Palaeogeography, Palaeoclimatology, Palaeoecology 372 (2013) 18–41

Contents lists available at SciVerse ScienceDirect

Palaeogeography, Palaeoclimatology, Palaeoecology

journal homepage: www.elsevier.com/locate/palaeo

The first half of tetrapod evolution, sampling proxies, and fossil record quality

Michael J. Benton [a,*], Marcello Ruta [a,b], Alexander M. Dunhill [a], Manabu Sakamoto [a]

[a] School of Earth Sciences, University of Bristol, Bristol, BS8 1RJ, UK
[b] School of Life Sciences, University of Lincoln, Lincoln, LN6 7TS, UK

ARTICLE INFO

Article history:
Received 17 March 2012
Received in revised form 9 September 2012
Accepted 10 September 2012
Available online 26 September 2012

Keywords:
Tetrapoda
Amniota
Diversification
Devonian
Carboniferous
Permian
Triassic
Jurassic

ABSTRACT

The first half of tetrapod evolution witnessed substantial diversification of the clade and several major turnovers and mass extinctions. In the time since their origin, more than 380 Myr ago, to the beginning of the Middle Jurassic 175 Myr ago, tetrapods apparently diversified fitfully, reaching their highest level in the Middle Permian, and showing major diversity declines in the late Moscovian, Early Permian, Wordian, lower Wuchiapingian, end-Permian, lower Anisian, lower Ladinian, Late Triassic (lower Norian to upper Rhaetian), end-Triassic, and Early Jurassic (upper Sinemurian, lower Pliensbachian). Of these diversity drops, only the end-Permian and end-Triassic correspond to recognised mass extinctions, and the late Moscovian and early Norian drops to other previously identified environmental crises. The remainder could be real extinction or turnover events, or partially artefacts of biased sampling. There are strong correlations between formation counts and tetrapod palaeodiversity, suggesting a sampling component in the raw data, but the covariation is not uniform through the whole time span, being poor from Devonian to Middle Permian, and better from Late Permian to Early Jurassic. There is limited evidence for covariation between the tetrapod palaeodiversity time series and other putative sampling metrics, such as specimen completeness, numbers of publications, map areas, gap-bounded sedimentary units, rock volumes, formations, and fossil collections. Modelling by multiple correlations shows that formation count is generally the best explanatory model, either on its own, or combined with other 'sampling' time series. However, it is not clear that formation count is independent of the palaeodiversity time series, because rises and falls in both signals could reflect variations in original diversity or in preservation or in sampling.

1. Introduction

The limbed vertebrates, or tetrapods, colonised land by 380 Myr ago, and diversified in several evolutionary stages, from early 'amphibian'-grade tetrapods (including stem-tetrapods, stem-amphibians, and stem-amniotes), generally associated closely with water and/or damp habitats during the Late Devonian and Carboniferous, to crown amniotes (= 'reptiles', birds, and mammals) from the Late Carboniferous onwards. Further major turnovers during the first half (some 215 Myr) of tetrapod history include: (1) the demise of many early tetrapod groups at the time of the rapid decline of the North American–European rain forests near the end of the Carboniferous, some 305 Myr ago, and their replacement by dryland-adapted amniotes (Sahney et al., 2010); (2) the diversification of terrestrial vertebrate ecosystems, with the origin of herbivory in the Permian (Sues and Reisz, 1998); (3) the crash of diverse Middle and Late Permian therapsid–pareiasaur ecosystems during the end-Permian mass extinction (Benton et al., 2004); (4) the rise and diversification of crurotarsan- and dinosauromorph-dominated ecosystems in the Triassic (Benton, 1983; Brusatte et al., 2008a, 2010, 2011); (5) the end-Triassic mass extinction; and (6) the further radiation of dinosaurs in the Early and Middle Jurassic (Brusatte et al., 2008b). This first half of the history of the clade Tetrapoda, then witnessed a number of striking diversifications and extinction events, and the foundations of most modern tetrapod subclades—Lissamphibia, Testudinata, Crocodylomorpha, Lepidosauria, Dinosauria (including birds), and Mammalia. However, the quality of the fossil data is often questioned.

As many recent publications have highlighted (e.g. Smith, 2007a; McGowan and Smith, 2011), the fossil record is of uncertain quality and biased by incomplete sampling. This incompleteness results from three main factors, namely variable rock volume, variable rock availability, and variable human effort. Together, these sources of error mean that current knowledge underestimates the fossil record (= all fossils preserved in the rocks) and the fossil record underestimates past life (= all species that ever lived). As a result, it has long been debated whether the available data give a good enough signal to document the history of life more or less reliably (e.g. Sepkoski et al., 1981; Benton, 1995; Benton et al., 2000; Peters, 2005; Stanley, 2007) or not (e.g. Raup, 1972; Alroy et al., 2001; Peters and Foote, 2002; Smith, 2007a; Alroy et al., 2008; Alroy, 2010a).

A repeated finding is that palaeodiversity curves and sampling proxies generally covary (McGowan and Smith, 2011). A *palaeodiversity curve* is a time series of taxon counts (typically, species, genera, or

* Corresponding author. Tel.: +44 117 9545433; fax: +44 117 9253585.
E-mail address: mike.benton@bristol.ac.uk (M.J. Benton).

0031-0182/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.palaeo.2012.09.005

families) plotted against geological time bins, and assessed at a global, regional, or local scale. A *sampling proxy* is a metric that represents 'collecting effort' in some way, and it should have two properties (Smith, 2007a; Benton et al., 2011): (1) it should represent some or all of the geological and human factors that can introduce error into interpretations of data from the fossil record; and (2) it should be independent of the signal it seeks to correct, namely the documented fossil record.

Three hypotheses can explain the covariation of palaeodiversity curves and sampling proxies, namely: (1) the *bias hypothesis*, that the palaeodiversity curve is largely driven by geological and human biases (Raup, 1972; Smith, 2001, 2007a; Alroy, 2010a); (2) the *common cause hypothesis*, that both signals are dependent on a third factor, such as sea level, topography, temperature, or isotopic signals (Sepkoski et al., 1981; Peters, 2005; Peters and Heim, 2010; Hannisdal and Peters, 2011); and (3) the *redundancy hypothesis*, that in some cases the fossil record and sampling proxy signals may be partially redundant with each other (Benton et al., 2011). None of these explanations is exclusive, and in reality some component of all three hypotheses doubtless pertains in each case investigated (Smith, 2007a; Benton et al., 2011; Hannisdal and Peters, 2011).

Many recent studies have suggested that the records of various vertebrate clades are tightly correlated with sampling (e.g. Fröbisch, 2008; Lloyd et al., 2008; Barrett et al., 2009; Butler et al., 2009; Benson et al., 2010; Butler et al., 2011; Mannion et al., 2011; Upchurch et al., 2011; Benson and Mannion, 2012; Lloyd, 2012). If correct, these studies confirm a long-held assumption that terrestrial fossil records in general, and those of vertebrates in particular, are poor, perhaps dominated by so-called 'megabiases' (Barrett et al., 2009; Benson et al., 2010). In many of the recent studies, the bulk of the palaeodiversity curve appears to correlate closely with the sampling proxy, and the residuals (i.e. palaeodiversity signal minus sampling signal) often provide a very different picture of macroevolution, in which the 'corrected' palaeodiversity curve shows bursts of diversification and extinctions in different places from those in the 'uncorrected' empirical curve. It would be interesting to know in each case whether either the empirical or the 'corrected' curve is closer to the true diversity signal (Benton, 2009).

The use of formation counts as a sampling proxy has been core to many recent studies. Despite criticism (Crampton et al., 2003; Smith, 2007a; Benton et al., 2011), this proxy represents a rather simple tool because fossil counts and formation counts may be extracted from online databases such as the Paleobiology Database (PaleoDB) or such data may be collected in combined fossil and stratigraphic distribution data searches, as in this paper. Evidence for the usefulness or validity of formation counts as a sampling metric is that (1) the number of formations often correlates strongly with other sampling proxies such as number of rock sections or estimates of total rock volume (Peters and Foote, 2001; Upchurch et al., 2011) and (2) at least in the case of dinosaurs, the cumulative sum of dinosaur-bearing formations through research time correlates strongly with the cumulative number of valid new dinosaurian species named (Benton, 2008a). This latter point has not been used before as evidence that the number of dinosaur-bearing formations is a direct metric of sampling, but it might appear compelling (see Section 6.2).

The aim of this paper is to present the first outputs of investigations of a new early tetrapod database, and to explore by steps the macroevolutionary signals, if any, that might be extracted from current knowledge of the tetrapod fossil record. In particular, we wish to determine whether apparent rises in palaeodiversity document diversifications, and whether drops correspond to extinction events—or whether such rises and falls might be better accounted for as resulting in part or entirely from uneven sampling.

2. The database

The early tetrapod database (ETD), compiled by M.J.B., spans from Middle Devonian (late Givetian) to terminal Early Jurassic (late Toarcian), approximately 390 to 175 Myr ago. This stratigraphic range was chosen to cover at least the first half of the history of tetrapods as a whole, and to encompass substantial amounts of the history of 'amphibian'-grade tetrapods and amniotes. Furthermore, among amniotes, the ETD covers Parareptilia, basal Eureptilia, basal Diapsida, and basal Synapsida. Finally, the selected time span encompasses episodes of major climatic and vegetational change in the Devonian and Carboniferous, the end-Permian and end-Triassic mass extinctions, and the initial rise of both dinosaurs and a cross-section of 'modern' tetrapod groups (e.g. lissamphibians, turtles, lepidosaurs, crocodylomorphs, and mammals) in the Late Triassic and Early Jurassic. Insofar as the analyses are restricted to that time period, the data set encompasses a single monophyletic group or clade, with an upper temporal boundary placed at 175 Myr ago.

The ETD was compiled by scanning all appropriate journals up to the end of February, 2012, and documenting all cases of new taxa and revisions to taxa named in the past. Where clades have been revised in detail at the alpha taxonomic level (e.g. Kammerer et al., 2011), the exact taxon lists and stratigraphic ranges are taken from that recent synoptic work. More commonly, such revisions have not been done, and information on taxon validity is summarised to represent as far as possible published current opinion.

Stratigraphic ranges were determined by compiling separate lists of named geological formations for each time period. Searches were made for the latest geological accounts of each of those named formations to check for alternative names and/or definitions, environmental interpretations (primarily continental vs. marine), and for independent dating evidence. The age range then is identical for each genus reported from any geological formation, and revisions to the estimated ages of any particular geological formation are disseminated through the database for uniformity.

In total, we identify 542 *geological formations*, from Middle Devonian to Lower Jurassic, that have yielded tetrapod fossils. These include 459 predominantly continental formations and 83 predominantly marine formations, and these are distributed variably during each geological period (Table 1). Further, we were able to assign 409 of the 542 geological formations (75.5%) to a single time bin, 96 (17.7%) to two time bins, and 37 (6.8%) to three time bins (Table 1). The multiple-bin geological formations arise either because they are extensive and thus represent a long span of geological time or because their dating is uncertain. An example is the Burgersdorp Formation of the Karoo Basin, South Africa (Johnson and Hiller, 1990). This unit overlies the Katberg Formation, and its base is partially laterally equivalent to the top of that unit, and is overlain, apparently conformably, by the Molteno Formation. Its thickness varies from 400–1000 m across the basin, and its overall age span is currently estimated as late Olenekian to late Anisian (Abdala et al., 2005). The Burgersdorp Formation is approximately equivalent to the *Cynognathus* Assemblage Zone, which was traditionally dated as 'Early Triassic' or 'Early to Middle Triassic'. This biozone is now divided into three subzones, A, B, and C, each with its own unique tetrapod fauna (Abdala et al., 2005), and these are dated respectively as late Olenekian, early Anisian, and late Anisian. If a fossil is labelled simply as '*Cynognathus* Assemblage Zone,' then its age spans three time bins,

Table 1
Distribution of the ETD geological formations that have produced tetrapod fossils, summarised by geological period, and indicating marine and non-marine formations, and those that are reported from one, two, or three time bins (representing either long-lasting, or poorly dated, geological units).

Period	Marine	Non-marine	Total	1-bin	2-bin	3-bin
Devonian	0	16	16	16	0	0
Carboniferous	11	80	91	85	5	1
Permian	4	128	132	84	32	16
Triassic	57	194	251	205	36	10
Early Jurassic	11	41	52	19	23	10
Totals	83	459	542	409	96	37

either in reality or by default; if the fossil is identified more precisely as belonging to one or two of the divisions A, B, and C, then its age is constrained.

We use the *standard international marine geological time scale* (Gradstein et al., 2004; Ogg et al., 2008) as the basis for all assignments of fossil genera to time bins. These standard time scales have been substantially revised, especially by subsequent studies of the redbed successions, many based on new biostratigraphic, magnetostratigraphic, and radiometric dating work on the Permian (e.g. Rubidge, 2005; Lozovskiy et al., 2009; Taylor et al., 2009; Benton et al., 2012; Benton, 2012) and Triassic (e.g. Muttoni et al., 2004; Hounslow and Muttoni, 2010; Muttoni et al., 2010; Ramezani et al., 2011). We use these latest revisions in age determinations.

The standard time unit used in the ETD is the *stratigraphic stage*. Some are short and are kept undivided; others are longer, and so are divided into two (early, late) or three (early, middle, late) subdivisions (Table 2). In all analyses, stage names are given their standard abbreviations (Gradstein et al., 2004). Stages are listed in stratigraphic order, starting with the earliest (see Table 2 for durations). Seven stages are left undivided: Kasimovian, Gzelian, Roadian, Wordian, Changhsingian, Induan, and Hettangian. Eighteen stages are divided into two halves: Givetian, Serpukhovian, Bashkirian, Moscovian, Asselian, Sakmarian, Artinskian, Kungurian, Capitanian, Wuchiapingian, Olenekian, Anisian, Ladinian, Carnian, Rhaetian, Sinemurian, Pliensbachian, and Toarcian. Finally, five stages are divided into three: Frasnian, Famennian, Tournaisian, Viséan, and Norian. The stages are those used in the international time scale (Gradstein et al., 2004; Ogg et al., 2008). Unfortunately, there is no internationally agreed system of substages for the Devonian, Carboniferous, or Permian, but only parallel schemes for different continents (Gradstein et al., 2004). The standard substages for the Triassic and Jurassic were used.

We used two systems of time bins for different elements of the ETD study. The initial analyses used all 58 substage-level time bin divisions (Table 2), which together span 210.8 Myr, with a mean duration of 3.634 Myr and a standard deviation of 1.348 Myr. However, most studies hitherto have considered diversity counts by stratigraphic stages rather than substages. Therefore, for comparative purposes, we also ran all analyses after grouping the data into stage-level time bins. This gave rise to exactly half the number of time bins, 29, with a mean duration of 7.269 Myr and a standard deviation of 4.620 Myr.

The database comprises (February 2012) 1388 valid genera, of which 391 are classed as 'Amphibia' (namely stem-tetrapods, Temnospondyli, Lepospondyli, Reptiliomorpha, Lissamphibia) and 997 are classed as Amniota (namely Parareptilia, Eureptilia [i.e. Diapsida plus basal forms], Synapsida). These totals comprise taxa that are currently regarded as valid according to recent taxonomic revisions, and the many hundreds of synonymous genera are sunk into those valid taxa; decisions about taxon content and validity can only reflect current published opinion, and some invalid taxa, including synonyms, doubtless remain in tetrapod families that have not undergone recent cladistic revision. A further 42 'amphibian' genera and 127 amniote genera are classed in the ETD as nomina nuda and nomina dubia, and they are not included in any analyses. The 1388 individual genera, when listed as occurrences by substage time bins, generate a total of 2453 records (798 'amphibian' and 1755 amniote occurrence records), based on the occurrences of named species in different stratigraphic substages.

These 2453 records require further investigation to determine the true balance between singletons and long-lived genera. Among the 1388 genera in the ETD are 1040 (74.9%) singletons, i.e. genera represented by one species occurring in one stratigraphic formation (Table 3). Most of the remaining taxa (2453 occurrences minus 1040 singleton genera; 1413 occurrences) are recorded in more than one time bin because the dating of the host geological formation itself includes multiple time bins. In turn, this dating reflects either the amount of time actually implied by the extension of the formation, or uncertainty over dating, as noted above. In rare cases, multiple species in a genus occur in different geological formations and so the generic range really does span more than one time bin. Singletons (Table 3) are more common among amniote genera (78.2%) than among 'amphibian' genera (66.5%), and this could be either a real phenomenon or the result of different taxonomic practice both in terms of identifying newly discovered specimens as new species and in terms of synonymisation of taxa in taxonomic revision.

In addition, for each taxon we recorded the available material in broad terms as a means of assessing approximate specimen quality; this term was deliberately generalised to avoid difficulties

Table 2
The stratigraphic stages, with abbreviations and durations, as used in the early tetrapod database (ETD). Information is taken from Gradstein et al. (2004), and revisions to Triassic stage dating from Muttoni et al. (2004, 2010). Abbreviations: l, lower; m, middle; u, upper.

Subsystem	Substage	Acronym	Duration (Myr)
Middle Devonian	Givetian	GIV-l	3.2
		GIV-u	3.2
Upper Devonian	Frasnian	FRS-l	3.5
		FRS-m	3.5
		FRS-u	3.5
	Famennian	FAM-l	5.1
		FAM-m	5.1
		FAM-u	5.1
Mississippian	Tournaisian	TOU-l	4.6
		TOU-m	4.6
		TOU-u	4.6
	Viséan	VIS-l	6.3
		VIS-m	6.3
		VIS-u	6.3
	Serpukhovian	SPK-l	4.1
		SPK-u	4.1
Pennsylvanian	Bashkirian	BSH-l	3.2
		BSH-u	3.2
	Moscovian	MOS-l	2.6
		MOS-u	2.6
	Kasimovian	KAS	2.6
	Gzelian	GZE	4.9
Cisuralian	Asselian	ASS-l	2.8
		ASS-u	2.8
	Sakmarian	SAK-l	5.1
		SAK-u	5.1
	Artinskian	ART-l	4.4
		ART-u	4.4
	Kungurian	KUN-l	2.5
		KUN-u	2.5
Guadalupian	Roadian	ROA	2.5
	Wordian	WOR	2.5
	Capitanian	CAP-l	2.5
		CAP-u	2.5
Lopingian	Wuchiapingian	WUC-l	3
		WUC-u	3
	Changhsingian	CHX	2
Lower Triassic	Induan	IND	1.5
	Olenekian	OLE-l	1.5
		OLE-u	1.5
Middle Triassic	Anisian	ANS-l	3.5
		ANS-u	3.5
	Ladinian	LAD-l	1.5
		LAD-u	1.5
Upper Triassic	Carnian	CRN-l	3.5
		CRN-u	3.5
	Norian	NOR-l	6.5
		NOR-m	6.5
		NOR-u	6.5
	Rhaetian	RHT-l	3
		RHT-u	3
Lower Jurassic	Hettangian	HET	3
	Sinemurian	SIN-l	3.5
		SIN-u	3.5
	Pliensbachian	PLB-l	3.3
		PLB-u	3.3
	Toarcian	TOA-l	3.7
		TOA-u	3.7
Total			210.8

Table 3
Numbers and proportions of genera in the Early Tetrapods Database (ETD) that are represented by more than one species or by specimens from more than one geological formation, but that are all dated to the same time bin (>1 record), by multiple records or species in more than one time bin (>1 time bin), and by singletons (total genera minus genera with >1 record). Occurrences are the number of records in total based on records of species per stratigraphic time bin (Table 2).

Clade	Genera	>1 record	>1 time bin	Singletons	Occurrences
'Amphibia'	391	131 (33.5%)	64 (16.4%)	260 (66.5%)	798
Amniota	997	217 (21.8%)	101 (10.1%)	780 (78.2%)	1755
Tetrapoda	1388	348 (25.1%)	165 (11.9%)	1040 (74.9%)	2453

of prevision. The *specimen completeness categories* (Table 4) are also determined as classes rather than exact measures, and these completeness classes are of very different scopes—in summary the classes include: scrap, skull, skeleton, and multiple skeletons. At one end of the scale, a genus might be represented by only a jaw fragment or femur, and at the other end by dozens of complete skeletons. However, in terms of confidence of identification and discrimination from other taxa, a single skull or single skeleton may be just as useful taxonomically as 100 skeletons. The first category, 'scrap', includes genera based on a single specimen, as noted, but also genera based on numerous incomplete remains that do not include anything close to a complete skull or any articulated remains. The other categories include single and multiple skulls and skeletons, and could be subdivided more finely than we did; however, at this stage we have not sought to determine, for the more completely represented taxa, whether they are known from two skeletons or 46 skeletons. The present scheme is sufficient to determine, for example, whether a taxon is represented by incomplete or substantial material, and whether any particular geological formation or time bin has yielded only disassociated and incomplete remains, or numerous rather complete skulls and skeletons.

The ETD database will be available from Dryad one year after publication: http://dx.doi.org/10.5061/dryad.44b50.

3. Sampling proxies

As noted above, the amount of bias or error in a palaeodiversity record can be assessed by means of sampling proxies or sampling metrics. A sampling proxy should document some elements of the sampling bias arising from variable rock volume, variable rock accessibility, and variable human effort affecting different fossil taxa and different time bins. The ideal sampling proxy ought to document all three contributors to error, but in reality most sampling proxies document aspects of only one of the three confounding variables, and then perhaps only incompletely. In recent studies, sampling proxies such as formation counts (Peters and Foote, 2001, 2002; Fröbisch, 2008; Barrett et al., 2009; Butler et al., 2009; Benson et al., 2010; Benson and Butler, 2011; Butler et al., 2011; Mannion et al., 2011; Benson and Mannion, 2012; Lloyd, 2012), map areas (Smith, 2001; Crampton et al., 2003; Smith, 2007a; Smith and McGowan, 2007, 2008; Wall et al., 2009, 2011), and counts of published papers (Sheehan, 1977) have been used variously as metrics of geological and human factors, but without clear evidence that these metrics do represent effective proxies. All commonly used sampling proxies have been criticised as inadequate in one way or another (Raup, 1977; Wignall and Benton, 1999; Crampton et al., 2003; Peters, 2006; Smith, 2007a; Benton, 2010; Peters and Heim, 2010; Benton et al., 2011; Dunhill, 2011, 2012), and this is a key theme for further consideration (see Section 6).

Table 4
Class definitions for specimen completeness metrics.

Specimen availability (for each genus)	
Class 1	One element, or a number of disarticulated elements
Class 2	One or more complete skull(s)
Class 3	One complete skeleton
Class 4	Several complete skeletons

In our study, we employed a number of such sampling proxies, in line with current practice, as a means of exploring the quality of our data and the effects of sampling (Table 5). First, we made counts of all tetrapod-bearing geological formations from within our database: these include all named units that have yielded tetrapod fossils that we include in our records. This is a 'strict fossiliferous formation count (FFC)' in the terms indicated by Benton et al. (2011, p. 71), consisting of 'only those formations that have produced named fossils included in the diversity measure'. Those authors recommended the use of a wider FFC, consisting of all formations that have ever produced any kind of fossil of the group in question, whether a named taxon, or unidentified elements, or trace fossils, or a comprehensive FFC that includes all geological formations of the correct facies that have produced, or might produce, fossils of the group in question. The use of such wider formation counts allows for inclusion of future fossil finds, but also documents formations that could potentially contain fossils of the group in question, even though these have not been found, and thus accommodates failure of sampling (Wignall and Benton, 1999). In our case, there is no such listing of worldwide, adequately dated geological formations, indicating facies and interpreted palaeoenvironment, and we did not attempt to construct such a listing.

We devised two new sampling metrics that reflect the completeness of the known fossils of each identified genus. The idea here was to seek to determine whether any time bins might be characterised by unusually incomplete fossils when compared to others. We term these metrics fossil completeness and the proportion of good fossils.

(A) *Mean fossil completeness* is the mean measure of values for all genera represented within a time bin, ranging from 1.0 (all taxa represented by incomplete remains) to 4.0 (all taxa represented by >2 complete skeletons each).

Table 5
Sampling metrics used for the fine-time scale (substage time bins) and coarse-time scale studies (stage time bins).

Fine time scale
1. Tetrapod-bearing formations
2. Fossil completeness (mean value of completeness classes 1–4 for all taxa in time bin)
3. Proportion of good fossils: ratio of good to poor material (ratio of numbers of taxa in completeness classes 2–4 to numbers in completeness classes 1–4).

Coarse time scale
1. Terrestrial formation count from the Paleobiology Database (PaleoDB, Alroy et al., 2001) downloaded in January 2012.
2. Terrestrial collections count from the PaleoDB. This is a measure of the number of individual collection records, each representing a single locality or geological formation.
3. Terrestrial occurrence count from the PaleoDB. This is a measure of all individual mentions of a genus, whether from different or identical collection records.
4. North American rock unit count from the Macrostrat database (Peters and Heim, 2010), downloaded in February 2012.
5. North American rock column count from Macrostrat.
6. North American total area measure from Macrostrat, based on occurrences of rocks of a particular age within predefined map area polygons.
7. North American fossiliferous units from Macrostrat and the PaleoDB.
8. North American fossil collections from Macrostrat and the PaleoDB.
9. North American fossil occurrences from Macrostrat and the PaleoDB.
10. Northwest European map areas from Smith and McGowan (2007): numbers taken directly from the published database, except for the time span from Serpukhovian to Olenekian ('Namurian' to 'Lower Triassic') where the time bins differed. In these cases, numbers of maps were divided roughly in proportion to the international stage durations.
11. Number of publications concerning tetrapods from each stage, based on a search of Google-Scholar in February 2012, using the search terms 'stage name AND tetrapod'.
12. Number of publications concerning tetrapods from each stage, based on a search of Web of Science in February 2012, using the search terms 'stage name AND tetrapod'.

(B) The *proportion of good fossils* is the ratio of good material (completeness classes 2–4) to all material (completeness classes 1–4), and it ranges from 0.0 (all taxa represented by category-1 material) to 1.0 (all taxa represented by category-2 to −4 material). The division is chosen between completeness classes 1 and 2–4 to discriminate the many taxa that have been based on disarticulated and isolated remains (class 1) versus those based on one or more complete skulls or skeletons (classes 2–4). This is probably a more useful discrimination of the data - in terms of the possibilities for valid, character-based species and generic discrimination—than, say, a discrimination of taxa based on scrap plus complete skulls (classes 1–2) versus those based on one or more complete skeletons (classes 3–4).

Such fossil completeness/preservation metrics have only rarely been used before in either regional- or global-scale studies. Benton et al. (2004) used counts of the numbers of collections, numbers of sampled localities, numbers of individual fossils, and crude measures of specimen completeness and quality, similar to those proposed here, as sampling proxies in their study of tetrapods from the Russian Permo-Triassic redbed basins. Further, Smith (2007b) applied such completeness measures to a global database of 352 species of Triassic and Lower Jurassic echinoids, identifying, as here, similar kinds of completeness classes, namely 1, complete test with spines; 2, complete test without spines; 3, partial test fragment (preserving both ambulacral and interambulacral plating); 4, interambulacral fragments and isolated plates; 5, isolated spines; 6, spine and test debris. At a global scale, Mannion and Upchurch (2010) proposed a 'Skeletal completeness metric' and a 'Character completeness metric' that allow for comparisons of data quality across time bins or phylogenies; these were not employed here because of the numbers of taxa involved and their wide phylogenetic spread, but it would be useful to compute such metrics, and compare their occurrences through time.

Other commonly used sampling proxies are hard to assemble at the chosen stratigraphic scale, so we acquired sampling proxy data at stage level for use with the data binned by coarse time scale. These include 12 potential sampling proxies representing aspects of rock volume, rock availability, and human effort (Table 5). Half of these 12 metrics are at continental scale (NW European map area; North American Macrostrat metrics) and so ideally should not be used as proxies for sampling of a global palaeodiversity record. However, this has commonly been done in the past (e.g. Peters and Foote, 2001; Smith, 2001; Peters and Foote, 2002; Peters, 2005; Smith, 2007a; Smith and McGowan, 2007) based on an argument that these proxies come from continents that dominate the global record. Hence these are included for comparison with previous studies, but can be disregarded. The inclusion of six out of 12 regional-based metrics could be said to swamp the global-scale metrics, and so reduce the pass level for significance after multiple-sample correction. In order to avoid this potential problem, multiple-sample correction is done for all 12 and for the six global-only sampling proxies, and differences in pass level are recorded for both sets of comparisons.

4. Statistical analysis

4.1. Time series modification, including generalised differencing

The palaeodiversity signals and all putative sampling metrics were considered first in their raw form, and then modified in various ways for purposes of comparison—calculated to logarithm base-10 (to normalise value distributions), time-standardised (to control for variable durations of time bins), first-differenced and generalised-differenced (to detrend the data and focus on rises and falls about the mean). The last two manipulations seek to remove general background trends, and although both methods have been called 'detrending', only generalised differencing (GD) has this attribute. First differencing consists simply of subtracting each value from its immediate precursor, and so it merely removes short-term serial correlations. GD, on the other hand, is a method (McKinney, 1990; Benson and Butler, 2011) that explicitly incorporates both identifying any trends in the data (using a linear model) and the differences from it. As outlined by Benson and Butler (2011), time series were assessed first for the presence of a trend, by a significant fit of the least squares regression against the midpoint ages of the time bins. In those cases where a trend existed, the slope of the best-fitting line represents the autocorrelation coefficient (a), and autocorrelation (the correlation of values at different points within the time series) was removed to produce the generalised differenced values (tGD) according to the formula:

$$t_{GD} = t_i - a t_{i-1}$$

All data transformations were carried out in Excel, except generalised differencing which was completed using Graeme Lloyd's 'functions_2.r' script in R v.2.14.1 (R Development Core Team, 2011; see www.graemetlloyd.com), together with an automated R code, 'gd.R' written by M.S. and designed to process large numbers of time series from single .csv files (Appendix A). These are standard approaches, used in many such prior studies (e.g. Benson et al., 2010; Benson and Butler, 2011; Benton et al., 2011; Butler et al., 2011; Lloyd et al., 2011; Dunhill et al., in press-a, in press-b).

4.2. Pairwise tests of correlation

Spearman rank correlation tests were carried out between the palaeodiversity signals at fine and coarse stratigraphic scales and the different sampling proxies, in order to determine whether any showed convincing correlation and so could explain some of the signal variance. Further, all the sampling proxies were tested against each other for correlation in order to determine how they might or might not represent elements of the same sampling bias signal. All statistical tests were carried out using R v.2.14.1 (R Development Core Team, 2011) and corrections for multiple correlation tests were carried out with the False Discovery Rate (FDR) approach of Benjamini and Hochberg (1995), using an R script developed by M.S., pair.cor.R (see Appendix A), which runs numerous correlation tests with unadjusted and adjusted p values.

The problem of Type II statistical errors, where one fails to reject a false null hypothesis, is a risk for multiple pairwise comparisons that is well recognised in studies such as this (e.g. Benson et al., 2010; Benson and Butler, 2011; Butler et al., 2011; Dunhill, 2011; Dunhill et al., in press-a, in press-b) where the sheer number of comparisons means that correlations might be recognised erroneously (false positives). The Bonferroni correction has been commonly used, but this is seen as a rather punitive measure that lowers the pass level for statistical significance perhaps too harshly (Benjamini and Hochberg, 1995), and so we prefer to use the less conservative FDR metric, as noted.

The FDR corrections were performed on families of comparisons (for example all raw data sets; all log-transformed data sets), rather than on the whole set of comparisons—for example, for the coarse (stage-level) time scale, 56 metrics were cross-compared, so generating 3136 Spearman's *rho* values, and the figure was 1225 comparisons for the fine (substage-level) timescale data set. FDR correction across all these correlations would drive the pass rate for correlation far too low. The 'families' of comparisons consisted of totals from 7 to 15 time series in each set.

4.3. Multiple linear models

In the face of many potential explanatory variables, it is rather unsatisfactory simply to compare numerous independent pairwise comparisons. A modelling technique involving multiple regressions can

assess the explanatory power of each such explanatory variable independently, but importantly also in every possible combination. Such modelling approaches have been applied to palaeontological time series by several authors (e.g. Marx and Uhen, 2010; Benson and Butler, 2011; Butler et al., 2011; Benson and Mannion, 2012), and there are several approaches.

Here we use multiple linear regression, as executed by the standard 'lm' function in R, which produces regression fits by simple and multiple least-squares regression, specifying the goodness of fit by the coefficient of determination (r^2), adjusted r^2 (which applies a penalty for more fitted parameters), and by Akaike information criterion (AIC) as a measure of the relative goodness of fit of the specified model (again, with a penalty proportional to the number of model parameters). The initial routine considers all putative explanatory variables together.

Evidently, some of the putative explanatory variables fit the response variable time series better than others, and their goodness of fit can be determined simply by ranking them in terms of r^2 and probability. However, this does not allow for the possibility that pairs and larger groupings of explanatory variables might function together to provide an even closer fit, or more convincing explanation of the response variable. Model fitting can be carried out in a stepwise manner using the standard R function 'step'. This can operate either forwards by adding likely predictor variables one at a time in order to improve the overall goodness of fit of the composite model, or backwards by eliminating poorly fitting explanatory variables. The R 'step' function also allows both forwards and backwards approaches. The best-fitting model is identified by step-wise experimentation, and overall improvements to the AIC and adjusted r^2. Final 'best' models can consist of anything from one to all explanatory variables, but the totals were usually 2–5 variables. We carried out these manipulations using the R routine 'mlm.R' (Appendix A).

5. Results

5.1. Palaeodiversity

The raw palaeodiversity data from the ETD (Fig. 1A) show a classic 'spiky' curve, which suggests that the record reflects a mix of real global diversifications and extinctions overlain by major rises and falls in sampling. The total generic sample in the ETD, although comprising more than 1300 genera, is still small enough that apparent global diversity can plunge to levels as low as 20–40 genera in many time bins. Ignoring the Devonian and Early Carboniferous parts of the curve because of very low apparent diversity, two widely accepted mass extinction events stand out, the end-Permian and end-Triassic (Fig. 1A, arrows 5, 8).

Apart from issues of sampling, these plots cannot be read literally as a source of data on extinction rates. For example, from the Changhsingian to Induan stages, crossing the Permo-Triassic boundary, tetrapod generic diversity rose from 94 to 112, but those totals encompass major losses in the Changhsingian, and an even higher rate of origination in the subsequent Induan stage. So, the actual generic loss across the Permo-Triassic boundary was 84 out of 94 genera present in the Changhsingian stage (89% loss), and across the Triassic-Jurassic boundary, 9 out of 22 genera in the upper Rhaetian (41% loss). The recoveries after the end-Permian and end-Triassic mass extinctions are shown by the origin of 102 out of 112 taxa in the Induan (91%) and 26 out of 39 taxa in the Hettangian (67%).

There are seven further diversity drops in the tetrapod generic palaeodiversity curve (Fig. 1A, numbers 1–4, 6, 7, 9), some matching recognised events and others possibly representing merely data loss. The first is at the end-Moscovian (Fig. 1A, arrow 1), 306.5 Myr ago (51 to 24 genera, 47% reduction), generated almost entirely by a major decline among the 'amphibians', with a loss of 40 out of 44 genera (91%) – this has been identified (Sahney et al., 2010) as a

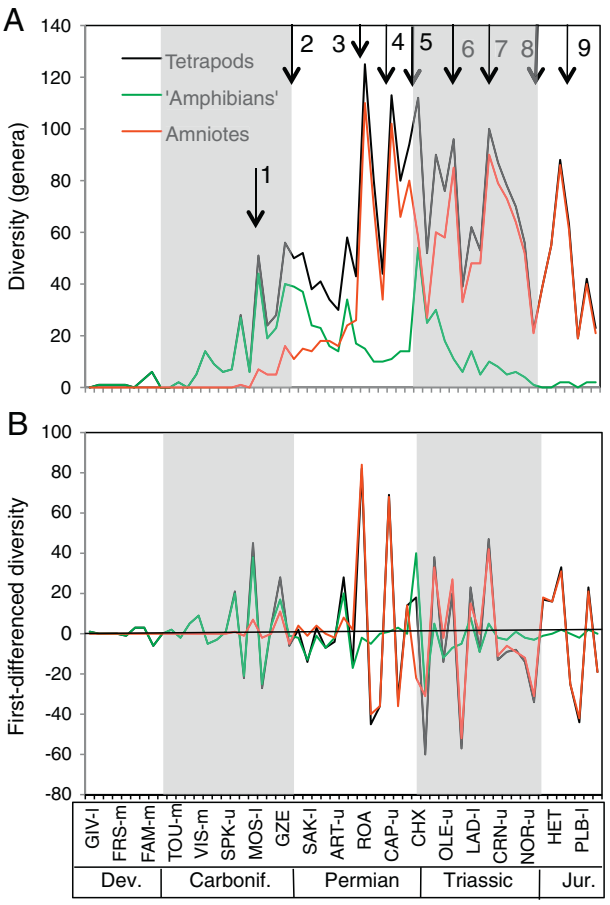

Fig. 1. Diversity of early tetrapods through the first half of their evolution, from the Middle Devonian (Givetian, 380 Myr ago) to Early Jurassic (Toarcian, 190 Myr ago), showing raw generic palaeodiversity data (A) and first differenced data (B), against the 'fine' substage-level time scale. Curves for 'amphibians' (= non-amniote tetrapods), amniotes, and tetrapods (= 'amphibians' + amniotes) are shown separately. Key events are marked by arrows (see text), including the end-Permian and end-Triassic mass extinctions (arrows 5 and 8 respectively). Stratigraphic acronyms are given in Table 2.

time of profound aridification of the Earth when the characteristic lush, tropical 'coal forests' declined rapidly, and with them their faunas of large arthropods and 'amphibians'.

Next is the apparently long decline in tetrapod diversity during the Early Permian (Fig. 1A, arrow 2), from the late Sakmarian to early Kungurian, approximately 290–273 Myr ago, when tetrapod generic diversity fell, in fits and starts, from 52 genera to 30 (42% loss). These losses were mainly among the 'amphibians', among which certain clades had re-radiated after the end-Moscovian climate crisis, before beginning a steady decline through the remainder of the Permian (Ruta and Benton, 2008; Carroll, 2009; Sahney et al., 2010). As they apparently dwindled, amniote diversity rose through the Middle Permian, and outstripped that of the non-amniote tetrapods. The Sakmarian to Kungurian diversity fall among tetrapods had long been posited to result purely from sampling failure, and it was termed Olson's Gap by Lucas (2004), to highlight the dwindling of the rich North American Early Permian tetrapod-bearing redbed rocks at the end of the Lower Permian, and the supposed temporal gap before the Russian and South African Middle and Upper successions began. This, however, appears to be an artefact of incorrect dating, and the upper parts of the North American successions in the upper Kungurian and Roadian squarely overlap the basal units (Ufimian, Kazanian) of the Russian tetrapod-bearing successions (Reisz and Laurin, 2001; Lozovskiy et al., 2009; Benton, 2012). Nonetheless, the long Early Permian decline among tetrapods, may be partly a real

decline among 'amphibians', but also partly a bias introduced by the loss of richly fossiliferous rock in North America.

There are two diversity peaks and subsequent declines in the Middle Permian (Fig. 1A, arrows 3, 4). During the Wordian, 110 out of 125 genera disappeared (88% loss), and these losses are primarily among amniotes, including genera of parareptiles (bolosaurids, nycteroleterids), basal synapsids (biarmosuchians as well as anteosaurid and estemmenosuchid dinocephalians). A similar diversity reduction, 92 out of 113 genera (81%), occurred in the lower Wuchiapingian, with the loss of further genera of parareptiles (millerettids, pareiasaurids), dicynodonts, gorgonopsians, and therocephalians (Benton et al., 2004; Rubidge, 2005; Benton, 2012). Neither of these apparent extinction events matches the Middle Permian (mid-Capitanian) mass extinction identified among marine organisms (Bond et al., 2010): this corresponds to a significant replacement of dinocephalians by anomodonts, but it does not stand out numerically in the palaeodiversity curves.

Immediately following the end-Permian mass extinction (Fig. 1A, arrow 5), 'amphibians' show a remarkable diversity peak in the Induan, with the origin of 50 of 58 genera (86% rise), but followed by a rate of loss of these genera just as rapid as their rise. This earliest Triassic diversity peak among temnospondyls and other non-amniote tetrapods has long been noted (Milner, 1990; Benton et al., 2004; Ruta and Benton, 2008), but has yet to be fully understood. Was this a real burst of radiation following the end-Permian crisis in which temnospondyls were acting in some way as disaster taxa or in which climatic conditions somehow favoured these primarily water-living tetrapods, or is this an artefact of a major change in sedimentation across the Permo-Triassic boundary towards higher-energy streams (Newell et al., 1999; Ward et al., 2000) and so perhaps excessive preservation of detrital remains of water-living animals?

Following a late Olenekian peak of 90 tetrapod genera, diversity fell to 76 in the lower Anisian, rose again to 96 in the late Anisian, and then plummeted to 39 in the lower Ladinian (Fig. 1A, arrow 6). In earlier works (e.g. Benton, 1985, 1995) a terminal Early Triassic peak of extinction among tetrapods had already been noted, but re-dating of redbed units in South Africa (e.g. *Cynognathus* Assemblage Zone) and elsewhere has extended many formerly late Olenekian ranges into the Middle Triassic. Whether, however, any of these declines are real, and match the turbulence in atmospheric and oceanic conditions that continued through the Olenekian and early Anisian (Payne et al., 2004; Chen and Benton, 2012), or are artefacts of sampling is to be resolved.

The Late Triassic apparently witnessed a long decline among tetrapod genera (Fig. 1A, arrows 7 to 8). The peak of diversity, with 100 genera, is in the late Carnian, and global tetrapod diversity fell steadily through the five time bins of the Norian and Rhaetian, with small losses among the already rare 'amphibians', but primarily among amniotes (mainly among synapsids and crurotarsan archosaurs). These declines, especially those in the lower and middle Norian, have been identified before (Benton, 1983, 1993; Brusatte et al., 2008a,b) as the Carnian/Norian turnover among amniotes, when dicynodonts and rhynchosaurs disappeared following major climatic and floral changes—the event was expressed more in terms of changes in relative abundances and ecosystem structure than in raw palaeodiversity. The highest percentage loss was from 56 to 22 taxa (39% loss) between lower and upper Rhaetian, but this might be partly an artefact of the loss of small derived cynodonts (traversodontids, dromatheriids), many based on teeth and jaws, combined with a major reduction in appropriate fossiliferous units.

The final diversity drop to be considered, in the Early Jurassic (Fig. 1A, arrow 9), saw a reduction from 88 genera to 19 (78% loss) from the late Sinemurian to late Pliensbachian. These losses were mainly among sphenosuchid crocodylomorphs, basal sauropodomorph ('prosauropod') dinosaurs, basal ornithischian dinosaurs, tritylodontids, and mammals, and may reflect a steady reduction in suitable terrestrial formations.

Overall then, the tetrapod diversity signal (Fig. 1A) is identical at first (Devonian-mid Carboniferous) with the 'amphibian' curve, then the steady rise of amniotes through the Late Carboniferous and Early Permian coincides with a switch to near-100% dominance of the signal by amniotes through the Middle Permian to Early Jurassic interval, during which 'amphibians' dwindled substantially in diversity, except for the Induan peak.

The all-tetrapod diversity curve (Fig. 1A) may be represented by a single straight line ($y = 1.3471x + 2.5535$, Spearman's $rho = -0.382$, $p = 0.003$; *rho* value negative because time decreases towards the present), which permits detrending. In order to explore the effects of short-term fluctuations in the curves, we remove short-term serial correlations by taking first differences (Fig. 1B) and by generalised differencing (Tables 6, 7). These curves highlight the bin-by-bin changes in apparent diversity through time, with vicissitudes in the tetrapod record dominated first by the 'amphibian' data, and then by the amniote data. The extinction events and diversifications, numbered 1–9 (Fig. 1A) are all identifiable in the first-differenced time series (Fig. 1B).

5.2. Sampling proxies and substage-scale time bins

When the raw palaeodiversity curve plotted against substage time bins is compared with the raw formations count (Fig. 2A), there appears to be generally good matching, especially in the Late Permian to Early Jurassic portion. Tracking through time, the highs and lows

Table 6
Comparisons of the raw ETD palaeodiversity data on a fine time scale (58 stratigraphic substages) with putative sampling proxies. The comparisons, Spearman's *rho*, probability, and Benjamini–Hochberg adjusted probability values are given. The first three correlations are run for the whole data set, and then for two subdivisions, the first ('1') running from Middle Devonian to Middle Permian, and the second ('2') from Late Permian to Early Jurassic, both inclusive. Significant ($p<0.05$)* and highly significant ($p<0.005$)** correlations are indicated. Abbreviations: FD, first-differenced; GD = generalised-differenced (detrended); Log, logarithm to base 10.

Comparison	Spearman's *rho*	*p*	adjusted *p*
1. ETD formation count			
'Amphibian' total vs. Formations	0.398	0.002**	0.171
Amniote total vs. Formations	0.847	0.000**	0.000**
Tetrapod total vs. Formations	0.867	0.000**	0.000**
1: 'Amphibian' total vs. Formations	0.392	0.043*	1
1: Amniote total vs. Formations	0.542	0.003**	0.595
1: Tetrapod total vs. Formations	0.633	0.000**	0.067
2: 'Amphibian' total vs. Formations	0.878	0.000**	0.000**
2: Amniote total vs. Formations	0.765	0.000**	0.000**
2: Tetrapod total vs. Formations	0.868	0.000**	0.000**
Log ('Amphibian' total vs. Formations)	0.432	0.001**	0.036*
Log (Amniote total vs. Formations)	0.843	0.000**	0.000**
Log (Tetrapod total vs. Formations)	0.881	0.000**	0.000**
('Amphibian' total vs. Formations)/time	0.604	0.000**	0.000**
(Amniote total vs. Formations)/time	0.880	0.000**	0.000**
(Tetrapod total vs. Formations)/time	0.921	0.000**	0.000**
FD ('Amphibian' total vs. Formations)	0.479	0.000**	0.016*
FD (Amniote total vs. Formations)	0.653	0.000**	0.000**
FD (Tetrapod total vs. Formations)	0.700	0.000**	0.000**
GD ('Amphibian' total vs. Formations)	0.439	0.001**	0.052*
GD (Amniote total vs. Formations)	0.489	0.000**	0.010*
GD (Tetrapod total vs. Formations)	0.589	0.000**	0.000**
FD ('Amphibian' total vs. Formations)/time	0.569	0.000**	0.000**
FD (Amniote total vs. Formations)/time	0.727	0.000**	0.000**
FD (Tetrapod total vs. Formations)/time	0.745	0.000**	0.000**
2. Fossil completeness counts and ratios			
'Amphibian' total vs. Fossil completeness	0.232	0.083	1
Amniote total vs. Fossil completeness	0.553	0.000**	0.000**
Tetrapod total vs. Fossil completeness	0.098	0.469	1
'Amphibian' total vs. Proportion good fossils	0.222	0.096	1
Amniote total vs. Proportion good fossils	0.670	0.000**	0.000**
Tetrapod total vs. Proportion good fossils	0.239	0.074	1
Fossil completeness vs. Proportion good fossils	0.831	0.000**	0.000**
Fossil completeness vs. Formations	0.053	0.694	1
Proportion good fossils vs. Formations	0.029	0.831	1

Table 7
Comparisons of the raw ETD palaeodiversity data on a coarse time scale (29 stratigraphic stages) with putative sampling proxies. The comparisons, Spearman's *rho*, probability, and Benjamini–Hochberg adjusted probability values are given. The first correlation is run for the whole data set, and then for two subdivisions, the first ('1') running from Middle Devonian to Early Permian, and the second ('2') from Middle Permian to Early Jurassic, both inclusive. Significant ($p<0.05$)* and highly significant ($p<0.005$)** correlations are indicated. The 12-sample and 6-sample adjusted *p* values refer to adjustments for all 12 sampling proxies, and for the global-scale proxies only (i.e. excluding Macrostrat variants and European map areas). Abbreviations: coll., collections; FD, first-differenced; GD, generalised-differenced (detrended); Log, logarithm to base 10; occ., occurrences.

Comparison	Spearman's *rho*	*p*	12-sample adjusted *p*	6-sample adjusted *p*
1. ETD formation count				
Tetrapod total vs. Formations	0.860	0.000**	0.000**	0.000**
1: Tetrapod total vs. Formations	0.687	0.005**	0.425	0.098
2: Tetrapod total vs. Formations	0.778	0.002**	0.157	0.006*
Log (Tetrapod total vs. Formations)	0.860	0.000**	0.000**	0.000**
(Tetrapod total vs. Formations)/time	0.868	0.000**	0.000**	0.000**
FD (Tetrapod total vs. Formations)	0.743	0.000**	0.000**	0.000**
GD (Tetrapod total vs. Formations)	0.632	0.000**	0.038*	0.009*
2. PaleoDB counts				
Tetrapod total vs. PaleoDB formations	0.424	0.022*	1	0.462
Tetrapod total vs. PaleoDB collections	0.484	0.008*	0.707	0.163
Tetrapod total vs. PaleoDB occurrences	0.242	0.207	1	1
Log (Tetrapod total vs. PaleoDB formations)	0.424	0.022*	1	0.462
Log (Tetrapod total vs. PaleoDB collections)	0.484	0.008*	0.707	0.163
Log (Tetrapod total vs. PaleoDB occurrences)	0.242	0.207	1	1
(Tetrapod total vs. PaleoDB formations)/time	0.705	0.000**	0.002**	0.000**
(Tetrapod total vs. PaleoDB collections)/time	0.689	0.000**	0.003**	0.000**
(Tetrapod total vs. PaleoDB occurrences)/time	0.527	0.003**	0.302	0.070
FD (Tetrapod total vs. PaleoDB formations)	0.446	0.015*	1	0.323
FD (Tetrapod total vs. PaleoDB collections)	0.464	0.011*	1	0.237
FD (Tetrapod total vs. PaleoDB occurrences)	0.246	0.198	1	1
GD (Tetrapod total vs. PaleoDB formations)	0.431	0.023*	1	0.480
GD (Tetrapod total vs. PaleoDB collections)	0.310	0.108	1	1
GD (Tetrapod total vs. PaleoDB occurrences)	0.205	0.295	1	1
3. Macrostrat counts				
Tetrapod total vs. Macrostrat units	−0.118	0.543	1	NA
Tetrapod total vs. Macrostrat columns	0.030	0.879	1	NA
Tetrapod total vs. Macrostrat areas	0.025	0.896	1	NA
Tetrapod total vs. Macrostrat fossil units	0.032	1	1	NA
Tetrapod total vs. Macrostrat fossil collections	0.184	1	1	NA
Tetrapod total vs. Macrostrat fossil occurrences	0.219	1	1	NA
Log (Tetrapod total vs. Macrostrat units)	−0.118	0.543	1	NA
Log (Tetrapod total vs. Macrostrat columns)	0.030	0.879	1	NA
Log (Tetrapod total vs. Macrostrat areas)	0.025	0.896	1	NA
Log (Tetrapod total vs. Macrostrat fossil units)	0.039	0.869	1	NA
Log (Tetrapod total vs. Macrostrat fossil coll.)	0.192	0.339	1	NA
Log (Tetrapod total vs. Macrostrat fossil occ.)	0.219	0.254	1	NA
(Tetrapod total vs. Macrtostrat units)/time	0.218	0.255	1	NA
(Tetrapod total vs. Macrostrat columns)/time	0.370	0.048*	1	NA
(Tetrapod total vs. Macrostrat areas)/time	0.370	0.048*	1	NA
(Tetrapod total vs. Macrostrat fossil units)/time	0.239	0.212	1	NA
(Tetrapod total vs. Macrostrat fossil coll.)/time	0.163	0.397	1	NA
(Tetrapod total vs. Macrostrat fossil occ.)/time	0.210	0.274	1	NA
FD (Tetrapod total vs. Macrostrat units)	0.408	0.028*	1	NA
FD (Tetrapod total vs. Macrostrat columns)	0.415	0.025*	1	NA
FD (Tetrapod total vs. Macrostrat areas)	0.375	0.045*	1	NA
FD (Tetrapod total vs. Macrostrat fossil units)	0.313	0.098	1	NA
FD (Tetrapod total vs. Macrostrat fossil coll.)	0.434	0.019*	1	NA
FD (Tetrapod total vs. Macrostrat fossil occ.)	0.436	0.018*	1	NA
GD (Tetrapod total vs. Macrostrat units)	0.013	0.948	1	NA
GD (Tetrapod total vs. Macrostrat columns)	−0.050	0.799	1	NA
GD (Tetrapod total vs. Macrostrat areas)	−0.058	0.769	1	NA
GD (Tetrapod total vs. Macrostrat fossil units)	0.001	0.997	1	NA
GD (Tetrapod total vs. Macrostrat fossil coll.)	0.081	0.681	1	NA
GD (Tetrapod total vs. Macrostrat fossil occ.)	0.125	0.525	1	NA
4. European map areas				
Tetrapod total vs. European map area	0.020	1	1	NA
Log (Tetrapod total vs. European map area)	0.020	0.918	1	NA
(Tetrapod total vs. European map area)/time	0.236	0.217	1	NA
FD (Tetrapod total vs. European map area)	0.234	0.221	1	NA
GD (Tetrapod total vs. European map area)	0.004	0.986	1	NA
5. Publication counts				
Tetrapod total vs. Papers-Google	0.297	1	1	1
Tetrapod total vs. Papers-WoS	0.102	1	1	1
Log (Tetrapod total vs. Papers-Google)	0.297	0.118	1	1
Log (Tetrapod total vs. Papers-WoS)	0.102	0.599	1	1

(continued on next page)

Table 7 (*continued*)

Comparison	Spearman's *rho*	*p*	12-sample adjusted *p*	6-sample adjusted *p*
(Tetrapod total vs. Papers-Google)/time	0.599	0.000**	0.054	0.012*
(Tetrapod total vs. Papers-WoS)/time	0.315	0.097	1	1
FD (Tetrapod total vs. Papers-Google)	0.325	0.085	1	1
FD (Tetrapod total vs. Papers-WoS)	0.356	0.058	1	1
GD (Tetrapod total vs. Papers-Google)	0.127	0.518	1	1
GD (Tetrapod total vs. Papers-WoS)	0.240	0.217	1	1

between palaeodiversity and formation count appear to be well matched in the Devonian and Early Carboniferous, although counts for both time series are low and perhaps comparisons through this portion of the time series are meaningless. There is a major mismatch in the Late Carboniferous to Early Permian interval, with diversity peaks in the upper Bashkirian and lower Asselian, but a single formations peak in between, in the Gzhelian. The two time series go somewhat out of close matching through the Early and Middle Permian, and then appear to be much more closely correlated through the Late Permian, Triassic, and Early Jurassic. Note that the diversity lows (20–40 genera) often correspond to time intervals with small numbers of fossiliferous formations (5–10).

When the three diversity signals, for 'amphibians', amniotes, and all tetrapods, are compared with formation counts, there are highly significant correlations (Table 6, part 1), except for the 'amphibian' totals when corrected for multiple comparisons. When the data set is split at the Early/Middle Permian boundary, and each half is considered separately, correlations are lower for the first half, and higher for the second half, matching expectations from a visual inspection of the two curves (Fig. 2A). Even so, although matching of the time series in the first half of the time span is not significant for either 'amphibians' ($rho = 0.392$, $p = 1$) or amniotes ($rho = 0.542$, $p = 0.595$) separately, the combined tetrapod signal shows only marginally nonsignificant correlation ($rho = 0.633$, $p = 0.067$). Values are highly significant for all metrics in the second half of the time span.

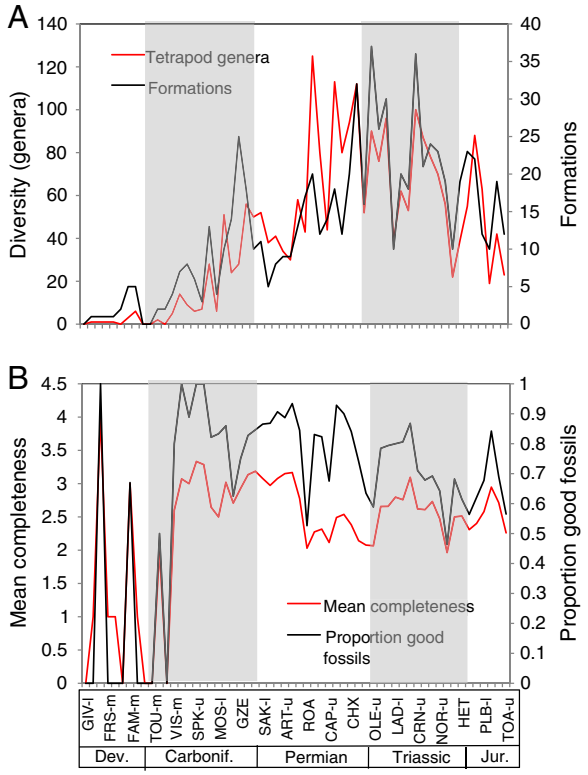

Fig. 2. Diversity of early tetrapods through the first half of their evolution, from the Middle Devonian (Givetian, 380 Myr ago) to Early Jurassic (Toarcian, 190 Myr ago), compared to some key sampling proxies, tetrapod-bearing formation count (A), and two measures of fossil completeness, mean completeness of all genera in each time bin, and the proportion of good fossils in each time bin (B). These metrics are defined in Table 5.

The evidence for tight covariation between the palaeodiversity and formation counts is borne out when the data are transformed, as logarithms to base-10, time-standardised, first-differenced, first-differenced and time-standardised, and generalised differenced (Table 6, part 1). In all cases, the 'amphibian' time series shows weaker correlation (lower Spearman's *rho*) with the formations time series than the amniote or tetrapod time series, but correlations are all significant or highly significant.

The results of comparisons are rather different for fossil completeness. Only the amniote time series shows correlation, and it is a highly significant correlation for both the fossil completeness metric and the proportion of good fossils metric, whereas 'amphibians' on their own and tetrapods as a whole show no evidence of correlation with either fossil completeness metric. These two completeness metrics correlate highly significantly with each other, but not at all with formation count (Table 6, part 2), and this is evident when the two time series are inspected visually (Fig. 2B).

The fossil completeness curves (Fig. 2B) provide novel data on the materials available for study, and they show substantial variations in the mean and proportional versions of the same genus-based completeness class codings. Certain time bins, notably the Kasimovian, Roadian, Capitanian (upper), Olenekian (lower, upper), Rhaetian (lower), Sinemurian (lower), and Toarcian (upper) are all marked by substantial drops in completeness scores, meaning that relatively high numbers of genera of those ages were named on the basis of incomplete fossil remains. This could reflect unusual practice by palaeontologists, but more likely indicates times of particularly poor fossils. Each comparison is between time bin samples of forty or more genera, so the variations in means and proportions are probably meaningful; however, formation counts, at 10–20 per time bin, are much lower, and so low values of specimen completeness in some or all of these time bins could arise from small numbers of geological formations that offer particularly incomplete fossils. Either way, these metrics might represent some aspect of sampling bias, and they certainly highlight time bins in which the fossil materials are likely to be less well understood than in others.

The relationship is not so simple, however. It might have been expected that the times of poor fossil completeness would correspond to diversity lows, but when each completeness proxy is compared with diversity (Fig. 3), there is only a weak positive relationship between these time series, with only 9% and 17% of the diversity signal explained either by mean completeness (A) or proportion of good fossils (B). Most values cluster around the mean values of 2.51 (A) and 0.70 (B) respectively, and the times of mass extinction (CHX, RHT-u) and recovery (IND, HET) show, for example, no special characteristics in terms of sample completeness and observed diversity. These weakly positive trend lines become more-or-less horizontal when the four substages with zero diversity (lower Famennian, lower Tornaisian, middle Tournaisian, lower Viséan) are excluded, and this tends to confirm the absence of compelling evidence for correlation of fossil quality and diversity.

5.3. Palaeodiversity and sampling proxies

In order to compare the present palaeodiversity curve with those derived from earlier data compilations, and with a greater range of putative sampling proxies, the time bins were generalised to coarser

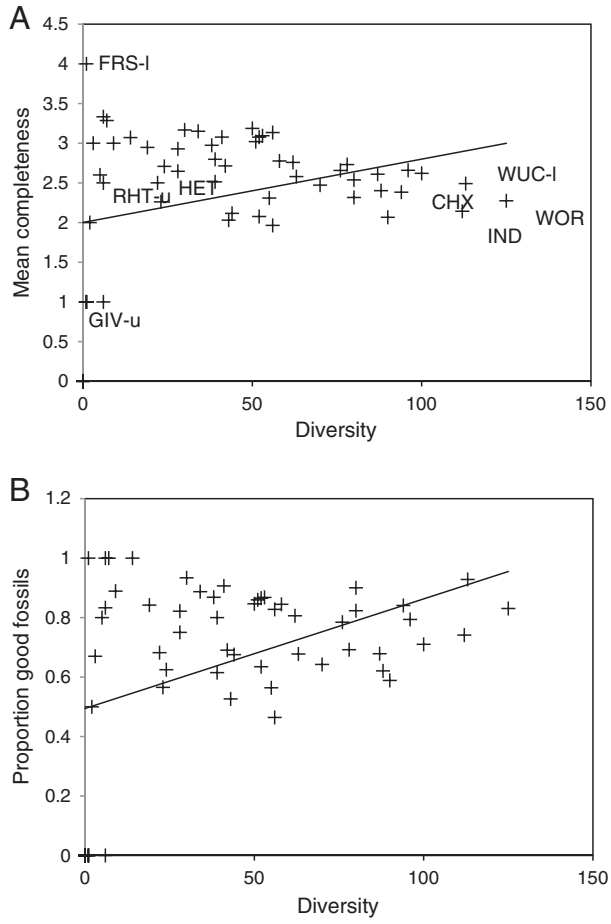

Fig. 3. Comparison of specimen completeness metrics with early tetrapod diversity, plotted for the 58 time bins from lower Givetian to upper Toarcian, shown as (A) mean completeness value, ranging from 0 (worst) to 4 (best), and (B) the proportion of 'good' fossils (i.e. complete skull or skulls or complete skeleton or skeletons). There is a very weakly positive relationship in each case between overall diversity and the completeness metric, with coefficients of determination (r^2) of 0.09 (A) and 0.17 (B). Some time intervals are indicated, using standard acronyms for stratigraphic stages (Table 1).

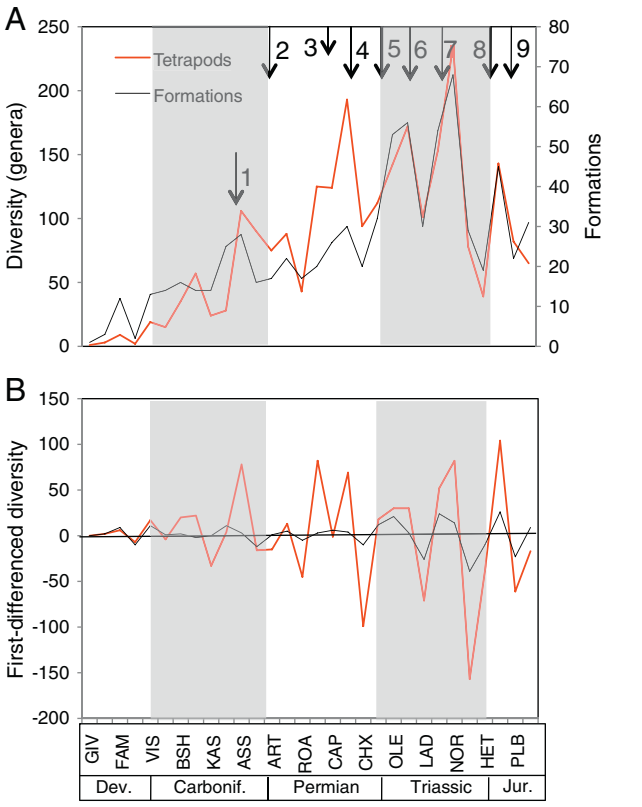

Fig. 4. Diversity of early tetrapods through the first half of their evolution, from the Middle Devonian (Givetian, 380 Myr ago) to Early Jurassic (Toarcian, 190 Myr ago), showing raw tetrapod generic palaeodiversity data (A) and first differenced data (B) against the 'coarse' stage-level time scale. In both cases, the tetrapod generic curve is compared with the counts of tetrapod-bearing formations. Key events are marked by arrows (see text), including the end-Permian and end-Triassic mass extinctions (arrows 5 and 8 respectively). Stratigraphic acronyms are given in Table 1.

time divisions, generally stages, averaging 7.269 Myr in length. At this level, many of the distinctive events identified above (Fig. 1) are harder to discern. For example, the late Moscovian and Early Permian diversity falls are partly blended together (Fig. 4A, arrows 1, 2). The Wordian and lower Wuchiapingian drops (Fig. 4A, arrows 3, 4) appear as a step and a major decline respectively, and the latter combines with the end-Permian mass extinction decline (Fig. 4A, arrow 5). The Anisian-Ladinian drop and the long decline in the Late Triassic (Fig. 4A, arrows 6, 7) are clear, but the latter extends into the terminal Triassic mass extinction (Fig. 4A, arrow 8), and the Early Jurassic peak and decline (Fig. 4A, arrow 9) are also present.

When modified by first differencing, the diversity curve shows how the fluctuations about the mean increase through geological time (Fig. 4B), as was seen also at finer time scales (Fig. 1B), although the effect appears to be enhanced for the coarser, stage-level time bins.

As noted earlier, there is generally strong correlation between the palaeodiversity and formation count signals, whether the data are raw or first-differenced (Fig. 4). As with the finer time scale (Fig. 2A), the matching of peaks and troughs is better in the second half of the time series, with apparently poor correspondence through the Devonian to Late Carboniferous time interval, but close matching of peaks in palaeodiversity and formation counts in the Early Permian, Middle Permian, Early Triassic, Late Triassic, and Early Jurassic. When correlations are assessed (Table 7, part 1), there is a highly significant relationship between the tetrapod totals and formation counts, but this disappears (12-sample correction), or is reduced for the second time span only (6-sample correction) when the data are split into two time segments, Middle Devonian to Early Permian and Middle Permian to Early Jurassic. The loss or reduction of correlation when the time series is split arbitrarily suggests need for caution, and this has been noted before for the Permian alone (Benton, 2012). The generic palaeodiversity curve and ETD formation count also correlates highly significantly when first-differenced and significantly when generalised-differenced (Table 7, part 1).

A substantial exercise of comparison with numerous putative rock-volume proxies was then performed (Table 7). Whereas there was generally significant to highly significant correlation between the palaeodiversity signal and the ETD formation count, according to all versions of data treatment, comparisons with other metrics of rock volume showed limited to no correlation (Table 7, parts 2–5). Comparisons were made with PaleoDB global formations, collections, and occurrences, and the only significant correlations after correction for multiple comparisons were with the time-adjusted (divided by bin length) formations and collections counts (Table 7, part 2). Remarkably, no correlations were found with any of the variants of the North American rock volume data from Macrostrat, or with the European map-area data (Table 7, parts 3, 4).

The general absence of correlation of the tetrapod palaeodiversity time series with global and regional rock-volume proxies might not be surprising, but some covariance with metrics of human effort might be expected. The counts of numbers of papers published on tetrapods from each stratigraphic stage, from both Google-Scholar and Web of Science show somewhat variable patterns through geological

time (Fig. 5). The totals from Google-Scholar are generally much higher than from Web of Science, using the same search criteria, but when plotted on a double-y graph (Fig. 5), the shapes of the curves line up, except for a huge peak in the Google count of publications for 'Induan tetrapods'—searches with different permutations of key words did not break down this unusually high peak, so it cannot be further explained (why for example is there not a similar peak for 'Olenekian tetrapods', if the Induan peak relates to the plethora of publications on recovery from the end-Permian mass extinction?). Otherwise, both censuses of papers (Fig. 5A) show modest peaks in the Late Devonian and Early Carboniferous, perhaps reflecting intense research on the rather rare basal tetrapods. The low publication levels through the Late Carboniferous and Permian are surprising—although these are represented by 100–250 papers per stratigraphic stage from the Google counts—one might have expected a steady increase through the Permian. The anomalous Induan peak is followed by a low value in the Olenekian, followed by a peak in the Middle and Late Triassic and, again perhaps surprisingly, quite a substantial drop in publication numbers through the Early Jurassic. The first-differenced data (Fig. 5B) show these general similarities between both records of publications, except for the Induan peak. Both plots (Fig. 5A, B) show little evident correlation with the palaeodiversity curve, and this is confirmed by the general absence of correlation in various permutations of the time series (Table 7, part 5), a rather surprising result. The only human effort metric that shows correlation is the time-adjusted count of papers from Google-Scholar ($p = 0.054$, when adjusted for 12 comparisons; $p = 0.012$, significant, when adjusted for six comparisons).

When plotted on an x–y plot of paper counts against diversity (Fig. 6), there is only a weak positive relationship between both measures, with only 17% (A) and 10% (B) of the diversity count explained by the two versions of the paper counts. Values scatter quite widely, and there does not seem to be a temporal pattern of increasing numbers of papers, nor any especially high numbers for the times of maximum extinction (CHX, RHT), nor for the immediately subsequent recovery times (IND, HET).

5.4. Interrelations of sampling proxies

Presumably many of the putative sampling proxies employed here, all of which are the same as sampling proxies used in similar studies recently, ought to reflect certain shared geological and sampling bias signals. If the various counts of formations, rock packets, map areas, and papers are dominated by the biases imposed on the fossil record by unequal rock volume, unequal rock availability, and unequal effort, then these putative sampling proxies ought to correlate, as has been found in some previous studies (e.g. Upchurch et al., 2011). This does not appear to be the case here, except among families of co-dependent metrics from the same databases (including correlation of various PaleoDB measures with each other, and correlation of various Macrostrat measures with each other).

In our study, we correlated everything against everything (see Supplementary material), and reproduce here the comparisons of proxies from external sources with the internal ETD formation count metric, taken in a variety of raw and modified forms (Table 8). Whereas the ETD formation count correlates generally highly significantly with

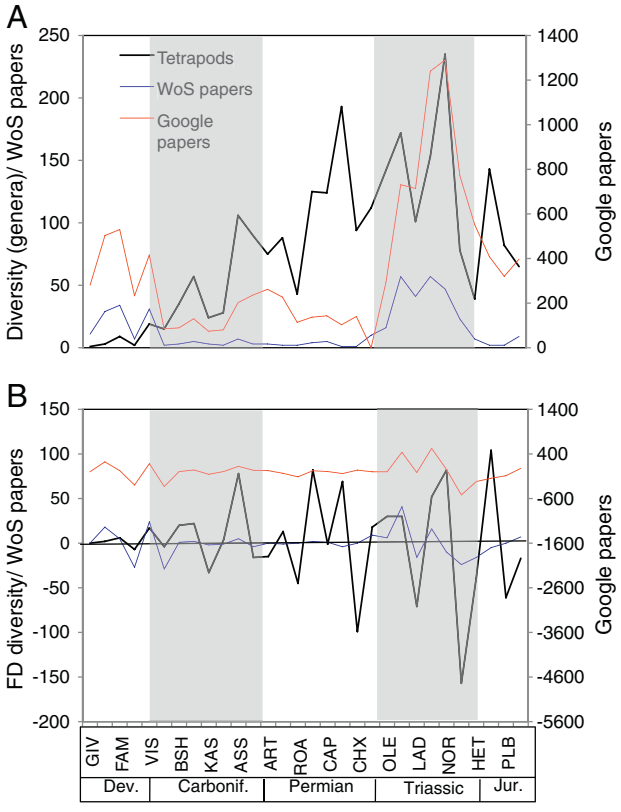

Fig. 5. Comparisons of metrics of human effort, counts of papers published concerning each stratigraphic stage and 'tetrapods', as assessed from Google-Scholar and from Web of Science, shown as raw data (A) and first-differenced (B), and in comparison to the tetrapod diversity signal. Note that the time series for Google papers (red; scale on right-hand y-axis) is offset upwards so it can be seen.

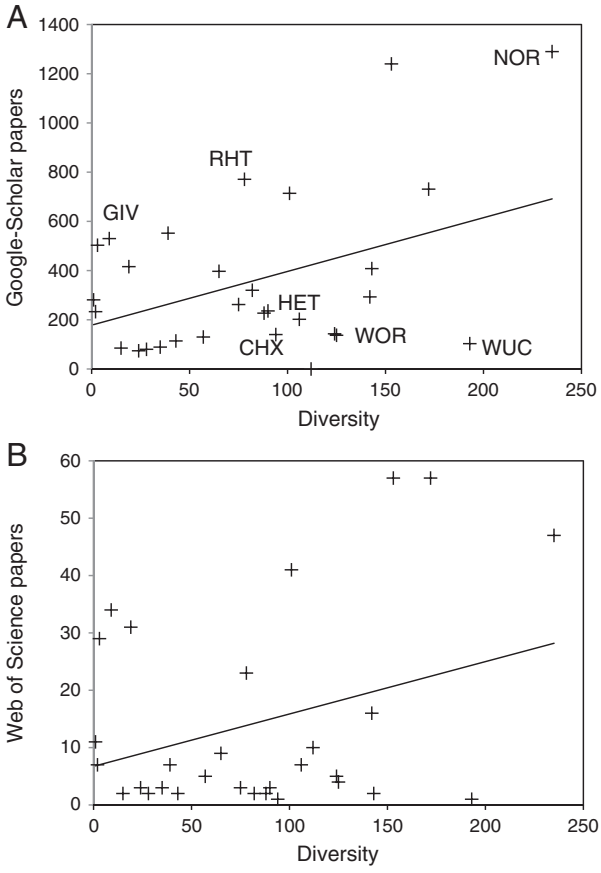

Fig. 6. Comparison of counts of published papers with early tetrapod diversity, plotted for the 58 time bins from lower Givetian to upper Toarcian, shown as counts from (A) Google-Scholar and (B) Web of Science. There is a very weakly positive relationship in each case between overall diversity and the completeness metric, with coefficients of determination (r^2) of 0.17 (A) and 0.10 (B). In (A), the anomalous high point for the Induan ($n = 3980$ papers) is omitted. Some time intervals are indicated, using standard acronyms for stratigraphic stages (Table 1).

Table 8
Comparisons of the putative sampling proxies with each other, comparing the ETD formation count with sampling time series from various sources, on the coarse time scale (20 stages). The comparisons, Spearman's *rho*, probability, and Benjamini-Hochberg adjusted probability values are given. Significant ($p<0.05$)* and highly significant ($p<0.005$)** correlations are indicated. Abbreviations: coll., collections; FD, first differenced; Log, logarithm to base 10; occ., occurrences.

Comparison	Spearman's *rho*	*p*	adjusted *p*
1. Raw data			
Formations vs. PaleoDB formations	0.438	0.017*	1
Formations vs. PaleoDB collections	0.479	0.008**	0.778
Formations vs. PaleoDB occurrences	0.288	0.129	1
Formations vs. Macrostrat units	0.039	0.841	1
Formations vs. Macrostrat columns	0.217	0.259	1
Formations vs. Macrostrat areas	0.170	0.377	1
Formations vs. Macrostrat fossil units	0.207	1	1
Formations vs. Macrostrat fossil coll.	0.401	1	1
Formations vs. Macrostrat fossil occ.	0.424	1	1
Formations vs. European map area	0.325	1	1
Formations vs. Papers-Google	0.434	1	1
Formations vs. Papers-WoS	0.233	1	1
2. Log-transformed data			
Log (Formations vs. PaleoDB formations)	0.438	0.017*	1
Log (Formations vs. PaleoDB collections)	0.479	0.008*	0.778
Log (Formations vs. PaleoDB occurrences)	0.288	0.129	1
Log (Formations vs. Macrostrat units)	0.039	0.841	1
Log (Formations vs. Macrostrat columns)	0.217	0.259	1
Log (Formations vs. Macrostrat areas)	0.170	0.377	1
Log (Formations vs. Macrostrat fossil units)	0.213	0.268	1
Log (Formations vs. Macrostrat fossil coll.)	0.407	0.028*	1
Log (Formations vs. Macrostrat fossil occ.)	0.424	0.022*	1
Log (Formations vs. European map area)	0.325	0.085	1
Log (Formations vs. Papers-Google)	0.434	0.019*	1
Log (Formations vs. Papers-WoS)	0.232	0.225	1
3. Time-normalised data			
(Formations vs. PaleoDB formations)/time	0.671	0.000**	0.006*
(Formations vs. PaleoDB collections)/time	0.678	0.000**	0.007*
(Formations vs. PaleoDB occurrences)/time	0.456	0.014*	1
(Formations vs. Macrostrat units)/time	0.074	0.702	1
(Formations vs. Macrostrat columns)/time	0.213	0.265	1
(Formations vs. Macrostrat areas)/time	0.246	0.197	1
(Formations vs. Macrostrat fossil units)/time	0.099	0.610	1
(Formations vs. Macrostrat fossil coll.)/time	0.083	0.668	1
(Formations vs. Macrostrat fossil occ.)/time	0.114	0.555	1
(Formations vs. European map area)/time	−0.006	0.976	1
(Formations vs. Papers-Google)/time	0.532	0.003**	0.312
(Formations vs. Papers-WoS)/time	0.176	0.359	1
4. First-differenced data			
FD (Formations vs. PaleoDB formations)	0.338	0.073	1
FD (Formations vs. PaleoDB collections)	0.303	0.110	1
FD (Formations vs. PaleoDB occurrences)	0.100	0.605	1
FD (Formations vs. Macrostrat units)	0.173	0.370	1
FD (Formations vs. Macrostrat columns)	0.227	0.237	1
FD (Formations vs. Macrostrat areas)	0.072	0.709	1
FD (Formations vs. Macrostrat fossil units)	0.084	0.664	1
FD (Formations vs. Macrostrat fossil coll.)	0.351	0.062	1
FD (Formations vs. Macrostrat fossil occ.)	0.379	0.043*	1
FD (Formations vs. European map area)	0.310	0.102	1
FD (Formations vs. Papers-Google)	0.350	0.063	1
FD (Formations vs. Papers-WoS)	0.466	0.011*	0.993
5. Generalised-differenced data			
FD (Formations vs. PaleoDB formations)	0.431	0.049*	1
FD (Formations vs. PaleoDB collections)	0.310	0.091	1
FD (Formations vs. PaleoDB occurrences)	0.205	0.505	1
FD (Formations vs. Macrostrat units)	0.013	0.402	1
FD (Formations vs. Macrostrat columns)	−0.050	0.520	1
FD (Formations vs. Macrostrat areas)	−0.058	0.700	1
FD (Formations vs. Macrostrat fossil units)	0.001	0.638	1
FD (Formations vs. Macrostrat fossil coll.)	0.081	0.115	1
FD (Formations vs. Macrostrat fossil occ.)	0.125	0.030*	1
FD (Formations vs. European map area)	0.004	0.133	1
FD (Formations vs. Papers-Google)	0.127	0.207	1
FD (Formations vs. Papers-WoS)	0.240	0.072	1

palaeodiversity, whether considered on a fine or coarse time scale (Tables 6, 7), the other rock record and effort metrics do not apparently explain much, if anything, of the palaeodiversity signal. Without adjustment for multiple comparisons, the ETD formation count correlates highly significantly with PaleoDB collections and significantly with PaleoDB formations (Table 8, part 1). When log-transformed (Table 8, part 2), there are significant correlations with PaleoDB formations, PaleoDB collections, Macrostrat fossil collections, Macrostrat fossil occurrences, and Google Papers. When normalised for time (Table 8, part 3), there are highly significant correlations with PaleoDB formations, PaleoDB collections, and Google papers, and a significant correlation with PaleoDB occurrences. First-differenced data (Table 8, part 4) shows significant correlations only between ETD formation count and Macrostrat fossil occurrences and Web of Science publications, and generalised-differenced data show significant correlations with PaleoDB formations and Macrostrat fossil occurrences. When corrected for multiple comparisons, most of these apparent correlations disappear, and only the highly significant correlations between time-corrected ETD formation count and PaleoDB formations and PaleoDB collections remain (Table 8).

All other metrics were compared amongst each other, and results are listed in the Supplementary Information. After adjustment for multiple comparisons, there are highly significant correlations between PaleoDB formations, PaleoDB collections, and PaleoDB occurrences (and transformed versions). The same is true for Macrostrat units, columns, areas, fossil unit, fossil collections, and fossil occurrences (and transformed versions), but none of these North American measures correlates with any of the PaleoDB metrics, which are global. Interestingly, the NW European map area metrics from Smith and McGowan (2008) correlate highly significantly with Macrostrat units, columns, fossil units, fossil collections, and fossil occurrences, and significantly with Macrostrat map areas; this is true also for log-transformed data, and the time-standardised NW European map areas correlate highly significantly with Macrostrat fossil units/collections/occurrences, but the first-differenced data show no relationships. The NW European map area metrics do not correlate with any of the other global metrics, such as ETD formations, all PaleoDB metrics, and counts of papers. Finally, and not unexpectedly, the counts of papers from Google-Scholar and from Web of Science correlate highly significantly with each other, both as raw measures, and in all transformations.

5.5. Multiple linear models

In the multiple linear modelling, the response variable was tetrapod diversity (or 'amphibian' diversity or amniote diversity), and putative explanatory variables were the ETD formation counts, the various PaleoDB and Macrostrat metrics, European map areas, and publication counts. Modelling was applied to 20 data sets (Table 9), the five variants of data considered against the coarse time scale, and 15 variants of data considered against the fine time scale (5 variants considered for each of 'amphibians', amniotes, and tetrapods).

The first overview of the linear modelling exercise (Table 9) presents summaries of all explanatory variables, those that were selected for the model, and the ranking of the others in terms of goodness of fit. In the forward step modelling, this represents the order as variables were added to the model, and in the backward step modelling, this is the inverse of the order in which variables were eliminated.

For the coarse time scale, models consisted of one to seven explanatory variables (mean, 3.3; Table 9, part A). By far the most informative explanatory character is ETD formation count, which featured in all models. Next comes European map area, which featured in three of the five models, and PaleoDB collections which features in two. The order of goodness of fit for the others varies according to the data treatment, but PaleoDB formations, Macrostrat units, Macrostrat fossil collections, Google papers, and Web of Science papers were all ranked poorly across several modelling exercises.

Table 9
Results of multiple linear modelling, for 20 data runs. Results are presented for both the coarse time scale (Part A), with the data transformed in various ways (runs 1–5), and for the fine time scale (Part B), with the data transformed in various ways, and for 'amphibians'. Amniotes, and all tetrapods in each case (runs 1A–5C). In initial runs against the coarse time scale, the counts of terrestrial, marine, and total formations were interdependent, and so one count (terrestrial formations) was disregarded in the linear modelling. Because the counts of marine formations were modest, and inconsequential for many time bins, these data columns were excluded from most of the models. Abbreviations: Amn., amniotes; Amph., amphibians; diff., differenced; gen., generalised; NA, not applicable; Tet., tetrapods; transf., transformed.

Explanatory variables	ETD Formations	PaleoDB Formations	PaleoDB Collections	PaleoDB Occurrences	Macrostrat Units	Macrostrat Columns	Macrostrat Area	Macrostrat Fossilif. units	Macrostrat Fossil collections	Macrostrat Fossil occurences	European Map area	Papers Google-Scholar	Papers WoS
Number in file	1	2	3	4	5	6	7	8	9	10	11	12	13
A. Coarse time scale													
1. Raw data	Model	11th	4th	7th	10th	8th	9th	6th	13th	5th	Model	12th	3rd
2. Log-transformed	Model	11th	Model	8th	10th	6th	7th	5th	13th	9th	Model	12th	Model
3. By time	Model	8th	9th	Model	Model	Model	10th	Model	11th	Model	Model	13th	12th
4. First-differenced	Model	6th	Model	9th	4th	7th	8th	12th	5th	11th	3rd	10th	13th
5. Gen.-differenced	Model	11th	3rd	6th	10 h	8th	7th	5th	4th	13th	2nd	9th	12th

B. Fine time scale

Explanatory variables	Formations All	Formations Marine	Formations Terrestrial	Completeness Amph.	Completeness Amniotes	Completeness Tetrapods	Fossil ratio Amph.	Fossil ratio Amniotes	Fossil ratio Tetrapods	Number of variables in model
Number in file		1	2	3	4	5	6	7	8	9
1A. Raw data—Amph.	4th	3rd	NA	6th	Model	7th	5th	8th	Model	2
1B. Raw data—Amn.	Model	8th	NA	7th	Model	6th	Model	Model	Model	5
1C. Raw data—Tet.	Model	6th	NA	7th	Model	8th	Model	Model	Model	5
2A. Log-transf.—Amph.	Model	Excluded	Excluded	Model	6th	7th	Model	Model	Model	5
2B. Log-transf.—Amn.	Model	Excluded	Excluded	5th	Model	Model	7th	6th	Model	4
2C. Log-transf.—Tet.	Model	Excluded	Excluded	Model	Model	7th	Model	5th	6th	4
3A. By time—Amph.	Model	Excluded	Excluded	Model	Model	Model	7th	Model	Model	6
3B. By time—Amn.	Model	Excluded	Excluded	6th	Model	5th	7th	4th	Model	3
3C. By time—Tet.	Model	Excluded	Excluded	Model	4th	3rd	6th	5th	7th	2
4A. First-diff.—Amph.	Model	Excluded	Excluded	Model	7th	Model	Model	Model	Model	6
4B. First-diff.—Amn.	Model	Excluded	Excluded	7th	Model	Model	Model	Model	6th	5
4C. First-diff.—Tet.	Model	Excluded	Excluded	Model	Model	7th	Model	Model	Model	6
5A. Gen.-diff.—Amph.	6th	5th	Model	Model	Model	8th	Model	9th	7th	4
5B. Gen.-diff.—Amn.	8th	Model	Model	Model	Model	Model	9th	Model	7th	6
5C. Gen.-diff.—Tet.	8th	Model	Model	9th	Model	Model	Model	Model	7th	6

For the fine time scale, models consisted of two to six explanatory variables (mean 4.6; Table 9, part B), suggesting rather less clarity in identifying dominant explanatory variables than with the coarse time scale study. None of the explanatory variables was included in all models for all data transformations, but the most frequent in sequence were Completeness amniotes (12), ETD formations (11), Fossil ratio amphibians (9), Fossil ratio amniotes (9), Fossil ratio tetrapods (9), and Completeness amphibians (8), with the number of models in which each variable participates in parentheses.

The summary of the models (Table 10) confirms these results. For the coarse time scale (Table 10, part A), ETD formations represented a highly significant explanatory variable ($p<0.0001$) in all cases. Next in sequence were Macrostrat fossil occurrences for the time-standardised data (Table 10, part A3), European map areas for the raw data (Table 10, part A1), PaleoDB collections for the log-transformed data (Table 10, part 1B), European map areas for the time-standardised data (Table 10, part A3), Macrostrat fossiliferous units and PaleoDB occurrences for the time-standardised data (Table 10, part A3), and European map areas and Web of Science papers for the log-transformed data (Table 10, part A2).

For the fine time scale, the results (Table 10, parts B1A–B5C) are more mixed, with two rather dominant explanatory variables. Fossil completeness of amniotes proved to be a highly significant or significant explanatory variable in all but the log-transformed and first-differenced data for amphibians, and the time-standardised data for tetrapods (Table 10, parts B2A, B3C, B4A), and ETD formations being a highly significant explanatory variable in all but the raw data for amphibians and in all three generalised-differenced runs (Table 10, parts B1A, B5A–C). As noted above, the three variants of the Fossil completeness ratio metric were also significant or highly significant explanatory variables in the majority of runs.

6. Discussion

6.1. Do sampling proxies measure sampling?

Sampling, in statistical terms, refers to procedures that seek to use a subset of individuals from within a population to estimate characteristics of the whole population. To be valid, the subsample must be an unbiased representative of the whole, and this can be determined when the whole population is known. In the present context, palaeontologists have long understood (Darwin, 1859; Raup, 1972) that they have access to only a very small sample of the fossil record, and that the fossil record falls far short of a correct and unbiased view of all life that ever existed (Foote, 2001, 2003; Smith, 2007a; Benton et al., 2011; Smith and McGowan, 2011). Further, it is evident that the fossil record as a whole is biased towards the preservation of shallow marine organisms with hard skeletons (Valentine, 1969; Raup, 1972). Whether these rather obvious statements then mean that the fossil record is biased along every axis (temporal, latitudinal, environmental, taxonomic, body size) and so is incapable of yielding much biological signal without thorough investigation and correction, or whether it might be *adequate* (Paul, 1998) for certain constrained studies, is a matter of real concern for palaeobiologists.

The known fossil record is limited by a number of geological and human factors (Raup, 1972) that may be referred to 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 ease of access to it; and (3) *effort*, human factors, such as geographical location and subject interest (by age, location, or fossil group). Some would distinguish geological bias from sampling, using the latter term to refer only to human factors.

Sampling proxies, or sampling metrics, such as formation counts, map areas, measures of specimen completeness, or counts of publications are intended to reflect aspects of geological bias and human factors. Further, and evidently, those sampling proxies ought to be *independent of the signal they seek to correct*: if they are not, then any covariation between response variable (e.g. palaeodiversity time series) and explanatory variable (e.g. formation count or map area) may simply reflect redundancy between the two signals (Benton et al., 2011). Redundancy could arise from the fundamentals of the two variables, in that they both share common numerical components, or from a shared response in nature to a third variable, the common cause hypothesis of Peters (2005). In either case, covariation of response and explanatory variables need not automatically prove bias and the independence of the preferred sampling metric becomes an issue for further consideration.

The problem in using putative sampling metrics is that their covariation with the palaeodiversity signal can reflect either evidence for bias or for redundancy and/or the common cause model (Peters, 2005; Smith, 2007a; Benson and Butler, 2011; Benton et al., 2011; Hannisdal, 2011). Distinguishing these two explanations has often been difficult. The impasse can be breached in one of three ways: (1) by constructing a plausible primary argument for how the putative sampling proxy causes bias and cannot be explained by common cause or redundancy; (2) by using statistical methods that not only detect correlation, but also indicate causation (i.e. directionality from cause to effect), as has been done by the use of information transfer statistical methods (Hannisdal, 2011; Hannisdal and Peters, 2011); or (3) by using comparative modelling approaches in which the palaeodiversity signal is compared with multiple signals of causation (e.g. sea level, oxygen, productivity) with and without the putative sampling proxy. Many authors (e.g. Benton et al., 2011; Hannisdal and Peters, 2011; Benson and Mannion, 2012) have recommended the modelling approach, but this suffers from a lack of decisiveness: model comparisons can work only with the models presented, and the true model may not be included. Nonetheless, some such studies (e.g. Marx and Uhen, 2010) suggest that sampling is not a major factor, whilst others (e.g. Benson and Butler, 2011; Benson and Mannion, 2012) suggest that it is.

6.2. Formation counts

Time series of formation counts have been widely used as sampling proxies. Peters and Foote (2001, p. 587) suggested that a formation count might be a good sampling proxy because it 'reflects both the areal extent of sedimentary rock and the total thickness and lithologic heterogeneity (i.e., environmental diversity) captured by the stratigraphic record. The amount of research conducted on a region may also be correlated with the number of named formations… The number of formations may thus provide a joint measure of the quantity of the record (Raup, 1976), research effort (Sheehan, 1977), and lithologic variability.' These assumptions, and their tractability, lie behind their wide use as universal sampling proxies (e.g. Peters and Foote, 2001, 2002; Fröbisch, 2008; Barrett et al., 2009; Butler et al., 2009, 2011; Benson et al., 2010; Benson and Butler, 2011; Mannion et al., 2011; Upchurch et al., 2011; Benson and Mannion, 2012; Lloyd, 2012), and it has been assumed to be the decisive *causal argument*.

Others, however, find the linkage between formation count and sampling hard to justify. For example, Crampton et al. (2003, p. 359) stated that, in their studies, 'the number of named formations is a poor proxy for outcrop area, whereas outcrop area is a comparatively robust proxy for collection effort.' These and other authors (Wignall and Benton, 1999; Peters, 2006; Smith, 2007a; Peters and Heim, 2010; Benton et al., 2011; Dunhill, 2011, 2012) have provided six reasons why formation counts are problematic: (1) their definitions are arbitrary and vary enormously by stratigraphic age, geography, and environment; (2) they do not generally correlate with rock volume or accessibility; (3) they do not necessarily correlate with collection effort; (4) they reflect rock heterogeneity; (5) they are partly redundant with palaeodiversity, both of which reflect rock

Table 10
Results of multiple linear modelling, for 20 data runs. Results are presented for both the coarse time scale (Part A), with the data transformed in various ways (runs 1–5), and for the fine time scale (Part B), with the data transformed in various ways, and for 'amphibians'. Amniotes, and all tetrapods in each case (runs 1A–5C). Data runs are numbered according to the scheme in Table 9. Under each heading, the intercept and the explanatory variables retained in the model are listed, with their coefficient estimate, standard error, t-value, and probability (P). Significance levels are indicated as *** $P<0.001$, ** $P<0.01$, * $P<0.05$. Abbreviations: amn., amniotes; amph., amphibinas; foss., fossil; fossilif., fossiliferous; tet., tetrapods.

	Estimate	Standard error	t-value	P
A1. Coarse time scale—raw data				
Intercept	9.8260	10.5367	0.933	0.3596
Formations	3.7789	0.3864	9.780	<0.0001***
European map areas	−0.1600	0.0544	−2.940	0.0068**
A2. Coarse time scale—log-transformed data				
Intercept	0.0293	0.1904	0.154	0.8790
Formations	1.3928	0.1007	13.829	<0.0001***
PaleoDB collections	0.2585	0.0884	2.924	0.0074**
European map areas	0.1968	0.0793	−2.482	0.0204*
Papers—Web of Science	−0.1730	0.0754	−2.294	0.0308*
A3. Coarse time scale—time-standardised data				
Intercept	0.5918	0.4840	1.223	0.2350
Formations	0.1464	0.0229	6.392	<0.0001***
PaleoDB occurrences	0.0157	0.0065	2.423	0.0245*
Macrostrat units	−0.3566	0.2527	−1.411	0.1729
Macrostrat columns	0.7269	0.3823	1.901	0.0711
Macrostrat fossilif. units	−2.2906	0.8526	−2.687	0.0138*
Macrostrat foss. occurrences	0.0190	0.0044	4.291	0.0003***
European map areas	0.0653	0.0227	2.873	0.0091**
A4. Coarse time scale—first-differenced data				
Intercept	−0.4261	6.2454	−0.068	0.946
Formations	2.6223	0.5486	4.780	<0.0001***
PaleoDB collections	0.1652	0.0975	1.694	0.102
A5. Coarse time scale—generalised-differenced data				
Intercept	1.5773	5.4473	0.290	0.774
Formations	2.9702	0.4784	6.208	<0.0001***
B1A. Fine time scale—raw data/amphibians				
Intercept	−0.9582	3.7242	−0257	0.7979
Fossil completeness—amn.	3.7727	1.4045	2.686	0.0096**
Fossil ratio—tetrapods	9.5419	5.8336	1.636	0.1077
B1B. Fine time scale—raw data/amniotes				
Intercept	−1.5616	5.8990	−0.265	0.7923
Formations—all	2.3771	0.3239	7.339	<0.0001***
Fossil completeness—amn.	−39.8019	8.3269	−4.780	<0.0001***
Fossil ratio—amphibians	20.6941	10.3693	1.996	0.0513
Fossil ratio—amniotes	156.0857	28.8085	5.418	<0.0001***
Fossil ratio—tetrapods	−31.4889	13.8071	−2.281	0.0268*
B1C. Fine time scale—raw data/tetrapods				
Intercept	−2.6474	4.5281	−0.585	0.5614
Formations—all	2.4639	0.2486	9.910	<0.0001***
Fossil completeness—amn.	−39.3183	6.3918	−6.151	<0.0001***
Fossil ratio—amphibians	22.0733	7.9596	2.773	0.0077**
Fossil ratio—amniotes	166.5512	22.1137	7.532	<0.0001***
Fossil ratio—tetrapods	−23.9827	10.5985	−2.263	0.0279*
B2A. Fine time scale—log-transformed data/amphibians				
Intercept	0.0295	0.1153	0.256	0.7989
Formations—all	0.5409	0.1434	3.772	0.0004***
Fossil completeness—amph.	0.9858	0.2386	4.132	0.0001***
Fossil ratio—amphibians	−2.9436	0.5237	−5.621	<0.0001***
Fossil ratio—amniotes	−1.4330	0.8666	−1.654	0.104367
Fossil ratio—tetrapods	4.2359	0.9258	4.576	<0.0001***
B2B. Fine time scale—log-transformed data/amniotes				
Intercept	−0.0908	0.1373	−0.661	0.5115
Formations—all	0.8336	0.2047	4.073	0.0002***
Fossil completeness—amn.	1.9522	0.4193	4.656	<0.0001***
Fossil completeness—tet.	−0.7515	0.3753	−2.002	0.0505
Fossil ratio—tetrapods	−0.9305	0.6689	−1.391	0.1702
B2C. Fine time scale—log-transformed data/tetrapods				
Intercept	−0.0250	0.0605	−0.413	0.6813
Formations—all	0.9534	0.0910	10.481	<0.0001***
Fossil completeness—amph.	0.3632	0.1278	2.842	0.0064**
Fossil completeness—amn.	1.1552	0.1939	5.958	<0.0001***
Fossil ratio—amphibians	−0.3813	0.2338	−1.631	0.1089
B3A. Fine time scale—time-standardised data/amphibians				
Intercept	−1.0231	1.0643	−0.961	0.341056
Formations—all	1.0315	0.1673	6.167	<0.0001***
Fossil completeness—amph.	6.9577	1.5069	4.617	<0.0001***
Fossil completeness—amn.	−222.7693	126.1657	−1.766	0.0836
Fossil completeness—tet.	−1.7898	0.6412	−2.791	0.0074**
Fossil ratio—amniotes	−18.4402	5.0283	−3.667	0.0006***
Fossil ratio—tetrapods	4.0505	2.4643	1.644	0.1065
B3B. Fine time scale—time-standardised data/amniotes				
Intercept	−1.4909	1.6749	−0.89	0.3774
Formations—all	2.2519	0.2878	7.825	<0.0001***
Fossil completeness—amn.	541.434	224.1181	2.416	0.0192*
Fossil ratio—tetrapods	−6.8358	4.8718	−1.403	0.1664
B3C. Fine time scale—time-standardised data/tetrapods				
Intercept	−2.4066	1.6407	−1.467	0.1482
Formations—all	3.7052	0.2254	16.439	<0.0001***
Fossil completeness—amph.	4.1131	2.1107	1.949	0.0565
B4A. Fine time scale—first-differenced data/amphibians				
Intercept	−0.4094	1.1725	−0.349	0.7284
Formations—all	0.5874	0.1939	3.03	0.0039**
Fossil completeness—amph.	−8.0546	3.6897	−2.183	0.0338*
Fossil completeness—tet.	8.6825	5.1689	1.68	0.0992
Fossil ratio—amphibians	34.3824	13.7205	2.506	0.0155*
Fossil ratio—amniotes	23.7932	5.0523	4.709	<0.0001***
Fossil ratio—tetrapods	−34.2717	17.7792	−1.928	0.0596
B4B. Fine time scale—first-differenced data/amniotes				
Intercept	−0.1273	2.4008	−0.053	0.9579
Formations—all	1.9096	0.395	4.834	<0.0001***
Fossil completeness—amn.	−28.3726	9.7773	−2.902	0.0055**
Fossil completeness—tet.	−6.4761	3.3454	−1.936	0.0584
Fossil ratio—amphibians	21.784	8.0099	2.72	0.0089**
Fossil ratio—amniotes	100.2451	35.5004	2.824	0.0068**
B4C. Fine time scale—first-differenced data/tetrapods				
Intercept	−0.6225	2.3929	−0.26	0.7958
Formations—all	2.506	0.3946	6.351	<0.0001***
Fossil completeness—amph.	−11.5952	5.7594	−2.013	0.0495*
Fossil completeness—amn.	−27.7783	10.1483	−2.737	0.0086**
Fossil ratio—amphibians	67.1083	22.2774	3.012	0.0041**
Fossil ratio—amniotes	121.7469	37.152	3.277	0.0019**
Fossil ratio—tetrapods	−24.7236	11.9942	−2.061	0.0445*
B5A. Fine time scale—generalised-differenced data/amphibians				
Intercept	−0.1458	1.017	−0.143	0.8866
Formations—terrestrial	0.637	0.1988	3.204	0.0023**
Fossil completeness—amph.	−5.4036	3.038	−1.779	0.0812
Fossil completeness—amn.	7.1662	1.4122	5.074	<0.0001***
Fossil ratio—amphibians	22.0618	10.5573	2.09	0.0416*
B5B. Fine time scale—generalised-differenced data/amniotes				
Intercept	0.6069	1.9996	0.304	0.7628
Formations—marine	1.8184	1.1149	1.631	0.1093
Formations—terrestrial	1.5856	0.4258	3.724	0.0005***
Fossil completeness—amph.	6.2445	2.4716	2.527	0.0148*
Fossil completeness—amn.	−36.0125	8.9693	−4.015	0.0002***
Fossil completeness—tet.	−7.42	3.5884	−2.068	0.0440*
Fossil ratio—amniotes	123.3122	31.431	3.923	0.0003***
B5C. Fine time scale—generalised-differenced data/tetrapods				
Intercept	0.5001	1.9744	0.253	0.8011
Formations—marine	1.9027	1.0997	1.73	0.0899
Formations—terrestrial	2.2458	0.4202	5.344	<0.0001***
Fossil completeness—amn.	−27.0756	8.8303	−3.066	0.0035**
Fossil completeness—tet.	−8.1264	3.5921	−2.262	0.0281*
Fossil ratio—amphibians	28.5827	8.5687	3.336	0.0016**
Fossil ratio—amniotes	117.4295	30.9233	3.797	0.0004***

heterogeneity; and (6) they vary in volume over at least eight orders of magnitude.

These are all issues of concern, but as Upchurch et al. (2011) say, abundant dinosaurs do not always mean that more formations are named. The point of the redundancy argument is the close linkage between the discovery of formations and dinosaurs through time (Benton, 2008a). As more formations (indicated as the roughly synonymous 'basins' here) are found, so too are more dinosaurs (Fig. 7). The discovery signals for formations and dinosaurs run in parallel, and the more formations that are explored, the more dinosaur taxa are discovered. Note that the term 'explored' here does not distinguish between new formations identified by palaeontological field campaigns and

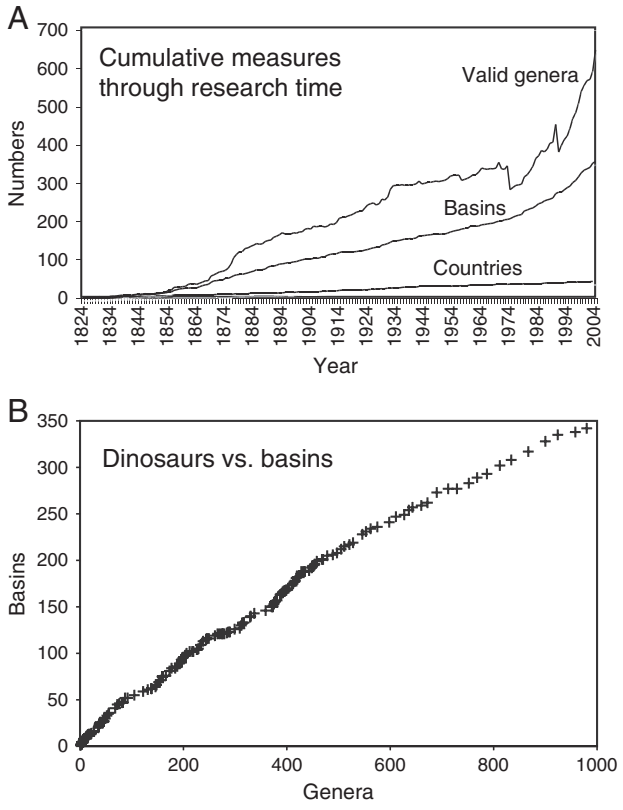

Fig. 7. The relationship between new basins (≈formations) and discovery of valid new genera. Data here is for dinosaurs, based on Benton (2008a), but similar relationships presumably apply for early tetrapods. As researchers added new countries and basins to the roster from which dinosaurs had been found, new genera accumulated (A). The best correlation for cumulative new genera or species is with new basins ($y = 0.600x - 1.114$; $r2 = 0.972$), suggesting that formation count is an excellent predictor of taxon count.

those identified by chance by survey geologists, or indeed by trawling museum drawers.

The parallel discovery of formations and dinosaurs may sound like a perfect primary explanation of why formations make a good sampling proxy, but it is not—it is a statement of redundancy. If each formation is likely to yield 1–10 new dinosaurian species, then the global roster of dinosaurs at any time depends on the energy and thoroughness of palaeontologists in exploring new regions and revealing their dinosaurs. As the relationship between number of formations and number of dinosaurs remains roughly constant through time (13 formations and 11 genera in 1850; 98 and 173 in 1900; 168 and 307 in 1950; 335 and 557 in 2000; Benton, 2008a), which is the signal and which the sampling proxy? Dividing dinosaur count by formation count gives a constant figure through research time (after 1850; 0.85; 1.76, 1.83, 1.66), so dinosaurs were evidently as well sampled in 1900 as they are today, according to the formations vs. taxa assumption. When two time series, such as DBFs and dinosaur species, vary through research time in lockstep, they represent linked metrics of effort. It cannot be declared that one is an independent yardstick of the other.

There are two kinds of dinosaur discovery: either in a known formation or in a new formation. Evidently, we are still finding many dinosaurs in new formations, and perhaps this close lock-step relationship between DBFs and dinosaurian taxa will continue until all possible DBFs have been identified. Once all possible DBFs have been identified, then all new dinosaurian taxa will come from known DBFs and these additional taxa (providing they are valid new taxa) will represent a combination of search effort and initial diversity in each formation.

In cartoon form (Fig. 8), comparing two redundant signals means that it is impossible to distinguish sampling from diversity. A poorly sampled, but diverse world (Fig. 8B) looks just the same as a well sampled but rare world (Fig. 8C). Sampling can be assessed in terms of the cumulative acquisition of new taxa per stratigraphic formation, the standard collector curve, but this is entirely missed in a raw global or regional DBF count. To globalise such sampling data would require specimen counts and locality counts to show how intensively sampled particular formations, basins, or regions might be.

Upchurch and Barrett (2005), Benton (2010), and Benton et al. (2011) argued that, if formation counts are to be used, they should be as broad as possible to allow for all possible sampling opportunities, including poorly identifiable remains and even no remains at all, and this has been the view of most practitioners (e.g. Lloyd et al., 2008; Benson et al., 2010; Benson and Butler, 2011; Butler et al., 2011, 2012). Clearly, 'all possible sampling opportunities' covers appropriate sedimentary formations that may have yielded any kind of fossil bone, and would exclude entirely barren redbeds or igneous rocks. However, this point has been debated (Barrett et al., 2009; Upchurch et al., 2011).

In statistical or ecological sampling, absence of evidence (i.e. non discovery) might be evidence of absence. The analyst ought to record both successes and failures in sampling, not ignore the fails. For example, ecologists throw their quadrats over richly and sparsely populated areas when mapping the occurrence of species, because they must know when the taxa of interest are abundant, rare, or non-existent. Absent taxa, indicated by null quadrat returns, are recorded as absent, not ignored. The opposite case has been made by Barrett et al. (2009, p. 2668), who argued in favour of using a strict dinosaur-bearing formation (DBF) count, because 'this is a more credible subset of formations to use than the total number of formations available (similarly, modern ecological surveys do not expend significant search effort in habitats unfrequented by the target group of organisms), and the use of DBF provides some taphonomic control. Moreover, it has been demonstrated that relationships between the rock record and diversity are not strongly affected if unfossiliferous formations are also taken into account.' However, as Benton et al. (2011) argued, ignoring non-occurrences of dinosaurs increases the risk of redundancy between palaeodiversity and formation count time series. As an example, it could be the case that true dinosaurian diversity varied over an order of magnitude between adjoining time bins, perhaps as a result of major climatic or topographic change. All other things being equal (i.e. probability of fossilisation of each skeleton, mean formation rock volume, human effort), the strict DBF count is also likely to vary over an order of magnitude as fewer dinosaur fossils mean fewer DBF. The analyst then cannot distinguish whether the order-of-magnitude drop in apparent dinosaurian diversity was real or not. Cross-testing across the plunge in apparent palaeodiversity and in DBF could include assessing whether the probability of fossilisation of dinosaurian skeletons had reduced (look at ghost ranges and Lazarus taxa, mean specimen quality, and mean likelihood of occurrence through time), whether there are major changes in rock volume represented by named geological formations, or whether human effort is the driver (Smith, 2007a,b; Benton et al., 2011). Variable likelihood of occurrence through time of a target fossil group, such as dinosaurs, requires a wider than strict DBF, such as a count of all-tetrapod formations. This would highlight times when suitable rocks occurred but no dinosaurs were found.

Our ETD formation counts include only the named rock units that have produced tetrapod remains, and they exclude formations that produced unidentified tetrapod fossils as well as those that produced none at all. Therefore, these metrics suffer all the criticisms just noted, and correlations (Tables 6, 7) should be treated with extreme caution. Further, the ETD formation count does not correlate well with other plausible formation counts, such as those from the PaleoDB (correlation only for time-adjusted measures) or Macrostrat (Table 8). This is

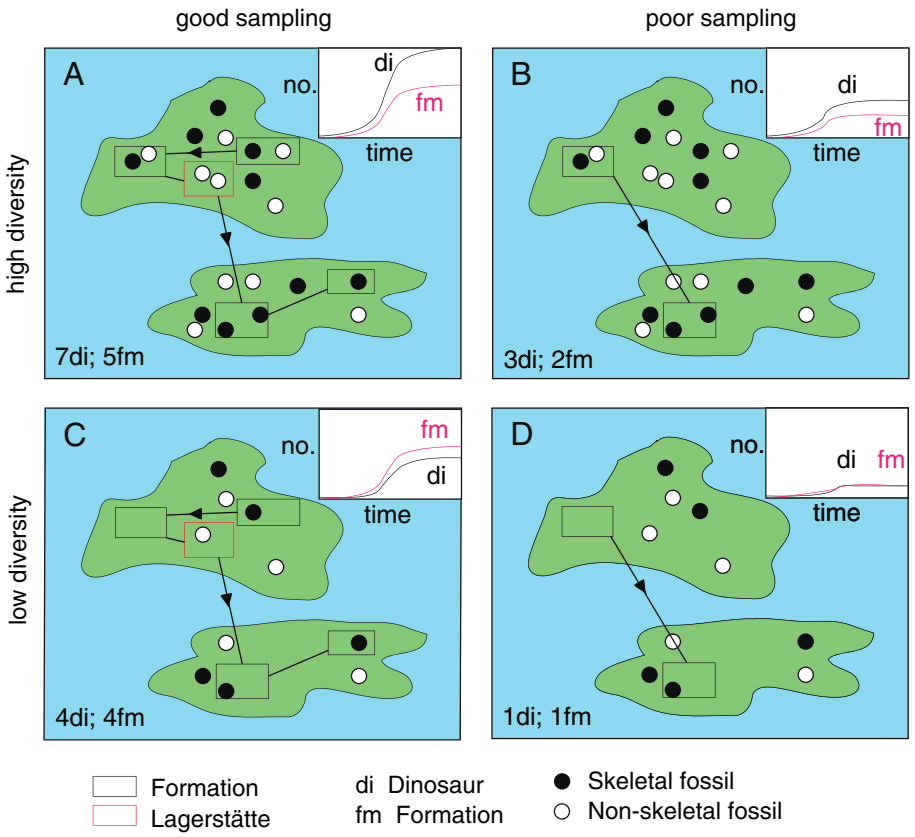

Fig. 8. Exploration of why formation count and dinosaur taxon count are so intimately linked. An imaginary world is explored, and dinosaur-bearing formations (DBFs; boxes) are identified sequentially through research time (search track indicated by arrows). The model assumes equal abundance of specimens in each formation, and formations are visualised as equal-sized exposures of rock (neither of which is true in nature). Dinosaurs (circles) are distributed over the landmasses, and these are encountered from time to time; some are normal skeletal fossils, others are more subtle traces only found in special conditions (Lagerstätten). Over research time, formations (fm) and dinosaurian taxa (di) accumulate following a sigmoid (collector) curve (top right). Four conditions are imagined, two where dinosaurs occur commonly (A, B) and two where specimens in the rocks are rare (C, D). These are intersected by two search regimes, one representing good sampling (A, C) and one poor (B, D). In all cases, formation counts and dinosaurian diversity are closely linked, and it is not possible to determine whether low and high values represent either poor or good sampling or initial low or high diversity. The close covariance of DBFs and taxon counts could indicate either close linkage between sampling and discovery rate or between initial diversity and discovery rate: sampling and initial diversity cannot be distinguished because the formation count and palaeodiversity counts are intimately interlinked, or redundant, and a strict 'formation count' cannot be used as an independent sampling proxy.

not unexpected because each formation count represents a very different mix of aspects of rock heterogeneity, regional bias, and arbitrariness—there is sampling signal buried in the various formation counts, but it is perhaps overprinted by all the other variables.

The linear modelling study gave mixed messages. For the coarse time scale, the ETD formations count dominated all models (Tables 9, 10), and European map area was a convincing correlate in several. However, when the data were considered at finer time resolution, the models became larger, with typically 4–6 explanatory variables, and mixed messages about the relative importance of the ETD formation count and the various skeletal completeness metrics (Tables 9, 10). Future work, in which a variety of putative environmental drivers will also be included in the model comparison, may provide clearer results.

6.3. Rock volume and area

Geologically minded analysts (e.g. Raup, 1972; Crampton et al., 2003; Peters, 2005, 2008; Smith, 2007a; Smith and McGowan, 2008; Wall et al., 2009; Peters and Heim, 2010; Hannisdal and Peters, 2011; Heim and Peters, 2011; Wall et al., 2011) have generally been in favour of assessing rock volume and rock accessibility by other means. Two current approaches are to use either map areas (= outcrop areas), intended to represent rock accessibility (Smith, 2001, 2007a; Smith and McGowan, 2008; Wall et al., 2009, 2011), or gap-bounded rock packages, the approach of the Macrostrat project (Peters, 2005, 2008; Peters and Heim, 2010; Hannisdal and Peters, 2011; Heim and Peters, 2011). Some more direct measures of rock area and rock volume are now possible using Geographic Information Systems (GIS), especially for regional-scale studies. These include exact measurements of outcrop areas, exposure areas, mean unit thicknesses, and calculated unit-scale rock volumes, as used by Dunhill et al. (in press-b), and they are surely preferable to rough tallies of numbers of maps or generalised areas taken from global- and continent-scale published maps. These more precise metrics of rock volume and accessibility derive from new online databases such as Macrostrat and georeferenced geological maps, as well as data on locations of quarries and coastal exposures, and georeferenced named fossil sites. When these are combined with GIS software, they provide an order-of-magnitude improvement in the potential of this field.

In reality, simple rock volume and area measures are likely to be overprinted by sediment heterogeneity, such as switches from nonmarine to marine deposition. In applying a variety of sampling metrics to the British Triassic, Dunhill et al. (in press-a) found that proxies for sedimentary rock volume and accessibility did not correlate with palaeodiversity until the removal of facies-related preservational and palaeoecological factors. So, for example, changes between mudstone-dominated and sandstone-dominated continental redbed sediments had huge effects on the preservation and

potential to collect fossils. Even more striking of course is the rapid switch from continental redbeds to shelf-marine sediments at the Rhaetic transgression. The Dunhill et al. (in press-a) study has shown that a weak sampling signal may be present, but the effects of changing palaeoenvironments are far more important.

All these approaches go to the fundamentals of the rock record and its effects on sampling (Raup, 1972), and they avoid the arbitrariness and potential redundancy of formation counts (Crampton et al., 2003; Smith, 2007a; Benton et al., 2011). The *causal argument* is that palaeontologists can only find fossils if the rocks are there, or at least accessible, and the opportunities for sampling from any time bin and facies can therefore range from zero to substantial (Raup, 1972).

This is not to say that there is a guaranteed and easy geological metric of sampling. The analyst has to choose between using poor-quality global signals derived from synoptic geological maps, with all their vagaries (Wall et al., 2009, 2011), or the criticism that a more precise regional map area data set cannot be guaranteed to represent global-scale sampling (Smith and McGowan, 2011). Further, outcrop area (= map area) and exposure area (= rock accessibility) do not always correlate: in some of the most intensively palaeontologically sampled areas of Europe and North America, for example, bedrock is so obscured by Pleistocene deposits, soil and human activity, that there is no relationship at all (Dunhill, 2011, 2012). Fossils are found on coastlines and in quarries, and map areas do not predict the proportional scaling of these resources: therefore there is no demonstrable causal link between the two, and covariation of map areas and palaeodiversity is possibly a result of the common cause model than sampling bias—more rock means more map area, but also indicates more habitat in the past, especially in the case of shallow marine organisms (Peters and Heim, 2010; Benton et al., 2011; Dunhill, 2012; Dunhill et al., in press-a, in press-b).

In our study, we found almost no evidence for correlation among the various geological proxies: the only significant correlation was for PaleoDB formations/ collections with the ETD formations count after time-normalisation (Table 8). The regionally-based metrics from the Macrostrat database (North America) did not correlate with the NW European map areas (Smith and McGowan, 2008), and the latter did not correlate with the ETD formation counts. We did not include a global-scale outcrop measure, whether from the Ronov maps (Ronov, 1994) or the more recent data of Wall et al. (2009, 2011), because these were compiled in broader time bins than those we used for the ETD data, and so we cannot say whether either the North American or European rock volume/area metrics correlate with the global-scale data.

6.4. Specimen completeness and quality

Surprisingly little used in considerations of the quality of the fossil record have been the quality and completeness of specimens themselves. For common fossils, this may be a meaningless measure, but for rarer fossils, such as vertebrates, it might be helpful to know whether some time bins or formations are characterised by predominantly complete or incomplete skeletons, and whether the fossils are in good condition or have been altered by taphonomic and diagenetic processes. At an extreme, the occurrence of one or more Lagerstätten within a time bin ought to provide a dramatic peak in knowledge when compared to neighbouring time bins without Lagerstätten (e.g. Butler et al., 2009, 2012). The *causal argument* for the use of metrics of fossil completeness or quality is similar to that for rock volume or accessibility, that the opportunities to identify species and genera depend on the quality and completeness of specimens. However, it should be noted that this could work both ways: large numbers of fragmentary specimens might cause palaeontologists either to undercount because they are behaving conservatively, or to exaggerate the species count by optimistically naming fragmentary specimens that cannot be definitively synonymised (Benton, 2008a,b; Mannion and Upchurch, 2010).

Completeness of fossils has been used in several studies as a metric of the quality of the fossil record, including Benton et al. (2004) on tetrapods across the Permo-Triassic boundary, Fountaine et al. (2005) and Brocklehurst et al. (2012) on Mesozoic birds, Smith (2007b) on Triassic-Jurassic echinoids, Benton (2008a,b) on dinosaurs, and Mannion and Upchurch (2010) on sauropodomorph dinosaurs.

Benton et al. (2004) documented quality classes of fossils, as well as numbers of specimens and numbers of localities (= individual find spots) for tetrapods across the Permo-Triassic boundary in Russia, and they found that the time of apparently very low diversity in the five successive stratigraphic divisions of the Early Triassic was characterised by unusually high levels of sampling, as documented by numbers of specimens and localities. However, the quality of specimens collected through this time span was uniformly poor. Using a metric of fossil completeness, based on the allocation of fossils to four completeness categories, ranging from isolated remains to multiple complete skeletons (as in the current study), they showed that completeness and apparent diversity did not covary in the Permian, but they did so in the Triassic (Benton et al., 2011).

Smith (2007b) assessed the completeness of echinoid specimens in the Triassic and Jurassic, and found a major evolutionary improvement in the robustness of the skeleton, and so also in sampling. The effect, as with our tetrapod studies, is direct: scrappy remains (whether isolated echinoid skeletal debris or isolated vertebrate teeth and ribs) are nearly impossible to identify, whereas more complete fossils can be assigned to genera and species. An abundance of scrappy fossils leads to undersampling of species, not by the absence of fossils, but by our inability to identify what we have in front of us.

Mannion and Upchurch (2010) argued that their Skeletal completeness metric (SCM) could function as a useful sampling proxy because it records times when data on skeletons is sufficient for species-level determination versus times when the SCM of all materials is low and so the specimens might be non-diagnostic. However, these authors note that the SCM records only the proportions of bones present in skeletons, and does not report the numbers of specimens or of taxa, and so it is not a comprehensive sampling proxy. Information is required both on the mean quality or completeness of specimens as well as their quantity.

The absence of correlation between the fossil completeness metrics and formation counts used in our study (Fig. 2B, Tables 6, 7) confirms points made by Mannion and Upchurch (2010), that both metrics assess different aspects of the data, and so may report different aspects of sampling bias. A direct causal link can be made between specimen completeness and identified diversity count because low-quality and incomplete specimens cannot be named, whereas complete specimens can be named. Therefore, unlike formation counts and outcrop areas, where such causal links are harder to make, fossil completeness may actually record a useful aspect of sampling bias. There could be a partial explanation by common cause: perhaps sea level change, particularly regression, destroys rocks and habitats in coastal areas (Smith, 2007a), but in terrestrial sediments, the reasons for incompleteness are hard to link to more widely operating drivers. Whether fossils are retrieved from a lake or overbank deposit, and hence are complete, or from a channel lag, and hence are disarticulated and abraded, probably depends more on local circumstances than on any wider pattern of sea level or global climate change. Nonetheless, such facies differences may dominate particular time bins—for example channel lags and poor tetrapod fossils in the Early Triassic of Russia—and so may bias perceptions of palaeodiversity.

6.5. Human effort

For a time, human effort was seen as potentially an excellent metric of bias. Raup (1972) noted the enormously variable number of palaeontologists and of publications on different geological periods (e.g. Cambrian vs. Miocene), and suggested that this demonstrated the

causal link from palaeontological interest to reported knowledge. Sheehan (1977) provided the data, and yet Raup (1977) pointed out that 'systematists follow the fossils'. The causal link is highly plausible, but the directionality is questioned: palaeontologists do not search for fossils evenly over the landscape, or evenly through geological time, but they focus their efforts around known rich fossil sites. Further, palaeontologists undoubtedly bias their work by focusing on taxa and times of particular evolutionary importance, so producing peaks in numbers of papers and in fossil diversity at key times (e.g. origin of tetrapods, origin of birds, times before and after mass extinction events). Therefore, metrics of human effort covary with palaeodiversity time series because of massive redundancy of the two signals (Benton et al., 2011), except perhaps in the case of some microfossil groups where commercial palaeontologists search doggedly through every rock horizon within boreholes, and so demonstrably apply uniform effort equably to both fossil-rich and fossil-poor time intervals.

Systematic effort depends on areal exposure of rocks (Sheehan, 1977), and so map area proxies relate closely to metrics of human effort rather than reflecting a direct control of apparent diversity by rock area. Further, metrics of human effort provide the best matches with palaeodiversity curves in local-scale studies. For example, Dunhill et al. (in press-b) explored all potential sampling metrics in an intensive study of one of the best-sampled classic palaeontological areas in the world, the Lower Jurassic of Dorset, SW England. Here, geologists and palaeontologists have been tramping over the primarily coastal exposures for 200 years, and many hundreds of thousands of fossil specimens have been collected, and many hundreds of papers have been published. Dunhill et al. (in press-b) found that the proxies for rock volume and accessibility did not correlate well with either other sampling proxies, or with apparent diversity, suggesting that the total amount of sedimentary rock preserved does not influence apparent diversity at a local scale. However, they did find some correlations between apparent diversity and proxies for worker effort. The fact that the proxies did not correlate significantly with each other suggests that none can be regarded as an all-encompassing sampling proxy that covers all aspects of bias. Further, the presence of some correlations between sampling proxies and diversity most probably indicates bonanza (Raup, 1977), as palaeontologists have preferentially sampled the richest rock units.

In the ETD study here, the lack of correlation between counts of publications and observed palaeodiversity (Fig. 5) is surprising because these signals are commonly found to covary strongly (e.g. Sheehan, 1977; Purnell and Donoghue, 2005; Butler et al., 2011). The absence of correlation could arise from the nature of this data set. Unlike the Dorset coast study (Dunhill et al., in press-b), where the fossil supply dominates collecting focus and the literature, palaeontologists who work on early tetrapods have devoted varying amounts of effort to known fossils in proportion to their perceived relative phylogenetic significance: for example, the first tetrapods or the first amniotes receive excessive attention, to the extent that any linkage between overall effort, number of finds, and richness of the record is lost. It could also be that the chosen metrics of effort, counts of published papers, are themselves incomplete or biased samples.

The other metrics of effort included in the ETD study are the counts of PaleoDB collections and occurrences – although these are unfiltered and include all fossils, both tetrapods and others. PaleoDB collections document species lists from published localities, and so incorporate a large element of 'human effort'. In our study, these PaleoDB metrics generally did not correlate with tetrapod palaeodiversity (Table 7, part 5), except when corrected for time bin duration, so there is a hint of a linkage between effort and diversity here, and this would clearly improve if the collections and occurrence data were filtered to include only collections with tetrapods or with continental taxa.

6.6. Correction of diversity curves

If a suitable, comprehensive sampling proxy could be determined, then it could be used to correct a palaeodiversity curve. This was first suggested by Raup (1976), who used residuals from his measures of rock volume to calculate a 'sampling-free' curve for the global-scale diversification of marine animals. In separate studies, this idea was taken further by Smith and McGowan (2008) and Wall et al. (2009) who used, respectively, NW European map areas and global map areas to correct the global palaeodiversity curve of marine animals. The argument is not necessarily that the 'corrected' curve is perfect, simply that a particular bias, namely variable outcrop area, has been removed. Residuals above or below the mean line are interpreted as times when palaeodiversity was higher or lower than predicted from rock volume and accessibility metrics and so might be said to exhibit some true biodiversity signal and so deserve additional explanation. The idea was extended to correct the overall terrestrial fossil record by Wall et al. (2011), who used generalised global map areas, and the dinosaurian fossil record by Lloyd et al. (2008), who used locality counts, and by Barrett et al. (2009), Mannion et al. (2011), and Lloyd (2012), who used formation counts. In all cases, the 'corrected' palaeodiversity signal was understood to lie closer to the truth than the uncorrected time series. However, as Butler et al. (2011, p. 1165) note, 'if [the common cause hypothesis] is true, then attempts to 'correct' palaeodiversity curves may actually distort genuine palaeodiversity signals.'

The assumption that the residuals (= deviations from the sampling-corrected curve) indicate higher- or lower-than-expected diversity (e.g. Smith, 2001; Smith and McGowan, 2008; Barrett et al., 2009; Lloyd, 2012; Lloyd and Friedman, in press) is of course true only if the sampling proxy really is a sampling proxy. However, if a strict formation count is employed, for the reasons noted above (Section 6.2), then positive residuals could indicate one of three things, or a combination of these: (1) unusually high global diversity; (2) unusually good preservation (whether associated with Lagerstätten or not), or (3) intense worldwide elevated human effort. Likewise, a negative residual could indicate (1) unusually low global diversity, (2) the absence of any Lagerstätten or geologically favourable sites for preservation, or (3) diminished human search effort, or some combination of these factors.

The claim to be able to correct a palaeodiversity time series by the use of a sampling proxy depends on a series of assumptions, primarily (1) that a global sampling metric can be identified that overrides regional effects, (2) that the sampling metric is independent of (= not redundant with) the response variable, and (3) that the common cause does not prevail.

The first key issue is basin-scale regional variability, and yet this has been generally ignored. Crampton et al. (2003, p. 360) noted: 'It is inappropriate to assume a single predictor or correction for the rock volume bias, such as a sea-level curve, across all regions, or to group data from different tectonic regimes within a single analysis.' In other words, patterns of sediment supply and basin deepening vary from basin to basin, and regional effects of latitude, climate, and topography vary in different ways over geographic areas. Each basin may show a very different sedimentary log through time, with rare correlations of specific horizons that reflect wider-scale phenomena such as tsunamis or volcanic ashfalls. It is unlikely then that all these variables can be combined and represented by a single sampling metric (Alroy, 2010b). Earlier attempts to determine a true global species palaeodiversity pattern by modelling multiple sampling biases (variable rock area and effort) together (e.g. Signor, 1982, 1985) were criticised (Sepkoski, 1994) for their sensitivity to modest changes in assumptions about collecting, namely that palaeontologists collect fossils evenly over available rock areas and do not particularly seek out rare fossils. Further, when continental- and global-scale are compared, wide differences may emerge, and such studies (e.g. Upchurch et al., 2011) highlight the inevitable problems of comparing continental and global palaeodiversity and sampling proxy time series, as we have also found in comparing

the global ETD palaeodiversity curve with European-only or North American-only rock volume metrics.

The second issue, potential redundancy is discussed by Benton et al. (2011) and above (Sections 6.1–6.5).

The third requisite for confident use of sampling proxies as correcting factors is that the common cause model does not prevail. This, however, is very hard to exclude, and increasing evidence now strengthens the likelihood of close linkage between aspects of change on the Earth's surface and the diversity of life, in some major settings at least (Peters, 2005; Benson and Butler, 2011; Hannisdal and Peters, 2011). This is discussed further in Section 6.7.

If formation count is to be used as the key sampling metric, then rises should correspond to favourable facies that permitted unusually good fossil preservation (including Lagerstätten), as well as a high input of human effort. Highlighting Lagerstätten as positive residuals, as is especially the case for the pterosaur fossil record for example (Butler et al., 2009; Benton et al., 2011; Butler et al., 2012, 2013) does not tell us that pterosaurs or birds or dinosaurs were unusually globally diverse at the times of the Solnhofen Limestone Formation and the Yixian Formation. These issues are all recognised in model-based approaches that discriminate Lagerstätten occurrence and that compare formation counts with other sampling proxies (e.g. Marx and Uhen, 2010; Benson and Butler, 2011; Lloyd et al., 2011; Benson and Mannion, 2012; Butler et al., 2012, 2013; Dunhill et al., in press-a, in press-b; Lloyd and Friedman, in press).

The conclusion about attempts to correct palaeodiversity curves for sampling may be that there is no reliable one-hit sampling proxy for global-scale studies, a point made by others (e.g. Raup, 1976; Crampton et al., 2003; Alroy, 2010b; Benton et al., 2011). This means that a sampling-corrected curve of the diversification of life may not be possible at global scale, or at least would require a complex evaluation of numerous potential drivers that might vary in their relative influences through geological time.

6.7. Common cause models

Until recently, the debate for and against the common cause model remained open, with suggestive evidence pointing in both directions (e.g. Peters, 2005; Smith, 2007a). For example, Peters and Heim (2011) showed that last fossil occurrences are concentrated towards the tops of gap-bounded sedimentary packages, whereas first appearances are more evenly distributed; this suggests that rock units and species often terminate together, but that species originate anywhere through the thickness of a unit. Originations then are unrelated to facies changes, but extinctions might be, and the relationships between marine bedrock area, lithofacies diversity (heterogeneity), and recorded species diversity change during the transgressive and regressive phases in the 80 Myr sea-level cycle (Smith and Benson, in press). Hannisdal and Peters (2011) have shown by the use of information transfer statistics, which provide evidence for directionality, that there are causal links between continental flooding, sulfur and carbon cycling, and macroevolution, and that covariation of marine rock and fossil signals arises from the common cause. Marine shelf diversity is best predicted by a combination of bedrock area and number of lithofacies units (Smith and Benson, in press).

A terrestrial common cause is harder to frame. Butler et al. (2011) found little support that dinosaurian palaeodiversity was driven by sea level change, confirming earlier results from empirical work on small terrestrial vertebrates by Fara (2002). However, the Butler et al. (2011) study has not rejected a terrestrial common cause, just that dinosaurian diversity probably was not driven by sea level change or changes in estimated land area. As Butler et al. (2011) argue, sea level could affect terrestrial diversity in opposing ways: high sea level could reduce terrestrial biodiversity by reducing the overall global land area, or it could increase terrestrial biodiversity by splitting land masses and generating endemicity.

An interesting recent suggestion (Benson and Butler, 2011) is that diversity fluctuations among marine shelf animals might be driven to a large extent by sea level change (the common cause), whereas the deep sea (Lloyd et al., 2011) and terrestrial fossil records might be much more subject to the vagaries of sampling. The deep oceanic realm is rather depauperate, whereas sampling by means of boreholes can be remarkably productive. On the other hand, the terrestrial realm was initially unpopulated, but terrestrial biodiversity apparently overtook marine biodiversity at the time of the Cretaceous Terrestrial Revolution (KTR, 100–125 Myr ago; Lloyd et al., 2008; Vermeij and Grosberg, 2011) and rose since then to represent 85–95% of all diversity on Earth today; and yet sampling has been incomplete because of the sporadic nature of continental sedimentary deposition.

An interesting conundrum in this contrast in explanations for palaeodiversity fluctuations, common cause (shelf) vs. sampling bias (deep sea; terrestrial), is whether such a distinction is real or not, or whether it reflects the nature of the systems and our capacity to model them. For example, it might be easier to detect a common cause model on the shelf because ecosystems are predominantly driven by sea level change (Peters, 2005) whereas it is harder to formulate a meaningful physical environmental model that drives oceanic and terrestrial systems. The deep ocean faunas are a combination of plankton, nekton, and abyssal benthos, and there is no reason that temporal variations in sea level, oxygen, carbon dioxide, sulfur—either singly or in combination—would provide a common cause. Indeed, deep-sea faunas are sparse and might be subject to local- and regional-scale drivers peculiar to particular water masses. On land, it is unlikely that sea level drives biodiversity (Fara, 2002; Butler et al., 2011), and it would be hard to model an intricate combination of the more likely environmental drivers of terrestrial biodiversity (Benton, 2010) such as topographic diversity, continentality, latitudinal temperature gradients, relative latitudinal land areas, palaeotemperatures, oxygen levels, and carbon dioxide levels. However, our difficulties in modelling many intersecting variables that may be dominated by regional-scale effects, does not mean there is no model.

7. Conclusions and prospects

The quest to understand the role of sampling in palaeodiversity estimation is not mere navel-gazing for palaeontologists. It is important to know whether fossil data can provide useful information on the shape of the history of life both for studies of global change and for understanding the origins of modern biodiversity. Are we any closer to resolving this long-running debate than when Raup (1972) discriminated between biological and geological explanations for the global palaeodiversity signal?

One approach has been sampling standardisation (SS), pioneered by Alroy et al. (2001, 2008) and Alroy (2010a), and presented as a means of correcting global palaeodiversity signals. It is likely, however, that SS methods fail because they depend on an assumption that the number of collections in any geological time bin, and the reported diversity of taxa within each collection, both represent measures of sampling. An alternative is that large samples and abundant samples within a time bin arise from high initial diversity and abundance. As with the formations vs. palaeodiversity comparisons (see Section 6.2), peaks and troughs in collection numbers and sizes might reflect fluctuations in sampling (preservation and human effort) or diversity. If the last is even partially correct, then SS procedures would truncate genuine peaks and troughs in diversity. Inevitably, a time of genuinely low diversity, perhaps following a mass extinction, will yield few collections, and fossils may be sparse within those collections (Wignall and Benton, 1999; Bush and Bambach, 2004; Smith and McGowan, 2011), so to downsample neighbouring richer time bins removes a true signal. Further, Hannisdal and Peters (2011, p. 1123) found that 'the number of fossil collections may carry a significant environmental signal related to changes in sea level… that may be removed by some sampling-standardization techniques', evidence for a common cause driver and a good reason to be cautious

about the meaning of SS-corrected global palaeodiversity curves (e.g. Alroy et al., 2001, 2008; Alroy, 2010a).

It may also be difficult to correct global palaeodiversity time series by the use of sampling metrics. Unless a sampling metric captures the complexities of multiple intersecting regional-scale biases (differential preservation, facies variation, differing geological histories, variable human effort), and distinguishes them from potential environmental drivers of diversity (topography, sea level, climate, atmospheric composition, palaeotemperature) it will remain uncertain exactly what the sampling-corrected curve, or the residuals, really mean. Fruitful ways forward may include regional-scale studies, model-based studies of particular clades, and comparisons of cladistic and fossil distribution data.

At regional scale, such as a sedimentary basin or province, it becomes possible to focus on direct measures of sampling, such as numbers of fossiliferous localities or numbers of collections (Alroy et al., 2001; Benton et al., 2004; Peters and Heim, 2010; Benton et al., 2011; Dunhill, 2012; Dunhill et al., in press-b), rather than continent-wide or global sampling proxies. Matching geographic scales between palaeodiversity signal and sampling proxy is rather evident (Benton et al., 2011; Smith and McGowan, 2011; Upchurch et al., 2011). Studies of individual clades may also circumvent difficulties of conflicting signals that may confuse broader-scale studies, at least in cases where the organisms occupied comparable habitats and were preserved in limited sedimentary facies, and where environmental impacts might be constrained according to their biology (e.g. Smith, 2007a; Lloyd et al., 2012).

Model-based studies (e.g. Marx and Uhen, 2010; Benson and Butler, 2011; Lloyd et al., 2011; Benson and Mannion, 2012; Butler et al., 2012, 2013; Lloyd and Friedman, in press) permit evaluation of numerous potential explanatory variables, including both sampling proxies and environmental drivers (and some that might be either or both). Such studies may be most informative when conducted at regional scale or for a particular clade because model construction is likely to be more reliable than for 'all life' or 'all marine life'. Directional statistical methods (e.g. Hannisdal and Peters, 2011) offer an exciting addition to such approaches in discriminating causality in cases of covariation.

For tetrapods, phylogenetically corrected diversity estimates have been widely recommended and used (e.g. Barrett et al., 2009; Benson et al., 2010; Benton et al., 2011), especially because of the ready availability of well-tested and scrutinised cladograms. This is an important next step for the ETD study. Once the relevant supertree/composite tree is compiled, it will be fascinating to determine how ghost ranges are distributed, how they change apparent palaeodiversity signals, and what, if anything, they indicate about relative sampling levels in different time bins.

Acknowledgements

First, we thank Roger Benson, Phil Mannion, and Richard Butler for organising the 2011 Society of Vertebrate Paleontology symposium, "Vertebrate Diversity Patterns and Sampling Bias", and this special volume. We thank Graeme Lloyd (University of Oxford) for his generalised differencing R code. We are also grateful to the referees for their challenging critiques of the first version of this paper, and to Roger Benson, Richard Butler, Graeme Lloyd, Al McGowan, Andrew Smith, and Paul Upchurch for some intense debates about the redundancy arguments we present here. This work was supported by NERC grant NE/C518973/1 to M.J.B.

Appendix A

R codes, written by Manabu Sakamoto, to carry out calculations for generalised differencing, pairwise correlations, and multiple linear modelling, in batch mode. All require .csv (comma-separated value) files that can be constructed and saved in Excel.

(1) Generalised differencing ('gd.R')

Data can be entered manually into R, or supplied as a .csv file in which the first column is midpoint ages of the time bins, and the remaining columns are data time series. The program generates GDs of every column following, in sequence, using Graeme Lloyd's R function gen.diff (http://www.graemetlloyd.com/methgd.html). Note also that gen.diff assesses each time series for significant trend against a least squares regression, and warns the user if such a trend does not exist.

```
gd <- function(x){
X <- x[,-1]
    t <- x[,1]
M <- matrix(nrow = length(X[,1])-1,ncol = length(X[1,]))
for(i in 1:length(X[1,])){
    G <- gen.diff(X[,i],t)
    M[,i] <- G
    }
colnames(M) <- colnames(x)[-1]
return(M)
}
gen.diff <- function(x,time)
{
#if(cor.test(time,x)$p.value > 0.05) print("Warning: variables not
significantly correlated, generalised differencing not recommended")
dt      <-x-((lsfit(time,x)$coefficients[2]*time)+lsfit(time,x)
$coefficients[1])
m<-lsfit(dt[1:(length(dt)-1)],dt[2:length(dt)])$coefficients[2]
gendiffs<-dt[1:(length(dt)-1)]-(dt[2:length(dt)]*m)
gendiffs
}
```

(2) Pairwise correlation ('pair.cor')

This program performs multiple pairwise correlation tests, allowing a range of standard methods (Pearson, Spearman, Kendall) as well as, importantly, correction for multiple comparisons (p.adj.method = Holm, Hochberg, Hommel, Bonferroni, False discovery rate, etc.)

```
pair.cor <- function(x, test = c("pearson", "kendall", "spearman"),
p.adj.method = c("holm", "hochberg", "hommel", "bonferroni",
"BH", "BY", "fdr", "none")){
mat <- as.matrix(x)
n <- length(mat[1,])
N <- n*(n-1)/2
cor.stats <- matrix(nrow = length(mat[1,]),ncol = length(mat
[1,]))
p.value <- matrix(nrow = length(mat[1,]),ncol = length(mat
[1,]))
p.adj <- matrix(nrow = length(mat[1,]),ncol = length(mat[1,]))
for( i in 1:length(mat[1,]) ){
    v1 <- mat[,i]
    cor <- vector()
    p <- vector()
    padj <- vector()
    for( j in 1:length(mat[1,]) ){
        v2 <- mat[,j]
        COR <- cor.test(v1,v2, method = test)
        cor <- append(cor,COR$estimate)
        p <- append(p,COR$p.value)
        padj <- append(padj,p.adjust(p[j], method = p.adj.method,
n = N ))
        }
```

```
    cor.stats[i,] <− cor
    p.value[i,] <− p
    p.adj[i,] <− padj
    colnames(cor.stats) <−colnames(mat)
    rownames(cor.stats) <−colnames(mat)
    colnames(p.value) <−colnames(mat)
    rownames(p.value) <−colnames(mat)
    colnames(p.adj) <−colnames(mat)
    rownames(p.adj) <−colnames(mat)
    }
result <− list(stats = cor.stats, p.values = p.value, p.adjust =
p.adj)
return(result)
    }
```

(3) Multiple linear modelling ('mlm.R')

This program compares a number of explanatory variables (x1->) with one or more response variables (y1 ->), calculates the goodness of fit of a model consisting of all explanatory variables (the 'lm' routine), and then compares this (the 'step' routine) with the fit of models consisting of variable combinations from one to all explanatory variables, determining the best model (in terms of overall adjusted coefficient of determination, r^2). The mlm.step method can work by forward selection, backward elimination, or by both methods (direction = "forward", "backward", "both"). In the example below, the data file, in .csv format, is 'filename', the response variable is 'y', and the explanatory variables, x1 to x13.

```
y <− filename[,c(1)]
x1 <− filename[,c(2)]
x2 <− filename[,c(3)]
x3 <− filename[,c(4)]
x4 <− filename[,c(5)]
x5 <− filename[,c(6)]
x6 <− filename[,c(7)]
x7 <− filename[,c(8)]
x8 <− filename[,c(9)]
x9 <− filename[,c(10)]
x10 <− filename[,c(11)]
x11 <− filename[,c(12)]
x12 <− filename[,c(13)]
x13 <− filename[,c(14)]
mlm1 <− lm(y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9
+ x10 + x11 + x12 + x13)
summary(mlm1)
mlm.step <− step (mlm1, direction = "both")
summary(mlm.step)
```

Appendix B. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.palaeo.2012.09.005.

References

Abdala, F., Hancox, P., Neveling, J., 2005. Cynodonts from the uppermost Burgersdorp Formation, South Africa, and their bearing on the biostratigraphy and correlation of the Triassic *Cynognathus* assemblage zone. Journal of Vertebrate Paleontology 25, 192–199.

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

Alroy, J., 2010b. Geographical, environmental and intrinsic biotic controls on Phanerozoic marine diversification. Palaeontology 53, 1211–1235.

Alroy, J., Marshall, C.R., Bambach, R.K., Bezusko, K., Foote, M., Fürsich, F.T., Hansen, T.A., Holland, S.M., Ivany, L.C., Jablonski, D., Jacobs, D.K., Jones, D.C., Kosnik, M.A., Lidgard, S., Low, S., Miller, A.I., Novack-Gottshall, P.M., Olzewski, T.D., Patzkowsky, M.E., Raup, D.M., Roy, K., Sepkoski Jr., J.J., Sommers, M.G., Wagner, P.J., Webber, A., 2001. Effects of sampling standardization on estimates of Phanerozoic marine diversification. Proceedings of the National Academy of Sciences of the United States of America 98, 6261–6266.

Alroy, J., Aberhan, M., Bottjer, D.J., Foote, M., Fürsich, F.T., Harries, P.J., Hendy, A.J.W., Holland, S.M., Ivany, L.C., Kiessling, W., Kosnik, M.A., Marshall, C.R., McGowan, A.J., Miller, A.I., Olzewski, T.D., Patzkowsky, M.E., Peters, S.E., Villier, L., Wagner, P.J., Bonuso, N., Borkow, P.S., Brenneis, B., Clapham, M.E., Fall, L.M., Ferguson, C.A., Hanson, V.L., Krug, A.Z., Layou, K.M., Leckey, E.H., Nürnberg, S., Powers, C.M., Sessa, J.A., Simpson, C., Tomasovych, A., Visaggi, C.C., 2008. Phanerozoic trends in the global diversity of marine invertebrates. Science 321, 97–100.

Barrett, P.M., McGowan, A.J., Page, V., 2009. Dinosaur diversity and the rock record. Proceedings of the Royal Society of London. Series B 276, 2667–2674.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate—a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B: Methodological 57, 289–300.

Benson, R.B.J., Butler, R.J., 2011. Uncovering the diversification history of marine tetrapods: ecology influences the effect of geological sampling biases. In: McGowan, A.J., Smith, A.B. (Eds.), Comparing the Geological and Fossil Records: Implications for Biodiversity Studies. Geological Society, London, pp. 191–208.

Benson, R.B.J., Mannion, P.D., 2012. Multi-variate models are essential for understanding vertebrate diversification in deep time. Biology Letters 8, 127–130.

Benson, R.B.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 of London. Series B 277, 829–834.

Benton, M.J., 1983. Dinosaur success in the Triassic: a noncompetitive ecological model. The Quarterly Review of Biology 58, 29–55.

Benton, M.J., 1985. Mass extinction among non-marine tetrapods. Nature 316, 811–814.

Benton, M.J., 1993. Late Triassic extinctions and the origin of the dinosaurs. Science 260, 769–770.

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

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

Benton, M.J., 2008b. Fossil quality and naming dinosaurs. Biology Letters 4, 729–732.

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., 2010. The origins of modern biodiversity on land. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 365, 3667–3679.

Benton, M.J., 2012. No gap in the Middle Permian record of vertebrates. Geology 40, 339–342.

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., Dunhill, A.M., Lloyd, G.T., Marx, F.G., 2011. Assessing the quality of the fossil record: insights from vertebrates. In: McGowan, A.J., Smith, A.B. (Eds.), Comparing the geological and fossil records: implications for biodiversity studies. Geological Society, London, pp. 63–94.

Benton, M.J., Newell, A.J., Khlyupin, A.Y., Shumov, I.S., Price, G.D., Kurkin, A.A., 2012. Preservation of exceptional vertebrate assemblages in Middle Permian fluviolacustrine mudstones of Kotel'nich, Russia: stratigraphy, sedimentology, and taphonomy. Palaeogeography, Palaeoclimatology, Palaeoecology 319–320, 58–83.

Bond, D., Wignall, P., Wang, W., Izon, G., Jiang, H., Lai, X., Sun, Y., Newton, R., Shao, L., Vedrine, S., Cope, H., 2010. The mid-Capitanian (Middle Permian) mass extinction and carbon isotope record of South China. Palaeogeography, Palaeoclimatology, Palaeoecology 292, 282–294.

Brocklehurst, N., Upchurch, P., Mannion, P.D., O'Connor, J., 2012. The completeness of the fossil record of Mesozoic birds: implications for early avian evolution. PLoS One 7 (6), e39056. http://dx.doi.org/10.1371/journal.pone.0039056.

Brusatte, S.L., Benton, M.J., Ruta, M., Lloyd, G.T., 2008a. Superiority, competition, and opportunism in the evolutionary radiation of dinosaurs. Science 321, 1485–1488.

Brusatte, S.L., Benton, M.J., Ruta, M., Lloyd, G.T., 2008b. The first 50 Myr of dinosaur evolution: macroevolutionary pattern and morphological disparity. Biology Letters 4, 733–736.

Brusatte, S.L., Nesbitt, S.J., Irmis, R.B., Butler, R.J., Benton, M.J., Norell, M.A., 2010. The origin and early radiation of dinosaurs. Earth-Science Reviews 101, 68–100.

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, 367–382.

Bush, A.M., Bambach, R.K., 2004. Did alpha diversity increase during the Phanerozoic? Lifting the veils of taphonomic, latitudinal, and environmental biases. Journal of Geology 112, 625–642.

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 of London, Series B: Biological Sciences 278, 1165–1170.

Butler, R.J., Brusatte, S.L., Andres, B., Benson, R.B.J., 2012. How do geological sampling biases affect studies of morphological evolution in deep time? A case study of the Pterosauria (Reptilia: Archosauria). Evolution 66, 147–162.

Butler, R.J., Benson, R.B.J., Barrett, P.M., 2013. Pterosaur diversity: untangling the influence of sampling biases, Lagerstätten, and genuine biodiversity signals.

Palaeogeography, Palaeoclimatology, Palaeoecology. http://dx.doi.org/10.1016/j.palaeo.2012.08.012 online ahead of print.
Carroll, R., 2009. The Rise of Amphibians: 365 Million Years of Evolution. The Johns Hopkins University Press, Baltimore . 360 pp.
Chen, Z.Q., Benton, M.J., 2012. The timing and pattern of biotic recovery following the end-Permian mass extinction. Nature Geoscience 5, 375–383.
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.
Dunhill, A.M., 2011. Using remote sensing and a geographic information system to quantify rock exposure area in England and Wales: implications for paleodiversity studies. Geology 39, 111–114.
Dunhill, A.M., 2012. Problems with using rock outcrop area as a paleontological sampling proxy: rock outcrop and exposure area compared with coastal proximity, topography, land use, and lithology. Paleobiology 38, 126–143.
Dunhill, A.M., Benton, M.J., Newell, A.J., Twitchett, R.J., in press. Completeness of the fossil record and the validity of sampling proxies: a case study from the Triassic of England and Wales. Journal of the Geological Society 169.
Dunhill, A.M., Benton, M.J., Twitchett, R.J., Newell, A.J., in press. Completeness of the fossil record and the validity of sampling proxies at outcrop level. Palaeontology 55.
Fara, E., 2002. Sea-level variations and the quality of the continental fossil record. Journal of the Geological Society of London 159, 489–491.
Foote, M., 2001. Inferring temporal patterns of preservation, origination, and extinction from taxonomic survivorship analysis. Paleobiology 27, 602–630.
Foote, M., 2003. Origination and extinction through the Phanerozoic: a new approach. Journal of Geology 111, 125–148.
Fountaine, T.M.R., Benton, M.J., Dyke, G.J., Nudds, R.L., 2005. The quality of the fossil record of Mesozoic birds. Proceedings of the Royal Society of London. Series 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.
Gradstein, F., Ogg, J., Smith, A., 2004. A Geologic Time Scale 2004. Cambridge University Press, Cambridge.
Hannisdal, B., 2011. Detecting common-cause relationships with directional information transfer. In: McGowan, A.J., Smith, A.B. (Eds.), Comparing the Geological and Fossil Records: Implications for Biodiversity Studies. Geological Society, London, pp. 19–29.
Hannisdal, B., Peters, S.E., 2011. Phanerozoic Earth System Evolution and Marine Biodiversity. Science 334, 1121–1124.
Heim, N., Peters, S., 2011. Covariation in macrostratigraphic and macroevolutionary patterns in the marine record of North America. Geological Society of America Bulletin 123, 620–630.
Hounslow, M.W., Muttoni, G., 2010. The geomagnetic polarity timescale for the Triassic: linkage to stage boundary definitions. In: Lucas, S.G. (Ed.), The Triassic Time Scale. Geological Society, London, pp. 61–102.
Johnson, M.R., Hiller, N., 1990. Burgersdorp Formation. Catalogue of South African Lithostratigraphic Units 2–10, 1–2.
Kammerer, C.F., Angielczyk, K.D., Fröbisch, N.J., 2011. A comprehensive taxonomic revision of Dicynodon (Therapsida, Anomodontia) and its implications for dicynodont phylogeny, biogeography, and biostratigraphy. Journal of Vertebrate Paleontology 31, 1–158 (Supplement 1 31, Supplement 1).
Lloyd, G.T., 2012. A refined modelling approach to assess the influence of sampling on palaeobiodiversity curves: new support for declining Cretaceous dinosaur richness. Biology Letters 8, 123–126.
Lloyd, G.T., Friedman, M., in press. A survey of palaeontological sampling biases in fishes based on the Phanerozoic record of Great Britain. Palaeogeography, Palaeoclimatology, Palaeoecology. http://dx.doi.org/10.1016/j.palaeo.2012.07.023.
Lloyd, G.T., Davis, K.E., Pisani, D., Tarver, J.E., Ruta, M., Sakamoto, M., Hone, D.W.E., Jennings, R., Benton, M.J., 2008. Dinosaurs and the Cretaceous Terrestrial Revolution. Proceedings of the Royal Society of London. Series 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, pp. 167–177.
Lloyd, G.T., Pearson, P.N., Young, J.R., Smith, A.B., 2012. Sampling bias and the fossil record of planktonic foraminifera on land and in the deep sea. Paleobiology 38, 569–584.
Lozovskiy, V.R., Minikh, M.G., Grunt, T.A., Kukhtinov, D.A., Ponomarenko, A.G., Sukacheva, I.D., 2009. The Ufimian Stage of the East European scale: status, validity, and correlation potential. Stratigraphy and Geological Correlation 17, 602–614.
Lucas, S.G., 2004. A global hiatus in the Middle Permian tetrapod fossil record. Stratigraphy 1, 47–64.
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.
Marx, F.G., Uhen, M.D., 2010. Climate, critters, and cetaceans: Cenozoic drivers of the evolution of modern whales. Science 327, 993–996.
McGowan, A.J., Smith, A.B. (Eds.), 2011. Comparing the Geological and Fossil Records: Implications for Biodiversity Studies. Geological Society, London. 247 pp.
McKinney, M.L., 1990. Classifying and analysing evolutionary trends. In: McNamara, K.J. (Ed.), Evolutionary Trends. Belhaven Press, London, pp. 28–58.
Milner, A.R., 1990. The radiations of temnospondyl amphibians. In: Taylor, P.D., Larwood, G.P. (Eds.), Major Evolutionary Radiations. Clarendon Press, Oxford, pp. 321–349.
Muttoni, G., Kent, D.V., Olsen, P.E., Di Stefano, P., Lowrie, W., Bernasconi, S.M., Hernández, F.M., 2004. Tethyan magnetostratigraphy from Pizzo Mondelo (Sicily) and correlation to the Late Triassic Newark astrochronological polarity time scale. Bulletin of the Geological Society of America 116, 1043–1058.
Muttoni, G., Kent, D.V., Jadoul, F., Olsen, P.E., Rigo, M., Galli, M.T., Nicora, A., 2010. Rhaetian magneto-biostratigraphy from the Southern Alps (Italy): constraints on Triassic chronology. Palaeogeography, Palaeoclimatology, Palaeoecology 285, 1–16.
Newell, A.J., Tverdokhlebov, V.P., Benton, M.J., 1999. Interplay of tectonics and climate on a transverse fluvial system, Upper Permian, southern Uralian foreland basin, Russia. Sedimentary Geology 127, 11–29.
Ogg, J.G., Ogg, G., Gradstein, F.M., 2008. The Concise Geologic Time Scale. Cambridge University Press, Cambridge.
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, pp. 1–22.
Payne, J.L., Lehrmann, D.J., Wei, J.Y., Orchard, M.J., Schrag, D.P., Knoll, A.H., 2004. Large perturbations of the carbon cycle during recovery from the end-Permian extinction. Science 305, 506–509.
Peters, S.E., 2005. Geologic constraints on the macroevolutionary history of marine animals. Proceedings of the National Academy of Sciences of the United States of America 102, 12326–12331.
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.
Peters, S.E., Heim, N.A., 2011. Stratigraphic distribution of marine fossils in North America. Geology 39, 259–262.
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.
R Development Core Team, 2011. The R project for statistical computing, version 2.14.1.
Ramezani, J., Hoke, G., Fastovsky, D., Bowring, S., Therrien, F., Dworkin, S., Atchley, S., Nordt, L., 2011. High-precision U– Pb zircon geochronology of the Late Triassic Chinle Formation, Petrified Forest National Park (Arizona, USA): temporal constraints on the early evolution of dinosaurs. Geological Society of America Bulletin 123, 2142–2159.
Raup, D.M., 1972. Taxonomic diversity during the Phanerozoic. Science 177, 1065–1071.
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.
Reisz, R.R., Laurin, M., 2001. The reptile Macroleter: first vertebrate evidence for correlation of Upper Permian continental strata of North America and Russia. Geological Society of America Bulletin 113, 1229–1233.
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.
Rubidge, B.S., 2005. Re-uniting lost continents—fossil reptiles from the ancient Karoo and their wanderlust. South African Journal of Geology 108, 135–172.
Ruta, M., Benton, M.J., 2008. Calibrated diversity, tree topology and the mother of all mass extinctions: the lesson of the temnospondyls. Palaeontology 51, 1261–1288.
Sahney, S., Benton, M.J., Falcon-Lang, H.J., 2010. Rainforest collapse triggered Carboniferous tetrapod diversification in Euramerica. Geology 38, 1079–1082.
Sepkoski Jr., J.J., 1994. Limits to randomness in paleobiologic models: the case of Phanerozoic species diversity. Acta Palaeontologica Polonica 38, 175–198.
Sepkoski Jr., J.J., 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.
Signor, P.W., 1982. Species richness in the Phanerozoic: compensating for sampling bias. Geology 10, 625–628.
Signor, P.W., 1985. Real and apparent trends in species richness through time. In: Valentine, J.W. (Ed.), Phanerozoic Diversity Patterns; Profiles in Macroevolution. Princeton University Press, Princeton, New Jersey, pp. 129–150.
Smith, A.B., 2001. Large-scale heterogeneity of the fossil record: implications for Phanerozoic diversity studies. Philosophical Transactions of the Royal Society of London, Series B 356, 351–367.
Smith, A.B., 2007a. Marine diversity through the Phanerozoic: problems and prospects. Journal of the Geological Society 164, 731–745.
Smith, A.B., 2007b. 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., Benson, R.B.J., in press. Marine diversity in the geological record and its relationship to surviving bedrock area, lithofacies diversity, and original marine shelf area. Geology.
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.
Smith, A.B., McGowan, A.J., 2011. The ties linking rock and fossil records and why they are important for palaeobiodiversity studies. In: McGowan, A.J., Smith, A.B. (Eds.), Comparing the Geological and Fossil Records: Implications for Biodiversity Studies. Geological Society, London, pp. 1–7.

Stanley, S.M., 2007. An analysis of the history of marine animal diversity. Paleobiology 33 (Suppl. 4), 1–55.
Sues, H.-D., Reisz, R.R., 1998. Origins and early evolution of herbivory in tetrapods. Trends in Ecology & Evolution 13, 141–145.
Taylor, G.K., Tucker, C., Twitchett, R.J., Kearsey, T., Benton, M.J., Newell, A.J., Surkov, M.V., Tverdokhlebov, V.P., 2009. Magnetostratigraphy of Permian/Triassic boundary sequences in the Cis-Urals, Russia: no evidence for a major temporal hiatus. Earth and Planetary Science Letters 281, 36–47.
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, pp. 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: McGowan, A.J., Smith, A.B. (Eds.), Comparing the Geological and Fossil Records: Implications for Biodiversity Studies. Geological Society, London, pp. 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., 2011. The great divergence: when did diversity on land exceed that in the sea. Integrative and Comparative Biology 50, 675–682.
Wall, P.D., Ivany, L.C., Wilkinson, B.H., 2009. Revisiting Raup: exploring the influence of outcrop area on diversity in light of modern sample-standardization techniques. Paleobiology 35, 146–167.
Wall, P.D., Ivany, L.C., Wilkinson, B.H., 2011. Impact of outcrop area on estimates of Phanerozoic terrestrial biodiversity trends. In: McGowan, A.J., Smith, A.B. (Eds.), Comparing the Geological and Fossil Records: Implications for Biodiversity Studies. Geological Society, London, pp. 53–62.
Ward, P.D., Montgomery, D.R., Smith, R.M.H., 2000. Altered river morphology in South Africa related to the Permian–Triassic extinction. Science 289, 1741–1743.
Wignall, P.B., Benton, M.J., 1999. Lazarus taxa and fossil abundance at times of biotic crisis. Journal of the Geological Society 156, 453–456.