

## Original Article

# Revealing patterns of homoplasy in discrete phylogenetic datasets with a cross-comparable index

Elizabeth M. Steell<sup>1,2,\*</sup>, Allison Y. Hsiang<sup>3</sup> and Daniel J. Field<sup>1,4</sup>

<sup>1</sup>Department of Earth Sciences, University of Cambridge, Downing Street, Cambridge, CB2 3EQ, United Kingdom

<sup>2</sup>Girton College, Huntingdon Road, Cambridge, CB3 0JG, United Kingdom

<sup>3</sup>Department of Geological Sciences, Stockholm University, 106 91 Stockholm, Sweden

<sup>4</sup>University Museum of Zoology, David Attenborough Building, University of Cambridge, Downing Place, Cambridge, CB2 3EJ, United Kingdom

\*Corresponding author. Department of Earth Sciences, University of Cambridge, Downing Street, Cambridge, CB2 3EQ, United Kingdom. E-mail: [ems207@cam.ac.uk](mailto:ems207@cam.ac.uk)

## ABSTRACT

Investigating patterns of homoplasy can improve our understanding of macroevolutionary processes by revealing evolutionary constraints on morphology and highlighting convergent form–function relationships. Here, we test the performance of several widely-used methods that provide measures of homoplasy, including the consistency (CI) and retention indices (RI), using simulated and empirical discrete morphological datasets. In addition, we describe and test a new method employing a novel randomization protocol, which we term the relative homoplasy index (RHI). RHI outperforms other methods in a range of situations for measuring relative homoplasy and allows comparisons between different datasets. In line with some previous work, we show that relative levels of homoplasy remain constant with the addition of characters and decrease with the addition of taxa. We also show that the extent of homoplasy strongly influences the distribution of taxa in morphospace. Low homoplasy results in highly partitioned morphospace, while high homoplasy leads to clades overlapping in morphospace. Our results help illuminate the properties of relative homoplasy in morphological phylogenetic matrices, opening new potential avenues for research on homoplasy quantification in macroevolutionary studies.

**Keywords:** homoplasy; discrete characters; phylogenetics; morphology; parsimony

## INTRODUCTION

Homoplasy is an almost universal occurrence in discrete character matrices constructed for phylogenetic inference. The term homoplasy, introduced by [Lankester \(1870\)](#), refers to non-homologous similarity. In discrete character matrices, homoplasy describes the independent acquisition of or reversal to the same character state, and is often considered problematic as it can obscure phylogenetic relationships ([Kluge and Farris 1969](#), [Farris 1989, 1991](#), [Sanderson and Donoghue 1989](#), [Goloboff 1991a, b](#), [Huelsenbeck 1991](#), [Klassen \*et al.\* 1991](#), [Wilkinson 1991](#), [Hillis and Huelsenbeck 1992](#), [Archie 1996](#), [Davis \*et al.\* 1998](#)). However, with ever-improving methods to infer phylogenetic relationships and the increasing availability of genomic data for extant clades, accurately quantifying the extent of homoplasy in phylogenetic datasets may be useful for informing estimates of evolvability and constraints on morphological evolution ([Sanderson and Donoghue 1989](#), [Wagner 2000](#), [Sidlauskas 2008](#), [Brocklehurst and Benson 2021](#), [Brocklehurst \*et al.\* 2021](#)).

Methods to estimate homoplasy received considerable attention in the late 20<sup>th</sup> century with an extensive body of research exploring parsimony-based methods with cladistic matrices [summarized in [Archie \(1996\)](#)]. In recent years, however, few studies have explored patterns of homoplasy in discrete datasets, and the development of new homoplasy estimation methods have lagged behind methodological advances in phylogenetic inference. Although their original functions were not aimed at estimating homoplasy, two of the most widely implemented and easily calculated indices are the parsimony-based consistency index (CI) ([Kluge and Farris 1969](#)) and the retention index (RI) ([Farris 1989](#)). These indices can be calculated in phylogenetic inference programmes such as TNT ([Goloboff \*et al.\* 2008](#)) and the *phangorn* R package ([Schliep 2011](#)), and have been the focus of numerous comparative analyses on their efficacy for detecting homoplasy, and used to explore patterns of homoplasy within discrete character matrices (e.g. [Archie 1989, 1996](#), [Sanderson and Donoghue 1989](#), [Goloboff 1991b](#), [Klassen \*et al.\* 1991](#), [Meier](#)

*et al.* 1991, Naylor and Kraus 1995, Hauser and Boyajian 1997, Sookias 2020, Murphy *et al.* 2021).

However, CI and RI are not cross-comparable methods; that is, they cannot be directly compared across alternative datasets and phylogenetic trees, limiting attempts to infer generalizable conclusions about patterns of homoplasy in phylogenetic datasets (Archie 1989, Klassen *et al.* 1991, Meier *et al.* 1991). For instance, CI is highly sensitive to variable numbers of taxa (but not characters) in phylogenetic matrices, with a negative relationship detected between CI and the number of taxa included in a dataset (Archie 1989, Sanderson and Donoghue 1989, Hauser and Boyajian 1997, Murphy *et al.* 2021). By contrast, RI represents a more robust measure with respect to matrix dimensions, although it is nonetheless sensitive to taxon number (Archie 1989, Naylor and Kraus 1995, Hauser and Boyajian 1997, Murphy *et al.* 2021), and represents a measure of relative homoplasy as opposed to absolute homoplasy, like CI (see below for details) (Farris 1989, Goloboff 1991b). Nonetheless, for the majority of cases, values of RI are not scaled between zero and one if nodes within a tree are partially or fully resolved; instead, the observed minimum value (which indicates the maximum amount of homoplasy possible) is an arbitrary value somewhere within that range, an artefact arising from the theoretical maximum tree length ( $L_{max}$ ; Fig. 1; Table 1) used when calculating RI (Archie 1989, 1996, Farris 1989, Goloboff 1991b).

To mitigate this scaling issue, additional methods incorporating a randomization step such as the homoplasy excess ratio (HER) (Archie 1989) and the homoplasy slope ratio (HSR) (Meier *et al.* 1991) were developed to measure relative homoplasy that can be compared across datasets. However, these methods are limited due to long computation times, and by the fact that the method of estimating the maximum value of homoplasy complicates interpretations of relative homoplasy values with the use of particular datasets and topologies (Fig. 1; Table 1). Although HER and HSR have been applied in several studies (e.g. Wills 1999, Hughes 2013, Hoyal Cuthill 2015b, Oyston *et al.* 2015, 2022, Mendler *et al.* 2019, Murphy *et al.* 2021), they are generally challenging to implement and require several time-consuming steps to calculate.

Towards improving the rigour of systematic investigations of patterns of homoplasy, we present a method with a novel randomization protocol in the context of homoplasy indices, the relative homoplasy index (RHI), for estimating the amount of relative homoplasy in phylogenetic datasets compared to an estimated maximum value of homoplasy for the input tree. Thus, the homoplasy that RHI estimates is a property of a specific matrix and tree combination. Like HER and HSR, the algorithmic basis for RHI is similar to RI, and the resulting index is consistently scaled between zero and one for empirical datasets and associated trees, thus facilitating meaningful comparisons across datasets with differing numbers of taxa and characters. Compared to HSR and HER, the randomization method implemented in RHI ensures the null distribution of maximum homoplasy estimates is appropriate for any input matrix and tree combination that contains at least some phylogenetic signal with respect to each other. We test the performance of RHI against variation in numbers of taxa, characters, and character state transition rates. Our method is quick to implement in the

R programming environment (R Core Team 2021) and includes a range of functionalities such as per-character calculation, calculations for a distribution of trees, and incorporating specific ordered characters.

### Quantifying homoplasy

When measuring homoplasy, it is possible to measure either *absolute* homoplasy (e.g. by calculating number of extra steps) or *relative* homoplasy. Relative homoplasy is estimated by measuring the observed homoplasy and comparing it to an estimate of the maximum possible value of homoplasy, which is calculated in different ways depending on the index used. Out of the methods summarized below, CI is the only index that measures an absolute value of homoplasy (i.e. observed homoplasy that is not compared to an estimated maximum), whereas the other indices presented here measure relative homoplasy (except data decisiveness (DD), but see below).

#### Consistency and retention indices

The methodologies of the consistency (CI) and retention (RI) indices (Kluge and Farris 1969, Farris 1989) are summarized in the [Supporting Information](#), and their formulas and terms are summarized in Table 1. See Archie (1996) for further discussion.

CI and RI can be expressed with respect to each other:

$$CI = \frac{L_{min}}{L_{max} - RI(L_{max} - L_{min})}$$

$$RI = \frac{(CI * L_{max}) - L_{min}}{CI * (L_{max} - L_{min})}$$

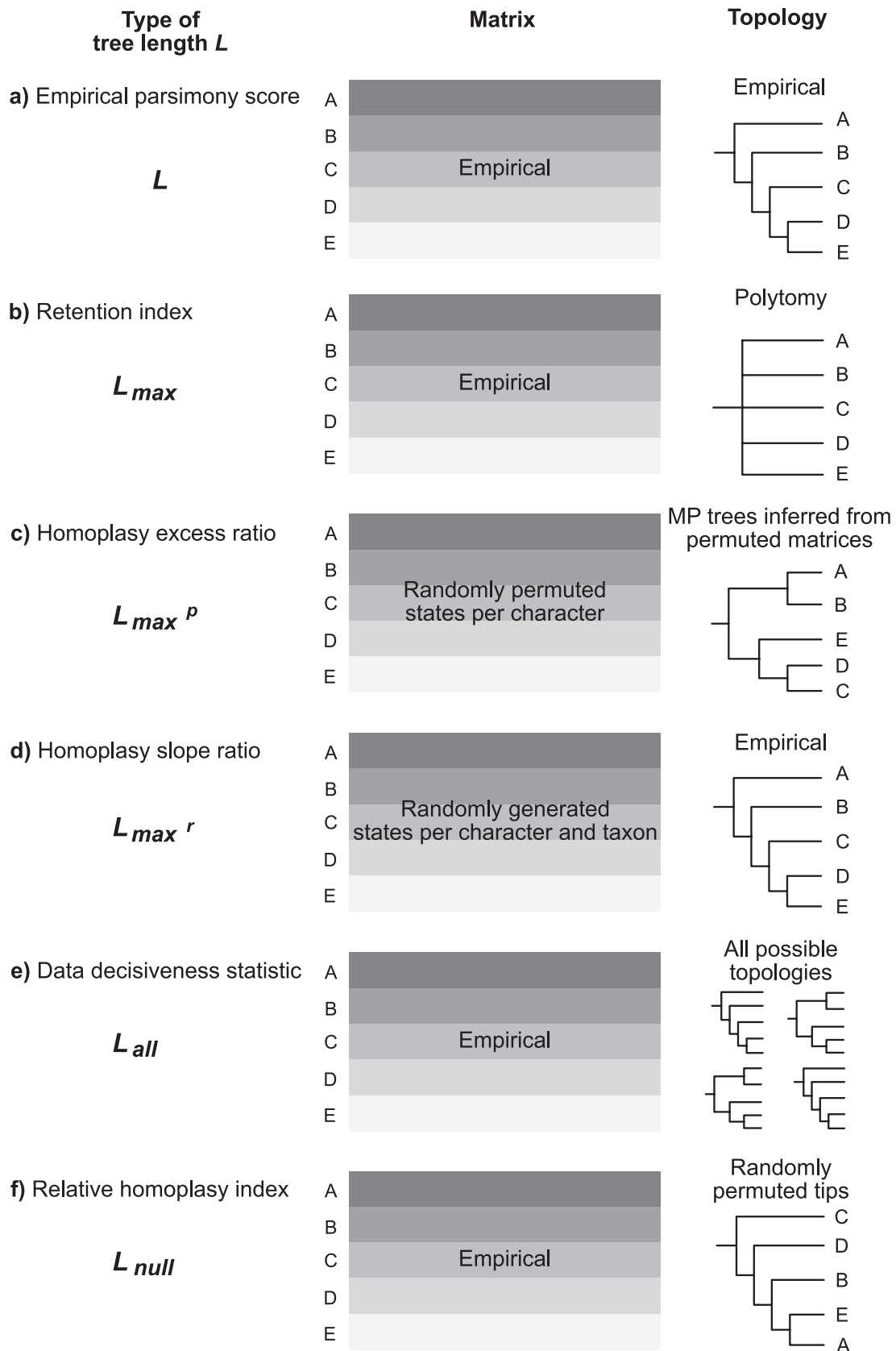
RI is equal to 0 when  $CI = L_{min}/L_{max}$ , i.e. when  $L = L_{max}$  (Farris 1989). It should be noted that the complement of CI is referred to as the homoplasy index in Kluge and Farris (1969), and the complement of RI is referred to as the distortion index in Farris (1989).

#### Homoplasy excess ratio

The formula format for the homoplasy excess ratio (HER) is identical to RI, but  $L_{max}$  is replaced by  $L_{max}^p$ , which represents the mean tree length of minimum length trees inferred through maximum parsimony for a set of  $n$  matrices that have been randomized through permutations of the character states within each character for all taxa (Archie 1989).

$$HER = 1 - \frac{L - L_{min}}{L_{max}^p - L_{min}} = \frac{L_{max}^p - L}{L_{max}^p - L_{min}}$$

A limitation of HER is that its value is difficult to interpret for non-minimum length empirical trees; for example, empirical trees that have been inferred through any other method than maximum parsimony, such as trees inferred using model-based inference methods or with forced topological constraints. If any empirical tree is not derived strictly from maximum parsimony, then the mean of the null distribution of tree lengths is not a viable estimate of maximum homoplasy due to the fact that the null trees are inferred through maximum parsimony themselves, and this may lead to HER values exceeding unity. Furthermore, HER is often impractical to calculate due to high computational demands at the phylogenetic inference step.



**Figure 1.** Summary of tree length calculations per method. Schematic to summarize randomization protocols showing how tree length ( $L$ ) and its variations ( $L_{max}$ ,  $L_{max}^p$ ,  $L_{max}^r$ ,  $L_{all}$ , and  $L_{null}$ ) are calculated with respect to input matrices and topologies. MP refers to maximum parsimony inference.  $L_{max}^p$ ,  $L_{max}^r$ , and  $L_{null}$  are each calculated for a distribution of matrices and/or trees (see Table 1).

Table 1. Summary of homoplasy indices.

Method	Formula	Terms	Typical range of values	Relationship to homoplasy	Notes	Reference
Consistency index	$CI = \frac{L_{min}}{L}$	$L_{min}$ = theoretical minimum tree length; $L$ = tree length for a given tree and matrix (parsimony score).	$0 < CI \leq 1$	Negative	Minimum value will always be above zero.	<a href="#">Kluge and Farris (1969)</a>
Retention index	$RI = \frac{L_{max} - L}{L_{max} - L_{min}}$	$L_{max}$ = theoretical maximum tree length for a matrix (parsimony score for a matrix on a bush phylogeny).	$0 \leq RI \leq 1$	Negative	Minimum value will usually be above zero if the tree has resolved nodes.	<a href="#">Farris (1989)</a>
Homoplasy excess ratio	$HER = \frac{L_{max}^p - L}{L_{max}^p - L_{min}}$	$L_{max}^p$ = mean tree length of minimum length trees inferred for a set of $n$ permuted matrices.	$0 \leq HER \leq 1$	Negative	Difficult to interpret for empirical trees not inferred through maximum parsimony. Value can be negative in certain conditions.	<a href="#">Archie (1989)</a>
Homoplasy slope ratio	$HSR = \frac{L - L_{min}}{L_{max} - L_{min}}$	$L_{max}^r$ = mean tree length for a set of $n$ randomly generated matrices.	$0 \leq HSR \leq 1$	Positive	Only suitable for empirical matrices with no missing data, polymorphisms, or uncertainties.	<a href="#">Meier <i>et al.</i> (1991)</a>
Modified homoplasy slope ratio	$HSRm = \frac{L - L_{min}}{L_{max}^r - L_{min}}$	See HSR formula.	$0 \leq HSRm \leq 1$	Positive	The modified version incorporates non-state scorings into the random matrices, such as missing data ('?') or polymorphisms ('(01)') that are present in the empirical matrix.	Present study
Data decisiveness statistic	$DD = \frac{L_{all} - L}{L_{all} - L_{min}}$	$L_{all}$ = mean length of all possible trees for the empirical matrix.	$0 \leq DD \leq 1$	Negative	$L_{all}$ is not strictly an estimate of maximum possible homoplasy, therefore $DD$ does not measure relative homoplasy.	<a href="#">Goloboff (1991a)</a>
Relative homoplasy index	$RHI = \frac{L - L_{min}}{L_{null} - L_{min}}$	$L_{null}$ = median tree length for a set of $n$ trees with tip labels permuted and the empirical matrix.	$0 \leq RHI \leq 1$	Positive	Conserves the properties of the empirical matrix and empirical tree shape for the null distribution. Value can be more than one in certain conditions.	Present study

### Data decisiveness

An additional index, data decisiveness (DD) (Goloboff 1991a), was originally introduced to measure the decisiveness of a matrix (information that enables a choice of one topology over another; for instance, there being only a few most-parsimonious topologies compared with many non-minimum length topologies). Decisive matrices exhibit a skewed distribution of all possible tree lengths, with non-minimum length trees being more frequent than minimum-length trees and a high degree of variance in tree length across the distribution, whereas undecisive matrices show low variance in tree length across the distribution of all possible trees (Goloboff 1991a). Although not intended to measure homoplasy, DD follows the same general formula of RI and HER:

$$DD = 1 - \frac{L - L_{\min}}{L_{\text{all}} - L_{\min}} = \frac{L_{\text{all}} - L}{L_{\text{all}} - L_{\min}}$$

where  $L_{\text{all}}$  is the mean tree length for all possible trees for the dataset (denoted by  $\bar{S}$  in Goloboff 1991a). In theory,  $L_{\text{all}}$  can be calculated analytically using Theorem 1 from Carter *et al.* (1990) (see additional comments on homoplasy indices in Supporting Information). However, many programmes, including R (R Core Team 2021), have an upper limit for large numbers which will be exceeded once a threshold number of taxa is exceeded; as such,  $L_{\text{all}}$  is challenging to calculate due to the exponential nature of increasing possible topologies as the number of taxa in a dataset increases (Felsenstein 1978, Carter *et al.* 1990, Davis *et al.* 1998). When attempting to calculate DD in R, we were unable to calculate the statistic for more than 85 taxa. It is possible to estimate  $L_{\text{all}}$  from a large enough sample of possible trees inferred through maximum parsimony (Hillis and Huelsenbeck 1992, Davis *et al.* 1998). Additionally, DD is only applicable to datasets with a fully-bifurcating empirical tree for binary characters [which requires many matrices to be modified extensively to adjust all multistate characters to binary (Goloboff 1991a)], and without any ambiguous character states (Goloboff 1991a). Furthermore,  $L_{\text{all}}$  does not represent an estimate of maximum homoplasy, as much of the sample of possible tree lengths is represented by minimum-length or near-minimum-length trees (Goloboff 1991a).

### Homoplasy slope ratio

The homoplasy slope ratio (HSR) estimates homoplasy based on the homoplasy slope (HS) of empirical trees and datasets compared to the HS of a null distribution of randomly generated matrices (Meier *et al.* 1991). These random matrices are generated with the same dimensions as the empirical matrix (same numbers of characters and taxa). In the original description of this method (Meier *et al.* 1991), all characters were binary. Character states for each character and taxon are sampled with equal probability, independent of states for other characters or taxa.

The HSR method uses the CI to calculate a value for extra steps (ES) per character in the matrix as:

$$ES = \frac{1}{CI} - 1 = \frac{L - L_{\min}}{L_{\min}}$$

This value is then used to estimate HS as:

$$HS = \frac{ES}{t - 3} = \frac{L - L_{\min}}{L_{\min} * (t - 3)}$$

where  $t$  is the total number of taxa in the dataset. HSR can be calculated as follows:

$$HSR = \frac{HS_{\text{empirical}}}{HS_{\text{random}}}$$

where the HS of a set of random matrices is calculated and averaged to obtain  $HS_{\text{random}}$ . The HSR is therefore similar to RI and HER, except that HSR scales positively with increasing homoplasy as opposed to negatively. Meier *et al.* (1991) attempted to take number of taxa into account so that the resulting index would not be sensitive to changes in matrix size when comparing homoplasy across datasets. However, the HSR equation can be simplified as follows:

$$HSR = \frac{ES_{\text{empirical}}}{t - 3} \div \frac{ES_{\text{random}}}{t - 3}$$

$$HSR = \frac{L}{L_{\min} * (t - 3)} \div \frac{L_{\text{random}}}{L_{\min} * (t - 3)}$$

$$HSR = \frac{L - L_{\min}}{L_{\text{random}} - L_{\min}} = \frac{L - L_{\min}}{L_{\text{max}}^r - L_{\min}}$$

When the ratio is calculated, the  $(t - 3)$  term is cancelled out and the equation resembles the format of  $1 - \text{HER}$  and  $1 - \text{RI}$ . We will refer to  $L_{\text{random}}$  as  $L_{\text{max}}^r$  for simplicity (Table 1). This also means that although number of taxa are accounted for in each individual HS value, the HSR is not any more robust against changes to taxon number than RI or HER.

### Modified homoplasy slope ratio

In the original description of the HSR method (Meier *et al.* 1991), only binary character matrices with no polymorphisms or missing data were investigated. With datasets that include polymorphisms or missing data, strictly following the randomization protocol of the original method results in unusually low HSR values for empirical datasets (see Table 4). To mitigate this effect, we include a modified version of HSR (hereon named HSRm; Table 1) that follows an identical formula to HSR but which generates random matrices based on all the scorings within the input matrix instead of only the specified states, including polymorphisms and missing data.

### Relative homoplasy index

We introduce a new method to calculate homoplasy, the relative homoplasy index (RHI), which is defined as follows:

$$RHI = \frac{L - L_{\min}}{L_{\text{null}} - L_{\min}}$$

Where  $L_{\text{null}}$  (null tree length) represents the tree length for a given matrix when no characters are phylogenetically informative with respect to the topology except those that are informative due to chance.  $L_{\text{null}}$  is calculated by taking the median tree length for the empirical matrix of a set of  $n$  randomized trees. In these randomized (null) trees, the topology of the empirical tree (i.e. tree shape, including the root) is maintained, but the



tip names are randomized, including outgroups. In this way, observed homoplasy ( $L - L_{\min}$ ) is compared to an estimated maximum value of homoplasy ( $L_{\text{null}} - L_{\min}$ ) for a given matrix and given topology based on a sample of random trees of the same shape. Therefore, the properties of the empirical tree—such as number of polytomies, tree balance, and tree shape—are retained in the set of null trees, avoiding systematic biases associated with these properties (Felsenstein 2004, Simmons *et al.* 2004). The distribution of null tree lengths in the datasets we tested follows a normal distribution (Supporting information, Fig. S1), as can be expected when no phylogenetic information is present except that due to chance (Le Quesne 1989, Huelsenbeck 1991, Hillis and Huelsenbeck 1992). This differs from the dataset-dependent variable shape in distribution for all possible tree lengths ( $L_{\text{all}}$  in DD), whereby highly structured (i.e. decisive) datasets show a more skewed distribution and random or unstructured (i.e. undecisive) datasets follow a more normal distribution (Le Quesne 1989, Goloboff 1991a, Huelsenbeck 1991, Hillis and Huelsenbeck 1992).

Although the equation for RHI is similar in format to the HSR, the randomization procedure to calculate RHI is unique among homoplasy indices described to date and does not rely on any randomization of the empirical matrix, unlike the HSR or the HER (Fig. 1), thus avoiding the computationally expensive tree inference step necessitated by each matrix randomization for HER or time-consuming matrix permutations for both HSR and HER. Further, by only randomizing tips in a tree, the RHI upholds the assumption that the null distribution of trees does not exhibit any phylogenetic structure with respect to the original data except that due to chance. Random tip permutation has been employed in studies previously, for example by Laurin (2004), to test the phylogenetic signal of individual characters and provides an efficient way to generate a distribution of trees that have no phylogenetic information.

The RHI differs from the RI and HER indices in that 0 indicates an absence of relative homoplasy within a matrix (when  $L = L_{\min}$ ), and 1 indicates that a matrix shows as much homoplasy as possible (maximum relative homoplasy) when the corresponding topology is resolved beyond a polytomy (when  $L = L_{\text{null}}$ ). In this case, a value of 1 would occur if the matrix contains no phylogenetic information with respect to the empirical tree, beyond that due to chance, and so the empirical tree is no better suited to the matrix than a random tree of the same shape.

It is possible for the value of RHI to exceed 1 under certain conditions (for example, if the empirical tree has a much higher tree length than the minimum length tree). This would happen if the observed homoplasy was higher than that of the estimated maximum homoplasy given the empirical tree shape and empirical matrix, and therefore the matrix and tree combination contains more homoplasy than a matrix and tree with no phylogenetic signal with respect to each other. This is generally unlikely to occur in real phylogenetic datasets, and the aim of introducing RHI is to provide an alternative method to calculating relative homoplasy in empirical datasets. It remains to be shown how limiting this feature of RHI is, and would be best investigated with a combination of theoretical simulated datasets and a meta-analysis of empirical datasets in future work. In datasets that do exceed unity, RHI would not be a useful measure of relative homoplasy and an alternative method, such as the

retention index, may be more appropriate. Throughout the analyses in this study, RHI does not exceed unity even at the highest levels of homoplasy in simulated datasets with realistic parameters, whereas HER and HSRm do exceed unity (Supporting information, Table S2). Due to randomizing the tips as opposed to the matrix, RHI is unique among homoplasy indices in that its value can be different for two minimum-length trees with different tree shapes inferred from the same matrix. Therefore, one tree shape may have a higher estimated maximum homoplasy value than an alternative tree shape of the same length.

RHI can be calculated for any discrete matrix with a given phylogeny, regardless of the method or matrix that the phylogeny was inferred from. RHI is a cross-comparable index that can be applied to matrices with accompanying phylogenetic trees that may have been generated from independent matrices (e.g. molecular data), when the topology is a consensus between multiple datasets, or when the topology is partially constrained by another phylogenetic hypothesis.

We provide novel functions in the R programming language (R Core Team 2021) to carry out all calculations and analyses (see Supporting Information for all R code associated with this work). Our methodology relies on the *phangorn* (Schliep 2011) and *APE* (Paradis *et al.* 2004) packages and includes functions to make morphological matrices compatible for use with *phangorn*. RHI can be calculated for whole matrices, for individual characters, when incorporating specific characters as ordered, or for a distribution of trees (e.g. a set of most parsimonious trees).

## MATERIAL AND METHODS

### Data treatment

We selected six published discrete morphological datasets with a focus on birds to investigate homoplasy: Telluraves (Ksepka *et al.* 2019), Passeriformes *a* and *b* (Steell *et al.* 2023), Neornithes (Field *et al.* 2020), and Avialae *a* and *b* (Benito *et al.* 2022) (see Supporting Information for all empirical matrix and tree files used in this study). Datasets were selected based on their properties such as tree inference method, taxon and character dimensions, and availability of data from public repositories. Datasets from Benito *et al.* 2022 and Steell *et al.* (2023) were formatted so that two different treatments of the matrix and/or empirical tree could be analysed. For Benito *et al.* 2022 two datasets were analysed (Avialae *a* and Avialae *b*). Both matrices are the same, but the topologies are based on different phylogenetic inference methods; the topology for Avialae *a* was inferred under maximum parsimony, and the topology for Avialae *b* was inferred under Bayesian inference. For Steell *et al.* (2023) two datasets were analysed with the same characters but with slightly different taxon samples and topologies (Passeriformes *a* and Passeriformes *b*). Passeriformes *a* includes only taxa present in the full topological constraint used in the original publication that was inferred from molecular datasets from other studies (Prum *et al.* 2015, Oliveros *et al.* 2019, Harvey *et al.* 2020). Passeriformes *b* includes all topologically constrained taxa from Passeriformes *a* in addition to fossil specimens analysed in the original publication. The topologies used in this study were taken as previously published; as such, some are unrooted and all except Passeriformes *a* are incompletely resolved. We chose to leave the trees in this condition as we were

interested in investigating patterns of homoplasy across a range of real datasets and phylogenetic trees, as opposed to reanalysing phylogenetic relationships. For simplicity, all characters were treated as unordered in our analyses with the same step cost per state transition, although we acknowledge there are numerous ways that different character types regarding cost per state transition can affect homoplasy levels per character (Hoyal Cuthill and Lloyd 2024).

### Analyses

We carried out five sets of analyses (Table 2) to investigate patterns of homoplasy across a range of methods, utilizing empirical matrices, randomly generated matrices, and matrices simulated under an equal rates model of character state evolution. All analyses were carried out in R v.4.3.2 (R Core Team 2021), primarily using functions within the *APE* (Paradis *et al.* 2004), *phangorn* (Schliep 2011), *phytools* (Revell 2012), and *dispRity* (Guillerme 2018) packages, as well as custom functions (Supporting Information).

#### Empirical and random matrices (Analysis 1)

Analysis 1 investigated homoplasy across the aforementioned methods (consistency index, retention index, homoplasy excess ratio, homoplasy slope ratio, modified homoplasy slope ratio, and relative homoplasy index) to compare the performance of each index in empirical matrices and randomly generated matrices. We excluded DD from our analyses for two reasons: firstly, DD is not strictly a measure of relative homoplasy in the same sense as the other methods. Secondly, although the conceptual basis for  $L_{all}$  is similar to  $L_{null}$  (the mean of all possible tree lengths,  $L_{all}$ , and the median of a sample of all tree lengths

for the same tree shape,  $L_{null}$ ), calculating DD for the empirical datasets in this study was not feasible due to the nature of the chosen matrices and trees, with several datasets exceeding the taxon number that R (R Core Team 2021) can accommodate (approximately 85 taxa, which generates  $> 10^{300}$  possible topologies). Additionally, as our aim was not to infer topologies from these matrices, sampling enough possible trees was beyond the scope of this study in order to estimate DD. A more detailed comparison between RHI and DD is warranted in future work.

For the HER and HSR, we created functions to calculate homoplasy following the original method descriptions as accurately as possible (see Supporting Information). We used the benchmarking package *tictoc* (Izrailev 2024) to time each run on empirical datasets for each method when the number of randomizations ( $n$ ) was set to 100 for HSR, HER, and RHI, as well as an additional run of 1000 for RHI (Table 3). Generally, we recommend  $n = 1000$ , but  $n = 100$  may be adequate based on the simulations and the close similarity of the  $L_{null}$  distributions at different values of  $n$  (Supporting information, Fig. S1).

Finally, in Analysis 1, we generated random matrices with the same character and taxon dimensions and approximate state frequencies as each empirical matrix. Each taxon was assigned character states by sampling with replacement from the available states for each character, so that character states per taxon were not an exact permutation of the empirical matrix. We generated random matrices for Telluraves, Passeriformes *a*, Neornithes, and Avialae. Then, we ran a maximum parsimony analysis using the *phangorn* function ‘pratchet’ (Schliep 2011) with ‘maxit’ = 1000, ‘minit’ = 50, and ‘k’ = 10. The resulting trees were used to calculate homoplasy across each method for each random matrix. The aim for this step was to compare efficacy of methods at detecting

**Table 2.** Summary of analyses.

Analysis	Independent variable (x)	x values	Rate categories ( $\mu$ )	Trees per sample	Matrices per tree	No. samples	Total no. simulations
1. Homoplasy in empirical and random matrices	N/A	N/A	N/A	N/A	N/A	N/A	N/A
2. Variable transition rate	Mean log ( $\mu$ )	-8:2 by 0.5	N/A	1	100	21	2100
3. Variable number of taxa	Number of taxa	Telluraves: 10:55 by 5	-6, -4, -2	100	1	10	1000
		Passeriformes (a): 10:135 by 15	-6, -4, -2	100	1	9	900
		Neornithes: 10:37 by 3	-6, -4, -2	100	1	9	900
		Avialae (a): 10:82 by 8	-6, -4, -2	100	1	9	900
4. Variable number of characters	Number of characters	Telluraves: 10:130 by 15	-6, -4, -2	1	100	9	900
		Passeriformes (a): 10:45 by 5	-6, -4, -2	1	100	8	800
		Neornithes: 10:280 by 30	-6, -4, -2	1	100	10	1000
		Avialae (a): 10:280 by 30	-6, -4, -2	1	100	10	1000
5. Principal coordinates in phylomorphospace	Mean log ( $\mu$ )	-6:0 by 2	N/A	1	1	4	4

**Table 3.** Summary of run times of each index per empirical dataset. For methods involving a randomization step, number of randomizations ( $n$ ) = 100 unless stated otherwise.

Dataset	No. taxa	No. characters	Run time (s)						
			CI	RI	HER	HSR	HSRm	RHI	RHI ( $n = 1000$ )
Telluraves	62	146	0.01	0.01	366	1.82	1.99	0.08	0.32
Passeriformes (a)	143	49	0.01	0.01	1504	1.22	1.34	0.08	0.33
Passeriformes (b)	154	49	0.01	0.00	1760	1.35	1.46	0.05	0.20
Neornithes	39	295	0.01	0.02	179	2.33	2.59	0.08	0.27
Avialae (a)	85	282	0.02	0.02	710	4.97	6.28	0.16	0.68
Avialae (b)	85	282	0.02	0.01	727	4.82	6.27	0.13	0.46

homoplasy in highly homoplastic, worst-case scenario matrices with no missing data or polymorphisms.

### Simulated matrices (Analyses 2–4)

Analyses 2–4 investigated patterns of homoplasy in simulated matrices for topologies of Telluraves, Passeriformes *a*, Neornithes, and Avialae *a*; see extended methodology in [Supporting Information](#) for details on the simulation protocol. Due to the substantial run times for the HER and HSR ([Table 3](#)), we decided not to include these indices in our simulations due to the expected computational time being orders of magnitude longer than when calculating CI, RI, and RHI for the 800–2100 simulated matrices per topology per analysis ([Table 2](#)). We also believe Analysis 1 is sufficient to determine the efficacy of each method at detecting homoplasy in a range of datasets, and instead Analyses 2–4 focus on the behaviour of homoplasy in discrete matrices with changing variables.

### Character state transition rates (Analysis 2)

We investigated homoplasy across different character state transition rates by varying mean log ( $\mu$ ) values when generating simulated datasets. We used the rate of character state change (hereon referred to as transition rate) as a proxy for homoplasy in simulated matrices. Higher transition rates lead to increased homoplasy because a greater number of character state transitions may result in independent acquisitions of or reversals to the same state ([Simmons \*et al.\* 2004](#), [Harrison and Larsson 2015](#)). However, the efficacy of increasing transition rates as a proxy for homoplasy will decrease with increased numbers of character states present in a matrix (i.e. more multistate characters with more states per character) and may produce unexpected effects on homoplasy detection in datasets with many multistate characters ([Hoyal Cuthill 2015b](#)). Moreover, the assumption of homoplasy increasing in tandem with transition rates may not be valid if transitions between states do not occur with equal probability. Therefore, all our simulations assume a conservative 9:1 ratio of binary to three-state characters and equal probabilities of character state change directionality. We varied transition rates by employing a range of values for the mean log ( $\mu$ ) transition rate and simulated 100 matrices per rate category ([Table 2](#)). The range of values (–8 to 2) for  $\mu$  was selected based on trialling multiple scenarios; the selected range was appropriate for the empirical datasets investigated here, as it covers very low transition rates up to very high rates. The median, 5%, and 95% quantiles of RHI, 1 – CI, and 1 – RI were plotted against mean log transition

rate for the Telluraves, Passeriformes *a*, Neornithes, and Avialae *a* topologies. We also normalized median treelength ( $L$ ) between 0 and 1 with the following formula:

$$\frac{L - \min(L)}{\max(L) - \min(L)}$$

where  $\min(L)$  and  $\max(L)$  represent the minimum and maximum treelengths from the simulation.

### Number of taxa (Analysis 3)

Homoplasy across datasets with varying numbers of taxa was investigated by sampling tips from each topology without replacement in decreasing increments ([Table 2](#)). We sampled 100 reduced-taxa trees per taxon-number increment and simulated one matrix per reduced-taxa tree. We also simulated 100 matrices for each topology with the full set of tips. Analysis 3 was reproduced three times with different character state transition rate categories:  $\mu = -6$  (slow),  $\mu = -4$  (medium), and  $\mu = -2$  (fast).

### Number of characters (Analysis 4)

Homoplasy across datasets with varying numbers of characters was investigated by sampling characters in decreasing increments ([Table 2](#)). We simulated 100 matrices with specified character numbers for each increment for each topology. This was repeated three times with the same rate categories as Analysis 3.

### Effect of homoplasy on morphospace occupation (Analysis 5)

To investigate the behaviour of datasets with different levels of homoplasy in morphospace, we simulated matrices with varying transition rates for each empirical topology, carried out a principal coordinate analysis for each matrix and plotted the first two axes of variation in a phylomorphospace. We chose four character state transition rate categories by varying mean log values ( $\mu = -6, -4, -2, 0$ ), following the same procedure as Analyses 2–4 ([Table 2](#)). One matrix per empirical topology per rate category was generated. Relative corrected eigenvalues for the first five axes, which indicate the proportional variation accounted for per principal coordinate, as well as CI, RI, HER, HSR, HSRm, and RHI for each simulated matrix are reported in [Supporting information, Table S2](#).

## RESULTS

We quantified homoplasy in empirical and randomly generated datasets across a range of methods (Analysis 1, [Table 2](#)). Results



**Table 4.** Homoplasy index results from Analysis 1 (homoplasy across empirical and random matrices). MP refers to maximum parsimony.

Dataset	Input tree	'?' (%)	Tree length	1 – CI	1 – RI	1 – HER	HSR	HSRm	RHI
<b>Empirical</b>									
Telluraves	MP	27	702	0.74	0.40	0.58	0.16	0.44	0.45
Passeriformes (a)	-	8	995	0.94	0.49	0.93	0.39	0.62	0.62
Passeriformes (b)	Bayesian	11	1056	0.95	0.51	0.98	0.37	0.62	0.62
Neornithes	Bayesian	18	1498	0.73	0.42	0.57	0.28	0.51	0.50
Avialae (a)	MP	51	1407	0.72	0.35	0.51	0.11	0.40	0.41
Avialae (b)	Bayesian	51	1616	0.76	0.42	0.61	0.13	0.46	0.47
<b>Random</b>									
Telluraves	MP	0	2815	0.93	0.63	1.00	0.85	0.88	0.86
Passeriformes (a)	MP	0	1760	0.97	0.50	1.00	0.71	0.72	0.71
Neornithes	MP	0	3964	0.90	0.68	1.00	0.91	0.94	0.91
Avialae	MP	0	8145	0.95	0.66	1.00	0.89	0.91	0.89

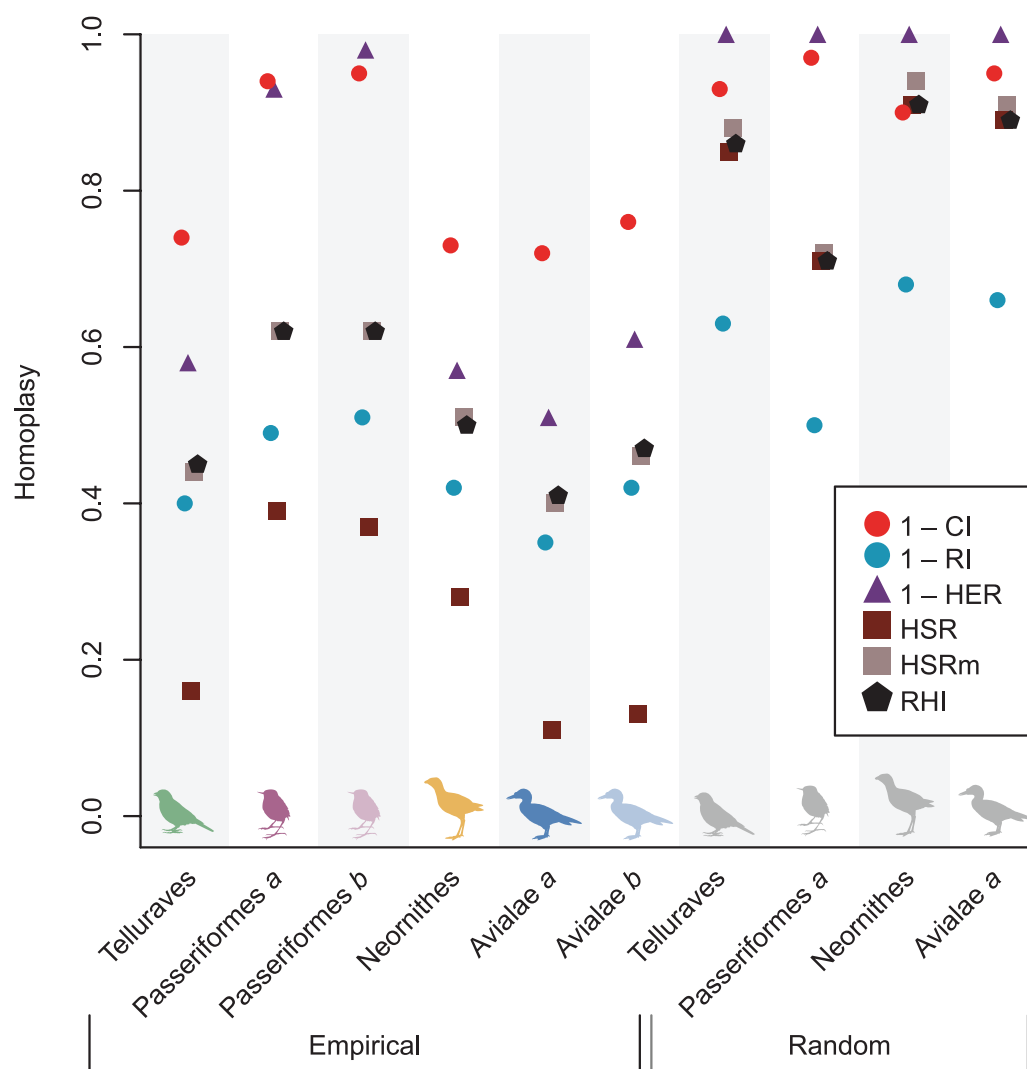
are summarized in Table 4 and Figure 2 (note that Fig. 2 shows CI, RI, and HER values as 1 minus each value for comparison to other methods). We found that results varied greatly depending on whether the matrix was empirical or randomly generated, revealing that some methods were sensitive to properties within empirical matrices that are not present in randomly generated matrices. The HSR estimated very low levels of homoplasy for all empirical matrices, but higher levels of homoplasy for random matrices that were also near-identical in value to the HSRm and RHI for those datasets. HER estimated maximum or near-maximum levels of homoplasy for all randomly generated matrices as well as for the empirical Passeriformes datasets. We also found that HSRm and RHI values were identical to two decimal points or near identical across all datasets.

We varied the rate of character state change as a proxy for varying levels of homoplasy in simulated datasets in Analysis 2. Results for CI, RI, and RHI calculations as well as normalized tree length ( $L^*$ ) across these simulations are summarized in Figure 3. The results indicated that as transition rate increased (by increasing mean  $\log \mu$ ), homoplasy increased (Fig. 3), although the extent of homoplasy detected was dependent on the index applied. Normalized tree length for each dataset showed that at approximately  $\mu = -2$ , tree length reached its maximum value for each dataset (Fig. 3). The CI rapidly indicated higher levels of absolute homoplasy with a small increase in transition rate, and subsequently asymptotes at a consistently lower  $\mu$  value than the other indices. Ranges of y-axis values are reported in Table 5, as well as area under the curve values, the difference between areas under the  $L^*$  curve, and areas under curves for 1 – CI, 1 – RI, and RHI. RHI consistently yielded the greatest y-axis range across each dataset, and RI consistently yielded the smallest range of homoplasy values (i.e. 1 – RI always asymptotes at a homoplasy value lower than one). Normalized tree length, 1 – RI, and RHI all asymptote at approximately the same  $\mu$  values for each dataset. The area difference between RHI and  $L^*$  for each dataset was the smallest across the three methods (Table 5), which can be seen in Figure 3 in which the RHI curve closely tracked the  $L^*$  curve. The y-axis range difference for each index is reported in Supporting Information, Table S1.

We tested the effects of varying numbers of taxa and characters using simulated matrices in Analyses 3 and 4, respectively.

Results for Analysis 3 are summarized in Figure 4 and showed that as taxa were added to datasets (i.e. as tips are added to a phylogeny), the levels of homoplasy detected by CI, RI, and RHI changed. CI detected more absolute homoplasy as number of taxa was increased, and RI and RHI both detected less relative homoplasy as number of taxa was increased. These effects were also variable depending on the underlying transition rate of character state change (Fig. 4; Supporting Information, Table S1); a greater difference in detected homoplasy was observed for a mid to high transition rate ( $\mu = -4$ ) than for either the low- or high-rate categories ( $\mu = -6$  and  $\mu = -2$ , respectively). Results for Analysis 4 are summarized in Figure 5 and show that as characters are added to a matrix, no change in levels of homoplasy was detected across all methods tested (Supporting Information, Table S1). We found that in small datasets (either low numbers of taxa or low numbers of characters), the variance in detected homoplasy was greater (Figs 4, 5).

Finally, we generated phylomorphospace plots (Sidleuskas 2008) using Principal Coordinate (PCo) analysis for four rate categories of simulated matrices (Fig. 6). In general, matrices at low rates of character state change showed distinct morphospace partitioning between clades, with a trend towards more overlap between clades as rates were increased. Visually, internal branches at low transition rates were longer than terminal branches, whereas at high transition rates (and therefore high homoplasy), branch lengths across the whole tree appeared equivalent, or with longer terminal branches. The amount of partitioning and overlap in morphospace depended to some extent on properties of the tree and character matrix. For instance, the dataset with the most taxa (Passeriformes a) showed less partitioning of morphospace even at very low levels of homoplasy (RHI = 0.08; Fig. 6b; Supporting information, Table S2) compared to Telluraves or Avialae a at equivalent homoplasy levels, whereas the dataset with the fewest taxa (Neornithes) showed substantial partitioning even at more moderate levels of homoplasy (RHI = 0.25; Fig. 6c; Supporting information, Table S2). There was a general trend of decreasing morphological variation accounted for in PCo axes 1 and 2 from low levels to high levels of relative homoplasy (Fig. 6; Supporting information, Table S2).



**Figure 2.** Homoplasiness indices for empirical and random matrices (Analysis 1). Homoplasiness values (Table 4) for consistency index (CI), retention index (RI), homoplasiness excess ratio (HER), homoplasiness slope ratio (HSR), modified homoplasiness slope ratio (HSRm), and relative homoplasiness index (RHI) are plotted for Analysis 1 for empirical datasets (left-hand side) and randomly generated datasets (right-hand side, corresponding grey bird silhouettes). Note that values for consistency and retention indices and homoplasiness excess ratio are subtracted from 1 for ease of comparison to other methods. A value of 1 indicates the maximum amount of detectable homoplasiness for each method.

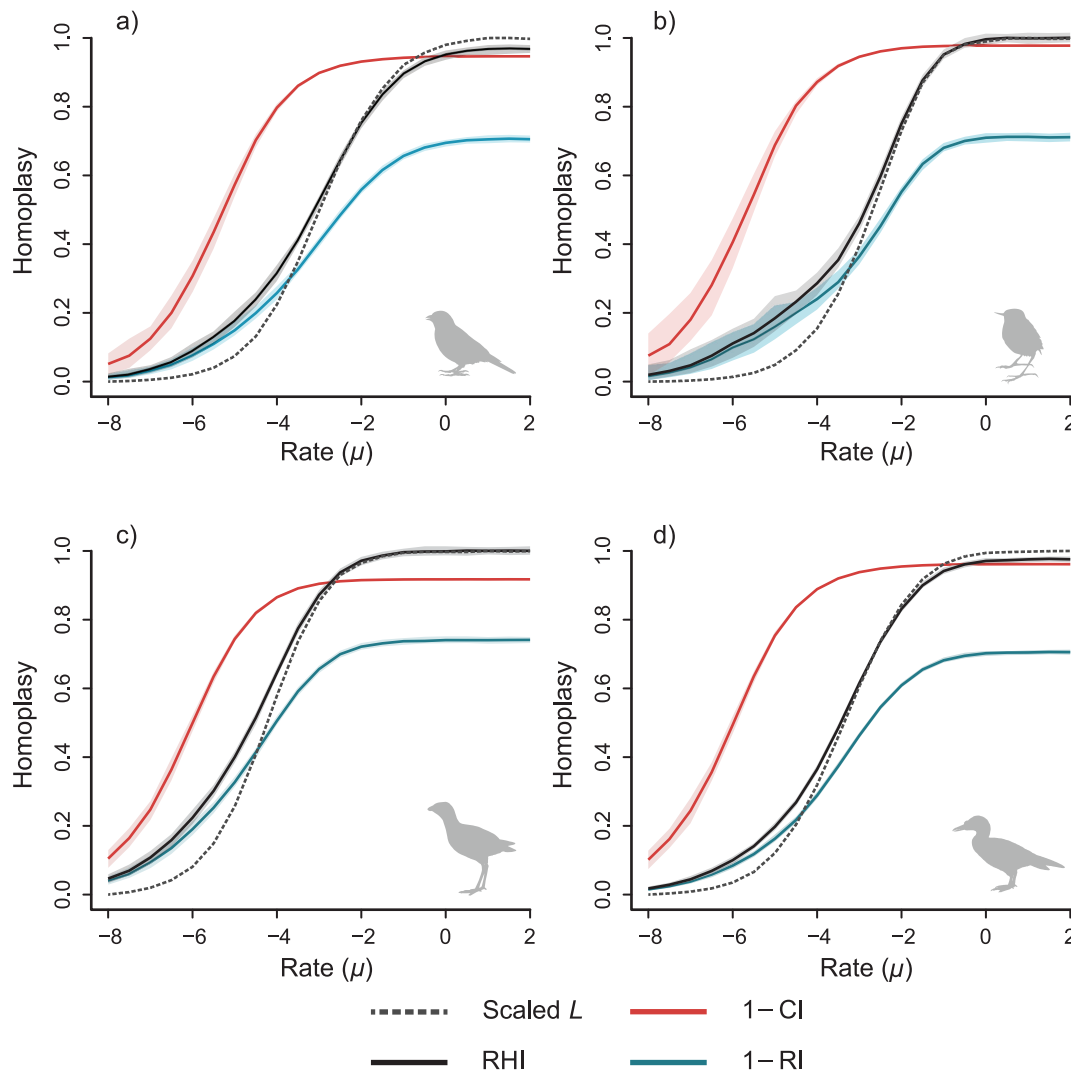
## DISCUSSION

We investigated the behaviour of homoplasiness in discrete matrices by testing different methods with empirical, random, and simulated matrices (Tables 1, 2). In general, levels of relative homoplasiness decrease as taxa are added to a dataset (Fig. 4) but remain constant when characters are added to a dataset (Fig. 5). Our results show that the commonly used consistency (CI) and retention (RI) indices are problematic with respect to appropriate scaling between zero and one (Figs 2 and 3, Tables 4 and 5), hampering meaningful comparisons among different datasets. The homoplasiness excess ratio (HER), first developed to overcome the scaling issue with CI (Archie 1989, 1996), is usually not a suitable method when applied to modern datasets due to its maximum parsimony tree inference step and the time it takes to execute a single calculation (Table 3). The homoplasiness slope ratio (HSR) is also not appropriate with datasets that contain missing data or other non-state scorings due to its randomization step (Figs 1 and 2). The modified homoplasiness slope ratio (HSRm, described here) proves to be more appropriate

for empirical datasets. The relative homoplasiness index (RHI), our new method with a novel randomization step (Fig. 1 and Table 1), overcomes the above-mentioned issues by being scaled between zero and one for realistic datasets (Fig. 3) and maintaining the characteristics unique to the input matrix when estimating maximum homoplasiness, ensuring that the null distribution of trees is appropriate for the dataset at hand. Furthermore, the RHI is computationally faster than all other methods that involve a randomization step (i.e. HER, HSR, and HSRm; Table 3) due to the simplicity of permuting the tree tip order and maintaining the input dataset and tree shape (Fig. 1), yielding similar results to the HSRm while exhibiting a reduction of computation time in the range of ~16–48×.

## Comparison of methods

The major limitation of the RI is that it underestimates the proportional extent of homoplasiness in a given dataset and cannot be compared across alternative datasets because it is not scaled between zero and one (Table 5), but rather between one and an



**Figure 3.** Levels of homoplasy when varying character state transition rate in simulated matrices (Analysis 2). Results for 1 – CI (consistency; red), 1 – RI (retention; blue) and RHI (relative homoplasy; black) indices, as well as scaled L (tree length normalized between 0 and 1; grey dotted line), when the mean log for character state transition rate ( $\mu$ ) is varied for simulated matrices for the Telluraves (a), Passeriformes *a* (b), Neornithes (c), and Avialae *a* (d) topologies. Envelopes around lines represent 95% and 5% quantile ranges.

arbitrary value conditioned on each particular dataset (Archie 1996). This becomes particularly problematic when datasets exhibit elevated levels of homoplasy, as RI underestimates levels of relative homoplasy to a greater extent than RHI (Fig. 3; Table 5). We see this when comparing RI values for empirical and random matrices of Passeriformes *a* in Analysis 1 (Fig. 2; Table 4); RI is unable to discriminate between the homoplasy levels in these matrices with identical numbers of characters, taxa, and character state frequencies, but HER, HSR, HSRm, and RHI do detect a difference between these matrices (Table 4).

Our results suggest that HER is not an appropriate method to estimate homoplasy unless specific conditions are met in the tree inference step for the input topology, due to the null model of maximum homoplasy not being suitable for most empirical matrices that have a tree inferred through any method except maximum parsimony. We find that HER overestimates homoplasy in many cases; HER detects maximal levels of homoplasy in matrices where other methods detect more

moderate levels (e.g. empirical Passeriformes *a* and *b*, Fig. 2). HER assumes that empirical matrices cannot exhibit more homoplasy than a random matrix, but negative HER values indicate that this scenario is possible. Negative HER values were reported by Farris (1991) and Murphy *et al.* (2021), as well as Hoyal Cuthill (2015b) during simulations with very high levels of inferred homoplasy, and Goloboff (1991b) suggested that all undecisive matrices would yield negative HER values. Indeed, HER values for simulated datasets in Analysis 5 show negative values at high rates of character state change for all simulated datasets (Supporting information, Table S2), and these values are not all close to zero; for Passeriformes *a*, HER is as low as -0.41 in the highest rate category ( $\mu = 0$ ), suggesting that tree length for the input tree and simulated matrix far exceeds the average tree length for permuted matrices in the HER randomization step.

We found that HSR greatly underestimates relative homoplasy in empirical matrices due to the randomly generated matrices

**Table 5.** Results from Analysis 2 (variable transition rate).

Topology	y-axis range				Area under curve				Area difference from $L^*$		
	$L^*$ <sup>a</sup>	1 – CI	1 – RI	RHI	$L^*$	1 – CI	1 – RI	RHI	1 – CI	1 – RI	RHI
Telluraves	0–1	0.05–0.95	0.01–0.71	0.01–0.97	4.98	6.97	3.89	5.20	-1.99	1.09	-0.22
Passeriformes (a)	0–1	0.08–0.98	0.02–0.71	0.02–1.00	4.79	7.52	3.92	5.29	-2.73	0.87	-0.50
Neornithes	0–1	0.10–0.92	0.04–0.74	0.05–1.00	6.25	7.45	5.10	6.70	-1.20	1.15	-0.45
Avialae (a)	0–1	0.10–0.96	0.02–0.71	0.02–0.98	5.39	7.69	4.10	5.54	-2.30	1.29	-0.15

<sup>a</sup> $L^*$  denotes normalized tree length.

not being an appropriate null model to estimate maximum homoplasy, but that HSRm does appear to be a suitable alternative when all character scorings are taken into consideration. Although not identical, the matrix randomization protocol in HSRm (Fig. 1) is almost numerically equivalent to the tip randomization in RHI, and the two indices consistently yield very similar values (Table 4; Supporting information, Table S2). However, generating randomized matrices is more computationally demanding than randomizing tips on a tree, so we recommend the use of the more efficient RHI method (Table 3).

### Homoplasy in discrete matrices

Analysis 3 shows that, with the addition of taxa, relative homoplasy detected by RHI and RI decreases, whereas absolute homoplasy detected by CI increases considerably (Fig. 4). This pattern in CI mirrors early work investigating homoplasy in discrete matrices, which concluded that adding more taxa into a dataset would increase estimated levels of homoplasy (Sanderson and Donoghue 1989). This longstanding hypothesis has been corroborated by numerous subsequent studies (Meier et al. 1991, Hauser and Boyajian 1997, Murphy et al. 2021), and is explained by the increased detection of individual homoplastic character state changes with the addition of taxa and/or characters, thereby resulting in more absolute homoplasy (Sanderson and Donoghue 1989, Hauser and Boyajian 1997, Murphy et al. 2021).

CI decreases due to cumulative homoplastic events in a dataset resulting in a greater tree length. However, when considering CI in this format:

$$CI = \frac{L_{min}}{H + L_{min}}$$

$H$  represents the number of extra steps for a tree with a given matrix, and can be considered an unscaled, arbitrary value of homoplasy (Kluge and Farris 1969).  $L_{min}$  does not change with respect to additional taxa, and only varies with changes in number of characters or the composition of character states in a matrix. Therefore, although a tree has a greater probability of gaining homoplastic state changes with the addition of more branches (i.e.  $H$  increases with additional taxa), this absolute quantification of homoplasy in the CI equation is arbitrary and not comparable across datasets (Klassen et al. 1991), as in other indices that attempt to quantify relative homoplasy levels (observed homoplasy compared to an estimated or theoretical maximum). Thus, CI is not useful when attempting to compare estimates of homoplasy from alternative datasets. The original purpose of CI was to measure the consistency of a given dataset with a given

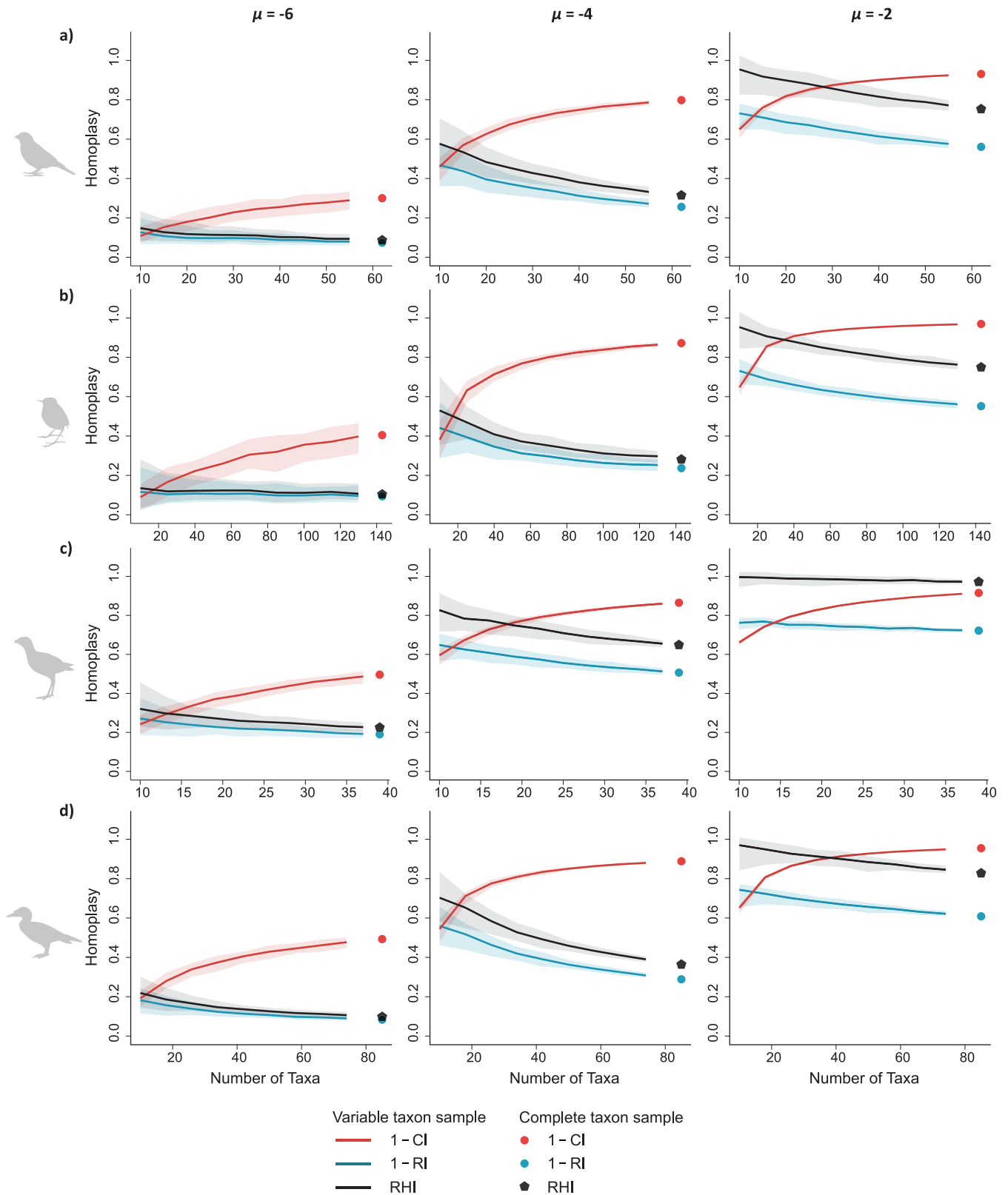
tree in order to illustrate how far an empirical topology diverged from a theoretical zero-homoplasy topology derived from a version of the matrix where each character underwent the minimal number of changes (i.e.  $L = L_{min}$ ; Farris 1989, Kluge and Farris 1969). The use of CI to quantify homoplasy has been warned against previously due to its high sensitivity towards increasing numbers of taxa (Archie 1989, 1996).

RI and RHI, on the other hand, are expected to measure a decrease in relative homoplasy with the addition of taxa to a given dataset, as demonstrated in our results (Fig. 4). The addition of branches to a topology leads to an increase in the total potential amount of homoplasy in a dataset ( $L_{max}$  in the case of RI and  $L_{null}$  in the case of RHI; Fig. 1; Table 1) due to the exponentially increasing number of possible arrangements of tips as taxa are added (Felsenstein 1978) and the increasing number of possible character state combinations available in character state space (Foote 1994, Hoyal Cuthill 2015a). However, provided the rate distribution of character state transitions for the matrix is maintained, there is no reason to expect more relative homoplasy in a dataset with more taxa. On the contrary, observed homoplasy values will be proportionally lower compared to the increasing potential homoplasy of the dataset as a taxon sample increases.

We find that CI, RI, and RHI do not vary with increasing numbers of characters (Fig. 5). For CI, provided the rate distribution of character state transitions is the same regardless of number of characters,  $L$  and  $L_{min}$  (and therefore  $H$ ) increase proportionally as characters are added. For RI and RHI, again assuming transition rates are maintained,  $L$  and  $L_{max}$  and  $L$  and  $L_{null}$ , respectively, also increase proportionally as characters are added.

Previous studies investigating patterns of detected homoplasy with CI, RI, and additional parsimony indices such as the HER (Archie 1989) and the HSR (Meier et al. 1991) report conflicting results. Archie (1989) also observed that RI (therein referred to as the ‘homoplasy excess ratio maximum’) detects less relative homoplasy in subsets of datasets with more taxa, although a slight increase in relative homoplasy with the addition of characters is observed (Archie 1989). Other studies using multiple non-subsetted empirical datasets report mixed results regarding RI. For example, Hauser and Boyajian (1997) found a slight decrease in relative homoplasy with increasing taxa in their meta-analysis of 600 datasets, but Murphy et al. (2021) showed the opposite trend in an analysis of 364 datasets. These variable patterns are likely explained by idiosyncrasies specific to the datasets analysed; each dataset exhibits a particular amount of estimated homoplasy dependent on factors such as tree shape, character state space (Simmons et al. 2004), and

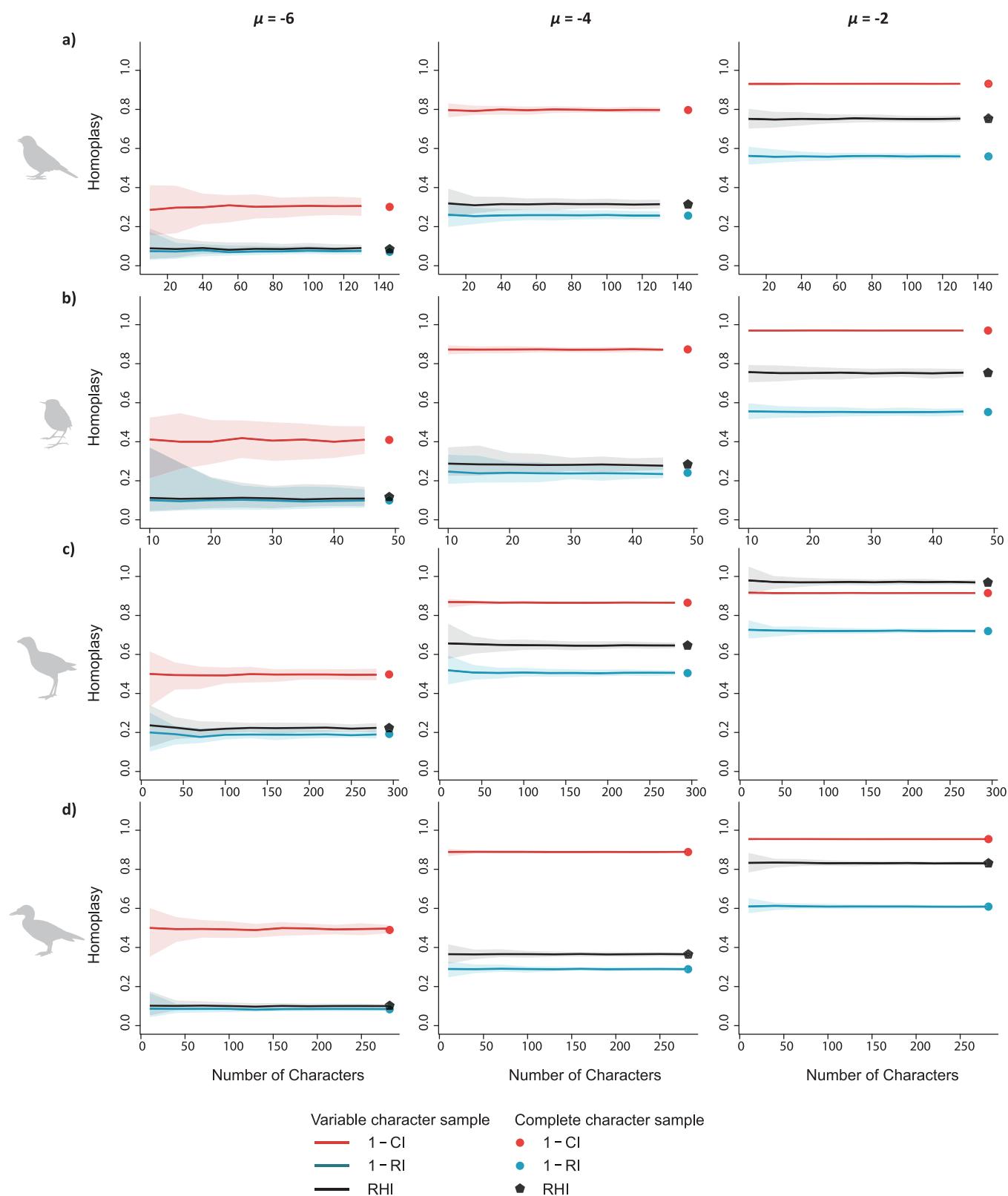




**Figure 4.** Levels of homoplasy when varying numbers of taxa in simulated matrices (Analysis 3). Results for varying number of taxa in simulated matrices at low ( $\mu = -6$ ), medium ( $\mu = -4$ ), and high ( $\mu = -2$ ) rates of character state change for the Telluraves (a), Passeriformes (b), Neornithes (c), and Avialae (d) topologies. Envelopes represent 95% and 5% quantile ranges. Points represent index values with the full set of tips. y-axis ranges are reported in [Supporting Information, Table S1](#).

biological drivers of homoplasy. As our results show, larger datasets do not show more relative homoplasy based on their dimensions alone.

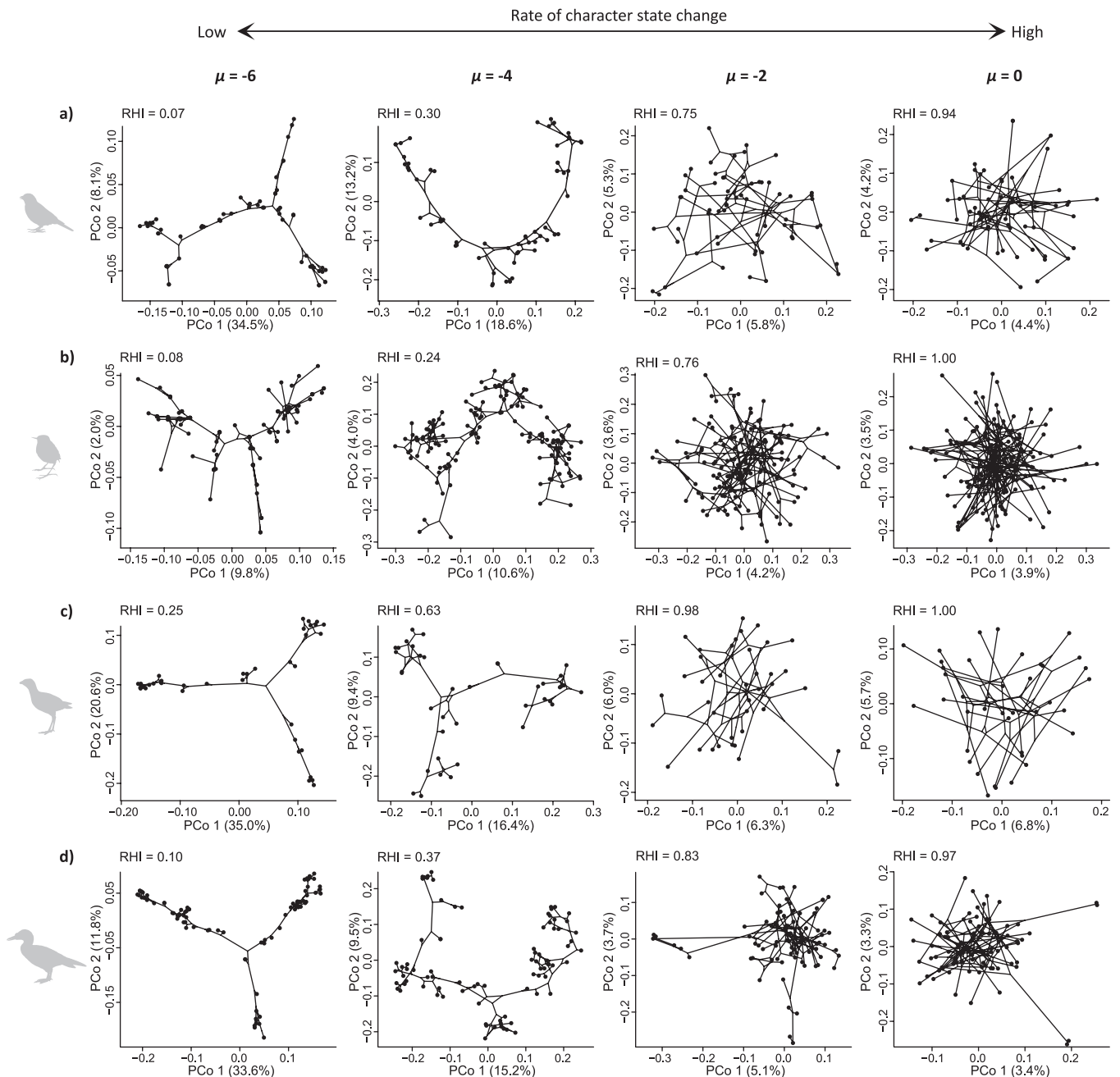
Tree shape is expected to impact patterns of homoplasy estimation ([Simmons et al. 2004](#)), which is why an advantage of the RHI is the novel tip randomization process. We randomized



**Figure 5.** Levels of homoplasy when varying numbers of characters in simulated matrices (Analysis 4). Results for varying number of characters in simulated matrices at low ( $\mu = -6$ ), medium ( $\mu = -4$ ), and high ( $\mu = -2$ ) rates of character state change for the Telluraves (a), Passeriformes (b), Neornithes (c), and Avialae (d) topologies. Envelopes represent 95% and 5% quantile ranges. Points represent index values with the full set of characters. y-axis ranges are reported in [Supporting Information, Table S1](#).

tips to retain the overall tree shape of each empirical tree, rather than sampling randomly generated trees that may come from a distribution of variable tree shapes with unequal frequencies

(Felsenstein 2004). Simmons *et al.* (2004) explored the effect of tree shape on RI and CI with simulated datasets and found that with increasing transition rate, homoplasy detected by RI



**Figure 6.** Phylomorphospace plots of simulated matrices at different character state transition rates from Analysis 5. Axes 1 and 2 from each Principal Coordinate (PCo) analysis for simulated matrices based on the empirical topologies for Telluraves (a), Passeriformes (b), Neornithes (c), and Avialae (d) are plotted with percentage of variation accounted for in each axis. Relative homoplasy index (RHI) is reported for each matrix. Values along each axis represent eigenvectors.

is affected substantially by tree shape. At low transition rates, levels of homoplasy increase significantly more for a completely asymmetrical, unequal branch-length tree than they do for a completely symmetrical, equal branch-length tree (Simmons *et al.* 2004). We expect that tree shape affects relative homoplasy measured by RHI in cases with multiple most-parsimonious trees for the same dataset. Trees of different shapes but the same length may have different RHI values, as we expect that some tree shapes are associated with a greater amount of possible homoplasy than others, which warrants further exploration in future work. Investigating homoplasy patterns with random

trees generated by different methods, for example those implemented in APE (Paradis *et al.* 2004), was beyond the scope of the current study but could be useful for the development of new, non-parsimony-based methods for homoplasy quantification.

### Using homoplasy in macroevolutionary studies

Estimating relative homoplasy for discrete matrices is not only useful for estimating the overall phylogenetic informativeness of characters but has wider implications when considering macroevolutionary patterns of morphological disparity and evolution. Although some phylogenetic matrices may not be an ideal source

of data for disparity analysis (Lloyd 2016), particularly at very broad taxonomic scales, many datasets provide sufficient information to explore morphological evolution with discrete data (Gerber 2019). Quantifying homoplasy in discrete datasets may therefore become more commonplace for evaluating macroevolutionary patterns across different clades, particularly if this is combined with additional analyses investigating disparity and character state space occupation (Oyston *et al.* 2015, Lloyd 2016, 2018, Gerber 2019).

Our phylomorphospace results from simulated matrices show decreasing morphospace partitioning and longer terminal branch lengths for principal coordinate (PCo) axes 1 and 2 as character state transition rates (and therefore absolute and relative homoplasy) increase (Fig. 6; Supporting information, Table S2). Our results appear to be analogous to snapshots of clade evolution through time; high morphospace partitioning and low homoplasy are consistent with the idea that clades explore morphospace early in their evolutionary history and, as a result, disparity rapidly increases as taxa occupy new combinations of character states (Wagner 2000, Harmon *et al.* 2003, Hughes *et al.* 2013, Oyston *et al.* 2015). Similarly, higher levels of homoplasy corresponding with weak morphospace partitioning are consistent with character state exhaustion that is expected to occur later in the evolutionary history of a clade when biological constraints are acting on morphological evolution, resulting in convergent character state combinations (Wagner 2000, Oyston *et al.* 2015, Brocklehurst and Benson 2021). We also see that increasing relative homoplasy is associated with a general trend of decreasing amounts of variation explained per principal coordinate axis (Fig. 6; Supporting information, Table S2). This is potentially explained by varying levels of integration among characters across different relative homoplasy levels, where it is expected that trait integration is stronger if the standard deviation of eigenvalues is high (i.e. if most variation can be explained by the first few axes of the PCo; Gerber 2013).

Simulated datasets with varying amounts of homoplasy may prove to be a helpful tool for emulating different stages of clade evolution to further investigate the tempo and mode of evolution of discrete characters and the role of integration in shaping morphospace. However, interpreting phylomorphospaces of discrete characters is more challenging due to the geometric nature of morphospace when discrete traits are ordinated into a multi-dimensional character state space (Gerber 2019). Therefore, it is increasingly recommended to exercise caution when using these techniques, particularly as the tools to visualize and interpret discrete morphospaces are less developed than those of continuous traits (Lloyd 2018, Gerber 2019). Using relative homoplasy quantification in tandem with analyses of character state exhaustion, integration, and disparity across discrete datasets will help advance our understanding of how homoplastic traits accumulate within clades, potentially shedding further light on the properties of convergent evolution in nature.

## SUPPORTING INFORMATION

Supplementary data is available at *Zoological Journal of the Linnean Society* online.

Extended methodology. Additional comments on homoplasy indices and analyses.

Figure S1. Kernel density plots to show distribution of the null tree length ( $L_{\text{null}}$ ) for empirical datasets at different  $n$  values.

Table S1. y-axis ranges for homoplasy indexes for Analyses 2–4.

Table S2. Summary of homoplasy indexes and variation accounted for in phylomorphospace axes for simulated matrices in Analysis 5.

Usage notes for custom functions. A guide to using custom functions associated with analyzing levels of homoplasy.

List of R scripts. File names for associated functions and analyses for use in R.

List of data files. File names for trees and matrices used in the analyses.

## ACKNOWLEDGEMENTS

We thank D. Ksepka for providing a tree file of the published Telluraves topology for analyses. We are grateful to R. Benson, G. Thomas, R. Asher, D. Brennan, J. Benito, G. Navalón, M.G. Burton, and A. Chen for insightful discussion regarding this methodological approach. This manuscript was substantially improved by comments from P. Goloboff, P. Wagner, M. Wilkinson, and two additional anonymous reviewers. This work was supported by an UK Research and Innovation Future Leaders Fellowship MR/X015130/1 to D.J.F., a Natural Environment Research Council grant NE/S007164/1 and the Sarah Woodhead Research Fellowship (Girton College, Cambridge) to E.M.S., and a Swedish Research Council (Vetenskapsrådet) Starting Researcher Grant ÅR-NT 2020-03515 to A.Y.H. For the purposes of open access, the authors have applied a Creative Commons Attribution (CC BY) license to any Author Accepted Manuscript version arising.

## AUTHOR CONTRIBUTIONS

E.M.S. conceived the ideas, designed the methodology, compiled the data, and carried out the analyses. A.Y.H. and D.J.F. helped develop the methodology and supervised the project. E.M.S. wrote the first draft of the manuscript and made the figures. A.Y.H. and D.J.F. provided critical feedback on early drafts and contributed to writing the final version of the manuscript.

## CONFLICT OF INTEREST

The authors declare no conflict of interest.

## DATA AVAILABILITY

All raw data files, including tree and matrix files in PHYLIP and NEXUS format, are available in the [Supporting Information](#). All R code is provided in the [Supporting Information](#), and the development version of the code can be found at [github.com/LizzySteell/Homoplasy](https://github.com/LizzySteell/Homoplasy), which is regularly maintained by E.M.S.

## REFERENCES

- Archie JW. Homoplasy excess ratios: new indices for measuring levels of homoplasy in phylogenetic systematics and a critique of the consistency index. *Systematic Zoology* 1989;38:253–69. <https://doi.org/10.2307/2992286>
- Archie JW. Measures of homoplasy. In: Sanderson MJ, Hufford L (eds), *Homoplasy: the Recurrence of Similarity in Evolution*. San Diego, CA: Academic Press, 1996, 153–88.



- Benito J, Kuo PC, Widrig KE *et al.* Cretaceous ornithurine supports a neognathous crown bird ancestor. *Nature* 2022;**612**:100–5. <https://doi.org/10.1038/s41586-022-05445-y>
- Brocklehurst N, Benson RJ. Multiple paths to morphological diversification during the origin of amniotes. *Nature Ecology and Evolution* 2021;**5**:1243–9. <https://doi.org/10.1038/s41559-021-01516-x>
- Brocklehurst N, Panciroli E, Benevento GL *et al.* Mammaliaform extinctions as a driver of the morphological radiation of Cenozoic mammals. *Current Biology* 2021;**31**:2955–63.e4. <https://doi.org/10.1016/j.cub.2021.04.044>
- Carter M, Hendy M, Penny D *et al.* On the distribution of lengths of evolutionary trees. *SIAM Journal on Discrete Mathematics* 1990;**3**:38–47. <https://doi.org/10.1137/0403005>
- Davis JL, Simmons MP, Stevenson DW *et al.* Data decisiveness, data quality, and incongruence in phylogenetic analysis: an example from the monocotyledons using mitochondrial *atp A* sequences. *Systematic Biology* 1998;**47**:282–310. <https://doi.org/10.1080/106351598260923>
- Farris JS. The retention index and the rescaled consistency index. *Cladistics* 1989;**5**:417–9. <https://doi.org/10.1111/j.1096-0031.1989.tb00573.x>
- Farris JS. Excess homoplasy ratios. *Cladistics* 1991;**7**:81–91. <https://doi.org/10.1111/j.1096-0031.1991.tb00023.x>
- Felsenstein J. The number of evolutionary trees. *Systematic Zoology* 1978;**27**:27–33. <https://doi.org/10.2307/2412810>
- Felsenstein J. *Inferring Phylogenies*. Sunderland, MA: Sinauer Associates, 2004.
- Field DJ, Benito J, Chen A *et al.* Late Cretaceous neornithine from Europe illuminates the origins of crown birds. *Nature* 2020;**579**:397–401. <https://doi.org/10.1038/s41586-020-2096-0>
- Foote M. Morphological disparity in Ordovician-Devonian crinoids and the early saturation of morphological space. *Paleobiology* 1994;**20**:320–44. <https://doi.org/10.1017/s009483730001280x>
- Gerber S. On the relationship between the macroevolutionary trajectories of morphological integration and morphological disparity. *PLoS One* 2013;**8**:e63913. <https://doi.org/10.1371/journal.pone.0063913>
- Gerber S. Use and misuse of discrete character data for morphospace and disparity analyses. *Palaeontology* 2019;**62**:305–19. <https://doi.org/10.1111/pala.12407>
- Goloboff PA. Homoplasy and the choice among cladograms. *Cladistics* 1991a;**7**:215–32. <https://doi.org/10.1111/j.1096-0031.1991.tb00035.x>
- Goloboff PA. Random data, homoplasy and information. *Cladistics* 1991b;**7**:395–406. <https://doi.org/10.1111/j.1096-0031.1991.tb00046.x>
- Goloboff PA, Farris JS, Nixon KC. TNT, a free program for phylogenetic analysis. *Cladistics* 2008;**24**:774–86. <https://doi.org/10.1111/j.1096-0031.2008.00217.x>
- Guillaume T. dispRity: a modular R package for measuring disparity. *Methods in Ecology and Evolution* 2018;**9**:1755–63.
- Harmon LJ, Schulte JA, Larson A *et al.* Tempo and mode of evolutionary radiation in iguanian lizards. *Science* 2003;**301**:961–4. <https://doi.org/10.1126/science.1084786>
- Harrison LB, Larsson HCE. Among-character rate variation distributions in phylogenetic analysis of discrete morphological characters. *Systematic Biology* 2015;**64**:307–24. <https://doi.org/10.1093/sysbio/syu098>
- Harvey MG, Bravo GA, Claramunt S *et al.* The evolution of a tropical biodiversity hotspot. *Science* 2020;**370**:1343–8.
- Hauser DL, Boyajian G. Proportional change and patterns of homoplasy: Sanderson and Donoghue revisited. *Cladistics* 1997;**13**:97–100. <https://doi.org/10.1111/j.1096-0031.1997.tb00243.x>
- Hillis DM, Huelsenbeck JP. Signal, noise, and reliability in molecular phylogenetic analyses. *The Journal of Heredity* 1992;**83**:189–95. <https://doi.org/10.1093/oxfordjournals.jhered.a111190>
- Hoyal Cuthill J. The morphological state space revisited: what do phylogenetic patterns in homoplasy tell us about the number of possible character states? *Interface Focus* 2015a;**5**:20150049.
- Hoyal Cuthill J. The size of the character state space affects the occurrence and detection of homoplasy: modelling the probability of incompatibility for unordered phylogenetic characters. *Journal of Theoretical Biology* 2015b;**366**:24–32. <https://doi.org/10.1016/j.jtbi.2014.10.033>
- Hoyal Cuthill JF, Lloyd GT. Measuring homoplasy I: comprehensive measures of maximum and minimum cost under parsimony across discrete cost matrix character types. *Cladistics* 2024;**41**:1–27. <https://doi.org/10.1111/cla.12582>
- Huelsenbeck JP. Tree-length distribution skewness: an indicator of phylogenetic information. *Systematic Zoology* 1991;**40**:257–70. <https://doi.org/10.2307/2992321>
- Hughes M. Major evolutionary trends. Unpublished Thesis, University of Bath, 2013.
- Hughes M, Gerber S, Wills MA. Clades reach highest morphological disparity early in their evolution. *Proceedings of the National Academy of Sciences of the United States of America* 2013;**110**:13875–9. <https://doi.org/10.1073/pnas.1302642110>
- Izrailev S. *tictoc: functions for timing R scripts, as well as implementations of 'Stack' and 'Stacklist' structures*. R Package version 1.2, 2024. <https://github.com/jabiru/tictoc>
- Klassen GJ, Mooi RD, Locke A. Consistency indices and random data. *Systematic Biology* 1991;**40**:446–57. <https://doi.org/10.1093/sysbio/40.4.446>
- Kluge AG, Farris JS. Quantitative phyletics and the evolution of anurans. *Systematic Biology* 1969;**18**:1–32. <https://doi.org/10.1093/sysbio/18.1.1>
- Ksepka DT, Grande L, Mayr G. Oldest finch-beaked birds reveal parallel ecological radiations in the earliest evolution of passerines. *Current Biology* 2019;**29**:657–63.e1. <https://doi.org/10.1016/j.cub.2018.12.040>
- Lankster ER. On the use of the term homology in modern zoology, and the distinction between homogenetic and homoplastic agreements. *Annals and Magazine of Natural History* 1870;**6**:34–43. <https://doi.org/10.1080/00222937008696201>
- Laurin M. The evolution of body size, Cope's rule and the origin of amniotes. *Systematic Biology* 2004;**53**:594–622. <https://doi.org/10.1080/10635150490445706>
- Le Quesne WJ. Frequency distributions of lengths of possible networks from a data matrix. *Cladistics* 1989;**5**:395–407.
- Lloyd GT. Estimating morphological diversity and tempo with discrete character-taxon matrices: implementation, challenges, progress, and future directions. *Biological Journal of the Linnean Society* 2016;**118**:131–51. <https://doi.org/10.1111/bij.12746>
- Lloyd GT. Journeys through discrete-character morphospace: synthesizing phylogeny, tempo, and disparity. *Palaeontology* 2018;**61**:637–45. <https://doi.org/10.1111/pala.12380>
- Meier R, Kores P, Darwin S. Homoplasy slope ratio: a better measurement of observed homoplasy in cladistic analyses. *Systematic Zoology* 1991;**40**:74–88. <https://doi.org/10.2307/2992223>
- Mendler K, Chen H, Parks DH *et al.* Annotree: visualization and exploration of a functionally annotated microbial tree of life. *Nucleic Acids Research* 2019;**47**:4442–8. <https://doi.org/10.1093/nar/gkz246>
- Murphy JL, Puttick MN, O'Reilly JE *et al.* Empirical distributions of homoplasy in morphological data. *Palaeontology* 2021;**64**:S05–18. <https://doi.org/10.1111/pala.12535>
- Naylor G, Kraus F. The relationship between *s* and *m* and the retention index. *Systematic Biology* 1995;**44**:559–62. <https://doi.org/10.2307/2413661>
- Oliveros CH, Field DJ, Ksepka DT *et al.* Earth history and the passerine superradiation. *Proceedings of the National Academy of Sciences of the United States of America* 2019;**116**:7916–25. <https://doi.org/10.1073/pnas.1813206116>
- Oyston JW, Hughes M, Wagner PJ *et al.* What limits the morphological disparity of clades? *Interface Focus* 2015;**5**:20150042. <https://doi.org/10.1098/rsfs.2015.0042>
- Oyston JW, Wilkinson M, Ruta M *et al.* Molecular phylogenies map to biogeography better than morphological ones. *Communications Biology* 2022;**5**:521. <https://doi.org/10.1038/s42003-022-03482-x>
- Paradis E, Claude J, Strimmer K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 2004;**20**:289–90.
- Prum RO, Berv JS, Dornburg A *et al.* A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature* 2015;**526**:569–73. <https://doi.org/10.1038/nature15697>

- R Core Team. *R: a Language and Environment for Statistical Computing*. Vienna, Austria: Foundation for Statistical Computing, 2021.
- Revell LJ. phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* 2012;**3**:217–23.
- Sanderson MJ, Donoghue MJ. Patterns of variation in levels of homoplasy. *Evolution* 1989;**43**:1781–95. <https://doi.org/10.1111/j.1558-5646.1989.tb02626.x>
- Schliep KP. phangorn: phylogenetic analysis in R. *Bioinformatics* 2011;**27**:592–3. <https://doi.org/10.1093/bioinformatics/btq706>
- Sidlauskas B. Continuous and arrested morphological diversification in sister clades of characiform fishes: a phylomorphospace approach. *Evolution* 2008;**62**:3135–56. <https://doi.org/10.1111/j.1558-5646.2008.00519.x>
- Simmons MP, Reeves A, Davis JI. Character-state space versus rate of evolution in phylogenetic inference. *Cladistics* 2004;**20**:191–204. <https://doi.org/10.1111/j.1096-0031.2004.00014.x>
- Sookias RB. Exploring the effects of character construction and choice, outgroups and analytical method on phylogenetic inference from discrete characters in extant crocodilians. *Zoological Journal of the Linnean Society* 2020;**189**:670–99. <https://doi.org/10.1093/zoolinnean/zlz015>
- Steell EM, Nguyen JMT, Benson RBJ *et al.* Comparative anatomy of the passerine carpometacarpus helps illuminate the early fossil record of crown Passeriformes. *Journal of Anatomy* 2023;**242**:495–509. <https://doi.org/10.1111/joa.13761>
- Wagner PJ. Exhaustion of morphologic character states among fossil taxa. *Evolution* 2000;**54**:365–86. <https://doi.org/10.1111/j.0014-3820.2000.tb00040.x>
- Wilkinson M. Homoplasy and parsimony analysis. *Systematic Zoology* 1991;**40**:105–9. <https://doi.org/10.2307/2992227>
- Wills MA. Congruence between phylogeny and stratigraphy: randomization tests and the gap excess ratio. *Systematic Biology* 1999;**48**:559–80. <https://doi.org/10.1080/106351599260148>