Running head: TIME SUB-SAMPLES IN DISPARITY-THROUGH-TIME ANALYSES

Time for a rethink: time sub-sampling methods in disparity-through-time analyses

Thomas Guillerme[1] and Natalie Cooper[2]*

[1]School of Biological Sciences, University of Queensland, St. Lucia, Queensland, Australia.

[2]Department of Life Sciences, Natural History Museum, Cromwell Road, London, SW7 5BD, UK. natalie.cooper@nhm.ac.uk

*Corresponding author

Keywords: time bin, time-slice, disparity, sub-sampling

ABSTRACT

Disparity-through-time analyses can be used to determine how morphological diversity changes in response to mass extinctions, or to investigate the drivers of morphological change. These analyses are routinely applied to palaeobiological datasets, yet although there is much discussion about how to best calculate disparity, there has been little consideration of how taxa should be sub-sampled through time. Standard practice is to group taxa into discrete time bins, often based on stratigraphic periods. However, this can introduce biases when bins are of unequal size, and implicitly assumes a punctuated model of evolution. In addition, many time bins may have few or no taxa, meaning that disparity cannot be calculated for the bin and making it harder to complete downstream analyses. Here we describe a different method to complement the disparity-through-time tool-kit: time-slicing. This method uses a time-calibrated phylogenetic tree to sample disparity-through-time at any fixed point in time rather than binning taxa. It uses all available data (tips, nodes and branches) to increase the power of the analyses, specifies the implied model of evolution (punctuated or gradual), and is implemented in `R`. We test the time-slicing method on four example datasets and compare its performance in common disparity-through-time analyses. We find that the way you time sub-sample your taxa can change your interpretations of the results of disparity-through-time analyses. We advise using multiple methods for time sub-sampling taxa, rather than just time binning, to gain a better understanding disparity-through-time.

INTRODUCTION

Disparity-through-time analyses are common in palaeontology (Gould, 1991; Briggs *et al.*, 1992; Wills *et al.*, 1994; Foote, 1994). They reveal how the morphological diversity of clades has changed through time, and allow us to make inferences about the breadth of ecological niches extinct taxa occupied (Foote, 1997). Results from disparity-through-time studies also provide insights into the ecological impacts of mass extinctions, competitive replacements, and the drivers of morphological evolution (Brusatte *et al.*, 2008*b*; Foote, 1996; Friedman, 2010). Unfortunately, the way we perform these analyses may have profound effects on our conclusions.

Disparity-through-time analyses have two main analysis components: calculating disparity, and creating time sub-subsets of the data. Here we focus on the latter. The nature of disparity (i.e. it is a diversity metric), means it cannot be calculated using a single individual, so some way of sub-sampling taxa is required. Changes in disparity-through-time are generally investigated by calculating the disparity of taxa present during specific time intervals or time bins (e.g. Cisneros & Ruta, 2010; Prentice *et al.*, 2011; Hughes *et al.*, 2013; Hopkins, 2013; Benton *et al.*, 2014; Benson & Druckenmiller, 2014). These time bins are usually defined based on stratigraphy (e.g. Cisneros & Ruta, 2010; Prentice *et al.*, 2011; Hughes *et al.*, 2013; Benton *et al.*, 2014) but can also be arbitrarily chosen time bins of equal (or approximately equal) duration (Butler *et al.*, 2012; Hopkins, 2013; Benson & Druckenmiller, 2014). However, this approach has several limitations.

First, time bins defined by stratigraphy are not of equal size, biasing higher disparity towards longer stratigraphic periods. This can be dealt with using rarefaction methods, i.e. repeating the analysis while resampling the taxa to have the same number of taxa in each bin (e.g. using bootstrapping with limited resampling). This can, however, lead to large confidence intervals when there are stratigraphic periods with few species. Other studies split large time bins so they are of roughly equal size, but this is often an *ad hoc* procedure that can introduce more bias depending on where bins are split. Second, the time binning approaches (whether bins are equally sized or not) favour punctuated equilibrium modes of evolution. Whether the disparity represents an average across the interval (with no interpretation of if or how it varied within the time bin), or it is *effectively* postulated to be constant, when analysing the changes in disparity-through-time, this method will only allow changes in disparity to occur *between* intervals rather than also allowing for gradual changes within intervals (a pattern that is fairly common in the fossil record; Hunt *et al.*, 2015). Third, when investigating changes in disparity due to events at a specific time point (e.g. a mass extinction), time bins may have not have high enough resolution to resolve changes at the event, for example if time bins are every 20 million years it may be hard to capture the effects of an event five million years into the bin. Finally, time bin analyses are often limited by the number of taxa in each bin. If there are insufficient taxa in a time bin, disparity cannot be calculated, so further analyses, e.g. correlations of disparity with hypothesised drivers of morphological evolution, are not possible.

To address these issues, we propose a "time-slicing" approach that takes advantage of the wealth of palaeontological datasets which now have associated phylogenies. Time-slicing uses a phylogenetic tree and considers subsets of taxa at specific equidistant points in time, as opposed to considering subsets of taxa between two points in time (a similar approach is outlined in Halliday & Goswami, 2016). This results in even-sampling across time and permits us to define the underlying model of character evolution (punctuated or gradual). Time-slicing also includes any element present in the phylogeny (branches, nodes and tips) at the time-slice in question as part of the disparity calculation. This allows us to measure disparity at time points where there are no sampled terminal taxa, and increases the sample size at each time point, making downstream analyses of the drivers of disparity much more feasible.

Here we present our time-slicing methods using four datasets taken from the literature. We calculate disparity-through-time for each dataset using a range of time binning and time-slicing methods, and then compare these approaches with respect to the relative disparities calculated, but also investigate how the different approaches influence biological conclusions. We find that the choice of time sub-sampling method can have profound effects on the conclusions of disparity-through-time analyses.

MATERIALS AND METHODS

Overview

To test the different time sub-sampling methods, we followed the protocol below (Fig. 1). All the code needed to reproduce these analyses (along with detailed instructions) is provided on GitHub (`https://github.com/nhcooper123/time-slice`).

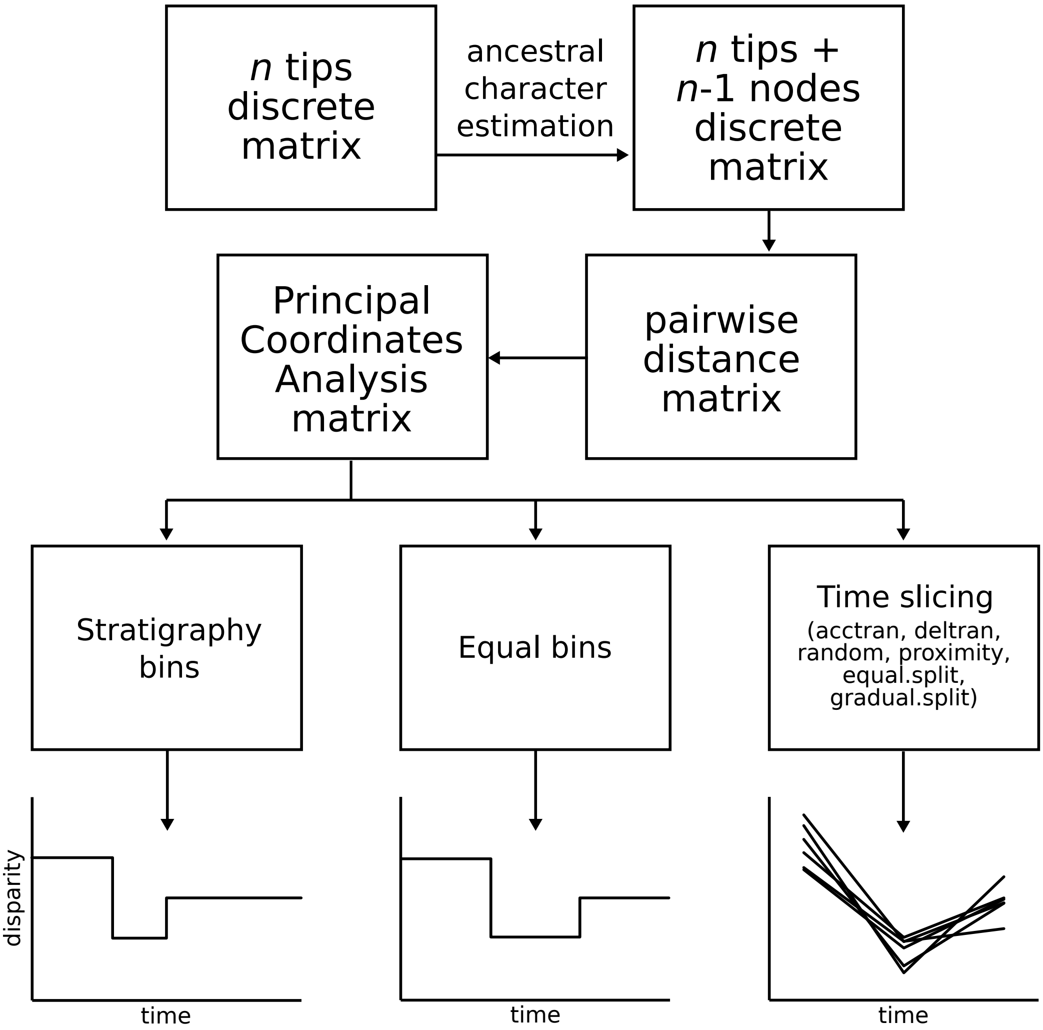

Figure 1: Outline of the disparity-through-time pipeline. 1) We use ancestral character estimation to infer nodal character states; 2) we measure the pairwise Gower distance between the tip character states and nodal character states; 3) we ordinate the distance matrix using principal coordinates analysis (PCoA/PCO); 4) we time sub-sample the PCoA matrix using time bins defined by stratigraphic periods, equally sized time bins and time-slices (using six methods to estimate ordination scores for branches); and finally 5) we measure disparity-through-time for each of these methods.

Example datasets

To test the different time binning/slicing methods we selected four datasets: a mammal dataset from Beck & Lee (2014), two theropod datasets from Brusatte *et al.* (2014*b*) and Bapst *et al.* (2016*b*), and a crinoid dataset from Wright (2017*b*). Table 1 and the Supporting Information Appendix S1 provide more details. Each dataset consists of first and last occurrence dates for all taxa, a matrix of morphological characters in NEXUS format, and a time-scaled phylogeny. These datasets are freely available with their accompanying papers (Table 1), but for reproducibility purposes we also provide the data we used on GitHub (`https://github.com/nhcooper123/time-slice/data`).

Table 1: Details of the datasets used in this study. Age ranges are root time to most recent tip taxon.

	Beck2014	**Brusatte2014**	**Bapst2016**	**Wright2017**
Group	mammals	theropods	theropods	crinoids
# taxa	106	152	89	42
# characters	421	853	374	87
Age range (MYA)	171.8 - 0	168.5 - 66	207.2 - 66	485.4 - 372.2
Mass extinction (MYA)	66 (K-Pg)	NA	NA	443 (O-S)
Reference	Beck & Lee (2014)	Brusatte *et al.* (2014*b*)	Bapst *et al.* (2016*b*)	Wright (2017*b*)
Data reference	Beck & Lee (2014)	Brusatte *et al.* (2014*a*)	Bapst *et al.* (2016*a*)	Wright (2017*a*)

Preparing the data for disparity-through-time analysis

Estimating ancestral character states. For each dataset we estimated the ancestral character states at each node using the `AncStatesEstMatrix` function from the `Claddis` `R` package (Lloyd, 2015; R Core Team, 2015). This function uses the re-rooting method (Yang *et al.*, 1996; Garland & Ives, 2000) to get Maximum Likelihood estimates of the ancestral states for each character at every node in the phylogeny (based on the `rerootingMethod` function in `phytools`; Revell, 2012). Inapplicable and missing characters for any taxon were treated as ambiguous characters (i.e. any possible observed state for the character). To prevent poor ancestral state estimations from biasing our results, especially when a lot of error is associated with the estimations, we only included ancestral state estimations with a scaled Likelihood ≥ 0.95. Ancestral state estimations with scaled Likelihoods below this threshold were recoded as missing data ("?"). This allowed our results to be less dependent on the quality (or the absence thereof) of the ancestral state estimations methods, especially in parts of the datasets where data were sparse. This approach is similar to Brusatte *et al.* (2011) but uses model based estimations (rather than parsimony) allowing us to control for ambiguous (i.e. poorly estimated) nodes.

Building morphospaces. To explore disparity-through-time in our datasets, we used a morphospace approach (e.g. Foote, 1994, 1996; Wesley-Hunt, 2005; Brusatte *et al.*, 2008*b*; Friedman, 2010; Toljagic & Butler, 2013; Hughes *et al.*, 2013). Morphospaces can be obtained from any multidimensional morphological data set but can differ in the data used (e.g. discrete or continuous), and whether they include phylogenetic data or not. Although empirical morphospaces from discrete or continuous data have been shown to have similar properties (Foth *et al.*, 2012; Hetherington *et al.*, 2015), our morphospaces are based on discrete morphological data (originally collected for phylogenetic analysis; *c.f.* geometric morphometric data) and include some phylogenetic information (see above). Mathematically, our morphospaces are n dimensional objects that summarise the distances between discrete morphological characters of the taxa present and their ancestors.

Constructing distance matrices. To estimate the morphospaces for each of our datasets we first constructed pairwise distance matrices of length n, where n is the total number of tips and nodes in the dataset. We calculated the $n \times n$ distances using the Gower distance (Gower, 1971), i.e. the number of mismatched characters over the number of shared characters. This allows us to correct for distances between two taxa that share many characters and could be closer to each other than to taxa with fewer characters in common (i.e. because some pairs of taxa share more characters in common than others, they are more likely to be similar). For discrete morphological matrices, using this corrected distance is preferable to the raw Euclidean distance because of its ability to deal with discrete or/and ordinated characters as well as with missing data (Anderson & Friedman, 2012). However, the Gower distance cannot calculate distances when taxa have no overlapping data. Therefore, we used the `TrimMorphDistMatrix` function from the `Claddis` `R` package to remove pairs of taxa with no cladistic characters in common. This led to us removing nine taxa from the Bapst *et al.* (2016*b*) dataset, and 19 from the Brusatte *et al.* (2014*b*) dataset, but none from the other two datasets (see Supporting Information Appendix S1 for details of which species).

Ordination. After constructing our distance matrices we transformed them using classical multidimensional scaling (MDS; Torgerson, 1965; Gower, 1966; Cailliez, 1983). This method (also referred to as PCO; e.g. Brusatte *et al.* 2015; or PCoA; e.g. Paradis *et al.* 2004; but distinguished in Legendre & Legendre 2012) is an eigen decomposition of the distance matrix. Because we used Gower distances instead of raw Euclidean distances, negative eigenvalues can be calculated. To avoid this problem, we first transformed the distance matrices by applying the Cailliez correction (Cailliez, 1983) which adds a constant c^* to the values in a distance matrix (apart from the diagonal) so that all the Gower distances become Euclidean ($d_{Gower} + c^* = d_{Euclidean}$; Cailliez 1983). We were then able to extract k eigenvectors for each matrix (representing the k dimensions of the morphospace) where k is equal to $n - 2$, i.e. the number of taxa in the matrix (n) minus the last two eigenvectors that are always null after applying the Cailliez correction. Contrary to previous studies (e.g Brusatte *et al.*, 2008*a*; Cisneros & Ruta, 2010; Prentice *et al.*, 2011; Anderson & Friedman, 2012; Hughes *et al.*, 2013; Benton *et al.*, 2014), we use all k dimensions of our morphospaces and not a sub-sample representing the majority of the variance in the distance matrix (e.g. selecting only x dimensions that represent up to 90% of the variance in the distance matrix; Brusatte *et al.* 2008*b*; Toljagic & Butler 2013). Note that our morphospaces represent an ordination of all possible morphologies coded in each study through time. It is unlikely that all morphologies will co-occur at each time point, therefore, the disparity of the whole morphospace is expected to be greater than the disparity at any specific point in time.

Disparity-through-time analyses

Disparity-through-time analyses were performed using the `dispRity` `R` package (Guillerme, 2016).

Calculating disparity. Disparity can be calculated in many different ways (e.g. Wills *et al.*, 1994; Ciampaglio, 2004; Thorne *et al.*, 2011; Hopkins, 2013; Huang *et al.*, 2015), however a majority of studies in palaeobiology estimate disparity using four metrics: the sum and products of ranges and variances, each of which gives a slightly different estimate of how the data fits within the morphospace (Foote, 1994; Wills *et al.*, 1994; Brusatte *et al.*, 2008*a*,*b*; Cisneros & Ruta, 2010; Thorne *et al.*, 2011; Prentice *et al.*, 2011; Brusatte *et al.*, 2012; Toljagic & Butler, 2013; Ruta *et al.*, 2013; Benton *et al.*, 2014; Benson & Druckenmiller, 2014). However, these metrics have limitations. First, the range metrics are affected by the uneven sampling of the fossil record (Butler *et al.*, 2012). Second, because we include all k dimensions in the analysis (see above), the products of ranges and variances will tend towards zero since the scores of the last dimension are usually really close to zero themselves. We therefore use the sum of variances metric to estimate disparity here:

$$\text{disparity} = \sum \sigma^2 k_i \tag{1}$$

where $\sigma^2 k_i$ is the variance for the k_i^{th} dimension ranging from n to $n-2$ with n being the number of taxa in the dataset. Note that there are still statistical issues with this metric (such as the co-variance between dimensions not being measured), but for the purposes of comparison with previous work we decided to use a standard metric for these analyses.

Time sub-sampling To estimate disparity-through-time we first need to split the data into time sub-samples. Here we use three time sub-sampling methods.

1. Stratigraphic time bins. This is the traditional method, where all the taxa within each stratigraphic period are included in the disparity calculation. This often leads to bins of unequal duration. Here we use stratigraphic stages and epochs.
2. Equally sized time bins. This is another commonly used method, where the time frame of interest is split into equally sized time bins, then all the taxa within each time bin are included in the disparity calculation.
3. Time-slicing. We describe this in more detail below, but in brief, time-slicing uses a phylogeny, and rather than binning the data, it takes slices through a phylogeny and includes all the taxa and nodes in that slice within the disparity calculation.

Time-slicing. The "time-slicing" approach considers subsets of taxa in the morphospace at specific equidistant points in time, as opposed to considering subsets of taxa between two points in time. This results in even-sampling of the morphospace across time and allows us to use different underlying models of character evolution (punctuated or gradual).

In practice, time-slicing considers the disparity of any element present in the phylogeny (branches, nodes and tips) at any point in time. When the phylogenetic elements are nodes or tips, the ordination scores for the nodes (estimated using ancestral state reconstruction as described above) or tips are directly used for calculating disparity. When the phylogenetic elements are branches we choose the ordination score for the branch using one of two evolutionary models:

1. **Punctuated evolution.** This model selects the ordination score from either the ancestral node or the descendant node/tip of the branch regardless of the position of the slice along the branch. Similarly to the time bin approach, this reflects a model of punctuated evolution where changes in disparity occur either at the start or at the end of a branch over a relatively short time period, and clades undergo long periods of stasis during their evolution (Gould & Eldredge, 1977; Hunt, 2007). We apply this model in four ways:
    (i) The "acctran" model, always selecting the ordination score of the descendant node/tip of the branch.
    (ii) The "deltran" model, always selecting the ordination score of the ancestral node of the branch.
    (iii) The "random" model, randomly selecting the ordination score of either the ancestor or the descendant of the branch.
    (iv) The "proximity" model, selecting the ordination score of the ancestor if the slice occurs in the first half of the branch, and the descendant if the slice occurs in the second half of the branch.

   The two first models assume that changes always occur early (*acc*elerated *tran*sition) or late along the branches (*del*ayed *tran*sition). The third model makes neither assumption and simply selects data from the ancestor or the descendant at random, and the fourth bases the selection of either the ancestor or the descendant on where the slice occurs along the branch. These punctuated models only select either the ordination score from the ancestor and the descendant once in the whole disparity analysis. For example, if using the "random" model, if the data of the ancestor has been randomly chosen, only this data will be used during the bootstrapping (see below) and for the disparity calculation.

2. **Gradual evolution.** Unlike the punctuated models, the following models do not select the ordination score of either the ancestor or the descendant but associate a probability to both. This reflects a model of gradual evolution where changes in disparity are gradual and cumulative along the branch.

    (v) The "equal splits" model (probabilistic), selects the ordination score from both the ancestor and the descendant with an equal probability:

$$p(\text{ancestor}) = p(\text{descendant}) = 0.5 \tag{2}$$

    (vi) The "gradual splits" model (probabilistic), selects the ordination score from both the ancestor and the descendant with a probability function of the distance between the nodes/tip at the ends of the branch and the slice:

$$p(\text{ancestor}) = \frac{d(\text{ancestor}, \text{slice})}{d(\text{ancestor}, \text{descendant})} \tag{3}$$

$$p(\text{descendant}) = 1 - p(\text{ancestor}) \tag{4}$$

where $d(x, y)$ is the distance between the two elements x, y (ancestor, slice or descendant) measured in units of branch length.

In these models, the ordination scores of both the ancestor and descendant contribute to the disparity calculation. For example, using the "gradual splits" model, if the slice occurs in the third quarter of a branch joining node A to node/tip B (75% of the total branch length), after bootstrapping, the disparity results will be based on 25% of the data from A and 75% of the data from B. Because of the probabilistic nature of these models, they are only meaningful when calculating disparity from bootstrapped datasets.

It is important to note that the time-slicing method is not an ancestral states estimation method *per se*. This method does not estimate values along a branch applying a model (*c.f.* methods described for ancestral character estimation in the "Preparing the data for disparity-through-time analysis" section above) but rather chooses between the two available pieces of information (the ordination score of the descendant or the ancestor) using the methods described above. This allows the method to be used in post-ordination analysis where the data used in each time-slice is data already present in the morphospace. In other words, this method does not require a re-ordination of the morphospace every time a slice goes through a branch, thus allowing the properties of the morphospace (e.g. distance between species, variance of each axis, etc.) to remain constant. For example, using the "equal.splits" model on an ancestor and a descendant with PCO1 values of respectively 0.04 and 0.03, after a sufficient number of bootstrap replicates (e.g. 100) the value along the branch will be close to $0.5 \times 0.04 + 0.5 \times 0.03 = 0.035$. By estimating this value rather than generating it (i.e. creating a new element mid-way along the branch that would be the average of the descendant and ancestor - 0.035) we obtain the same results without modifying the morphospace properties.

Comparing time sub-sampling methods

To compare the time binning and time-slicing approaches we applied the methods as follows (see Fig. **??**).

1. Stratigraphy: time sub-samples defined by stratigraphic periods (Fig. 2).
    (i) Time bins (unequal). We calculated disparity for the taxa in each stratigraphic period (stage or epoch). To reduce the influence of outliers on our disparity estimates, we bootstrapped each disparity measurement for each time bin by randomly resampling with replacement a new sub-sample of taxa from the observed taxa in the bin 100 times. We then calculated the median disparity value for each time bin along with the 50% and 95% confidence intervals.

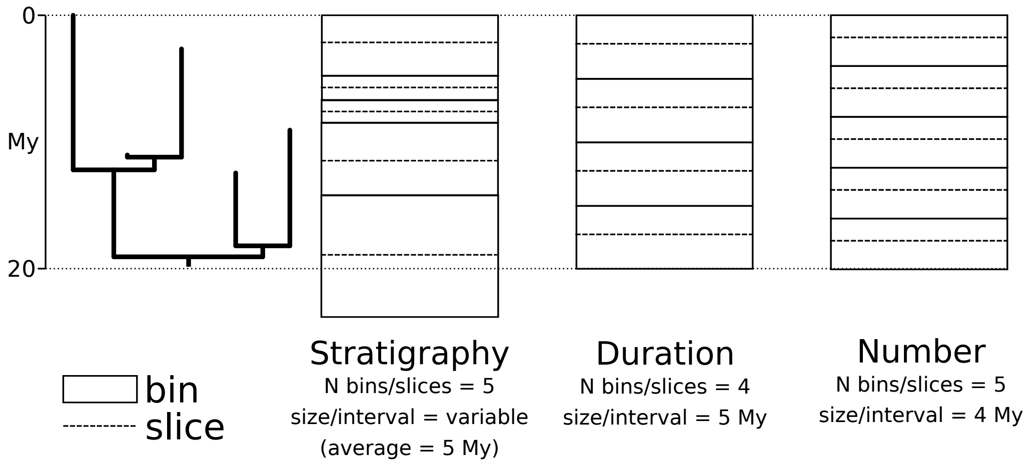

Figure 2: Outline of the three time sub-sampling methods. Stratigraphy: time sub-samples are defined by stratigraphic periods. Here there are five stratigraphic periods in the 20 My (million years) time frame of interest, i.e. five bins/slices with variable sizes/intervals. Duration: time sub-samples are defined based on the mean duration of stratigraphic periods in the time frame of interest. Here, the mean duration of stratigraphic periods is 5 My, so there are four bins/slices of 5 My duration (or four slices with 5 My intervals between them) in the 20 My time frame of interest. Number: time sub-samples are defined based on the number of stratigraphic periods in the time frame of interest. Here, there are five stratigraphic periods, so there are five bins/slices of 4 My duration (or five slices with 4 My intervals between them) in the 20 My time frame of interest.

    (ii) Time-slices (non-equidistant). We calculated disparity using our time-slicing approach with slices occurring at the midpoint of each stratigraphic period (stage or epoch), and using all six time-slicing methods (acctran, deltran, random, proximity, equal splits and gradual splits). To reduce the influence of outliers on our disparity estimates, we bootstrapped each disparity measurement as described above for the stratigraphic time bins.

2. Duration: time sub-samples defined by the duration of stratigraphic periods (Fig. 2).

    (i) Time bins (equal). We calculated disparity for the taxa in each time bin where time bin size was defined by the mean *duration* of the stratigraphic period (stage or epoch), and bootstrapped the disparity values as described above.

    (ii) Time-slices (equidistant). We calculated disparity using our time-slicing approach where the interval between slices, was defined by the mean *duration* of the stratigraphic period (stage or epoch). We used the six time-slicing methods and bootstrapped as described above.

3. Number: time sub-samples defined by the number of stratigraphic periods (Fig. 2).

    (i) Time bins (equal). We calculated disparity for the taxa in each time bin where the number of time bins was defined by the *number* of stratigraphic periods (ages or epochs) in the time frame of interest, and bootstrapped the disparity values as described above.

(ii) Time-slices (equidistant). We calculated disparity using our time-slicing approach where the number of slices, was defined by the *number* of stratigraphic periods (ages or epochs) in the time frame of interest. We used the six time-slicing methods and bootstrapped as described above.

We also recorded the number of taxa (or taxa and nodes for time-slicing methods) in each sub-sample as a proxy for taxonomic diversity.

Testing for differences in the time sub-sampling methods

Testing for statistical differences among the time sub-sampling methods described above is difficult, as we need to compare similar units, and also to tackle questions important to the interpretation of disparity-through-time analyses. We therefore present three different, simple ways of comparing the time sub-sampling methods as follows.

Systematic differences in disparity-through-time. To test whether using time bins or time-slices resulted in significantly different disparity values at common time points, we used paired Wilcoxon tests to compare the median bootstrapped disparities obtained in the stratigraphy (time sub-samples defined by stratigraphic periods), duration (time sub-samples defined by the duration of stratigraphic periods), and number (time sub-samples defined by the number of stratigraphic periods) analyses described above.

Due to the uneven spread of taxa across phylogenies, some time bins will contain one or no species, meaning that we cannot estimate disparity for that time bin. We first, therefore, removed the time bins, and corresponding time-slices, without disparity estimates. We then performed paired Wilcoxon tests with Bonferroni corrected p-values, so that bins and slices for the same time period are being compared. Significant results suggest that there is a systematic difference in disparity values at each time point, depending on whether bins or slices are used.

Disparity peaks. We are perhaps more interested in how the conclusions of disparity-through-time analyses are influenced by the choice of time sub-sampling method, rather than the disparities estimated by each method *per se*, especially as these will be influenced by the number of taxa (and/or nodes) included in each sub-sample. Therefore, we also investigated where peaks of disparity occurred in each of our datasets for the different time sub-sampling methods. We calculated the maximum bootstrapped disparities for each dataset and for each time sub-sampling method, along with their associated confidence intervals. Significant shifts in disparity peaks suggest that the choice of time sub-sampling method will influence our conclusions about relative changes in the disparity of our groups through time.

Effects of mass extinction events. Many analyses of disparity-through-time aim to demonstrate differences in disparity before and after mass extinction events. Two of our four datasets contain taxa before and immediately after a mass extinction (Cretaceous-Paleogene 66 MYA; Beck & Lee 2014; Ordovician-Silurian 455-430 MYA; Wright 2017*b*), so we used Wilcoxon tests with Bonferroni corrected p-values to compare disparity in the time bin/slice prior to the appropriate mass extinction, to that of the time bin/slice following the extinction event. Significant results suggest an effect of the mass extinction on disparity in the group. We then compare these results across the time sub-sampling methods to determine if our conclusions change depending on the method used. We repeated these analyses using the two time bins/slices after the one immediately following the mass extinction event to account for any lag effects of the mass extinction on disparity.

RESULTS

Disparity-through-time analyses

Disparity changes through time for each of our four datasets (Fig. 3, Supporting Information Appendix S2: Figs A1-A2). Relative disparities tend to be lower with time binning methods, likely because these contain fewer taxa than time-slicing methods. The six different time-slicing methods (acctran, deltran, random, proximity, equal splits and gradual splits) show similar patterns, so we focus only on the results for one method with a punctuated model of evolution (specifically the 'proximity' method), and one method with a gradual model of evolution (specifically the 'gradual splits' method). Results for all six methods can be found in Supporting Information Appendix S2: Figures A1-A2.

Testing for differences in the time sub-sampling methods

Systematic differences in disparity-through-time. There is no overall significant systematic difference among the disparities calculated using time bins and those calculated using the time-slicing methods (Table 2, Supporting

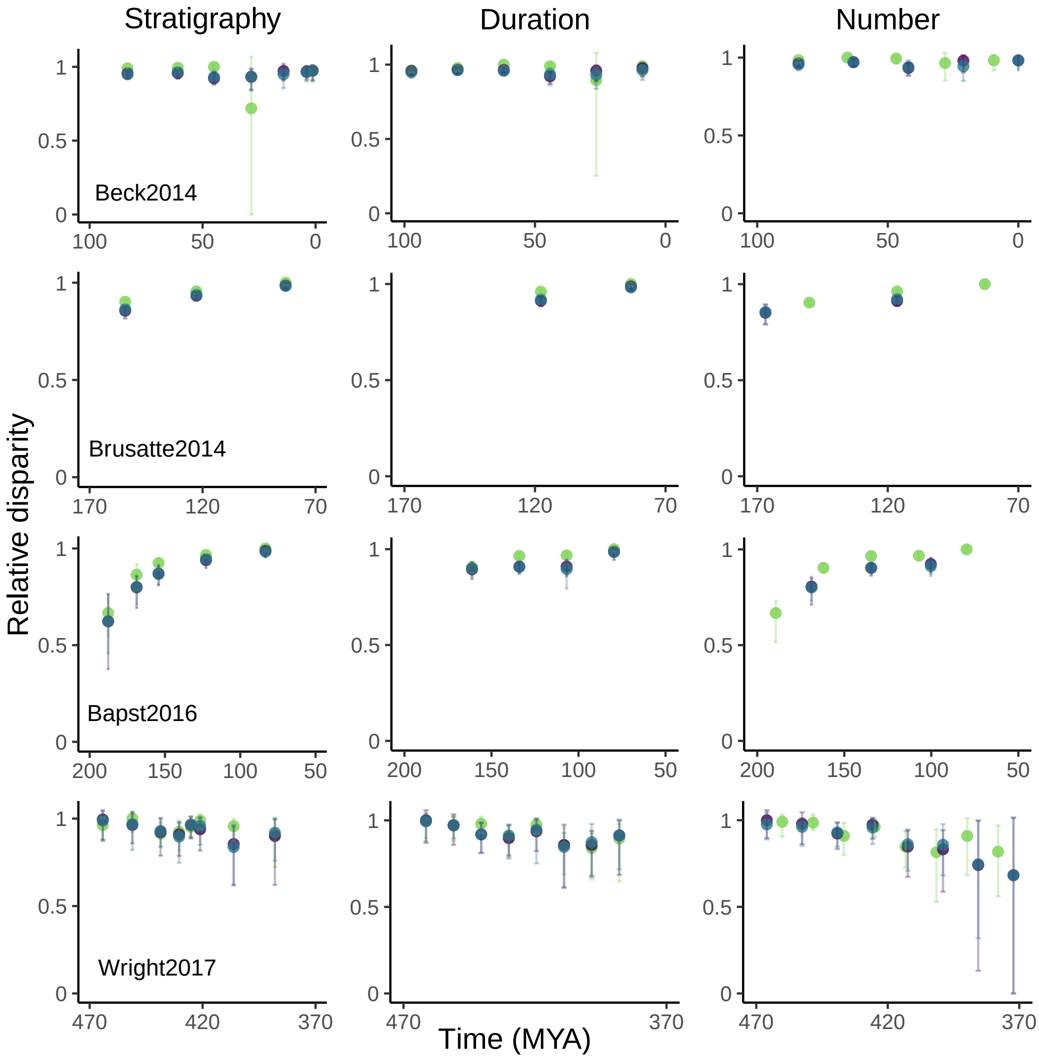

Figure 3: Relative disparity-through-time. Median bootstrapped disparities were calculated using time binning and time-slicing approaches. Green points represent time binning methods, purple points are time-slices with a punctuated model of evolution ('proximity' method), and blue points are time-slices with a gradual model of evolution ('gradual splits' method). Relative disparities (median bootstrapped disparity divided by the maximum median bootstrapped disparity for a dataset and analysis method) are presented so they can be compared across datasets/methods. Stratigraphy uses unequal time bins or non-equidistant time-slices, where the width of the bin, or the interval between slices, is equivalent to stratigraphic epochs. Duration uses equal time bins or equidistant time-slices, where the width of the bin, or the interval between slices, is the mean duration of stratigraphic epochs in the time frame of the dataset. Number uses equal time bins or equidistant time-slices, where the number of bins, or the number of slices, is the mean number of stratigraphic epochs in the time frame of the dataset. In all cases, time bin disparities are plotted at the midpoint of the bin, and error bars represent the 95% confidence intervals around the bootstrapped median disparity. The four dataset names are on the first plot for each dataset (see Table 1 for details). Results for stratigraphic stages, and for other time-slicing methods, are in the Supporting Information Appendix S2: Figures A1-A2.

Information Appendix S2: Table A1). Instead, the differences depend on the dataset and method in question. For example, the Brusatte *et al.* (2014*b*), Bapst *et al.* (2016*b*) and Wright (2017*b*) datasets, show significant differences when using bins versus time-slices defined by stratigraphy, but the Beck & Lee (2014) dataset appears robust to these different approaches. Likewise, the Beck & Lee (2014), Brusatte *et al.* (2014*b*) and Bapst *et al.* (2016*b*) datasets have different disparities when the number of bins or slices is the mean number of stratigraphic periods, but this is not seen in the Wright (2017*b*) dataset. Note that for epochs, we find fewer significant differences simply because the smaller numbers of bins and slices being compared means we have low power to detect a significant difference.

Disparity peaks. In the Beck & Lee (2014) and Bapst *et al.* (2016*b*) datasets, disparity peaks occur much at much older ages when time-slicing rather than time binning approaches are used (Fig. 4; Supporting Information Appendix S2: Figs A3-A4). This is also true for stratigraphic time bins in the Wright (2017*b*) dataset, although when using equal time bins the peaks are later than the time-slicing methods, or very similar (Fig. 4; Supporting Information Appendix S2: Figs A3-A4). Across the three time binning methods, the Brusatte *et al.* (2014*b*) dataset has similar disparity peaks whichever method is used, the Wright (2017*b*) dataset only had variation in peaks when using unequal time bins (stratigraphy), whereas in the Bapst *et al.* (2016*b*) and Beck & Lee (2014) datasets, stratigraphic (unequal) versus equally sized time bins make a large difference to where the disparity peak occurs (Fig. 4; Supporting Information Appendix S2: Figs A3-A4). Additionally, there seem to be small discrepancies within the

Table 2: Results of paired Wilcoxon tests investigating whether disparities calculated using time bins are significantly different to those calculated using time-slices. Time-slices used either a punctuated (proximity method) or gradual (gradual.split method) model of evolution. Stratigraphy uses unequal time bins or non-equidistant time-slices, where the width of the bin, or the interval between slices, is equivalent to stratigraphic ages or epochs. Duration uses equal time bins or equidistant time-slices, where the width of the bin, or the interval between slices, is the average duration of stratigraphic ages or epochs in the time frame of the dataset. Number uses equal time bins or equidistant time-slices, where the number of bins, or the number of slices, is the average number of stratigraphic ages or epochs in the time frame of the dataset. P-values were Bonferroni corrected. $***p < 0.001$. Results for other time-slicing methods are in the Supporting Information Appendix S2: Table A1.

Dataset	**Period**	**Method**	**Stratigraphy**	**Duration**	**Number**
Beck2014	Age	gradual.split	111	115***	65***
Beck2014	Age	proximity	105	83	68***
Beck2014	Epoch	gradual.split	21	39	43***
Beck2014	Epoch	proximity	21	36	32
Brusatte2014	Age	gradual.split	28***	61***	52***
Brusatte2014	Age	proximity	27***	31	28***
Brusatte2014	Epoch	gradual.split	3	6	6
Brusatte2014	Epoch	proximity	0	5***	5
Bapst2016	Age	gradual.split	93	153	165
Bapst2016	Age	proximity	57***	47	75***
Bapst2016	Epoch	gradual.split	4	6	12
Bapst2016	Epoch	proximity	2	0***	8
Wright2017	Age	gradual.split	152***	155	116
Wright2017	Age	proximity	160***	175***	101
Wright2017	Epoch	gradual.split	28	29	21
Wright2017	Epoch	proximity	23	28	18

time-slicing methods (gradual.split *vs.* proximity) except in the Beck & Lee (2014) dataset where the gradual split model recovered disparity peaks at younger ages than the proximity model (Fig. 4; Supporting Information Appendix S2: Figs A3-A4)

Effects of mass extinction events. Mass extinction events influence disparity in both the Beck & Lee (2014) and Wright (2017*b*) datasets (Fig. 5). However, whether this change in disparity is significant or not depends on the method used to create time sub-samples (Fig. 5), and whether stages or epochs are used. In general, for the Beck & Lee (2014) dataset, time binning tended to give more significant results than time-slicing methods, but this was not the case for the Wright (2017*b*) dataset.

DISCUSSION

Disparity-through-time analyses are influenced by the choice of time sub-sampling method used to divide the taxa. While differences in the relative disparities calculated among time sub-sampling methods may not be of much biological importance, these changes can have important implications for the conclusions of downstream analyses. For example, using stratigraphic epochs as our reference time period, there are 21 potential methods for time sub-sampling our data (splitting by stratigraphy, number and duration, see methods, and using time bins or one of six time-slicing methods). Of these 21 methods, in 16 (76%) we show that placental mammals (Beck & Lee, 2014) significantly increased in disparity in the time bin/slice immediately after the K-Pg mass extinction event, and in 20 (95%) we show that crinoids (Wright, 2017*b*) significantly decreased in disparity in the time bin/slice immediately after the O-S

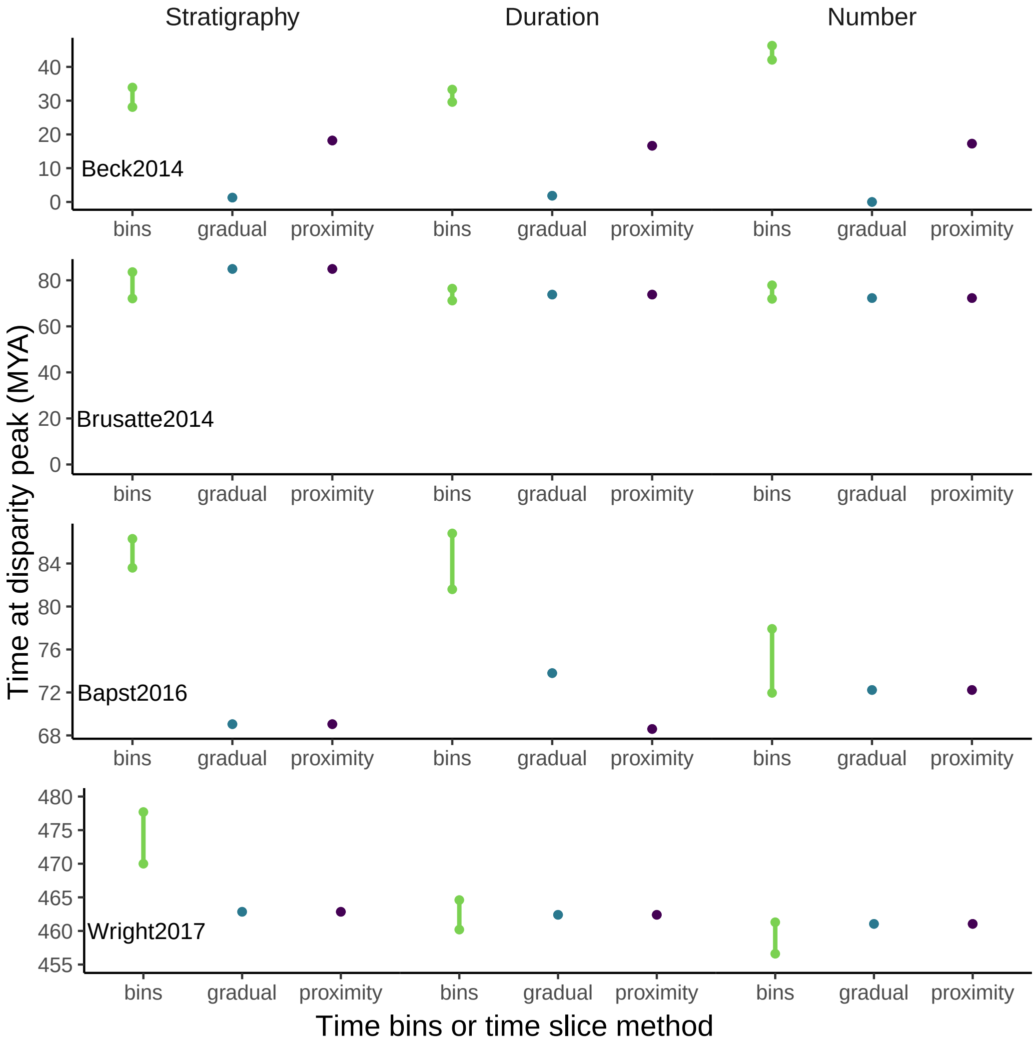

Figure 4: Timing of peak disparity. Median bootstrapped disparities were calculated using time binning and time-slicing approaches. Green points represent time binning methods, blue points are time-slices with a punctuated model of evolution ('proximity' method), and purple points are time-slices with a gradual model of evolution ('gradual splits' method). Stratigraphy uses unequal time bins or non-equidistant time-slices, where the width of the bin, or the interval between slices, is equivalent to stratigraphic epochs. Duration uses equal time bins or equidistant time-slices, where the width of the bin, or the interval between slices, is the mean duration of stratigraphic epochs in the time frame of the dataset. Number uses equal time bins or equidistant time-slices, where the number of bins, or the number of slices, is the mean number of stratigraphic epochs in the time frame of the dataset. For time bins the points indicate the maximum and minimum ages of the time bin within which peak disparities appeared. The four dataset names are on the first plot for each dataset (see Table 1 for details). Results for stratigraphic stages, and for other time-slicing methods, are in the Supporting Information Appendix S2: Figures A3-A4.

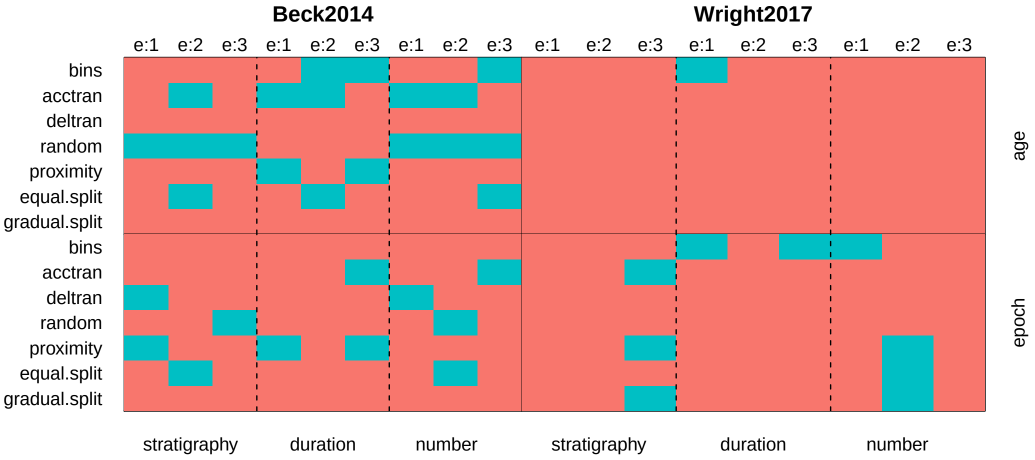

Figure 5: Effects of mass extinction events on disparity. Pink cells and blue cells indicate respectively a significant or non significant change in disparity before and after the mass extinction event (Cretaceous-Paleogene 66 MYA; Beck & Lee 2014; Ordovician-Silurian 455-430 MYA; Wright 2017*b*). *e:1*, *e:2*, and *e:3* denote whether the comparison was between the time bin or time-slice immediately after the mass extinction (*e:1*), or the second (*e:2*) or third (*e:3*) bin/slice after the mass extinction to account for any lag effect. The top seven rows use stratigraphic stages and the bottom seven rows use stratigraphic epochs. Labels on the left hand side indicate whether time bins ("bins") were used, or which of the six time-slicing methods was used.

mass extinction event. Given the high congruence (76% and 95%) of these results, one could argue that time-sub-sampling methods are not important. However, if we had chosen to investigate crinoid disparity only using time bins and splitting these so the number of time bins was equal to the number of epochs (number), we would have concluded that the O-S extinction had no effect on crinoid disparity. Likewise, the timing of peak disparity differs among methods. This is particularly evident when comparing stratigraphic time bins to time slicing methods, where for most of our datasets we see a much later time to peak disparity. This could have major implications for our understanding of how morphological diversity changes through time, for example in response to climate. These results highlight the sensitivity of disparity-through-time analyses to the choice of time sub-sampling method. Fortunately this issue is easy to solve; either disparity-through-time analyses should use, and report results from, multiple time sub-sampling methods (as demonstrated here), or great care should be taken in determining the appropriate time sub-samples to answer the question of interest.

Time-slicing has several advantages over time binning (using either equally or unequally sized bins) approaches. First, it allows us to use as much of the information available to us, in the form of phylogenetic relationships and ancestral taxa, as possible. This increases our ability to investigate key biological questions, such as how do various drivers influence morphological diversity through time, and how do mass extinctions influence disparity (Brusatte *et al.*, 2008*b*; Foote, 1996; Friedman, 2010), both by increasing the statistical power of analyses and through the availability of data at key time points in the history of our groups. Second, we are able to be more explicit about the mode of evolution in our clades; in time-slicing we can apply punctuated or gradual models of trait change rather than making an assumption of punctuated evolution. This may be important, as gradual change is a common pattern of trait evolution in the fossil record (Hunt, 2007).

Of course the method also has limitations. The main one of these is a practical one; it requires a time-calibrated phylogeny and these are not available for all palaeontological datasets. Furthermore, like most phylogeny based methods, time-slicing depends on ancestral state estimations. Care should be taken in interpreting these, as they are highly dependent on the data and models used for the estimations (Slater *et al.*, 2012; Ekman *et al.*, 2008). The difference between the time-binning and time-slicing results could also simply be due to the nature of the fossil record. Rates of sedimentation vary in time and space influencing the groups found within the rock record and their temporal distribution. In this case, different beds could represent different "packages" of fauna through time separated by gaps, resulting in natural "bins" rather than slices of the data. Slicing through such strata will yield similar results no matter where in time the slice occurs. It is important to note however, that the time slicing method also includes ancestral estimations (either through the nodes or the branches) that are by definition not available in the fossil record and thus are not influenced by its nature. Additionally, this effect is likely to be most obvious in groups where the fossil record is "patchy", e.g. vertebrates, but less problematic for groups with a more continuous record like Foraminifera. Finally, Hunt *et al.* (2015) found that time series are best characterized by gradual directional changes (biased random walks). In fact, homogeneous directional changes are more likely to be supported than heterogeneous ones (e.g. punctuated changes) in longer duration series with few samples in each series. In our implementation of time-slicing, the models are not selected based on any model fit criterion (e.g. AIC) but merely on researcher assumptions. We thus suggest that both types of models (punctuated and gradual) are tested during analysis, unless there is strong independent support for one or the other.

CONCLUSIONS

The choice of time sub-sampling method can alter the conclusions we obtain from disparity-through-time studies. Time-slicing methods, with explicit models of evolution, provide an alternative to traditional time binning approaches. Note that while we introduce the time-slicing methods here, and describe their advantages, we do not suggest time-slicing is necessarily the "best" method for time sub-sampling in all cases. As with all methods, the choice of methodology should be appropriate for the question and data at hand. However, we do strongly recommend performing disparity-through-time analyses using a series of appropriate time sub-sampling methods, and comparing these, to ensure that results are not merely a consequence of the time sub-sampling method employed.

DATA ARCHIVING STATEMENT

Data for this study are available from Bapst *et al.* (2016*a*); Wright (2017*a*); Brusatte *et al.* (2014*a*); Beck & Lee (2014) but for ease of reproducibility can also be found on GitHub (`https://github.com/nhcooper123/time-slice/data`). Supporting information can be found on Dryad at https://doi.org/10.5061/dryad.vp4q518 (Guillerme & Cooper, 2018).

SUPPORTING INFORMATION

Additional Supporting information can be found on Dryad at:
https://doi.org/10.5061/dryad.vp4q518 (Guillerme & Cooper, 2018).

Appendix S1: Extra details of datasets.

Appendix S2: Additional figures.

Appendix S3: Additional tables.

Acknowledgments. NC thanks Mark Sutton and Philip Mannion for the invitation to contribute to the 'Evolutionary Modelling' symposium at The Palaeontological Association Annual Meeting 2017. TG acknowledges support from the Australian Discovery Project Grant number DP170103227 awarded to Vera Weisbecker. We thank Dave Bapst, Graeme Lloyd, April Wright and David Wright for assistance in gathering data for the analyses and/or discussions about the approach; and Steve Brusatte, Sally Thomas and one anonymous reviewer for helpful comments on the manuscript.

REFERENCES

References

Anderson, P. S. & Friedman, M. 2012. Using cladistic characters to predict functional variety: experiments using early Gnathostomes. *Journal of Vertebrate Paleontology*, **32**, 1254–1270.

Bapst, D., Wright, A., Matzke, N. & Lloyd, G. 2016*a*. Data from: Topology, divergence dates, and macroevolutionary inferences vary between different tip-dating approaches applied to fossil theropods (dinosauria).

———. 2016*b*. Topology, divergence dates, and macroevolutionary inferences vary between different tip-dating approaches applied to fossil theropods (dinosauria). *Biology letters*, **12**, 20160237.

Beck, R. M. & Lee, M. S. 2014. Ancient dates or accelerated rates? Morphological clocks and the antiquity of placental mammals. *Proceedings of the Royal Society B: Biological Sciences*, **281**, 1–10.

Benson, R. B. J. & Druckenmiller, P. S. 2014. Faunal turnover of marine tetrapods during the Jurassic-Cretaceous transition. *Biological Reviews*, **89**, 1–23.

Benton, M. J., Forth, J. & Langer, M. C. 2014. Models for the rise of the dinosaurs. *Current biology*, **24**, R87–R95.

Briggs, D. E., Fortey, R. A. & Wills, M. A. 1992. Morphological disparity in the cambrian. *Science*, **256**, 1670–1673.

Brusatte, S., Butler, R. J., Prieto-Márquez, A. & Norell, M. A. 2012. Dinosaur morphological diversity and the end-Cretaceous extinction. *Nature Communications*, **3**, 804–804.

Brusatte, S., Lloyd, G., Wang, S. & Norell, M. 2014*a*. Data from: Gradual assembly of avian body plan culminated in rapid rates of evolution across dinosaur-bird transition.

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

———. 2008*b*. Superiority, competition, and opportunism in the evolutionary radiation of dinosaurs. *Science*, **321**, 1485–1488.

Brusatte, S. L., Butler, R. J., Barrett, P. M., Carrano, M. T., Evans, D. C., Lloyd, G. T., Mannion, P. D., Norell, M. A., Peppe, D. J., Upchurch, P. & Williamson, T. E. 2015. The extinction of the dinosaurs. *Biological Reviews*, **90**, 628–642.

Brusatte, S. L., Lloyd, G. T., Wang, S. C. & Norell, M. A. 2014*b*. Gradual assembly of avian body plan culminated in rapid rates of evolution across the dinosaur-bird transition. *Current Biology*, **24**, 2386–2392.

Brusatte, S. L., Montanari, S., Yi, H.-y. & Norell, M. A. 2011. Phylogenetic corrections for morphological disparity analysis: new methodology and case studies. *Paleobiology*, **37**, 1–22.

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

Cailliez, F. 1983. The analytical solution of the additive constant problem. *Psychometrika*, **48**, 305–308.

Ciampaglio, C. N. 2004. Measuring changes in articulate brachiopod morphology before and after the permian mass extinction event: do developmental constraints limit morphological innovation? *Evolution and Development*, **6**, 260–274.

Cisneros, J. C. & Ruta, M. 2010. Morphological diversity and biogeography of Procolophonids (Amniota: Parareptilia). *Journal of Systematic Palaeontology*, **8**, 607–625.

Ekman, S., Andersen, H. L. & Wedin, M. 2008. The limitations of ancestral state reconstruction and the evolution of the ascus in the lecanorales (lichenized ascomycota). *Systematic Biology*, **57**, 141–156.

Foote, M. 1994. Morphological disparity in Ordovician-Devonian crinoids and the early saturation of morphological space. *Paleobiology*, **20**, 320–344.

———. 1996. Ecological controls on the evolutionary recovery of post-Paleozoic Crinoids. *Science*, **274**, 1492–1495.

———. 1997. The evolution of morphological diversity. *Annual Review of Ecology and Systematics*, 129–152.

Foth, C., Brusatte, S. & Butler, R. 2012. Do different disparity proxies converge on a common signal? Insights from the cranial morphometrics and evolutionary history of Pterosauria (Diapsida: Archosauria). *Journal of evolutionary biology*, **25**, 904–915.

Friedman, M. 2010. Explosive morphological diversification of spiny-finned teleost fishes in the aftermath of the end-Cretaceous extinction. *Proceedings of the Royal Society B: Biological Sciences*, **277**, 1675–1683.

Garland, J., Theodore & Ives, A. R. 2000. Using the past to predict the present: Confidence intervals for regression equations in phylogenetic comparative methods. *The American Naturalist*, **155**, 346–364.

Gould, S. J. 1991. The disparity of the burgess shale arthropod fauna and the limits of cladistic analysis: why we must strive to quantify morphospace. *Paleobiology*, **17**, 411–423.

Gould, S. J. & Eldredge, N. 1977. Punctuated equilibria: The tempo and mode of evolution reconsidered. *Paleobiology*, **3**, 115–151.

Gower, J. C. 1966. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, **53**, 325–338.

———. 1971. A general coefficient of similarity and some of its properties. *Biometrics*, **27**, 857–871.

Guillerme, T. 2016. disprity: a package for measuring disparity in r.

Guillerme, T. & Cooper, N. 2018. *Appendices from: Time for a rethink: time sub-sampling methods in disparity-through-time analyses.* Dryad Digital Repository. https://doi.org/10.5061/dryad.vp4q518.

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

Hetherington, A. J., Sherratt, E., Ruta, M., Wilkinson, M., Deline, B. & Donoghue, P. C. 2015. Do cladistic and morphometric data capture common patterns of morphological disparity? *Palaeontology*, **58**, 393–399.

Hopkins, M. 2013. Decoupling of taxonomic diversity and morphological disparity during decline of the Cambrian trilobite family Pterocephaliidae. *Journal of Evolutionary Biology*, **26**, 1665–1676.

Huang, S., Roy, K. & Jablonski, D. 2015. Origins, bottlenecks, and present-day diversity: Patterns of morphospace occupation in marine bivalves. *Evolution*.

Hughes, M., Gerber, S. & Wills, M. A. 2013. Clades reach highest morphological disparity early in their evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **110**, 13875–13879.

Hunt, G. 2007. The relative importance of directional change, random walks, and stasis in the evolution of fossil lineages. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 18404–18408.

Hunt, G., Hopkins, M. J. & Lidgard, S. 2015. Simple versus complex models of trait evolution and stasis as a response to environmental change. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, 4885–4890.

Legendre, P. & Legendre, L. F. 2012. *Numerical ecology*, volume 24. Elsevier.

Lloyd, G. T. 2015. *Claddis: Measuring Morphological Diversity and Evolutionary Tempo.* R package version 0.1.

Paradis, E., Claude, J. & Strimmer, K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**, 289–290.

Prentice, K. C., Ruta, P. & Benton, M. J. 2011. Evolution of morphological disparity in Pterosaurs. *Journal of Systematic Palaeontology*, **9**, 337–353.

R Core Team. 2015. *R: a language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria.

Revell, L. J. 2012. phytools: An R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, **3**, 217–223.

Ruta, M., Angielczyk, K. D., Fröbisch, J. & Benton, M. J. 2013. Decoupling of morphological disparity and taxic diversity during the adaptive radiation of Anomodont Therapsids. *Proceedings of the Royal Society of London B: Biological Sciences*, **280**.

Slater, G. J., Harmon, L. J. & Alfaro, M. E. 2012. Integrating fossils with molecular phylogenies improves inference of trait evolution. *Evolution*, **66**, 3931–3944.

Thorne, P. M., Ruta, M. & Benton, M. J. 2011. Resetting the evolution of marine reptiles at the Triassic-Jurassic boundary. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 8339–8344.

Toljagic, O. & Butler, R. J. 2013. Triassic-Jurassic mass extinction as trigger for the Mesozoic radiation of Crocodylomorphs. *Biology Letters*, **9**.

Torgerson, W. S. 1965. Multidimensional scaling of similarity. *Psychometrika*, **30**, 379–393.

Wesley-Hunt, G. D. 2005. The morphological diversification of carnivores in North America. *Paleobiology*, **31**, 35–55.

Wills, M. A., Briggs, D. E. G. & Fortey, R. A. 1994. Disparity as an evolutionary index: A comparison of Cambrian and recent arthropods. *Paleobiology*, **20**, 93–130.

Wright, D. 2017*a*. Data from: Bayesian estimation of fossil phylogenies and the evolution of early to middle paleozoic crinoids (echinodermata).

Wright, D. F. 2017*b*. Bayesian estimation of fossil phylogenies and the evolution of early to middle paleozoic crinoids (echinodermata). *Journal of Paleontology*, 1–16.

Yang, Z., Kumar, S. & Nei, M. 1996. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics*, **141**, 1641–50.