Palaeontology
The Palaeontological Association
www.palass.org
[Palaeontology, Vol. 67, Part 5, 2024, e12729]

Standardizing fossil disparity metrics using sample coverage

by MENNA JONES[1,*] *and* ROGER CLOSE[2]

[1]University of Chicago Division of the Physical Sciences, 5734 S Ellis Avenue, Chicago, Illinois 60637-1403, USA; mjones07@uchicago.edu
[2]Department of Earth Sciences, University of Oxford, South Parks Road, Oxford OX1 3AN, UK
*Corresponding author

Typescript received 30 May 2024; accepted in revised form 17 September 2024

Abstract: Estimating past biodiversity using the fossil record is a central goal of palaeobiology. Because raw estimates of biodiversity are biased by variation in sampling intensity across time, space, environments and taxonomic groups, sampling standardization is routinely applied when estimating taxonomic diversity (e.g. species richness). However, sampling standardization is less commonly used when estimating alternative currencies of biological diversity, such as morphological disparity. Here, we show the effects of standardizing fossil time series of morphological disparity to equal sample completeness, or 'coverage', of the underlying taxon-frequency distribution. We apply coverage-based standardization to three published datasets of discrete morphological characters (echinoderms, ichthyosaurs and ornithischian dinosaurs), and quantify disparity using two metrics: weighted mean pairwise dissimilarity (WMPD) and the sum of variance (SOV). We also compare the effects of coverage-based and sample-size-based standardization. Our results show that coverage standardization can yield estimates of disparity through time that dramatically deviate from raw estimates, both in magnitude and direction of changes. These findings demonstrate that future studies of morphological disparity should control for variation in sampling intensity to enable more reliable inferences.

Key words: palaeobiology, disparity, standardization, sampling, biodiversity.

THE incompleteness of the fossil record means that it cannot be read literally. Because of this, a substantial body of research has addressed the impacts of sampling and preservational biases on our understanding of biodiversity change across time and space (e.g. Raup 1972; Koch 1978; Alroy *et al.* 2001; Smith 2001; Crampton *et al.* 2003; Jablonski *et al.* 2003; Smith & McGowan 2007; McGowan & Smith 2008; Close *et al.* 2020; Benson *et al.* 2021; Antell *et al.* 2024). When estimating taxonomic diversity, subsampling approaches are commonly applied to correct for variable sampling intensity. These sampling-standardized diversity estimates enable fairer comparisons of relative magnitudes of richness across time, space and environments than those from raw (i.e. 'uncorrected' or 'raw') taxon counts. Early studies used size-based standardization (commonly called classical rarefaction; Sanders 1968), which generates samples of a uniform size by drawing a fixed quota of items (individuals or taxon occurrences) from each sampling unit. However, classical rarefaction (hereafter 'CR') has been shown to be fundamentally flawed, because it undersamples richer assemblages, thereby compressing relative richness ratios and flattening diversity curves (Alroy 2010a, 2010b; Close *et al.* 2018). This flaw is rectified by standardizing diversity samples to equal sample completeness, or coverage of the underlying taxon-abundance frequency distribution of the species pool. Subsampling to a target coverage of 0.5, for example, yields the average number of taxa in a random sample drawn from 50% of individuals in the underlying species pool (Chao & Jost 2012). Coverage-based standardization is known to palaeobiologists as shareholder quorum subsampling (SQS; Alroy 2010a; Close *et al.* 2018) and to ecologists as coverage-based rarefaction (CBR; Chao & Jost 2012). The target coverage level used to standardize samples (analogous to the sample-size quota of CR) is called the quorum (SQS) or target coverage (CBR). Coverage-based sampling standardization of taxonomic diversity can either be implemented algorithmically (Alroy 2010a, 2010b, 2010c, 2014) or analytically, using mathematical equations (Chao & Jost 2012).

The application of coverage-based sampling standardization procedures have predominantly been restricted to taxonomic diversity (but see Chao *et al.* 2021, 2023 for recent exceptions). However, other biodiversity metrics may be valuable when studying evolutionary dynamics. One suite of metrics that has been very widely applied in palaeobiology is morphological disparity (Foote 1991, 1997; Wills *et al.* 1994; Roy & Foote 1997; Hopkins & Gerber 2017; Guillerme *et al.* 2020), which can be either estimated from discrete morphological characters or continuous data, such as landmarks in geometric

doi: 10.1111/pala.12729

TABLE 1. Description of published datasets used in this study; purpose of original study, along with taxonomic level, binning scheme, temporal resolution, and timespan of dataset are noted.

Dataset			PaleoDB occurrence data		
Taxa	Number of characters	Original study	Taxonomic level	Temporal resolution of binning scheme	Span
Echinoderms (n = 366)	413	Authors constructed a time-scaled 'supertree' consisting of 366 tips and 365 internal nodes which they used to plot temporal trends in echinoderm ecology and morphology throughout the early Palaeozoic (Novack-Gottshall *et al.* 2022b)	Genus	Epoch	Cambrian Series 2 – Late Ordovician
Ichthyosaurs (n = 108)	287	Authors identified a new ichthyosaur species (*Cymbospondylus deuler*). They used a morphological character matrix of 287 unordered characters for 108 ingroup taxa to reconstruct phylogenetic relationships and place the space within the *Cymbospondylus* clade (Klein *et al.* 2020)	Species	Epoch	Early Triassic – Early Cretaceous
Ornithischia (n = 194)	88	Authors examined the ecological diversity of herbivorous dinosaurs during the Cretaceous, using a character matrix representing 194 species (Nordén *et al.* 2018)	Species	Stage	Valanginian –Maastrichtian

morphometrics. A number of previous studies have applied sample-size-based rarefaction to the study of morphological disparity in the fossil record (e.g. Foote 1992; Ciampaglio *et al.* 2001; McClain *et al.* 2004; Prentice *et al.* 2011; Butler *et al.* 2012; Deline & Ausich 2016). However, size-based rarefaction of disparity must suffer from the same problems as size-based rarefaction of taxonomic richness; namely, that true diversity ratios will be compressed. To rectify this, we present an extension to the SQS algorithm (hereafter termed the '-coverage-standardized disparity' algorithm) to standardize estimates of morphological disparity to equal coverage. By applying this new technique to three published datasets, we demonstrate the potential future role of this approach for standardizing other biodiversity metrics through time.

METHOD

All computational work was conducted in R v4.3.1 (R Core Team 2023) using RStudio v4.3.1 (RStudio Team 2023). Full analysis scripts and data are available in Jones & Close (2024).

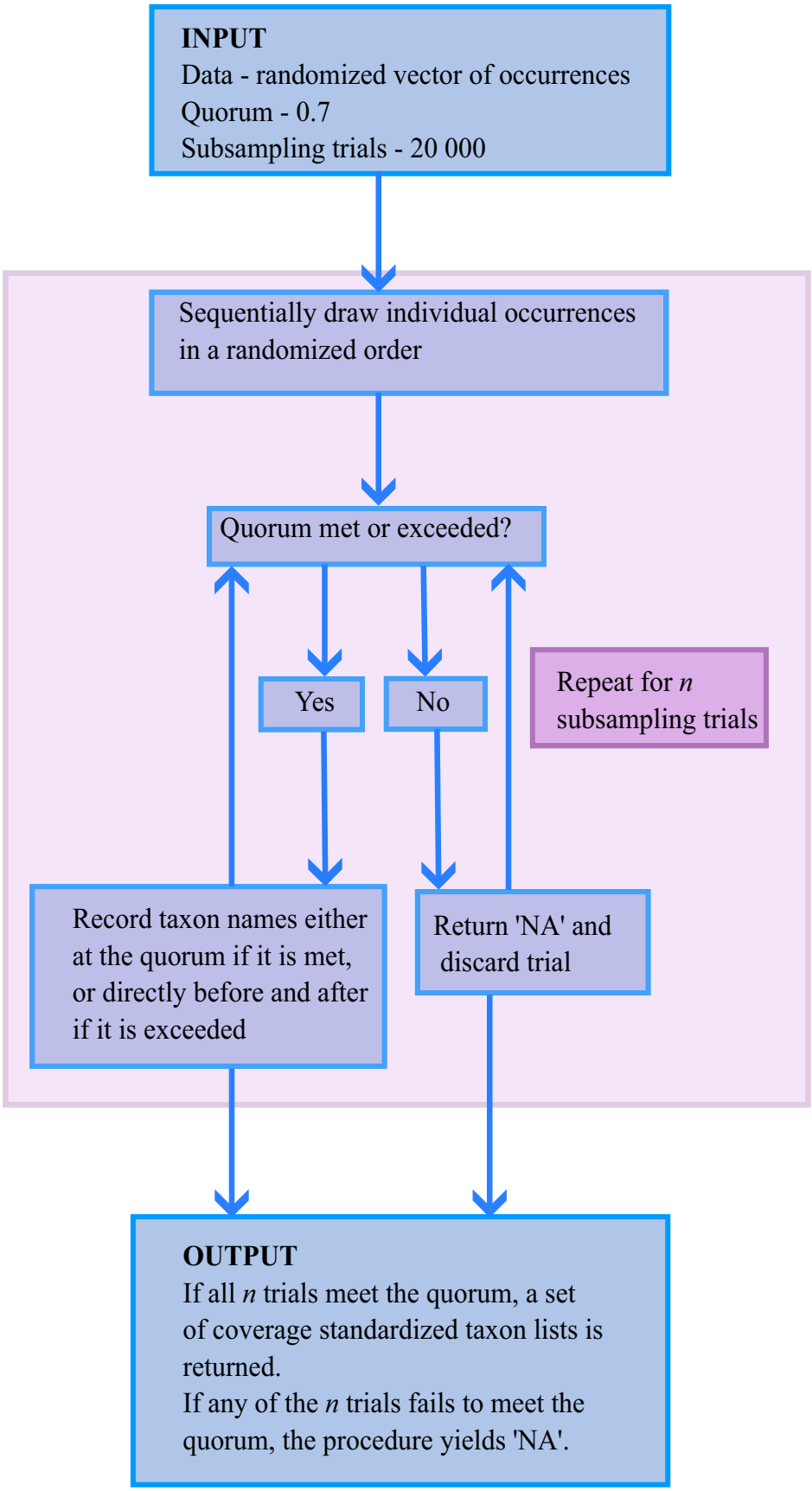
Discrete morphological character datasets

Three existing discrete morphological character datasets were obtained from the literature for this study: Cambro-Ordovician echinoderms (Novack-Gottshall *et al.* 2022a), Early Triassic and Early Cretaceous ichthyosaurs (Klein *et al.* 2020, suppl. material), and Valanginian–Maastrichtian ornithischians (Nordén *et al.* 2018). These were deemed to have both high numbers of taxa and a sufficient time span to allow clear time series to be constructed; relevant parameters are summarized in Table 1.

Fossil occurrence data download and binning

Fossil occurrence records and time interval data were simultaneously downloaded from the Paleobiology Database (PaleoDB; https://www.paleobiodb.org) to ensure correct alignment of stratigraphic bins and ages of occurrences (see Table 1 for the binning schemes used for each taxon set). The 'majority' binning rule was then used to assign each taxon to the bin in which over 50% of its range falls (see Close *et al.* 2020) due to calculating the proportion of its temporal range that lies within a specific interval. To

FIG. 1. Pipeline for the coverage-standardized disparity (CSD) developed in this study. The required inputs are occurrence data, a target quorum, and number of subsampling trials. The output is a set of coverage-standardized taxon lists representing the quorum-crossing points encountered throughout the subsampling trials. For a detailed description of the algorithm, see the main text.



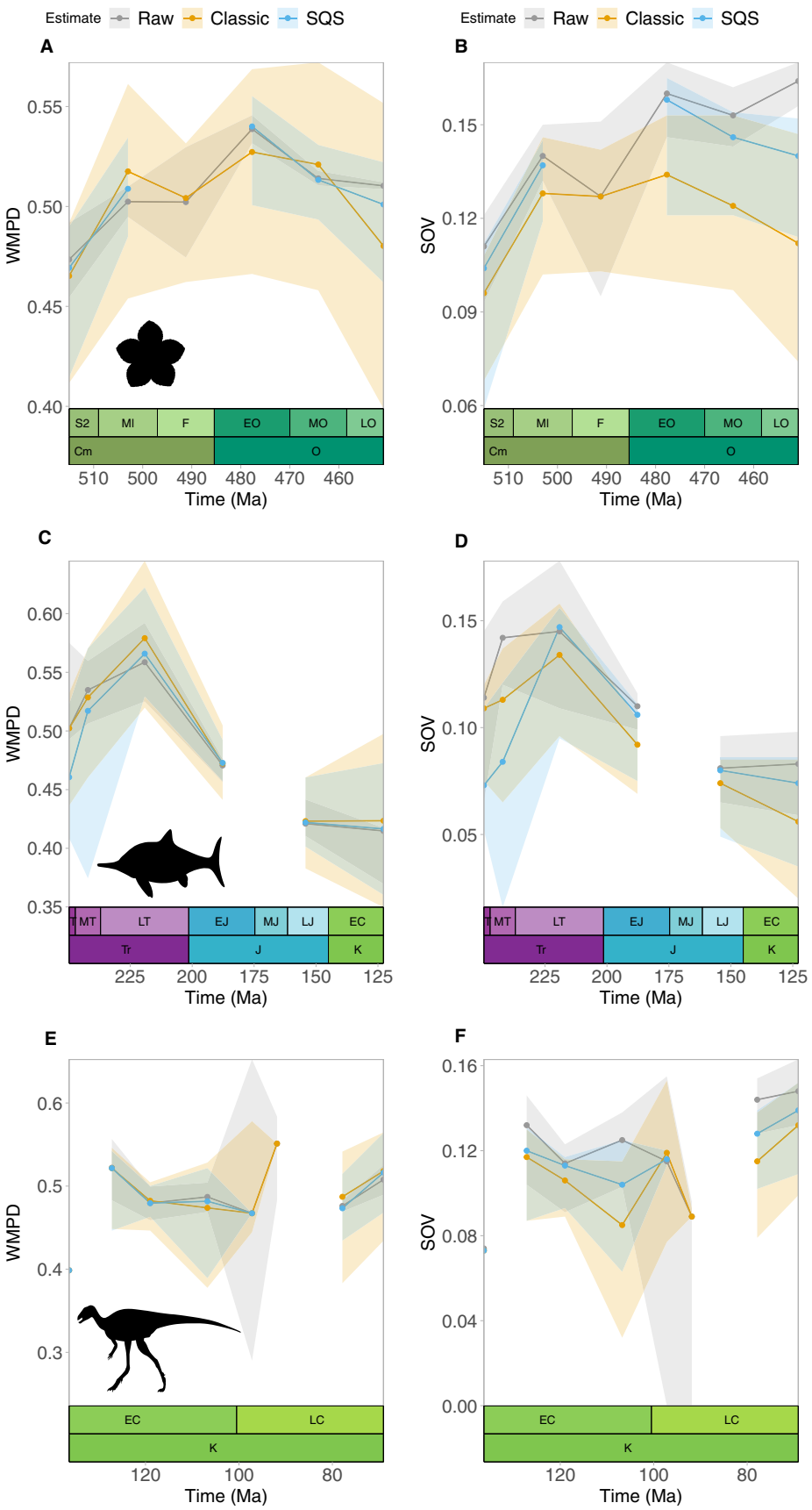


FIG. 2. Raw, size- and coverage-standardized disparity time series for echinoderms, ichthyosaurs, and ornithischian dinosaurs. Curves are plotted for WMPD (A, C, E) and SOV (B, D, F) metrics with 95% confidence intervals shaded. Coverage-standardized curves were generated using our coverage-standardized disparity algorithm for a quorum value of 0.7. Silhouette images are sourced from www.phylopic.org, courtesy of Michael Keesey, Jagged Fang Designs, and Scott Hartman (CC0 1.0).

exclude very poorly-sampled intervals, bins containing fewer than 10 occurrences were dropped from our analysis. Data-sparse bins may have spuriously high sample coverage and skew the nature of observed time series trends. For full details of the PBDB data download, see the Appendix S1.

The coverage-standardized disparity algorithm

In order to standardize morphological disparity estimates to equal coverage, we modified Alroy's 'EXACT' SQS algorithm (Alroy 2014). This new 'coverage-standardized disparity' algorithm takes the occurrence (or abundance) data, a target quorum (e.g. 0.7), and the number of subsampling trials as inputs, and outputs a set of coverage-standardized taxon lists (see Fig. 1 for a visual pipeline of the algorithm). In its original implementation of Alroy (2014), the EXACT algorithm performs subsampling trials in which all of the available occurrences or individuals are sequentially drawn, one at a time, in a random order. After each occurrence or individual is drawn, the coverage of the sample obtained up to that point is computed. The sampled richness is recorded either when the target quorum is met (if met exactly) or just before and after it is crossed (if it lies between the drawing of two sequential occurrences). The median richness for all of these crossing points is computed for each trial, before the overall median richness for all independently-randomized subsampling trials is computed to give the coverage-standardized diversity estimate.

To compute coverage-standardized disparity metrics, our modified algorithm retains all lists of unique taxon names drawn at the quorum-crossing points that are encountered during subsampling trials (either at, if the quorum is met exactly, or immediately before and after a crossing point, if it lies between drawing two occurrences). If a subsampling trial fails to reach the target quorum, the algorithm returns an 'NA' value to indicate missing data. The final output after repeating these steps for n subsampling trials is therefore a large set of coverage-standardized taxon lists, representing the crossing points encountered throughout all subsampling trials. Although it would be possible to directly compute disparity estimates at each crossing point, either in addition to or in place of taxonomic richness, our approach allows an arbitrary number of diversity metrics to be retrospectively computed from the same set of coverage-standardized taxon lists, which allows greater downstream flexibility.

Disparity calculations

Disparity was calculated on a per-bin basis to construct time series. Raw disparity estimates were calculated directly from the full set of temporally-binned taxa, and coverage-standardized values were calculated using the list of coverage-standardized taxon sets returned from the extended algorithm. Matrices of discrete morphological characters for the datasets obtained from the literature (Table 1) were converted into distance matrices using the *calculate_morphological_distances()* function of the Claddis package (v0.6.3; Lloyd 2016), using the default MORD distance metric. This metric divides the pairwise distances between taxa by the maximum realizable distance between them, thus restricting the resulting distances to between 0 and 1. This has the appeal of highlighting those that lie close to the maximum value (Lloyd 2016).

We quantified disparity using two metrics: weighted mean pairwise dissimilarity (WMPD; Close *et al.* 2015; also known in the literature as weighted mean pairwise distance) and the sum of variance (SOV; Van Valen 1974), two of the most commonly applied disparity metrics in the literature. As a measure of central tendency, WMPD is theoretically unbiased by sample size, whereas SOV is an additive metric that can scale with sample size. WMPD values were calculated as the sum of the product of the pairwise differences and the pairwise comparable characters divided by the pairwise comparable differences (Close *et al.* 2015). WMPD operates directly on distance matrices, rather than on ordinations of these matrices. This metric weights dissimilarities for pairs of taxa such that those based on more comparable characters contribute more to calculated disparity, eliminating the requirement to delete taxa with incomplete characters (Close *et al.* 2015). WMPD operates directly on distance matrices, rather than on ordinations of these matrices. This metric weights dissimilarities for pairs of taxa such that those based on a greater number of comparable characters contribute more to calculated disparity, eliminating the requirement to delete taxa with incomplete characters (Close *et al.* 2015). For raw estimates, 1000 bootstrapping trials were performed to produce weighted mean, 2.5%,

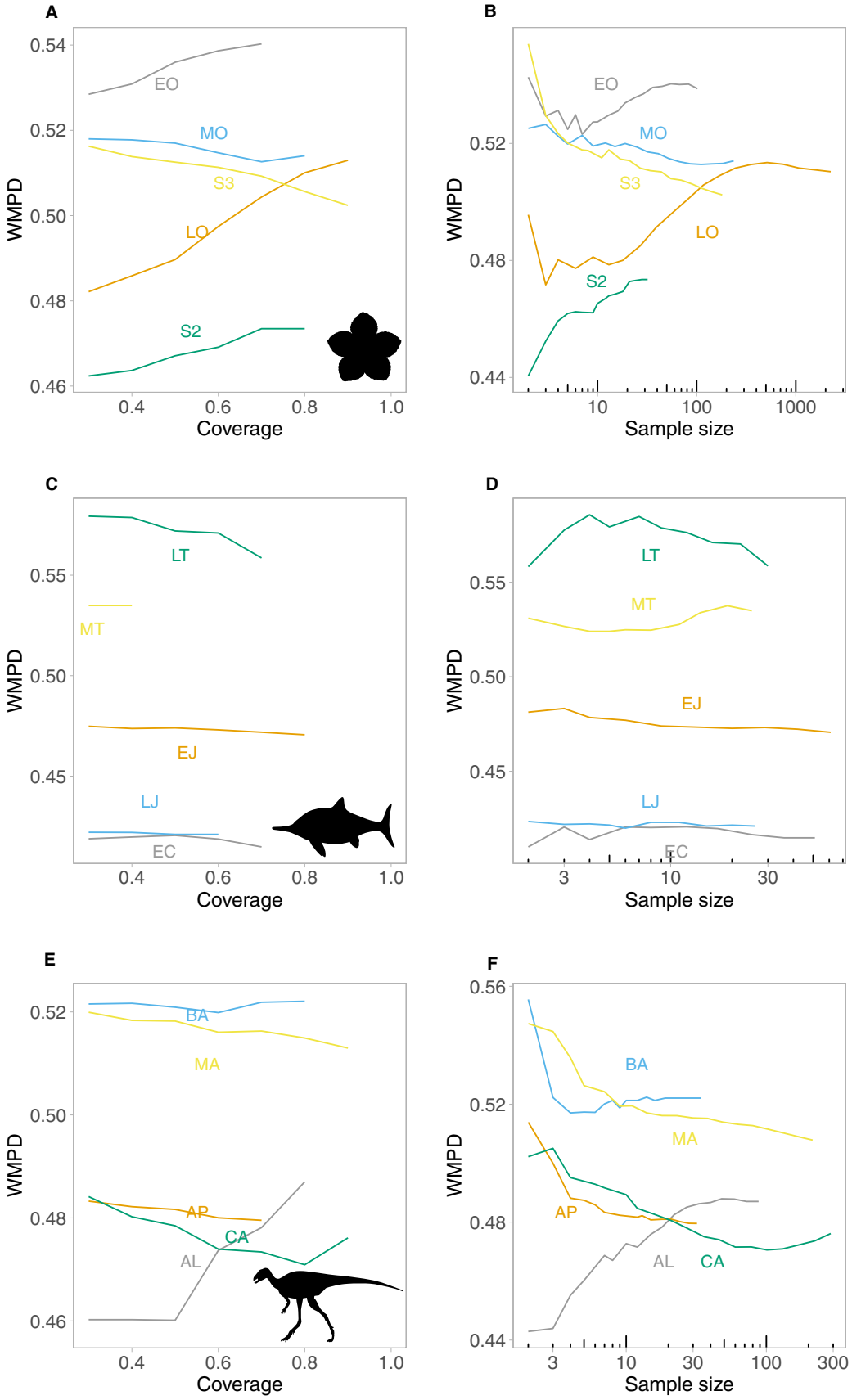


FIG. 3. Coverage- (A, C, E) and size-based (B, D, F) rarefaction curves for echinoderms, ichthyosaurs and ornithischian dinosaurs generated for the WMPD metric. The latter were generated using a vector of logarithmically-spaced sample sizes. *Abbreviations*: AL, Albian; AP, Aptian; BA, Barremian; CA, Campanian; EC, Early Cretaceous; EJ, Early Jurassic; EO, Early Ordovician; LJ, Late Jurassic; LO, Late Ordovician; LT, Late Triassic; MA, Maastrichtian; MO, Middle Ordovician; MT, Middle Triassic; S2, Series 2; S3, Series 3. Silhouette images are sourced from www.phylopic.org, courtesy of Michael Keesey, Jagged Fang Designs, and Scott Hartman (CC0 1.0).

and 97.5% values, which were taken as the average, lower and upper bounds, respectively. For standardized data, one WMPD value was calculated per subsampling trial, translating into 1000 values per bin. Subsequent values for the 2.5%, median and 97.5% values were obtained.

Sum of variance (SOV) is widely used in disparity studies (e.g. Halliday & Goswami 2016; Romano *et al*. 2018; Smith & Donoghue 2022) to measure the extent of morphospace occupation as the sum of the variance of each dimension. SOV values were calculated using the *DispRity()* function of the DispRity package (Guillerme 2018) which takes a distance matrix, calculates SOV, and returns the bootstrapped median values, 2.5% and 97.5% values over a default 100 trials. Note that the distance matrix was ordinated using the classical multidimensional scaling procedure via the *cmdscale()* function in the 'stats' package prior to these calculations. For the raw estimates, these values were retained for plotting. The treatment of the coverage-standardized estimates differs in that all of the individual subsampling trials for each bin were sorted, and the 2.5%, median and 97.5% values were extracted for each bin. These values constituted the mean, lower and upper bounds of 95% confidence intervals for each bin, respectively. Disparity estimates were normalized to allow direct comparisons of raw and standardized time series trends, because subsampled disparity estimates will always be lower than raw values.

Time series of sample coverage and spatial sampling

We also constructed time series of sample coverage and extent of spatial sampling to examine how these variables might influence estimates of disparity through time. For each time bin, sample coverage was calculated from the randomly sampled occurrence vector and quantified using the Good's *u* metric (Good 1953). We chose to use the more sophisticated formula of Chao & Shen (2010) that incorporates information from singletons and doubletons and does not require large sample sizes to produce accurate estimates. Calculations were performed over 1000 bootstrapping trials, and median values were retained for plotting: these represent the average sample coverage of each time bin. Spatial coverage was quantified using counts of equal-area hexagonal/pentagonal grid cells with 1000 km spacings between midpoints (using the R package dggridR; Barnes & Sahr 2017) that contained fossil data, also calculated on a per-bin basis.

Rarefaction curves

Both size- and coverage-based rarefaction curves were generated for the three datasets used in this study (see Foote 1999; Kotric & Knoll 2015). Coverage-based rarefaction analyses were performed for a set of quorum values from 0.4 to 0.9, spaced at intervals of 0.1, on a per-bin basis for the three datasets. Quorum values below 0.4 are likely to sample too few taxa and render disparity calculations unfeasible (this is a key difference between sampling-standardized estimates of disparity and taxonomic diversity, where one lone taxon can still be counted). Analyses were performed over 20 000 subsampling trials and any trial returning an occurrence vector containing <5 unique taxa was dropped to avoid the generation of a sparse or empty distance matrix. After subsetting the distance matrices, WMPD values for standardized data were calculated and the 2.5%, mean and 97.5% values from the trials were extracted for each quorum value. The same procedure was conducted for the SOV calculations. For the size-based rarefaction analyses, a vector of logarithmically-spaced sample sizes was generated for each bin within the three datasets. This vector size was tailored to the amount of occurrence data that was available for each dataset (i.e. more abundant occurrence data correlated with a larger sample size vector). WMPD and SOV values were then calculated and compared to those generated from coverage-based methods.

Example questions

Finally, we generated an example research question for each dataset to demonstrate how the choice to apply CR, SQS or neither, may influence a study's findings. These are as follows:

1. When did Cambro-Ordovician echinoderm disparity peak?

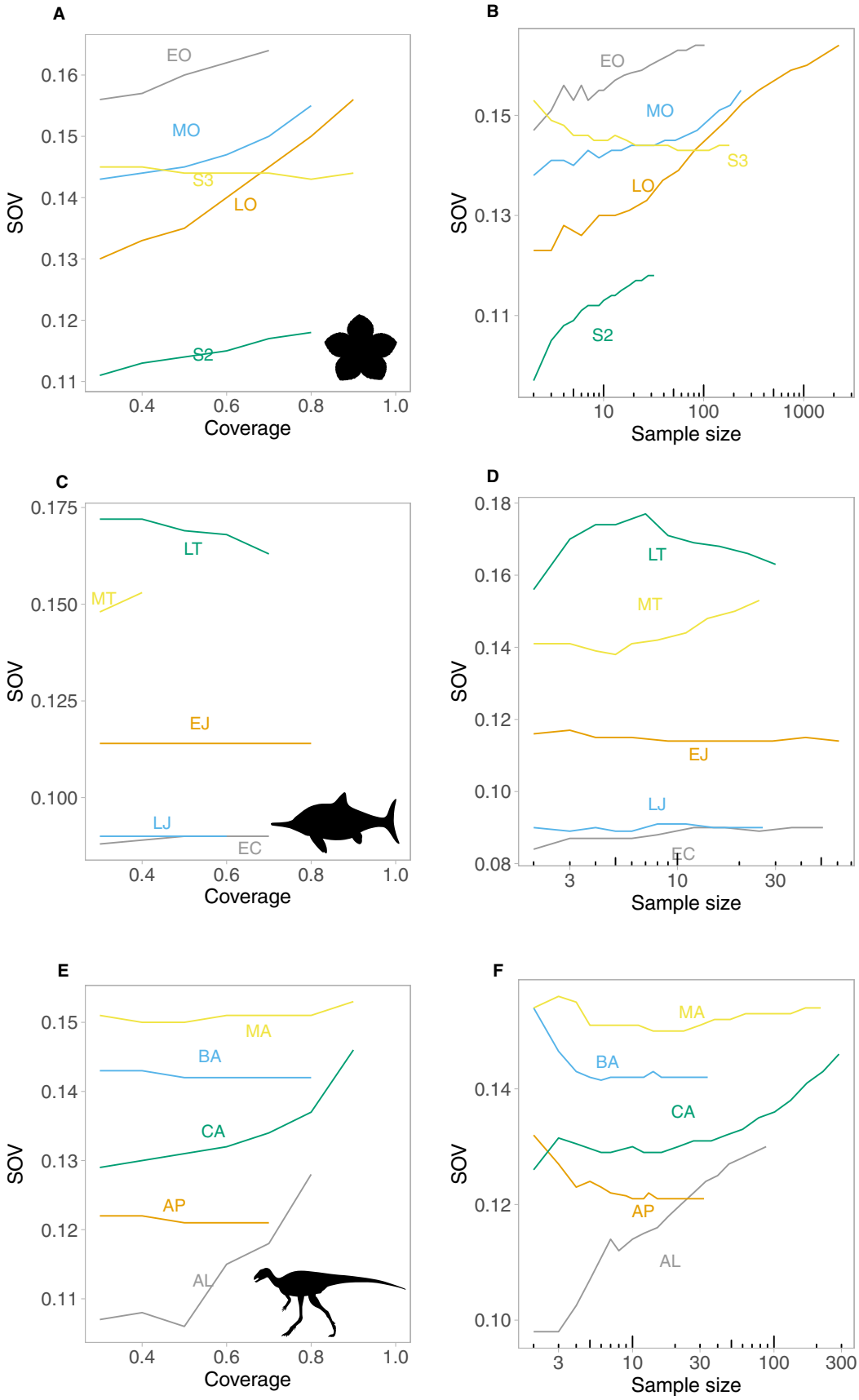


FIG. 4. Coverage- (A, C, E) and size-based (B, D, F) rarefaction curves for echinoderms, ichthyosaurs and ornithischian dinosaurs generated for the SOV metric. The latter were generated using a vector of logarithmically spaced sample sizes. Abbreviations as for WMPD plots (Fig. 3). Silhouette images are sourced from www.phylopic.org, courtesy of Michael Keesey, Jagged Fang Designs, and Scott Hartman (CC0 1.0).

2. Does ichthyosaur disparity stabilize after its decline across the Triassic–Jurassic boundary?
3. How dramatically did ornithischian disparity fluctuate throughout the Cretaceous?

The purpose of testing these questions is to demonstrate why researchers should take care in choosing their sampling method and disparity metric, both of which can influence the biological interpretations of results. How these influences manifest are explored during the results and later discussion.

RESULTS

Our results reveal that applying SQS to morphological disparity can alter both the magnitude and trend of estimates through time (Fig. 2); however, this is largely dependent on disparity metric choice. The SOV metric exhibits large deviations between raw and SQS disparity (echinoderms during Ordovician; Fig. 2B), yet the WMPD metric sometimes generates raw and standardized time series that are indistinguishable, as is seen for Triassic ichthyosaur data (Fig. 2C). Differences are also recovered between SQS and CR time series and are non-uniform between datasets. Whilst SQS consistently estimates higher SOV values than CR for the echinoderm dataset (Fig. 2B), the curves alternate in which recovers a higher WMPD estimate for Triassic ichthyosaur disparity (Fig. 2C). It appears that the only consistent difference between CR and SQS disparity estimates is the narrower of 95% confidence bands of the latter (e.g. see echinoderm disparity; Fig. 2A, B), which is also observed in the normalized time series (Fig. S1).

Traditionally, rarefaction curves show a predictable trend of increasing taxonomic diversity with sample size and coverage level. The same does not hold true for the rarefaction curves for disparity in this study, particularly those using WMPD, hence it is useful to consider the two metrics separately. For WMPD, some curves show no evidence of reaching an asymptote with increasing sample size and coverage, but the gradient of these curves differs between bins and datasets (Fig. 3). Several time bins produce static rarefaction curves that fail to respond to increasing sampling and coverage (e.g. ichthyosaurs in the Early Jurassic; Fig. 3C, D) whilst others estimate progressively lower disparity as rarefaction proceeds (e.g. ornithischians during the Maastrichtian; Fig. 3F). Furthermore, the size-based rarefaction curves for several bins (e.g. echinoderms in the Late Ordovician; Fig. 3B) follow non-monotonic trajectories, first decreasing, then increasing, before declining again. In contrast, SOV rarefaction curves for disparity are often seen to increase with sample size and coverage level (Fig. 4). This is especially apparent for the echinoderm dataset for which the most data-rich bin even exhibits a somewhat linear scaling relationship with increasing coverage (Late Ordovician; Fig. 4A). Despite the described lack of predictable scaling relationships for both WMPD and SOV, there is a common theme of narrowing confidence bands as rarefaction proceeds, which is exhibited by all bins across the three datasets, regardless of metric choice (Figs S2–S7). It appears the only consistent difference between CR and SQS disparity estimates is the narrower of 95% confidence bands of the latter (e.g. see echinoderm disparity, Fig. 2A, B), which is also observed in the normalized time series (Fig. S1).

Effect of sampling standardization on example questions

Finally, we return to our example questions to examine how choice of standardization approach influences interpretation of disparity-through-time trends.

1. *When did Cambro-Ordovician echinoderm disparity peak?*

WMPD values calculated from raw data support an Early Ordovician disparity peak, whilst raw SOV values support a Late Ordovician peak (Fig. 2A, B). CR and SQS time series both show that WMPD and SOV values peaked in the Early Ordovician, though the SQS time series illustrates a relatively more pronounced peak (Fig. 2A, B).

2. *Does ichthyosaur disparity stabilize after its decline across the Triassic–Jurassic boundary?*

The time series show remarkably similar trajectories for ichthyosaur disparity throughout the measured time intervals with almost fully overlapping confidence intervals. The main difference is present for the SOV metric: whilst raw data support a slight increase in disparity from the Jurassic into Cretaceous (Fig. 2D), both CR and SQS yield curves that continue to decline during this period. This decline is especially notable when CR is applied.

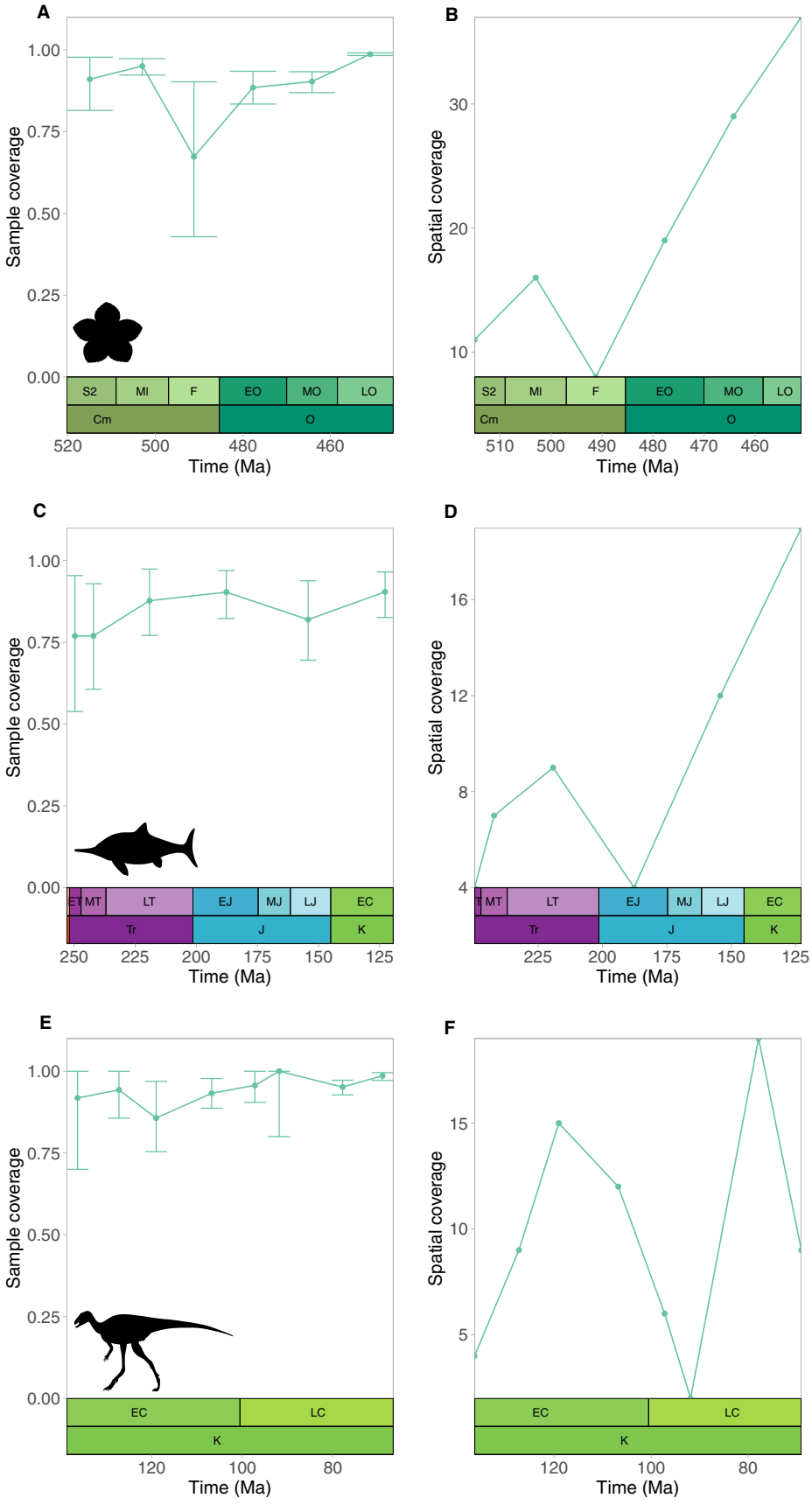


FIG. 5. Time series of sample coverage (Good's *u*) and spatial sampling (counts of occupied equal-area hexagonal/pentagonal grid cells with 1000 km spacings) for each of the three datasets in this study: echinoderms (A, B), ichthyosaurs (C, D), ornithischian dinosaurs (E, F). Silhouette images are sourced from www.phylopic.org, courtesy of Michael Keesey, Jagged Fang Designs, and Scott Hartman (CC0 1.0).

3. *How dramatically did ornithischian disparity fluctuate throughout the Cretaceous?*

Standardizing ornithischian data increases the volatility of disparity time series throughout the Cretaceous (Fig. 2E, F). This holds especially true for the CR SOV curves; however, the large confidence intervals associated with the raw time series cast doubt on this interpretation.

DISCUSSION

Estimates of SOV disparity and their associated evolutionary interpretation are strongly impacted by whether disparity is sampling standardized, and by choice of sampling standardization approach (CR vs SQS). For example, the apparent peak in Cambro-Ordovician echinoderm disparity occurs in either the Early or Late Ordovician depending on whether standardization procedures are applied. By contrast, the WMPD metric is more robust to sampling intensity, and so SQS rarefaction curves differ little from those using CR. The main distinction between the CR and SQS-generated WMPD curves is the smaller 95% confidence intervals when shifting from former to the latter approach, especially for the echinoderm dataset (Fig. 2A). This results from differences in how the two algorithms use each subsampling trial in disparity calculations. Whilst this finding is demonstrated for only the three datasets used in this study, it suggests that the metric is relatively robust to completeness of the taxonomic sample, especially compared to SOV. This is not necessarily an advantage (scaling with sample completeness might intuitively be expected to occur for a disparity metric) but it is nonetheless an important distinction between the two metrics.

Although our SQS approach is presented as an additional sampling-standardization method, the reliability of disparity estimates returned is influenced in part by the maximum achievable sampling coverage (a function of sampling intensity) at any point in time, and by the scope of the accessible sampling universe (distinct from sampling intensity; see Close *et al.* 2018, 2020). For example, the increase in raw SOV echinoderm disparity into the Late Ordovician (Fig. 2B) is driven by substantially higher sampling in that interval, as quantified by counts of occurrences, sample coverage and spatial coverage (Fig. 5A, B). By contrast, some bins fail to achieve sufficient sample coverage, so SQS disparity estimates for these bins are not available even at the lowest quorum used in this analysis (0.4). The Middle Jurassic ichthyosaurs are one example of this (Fig. 6C, D) and as a result, the reliability of corresponding raw disparity estimates is especially questionable.

SQS was originally developed for taxonomic diversity applications, so it is perhaps unsurprising that our application to morphological disparity reveals differences in how the two biodiversity metrics respond. For the datasets analysed in this study, SOV often shows no evidence of reaching an asymptote as more data is added to a size-based rarefaction curve (Fig. 4). This contrasts the general scenario in which increasing sample size consistently yields higher SOV values. We interpret this as reflecting a fundamental difference between taxonomic and morphological diversity: estimates of disparity do not always scale in a predictable and linear way with increased sampling and depend heavily on the choice of metric. Adding a single morphologically divergent species can increase disparity substantially, while adding many very similar species would not. Moreover, the presence of non-monotonic trajectories in the bin-by-bin size-based rarefaction curves illustrates how increasing the size of a random sample does not necessarily capture a more morphologically diverse set of taxa, which is another contrast between the behaviour of standardized taxonomic vs morphological diversity. Similar curve trajectories have been recovered in previous disparity studies (see fig. 1 in Ciampaglio *et al.* 2001 for a particularly clear SOV example), so we are confident that our results which reflect intrinsic characteristics of the two disparity metrics.

Previous work on standardizing disparity has utilized CR through subsampling datasets to a common number of either discrete morphological characters (e.g. Ciampaglio *et al.* 2001; Deline & Ausich 2016) or taxa (e.g. Ciampaglio *et al.* 2001; McClain *et al.* 2004; Prentice *et al.* 2011; Butler *et al.* 2012). Such studies have focused on the differing vulnerabilities of disparity metrics to sample size and effort: range-based metrics (e.g. sum of ranges; SOR) are strongly influenced by sampling effort and morphological outliers (Foote 1992), whilst variance-based metrics (e.g. SOV) can be heavily sensitive to the amount of data sampled, or 'sampling effort' (Bault *et al.* 2024). These are merely generalizations, however, with contradictions being rife in the literature. For example, Ciampaglio *et al.* (2001) used Ordovician

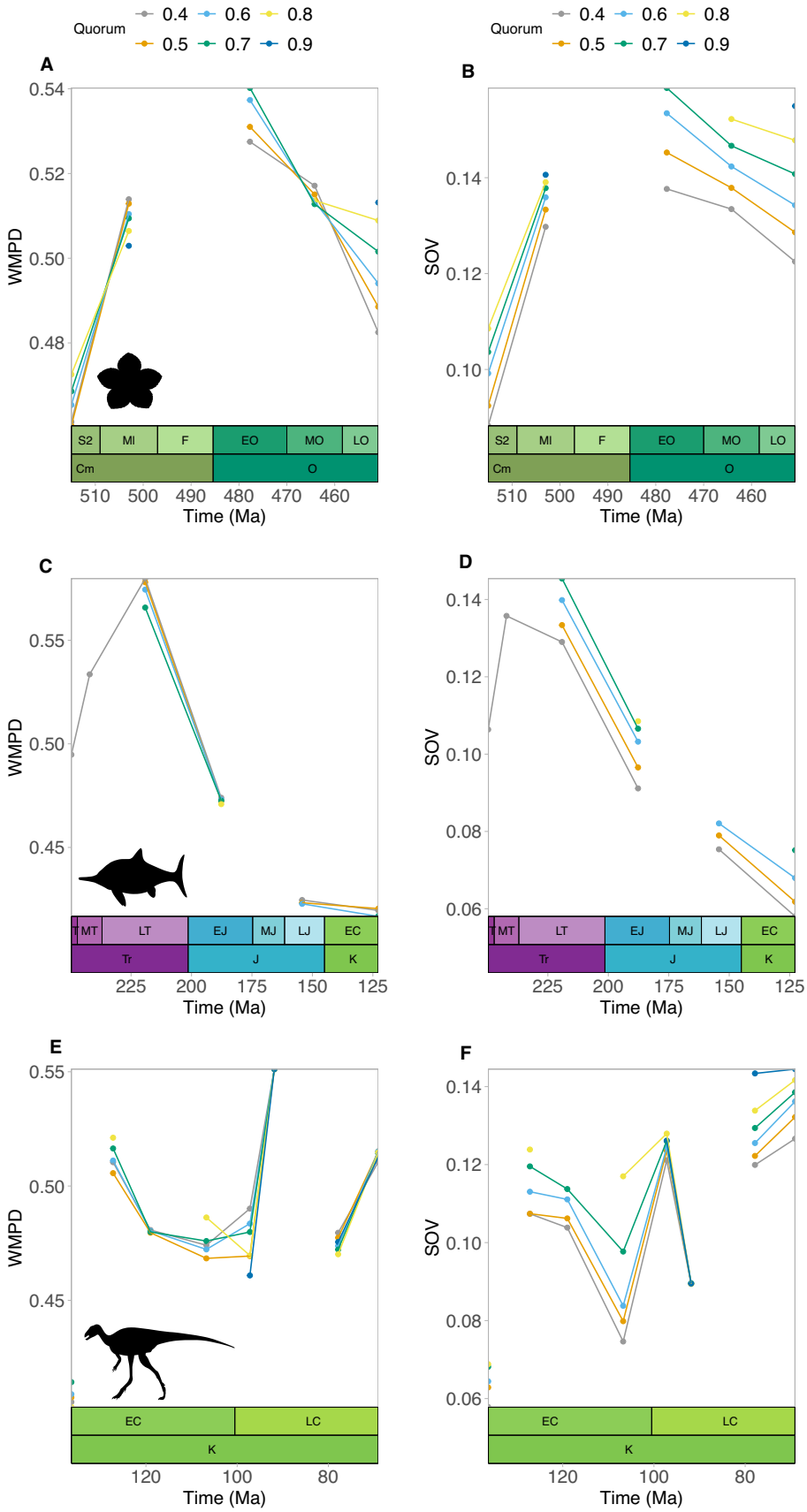


FIG 6. A comparison of SQS disparity time series for differing quorum values ranging from 0.4 to 0.9. See main text for a justification of the chosen values. Silhouette images are sourced from www.phylopic.org, courtesy of Michael Keesey, Jagged Fang Designs, and Scott Hartman (CC0 1.0).

crinoid data (Foote 1999) to demonstrate that SOV can recover the 'true' disparity estimate for a taxon pool from a mere 27% of the dataset, contradicting the accepted notion of its sample size sensitivity. The well-known shortcomings of CR are not restricted to its applications to taxonomic diversity, and we find that they are mirrored by metrics of morphological disparity. Relative undersampling of richer assemblages compresses disparity curves in the same manner observed for taxonomic richness curves, a phenomenon demonstrated for all datasets used in this study (Fig. 2). Such flattening of curves is markedly more pronounced for the SOV metric which probably reflects its sensitivity to sample size differences when estimating sample variances. These findings cast doubt on the ability of CR to accurately recover true biological signals present in discrete morphological character datasets and support the use of the coverage-based standardization.

Future studies of morphological disparity in the fossil record should refrain from only presenting raw or size-standardized morphological disparity curves and should instead include coverage-based standardization alongside them. Furthermore, how close each standardization approach comes to yielding the 'true' disparity curve has not been demonstrated; we recommend a simulation-based study as a potential way to explore this. Our study shows that controlling for variation in sampling intensity is important when comparing alternate measures of biodiversity such as morphological disparity. Our coverage-standardization algorithm could also be applied to other biodiversity metrics, such as phylogenetic lineage diversity. Nevertheless, variation in the scope of the accessible sampling universe (Fig. 5B, D, F) may have an even more profound effect on biodiversity estimates than variation in sample completeness (Alroy 2010c, 2014; Close *et al.* 2018, 2020), including for alternative metrics such as disparity. Ideally, future studies of morphological disparity in the fossil record would also attempt to control for variation in the scope of the sampling universe. However, greater limitations on data availability, vs that available for estimating taxonomic diversity, may make this substantially more challenging.

CONCLUSION

In this study, we applied our coverage-standardized disparity algorithm to existing morphological character datasets of echinoderms, ichthyosaurs and ornithischian dinosaurs. CR and SQS-generated time series differed from each other and the raw data equivalents. These differences were more apparent for SOV relative to WMPD which supports using the latter as a disparity metric. We emphasize the need to integrate coverage-based standardization techniques into studies of morphological disparity to generate more reliable estimates of biodiversity through time.

Acknowledgements. MLJ is funded by The University of Chicago. RAC is funded by a Royal Society University Research Fellowship (URF\R1\211571). We would like to thank T. Smith for interesting discussion regarding disparity metrics. We would also like to thank J. Alroy and M. Foote for their previous work that helped to inspire this study. T. Guillerme and C. Dean commented on an earlier draft of this manuscript. Finally, thank you to Michael Keesey, Jagged Fang Designs, and Scott Hartman for contributing the Phylopic images of echinoderms, ichthyosaurs, and ornithischian dinosaurs, respectively. This is Paleobiology Database official publication number 502.

Author contributions. **Conceptualization** Roger Close (RC), Menna Jones (MJ); **Data Curation** RC, MJ; **Formal Analysis** RC, MJ; **Funding Acquisition** RC; **Investigation** RC, MJ; **Methodology** RC, MJ; **Project Administration** RC; **Resources** RC; **Software** RC, MJ; **Supervision** RC; **Validation** RC, MJ; **Visualization** MJ; **Writing – Original Draft Preparation** RC, MJ; **Writing – Review & Editing** RC, MJ.

DATA ARCHIVING STATEMENT

Full analysis scripts and data for this study are available in the Dryad digital repository: https://doi.org/10.5061/dryad.wpzgmsbxt.

Editor. David Button

SUPPORTING INFORMATION

Additional Supporting Information can be found online (https://doi.org/10.1111/pala.12729):

Figure S1. Normalized raw, CR and SQS disparity time series for echinoderms, ichthyosaurs and ornithischian dinosaurs.
Figure S2. Individual sample-size and coverage-based rarefaction curves for the echinoderm dataset, as measured by SOV.
Figure S3. Individual sample-size and coverage-based rarefaction curves for the echinoderm dataset, as measured by WMPD.

Figure S4. Individual sample-size and coverage-based rarefaction curves for the ichthyosaur dataset, as measured by SOV.
Figure S5. Individual sample-size and coverage-based rarefaction curves for the ichthyosaur dataset, as measured by WMPD.
Figure S6. Individual sample-size and coverage-based rarefaction curves for the ornithischian dinosaur dataset, as measured by SOV.
Figure S7. Individual sample-size and coverage-based rarefaction curves for the ornithischian dinosaur dataset, as measured by WMPD.
Appendix S1. Data downloaded from the Paleobiology Database.

REFERENCES

Alroy, J. 2010a. Geographical, environmental and intrinsic biotic controls on Phanerozoic marine diversification. *Palaeontology*, **53** (6), 1211–1235.

Alroy, J. 2010b. Fair sampling of taxonomic richness and unbiased estimation of origination and extinction rates. *Palaeontology*, **53**, 1211–1235.

Alroy, J. 2010c. Shifting balance of diversity among major marine animal groups. *Science*, **329**, 1191–1194.

Alroy, J. 2014. Accurate and precise estimates of origination and extinction rates. *Paleobiology*, **40** (3), 374–397.

Alroy, J., Marshall, C. R., Bambach, R. K., Bezusko, K., Foote, M., Fürsich, F. T., Hansen, T. A., Holland, S. M., Ivany, L. C., Jablonski, D. and Jacob, D. K. 2001. Effects of sampling standardization on estimates of Phanerozoic marine diversification. *Proceedings of the National Academy of Sciences*, **98** (11), 6261–6266.

Antell, G. T., Benson, R. B. J. and Saupe, E. E. 2024. Spatial standardization of taxon occurrence data—a call to action. *Paleobiology*, **50** (2), 177–193.

Barnes, R. and Sahr, K. 2017. dggridR: discrete global grids for R. R package version 2.0.4. https://github.com/r-barnes/dggridR/

Bault, V., Crônier, C. and Monnet, C. 2024. Coupling of taxonomic diversity and morphological disparity in Devonian trilobites? *Historical Biology*, **36** (3), 473–484.

Benson, R. B., Butler, R., Close, R. A., Saupe, E. and Rabosky, D. L. 2021. Biodiversity across space and time in the fossil record. *Current Biology*, **31** (19), R1225–R1236.

Butler, R. J., Brusatte, S. L., Andres, B. and Benson, R. B. 2012. How do geological sampling biases affect studies of morphological evolution in deep time? A case study of pterosaur (Reptilia: Archosauria) disparity. *Evolution*, **66** (1), 147–162.

Ciampaglio, C. N., Matthieu, K. and McShea, D. W. 2001. Detecting changes in morphospace occupation patterns in the fossil record: characterization and analysis of measures of disparity. *Paleobiology*, **27** (4), 695–715.

Chao, A. and Jost, L. 2012. Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. *Ecology*, **93** (12), 2533–2547.

Chao, A. and Shen, T. J. 2010. SPADE (species prediction and diversity estimation) [software]. http://chao.stat.nthu.edu.tw/softwareCE.html

Chao, A., Henderson, P. A., Chiu, C., Moyes, F., Hu, K., Dornelas, M. and Magurran, A. E. 2021. Measuring temporal change in alpha diversity: a framework integrating taxonomic, phylogenetic and functional diversity and the iNEXT.3D standardization. *Methods in Ecology and Evolution*, **12** (10), 1926–1940.

Chao, A., Thorn, S., Chiu, C., Moyes, F., Hu, K., Chazdon, R. L., Wu, J., Magnago, L. F. S., Dornelas, M., Zelený, D., Colwell, R. K. and Magurran, A. E. 2023. Rarefaction and extrapolation with beta diversity under a framework of Hill numbers: the iNEXT.beta3D standardization. *Ecological Monographs*, **93** (4), e1588.

Close, R. A., Friedman, M., Lloyd, G. T. and Benson, R. B. 2015. Evidence for a mid-Jurassic adaptive radiation in mammals. *Current Biology*, **25** (16), 2137–2142.

Close, R. A., Evers, S. W., Alroy, J., Butler, R. J. and Cooper, N. 2018. How should we estimate diversity in the fossil record? Testing richness estimators using sampling-standardised discovery curves. *Methods in Ecology and Evolution*, **9** (6), 1386–1400.

Close, R. A., Benson, R. B. J., Saupe, E. E., Clapham, M. E. and Butler, R. J. 2020. The spatial structure of Phanerozoic marine animal diversity. *Science*, **368**, 420–424.

Crampton, J. S., Beu, A. G., Cooper, R. A., Jones, C. M., Marshall, B. and Maxwell, P. A. 2003. Estimating the rock volume bias in paleobiodiversity studies. *Science*, **301** (5631), 358–360.

Deline, B. and Ausich, W. I. 2016. Character selection and the quantification of morphological disparity. *Paleobiology*, **43** (1), 68–84.

Foote, M. 1991. Morphological patterns of diversification: examples from trilobites. *Palaeontology*, **34**, 461–485.

Foote, M. 1992. Rarefaction analysis of morphological and taxonomic diversity. *Paleobiology*, **18** (1), 1–16.

Foote, M. 1997. The evolution of morphological diversity. *Annual Review of Ecology and Systematics*, **28** (1), 129–152.

Foote, M. 1999. Morphological diversity in the evolutionary radiation of Paleozoic and post-Paleozoic crinoids. *Paleobiology*, **25** (Suppl), 1–115.

Good, I. J. 1953. The population frequencies of species and the estimation of population parameters. *Biometrika*, **40** (3/4), 237–264.

Guillerme, T. 2018. dispRity: A modular R package for measuring disparity. *Methods in Ecology and Evolution*, **9** (7), 1755–1763.

Guillerme, T., Cooper, N., Brusatte, S. L., Davis, K. E., Jackson, A. L., Gerber, S., Goswami, A., Healy, K., Hopkins, M. J., Jones, M. E., Lloyd, G. T., O'Reilly, J. E., Pate, A., Puttick, M. N., Rayfield, E. J., Saupe, E. E., Sherratt, E., Slater, G. J., Weisbecker, V., Thomas, G. H. and Donoghue, P. C. J. 2020. Disparities in the analysis of morphological disparity. *Biology Letters*, **16** (7), 20200199.

Halliday, T. J. D. and Goswami, A. 2016. Eutherian morphological disparity across the end-Cretaceous mass extinction. *Biological Journal of the Linnean Society*, **118** (1), 152–168.

Hopkins, M. J. and Gerber, S. 2017. Morphological disparity. 1–12. *In* Nuno de la Rosa, L. and Müller, G. (eds) *Evolutionary developmental biology*. Springer International Publishing.

Jablonski, D., Kaustuv, R., Valentine, J. W., Price, R. M. and Anderson, P. S. 2003. The impact of the pull of the recent on the history of marine diversity. *Science*, **300**, 1133–1135.

Jones, M. and Close, R. 2024. Data from: Standardising fossil disparity metrics using sample coverage [dataset]. Dryad. https://doi.org/10.5061/dryad.wpzgmsbxt

Klein, N., Schmitz, L., Wintrich, T. and Sander, P. M. 2020. A new cymbospondylid ichthyosaur (Ichthyosauria) from the Middle Triassic (Anisian) of the Augusta Mountains, Nevada, USA. *Journal of Systematic Palaeontology*, **18** (14), 1167–1191.

Koch, C. F. 1978. Bias in the published fossil record. *Paleobiology*, **4** (3), 367–372.

Kotric, B. and Knoll, A. H. 2015. A morphospace of planktonic marine diatoms. II. Sampling standardization and spatial disparity partitioning. *Paleobiology*, **41**, 68–88.

Lloyd, G. T. 2016. Estimating morphological diversity and tempo with discrete character–taxon matrices: implementation, challenges, progress, and future directions. *Biological Journal of the Linnean Society*, **118** (1), 131–151.

McClain, C. R., Johnson, N. A. and Rex, M. A. 2004. Morphological disparity as a biodiversity metric in lower bathyal and abyssal gastropod assemblages. *Evolution*, **58** (2), 338–348.

McGowan, A. J. and Smith, A. B. 2008. Are global Phanerozoic marine diversity curves truly global? A study of the relationship between regional rock records and global Phanerozoic marine diversity. *Paleobiology*, **34** (1), 80–103.

Nordén, K. K., Stubbs, T. L., Prieto-Márquez, A. and Benton, M. J. 2018. Data from: Multifaceted disparity approach reveals dinosaur herbivory flourished before the end-Cretaceous mass extinction [dataset]. Dryad. https://doi.org/10.5061/dryad.pp07dm0

Novack-Gottshall, P. M., Sultan, A., Smith, N. S., Purcell, J., Hanson, K. E., Lively, R., Ranjha, I., Collins, C., Parker, R., Sumrall, C. D. and Deline, B. 2022a. Nature Ecology & Evolution 2022: Morphological volatility precedes ecological innovation in early echinoderms [dataset]. https://dataverse.harvard.edu/dataverse/CamOrdEchinoderms

Novack-Gottshall, P. M., Sultan, A., Smith, N. S., Purcell, J., Hanson, K. E., Lively, R., Ranjha, I., Collins, C., Parker, R., Sumrall, C. D. and Deline, B. 2022b. Morphological volatility precedes ecological innovation in early echinoderms. *Nature Ecology & Evolution*, **6**, 263–272.

Prentice, K. C., Ruta, M. and Benton, M. J. 2011. Evolution of morphological disparity in pterosaurs. *Journal of Systematic Palaeontology*, **9** (3), 337–353.

Raup, D. M. 1972. Taxonomic diversity during the Phanerozoic: the increase in the number of marine species since the Paleozoic may be more apparent than real. *Science*, **177** (4054), 1065–1071.

R Core Team. 2023. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

RStudio Team. 2023. RStudio: Integrated Development for R. RStudio, PBC, Boston, MA. http://www.rstudio.com/

Romano, M., Brocklehurst, N., Manni, R. and Nicosia, U. 2018. Multiphase morphospace saturation in cyrtocrinid crinoids. *Lethaia*, **51** (4), 538–546.

Roy, K. and Foote, M. 1997. Morphological approaches to measuring biodiversity. *Trends in Ecology & Evolution*, **12** (7), 277–281.

Sanders, H. L. 1968. Marine benthic diversity: a comparative study. *The American Naturalist*, **102** (925), 243–282.

Smith, A. B. 2001. Large–scale heterogeneity of the fossil record: implications for Phanerozoic biodiversity studies. *Philosophical Transactions of the Royal Society B*, **356** (1407), 351–367.

Smith, A. B. 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.

Smith, T. J. and Donoghue, P. C. J. 2022. Evolution of fungal phenotypic disparity. *Nature Ecology & Evolution*, **6** (10), 1489–1500.

Van Valen, L. 1974. Multivariate structural statistics in natural history. *Journal of Theoretical Biology*, **45** (1), 235–247.

Wills, M. A., Briggs, D. E. and Fortey, R. A. 1994. Disparity as an evolutionary index: a comparison of Cambrian and Recent arthropods. *Paleobiology*, **20** (2), 93–130.