Palaeontology The Palaeontological Association www.palass.org

[Palaeontology, Vol. 61, Part 1, 2018, pp. 119–132]

ON FORMATION-BASED SAMPLING PROXIES AND WHY THEY SHOULD NOT BE USED TO CORRECT THE FOSSIL RECORD

by ALEXANDER M. DUNHILL[1], BJARTE HANNISDAL[2], NEIL BROCKLEHURST[3] *and* MICHAEL J. BENTON[4]

[1]School of Earth & Environment, University of Leeds, Leeds, LS2 9JT, UK; a.dunhill@leeds.ac.uk
[2]Centre for Geobiology, Department of Earth Science, University of Bergen, Bergen, Norway
[3]Museum für Naturkunde, Leibniz-Institut für Evolutions- und Biodiversitätsforschung, Berlin, Germany
[4]School of Earth Sciences, University of Bristol, Bristol, BS8 1TQ, UK

Typescript received 29 March 2017; accepted in revised form 13 September 2017

Abstract: The fossil record is a unique resource on the history of life, but it is well known to be incomplete. In a series of high-profile papers, a residual modelling technique has been applied to correct the raw palaeodiversity signal for this bias and incompleteness, and the claim is made that the processed time series are more accurate than the raw data. We apply empirical and simulation approaches to test for correlation and directionality of any relationships between rock and fossil data. The empirical data comprise samples of the global fossil record through the Phanerozoic, and we use simulations to assess whether randomly sampled subsets of modelled data can be improved by application of the residual modelling technique. Our results show that using formation counts as a sampling proxy to correct the fossil record via residual modelling is ill founded. The supposedly independent model of sampling is information-redundant with respect to the raw palaeodiversity data it seeks to correct, and so the outputs are generally likely to be further from the truth than the raw data. We recommend that students of palaeodiversity cease to use residual modelling estimates based on formation counts, and suggest that results from a substantial number of papers published in the past ten years require re-evaluation.

Key words: sampling proxy, bias, palaeodiversity, redundancy, residual modelling, Phanerozoic.

THE fossil record provides the only direct evidence which can be used to analyse biodiversity patterns over extended periods of geological time (Raup 1972; Smith 2007). However, it is generally accepted that the fossil record is compromised by incompleteness and bias, and therefore numerous methods have been employed to try to recover a bias-free, or corrected, palaeodiversity signal (Raup 1972, 1976; Alroy *et al.* 2001; Peters & Foote 2001; Wang & Dodson 2006; Smith & McGowan 2007; Alroy 2010; Lloyd 2012; Starrfelt & Liow 2016).

A commonly implemented technique for correcting for sampling biases in palaeodiversity studies is the residual modelling method, proposed by Smith & McGowan (2007) and refined by Lloyd (2012). The method employs a model-fitting approach using sampling proxies to identify times of poor and good sampling and to apply post hoc corrections. The residual modelling method is worth exploring in some detail because it has become the method of choice for a large number of high-profile papers, many of which make radical claims about macroevolutionary consequences after correction of the data; it has been cited over 175 times (Google Scholar, September 2017).

The most common sampling proxies used in conjunction with the residual modelling technique are counts of fossiliferous formations per time bin (Fröbisch 2008; Barrett *et al.* 2009; Benson *et al.* 2010; Benson & Butler 2011; Benson & Upchurch 2013; Dean *et al.* 2016). These are normally compiled in one of two ways, either as: (1) strict fossiliferous formation counts that contain only fossils of members of the clade of interest, i.e. clade-bearing formations (CBFs) (Fröbisch 2008; Barrett *et al.* 2009; Butler *et al.* 2009); or (2) total fossiliferous formation counts (TFFs) (Marx & Uhen 2010), which include all fossiliferous formations in which particular fossils of the group in question could have possibly occurred (see Table 1 for a list of acronyms complete with explanations). Although it has been claimed that formation counts are suitable sampling proxies, as they summarize aspects of rock volume, facies heterogeneity, geographical and temporal dispersion, and research effort (Benson & Upchurch 2013), many studies have criticized the use of

doi: 10.1111/pala.12331

TABLE 1. Summary table of formation count acronyms used in this manuscript.

Acronym	Definition	Explanation
CBF	Clade bearing formations	Formations that only contain fossils of the clade of interest
WCBF	Wider clade bearing formations	Formations that contain fossils of the clade of interest and members of the wider clade to which the clade of interest belongs
PCBF	Potential clade bearing formations	Formations that preserve environments that might be expected to contain the clade of interest but have yet to yield any fossils
TFF	Total fossiliferous formations	All sedimentary formations containing fossils

such approximations (Crampton *et al.* 2003; Dunhill 2011; Dunhill *et al.* 2014*b*; Benton 2015). Perhaps the most prominent criticism has been the identification of redundancy between formation counts and raw palaeodiversity data (Benton *et al.* 2011; Dunhill *et al.* 2014*b*; Benton 2015) in other words, that one time series is as likely to drive the other as vice versa. For sparsely occurring fossils such as dinosaurs, the two time series are essentially the same (Benton 2015). Dunhill *et al.* (2014*b*) provided a quantitative assessment of this issue using Information Transfer (IT) (Hannisdal 2011*a*, *b*; Hannisdal & Peters 2011) and concluded that, in the British marine fossil record, the strong association between formation counts and raw diversity is best explained as a result of redundancy.

Here, we further test the redundancy hypothesis on both observed global fossil data from the Ordovician–Neogene, and on simulated data, expanding on the approach of Brocklehurst (2015). Specifically, we confirm two predictions of the redundancy hypothesis, namely that: (1) raw palaeodiversity is more information-redundant with CBFs than TFFs; and (2) even in a random world, CBFs can be driven more by diversity than by sampling.

MATERIAL AND METHOD

Data

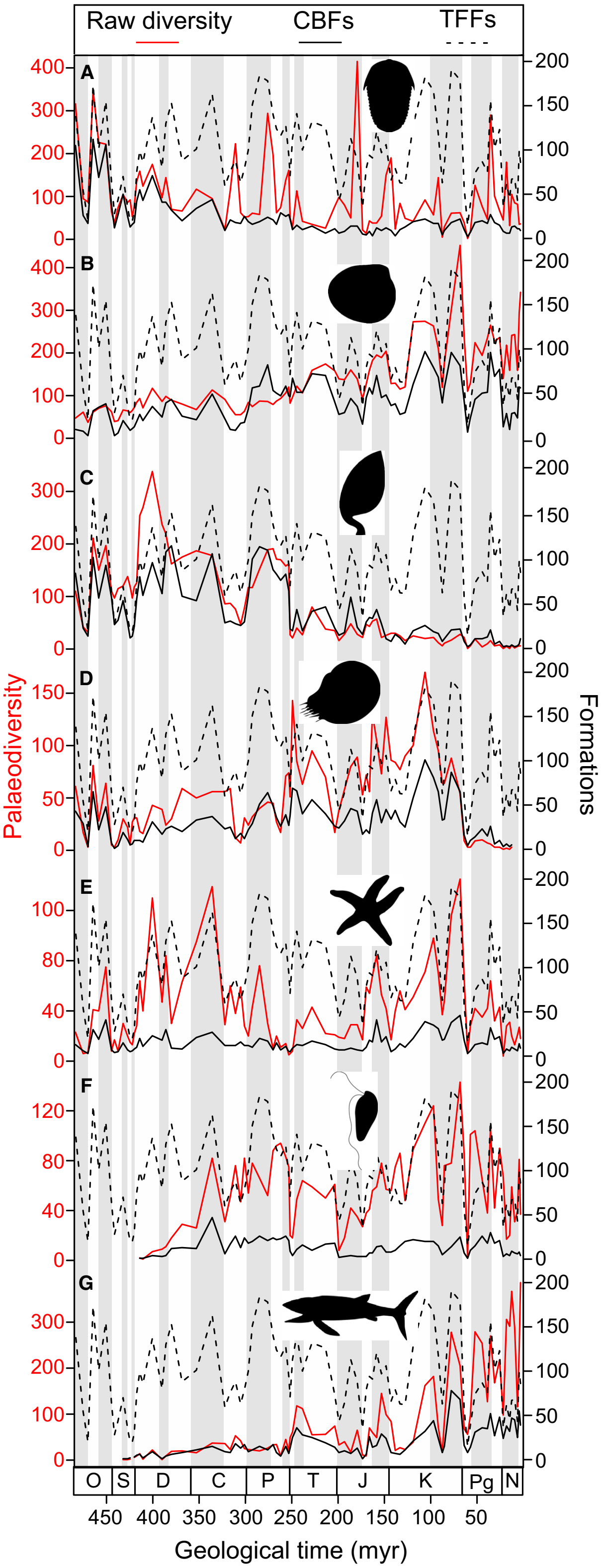
Global Phanerozoic generic occurrence databases for a number of major marine clades, Arthropoda, Bivalvia, Brachiopoda, Cephalopoda, Echinodermata, Foraminifera and Vertebrata, along with numbers of clade-bearing and total marine fossiliferous formations, were downloaded from the Paleobiology Database (PBDB) (http://paleobiodb.org; Clapham *et al.* 2015) (Fig. 1) (Dunhill *et al.* 2017).

Simulations

The simulations used in this study are an expansion of those presented by Brocklehurst (2015). The original simulation used a birth–death model with parameters for dispersal and local extinction to simulate the evolution of a clade over ten notional geographical regions. Within each region there were 10 localities, which would become the fossil-bearing localities sampled by palaeontologists. To simulate the taphonomic removal of specimens, each species was subject to random deletion from each locality in which it lived. The probability of a species being sampled at each locality is hereafter called PTAPH. To simulate the removal of fossil-bearing formations from the record, by erosion for example, being covered or simply not yet being found, entire regions were sampled with the probability PFORM. The number of regions sampled (not randomly deleted) in each time bin was stored to represent use of the number of fossil-bearing formations as a sampling proxy. Finally, individual localities within the sampled regions were sampled with a probability PLOC, and the number of localities sampled in each time bin was stored as a second sampling proxy. For further details of the original simulation see Brocklehurst (2015).

The simulation presented in this study includes two modifications to Brocklehurst (2015). First, having simulated a clade via the birth–death model, a smaller clade from within this was chosen at random to be the clade of interest. This allowed a comparison of two different classes of sampling proxy: formations bearing the clade of interest (CBFs), and formations bearing the larger clade containing the clade of interest, the wider clade-bearing formations (WCBFs). This latter class of proxy has been used in a number of studies (Marx & Uhen 2010; Mannion *et al.* 2011; Brocklehurst *et al.* 2012, 2013) in an attempt to circumvent the issues of redundancy and non-occurrences (palaeontologists having examined rocks of a particular age but not having found representatives of the clade of interest). As noted before (e.g. Benton *et al.* (2011), palaeontologists often used CBFs as yardsticks of sampling but, unlike in standard ecological sampling theory, they ignored null returns (non-occurrences), so choosing to exclude powerful evidence for relative sampling quality between temporal or spatial bins.

The second change was to add another ten regions to the simulated landscape, which the simulated clade would not be permitted to enter. However, these regions and the localities within were subjected to random deletion as described above, and could be counted towards a sampling proxy (TFFs). This allows assessment of whether all sampled formations/localities should be included in a sampling proxy, or whether the analyst should be more selective and only count those in which the clade of interest could have lived; for example, when examining a clade found only in shallow marine shelf environments, should deep marine formations be included?



With these additions to the simulation, four classes of proxy could be generated: (1) sampled formations/localities bearing the clade of interest (CBFs); (2) sampled formations/localities bearing the wider clade containing the clade of interest (WCBFs); (3) sampled formations/localities which the clade of interest could potentially have entered, i.e. potential clade-bearing formations (PCBFs); and (4) all sampled formations/localities (TFFs). Localities were not tested because the original simulations showed that they were a poor proxy for sampling (Brocklehurst 2015). The code for the simulations is presented in Dunhill *et al.* (2017).

Eight clades were simulated (Dunhill *et al.* 2017). Speciation and origination rates were equal, so the diversification of the clade proceeded via a random walk. Clades that did not survive for at least 100 time bins were discarded, as were clades containing fewer than 1000 and more than 3000 taxa. The birth–death model output was converted to a phylogeny, and nodes were selected at random one at a time. Once a node containing at least 25% of the total number of taxa but no more than 75% was selected, this became the clade of interest.

For each clade, we varied PFORM amongst time bins in order to capture fluctuating sampling levels through the time series. Therefore, we can test whether formation residual corrections perform better than raw diversity when sampling is heterogeneous and test the level of sampling at which raw taxonomic counts stop being more reliable proxies for true diversity. Hence, we are giving formation counts the best possible chance of being a good proxy for sampling (as varying sampling is dictated entirely by PFORM). Three sets of simulations were carried out incorporating different degrees of variation of PFORM. The first allowed PFORM to vary from 0.1 to 0.9, the second from 0.2 to 0.6, the third from 0.3 to 0.5. In each time bin in each simulation, a value of PFORM was selected at random from a uniform distribution covering the permitted range of values. PLOC and PTAPH did not vary across time bins within any of these simulations.

At each sampling level, 100 simulations were run. For each, a raw diversity estimate was calculated along with the four classes of formation-based sampling proxies.

FIG. 1. Marine palaeodiversity (generic richness), total fossiliferous formation counts (TFFs) and clade-bearing formation counts (CBFs) from the Ordovician to the Neogene as downloaded from the PBDB (Clapham *et al.* 2015) for: A, arthropods; B, bivalves; C, brachiopods; D, cephalopods; E, echinoderms; F, foraminifera; G, vertebrates. Colour online.

These were used to calculate four residual diversity estimates. The coefficient of determination (R^2) was used to quantify the shared variation between the residual diversity estimates and the true diversity estimates, and also shared variation between the sampling proxies and both the true diversity and the raw diversity estimate. Examples of the simulated diversity estimates and proxies are shown in Figure 2.

In the event, all eight clades showed extremely similar results across all sampling regimes, and so we report our statistical analyses for one example clade. For the lowest sampling level (0.1), the simulated raw diversity was in some cases too sparse for statistical analysis, hence we report our analyses of the example clade at three levels of PLOC and PTAPH: 0.2, 0.5 and 0.8.

Statistical analysis

Here we are primarily interested in quantifying the relative strength of statistical associations between sampling proxies and raw diversity, and, in the simulated cases, between sampling proxies, residual diversity estimates, true diversity, and true sampling. For the empirical data, we first carried out pairwise Spearman rank-order correlation tests between formation counts and raw diversity, detrended by generalized differencing, using an R script by G. Lloyd (http://www.graemetlloyd.com/pubdata/functions_2.r). False discovery rate corrections for multiple comparisons were applied using the method of Benjamini & Hochberg (1995).

Next, we evaluated the relative strength of statistical associations between pairs of time series using two quantities: (1) the coefficient of determination (R^2), calculated as the square of the Pearson product-moment correlation, equivalent to a linear regression with intercept term; (2) pairwise information transfer (IT), a more generalized measure of shared information, calculated in each direction, X→Y and Y→X (Hannisdal 2011*a*, *b*). To minimize directional bias due to differences in non-stationarity, time series were detrended (linearly, or, if necessary, using a higher-order polynomial fit) to satisfy a stationarity criterion (Kwiatkowski *et al.* 1992). All records were log transformed and normalized to mean zero and unit standard deviation prior to analysis. To quantitatively characterize the robustness of the IT and R^2, each time series was randomly subsampled across a spectrum of sample sizes, down to half the total number of time bins, with 200 iterations for each sample size. IT and R^2 results were then integrated across the subsampling spectrum. For each iteration, we computed the corresponding IT and R^2 values for 500 amplitude-adjusted Fourier transform surrogate time series.

Note that we use R^2 as a relative measure of shared variation, not as a basis for significance testing. We thus calculated R^2 on both raw time series and on the detrended series used in IT analysis. Similarly, we use IT primarily as a relative measure of generalized statistical dependence (mutual information), but we only report pairwise IT results for detrended data to avoid non-stationarity bias. For the residual diversity simulations, we evaluated the relative degree of association in sets of three variables using two quantities: (1) the partial rank (Spearman) correlation between X and Y, conditioned on a third variable Z; (2) conditional IT between X and Y when taking into account shared information with a third variable, Z. The analytical settings for conditional or partial analyses were identical to the pairwise analyses, except that the data were not detrended, iterations corresponded to simulation runs, and the surrogates were random shuffles of the original data.

To facilitate the comparison of IT and correlation-based results, all values are reported as relative to the 99th percentile of a distribution of values calculated for 500 surrogate time series (Fig. 3). The surrogates can be thought of as a null distribution unique to each combination of variables, but here we are not focusing on hypothesis (significance) testing. For the simulation results, significance testing is unwarranted. Instead, we are interested in the relative strength of statistical association as a measure of the degree of shared variation and information redundancy.

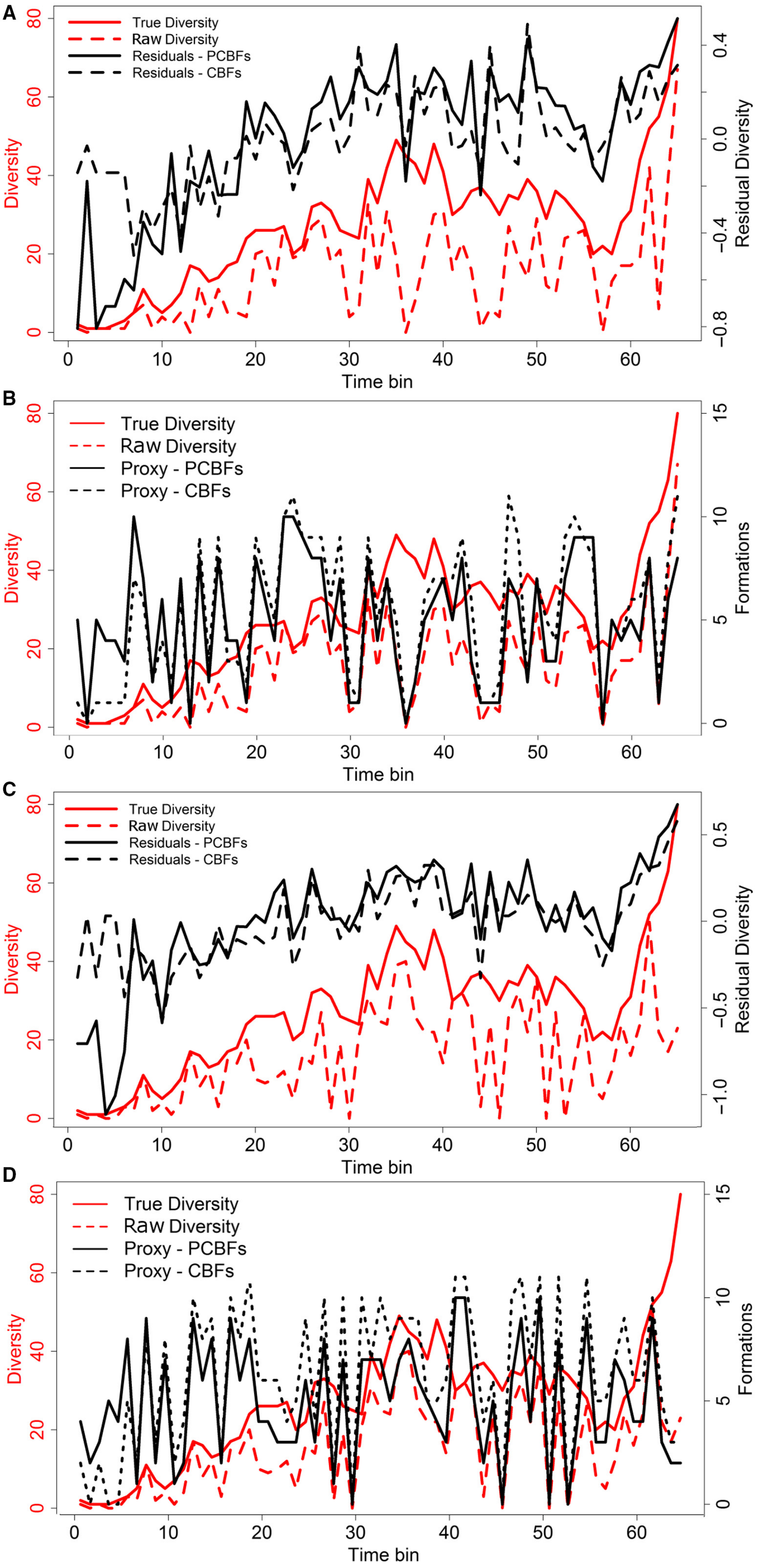
RESULTS

Empirical data

All clades in the analyses of empirical data show closer correlations between CBFs and raw diversity than between TFFs and raw diversity (Table 2). All clades show significant correlation with TFFs, although the correlations for arthropods and vertebrates become non-significant after correction for multiple comparisons (Table 2).

There is strong bidirectional IT between raw diversity and CBFs in all clade data sets (Fig. 4). In general, there is less IT between raw diversity and TFFs (Fig. 4). On average, only bivalves and echinoderms have detectable IT with TFFs (i.e. the median being above the zero line). The coefficient of determination (R^2) for the raw time series (Fig. 5A) and the detrended data (Fig. 5B) is consistent with the correlation and IT results in terms of the relative strength of CBFs and TFFs with respect to raw diversity within each clade. The R^2 on detrended records (Fig. 5B) is also in broad agreement with the pairwise IT analysis (Fig. 4), which is to be expected when the statistical associations are predominantly linear or monotonic: Within lineages, raw diversity is more strongly associated

FIG. 2. Examples of simulation output, illustrating the true diversity, the raw diversity, clade-bearing formation (CBF) and potential clade bearing formation (PCBF) proxies and residual diversity estimates calculated from these proxies. PFORM varies between 0.1 and 0.9 in each time bin. A–B, PLOC (random sampling of individual localities) and PTAPH (random sampling for taphonomic reasons) = 0.5; A, individual localities (PLOC). C–D, PLOC and PTAPH = 0.8. Colour online.



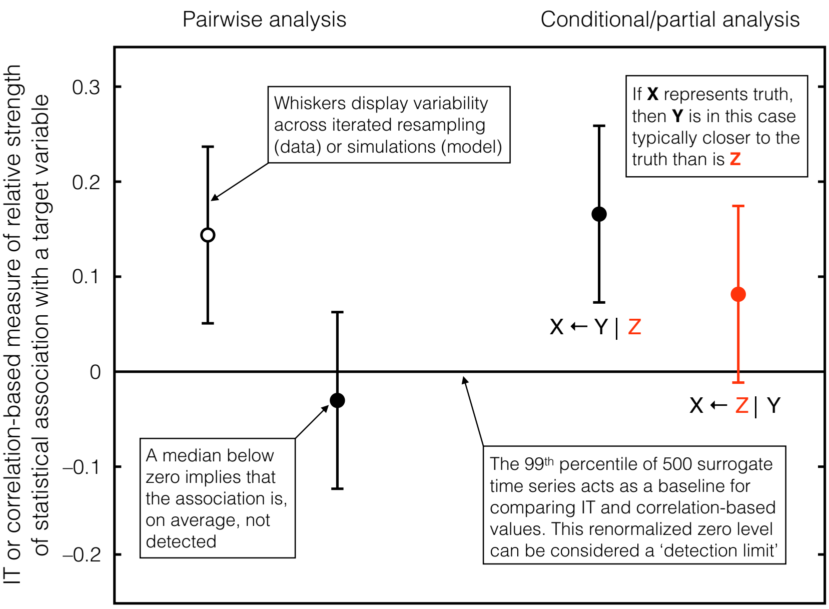

Pairwise analysis
Conditional/partial analysis
IT or correlation-based measure of relative strength of statistical association with a target variable
0.3
0.2
0.1
0
–0.1
–0.2
Whiskers display variability across iterated resampling (data) or simulations (model)
If X represents truth, then Y is in this case typically closer to the truth than is Z
X ← Y | Z
X ← Z | Y
A median below zero implies that the association is, on average, not detected
The 99th percentile of 500 surrogate time series acts as a baseline for comparing IT and correlation-based values. This renormalized zero level can be considered a 'detection limit'

FIG. 3. Schematic guide to interpreting the results shown in Figs 4, 5, 7, 8. For each combination of variables, the values for IT (pairwise or conditional) and the correlation-based measures (R^2 or partial correlation) are reported as distributions (dots and whiskers). These distributions are obtained, either by iterative resampling of the empirical data or by iterative simulation runs, to convey the variability of each statistical association. Filled or open dots are medians, and the whiskers represent the 99% range of values. All values are relative to the 99th percentile of the surrogates, as a reference line for comparing IT and correlation-based results. Colour online.

with CBFs than TFFs, and across lineages, TFFs tend to be below zero, on average (median).

Simulated data

The simulations demonstrate, as expected, that as sampling increases, the various residual diversity estimates display an increasingly better fit to true diversity, and the fit of the various proxies to the raw diversity decreases (Fig. 6A). CBFs show by the far the best fit to the raw diversity, but when used to calculate the residual diversity estimate, produce by the far the worst fit to the true diversity, even worse than the raw diversity estimate at low sampling levels. TFFs sampled in each time bin have the worst fit to the raw diversity. However, the best residual estimates are produced using formations that the clade of interest was 'allowed' to enter (i.e. PCBFs), despite the poor fit between this proxy and raw diversity (Fig. 6B). Residuals calculated using WCBFs fit the raw diversity estimate better than those calculated using CBFs, but not so well as those calculated using PCBFs. When we compare the four proxies to PFORM (our actual measure of true sampling heterogeneity) TFF shows the best correlation, in spite of performing extremely poorly when used to calculate residual diversity (Fig. 6C). This shows that the best proxy for indicating sampling heterogeneity is not the best proxy for producing residual diversity estimates.

Conditional IT analysis of the simulated data shows that residual diversity estimates based on CBFs are worse predictors of true diversity than raw diversity unless the sampling level is very high (i.e. PLOC and PTAPH values are high and PFORM variability is high) (Fig. 7). However, residual diversity estimates based on WCBFs, PCBFs and TFFs all show improved estimates of true diversity relative to raw diversity. The best predictor of true diversity comes from residual diversity estimates based on PCBFs (Fig. 7). The conditional IT and partial correlations agree when it comes to the relative strength of associations (Fig. 7). By looking at the relative strength of influence of true diversity and sampling on the different proxies, we can tease apart redundancy between the various proxy classes and true diversity. Conditional IT from true diversity to CBFs, beyond shared information with sampling (i.e. PFORM), demonstrates why residual diversity based on CBFs performs worse than the other proxies. At low sampling levels, CBFs are driven more by the true diversity of the clade of interest than by sampling (i.e. PFORM), therefore we can interpret this as information redundancy between CBFs and true diversity (Fig. 8). However, as PFORM variability increases, the sampling signal swamps all other signals in the proxy. As we

TABLE 2. Pairwise Spearman's rank-order correlation coefficients between palaeodiversity and formation counts.

	CBFs	TFFs
Arthropoda	Δ 0.64**	Δ 0.32*
Bivalvia	Δ 0.59**	Δ 0.52**
Brachiopoda	Δ 0.66**	Δ 0.56**
Cephalopoda	Δ 0.67**	Δ 0.44**
Echinodermata	Δ 0.78**	Δ 0.59**
Foraminifera	Δ 0.73**	Δ 0.52**
Vertebrata	Δ 0.8**	Δ 0.32*

*Significant at $p < 0.05$; **significant after FDR correction
Δ After generalized differencing.

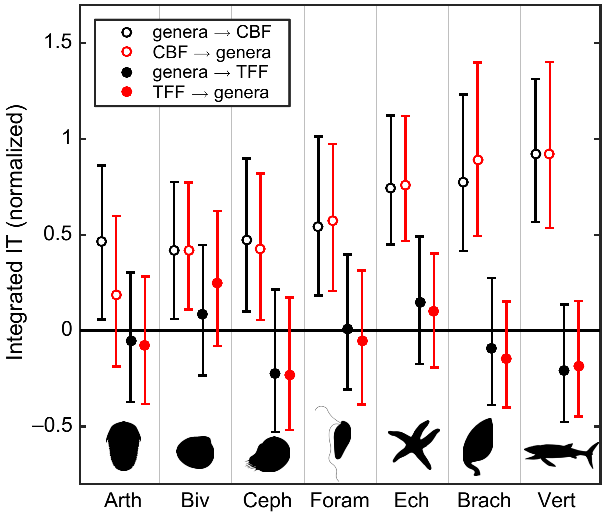


FIG. 4. Pairwise directional IT analysis of the time series in Fig. 1, comparing palaeodiversity (genera) to clade-bearing formation counts (CBFs; open symbols) and total fossiliferous formation counts (TFFs; closed symbols). Values are medians (circles) and 99% ranges (whiskers) of sampling iterations, integrated over a spectrum of sample sizes. Values are scaled relative to the 99th percentile of the surrogates (representing the zero level). See text for details. Colour online.

increase the phylogenetic scope of the formation counts, from WCBFs via PCBFs to TFFs, the sampling signal becomes more dominant, because any diversity signal in these proxies is less redundant with the true diversity of the sub-clade of interest (Fig. 8). Again, the conditional IT results agree with the partial correlation results (Fig. 8).

DISCUSSION

The strong correlations observed between CBFs and raw diversity could indicate severe temporal sampling bias in the fossil record. It could be said that the weaker correlations observed between TFFs and raw diversity (or complete lack of correlation in some clades) mean that TFFs are not as effective a sampling proxy as CBFs. However, when we consider that the IT between CBFs and raw diversity is strongly symmetrical (Fig. 4), it is not possible to claim that one signal drives the other more than vice versa. The relative strength of relationships between the CBFs/TFFs and raw diversity is confirmed by the R^2 values (Fig. 5). This result is supported by the simulation results, which not only show that, at low sampling levels, residual diversity estimates based on CBFs are worse predictors of true diversity than the raw data (e.g. Fig. 7B), despite showing strong correlation with raw diversity, but also that CBFs are driven more by true diversity than they are by sampling intensity, even when we give formations the best possible chance of being indicative of sampling (e.g. Fig. 8B). It is, therefore, a cause for concern that studies that have used CBFs as sampling proxies, such as Fröbisch (2008), Barrett *et al.* (2009), Butler *et al.* (2009), Benson & Upchurch (2013) and Dean *et al.* (2016) have reached macroevolutionary conclusions based on sampling-

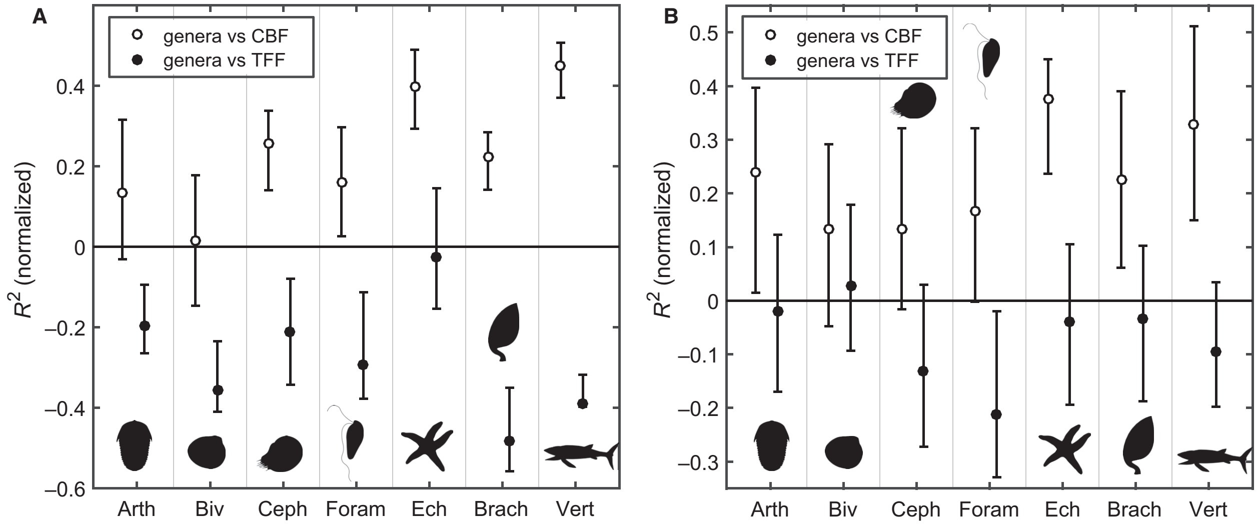


FIG. 5. R^2 values comparing palaeodiversity (genera) to clade-bearing formation counts (CBF) and total fossiliferous formation counts (TFF). A, raw time series. B, detrended time series. Data from Fig. 1. Values are medians (circles) and 99% ranges (whiskers) of sampling iterations, integrated over a spectrum of sample sizes. For comparison with the pairwise IT, the R^2 values are scaled relative to the 99th percentile of the surrogates (representing the zero level). See text for details.

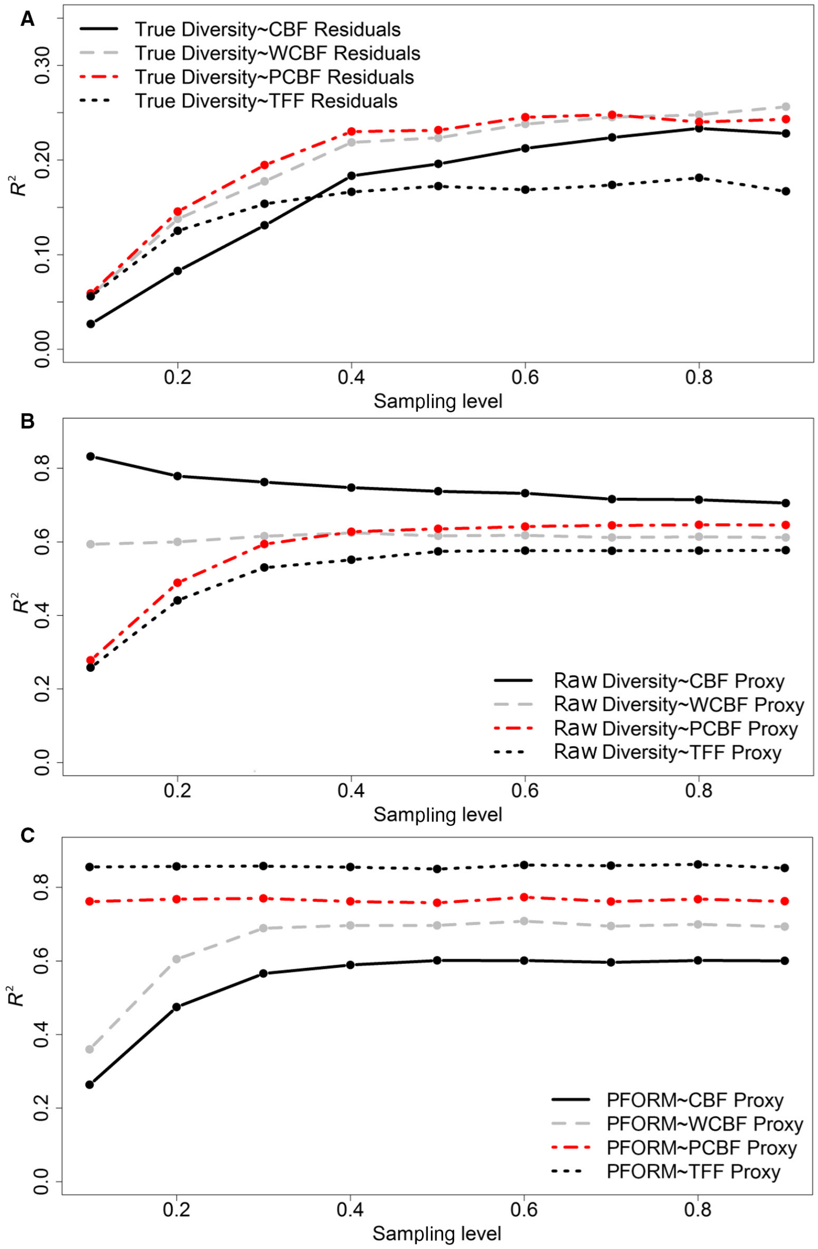
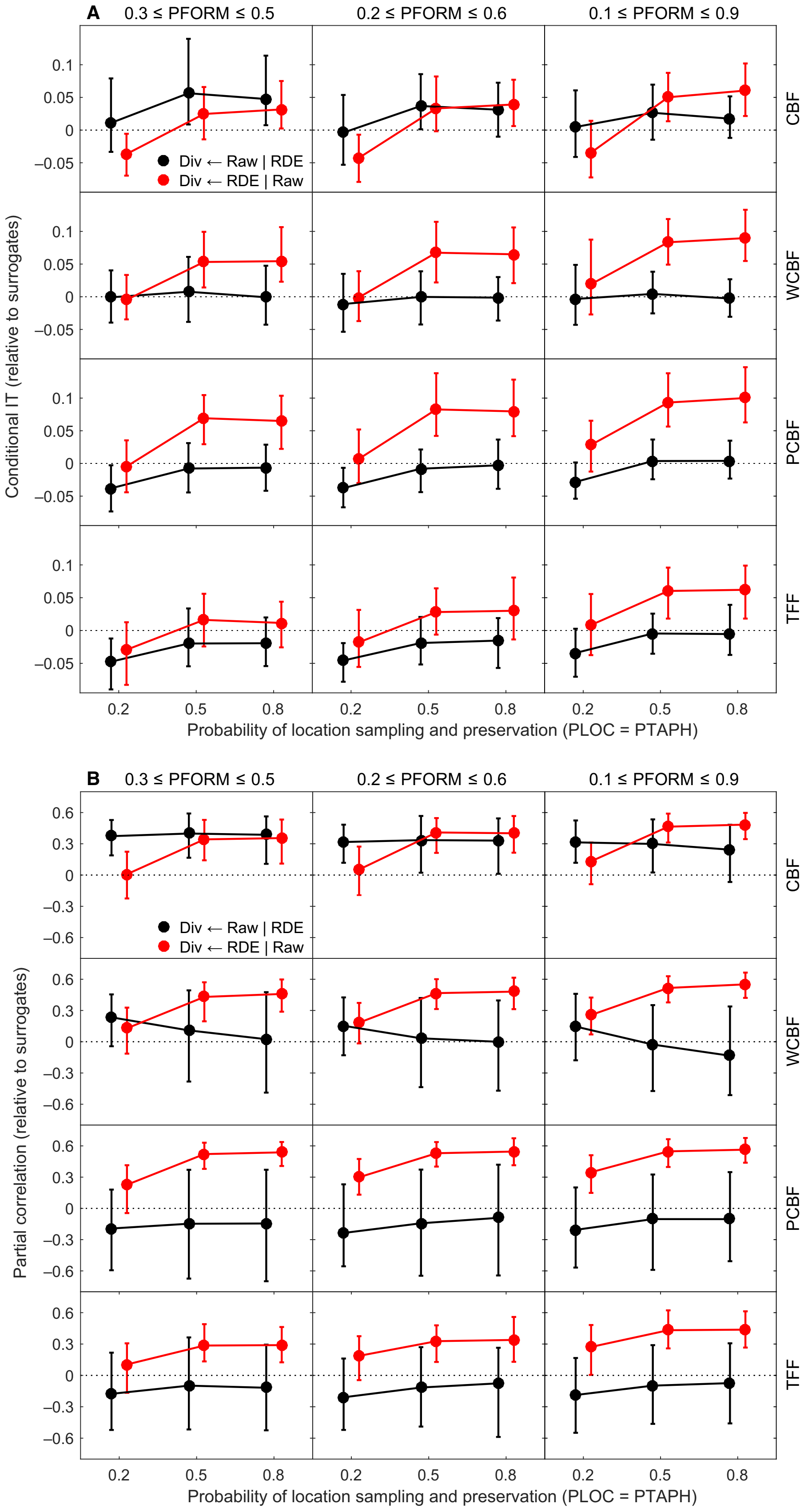
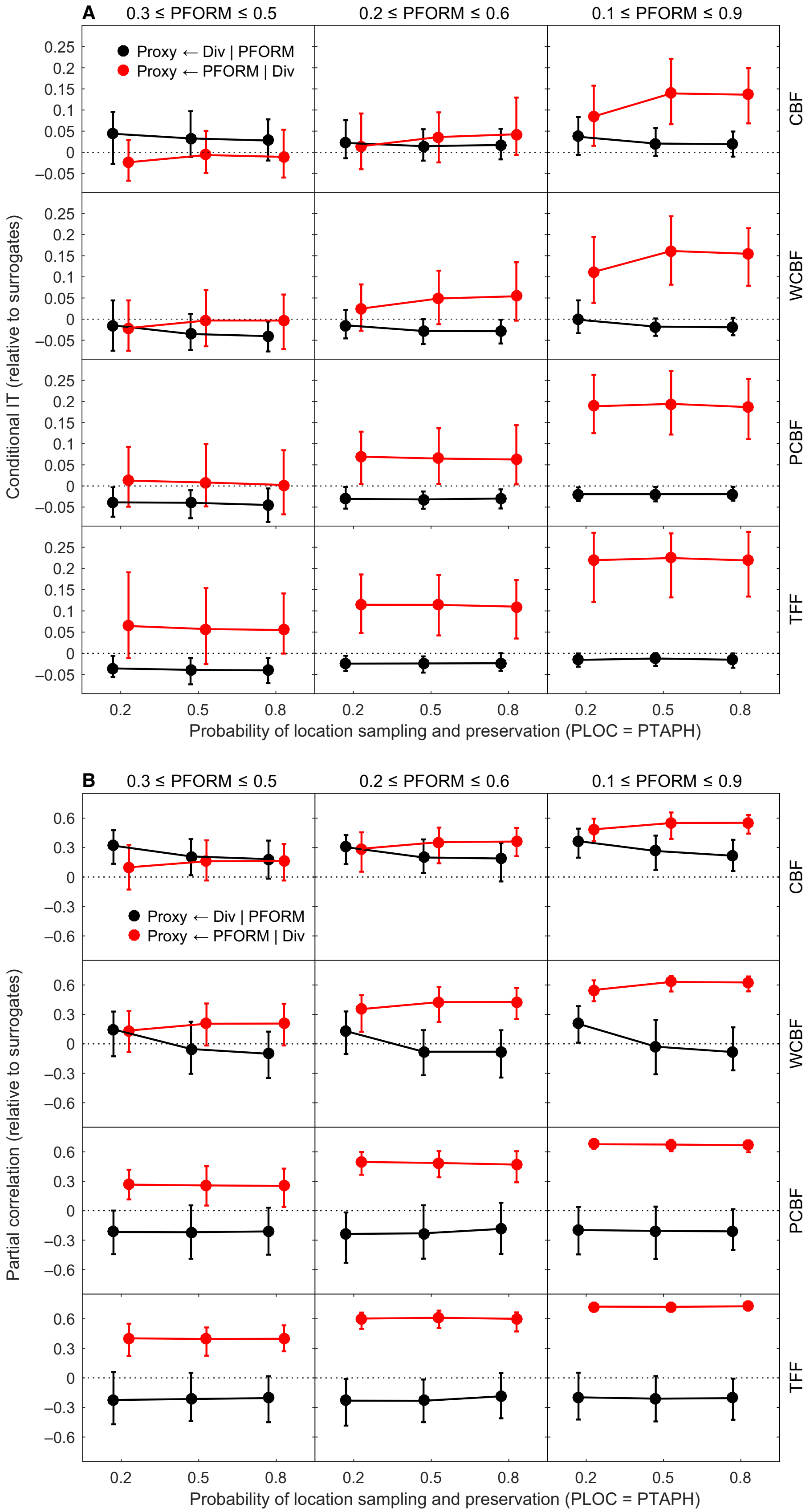


FIG. 6. Examples of simulation output, showing R^2 values comparing raw diversity to true and residual diversity estimates and values of PFORM. A, comparing true diversity to residual diversity estimates calculated using each of the four proxies. B, comparing raw diversity estimates to each of the four proxies. C, comparing each of the four proxies to PFORM. Colour online.

FIG. 7. At low sampling levels, residual diversity estimates using clade-bearing formation counts (CBFs) are further from the truth than raw data. Plotted values are integrated, conditional IT (A) and partial correlation coefficients (B) on simulated time series of true diversity (Div), raw sampled diversity (Raw), and residual diversity estimates (RDE). Panel rows correspond to the four different proxy classes (CBF, WCBF, PCBF and TFF) used to compute the RDE. Panel columns correspond to different levels of variability in sampling (PFORM), increasing from left to right. Values on the abscissa represent the probability of location sampling (PLOC) and preservation (PTAPH). Conditional IT from raw to true diversity, given RDE (Div ← Raw | RDE) quantifies how much information raw diversity provides on true diversity beyond the information already contained in RDE. Conversely, conditional IT from RDE to true diversity, given raw diversity (Div ← RDE | Raw) quantifies how much information RDE provides on true diversity beyond the information already contained in raw diversity. Partial correlations can be interpreted analogously. All IT and correlation values are relative to the 99th percentile of surrogate time series. Filled circles are medians, and error bars encompass the range of IT or correlation values obtained across the 100 simulation runs, with 500 surrogates for each run. Colour online.


A
0.3 ≤ PFORM ≤ 0.5
0.2 ≤ PFORM ≤ 0.6
0.1 ≤ PFORM ≤ 0.9
CBF
WCBF
PCBF
TFF
Conditional IT (relative to surrogates)
0.1
0.05
0
−0.05
● Div ← Raw | RDE
● Div ← RDE | Raw
0.2
0.5
0.8
Probability of location sampling and preservation (PLOC = PTAPH)
B
0.3 ≤ PFORM ≤ 0.5
0.2 ≤ PFORM ≤ 0.6
0.1 ≤ PFORM ≤ 0.9
Partial correlation (relative to surrogates)
0.6
0.3
0
−0.3
−0.6
● Div ← Raw | RDE
● Div ← RDE | Raw
CBF
WCBF
PCBF
TFF
0.2
0.5
0.8
Probability of location sampling and preservation (PLOC = PTAPH)


A
0.3 ≤ PFORM ≤ 0.5
0.2 ≤ PFORM ≤ 0.6
0.1 ≤ PFORM ≤ 0.9
● Proxy ← Div | PFORM
● Proxy ← PFORM | Div
Conditional IT (relative to surrogates)
0.25
0.2
0.15
0.1
0.05
0
−0.05
CBF
WCBF
PCBF
TFF
0.2
0.5
0.8
Probability of location sampling and preservation (PLOC = PTAPH)
B
0.3 ≤ PFORM ≤ 0.5
0.2 ≤ PFORM ≤ 0.6
0.1 ≤ PFORM ≤ 0.9
● Proxy ← Div | PFORM
● Proxy ← PFORM | Div
Partial correlation (relative to surrogates)
0.6
0.3
0
−0.3
−0.6
CBF
WCBF
PCBF
TFF
0.2
0.5
0.8
Probability of location sampling and preservation (PLOC = PTAPH)

corrected data that may be further from the truth than the original raw data. Our results emphasize that the strength of the correlation between sampling proxies and raw diversity is not a good indicator of the degree of sampling bias in the data.

It has previously been suspected that formation counts and raw palaeodiversity are information redundant, for two reasons: tallying and comparability. The palaeontological literature grows by the discovery of new fossils, and as those fossils are added to the catalogues of palaeodiversity, so too are the formations in which they occur. This is why, for sparsely occurring fossils at least, the tally of known fossils of group A (i.e. diversity), and the count of known formations yielding fossils of group A (the CBF, i.e. dispersal), both grow in tandem and both equally reflect intensity of sampling (Benton 2015). Therefore, they both equally reflect estimated palaeodiversity and sampling effort. The comparability argument can be put in two ways: first, geological formations vary in volume over eight orders of magnitude, from 0.073–225 000 km^3 (Benton *et al.* 2011; Dunhill *et al.* 2012) so they are a poor means of binning data and making comparisons of any kind. Second, their definition can depend on the richness of their fossil content (an aspect of redundancy that the simulations cannot account for) so greater environmental or faunal turnover may enable finer stratigraphic partitioning (Dunhill *et al.* 2014*a*, *b*). It is highly likely, especially in the case of CBFs, where formations that only contain the specific fossil group under study are used as sampling proxies, that raw palaeodiversity drives formation counts every bit as much as formation counts drive raw palaeodiversity. Even total formation counts may be inextricably linked to palaeodiversity, as exemplified by the Triassic–Jurassic fossil record of Great Britain, where formation boundaries are not independent of changes in fossil richness (Dunhill *et al.* 2014*b*). Benton (2015) showed that the discovery of new dinosaurs through research time has been closely linked to the discovery of new dinosaur-bearing formations (or the reverse), and the fact that they covary does not mean that one signal can be used to correct the other. Thus, CBFs only allow for the quantification of what has already been sampled and make no allowance for future possibilities to sample, whereas WCBFs, PCBFs, and TFFs offer some perspective on the ‘unknown’ (Benton *et al.* 2011, 2013). Therefore, sampling proxies that include formations that have not yielded fossils of the clade in question *a priori* represent a better sampling proxy than CBFs, because the former is a closer approximation of supposed total sampling effort and its underlying driver (i.e. availability of sedimentary rock) whilst the latter ignores all sampling that failed to find the group in question. However, the simulations assume that the definition of formations is random with respect to fossil diversity, which is likely to be violated in the real world due to mutual dependencies with facies changes. In addition to this, our empirical results show that IT is still symmetrical, although much weaker between TFFs and raw diversity, suggesting that this supposed sampling proxy may also be information redundant with raw diversity, despite not being as closely linked.

The simulations do show that including formations containing the wider clade of interest or, better still, the potential to include the clade of interest, produces residual diversity estimates that are more similar to true diversity than the raw data, albeit in a best-case scenario where formations equal true sampling. We also show that all sampling proxies perform better at higher levels of sampling. This gives some confidence in using residual modelling as a conservative method to correct for sampling bias in the fossil record, but only if the correct sampling proxy is used and our sampling of the fossil record is already very good. Therefore, the choice of sampling proxy is important and the problem therein is to ensure that the correct sampling proxy is employed. For example, the best performing sampling proxies are counts of formations that could potentially yield fossils of the clade of interest. How do we go about defining which formations might preserve our clade of interest? Whatever the answer, it is unavoidably highly subjective. It may be

FIG. 8. At low sampling levels, the clade-bearing formation (CBF) sampling proxy is driven more by diversity than by sampling. In the simulations, sampling variability is entirely driven by PFORM, and entirely random with respect to true diversity. Plotted values are integrated, conditional IT (A) and partial correlation coefficients (B) on simulated time series of a sampling proxy (Proxy), true diversity (Div), and true sampling (PFORM). Panel rows correspond to the four different proxy classes (CBF, WCBF, PCBF and TFF). Panel columns correspond to different levels of variability in sampling (PFORM), increasing from left to right. Values on the abscissa represent the probability of location sampling (PLOC) and preservation (PTAPH). Conditional IT from true diversity to the sampling proxy, given true sampling (Proxy ← Div | PFORM) quantifies how much information true diversity provides on the sampling proxy, beyond the information already contained in PFORM. Conversely, conditional IT from PFORM to the sampling proxy, given true diversity (Proxy ← PFORM | Div) quantifies how much information PFORM provides on the sampling proxy beyond the information already contained in true diversity. Partial correlations can be interpreted analogously. All IT and correlation values are relative to the 99th percentile of surrogate time series. Filled circles are medians, and error bars encompass the range of IT or correlation values obtained across the 100 simulation runs, with 500 surrogates for each run. Colour online.

easier to define a higher clade of interest, but there is still the question of how wide a clade is required to reach the optimum in estimating sampling intensity. Neither should we rely on a sampling correction method that cannot cope with poor sampling, particularly as the residual modelling approach has been widely used when the empirical fossil record (e.g. vertebrates) has been deemed too poor for sampling standardization approaches.

CONCLUSION

Our results indicate that the close correlations commonly observed between formation counts and raw palaeodiversity are often the result of information redundancy, rather than evidence of large-scale temporal sampling biases. It is therefore inadvisable to use metrics of formation counts to 'correct' raw fossil diversity using a residual modelling approach for two reasons: (1) easily definable formation counts produce inaccurate residual diversity estimates, i.e. CBFs, are information redundant with regard to raw sampled diversity and perform increasingly poorly as sampling levels are degraded; and (2) formation counts, shown to produce more accurate residual diversity estimates in simulated data, are difficult and subjective to define in empirical studies. Coupled with results from similar simulation studies showing that the residual modelling technique performs less well than phylogenetic diversity estimates and even more poorly than raw fossil diversity estimates when using CBFs (Brocklehurst 2015), the fact that formation counts have been repeatedly shown to be a poor proxy for actual sampling effort (Crampton *et al.* 2003; Dunhill 2011) and a recent critique of the methodology of residual modelling showing that there are objective statistical flaws in the residual modelling method as most commonly applied (Sakamoto *et al.* 2017), our results suggest that residual modelling using sampling proxies is not an appropriate method for correcting for temporal sampling biases in the fossil record.

Our study does not mean that the fossil record is not biased, it undoubtedly is, but there are much more appropriate methods available to palaeontologists to address this issue than residual modelling based on formation counts (Smith 2007; Alroy 2010; Benton *et al.* 2011; Hannisdal *et al.* 2012, 2017; Liow 2013; Starrfelt & Liow 2016; Walker *et al.* 2017). Second, any prior assumption that the fossil record is generally very poor and biased in a major way should be reconsidered on a case-by-case basis. It might just be that the fossil record is adequate, via the application of appropriate analytical techniques, for many of the macroevolutionary and palaeobiological questions that interest modern palaeontologists.

Acknowledgements. We thank James Crampton, Noel Heim and Andrew Smith for providing constructive critiques which greatly helped improve this manuscript. We thank Sally Thomas for providing editorial comments. This is Paleobiology Database official publication 294. AMD is funded by a Leverhulme Early Career Fellowship (ECF-2015-044) and a NERC research grant (NE/P013724/1). BH is funded by the Research Council of Norway grant 231259, and by the Bergen Research Foundation. NB is funded by Deutsche Forschungsgemeinschaft grant FR 2457/5-1, awarded to Professor Jörg Fröbisch.

Author contributions. AMD, BH and MJB conceived the project. AMD processed the data. BH and AMD performed the statistical analyses. NB performed the simulations. AMD led the writing of the manuscript. All authors reviewed and edited the final manuscript.

DATA ARCHIVING STATEMENT

Data for this study are available in the Dryad Digital Repository: https://doi.org/10.5061/dryad.rb86d

Editor. Andrew Smith

REFERENCES

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

—— 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., OLSZEWSKI, T. D., PATZKOWSKY, M. E., RAUP, D. M., ROY, K., SEPKOSKI, J. J. J., 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.

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

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

BENSON, R. B. J. and BUTLER, R. J. 2011. Uncovering the diversification history of marine tetrapods: ecology influences the effect of geological sampling biases. *Geological Society, London, Special Publications*, **358**, 191–208.

—— and UPCHURCH, P. 2013. Diversity trends in the establishment of terrestrial vertebrate ecosystems: interactions between spatial and temporal sampling biases. *Geology*, **41**, 43–46.

—— BUTLER, R. J., LINDGREN, J. and SMITH, A. S. 2010. Mesozoic marine tetrapod diversity: mass extinctions and temporal heterogeneity in geological megabiases affecting vertebrates. *Proceedings of the Royal Society B*, **277**, 829–834.

BENTON, M. J. 2015. Palaeodiversity and formation counts: redundancy or bias? *Palaeontology*, **58**, 1003–1029.

—— DUNHILL, A. M., LLOYD, G. T. and MARX, F. G. 2011. Assessing the quality of the fossil record: insights from vertebrates. *Geological Society, London, Special Publications*, **358**, 63–94.

—— RUTA, M., DUNHILL, A. M. and SAKAMOTO, M. 2013. The first half of tetrapod evolution, sampling proxies, and fossil record quality. *Palaeogeography, Palaeoclimatology, Palaeoecology*, **372**, 18–41.

BROCKLEHURST, N. 2015. A simulation-based examination of residual diversity estimates as a method of correcting for sampling bias. *Palaeontologia Electronica*, **18.3.7T**, 1–15.

—— UPCHURCH, P., MANNION, P. D. and O'CONNOR, J. 2012. The completeness of the fossil record of Mesozoic birds: implications for early avian evolution. *PLoS One*, **7**, e39056.

—— KAMMERER, C. F. and FRÖBISCH, J. 2013. The early evolution of synapsids, and the influence of sampling on their fossil record. *Paleobiology*, **39**, 470–490.

BUTLER, R. J., BARRETT, P. M., NOWBATH, S. and 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.

CLAPHAM, M. E., KIESSLING, W., UHEN, M. D., HENDY, A. J. W., ABERHAN, M., MILLER, A. I., KRÖGER, B., ALROY, J., FURSICH, F. T., HOLLAND, S. M., PATZKOWSKY, M. E., FOOTE, M., PÁLFY, J., BOTTJER, D. J., VILLIER, L., McGOWAN, A. J., WAGNER, P. J., IVANY, L. C., SESSA, J. A., NOVACK-GOTTSHALL, P. M., OLSZEWSKI, T. D., HOPKINS, M. J., SCHWEITZER, C. E., MANNION, P. D., BENSON, R. B. J., BUTLER, R. J., MUELLER, J., JARAMILLO, C. and CARRANO, M. T. 2015. Taxonomic occurrences of Ordovician to Neogene Arthropoda, Bivalvia, Brachiopoda, Cephalopoda, Echinodermata, Foraminifera, and Vertebrata. *Paleobiology Database*, accessed 21 April 2015. http://paleodb.org

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

DEAN, C. D., MANNION, P. D. and BUTLER, R. J. 2016. Preservational bias controls the fossil record of pterosaurs. *Palaeontology*, **59**, 225–247.

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

—— BENTON, M. J., TWITCHETT, R. J. and NEWELL, A. J. 2012. Completeness of the fossil record and the validity of sampling proxies at outcrop level. *Palaeontology*, **55**, 1155–1175.

—— —— —— —— 2014*a*. Testing the fossil record: sampling proxies and scaling in the British Triassic–Jurassic. *Palaeogeography Palaeoclimatology Palaeoecology* **404**, 1–11.

—— HANNISDAL, B. and BENTON, M. J. 2014*b*. Disentangling geological megabias and common-cause from redundancy in the British fossil record. *Nature Communications*, **5**, 4818.

—— —— BROCKLEHURST, N. and BENTON, M. J. 2017. Data from: On formation-based sampling proxies and why they should not be used to correct the fossil record. *Dryad Digital Repository*. https://doi.org/10.5061/dryad.rb86d

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

HANNISDAL, B. 2011*a*. Detecting common-cause relationships with directional information transfer. *Geological Society, London, Special Publications*, **358**, 19–29.

—— 2011*b*. Non-parametric inference of causal interactions from geological records. *American Journal of Science*, **311**, 315–334.

—— and PETERS, S. E. 2011. Phanerozoic earth system evolution and marine biodiversity. *Science*, **334**, 1121–1124.

—— HENDERIKS, J. and LIOW, L. H. 2012. Long-term evolutionary and ecological responses of calcifying phytoplankton to changes in atmospheric CO_2. *Global Change Biology*, **18**, 3504–3516.

—— HAAGA, K. A., REITAN, T., DIEGO, D. and LIOW, L. H. 2017. Common species link global ecosystems to climate change: dynamical evidence in the planktonic fossil record. *Proceedings of the Royal Society B*, **284**, 20170722.

KWIATKOWSKI, D., PHILLIPS, P. C. B. and SCHMIDT, P. 1992. Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, **54**, 159–178.

LIOW, L. H. 2013. Simultaneous estimation of occupancy and detection probabilities: an illustration using Cincinnatian brachiopods. *Paleobiology*, **39**, 193–213.

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.

MANNION, P. D., UPCHURCH, P., CARRANO, M. T. and 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. and UHEN, M. D. 2010. Climate, critters, and cetaceans: Cenozoic drivers of the evolution of modern whales. *Science*, **327**, 993–996.

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

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

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

SAKAMOTO, M., VENDITTI, C. and BENTON, M. J. 2017. 'Residual diversity estimates' do not correct for sampling bias in palaeodiversity data. *Methods in Ecology & Evolution*, **8**, 453–459.

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

—— and 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**, 765–774.

STARRFELT, J. and LIOW, L. H. 2016. How many dinosaur species were there? Fossil bias and true richness estimated using a Poisson sampling model. *Philosophical Transactions of the Royal Society of London B*, **371**, 20150219.

WALKER, F. M., DUNHILL, A. M., WOODS, M. A., NEWELL, A. J. and BENTON, M. J. 2017. Assessing sampling of the fossil record in a geographically and stratigraphically constrained dataset: the Chalk Group of Hampshire, southern UK. *Journal of the Geological Society*, **174**, 509–521.

WANG, S. C. and DODSON, P. 2006. Estimating the diversity of dinosaurs. *Proceedings of the National Academy of Sciences*, **103**, 13601–13605.