Palaeontology The Palaeontological Association www.palass.org

[Palaeontology, Vol. 60, Part 2, 2017, pp. 169–185]

BIAS IN PHYLOGENETIC MEASUREMENTS OF EXTINCTION AND A CASE STUDY OF END-PERMIAN TETRAPODS

by LAURA C. SOUL[1,2] *and* MATT FRIEDMAN[1,3]

[1]Department of Earth Sciences, University of Oxford, South Parks Road, Oxford, OX1 3AN, UK; soull@si.edu
[2]*Current address:* Department of Paleobiology, Smithsonian Institution National Museum of Natural History, [NHB, MRC 121], PO Box 37012, Washington, DC 20013-7012, USA
[3]*Current address:* Museum of Paleontology & Department of Earth & Environmental Science, University of Michigan, 1109 Geddes Ave, Ann Arbor, MI 48109-1079, USA

Typescript received 23 August 2016; accepted in revised form 17 December 2016

Abstract: Extinction risk in the modern world and extinction in the geological past are often linked to aspects of life history or other facets of biology that are phylogenetically conserved within clades. These links can result in phylogenetic clustering of extinction, a measurement comparable across different clades and time periods that can be made in the absence of detailed trait data. This phylogenetic approach is particularly suitable for vertebrate taxa, which often have fragmentary fossil records, but robust, cladistically-inferred trees. Here we use simulations to investigate the adequacy of measures of phylogenetic clustering of extinction when applied to phylogenies of fossil taxa while assuming a Brownian motion model of trait evolution. We characterize expected biases under a variety of evolutionary and analytical scenarios. Recovery of accurate estimates of extinction clustering depends heavily on the sampling rate, and results can be highly variable across topologies. Clustering is often underestimated at low sampling rates, whereas at high sampling rates it is always overestimated. Sampling rate dictates which cladogram timescaling method will produce the most accurate results, as well as how much of a bias ancestor–descendant pairs introduce. We illustrate this approach by applying two phylogenetic metrics of extinction clustering (Fritz and Purvis's *D* and Moran's *I*) to three tetrapod clades across an interval including the Permo-Triassic mass extinction event. These groups consistently show phylogenetic clustering of extinction, unrelated to change in other quantitative metrics such as taxonomic diversity or extinction intensity.

Key words: phylogenetic clustering, tetrapod, Permian–Triassic mass extinction, simulation.

COMPARISONS of palaeontological data on extinction from different time periods are complicated by profound contrasts in timescale, the volume and quality of available data, approaches to analysis, and the intensity with which different geographical areas and taxonomic groups have been studied (Jablonski 2008; Fritz *et al.* 2013; de Vos *et al.* 2014; Payne *et al.* 2016). These problems are especially acute for vertebrates, which are of considerable interest to biologists but have an incomplete palaeontological record in comparison to shelly marine invertebrates (Foote & Raup 1996; Foote & Sepkoski 1999). Despite these limitations, the fossil record can offer a natural laboratory for testing hypotheses about how extinction dynamics might change or be maintained in times of extreme ecological stress (Jablonski 1994, 2005; Finnegan *et al.* 2015). This deep-time perspective is becoming increasingly important to contemporary biological research as extinction rates increase and biodiversity declines (McKinney 1997; Erwin 2009; Barnosky *et al.* 2011).

Two approaches dominate studies of extinction: measuring selectivity with respect to different biological, life history or extrinsic traits (Bielby *et al.* 2006; Cardillo *et al.* 2008; Turvey & Fritz 2011; Harnik *et al.* 2012) and measuring extinction intensity and turnover rates. The latter has been the usual focus of quantitative analyses of extinction in the geological past (Raup 1994; Alroy 1996; Stanley 1998; Alroy *et al.* 2001, 2008; Jablonski 2008). Ideally, the fossil record might be used to identify traits which may make taxa vulnerable to extinction (Jackson & Erwin 2006; Purvis 2008; Fritz *et al.* 2013). Some high-resolution fossil records have indeed been used to investigate selection against a particular trait, or vulnerability to a particular pressure. Previous studies have shown extinction selectivity related to body size (Harnik 2011; Tomiya 2013), feeding strategy (Jeffery 2001), geographical range

doi: 10.1111/pala.12274

(Kiessling & Aberhan 2007; Payne & Finnegan 2007; Jablonski 2008), morphology (Liow 2007; Friedman 2009) and clade richness (Smith & Roy 2006), among others. Unfortunately even this basic level of trait data is not immediately accessible for much of the fossil record.

Phylogenetic approaches can lessen some of the biases introduced by imperfect sampling, while simultaneously providing results from different data and scales that can be directly compared across clades and through time (Purvis 2008; Fritz *et al.* 2013; Harnik *et al.* 2014). Many previous studies, focusing on a variety of different questions and methods, have demonstrated that application of phylogenetic data to study of the fossil record can be important in obtaining valid, statistically unbiased results (Felsenstein 1985; Grafen 1989; Norell 1992; Rabosky 2010; Pennell & Harmon 2013; Sakamoto *et al.* 2016). Studies of extinction can also be augmented by the incorporation of phylogeny, which provides additional information that cannot be accessed through taxonomic or stratigraphic approaches, or from measuring turnover rates alone (Hardy *et al.* 2012). For example, phylogenetic measurements of extinction can be used to find the presence or absence of taxon-independent selection against traits (Tomiya 2013), measure loss of evolutionary history (Huang *et al.* 2015), or understand the origin of phylogenetic community structure (Fraser *et al.* 2015).

Intuitively, we might expect that extinction is selective with respect to the relationships between taxa (i.e. phylogeny), given that some traits may make taxa vulnerable or resistant to extinction, and that these traits might be phylogenetically conserved (Hunt *et al.* 2005; Green *et al.* 2011; Smits 2015). In other words, due to their shared ancestry, closely related taxa are more likely to share similar characteristics, and the probability of a taxon becoming extinct might in turn be related to those characteristics (Fig. 1B). When this is the case the phylogenetic clustering of extinction (i.e. whether closely related taxa become extinct at the same time) might act as a proxy for selection for or against particular traits in the fossil record. This proxy could be studied in situations where a phylogeny is available, but detailed morphological or life history information is lacking. This approach broadly assumes that a Brownian motion-like model of trait evolution adequately reflects changes in the features that are relevant to extinction risk (Freckleton *et al.* 2002; Harmon *et al.* 2010). In such a case clustered extinction is indicative of selection with respect to phylogenetically conserved traits, whereas phylogenetically random extinction is indicative either of selection with respect to phylogenetically labile traits, or of extinction that is not selective with respect to particular traits (Fritz & Purvis 2010).

Although phylogenetic methods offer advantages over approaches based on taxonomy or extinction intensity, incorporating fossil taxa into phylogenies potentially introduces its own set of biases. For example, range

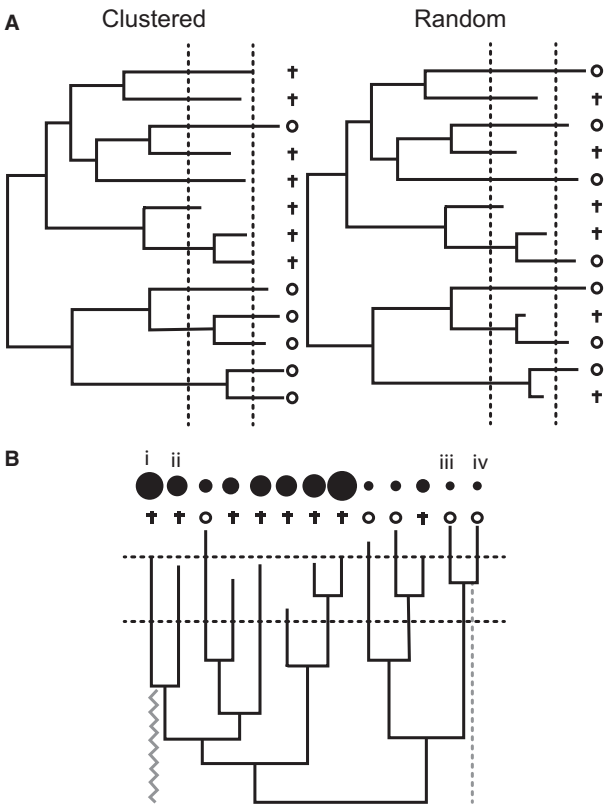

FIG. 1. Hypothetical phylogenies showing random and Brownian (clustered) expectations of extinction distributions across the tips. A, phylogenetically clustered extinction (left), and phylogenetically random extinction (right). The measurement is made for timeslices, shown by dashed lines. An extinction (cross) is any that occurs within that timeslice, a survival (open circle) is any taxon that survives past the end of the timeslice. B, extinctions and survivals represented as in A; size of filled circles represents the value of a continuous trait that has evolved under Brownian motion and that affects extinction probability (e.g. body size). The zig-zag grey line shows the shared evolutionary history between taxa i and ii, the dashed grey line shows the shared evolutionary history between taxa iii and iv. With a longer shared history and less time since diverging, iii and iv have closer values for this trait than do i and ii. In this example, large values of the trait increases extinction risk, shown by the higher proportion of extinctions in taxa with larger values. Brownian motion evolution of the trait generates clustering of similar values because of shared evolutionary history, and so generates a Brownian (clustered) distribution of extinctions.

extensions in a phylogeny are asymmetrical; they can predate fossil occurrences, thereby extending a taxon's range into the past, but the length of unsampled history after the last fossil occurrence of a taxon cannot easily be estimated. There have been studies on the effect on downstream analyses of several of the features that are more acute in phylogenies of fossil taxa than those of extant groups (e.g. uncertain divergence dates (Bapst 2014; Halliday & Goswami 2016), missing character data causing

tree misspecification (Stone 2011) and a higher proportion of soft polytomies (Garland & Diaz-Uriarte 1999; Housworth & Martins 2001; Davis *et al.* 2012)). However, the effect of the overall 'degraded' nature of a palaeontological phylogeny has not yet been fully investigated, particularly with respect to the phylogenetic structure of extinction.

Here we use simulations to examine the efficacy of Fritz and Purvis' *D* (Fritz & Purvis 2010; a metric of the clustering of binary traits across a phylogeny) when applied to phylogenies of fossil taxa to measure the phylogenetic clustering of extinction given evolution of relevant traits under a Brownian motion model of change. We investigate the ways in which results from this analysis of simulated fossil (i.e. degraded) data are biased with respect to true evolutionary patterns, and identify the likely causes of such bias. This provides a general guide for the use of these analyses on fossil data. We illustrate this approach to studying the clustering of extinction with an empirical example based on tetrapods during the Permian–Triassic mass extinction (PTME).

METHOD

All analyses were performed in R (v. 3.1.3; R Core Team 2015) using the packages *paleotree* (simulating palaeontological trees; Bapst 2012), *OUwie* (simulating traits; Beaulieu & O'Meara 2014) and *caper* (calculating clustering metrics; Orme *et al.* 2012).

Phylogenetic clustering of extinction

Both the simulation study and analysis of real data require measurement of the phylogenetic clustering of extinctions of lineages. Here we treat extinction and survival as a binary trait within a time bin (Fig. 1). There are several methods by which the phylogenetic or taxonomic clustering of a binary trait may be measured, but here we focus on Fritz and Purvis' *D* (Fritz & Purvis 2010). This metric is scaled to random and Brownian motion expectations of trait distribution. A random expectation is where extinctions and survivals are randomly scattered across the tips of the phylogeny within the time bin (Fig. 1A). The Brownian expectation is the pattern of extinctions and survivals across the tips that is obtained if a continuous trait evolves under a Brownian motion (random walk) model of evolution and is then converted into a binary trait using a threshold value. As outlined above, a longer shared ancestry means that under this model closely related taxa are more likely to have similar traits, leading to a pattern of clustering of the same trait values on the phylogeny (Fig. 1B).

The scaling of the test statistic *D* means that, unlike alternative metrics, it is robust to tree shape, tree size, and trait prevalence for trees containing more than 50 tips (Fritz & Purvis 2010). *D* can therefore be used to reliably compare values through time, and between clades, providing an advantage over other methods (Hardy *et al.* 2012). We also repeated all analyses on the real data using Moran's *I* (a test for spatial autocorrelation (Moran 1950) generalized for use to measure phylogenetic signal by Gittleman & Kot (1990)) to establish whether the same variation in extinction clustering through time was found with both measures.

D is calculated by scaling the observed sum of sister-clade differences (SSD) to sister-clade differences from 1000 iterations of Brownian and random models, using equation 1:

$$D = \frac{[\sum d_{\text{obs}} - \text{mean}(\sum d_{\text{b}})]}{[\text{mean}(\sum d_{\text{r}}) - \text{mean}(\sum d_{\text{b}})]} \quad (1)$$

where $\sum d_{\text{obs}}$ is the observed SSD and $\sum d_{\text{b}}$ and $\sum d_{\text{r}}$ are the Brownian and random SSD for each iteration. Once the value has been scaled, $D = 1$ corresponds to a random trait distribution, and $D = 0$ corresponds to a Brownian, or clustered, trait distribution. A p-value for *D* is calculated by comparing the estimated value to the distributions of values generated for $\sum d_{\text{b}}$ and $\sum d_{\text{r}}$ (see also Fritz & Purvis 2010, table 1).

Moran's *I* is a metric for spatial autocorrelation. It can be adapted for purpose here to measure the degree to which a binary trait (extinction) clusters in phylogenetic space (phylogenetic distance between taxa) (Gittleman & Kot 1990; Lockwood *et al.* 2002). It is calculated with equation 2:

$$I = \frac{\sum_i \sum_j z_i z_j w_{ij}}{\sum_i \sum_j w_{ij}} \times \frac{\text{n}}{\sum_i z_i^2} \quad (2)$$

where n is the number of observations, w_{ij} is a weighting that is calculated as 1 divided by the cophenetic distance between two species *i* and *j*, and z_i is the normalized value of the trait for the species *I* (Lockwood *et al.* 2002). In some previous studies, Moran's *I* correlograms have been used, which is possible when both extinction and taxonomic distance are binary traits. The generalized method for Moran's *I* used here has the advantages of providing one value for the entire tree, and including the additional information provided by phylogenetic branch duration (Hardy *et al.* 2012).

Timescaling

Phylogenetic comparative methods require a cladogram with branch durations scaled to time. The timescaling method may have an important influence on the outcome

of measurements of extinction clustering because it controls which taxa are included in each timeslice, as well as the phylogenetic distance between taxa. There are several *post hoc* methods for timescaling cladograms of fossil taxa, and here we applied four. First we used the *Hedman* algorithm (Hedman 2010; Lloyd *et al.* 2016*a*), which provides a distribution of estimates for the position of each internal node in the tree, based on the ages of the earliest representatives of consecutive sister groups. We performed this in R using code written by Graeme Lloyd and available in Lloyd *et al.* (2016*b*). We also tested the older and widely used *mbl* (minimum branch length; Laurin (2004)) and *equal* (Brusatte *et al.* 2008; Lloyd *et al.* 2012) methods. For the simulation study we additionally used the *cal3* timescaling method (Bapst 2013) which calibrates internal node positions according to three rates (origination, extinction and sampling) that can be estimated from occurrence data (Foote 2001). We could not use *cal3* on the real data because a majority of the taxa in our datasets are point occurrences, so we could not obtain reliable rate estimates (Bapst 2014).

Simulations

We used wrappers of functions in the *paleotree* package in R (Bapst 2012) to generate phylogenies that included episodic mass extinction events (scripts provided in Soul & Friedman 2017). These phylogenies were sampled to simulate fossil occurrence ranges, which were subsequently used to reconstruct and scale cladograms of the sampled fossil taxa according to time. We measured *D* for an identical timeslice, which included a mass extinction, through the 'true' phylogenetic histories and the sampled fossil cladograms, and compared the results.

In order to assess the way in which particular factors might bias measurements of clustering, we varied: (1) the method used to timescale the cladograms; (2) the degree to which extinction was phylogenetically clustered; and (3) the way in which sampled ancestral taxa were included within the timescaled cladograms.

Generating evolutionary histories. Phylogenies were generated using origination and sampling rates based on one simulation time unit representing 1 myr. Mass extinctions were generated by selecting 75% of taxa to go extinct. For clustered extinction we first simulated traits under Brownian motion. A low proportion of lineages with a trait value below a threshold were terminated, and a high proportion of those taxa with a trait value above the threshold were terminated. As discussed above (see Phylogenetic clustering of extinction) this leads to clusters of closely related tips on the phylogeny becoming extinct at the same time. For phylogenetically random extinction, the same overall proportion of lineages was terminated but terminations were selected randomly across the tree. The tree simulation continued from surviving lineages after each mass extinction event. We used three sets of five 'true' phylogenies, one set with clustered extinction, one with random extinction, and the final with bifurcating rather than budding origination (see Foote 1996, fig. 1). We sampled each of these 15 true phylogenies 50 times at three different *per-capita* rates: 0.01, 0.1 and 0.5 per-lineage time units. This sampling represents the combined processes of incomplete preservation and collection of fossil occurrence data.

Each of the sets of sampled ranges of taxa was used as the basis for timescaled cladograms (see Timescaling above). We tested three timescaling methods and implemented three different strategies for including sampled ancestral taxa. The options used in each set of simulations are detailed in Table 1. Overall this process yielded 15 simulated true phylogenies, 2250 sets of simulated taxon ranges and 5250 timescaled cladograms of sampled fossil taxa. Following generation of timescaled cladograms we measured Fritz and Purvis' *D* (Fritz & Purvis 2010) for the same, single, timeslice in each true phylogeny and each reconstructed fossil cladogram. This allowed assessment of which parameters were the most important controls on whether this measurement could recover the true signal for palaeobiological data.

Treatment of ancestors. Sampling taxa from ancestral lineages has been shown to be probable when dealing with data measured on long timescales (Foote 1996). In the majority of work estimating phylogenetic relationships, it has not been possible to identify which taxa might be ancestral to other sampled taxa (but see recent approaches e.g. Gavryushkina *et al.* 2014; Heath *et al.* 2014; Bapst *et al.* 2016). In commonly used methods of phylogenetic inference, sampled ancestral taxa are reconstructed as sister to their descendants. This may have an influence on the outcome of phylogenetic measures of extinction; the treatment of ancestors as they are incorporated into the phylogeny is therefore an important consideration. To simplify the test of how much of an influence sampled ancestral taxa might have on the outcome of the analysis, we used only a bifurcating model of origination (rather than budding or anagenetic origination, which can be simulated using *paleotree*). The first treatment of sampled ancestral taxa was to place them as sister taxa to their descendants and leave them in the cladogram (emulating the most likely result of a cladistic analysis where ancestors are sampled in real data (Wagner & Erwin 1995; Alroy *et al.* 2001)). This has two principal effects. First is the introduction of 'pseudoextinctions' where a taxon disappears from the fossil record and therefore appears to have become extinct, but actually the

TABLE 1. Parameters for sets of simulations.

Cladogram set	True phylogeny set	Model	Clustering	Timescaling	Ancestors
1	1	Budding	Yes	Hedman	Included as sister taxa to descendants
2	1	Budding	Yes	mbl	Included as sister taxa to descendants
3	1	Budding	Yes	cal3	Included as sister taxa to descendants
4	2	Budding	No	Hedman	Included as sister taxa to descendants
5	3	Bifurcating	Yes	Hedman	Included as sister taxa to descendants
6	3	Bifurcating	Yes	Hedman	Removed before timescaling
7	3	Bifurcating	Yes	Hedman	Removed after timescaling

Each 'Cladogram set' contains 750 timescaled cladograms, 50 for each different sampling rate of 0.01, 0.1 and 0.5 ltu^{-1} for each of the five phylogenies from the 'True phylogeny set'. 'Model' indicates the model of origination that was used to generate the phylogenies. 'Clustering' indicates whether or not the simulated true phylogeny had clustered or random extinction. 'Timescaling' refers to the method used to timescale cladograms. 'Ancestors' indicates how sampled ancestors were incorporated into the cladograms.

lineage has undergone morphological change. Second is the introduction of 'pseudosurvivals', which occur when an ancestor is sampled in an earlier time bin than its descendant. When they are reconstructed as sister taxa, the origin of the descendant must match the origin time of the ancestor and so a ghost range is inserted, crossing the boundary between time bins.

The second treatment of ancestors did not include sampled ancestral taxa, which where pruned from the cladograms before they were timescaled. This removes both pseudoextinctions and pseudosurvivals. The final treatment of ancestors was to remove sampled ancestral taxa only after the tree had been timescaled. As outlined above, this introduces ghost ranges into the phylogeny, so psuedosurvivals appear where these ghost ranges extend across the boundary into the previous timeslice. However, because the ancestors themselves are then pruned from the tree, pseudoextinctions are no longer present. The only treatment of ancestors available in reality is the first, because in the majority of cases we are unable to identify and remove ancestors from a phylogeny. Consequently, the second two treatments are performed only in order to understand the cause of any bias observed in the results, and do not represent real or reconstructed evolutionary trees. These scenarios, and their effects, are explored more fully in the discussion.

Caveats. The method used here can be viewed as optimistic, as only two factors (missing taxa and sampled ancestors) are investigated. We assume that cladograms recover true evolutionary relationships, which is unlikely to be the case. We also assume that there is no uncertainty in the ages of the fossil specimens, when in reality these are often only known as precisely as a geological stage, particularly for groups like terrestrial vertebrates where studies of phylogenetic clustering would most easily be conducted. Finally, we simulate the traits linked to extinction under a Brownian motion model of evolution, which leads to phylogenetically conserved trait patterns and phylogenetically clustered extinction. In reality, traits that are under selection may be best modelled by a different evolutionary regime (e.g. adaptive peak or early burst). We are therefore specifically investigating whether this approach can be used to detect selection with respect to traits that are adequately modelled by Brownian motion. The results of this simulation study do not fully represent our ability, or lack thereof, to correctly estimate this metric from fossil data. However, they do provide evidence of the way in which each cause of bias is likely to affect results and an indication of where problems are likely to arise. The code for all simulations and analyses can be found in Soul & Friedman (2017).

An empirical example: tetrapods at the PTME

As an illustration of this approach we quantified the phylogenetic clustering of extinctions in the fossil record of three major tetrapod clades (sauropsids, temnospondyls and synapsids) using two different metrics outlined in the phylogenetic clustering of extinction section above: Fritz and Purvis' *D* (Fritz & Purvis 2010) and Moran's *I* (Moran 1950; Gittleman & Kot 1990). The length of time over which we measured these metrics extended from the Pennsylvanian to the Late Triassic, divided into ten timeslices of similar length, each comprising one or two geological stages. We performed sensitivity analyses by varying the length of timeslices and the method used to scale cladogram branches to time.

Data. Phylogenies were composites constructed using published supertrees and cladistically inferred topologies for subgroups (cf. Soul & Friedman 2015). The topology for temnospondyls was a supertree taken directly from Ruta *et al.* (2007). The topologies for sauropsids and synapsids were composite trees constructed by combining

higher-level topologies for each clade that served as a 'backbone'; with the most recently available species-level topologies from studies of individual sub-clades. Source phylogenies are detailed in the supplementary information along with the set of 450 timescaled phylogenies used in the analyses and a plotted example tree for each clade (Soul & Friedman 2017, fig. S1). Occurrence data for each taxon were taken primarily from the Paleobiology Database (https://www.paleobiodb.org) except for parareptiles where these data were poorly covered in the database but available from the author of the published topology (Ruta *et al.* 2011).

To translate extinction to a binary trait, each timescaled cladogram was divided into successive timeslices of approximately the same length. If a taxon's last appearance fell within any one timeslice this was classified as an extinction; if the taxon's range included the end of the timeslice this was a survival because the taxon was present within the slice but survived into at least the next one. For the main analysis we used timeslices that began and ended at the start and end of geological stages, but combined some consecutive stages into single bins in order to generate intervals of more consistent length. It has been demonstrated previously that the intensity of the signal can be sensitive to temporal resolution of the timeslices (Hardy *et al.* 2012). Therefore, to test the effect of the length and timing of the timeslices we also conducted analyses using timeslices of exactly equal durations of 10 and of 15 myr.

The dates of occurrences of many fossil taxa, particularly vertebrates during the Palaeozoic and Mesozoic, are often only known to stage-level precision. To account for uncertainly in the actual times of first and last appearances of taxa in the record, a set of 50 stochastically generated fossil ranges was made for each taxon. First and last appearances were selected from a uniform distribution between the beginning and end of the most precise time period from which each taxon is known. The cladogram for each of the three groups was then timescaled using these sets of ranges. This can affect lineage divergence time estimates, and consequently the outcome of downstream analyses (Bapst 2014; Soul & Friedman 2015).

Sampling rate proxies. Variation between time bins in the rate of fossil preservation and discovery could have an important effect on the resulting signal (we test for this bias in the simulation section). In order to verify that preservation and sampling heterogeneity between bins was not the main driver of variation in extinction clustering results for our empirical data, we compared values of D to values for several proxies for fossil record quality. Due to the large proportion of point occurrences in the datasets (51%), and generally low number of occurrences per taxon, a sampling rate could not be directly estimated for the empirical data via any of several sophisticated and commonly used maximum likelihood or Bayesian estimators (e.g. Foote & Raup 1996; Alroy 2008; Liow & Finarelli 2014). Instead, we provide three proxies for the relative quality or heterogeneity of the fossil record through time: (1) the number of tetrapod bearing formations per bin; (2) the per-bin average number of formations in which each taxon occurring in that bin is represented; (3) a comparison of standard diversity (SD; a basic taxon count) with average duration of ghost lineage per taxon in each bin (average ghost lineage duration (AGLD); Cavin & Forey 2007). These proxies are only basic assessments of variation in fossil record quality through time, but are unfortunately the best methods currently available, given the nature of the data. They are adequate for their application here, which is to check whether sampled fossil record heterogeneity can be discounted as the main driver of the measured phylogenetic pattern in extinction.

For proxies 1 and 2 we performed a Pearson product–moment correlation test of first differences of D against the value for the proxy, a significant correlation would indicate that variation in D is an artefact of variation in fossil preservation and discovery potential through time. The method we used here for proxy 3 was developed by Cavin & Forey (2007) to distinguish between genuine and artefactual diversity peaks, by identifying time periods when the record comprises low numbers of highly productive horizons (Lagerstätten). A peak in SD that is not accompanied by a change in AGLD indicates that the record for that time bin is dominated by Lagerstätten. We use this method to identify time bins with particularly heterogeneous records, and compare this to times that extinction is particularly clustered or overdispersed.

RESULTS

Simulations

With the exception of Fig. 2, the figures in this section depict the median difference between D calculated on a simulated true phylogeny, and D calculated on the corresponding sampled cladograms. A positive value indicates that estimates of extinction were more strongly clustered on the sampled cladograms than on the true phylogeny.

Sampling rate. The baseline simulation demonstrates that accurate recovery of the strength of phylogenetic clustering of extinction is not guaranteed, whether or not extinction is clustered in (simulated) reality (Fig. 2). Correct recovery of the strength of phylogenetic clustering of extinction depends heavily on sampling rate (Figs 2, 3).

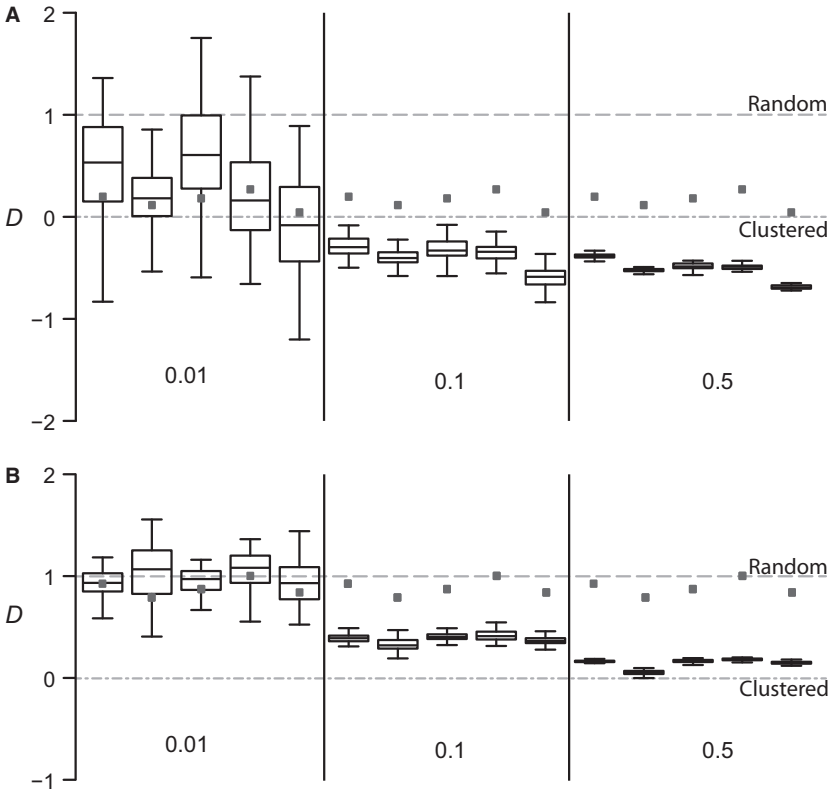

FIG. 2. Estimated values depend on sampling rate. Results of cladogram set 1 and 4. Five simulated phylogenies were sampled at three different sampling rates (0.01, 0.1 and 0.5, indicated at the bottom of the plot), filled squares are the true values for *D* for each phylogeny. Box and whisker plots show the range of values of *D* measured on 50 timescaled cladograms for each box. A, results when extinction in the simulation was phylogenetically clustered. B, results for extinction that was phylogenetically random.

At low sampling rates of 0.01 per lineage time unit (ltu) the value of *D* is on average higher (less clustered) than, or close to, the originally simulated value. A medium sampling rate of 0.1 ltu^{-1} usually resulted in overestimates of clustering (i.e. lower values of *D*), and a high sampling rate of 0.5 ltu^{-1} always leads to overestimates of the strength of clustering of extinction. In the simulations where extinction was not significantly clustered in the true phylogenies (Fig. 2B; Table 1: true phylogeny set 2), the analysis falsely rejected the possibility of phylogenetically random extinction at high sampling rates.

Timescaling method. The method used to timescale the trees of fossil taxa also had an important influence on recovery of accurate estimates of *D* (Fig. 3). At 0.01 ltu^{-1}, when the trees were timescaled using *mbl* and *cal3*, clustering was underestimated, but when the trees were timescaled using *Hedman*, estimates at 0.01 ltu^{-1} were closer to estimates of *D* from the real tree. However, these showed a large variance across measurements from different topologies. At higher sampling rates *Hedman* timescaled trees gave *D* values which implied a far greater strength of clustering than the original simulated phylogeny. When the trees were timescaled using *cal3*, estimates were more accurate overall, although low and high sampling rates did lead to a slight underestimate and overestimate of clustering respectively. Trees scaled using *mbl* did not give the most accurate estimates at any sampling rate, but were slightly better than *Hedman* at the two higher sampling rates.

Strength of clustering. Whether or not extinction in the simulation was phylogenetically clustered made a small difference in the mean accuracy of estimates of *D* (Fig. 4). When extinctions were phylogenetically clustered there was a larger variance in estimates from fossil trees than when extinction in the simulation was phylogenetically random. Medians of estimates for clustered and non-clustered extinctions showed approximately the same difference from the true value of *D*.

Ancestors. In the baseline simulation (Fig. 2), sampled ancestral taxa were placed in a polytomy with their descendants. When these were removed after timescaling (which removed pseudoextinctions but not pseudosurvivals) the measured signal shifted to lower values of *D* (more clustered); at high and medium sampling rates this lead to an overestimation of clustering, at low sampling rates clustering was still underestimated and showed large variation across topologies. When ancestors were removed before timescaling (removing both pseudoextinctions and pseudosurvivals) the measured signal at high sampling rates shifted from an overestimate of the strength of clustering to a more accurate estimate (Fig. 5).

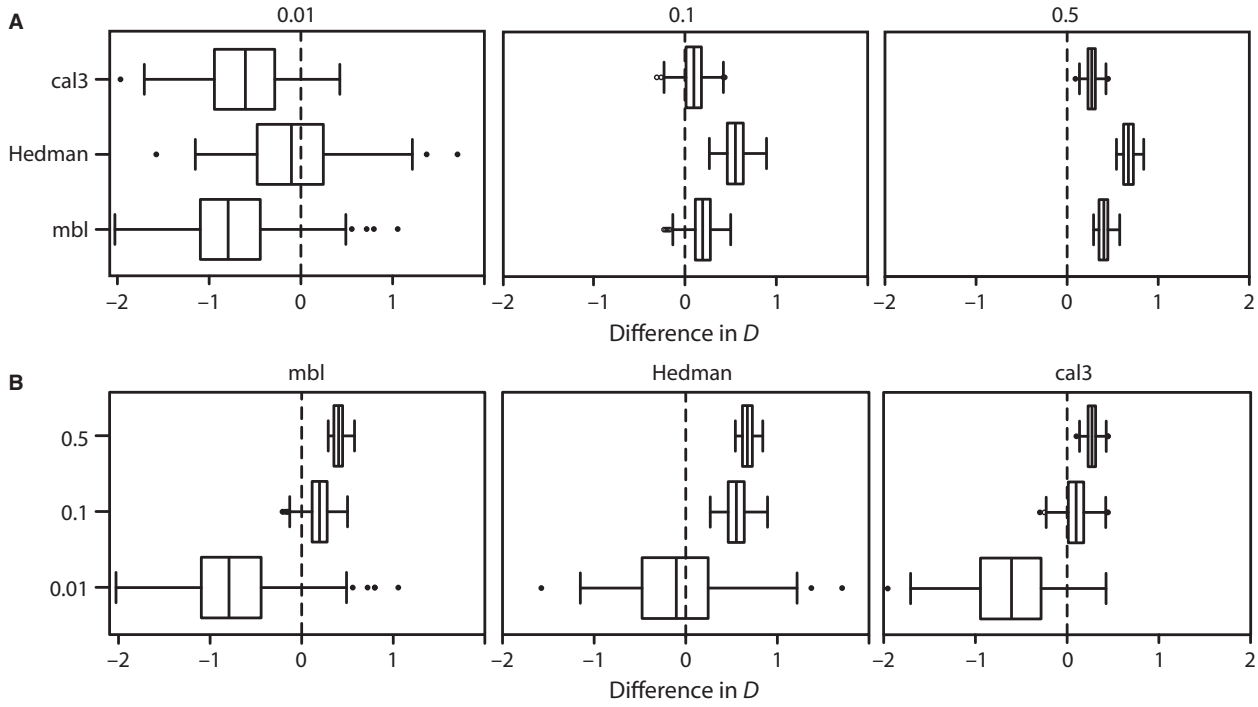

FIG. 3. Estimated values depend on timescaling method. Results of cladogram sets 1, 2 and 3. A, median and interquartile ranges of the difference in estimated value of *D* from the true value of *D* for three different sampling rates from left to right, using three different methods to timescale the cladogram; plotted to highlight the influence of sampling rate. B, the same data but arranged to highlight the influence of timescaling method. The methods increase in complexity and amount of input data required from left to right. Values close to the dashed line at 0 on the plots indicate that good estimates were made on the timescaled cladograms, with reference to the simulated true phylogeny. The narrower a box is, the more consistent results were across the iterations of cladograms.

Tetrapods at the PTME

Strength of clustering through time. Extinction was phylogenetically clustered in all three clades during the majority of the time bins investigated (Fig. 6), and fell within the distribution of the Brownian expectation. There is a greater spread in *D* values in time bins where the phylogenetic patterning is weak or random, showing that in these cases variation in both the topology and branch lengths of the tree has more of an effect on the result. All three clades show relatively random extinction in their early history; it is not clear whether this is a genuine signal or bias caused by proximity to the root of the tree or a small sample size. Extinctions are then consistently clustered in the last three timeslices of the Permian in all clades.

There does not seem to be an overall trend in changes in extinction clustering. It is not more likely for a decrease in signal strength between timeslices to follow an increase, or vice versa. Extinction intensity does not correlate significantly with strength of phylogenetic clustering for any of the clades (Pearson product–moment correlation: sauropsids $r = -0.6936$, $p = 0.0800$; synapsids $r = -0.5596$, $p = 0.1915$; temnospondyls $r = 0.2281$, $p = 0.6228$). Changing the algorithm used to timescale the cladograms lead to very similar estimates of *D* and did not affect the overall conclusions (Soul & Friedman 2017, fig. S2).

Measurements of Moran's *I* for sauropsids and synapsids showed similar patterns to *D,* with one exception in the Middle Triassic, during which a large proportion of taxa go extinct (72%). Moran's *I* for temnospondyls showed a slightly different pattern to *D* (Soul & Friedman 2017, fig. S3). Again this can most likely be attributed to the relative proportions of extinction; extinction intensity in temnospondyls correlates with the test statistic for *I* ($r = 0.8295$, $p = 0.02$).

D measured for timeslices of 15 and 10 myr in length was broadly similar to *D* obtained using combinations of stages as timeslices (Soul & Friedman 2017, fig. S4). The length of timeslices does not correlate with phylogenetic clustering (Pearson's *r*: sauropsids $r = -0.1518$, $p = 0.7740$; synapsids $r = 0.2469$, $p = 0.5935$; temnospondyls $r = 0.1034$, $p = 0.8254$).

Sampling rate proxies. Neither of the two formation-based proxies shows a significant correlation with *D* in any clade (Table 2). Average ghost lineage duration (AGLD)

shows a different pattern for each clade (Fig. 7). Sauropsids show an increase in heterogeneity of the record in the Middle Triassic, which does not correspond to an unusually high or low value of *D*. Synapsids have the same small increase in record heterogeneity in the Middle Triassic, preceded by a more dramatic increase in the Guadalupian that then declines in the end-Permian. These changes are not tracked by changes in *D*, which remains consistent and low throughout the Permian and Early to Middle Triassic. Temnospondyls show a very strong Lagerstätten effect in the Early Triassic but this time period is not distinguishable from others in the phylogenetic clustering analysis.

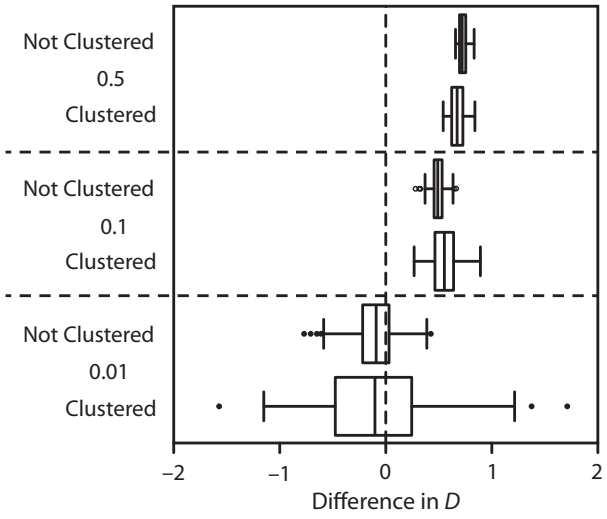

FIG. 4. Results of cladogram sets 1 and 4. The accuracy of estimates of *D* on the fossil trees compared to *D* on the true trees, at three different sampling rates when the simulated mass extinction events were, or were not, phylogenetically clustered when measured on the true tree.

DISCUSSION

Simulations

The results of the simulation analyses indicate that there are several important factors that need to be considered when interpreting phylogenetic clustering of extinction measured with fossil data. The effectiveness of different methods depends on the type of data being used for the analysis (Figs 2–6). The way in which taxa in the clade under investigation evolved and became extinct also has an effect on the accuracy and precision of results (Fig. 4), so caution must be taken when drawing conclusions from any one test. Although many factors have an influence on the bias in simulation outcomes, the sampling rate has the largest effect (Fig. 2). If the sampling rate can be estimated, at least approximately, the biases introduced by other factors can be anticipated.

Causes of bias. The two problems introduced in the simulation analyses were: (1) sampling rate variation (i.e. proportion of missing taxa); and (2) reconstruction of ancestors as sister taxa to their descendants. The second is linked to the first, as increased sampling rate increases the probability of sampling ancestors. Results suggest that the main bias at high sampling rates (towards overestimation of the strength of phylogenetic clustering) is a result of the second problem where pseudosurvivals result in an increased number of survivals at the end of each timeslice. This is demonstrated by the overestimation of clustering when only pseudosurvivals are included in the timescaled cladogram (Fig. 5). Situations where pseudosurvivals are likely to occur lead to clumps of closely related taxa surviving the end of timeslices (Fig. 8), which in turn lead to a lower phylogenetic distance between survivals on average. Extinctions and survivals are

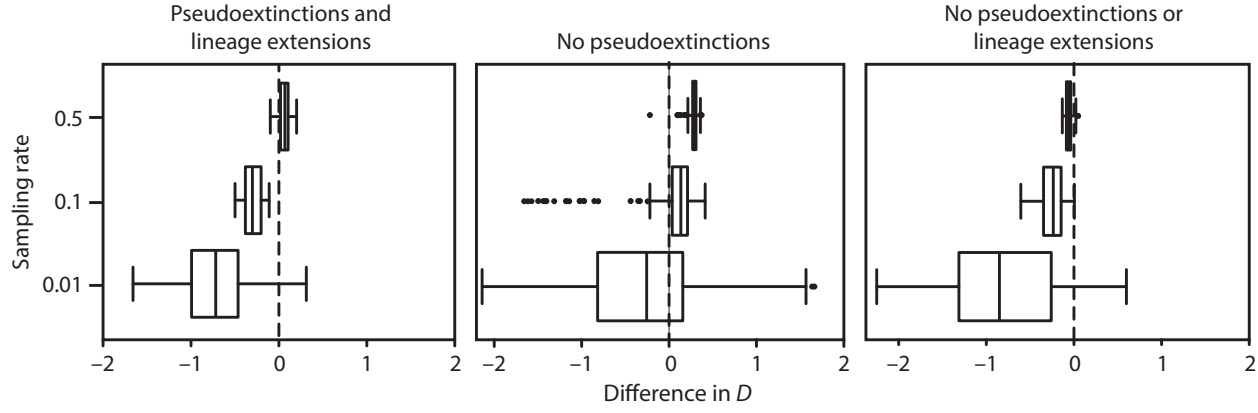

FIG. 5. Estimated values depend on treatment of sampled ancestors. Results of cladogram sets 5, 6 and 7, where ancestors were removed from the phylogenies at different points in the analysis. Removing sampled ancestors after timescaling the cladogram results in removal of pseudoextinctions (centre), removing sampled ancestors before timescaling results in removal of pseudoextinctions and lineage extensions (right).

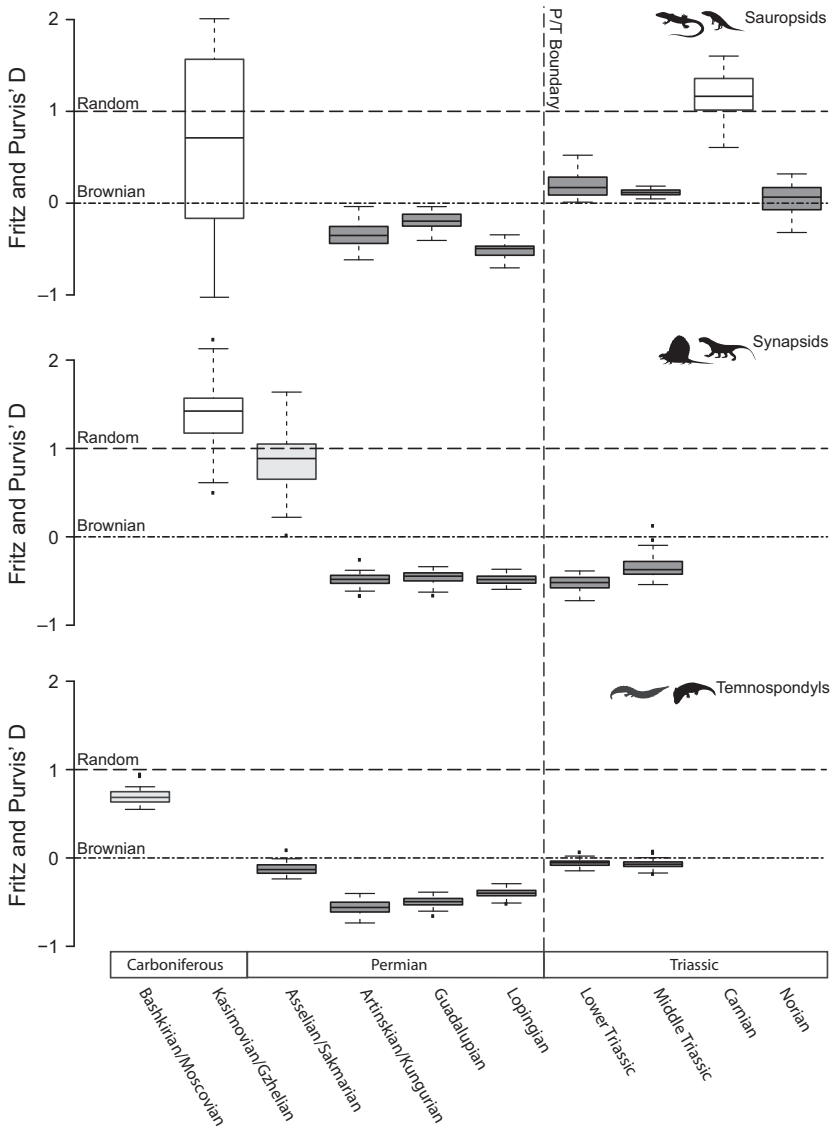

FIG. 6. Measurement of *D* through time on a set of 100 phylogenies timescaled using the *Hedman* method. The boxes encompass the middle 50% of the data and the line in each box is the median. Whiskers extend to the most extreme data point within 1.5 times the interquartile range. No shading indicates the values are within the distribution of the random expectation. Light grey shading indicates the values fall within both the random and normal expectations and dark grey that the values fall only within the Brownian expectation (i.e. extinction was phylogenetically clustered in the timeslices where boxes are shaded dark grey). Where there is a space for a particular timeslice rather than a box, the measurement for that timeslice did not fulfil the requirements of the method for *D* to provide a robust result, i.e. less than 25 tips, trait prevalence of less than 20% or more than 80%, or poor resolution. Silhouettes from http://phylopic.org by Nobu Tamura, Dmitry Bogdanov and Neil Kelley, vectorized by Michael Keesey.

symmetrical in the calculation of *D*, so an increase in survivals, where those survivals are in closely related taxa, has the same effect as an increase of extinctions in closely related taxa. When pseudoextinctions are also included they create an opposite bias, leading to an estimate closer to the originally simulated value of *D* (Figs 5, 8).

At low sampling rates the median estimate is rarely significantly clustered, even when the phylogeny that was originally simulated displayed highly clustered extinction. With fewer sampled taxa across the phylogeny overall, there is a lower probability of sampling closely related taxa, and a higher probability of sampling a taxon but not any of its descendants. For a poorly sampled tree, the most closely related taxa that have actually been sampled will not necessarily have been closely related in absolute terms, so the signal of very closely related taxa surviving or becoming extinct at the same time is lost. In addition, with smaller sample sizes the statistical power of the test to detect clustering is reduced.

Different timescaling methods changed the magnitude of bias in each case. The *mbl* method can be considered conservative because it does not assume large amounts of unsampled lineage history for which there is no direct evidence, but is unlikely to represent the true timings of lineage divergences accurately. The *cal3* method assigns branch durations in a less *ad hoc* manner and so tends to extend internal branches proportionally more than *mbl*, and the *Hedman* method extends internal branches even more so. This has the effect of drawing a greater number of divergences back into earlier timeslices, leading to more survivals and causing a more clustered signal to occur when compared to the signal measured on differently timescaled trees (Fig. 8).

TABLE 2. Test statistics for Pearson correlation test of first differences of estimates of *D* and sampling proxies.

Dependent	Independent	*r*	p-value
Sauropsid *D*	Tetrapod bearing formations per bin	−0.3218	0.5339
Sauropsid *D*	Average number of horizons per taxon per bin	−0.7282	0.1007
Synapsid *D*	Tetrapod bearing formations per bin	0.1173	0.8247
Synapsid *D*	Average number of horizons per taxon per bin	0.4505	0.3700
Temnospondyl *D*	Tetrapod bearing formations per bin	−0.4942	0.3974
Temnospondyl *D*	Average number of horizons per taxon per bin	0.4507	0.4462

Methodological recommendations. The correct way to implement and interpret measurements of the phylogenetic clustering of extinction is evidently a complex question. The nature of the data used for the analysis is important, as well as the way these data are subsequently treated. What does seem possible is that sampling rate can often be estimated (to the correct order of magnitude) and that an appropriate timescaling method can therefore be selected. However, all other biasing factors are either not possible to control, or difficult to estimate. With this in mind the most reasonable procedure is to begin by estimating sampling rate (in so far as that is possible), then to choose an appropriate timescaling method. At very low sampling rates, cladograms should be timescaled using the *Hedman* method to reduce bias, whereas at higher rates the *cal3* method should be used. Conveniently *cal3* is a method more suited to clades with higher sampling rates, as it requires additional information (origination, extinction and sampling rates) that can be more accurately measured for groups with a high sampling rate. Conversely the *Hedman* method can be used when sampling is low and the additional information on rates is not available (Lloyd *et al.* 2016*a*).

Following this, results should be interpreted in the context of the other biases that are probable given the dataset. For example if data are found to have a low sampling rate (*c.* 0.01 ltu^{-1}) and significantly clustered extinction, then this result can be expected to have a large error. If the data shows random extinction at a low sampling rate it can be considered more reliable. At high sampling rates (*c.* 0.5 ltu^{-1}) the analysis is consistently prone to overestimation of the strength of clustering, which means that a significantly clustered signal could be found that is in fact an artefact of the analysis and should be interpreted cautiously.

Ideally, to obtain an unbiased estimate of *D*, the phylogeny would be reconstructed using a method by which ancestors can be reliably inferred. New methods (e.g. Gavryushkina *et al.* 2014) which allow for sampled taxa to be directly ancestral to others in the estimated phylogeny hold possibilities for ancestral inference. These could be implemented for phylogenetic comparative methods in the future, but the data required for this kind of inference are unavailable for the clades studied here, and for many other clades of fossil taxa. Although the response of downstream analyses has not been quantified for phylogenies inferred in this way, there is great potential for improvement of phylogenetic comparative methods that are particularly vulnerable to bias caused by sampled ancestors, such as the method used here.

Tetrapods at the PTME

We provided an analysis of the phylogenetic clustering of extinction in three tetrapod clades during the PTME as an illustrative example of the application of this method to the fossil record. Tetrapod extinctions were phylogenetically clustered during the Pennsylvanian to Upper Triassic interval. This corroborates previous research that indicates that some degree of phylogenetic signal is a common feature of extinctions regardless of timescale, and can be considered a general rule (McKinney 1997; Janevski & Baumiller 2009; Roy *et al.* 2009). There is a large body of work to demonstrate that the nature and degree of extinctions during the PTME was different in each clade in measures such as diversity (Fröbisch 2008; Ruta & Benton 2008; Lucas 2009; Ruta *et al.* 2011). In combination with our results this indicates that variation in phylogenetic selectivity and variation in extinction intensity are not directly related, but may share a common driver of extreme values.

It has been suggested that the PTME represented a period of complete ecosystem restructuring for terrestrial tetrapods (Benton *et al.* 2004; Fröbisch 2013). Highly phylogenetically clustered extinction has a disproportionately large effect on biodiversity compared to random extinction (Davies & Yessoufou 2013), perhaps allowing for or requiring major ecosystem change (Krug & Patzkowsky 2015). The three focal clades show clustered extinction during the final timeslice of the Permian (Lopingian), but this is not unique; other intervals show clustering comparable to that of the PTME, indicating that phylogenetic selectivity may at times be decoupled from both extinction intensity and ecological impact (Droser *et al.* 2000; Hardy *et al.* 2012).

Geographically linked extinction. Fritz & Purvis (2010) suggest that phylogenetically random extinction can be attributed to geographical variation in the intensity of

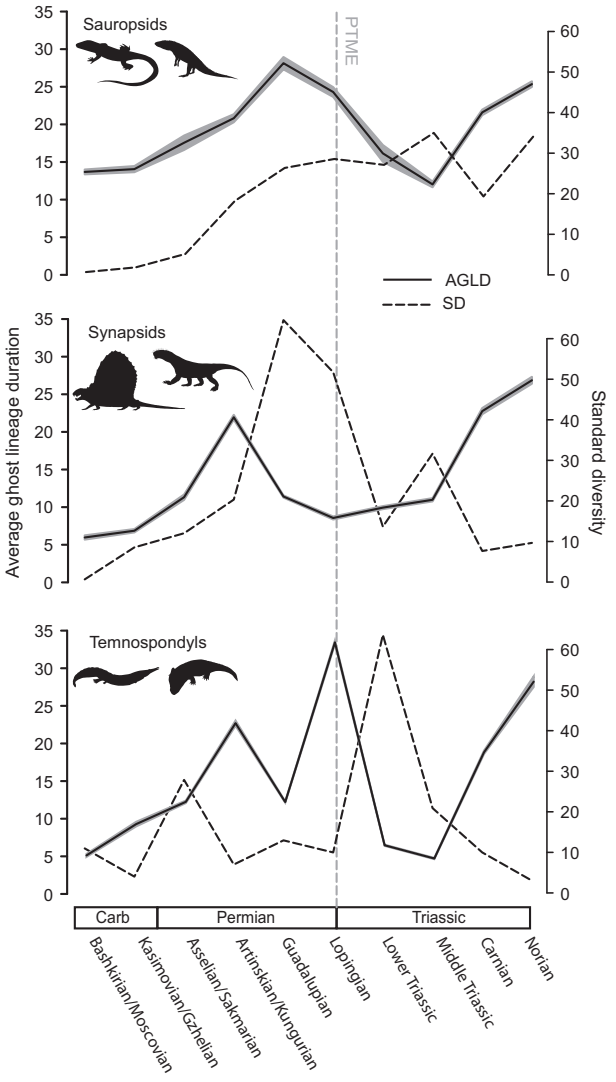

FIG. 7. Sampling heterogeneity did not drive estimates of *D* for the empirical data. Standard diversity (SD: a basic count of taxa without correction) and average ghost lineage duration (AGLD) through the time period of interest for each of the three clades tested. The dashed line and right hand axis show mean SD and the solid line and left hand axis show mean AGLD. Grey areas around the AGLD mean show the 95% confidence intervals based on the results on all 100 different topologies. Silhouettes from http://phylopic.org by Nobu Tamura, Dmitry Bogdanov and Neil Kelley, vectorized by Michael Keesey.

threat to survival in different regions that affects all the taxa living there (e.g. one region becomes very dry or hot). On small spatial scales, extinction (or extinction risk) is often phylogenetically clustered and this clustering can indeed be attributed to selection against particular phylogenetically conserved phenotypes (Roy *et al.* 2009; Hardy *et al.* 2012). However, across the large spatial and temporal scales of this study, geographical distribution may also be phylogenetically conserved, because many taxa are restricted to particular habitats or climatic zones, to which close relatives with whom they share a recent evolutionary history are also more likely to be restricted (Lieberman 2003; Krug & Patzkowsky 2015). This pattern may not occur in all taxa, particularly not in generalists with good dispersal ability, but overall the two factors which have control over a species' vulnerability to extinction on the timescales in this study, its phenotype and the extinction threat it experiences, are both expected to be phylogenetically conserved to some degree, particularly in the early history of a taxon. Future work could compare the phylogenetic clustering of extinction through time with the correlation between geographical and phylogenetic distance of sampled taxa to begin to tease apart these two factors.

Sensitivity tests. In agreement with the simulation study the sensitivity tests indicated that in some cases the method employed to perform the various steps required to obtain a result had an influence on the observed signal, but these effects were small.

Changing the algorithm used to timescale the trees had an effect because different timescaling methods add duration to internal branch lengths to varying degrees, which led to taxa being present in earlier time bins in *Hedman* timescaled trees. (see Fig. 8 and simulation results for the possible effect of this).

Implementing an alternative method (Moran's *I*) to measure clustering also gave a slightly different result, particularly for temnospondyls. However, the strong link between extreme values of trait prevalence and extreme values of *I* indicates that the method is not particularly robust to variation in trait prevalence, unlike Fritz and Purvis' *D*, which can give results that are comparable across timeslices even when they have a very high or low proportion of extinctions.

Changing the length of the timeslices changes which taxa survive in each timeslice, and therefore the strength of clustering. When the timeslices correspond to combinations of stages (as they do in the main analysis), a taxon will always go extinct in the same timeslice, even if its divergence date is in different stages for different iterations of the timescaling algorithm. This is not the case when 10 and 15 myr timeslices are used (except for the Lopingian because this boundary is used in all the alternative sets of timeslices) leading to the larger variance across different tree topologies in these results (Soul & Friedman 2017, fig. S4).

Sampling proxies. The simulation study demonstrated that the sampling rate has an important effect on whether estimates of *D* are biased towards phylogenetic clustering or overdispersion. With this in mind it was important to assess

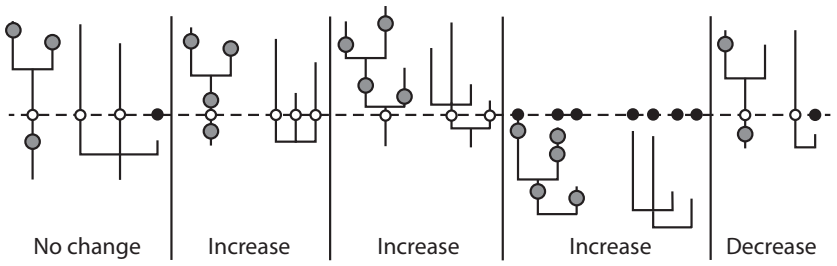

FIG. 8. Some examples of hypothetical sampling patterns and the consequent reconstructed cladograms, time increasing towards the present at the top of the figure. Shows inferred survivals and extinctions, along with whether they could be expected to lead to an increase or decrease in estimates of *D*. The focal timeslice is below the dashed line. In each section the 'real' phylogeny is on the left and the reconstructed tree is on the right. Samples of the real phylogeny are shown by grey filled circles (these samples represent fossil specimens when applied to real data). Extinctions are shown by filled circles and survivals by empty circles. Any survival or extinction present in the right hand tree but not the left can be considered a pseudosurvival or pseudoextinction respectively. These reconstructed cladograms have only short backwards range extensions, consistent with what would be seen when using the *mbl* method, if a different timescaling algorithm was used, divergences would move further back in time. These do not represent all possible scenarios but are illustrative of situations that could lead to localized increases or decreases in estimates of clustering.

whether variation in sampling probability between bins was driving increases and decreases in estimates of *D*. A record comprised mostly of singletons prevented direct estimation of sampling rate. However, two proxies for relative sampling through time based on formation counts showed no obvious correlation with extinction clustering metrics, demonstrating that preservation and discovery potential was not the main driver of differences in clustering between time bins. Likewise the average ghost lineage duration analysis showed that there are sections of the record of all three clades that are heterogeneous (taxa are sampled from one or a few horizons of exceptional preservation), but these do not correspond to unusually high or low values of *D*.

Implications of the simulation study for the tetrapod case study

The literature indicates that sampling rates that translate to the region of 0.1 per lineage million years (lmy) can be expected for marine invertebrate taxa. For Neogene mammals and well preserved marine invertebrate records a rate of 0.5 lmy^{-1} might be possible (Foote & Sepkoski 1999; Alba *et al.* 2001). For the majority of Palaeozoic and Mesozoic terrestrial vertebrate clades, particularly those which include many point occurrences, the sampling rate is likely to be on the order of 0.01 lmy^{-1} (Foote *et al.* 1999; Friedman & Brazeau 2011). Thus the simulation results at 0.01 ltu^{-1} are the most representative of our tetrapod dataset, and are therefore the best indicator of the bias that can be expected in our empirical results.

Results for 0.01 ltu^{-1} are on average biased towards underestimating the strength of clustering (Figs 2, 3). However, there is a large variation from results that were measured on different randomly selected samples of the record, and they include estimates of strong clustering. Given that the large majority of timeslices in the empirical analysis show clustered extinction, it is unlikely, but still possible, that each of these estimates is a biased result based on the sample of the record represented by the cladogram, and that extinction was not in fact phylogenetically clustered.

Strongly clustered median values of *D* were only produced in the simulation when fossil trees derived from records with high sampling rate were tested. At low sampling rates significant clustering was rarely observed within the simulations. This calls into question how, at the low sampling rates seen for terrestrial vertebrate clades, significant clustering was so commonly found in our real data. One possibility might be that there is a taphonomic bias caused by regional-scale ecological stress in combination with local scale preservation heterogeneity and taxon distributions. A further possibility is that the bifurcating constant rate birth–death model and univariate Brownian motion trait evolution used to simulate the true phylogenies was not an adequate model for the evolutionary process (Hagen *et al.* 2015). For example, there has been some previous support for the hypothesis that simultaneous phylogenetic selectivity with respect to multiple aspects of phenotype or ecology (i.e. a phylogenetically conserved ecological niche) should lead to strong clustering of extinction (Green *et al.* 2011).

CONCLUSIONS

The phylogenetic clustering of extinction is a useful measurement that can be made for clades where a robust phylogeny is available, but detailed trait information is lacking. In the absence of adequate data to identify the extinction risk associated with specific phenotypic or life history traits in the geological record,

phylogeny can act as a proxy for the effect of selection (or lack thereof) against the combination of these traits in a species, if those traits are phylogenetically conserved. In combination with previous studies, these results demonstrate that phylogenetic clustering of extinctions is common on all scales and that patterns in the short-term scale up over time to result in similar patterns in the long-term. There are several characteristics prevalent in phylogenies of fossil taxa that can introduce bias into the results of these measurements and they must be carefully considered before the analyses are performed and before conclusions are drawn. The following key points should be held in mind when measuring clustering of extinction in fossil groups:

1. The sampling rate for the clade of interest should be estimated as accurately as possible to provide a context for interpretation of results.
2. The cladogram should be timescaled with an algorithm appropriate for the sampling rate and incorporate estimation of ancestral relationships if possible.
3. Low preservation and discovery rate leads to a loss of information that causes a bias towards low estimates of the strength of clustering of extinction.
4. High preservation and discovery rate without consideration of ancestor–descendent relationships leads to topologies that cause a bias towards high estimates of the strength of clustering of extinction.

Despite the importance of these considerations, phylogenetic clustering of extinction can offer additional insight into macroevolutionary patterns associated with extinction events and how those patterns vary across time and clades.

Acknowledgements. Neil Brocklehurst, Allison Daley, Graeme Lloyd and Gene Hunt provided helpful discussion of the work. Mark Patzkowsky and Matt Wills provided reviews that improved the quality of the manuscript. LCS was supported by Natural Environment Research Council UK doctoral training grant NE/J500045/. This is Paleobiology Database publication number 275.

DATA ARCHIVING STATEMENT

Data for this study are available in the Dryad Digital Repository: https://doi.org/10.5061/dryad.208b1

Editor. Andrew Smith

REFERENCES

ALBA, D. M., AGUSTÍ, J. and MOYÀ-SOLÀ, S. 2001. Completeness of the mammalian fossil record in the Iberian Neogene. *Paleobiology*, **27**, 79–83.

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

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

—— MARSHALL, C. R., BAMBACH, R. K., BEZUSKO, K., FOOTE, M., FURSICH, 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., OLSZEWSKI, T. D., PATZKOWSKY, M. E., RAUP, D. M., ROY, K., SEPKOSKI, J. J. Jr, SOMMERS, M. G., WAGNER, P. J. and WEBBER, A. 2001. Effects of sampling standardization on estimates of Phanerozoic marine diversification. *Proceedings of the National Academy of Sciences*, **98**, 6261–6266.

—— 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., OLSZWESKI, T. D., PATZKOWSKY, M. E., PETERS, S. E., VILLIER, L., WAGNER, P. J., BONUSO, N., BORKOW, P. S., BERNNEIS, 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., TOMAŠOVÝCH, A. and VISAGGI, C. C. 2008. Phanerozoic trends in the global diversity of marine invertebrates. *Science*, **321**, 97–100.

BAPST, D. W. 2012. paleotree: an R package for paleontological and phylogenetic analyses of evolution. *Methods in Ecology & Evolution*, **3**, 803–807.

—— 2013. A stochastic rate-calibrated method for time-scaling phylogenies of fossil taxa. *Methods in Ecology & Evolution*, **4**, 724–733.

—— 2014. Assessing the effect of time-scaling methods on phylogeny-based analyses in the fossil record. *Paleobiology*, **40**, 331–351.

—— WRIGHT, A. M., MATZKE, N. J. and LLOYD, G. T. 2016. Topology, divergence dates, and macroevolutionary inferences vary between different tip-dating approaches applied to fossil theropods (Dinosauria). *Biology Letters*, **12**, 465–470.

BARNOSKY, A. D., MATZKE, N., TOMIYA, S., WOGAN, G. O. U., SWARTZ, B., QUENTAL, T. B., MARSHALL, C., McGUIRE, J. L., LINDSEY, E. L., MAGUIRE, K. C., *et al.* 2011. Has the Earth's sixth mass extinction already arrived? *Nature*, **471**, 51–57.

BEAULIEU, J. M. and O'MEARA, B. C. 2014. OUwie: analysis of evolutionary rates in an OU framework, v. 1.46. https://CRAN.R-project.org/package=OUwie

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

BIELBY, J., CUNNINGHAM, A. A. and PURVIS, A. 2006. Taxonomic selectivity in amphibians: ignorance, geography or biology? *Animal Conservation*, **9**, 135–143.

BRUSATTE, S. L., BENTON, M. J., RUTA, M. and LLOYD, G. T. 2008. Superiority, competition, and opportunism in the evolutionary radiation of dinosaurs. *Science*, **321**, 1485–1488.

CARDILLO, M., MACE, G. M., GITTLEMAN, J. L., JONES, K. E., BIELBY, J. and PURVIS, A. 2008. The predictability of extinction: biological and external correlates of decline in mammals. *Proceedings of the Royal Society B*, **275**, 1441–1448.

CAVIN, L. and FOREY, P. L. 2007. Using ghost lineages to identify diversification events in the fossil record. *Biology Letters*, **3**, 201–204.

DAVIES, T. J. and YESSOUFOU, K. 2013. Revisiting the impacts of non-random extinction on the tree-of-life. *Biology Letters*, **9**, 20130343.

DAVIS, S. P., FINARELLI, J. A. and COATES, M. I. 2012. Acanthodes and shark-like conditions in the last common ancestor of modern gnathostomes. *Nature*, **486**, 247–250.

DROSER, M. L., BOTTJER, D. J., SHEEHAN, P. M. and McGHEE, G. R. 2000. Decoupling of taxonomic and ecologic severity of Phanerozoic marine mass extinctions. *Geology*, **28**, 675–678.

ERWIN, D. 2009. A call to the custodians of deep time. *Nature*, **462**, 282–283.

FELSENSTEIN, J. 1985. Phylogenies and the comparative method. *The American Naturalist*, **125**, 1–15.

FINNEGAN, S., ANDERSON, S. C., HARNIK, P. G., SIMPSON, C., TITTENSOR, D. P., BYRNES, J. E., FINKEL, Z. V., LINDBERG, D. R., LIOW, L. H., LOCKWOOD, R., LOTZE, H. K., McCLAIN, C. R., McGUIRE, J. L., O'DEA, A., and PANDOLFI, J. M. 2015. Paleontological baselines for evaluating extinction risk in the modern oceans. *Science*, **348**, 567–570.

FOOTE, M. 1996. On the probability of ancestors in the fossil record. *Paleobiology*, **22**, 141–151.

—— 2001. Inferring temporal ptterns of preservation, origination and extinction from taxonomic survivorship analysis. *Paleobiology*, **27**, 602–630.

—— and RAUP, D. M. 1996. Fossil preservation and the stratigraphic ranges of taxa. *Palaeobiology*, **22**, 121–140.

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

—— HUNTER, J. P., JANIS, C. M. and SEPKOSKI, J. J. 1999. Evolutionary and preservational constraints on origins of biologic groups: divergence times of eutherian mammals. *Science*, **283**, 1310–1314.

FRASER, D., GORELICK, R. and RYBCZYNSKI, N. 2015. Macroevolution and climate change influence phylogenetic community assembly of North American hoofed mammals. *Biological Journal of the Linnean Society*, **114**, 485–494.

FRECKLETON, R. P., HARVEY, P. H. and PAGEL, M. 2002. Phylogenetic analysis and comparative data: a test and review of evidence. *The American Naturalist*, **160**, 712–726.

FRIEDMAN, M. 2009. Ecomorphological selectivity among marine teleost fishes during the end-Cretaceous extinction. *Proceedings of the National Academy of Sciences*, **106**, 5218–5223.

—— and BRAZEAU, M. D. 2011. Sequences, stratigraphy and scenarios: what can we say about the fossil record of the earliest tetrapods? *Proceedings of the Royal Society B*, **278**, 432–439.

FRITZ, S. A. and PURVIS, A. 2010. Selectivity in Mammalian extinction risk and threat types: a new measure of phylogenetic signal strength in binary traits. *Conservation Biology*, **24**, 1042–1051.

—— SCHNITZLER, J., ERONEN, J. T., HOF, C., BÖHNING-GAESE, K. and GRAHAM, C. H. 2013. Diversity in time and space: wanted dead and alive. *Trends in Ecology & Evolution*, **28**, 509–516.

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

—— 2013. Vertebrate diversity across the end-Permian mass extinction — separating biological and geological signals. *Palaeogeography, Palaeoclimatology, Palaeoecology*, **372**, 50–61.

GARLAND, T. and DIAZ-URIARTE, R. 1999. Polytomies and phylogenetically independent contrasts: examination of the bounded degrees of freedom approach. *Systematic Biology*, **48**, 547–558.

GAVRYUSHKINA, A., WELCH, D., STADLER, T. and DRUMMOND, A. 2014. Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration. *PLoS Computational Biology*, **10**, e1003919.

GITTLEMAN, J. L. and KOT, M. 1990. Adaptation: statistics and a null model for estimating phylogenetic effects. *Systematic Zoology*, **39**, 227–241.

GRAFEN, A. 1989. The phylogenetic regression. *Philosophical Transactions of the Royal Society B*, **326**, 119–157.

GREEN, W. A., HUNT, G., WING, S. L. and DIMICHELE, W. A. 2011. Does extinction wield an axe or pruning shears? How interactions between phylogeny and ecology affect patterns of extinction. *Paleobiology*, **37**, 72–91.

HAGEN, O., HARTMANN, K., STEEL, M. and STADLER, T. 2015. Age-dependent speciation can explain the shape of empirical phylogenies. *Systematic Biology*, **64**, 432–440.

HALLIDAY, T. J. D. and GOSWAMI, A. 2016. The impact of phylogenetic dating method on interpreting trait evolution: a case study of Cretaceous–Palaeogene eutherian body-size evolution. *Biology Letters*, **12**, 6–12.

HARDY, C., FARA, E., LAFFONT, R., DOMMERGUES, J., MEISTER, C. and NEIGE, P. 2012. Deep-time phylogenetic clustering of extinctions in an evolutionarily dynamic clade (Early Jurassic Ammonites). *PLoS One*, **7**, e37977.

HARMON, L. J., LOSOS, J. B., DAVIES, T. J., GILLESPIE, R. G., GITTLEMAN, J. L., JENNINGS, B. W., KOZAK, K. H., McPEEK, M. A., MORENO-ROARK, F., NEAR, T. J., PURVIS, A., RICKLEFS, R. E., SCHLUTER, D., SCHULTE, J. A. II, SEEHAUSEN, O., SIDLAUSKAS, B. L., TORRES-CARVAJAL, O., WEIR, J. T. and MOOERS, A. Ø., 2010. Early bursts of body size and shape evolution are rare in comparative data. *Evolution*, **64**, 2385–2396.

HARNIK, P. G. 2011. Direct and indirect effects of biological factors on extinction risk in fossil bivalves. *Proceedings of the National Academy of Sciences*, **108**, 13594–13599.

—— SIMPSON, C. and PAYNE, J. L. 2012. Long-term differences in extinction risk among the seven forms of rarity. *Proceedings of the Royal Society B*, **279**, 4969–4976.

—— FITZGERALD, P. C., PAYNE, J. L. and CARLSON, S. J. 2014. Phylogenetic signal in extinction selectivity in Devonian terebratulide brachiopods. *Paleobiology*, **40**, 675–692.

HEATH, T. A., HUELSENBECK, J. P. and STADLER, T. 2014. The fossilized birth–death process for coherent calibration of divergence-time estimates. *Proceedings of the National Academy of Sciences*, **111**, E2957–E2966.

HEDMAN, M. M. 2010. Constraints on clade ages from fossil outgroups. *Paleobiology*, **36**, 16–31.

HOUSWORTH, E. A. and MARTINS, E. P. 2001. Random sampling of constrained phylogenies: conducting phylogenetic analyses when the phylogeny is partially known. *Systematic Biology*, **50**, 628–639.

HUANG, D., GOLDBERG, E. E. and ROY, K. 2015. Fossils, phylogenies, and the challenge of preserving evolutionary history in the face of anthropogenic extinctions. *Proceedings of the National Academy of Sciences*, **112**, 4909–4914.

HUNT, G., ROY, K. and JABLONSKI, D. 2005. Species level heritability reaffirmed: a comment on 'On the heritability of geographic range sizes'. *The American Naturalist*, **166**, 129–135.

JABLONSKI, D. 1994. Extinctions in the fossil record. *Philosophical Transactions of the Royal Society B*, **344**, 11–17.

—— 2005. Mass extinctions and macroevolution. *Paleobiology*, **31**, 192–210.

—— 2008. Extinction and the spatial dynamics of biodiversity. *Proceedings of the National Academy of Sciences*, **105**, 11528–11535.

JACKSON, J. B. C. and ERWIN, D. H. 2006. What can we learn about ecology and evolution from the fossil record? *Trends in Ecology & Evolution*, **21**, 322–328.

JANEVSKI, G. A. and BAUMILLER, T. K. 2009. Evidence for extinction selectivity throughout the marine invertebrate fossil record. *Paleobiology*, **35**, 553–564.

JEFFERY, C. H. 2001. Heart urchins at the Cretaceous/Tertiary boundary: a tale of two clades. *Paleobiology*, **27**, 140–158.

KIESSLING, W. and ABERHAN, M. 2007. Geographical distribution and extinction risk: lessons from Triassic–Jurassic marine benthic organisms. *Journal of Biogeography*, **34**, 1473–1489.

KRUG, A. Z. and PATZKOWSKY, M. E. 2015. Phylogenetic clustering of origination and extinction across the Late Ordovician mass extinction. *PLoS One*, **10**, e0144354.

LAURIN, M. 2004. The evolution of body size, Cope's rule and the origin of amniotes. *Systematic Biology*, **53**, 594–622.

LIEBERMAN, B. S. 2003. Paleobiogeography: the relevance of fossils to biogeography. *Annual Review of Ecology, Evolution, & Systematics*, **34**, 51–69.

LIOW, L. H. 2007. Does versatility as measured by geographic range, bathymetric range and morphological variability contribute to taxon longevity? *Global Ecology & Biogeography*, **16**, 117–128.

—— and FINARELLI, J. A. 2014. A dynamic global equilibrium in carnivoran diversification over 20 million years. *Proceedings of the Royal Society of London B*, **281**, 20132312.

LLOYD, G. T., WANG, S. C. and BRUSATTE, S. L. 2012. Identifying heterogeneity in rates of morphological evolution: discrete character change in the evolution of lungfish (Sarcopterygii; Dipnoi). *Evolution*, **66**, 330–348.

—— BAPST, D. W., FRIEDMAN, M. and DAVIS, K. E. 2016*a*. Probabilistic divergence time estimation without branch lengths: dating the origins of dinosaurs, avian flight, and crown birds. *Biology Letters*, **12**, 20160609.

—— —— —— —— 2016*b*. Data from: Probabilistic divergence time estimation without branch lengths: dating the origins of dinosaurs avian flight and crown birds. *Dryad Digital Repository*. doi: 10.5061/dryad.p660m

LOCKWOOD, J. L., RUSSELL, G. J., GITTLEMAN, J. L., DAEHLER, C. C., McKINNEY, M. L. and PURVIS, A. 2002. A metric for analyzing taxonomic patterns of extinction risk. *Conservation Biology*, **16**, 1137–1142.

LUCAS, S. G. 2009. Timing and magnitude of tetrapod extinctions across the Permo-Triassic boundary. *Journal of Asian Earth Sciences*, **36**, 491–502.

McKINNEY, M. L. 1997. Extinction vulnerability and selectivity: combining ecological and paleontological views. *Annual Review of Ecology & Systematics*, **28**, 495–516.

MORAN, P. A. P. 1950. Notes on continuous stochastic phenomena. *Biometrika*, **37**, 17–23.

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

ORME, D., FRECKLETON, R. P., THOMAS, G., PETZOLDT, T. and FRITZ, S. A. 2012. caper: comparative analyses of phylogenetics and evolution in R, v. 0.5.2. https://cran.r-project.org/web/packages/caper/index.html

PAYNE, J. L. and FINNEGAN, S. 2007. The effect of geographic range on extinction risk during background and mass extinction. *Proceedings of the National Academy of Sciences*, **104**, 10506–10511.

—— BUSH, A. M., CHANG, E. T., HEIM, N. A., KNOPE, M. L. and PRUSS, S. B. 2016. Extinction intensity, selectivity and their combined macroevolutionary influence in the fossil record. *Biology Letters*, **12**, 20160202.

PENNELL, M. W. and HARMON, L. J. 2013. An integrative view of phylogenetic comparative methods: connections to population genetics, community ecology, and paleobiology. *Annals of the New York Academy of Sciences*, **1289**, 90–105.

PURVIS, A. 2008. Phylogenetic approaches to the study of extinction. *Annual Review of Ecology, Evolution, and Systematics*, **39**, 301–319.

RABOSKY, D. L. 2010. Extinction rates should not be estimated from molecular phylogenies. *Evolution*, **64**, 1816–1824.

R CORE TEAM. 2015. R: A language and environment for statistical computing, v. 3.1.3. R Foundation for Statistical Computing. http://www.R-project.org/

RAUP, D. M. 1994. The role of extinction in evolution. *Proceedings of the National Academy of Sciences*, **91**, 6758–6763.

ROY, K., HUNT, G. and JABLONSKI, D. 2009. Phylogenetic conservatism of extinctions in marine bivalves. *Science*, **325**, 733–737.

RUTA, M. and BENTON, M. J. 2008. Calibrated diversity, tree topology and the mother of mass extinctions: the lesson of temnospondyls. *Palaeontology*, **51**, 1261–1288.

—— PISANI, D., LLOYD, G. T. and BENTON, M. J. 2007. A supertree of Temnospondyli: cladogenetic patterns in the most species-rich group of early tetrapods. *Proceedings of the Royal Society B*, **274**, 3087–3095.

—— CISNEROS, J. C., LIEBRECHT, T., TSUJI, L. A. and MU, J. 2011. Amniotes through major biological crises: faunal turnover among parareptiles during the end-Permian mass extinction. *Palaeontology*, **54**, 1117–1137.

SAKAMOTO, M., VENDITTI, C. and BENTON, M. J. 2016. 'Residual diversity estimates' do not correct for sampling bias in palaeodiversity data. *Methods in Ecology & Evolution*, published online 24 October. doi: 10.1111/2041-210X.12666

SMITH, J. T. and ROY, K. 2006. Selectivity during background extinction: Plio-Pleistocene scallops in California. *Paleobiology*, **32**, 408–416.

SMITS, P. D. 2015. Expected time-invariant effects of biological traits on mammal species duration. *Proceedings of the National Academy of Sciences*, **112**, 13015–13020.

SOUL, L. C. and FRIEDMAN, M. 2015. Taxonomy and phylogeny can yield comparable results in comparative palaeontological analyses. *Systematic Biology*, **64**, 608–620.

—— —— 2017. Data from: Bias in phylogenetic measurements of extinction and a case study of end-Permian tetrapods. *Dryad Digital Repository*. doi: 10.5061/dryad.208b1

STANLEY, S. M. 1998. *Macroevolution, pattern and process*. Johns Hopkins University Press.

STONE, E. A. 2011. Why the phylogenetic regression appears robust to tree misspecification. *Systematic Biology*, **60**, 245–260.

TOMIYA, S. 2013. Body size and extinction risk in terrestrial mammals above the species level. *The American Naturalist*, **182**, E196–E214.

TURVEY, S. T. and FRITZ, S. A. 2011. The ghosts of mammals past: biological and geographical patterns of global mammalian extinction across the Holocene. *Philosophical Transactions of the Royal Society of London B*, **366**, 2564–2576.

VOS, J. M. DE, JOPPA, L. N., GITTLEMAN, J. L., STEPHENS, P. R. and PIMM, S. L. 2014. Estimating the normal background rate of species extinction. *Conservation Biology*, **29**, 452–462.

WAGNER, P. J. and ERWIN, D. H. 1995. Phylogenetic patterns as tests of speciation models. 87–122. *In* ERWIN, D. H. and ANSTEY, R. L. (eds). *New approaches to speciation in the fossil record*. Columbia University Press.