigraphy: randomization tests. *Systematic Biology*, **48**, 559–580.

WILLS, M. A. 2007. Fossil ghost ranges are most common in some of the oldest and some of the youngest strata. *Proceedings of the Royal Society, B*, **274**, 2421–2427.