Received: 3 January 2018 | Accepted: 8 March 2018

DOI: 10.1111/2041-210X.13006

RESEARCH ARTICLE

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

A global envelope test to detect non-random bursts of trait evolution

David J. Murrell[1,2]

[1]Department of Genetics, Evolution and Environment, University College London, London, UK

[2]Centre for Biodiversity and Environment Research, University College London, London, UK

Correspondence
David J. Murrell
Email: d.murrell@ucl.ac.uk

Funding information
Engineering and Physical Sciences Research Council, Grant/Award Number: EP/N007336/1

Handling Editor: William Pearse

Abstract

1. The joint analysis of species' evolutionary relatedness and their morphological evolution has offered much promise in understanding the processes that underpin the generation of biological diversity.
2. Disparity through time (DTT) is a popular method that estimates the relative trait disparity within and between subclades, and compares this to the null hypothesis that trait values follow Brownian evolution along the time-calibrated phylogenetic tree. To visualise the differences a confidence envelope is normally created by calculating, at every time point, the 97.5% minimum and 97.5% maximum disparity values from multiple simulations of the null model. The null hypothesis is rejected whenever the empirical DTT curve falls outside of this envelope, and these time periods may then be linked to events that may have sparked non-random trait evolution.
3. Here, simulated data are used to show this pointwise (ranking at each time point) method of envelope construction suffers from multiple testing and a poor, uncontrolled, false-positive rate. As a consequence it cannot be recommended. Instead, each DTT curve can be given a single rank based upon their most extreme disparity value, relative to all other curves, and across all time points. Ordering curves this way leads to a test that avoids multiple testing, but still allows construction of a confidence envelope. The null hypothesis is rejected if the empirical DTT curve is ranked within the most extreme 5% ranked curves from the null model. Comparison of the rank envelope curve to the Morphological Disparity Index and Node Height tests shows it to have generally higher power to detect non-Brownian trait evolution. An extension to allow simultaneous testing over multiple traits is also detailed.
4. Overall the results suggest the new rank envelope test should be used in null model testing for DTT analyses. The rank envelope method can easily be adapted into recently developed posterior predictive simulation methods used in model selection analyses. More generally, the rank envelope test should be adopted whenever a null model produces a vector of correlated values and the user wants to determine where the empirical data are different to the null model.

KEYWORDS
adaptive radiation, comparative methods, disparity through time, macroevolution, morphological diversification, null model, phenotypic diversification

1 | INTRODUCTION

Understanding the joint temporal dynamics of taxonomic and phenotypic diversity can provide tremendous insights into evolutionary success and its relationship with ecological opportunity, selective pressures, constraints, biotic interactions and environmental conditions. At the most basic level evolutionary biologists are often interested in detecting non-random evolution of biological traits within and across clades of species. Non-random bursts in evolution are often thought to be associated with events that open up ecological opportunities and enable a rapid increase in speciation rates and trait evolution, followed by slowdown in both processes as the ecological niches become filled. The evolutionary theory of adaptive radiation is the special case where the burst in speciation rate and trait evolution occur early in the clade's history (Schluter, 2000), but such bursts in trait evolution may occur at other times and can be triggered by other processes such as major events in the external environment.

A variety of null model, and model selection methods exist to look for the signature of evolutionary bursts in trait evolution (Freckleton & Harvey, 2006; Harmon, Schulte, Larson, & Losos, 2003; Harmon et al., 2010; Slater & Pennell, 2014; Slater, Price, Santini, & Alfaro, 2010). The model selection approach takes a variety of candidate models (Brownian evolution, early burst, selective peak) and fits these to the data using maximum likelihood methods before choosing the model that has the "best fit" (Harmon et al., 2010; Slater & Pennell, 2014). The null model approach remains more popular, partly because the methods have been established for longer, and the overall aim is to investigate if the data can be distinguished from the null model of Brownian evolution of trait values (Freckleton & Harvey, 2006; Harmon et al., 2003).

One of the commonly used null model approaches is to look at morphological traits to see if trait disparity increases, decreases or stays the same as species accumulate in evolutionary time, and also see whether this disparity is greater within or between clades. Convergent evolution of traits is implied if morphological disparity is predominantly found within one or more subclades; whereas adaptive radiations are expected to show divergence of traits between subclades, and in this scenario between clade morphological disparity should be greater than within subclade disparity. This analysis of between and within clade trait disparity has been championed by the disparity through time (DTT) approach introduced by Harmon et al. (2003). Here the empirical DTT curve is compared to the distribution of DTT curves generated on the same phylogenetic tree but under a specific model of how the trait diversity evolves. Generally the null model is an uncorrelated random walk, also referred to as Brownian evolution (i.e. a Brownian random walk over time in trait space). The method of comparison is critical in determining whether the empirical data can be distinguished from the null model. Early analyses used an integral deviation method called the Morphological Disparity Index (MDI) which sums the deviations of the empirical DTT curve from the median of the null model simulations (Harmon et al., 2003). The index can then be compared to the distribution of values produced by the simulation to test whether it is significantly different from the null model (Slater et al., 2010). Where MDI > 0, this implies within-clade trait variation is generally greater than expected under the null model, and MDI < 0 implies between-clade trait variation is more dominant than expected under the null model, and is suggestive of an adaptive radiation.

Since the MDI produces a number, it does not indicate the time periods when the empirical DTT curve deviates from the null model. Visualisation of when any non-random bursts might have occurred (e.g. early on in the radiation) can only proceed by plotting the empirical DTT curve against the DTT curves sampled from the null model. However, determining where *statistically significant* local deviations from the null model are occurring in the time series requires another test. The current go-to method is to simulate the null model n times (typically $n > 1{,}000$) and then construct a $(100 - 2\alpha)$ confidence interval by excluding the α largest, and α smallest relative disparity values at each time point. This method is also referred to as the *pointwise envelope method* because the ordering of the curves occurs at each time point (Myllymaki, Mrkvicka, Grabarnik, Seijo, & Hahn, 2017). The observed relative DTT curve can then be compared to this envelope and if it falls outside the null model is said to be rejected at the 2α level of significance.

The pointwise envelope method continues to be a popular method of inference, often as a diagnostic test in conjunction with the MDI test (e.g. Arbour & Lopez-Fernandez, 2016; Aristide et al., 2016; Blackburn et al., 2013; Dornburg et al., 2011; Feilich, 2016; Hlusko, Schmitt, Monson, Brasil, & Mahaney, 2016; Ingram, 2015; Johnson & Omland, 2004; Slater et al., 2010; Weber, Mitko, Eltz, & Ramirez, 2016). The visual/graphical interpretation of the DTT curve with an envelope test has extra appeal as it can be used to identify time points where the burst of non-Brownian evolution occurred, enabling correlation with known evolutionary or environmental events that have triggered the burst. However, the pointwise envelope method leads to weaker than expected statistical performance because multiple tests, one at each time point, are being performed simultaneously. Multiple testing leads to an increased type 1 statistical error rate (an elevated rate of rejection of the null hypothesis when it is true) that is no longer in line with the significance level being used to generate the confidence intervals of the envelope. Although multiple testing problems may be solved using a Bonferroni correction, it is not appropriate here because the assumption of independence of tests is violated by the correlation of disparity values between consecutive time points and also the (often) large number of time points being simultaneously evaluated (Loosmore & Ford, 2006).

The continued use of the pointwise envelope suggests its graphical interpretation is very appealing and it would therefore be worthwhile to circumvent its multiple testing issues. The recent development of the rank envelope test in spatial statistics (Myllymaki et al., 2017) holds promise to be useful in DTT analyses. Spatial analysis of ecological data often leads to the use of a nonparametric summary statistic such as Ripley's K that plots the tendency to cluster against the radial distance (e.g. Law et al., 2009; Flügge, Olhede, & Murrell, 2012), and the problem of pointwise envelopes for inference of non-random patterns is well established (Baddeley et al., 2014; Loosmore & Ford, 2006). As will be detailed below, the rank envelope test assigns each curve a single rank based on its most extreme deviation from the median curve from the null model simulation curves, and standard significance testing can then proceed by investigating if the empirical curve is found within the most extreme ranked curves. As shown by Myllymaki et al. (2017), the rank envelope method has good type 1 and type 2 error rates and is recommended for testing point pattern data against the null model of complete spatial randomness. The rank envelope can be developed and applied to any model that produces a vector (e.g. van Veen & Murrell, 2005), however, its performance needs to be tested since there are many ways of ordering curves based on how extreme they are compared to the null model and not all methods will produce desirable results.

In what follows, the rank envelope test will be developed for DTT null model analyses and its type 1 and type 2 statistical properties compared to the pointwise envelope, MDI and node height tests. The pointwise envelope test will be shown to have extremely poor type 1 error rates and should not be used for inference even as a diagnostic tool in conjunction with the MDI test. In contrast, the rank envelope method will be shown to possess desirable type 1 error rates, and the best overall power to detect accelerating or decelerating rates of trait evolution.

2 | MATERIALS AND METHODS

2.1 | Data simulation

Phylogenetic trees were generated within R (version 3.3.3) by implementing the pure birth (Yule) model using the *pbtree* function from the *phytools* library (version 0.6; Revell, 2012), and are rescaled so they run between 0 and 1 time units. These phylogenetic trees were then used to simulate quantitative trait evolution under a variety of scenarios including the null model of Brownian evolution. Specifically, trait evolution was simulated using the *fastBM* (in *phytools*) and *rescale* (in *geiger*, version 2.0.6, Pennell et al., 2014) functions. The *rescale* function allows the simulation of the accelerating-decelerating (ACDC) trait evolution model (Blomberg, Garland, & Ives, 2003) via an exponential rate change parameter, a. The null model of Brownian evolution is simulated when $a = 0$. When $a < 0$ the rate of evolution decelerates with time, and evolution accelerates over the phylogenetic tree when $a > 0$. The magnitude of a determines how quickly this burst of activity fades away or builds up, with large magnitudes delivering a rapid decay or late increase in evolutionary change (examples are given in Figure S1). However, as has been shown by Uyeda, Caetano, and Pennell (2015), assuming the phyologenetic tree is ultrametric, the ACDC model with $r > 0$ generates traits with a structure equivalent to those produced by a single optimum Ornstein–Uhlenbeck (OU) model. It is therefore not possible to distinguish between the OU and decelerating trait evolution models using the methods described here. Following Slater and Pennell (2014) results below are reported in terms of evolution half-life or doubling times, with half-life (doubling time) describing how much time is required for the trait evolution rate to fall to half (double) its initial value. The half-life or doubling time is defined as:

$$t_{1/2} = \frac{\log(2)}{a}$$

so half-life is defined when $a < 0$ and doubling time defined when $a > 0$.

2.2 | DTT analyses

DTT has proven to be one of the more popular approaches and uses the average pairwise Euclidean distance between species trait values as a measure of disparity. Following Harmon et al. (2003) relative disparity is calculated by dividing the disparity of each subclade by the disparity of the whole tree. At each time point (speciation event) the average relative disparity for that time point is calculated as the mean of the relative disparities for all subclades whose ancestral lineages are present at that time. Relative disparity values close to zero indicate that variation in the trait(s) is predominantly partitioned between subclades rather than within them. Relative disparity values larger than unity suggest that a clade contains a large amount of that variation, and that clades may overlap in trait space. By definition, disparity is 1 at the base of the phylogenetic tree, but is 0 at the present day.

Two methods that are currently used to search for the signal of bursts in morphological evolution using DTT are (1) the pointwise envelope test, and (2) an integral deviation test known as the MDI. As well as these a third test, the rank envelope test, was investigated. All make comparisons of the empirical DTT to the DTT taken from the ensemble of simulations generated by the null model of Brownian evolution, and all use the same measure of disparity defined above.

2.3 | The pointwise envelope test

The pointwise envelope test is a Monte Carlo simulation method that aims to produce a confidence interval, or envelope within which any part of the empirical DTT curve is said to be statistically indistinguishable from the null model. The method currently implemented in *geiger* (v2.0.6) constructs a (100 − 2α)% confidence interval by excluding at each time point the α largest, and α smallest relative disparity values across the entire ensemble of DTT curves simulated under the null model. Normally α = 2.5. More formally, the envelope is defined by the lower and upper bounding curves

$$T_{\mathrm{low}}^{(k)}(t) = \min\nolimits_{i=1,2,\ldots,s}^{k} T_i(t) \tag{1a}$$

$$T_{\mathrm{upp}}^{(k)}(t) = \max\nolimits_{i=1,2,\ldots,s}^{k} T_i(t), \tag{1b}$$

where $\min^k$ and $\max^k$ denotes the kth smallest and largest values of the DTT across all simulations s, of the null model at time (speciation event) t. If the empirical DTT curve falls outside of this envelope it is interpreted as being evidence for a departure from the null model of Brownian evolution at the 2α level of significance. The key message for the reader is that the confidence interval is determined by ordering the null model curves at each time point (hence the name *pointwise envelope*), and that each ordering is done independently of all other time points.

2.4 | The MDI test

Perhaps the simplest way to avoid multiple testing is to perform a deviation test that sums the deviations of the empirical DTT from the median DTT of the ensemble of null model simulations. Known as the MDI, negative values indicate the empirical DTT curve is below the null model median DTT for at least some of the range of time points, again pointing to the possibility of an early burst in diversity (Harmon et al., 2003). The test used by Slater et al. (2010) to assess statistical significance of a negative MDI is based on computing the proportion of cases in which the MDI for all null model simulated curves and the empirical curve were greater than 0. This is currently implemented using the *dtt* function in the *geiger* R library, but the user should note that the current version of *geiger* (v2.0.6) has an error which leads to inaccurate *p*-values (G. Slater, personal communication, October 16, 2017), and the results below use updated source code from https://github.com/mwpennell/geiger-v2/blob/master/R/disparity.R. The test as implemented in *geiger* is a one-tailed test to look for early bursts (i.e. the alternative hypothesis is that the empirical MDI is less than the null model expectation), but two-tailed tests are straightforward to implement with the same approach.

A disadvantage of this approach is that it is not possible to pinpoint the time periods when the empirical DTT deviates from the null model without plotting it against the null model simulations and then performing some sort of envelope test. Moreover, since the index sums up the deviations from the median of the ensemble of simulations of the null model, it is theoretically possible for time periods where the empirical DTT is above the median DTT to be cancelled out by time periods where it is below the median DTT, thus giving an MDI value close to that expected under Brownian evolution. However, as will be confirmed the MDI test has good statistical test properties and has been well used (e.g. Colombo, Damerau, Hanel, Salzburger, & Matschiner, 2015; Harmon et al., 2003; Ingram, 2015; Jonsson, Lessard, & Ricklefs, 2015; Slater et al., 2010).

2.5 | Rank envelope test

In this method each curve is given one ranking that summarises how extreme it is compared to all other curves. The more formal underpinnings of the test can be found in Myllymaki et al. (2017), but the process is quite straightforward. Each curve is given a single rank using the following steps:

1. Rank each DTT curve so
   i. $T_i^{\mathrm{asc}}(t)$ is the rank in ascending order of the disparity value of curve i at time t, against all other curves at time t. Higher ranks denote larger disparity values relative to all other curves (at time t). For example if there are s curves to be ranked, then the curve with the largest relative disparity value at time t is given the highest rank, that is $T_i^{\mathrm{asc}}(t) = s$.
   ii. $T_i^{\mathrm{des}}(t)$ is the rank in descending order of the disparity value of curve i at time t, against all other curves at time t. Higher ranks denote smaller relative disparity values at time t compared to all other curves. For example if there are s curves to be ranked, then the curve with the smallest disparity value at time t is given the highest rank, that is $T_i^{\mathrm{des}}(t) = s$.

   Since all curves are constrained to have disparity value of 1 at time 0, and disparity value 0 at the last time point, these are ignored in the rankings.

2. Obtain the global rank R_i for each curve by taking the highest ranking it has across all time points and across both ascending and descending sets:

$$R_i = \max\{T_i^{\mathrm{asc}}(t), T_i^{\mathrm{des}}(t), \text{for all } t\}.$$

   R_i is therefore a measure of how extreme the curve is compared to the rest of the curves, and is equivalent to ranking the curve according to its maximum deviation from the median DTT curve. An illustrative curve ranking is given in Figure 1.

3. The empirical DTT curve is then given its global rank, R_1 in exactly the same manner as for all other curves. Although the steps above describe the process for a two-tailed test, a one-tailed test can easily be implemented using only T_i^{asc} or T_i^{des} to determine global rank.

2.5.1 | Generation of *p*-values

If the empirical DTT curve is ranked outside of the (100 − 2α)th quantile of globally ranked curves then the null hypothesis can be rejected and the observed DTT curve can be said to be lower or

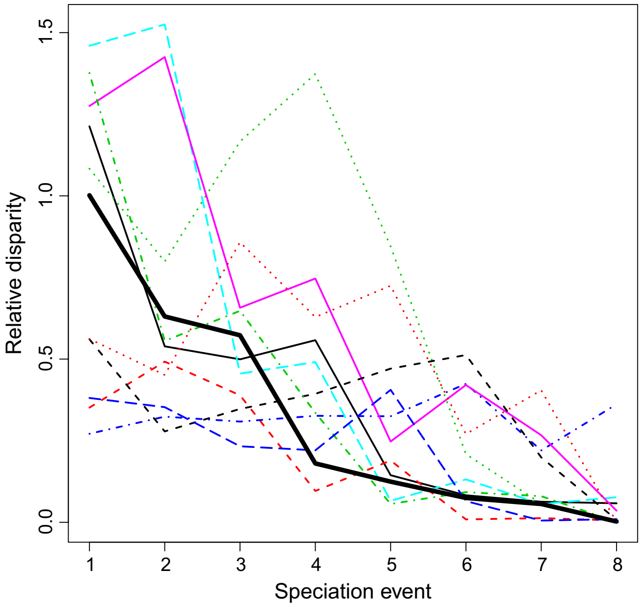

FIGURE 1 An example of the curve ranking process for the rank envelope test. Here there are 11 disparity through time curves and if we focus on the thick solid black line we see that in ascending order of disparity values it has the following rankings across the eight time points, $T^{\mathrm{asc}}_{\mathrm{black}} = \{6, 8, 7, 2, 3, 3, 4, 4\}$; that is at the first time point it has the sixth largest relative disparity value across all curves; at the second speciation event it has the eighth largest relative disparity value, etc. The corresponding rankings for this curve across the speciation events in the descending order of disparity values are, $T^{\mathrm{des}}_{\mathrm{black}} = \{6, 4, 5, 10, 9, 9, 8, 8\}$. The global ranking for this curve is the maximum ranking it takes in both sets, so in this case $R_{\mathrm{black}} = 10$

greater than expected under the null hypothesis, as appropriate. However, complications arise because ties in ranking are inevitable. For example the same curve cannot have both the highest and lowest disparity values across the whole ensemble of curves at time t, unless all curves have the same disparity value at that speciation event. In the two-tailed test there will almost always be at least two curves that take on the highest ranking s, at each time point. As such, the set of curves can only be weakly ordered. In order to generate a p-value a method for dealing with the ties needs to be used. Although one option is to use the mid-point to break the ties, following Myllymaki et al. (2017), a range of p-values is reported that encompasses the most liberal and most conservative p-values, respectively, defined as

$$p_{-} = \frac{1}{s+1} \sum_{i=1}^{s+1} 1(R_i < R_1) \tag{2a}$$

$$p_{+} = \frac{1}{s+1} \sum_{i=1}^{s+1} 1(R_i \leq R_1), \tag{2b}$$

where 1(…) is the indicator function that takes a value of 1 if the inequality is true, and takes value 0 otherwise; and where R_1 is the rank of the empirical DTT curve. This raises the problem that the interval defined by p_{-} and p_{+} could include the significance level α, leading to an ambiguous result. However, the likelihood of this happening is very small as long as s, the number of Monte Carlo simulations of the null model is sufficiently large. Myllymaki et al. (2017) recommend $s \geq 2,500$, and the results below use $s = 2,500$.

2.5.2 | Generation of confidence envelope for visualisation

From the above DTT curve ordering, it is straightforward to visualise the rank envelope determined by the significance level used. The upper and lower boundaries of the rank envelope are computed by taking the highest and lowest disparity values at each time point across the lowest (100 − 2α)th globally ranked curves. In other words, the rank envelope is the area bounded by the (100 − 2α)th globally ranked DTT curves simulated from the null model (more formally the confidence interval is the convex hull capturing the (100 − 2α) quantiles of the ranked curves). The user can then readily see where the empirical data falls outside of the rank envelope, and thus where the observed DTT is significantly different to Brownian evolution.

2.6 | The node height test

The final test does not use simulations of the null model to compare to the empirical data but instead relies upon the expectation that trait evolution should slow as niche space become packed. The node height test (Freckleton & Harvey, 2006) investigates if there is a significant correlation between the absolute magnitude of the standardised independent contrasts of the trait(s) and the height above the root of the node at which they were being compared to. The height of a node is defined as the absolute distance between the root and the most recent common ancestor of the pair from which the contrast is generated. A significant relationship between these indicates that the rate of trait evolution is changing systematically through the tree with early and late bursts in trait evolution being diagnosed by the sign of the slope. Graphical interpretation, and identification of key time periods of non-Brownian evolution is possible by looking to see which pairs of nodes are contributing to the non-zero slope. Since this is an established test, the analyses of the node height test were performed using the function *nh.test* within *geiger*.

3 | RESULTS

3.1 | False-positive rates (type 1 errors)

The false-positive rate was investigated by simulating an empirical dataset of trait evolution under the Brownian null model and testing how frequently each of the four tests described above incorrectly rejects the null hypothesis. Results for the DTT approach using the pointwise envelope test at the 5% level of significance show a disappointing, but unsurprising high rate of false positives (Figure 2). The multiple testing nature of this method means that the false-positive rate increases with the number of species in the comparison, and

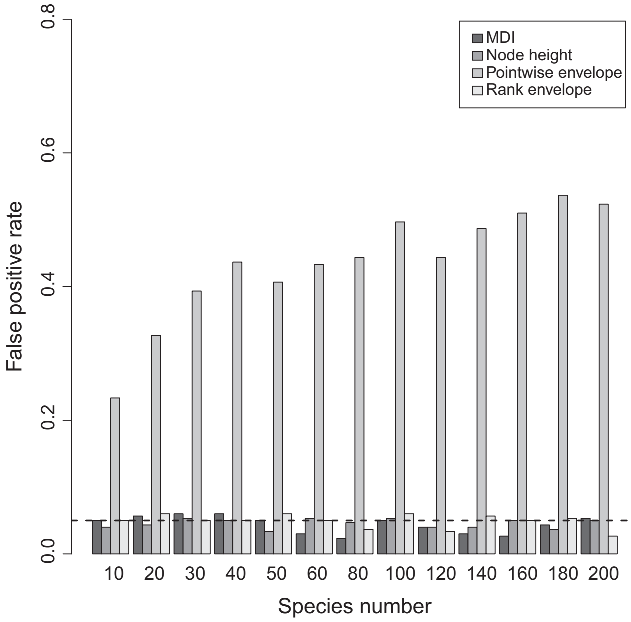

FIGURE 2 False-positive rates for the four tests for non-random disparity through time as a function of the number of species at the tips of the phylogenetic tree. False-positive rates are estimated from 300 simulated phylogenetic trees for each number of species using a pure birth model to generate the phylogenetic tree, and assuming Brownian evolution of the trait at each speciation event. All tests requiring Monte Carlo simulations were run with $s = 2,500$ trait evolution simulations. MDI, Morphological Disparity Index

in the simulations the rate of false positives ranges approximately between 0.25 and 0.5 for 10–200 species (Figure 2). That is to say, for comparisons using more than 100 species the pointwise envelope test is incorrectly rejecting the null hypothesis of Brownian evolution in *c*. 50% of cases. As such it is impossible to recommend this method for inference of non-Brownian bursts of trait evolution even as a diagnostic tool used with other tests that lack a visual interpretation. In comparison, all three other tests return consistent false-positive rates that hover around the significance level used (Figure 2).

3.2 | True-positive rates (type 2 errors)

Simulations for decelerating and accelerating rates of trait evolution confirm that the MDI, the node height, and the rank envelope tests can all successfully detect both early and late bursts in trait evolution (Figure 3), and the power of each is (unsurprisingly) positively related to the number of species and the strength of the early or late burst. However, other generalities do emerge. First, decelerating rates of trait evolution are easier to detect than accelerating rates for all tests. Second, the rank envelope generally shows the highest power to detect non-random trait evolution (see Figure S2). The MDI test appears to work best (relative to other tests) for very small phylogenies in early burst settings, but the increase in power with phylogeny size lags behind the node height and global envelope tests. In contrast, the node height test is generally intermediate to the other tests, but has similar power to the rank envelope tests for large phylogenies in the early burst model, and also for small phylogenies in the late burst scenario (Figures 3 and S2). The tests detailed here are all two-tailed, but the pattern remains unchanged when one-tailed tests are used instead (Figure S3).

3.3 | Data examples

Having established the rank envelope test possesses desirable type 1 and type 2 statistical error properties, three datasets were used to illustrate how inference of the rates of morphological evolution can change depending on whether the pointwise or global envelope test is used. Since the pointwise envelope test is too liberal in its rejection of the null model the expectation should be for a reduction in support for non-Brownian bursts in morphological evolution. All tests are two-tailed.

The first example uses the morphological and phylogenetic data on Darwin's finches (Geospiza) which is currently found in the *geiger* (version 2.0.6) R package. Re-analysis shows support for two time periods where the empirical DTT curve for culmen length sits above the pointwise envelope, consistent with both the accelerating and Orstein–Uhlenbeck models of trait evolution (Figure 4a). In contrast, there is no departure from the null model of Brownian evolution according to the rank envelope test (Figure 4b).

The second example uses a time-calibrated molecular phylogeny of extant cetaceans and a morphological dataset on body size from Slater et al. (2010) which is also available within *geiger* (version 2.0.6). The pointwise envelope test result would suggest an early burst in evolution of body size and that this occurred predominantly during the period 6–11 Ma, but that the rank envelope test fails to find any departure from the null model of Brownian evolution at the 5% level of statistical significance (Figure 4c,d). As noted by Slater et al. (2010), the MDI and node height tests also fail to find evidence for non-Brownian evolution in whale body length.

The final example is taken from (Feilich, 2016) who investigated the evolution of body shape, caudal fin shape, dorsal fin shape and anal fin shape in African cichlid fishes. Re-analysing the data for anal fin shape using the pointwise envelope (Figure 4e) confirms the spike in relative disparity coinciding with the Cichlinae–Pseudocrenilabrinae split 45–75 Ma reported in the original paper as well as the spike nearer to the present day that coincides with the haplochromine radiation (Feilich, 2016). In contrast, the rank envelope method finds no discernable departure (at the 5% level of significance) from the null model of Brownian evolution at any point in the evolutionary timeline (Figure 4f).

3.4 | Multiple traits

So far all tests and examples have considered just one trait of interest, but the rank envelope test can be readily extended to consider multiple traits. The first option, hereafter referred to as the multivariate disparity, is to compute a single DTT curve using all traits (or axes described by principle components) and then use the rank

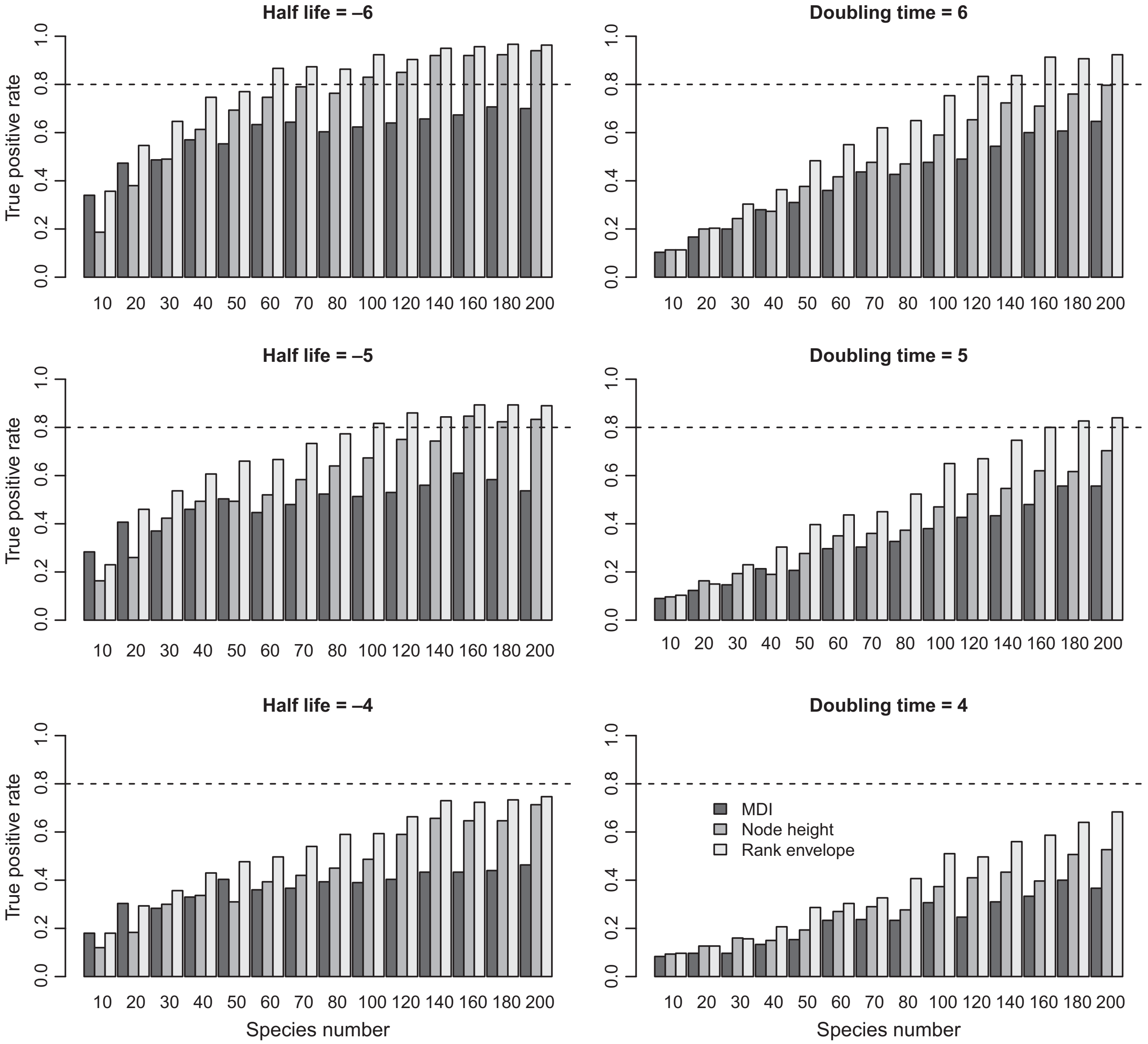

FIGURE 3 True-positive rates (statistical power) of the rank envelope test, the Morphological Disparity Index (MDI) test and the node height test under a range of simulated decelerating and accelerating evolution scenarios, and for a range of size of phylogenetic tree. Power is estimated from 300 simulated phylogenetic trees for each number of species using a pure birth model to generate the phylogenetic tree, and assuming trait evolution at each speciation event speeds ups or slow downs over evolutionary time. Null model tests on each tree were run with s = 2,500 trait evolution Monte Carlo simulations. Doubling time/half life is computed as a log 2, where a controls the time dependent change in rate of trait evolution. When a < 0, a decelerating rate in trait evolution occurs early in evolutionary time, and accelerating rate occurs when a > 0. Large magnitudes lead to the rate changes occurring over a smaller period of time. The dashed line is an arbitrary power threshold to allow easier comparison between parameter sets. Corresponding example plots of disparity through time in each scenario are given in Supporting Information

envelope test on that multivariate DTT curve. This is the extension of the disparity method to multiple traits, and involves computing the (normally Euclidean) distance in multiple dimensions by passing the multi-trait data to the *dtt* function in *geiger* R package. However, this approach potentially hides which of the traits are behind any non-random evolution, and there is also the possibility that computing a single DTT curve from all traits could lose vital information when some traits follow Brownian evolution (see below).

The user could instead perform a single (simultaneous) test across all traits by simply concatenating the DTT curves for all individual traits, that is for a two trait dataset the user would concatenate each pair of null model simulations of trait 1 and trait 2, and then rank the concatenated curves as before. The empirical DTT curves for each trait would also be concatenated in the same order. An example of this multivariate extension to the rank envelope method for two aspects of cichlid fish body shape is given in Figure S4, using data

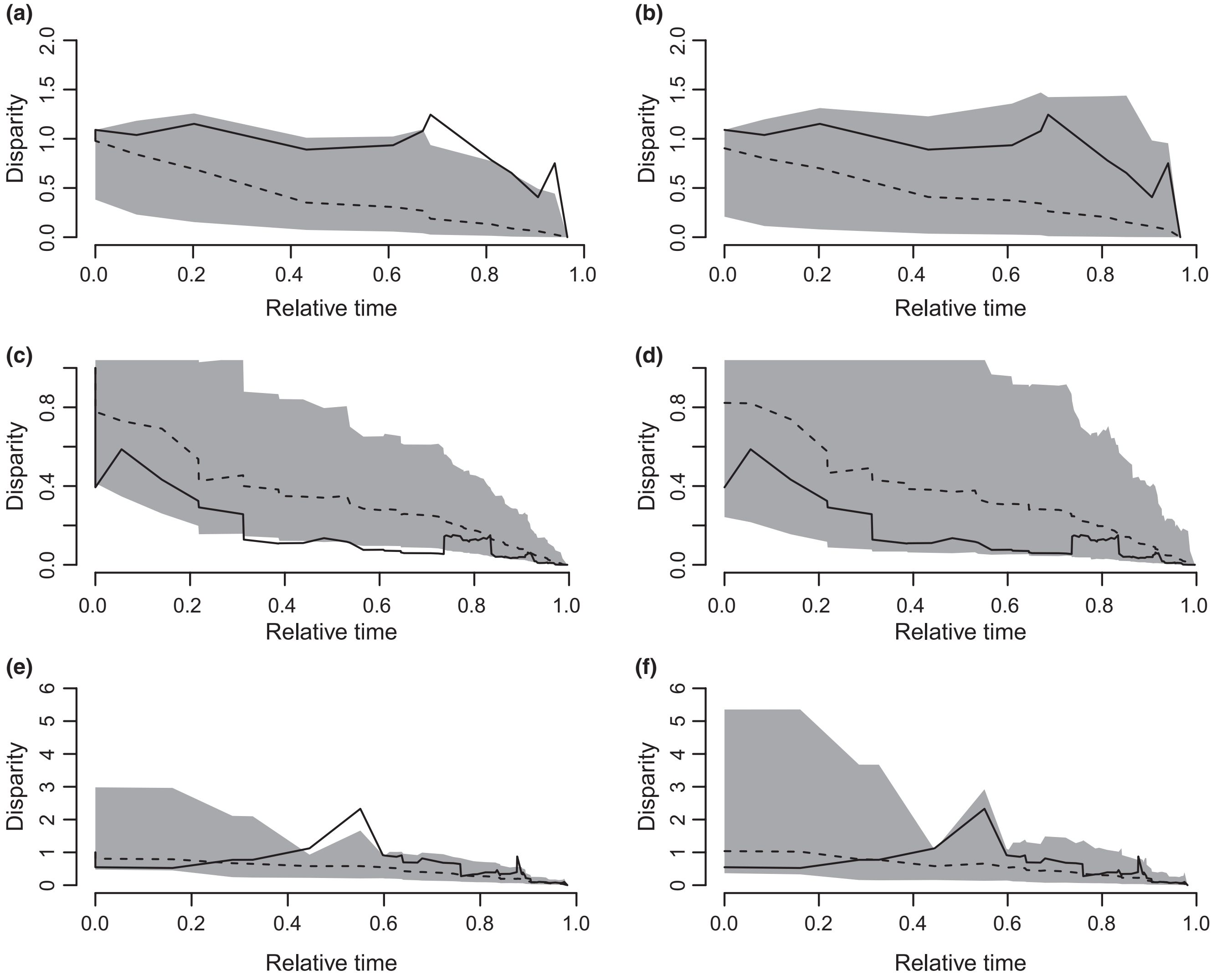

FIGURE 4 Comparisons of inference from using the pointwise envelope test (left hand column) and the rank envelope test (right hand column) for trait disparity through time for three showcase datasets. In each panel, the empirical pattern (solid black line) is compared to the median of 5,000 simulations of the null model of Brownian evolution (broken lines), and the shaded regions correspond to the 95% confidence intervals calculated using the pointwise (left column) and rank envelope (right column) methods. Top row (a,b) is for Darwin's finches (Geospiza) and the evolution of culmen length; middle row (c,d) is for the evolution of Cetacean body size (Slater et al., 2010); bottom row (e,f) is for the evolution of anal fin shape in African cichlids (Feilich, 2016)

taken from Feilich (2016). The null model is that the multiple traits follow independent Brownian evolution, and the key point is that there is one p-value for all traits.

Power analysis of the two options for investigating non-Brownian evolution of multiple traits shows that the rank envelope test using the multivariate DTT curve does well when both traits have the same rates of evolution (Figure S5a). However, power is lost as evolution of one of the traits becomes more like the null model (Figure S5b,c) because the signal from the strongly non-Brownian trait is lost in the averaging over the two traits. On the other hand, the concatenated rank envelope approach maintains higher power to detect deviations from the null model when there is greater difference between the evolutionary rates of the two traits. The reader should note that separate results confirm the concatenated rank envelope method retains desirable type 1 properties (results not shown), and in principle any number of traits could be considered.

4 | DISCUSSION

Envelope tests using the DTT pointwise envelope method continue to be a useful and popular way to pinpoint the time periods when trait evolution across a clade can be distinguished from being uncorrelated (e.g. Feilich, 2016; Hlusko et al., 2016; Slater et al., 2010). However, as shown here the current method of constructing the confidence envelope is prone to severe type 1 statistical errors due to multiple testing. An alternative method that ranks the DTT curves based upon their most extreme disparity value relative to the ensemble of null model DTT curves shows great promise to both avoid type 1 errors and also retain high power to detect true non-random time periods of trait evolution.

The pointwise envelope is often used in conjunction with the MDI (Harmon et al., 2003; Slater et al., 2010) or node height test (Freckleton & Harvey, 2006) and the introduced rank envelope test compares favourably to these alternatives. In general, all three tests show higher power to detect early bursts in trait evolution compared to the scenario where there is more trait convergence across clades (higher relative disparity) than under the null model (Figure 3). The node height test has the advantage of not requiring large numbers of simulations from the null model, and generally has greater power than the MDI test to detect non-Brownian trait evolution (Figure 3). However, the rank envelope test consistently outperforms both of these methods in all but the smallest of phylogenetic trees (Figure 3); is able to deal with multiple traits in a single test (Figure 4); and retains a simple visual interpretation that aids further inference of the processes that might be behind the bursts in trait evolution. On this basis, the rank envelope test can be recommended for the DTT approach to investigating trait evolution.

The methods investigated here all use the same hypothesis testing approach. That is to say we test our data against a suitable null model to see if there are detectable departures from the null model. A different approach is to consider a number of candidate models and ask which model best describes the data (Johnson & Omland, 2004). The advantage of this model selection approach is that multiple models are considered simultaneously, but of course there is no guarantee that the best model, usually determined by some information theoretic criterion, is a "good" descriptor of the data, and the method of model ranking is crucial to the outcome. Harmon et al. (2010) used maximum likelihood methods to fit models that could produce Brownian evolution, increasing or decreasing trait diversification rates, as well as selective peaks where the trait value has a tendency to return to a medial value. Using the likelihood ratio test, they found the Brownian evolution and the selective peak (OU) models to be the most frequently selected across 49 clades, implying early bursts in trait evolution are relatively rare. Slater and Pennell (2014) extended this method by employing a posterior predictive approach instead of the likelihood ratio test. The posterior predictive approach proceeds by fitting the parameters to the candidate models using maximum likelihood as in (Harmon et al., 2010), but model selection is based upon sampling the trait evolution from the fitted models and then comparing the fit of each model to the observed trait values. Slater and Pennell (2014) developed this method using either the MDI test, or the node height test and showed both of these posterior predictive methods can have a higher power to detect early bursts in trait evolution compared to the maximum likelihood ratio approach used in Harmon et al. (2010). Reanalysing the cetacean dataset with these methods led to the conclusion that an early burst model best described the evolution of whale body size (Slater & Pennell, 2014). This is not surprising given the rank envelope test clearly shows the empirical DTT curve is close to falling below the lower confidence interval (Figure 4d).

Ultimately, the user needs to choose between the null model testing and model selection methods. However, the rank envelope test developed here could easily be incorporated into the posterior predictive methods of Slater and Pennell (2014), since the ranking of the observed DTT curve in the ensemble of simulations from each of the candidate models generates a single metric, the global rank amongst the set of model curves, that could then be used to compare the models. However, for those who prefer null model testing, the rank envelope test appears to be a good starting point for investigating non-Brownian rates of trait evolution.

ACKNOWLEDGEMENTS

This work was supported by the Engineering and Physical Sciences Research Council grant EP/N007336/1. The author thank Alex Pigot, three anonymous referees, the handling editor and in particular Graham Slater, who all greatly improved earlier versions of this work.

DATA ACCESSIBILITY

R code to compute the rank envelopes and generate the results for all figures can be accessed via https://doi.org/10.5281/zenodo.1197535. The data used to compute the DTT analyses for cichlid fish body and fin morphology is taken from (Feilich, 2016) and can be accessed via https://datadryad.org//resource/doi:10.5061/dryad.h4k6f. All other data used are currently available within the R library *geiger* (Pennell et al., 2014), and can be accessed via downloading the library from https://CRAN.R-project.org/package=geiger.

ORCID

David J. Murrell http://orcid.org/0000-0002-4830-8966

REFERENCES

Arbour, J. H., & Lopez-Fernandez, H. (2016). Continental cichlid radiations: Functional diversity reveals the role of changing ecological opportunity in the Neotropics. *Proceedings of the Royal Society B-Biological Sciences*, *283*, 20160556. https://doi.org/10.1098/rspb.2016.0556

Aristide, L., dos Reis, S. F., Machado, A. C., Lima, I., Lopes, R. T., & Perez, S. I. (2016). Brain shape convergence in the adaptive radiation of New World monkeys. *Proceedings of the National Academy of Sciences of the United States of America*, *113*, 2158–2163. https://doi.org/10.1073/pnas.1514473113

Baddeley, A., Diggle, P. J., Hardegen, A., Lawrence, T., Milne, R. K., & Nair, G. (2014). On tests of spatial pattern based on simulation envelopes. *Ecological Monographs*, *84*, 477–489. https://doi.org/10.1890/13-2042.1

Blackburn, D. C., Siler, C. D., Diesmos, A. C., McGuire, J. A., Cannatella, D. C., & Brown, R. M. (2013). An adaptive radiation of frogs in a southeast Asian island archipelago. *Evolution*, *67*, 2631–2646. https://doi.org/10.1111/evo.12145

Blomberg, S. P., Garland, T., & Ives, A. R. (2003). Testing for phylogenetic signal in comparative data: Behavioral traits are more labile. *Evolution*, *57*, 717–745. https://doi.org/10.1111/j.0014-3820.2003.tb00285.x

Colombo, M., Damerau, M., Hanel, R., Salzburger, W., & Matschiner, M. (2015). Diversity and disparity through time in the adaptive radiation of Antarctic notothenioid fishes. *Journal of Evolutionary Biology*, *28*, 376–394. https://doi.org/10.1111/jeb.12570

Dornburg, A., Sidlauskas, B., Santini, F., Sorenson, L., Near, T. J., & Alfaro, M. E. (2011). The influence of an innovative locomotor strategy on the phenotypic diversification of triggerfish (family: Balistidae). *Evolution*, *65*, 1912–1926. https://doi.org/10.1111/j.1558-5646.2011.01275.x

Feilich, K. L. (2016). Correlated evolution of body and fin morphology in the cichlid fishes. *Evolution*, *70*, 2247–2267. https://doi.org/10.1111/evo.13021

Flügge, A. J., Olhede, S. C., & Murrell, D. J. (2012). The memory of spatial patterns: Changes in local abundance and aggregation in a tropical forest. *Ecology*, *93*, 1540–1549. https://doi.org/10.1890/11-1004.1

Freckleton, R. P., & Harvey, P. H. (2006). Detecting non-Brownian trait evolution in adaptive radiations. *PLOS Biology*, *4*, e373. https://doi.org/10.1371/journal.pbio.0040373

Harmon, L. J., Losos, J. B., Davies, T. J., Gillespie, R. G., Gittleman, J. L., Jennings, W. B., … Mooers, A. O. (2010). Early bursts of body size and shape evolution are rare in comparative data. *Evolution*, *64*, 2385–2396.

Harmon, L. J., Schulte, J. A., 2nd, Larson, A., & Losos, J. B. (2003). Tempo and mode of evolutionary radiation in iguanian lizards. *Science*, *301*, 961–964. https://doi.org/10.1126/science.1084786

Hlusko, L. J., Schmitt, C. A., Monson, T. A., Brasil, M. F., & Mahaney, M. C. (2016). The integration of quantitative genetics, paleontology, and neontology reveals genetic underpinnings of primate dental evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *113*, 9262–9267. https://doi.org/10.1073/pnas.1605901113

Ingram, T. (2015). Diversification of body shape in Sebastes rockfishes of the north-east Pacific. *Biological Journal of the Linnean Society*, *116*, 805–818. https://doi.org/10.1111/bij.12635

Johnson, J. B., & Omland, K. S. (2004). Model selection in ecology and evolution. *Trends in Ecology & Evolution*, *19*, 101–108. https://doi.org/10.1016/j.tree.2003.10.013

Jonsson, K. A., Lessard, J. P., & Ricklefs, R. E. (2015). The evolution of morphological diversity in continental assemblages of passerine birds. *Evolution*, *69*, 879–889. https://doi.org/10.1111/evo.12622

Law, R., Illian, J., Burslem, D. F., Gratzer, G., Gunatilleke, C., & Gunatilleke, I. (2009). Ecological information from spatial patterns of plants: Insights from point process theory. *Journal of Ecology*, *97*, 616–628. https://doi.org/10.1111/j.1365-2745.2009.01510.x

Loosmore, N. B., & Ford, E. D. (2006). Statistical inference using the G or K point pattern spatial statistics. *Ecology*, *87*, 1925–1931. https://doi.org/10.1890/0012-9658(2006)87[1925:SIUTGO]2.0.CO;2

Myllymaki, M., Mrkvicka, T., Grabarnik, P., Seijo, H., & Hahn, U. (2017). Global envelope tests for spatial processes. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, *79*, 381–404. https://doi.org/10.1111/rssb.12172

Pennell, M. W., Eastman, J. M., Slater, G. J., Brown, J. W., Uyeda, J. C., FitzJohn, R. G., … Harmon, L. J. (2014). geiger v2.0: An expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. *Bioinformatics*, *30*, 2216–2218. https://doi.org/10.1093/bioinformatics/btu181

Revell, L. J. (2012). phytools: An R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, *3*, 217–223. https://doi.org/10.1111/j.2041-210X.2011.00169.x

Schluter, D. (2000). *The ecology of adaptive radiation*. Oxford, UK: Oxford University Press.

Slater, G. J., & Pennell, M. W. (2014). Robust regression and posterior predictive simulation increase power to detect early bursts of trait evolution. *Systematic Biology*, *63*, 293–308. https://doi.org/10.1093/sysbio/syt066

Slater, G. J., Price, S. A., Santini, F., & Alfaro, M. E. (2010). Diversity versus disparity and the radiation of modern cetaceans. *Proceedings of the Royal Society B: Biological Sciences*, *277*, 3097–3104. https://doi.org/10.1098/rspb.2010.0408

Uyeda, J. C., Caetano, D. S., & Pennell, M. W. (2015). Comparative analysis of principal components can be misleading. *Systematic Biology*, *64*, 677–689. https://doi.org/10.1093/sysbio/syv019

van Veen, F., & Murrell, D. (2005). A simple explanation for universal scaling relations in food webs. *Ecology*, *86*, 3258–3263. https://doi.org/10.1890/05-0943

Weber, M. G., Mitko, L., Eltz, T., & Ramirez, S. R. (2016). Macroevolution of perfume signalling in orchid bees. *Ecology Letters*, *19*, 1314–1323. https://doi.org/10.1111/ele.12667

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Murrell DJ. A global envelope test to detect non-random bursts of trait evolution. *Methods Ecol Evol*. 2018;9:1739–1748. https://doi.org/10.1111/2041-210X.13006