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

Elizabeth M. Steell[1,2], Allison Y. Hsiang[3], Daniel J. Field[1,4]

[1]Department of Earth Sciences, University of Cambridge, Downing Street, Cambridge, CB2 3EQ, United Kingdom

[2]Girton College, Huntingdon Road, Cambridge, CB3 0JG, United Kingdom

[3]Department of Geological Sciences, Stockholm University, 106 91 Stockholm, Sweden

[4]University Museum of Zoology, University of Cambridge, David Attenborough Building, Downing Place, Cambridge, CB2 3EJ, United Kingdom

Corresponding author: Elizabeth M. Steell

Address: *[1]Department of Earth Sciences, University of Cambridge, Downing Street, Cambridge, CB2 3EQ, United Kingdom*

Email: ems207@cam.ac.uk; lizzysteell@gmail.com

Running headline

Relative homoplasy in phylogenetic datasets

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/S032177/1 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 ÄR-NT 2020-03515 to A.Y.H.

Data availability

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

Conflict of interest

The authors declare no conflict of interest.

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.

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 randomisation 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, Click or tap here to enter text.and is often considered problematic as it can obscure phylogenetic relationships (Kluge and Farris, 1969; Farris, 1989; Sanderson and Donoghue, 1989; Farris, 1991; Goloboff, 1991a; Goloboff, 1991b; 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 *et al.*, 2021; Brocklehurst and Benson, 2021).

Methods to estimate homoplasy received considerable attention in the late 20th century with an extensive body of research exploring parsimony-based methods with cladistic matrices (summarised 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 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., Sanderson and Donoghue, 1989; Archie, 1989; Goloboff, 1991b; Klassen *et al.*, 1991; Meier *et al.*, 1991;

97 Naylor and Kraus, 1995; Archie, 1996; Hauser and Boyajian, 1997; Sookias, 2020; Murphy *et*
98 *al.*, 2021).

99 However, CI and RI are not cross-comparable methods; that is, they cannot be
100 directly compared across alternative datasets and phylogenetic trees, limiting attempts to
101 infer generalisable conclusions about patterns of homoplasy in phylogenetic datasets
102 (Archie, 1989; Meier *et al.,* 1991; Klassen *et al.*, 1991). For instance, CI is highly sensitive to
103 variable numbers of taxa (but not characters) in phylogenetic matrices, with a negative
104 relationship detected between CI and the number of taxa included in a dataset (Archie,
105 1989; Sanderson and Donoghue, 1989; Hauser and Boyajian, 1997; Murphy *et al.*, 2021). By
106 contrast, RI represents a more robust measure with respect to matrix dimensions, although
107 it is nonetheless sensitive to taxon number (Archie, 1989; Naylor and Kraus, 1995; Hauser
108 and Boyajian, 1997; Murphy *et al.*, 2021), and represents a measure of relative homoplasy as
109 opposed to absolute homoplasy, like CI (see below for details) (Farris 1989; Goloboff 1991b).
110 Nonetheless, for the majority of cases, values of RI are not scaled between zero and one if
111 nodes within a tree are partially or fully resolved; instead, the observed minimum value
112 (which indicates the maximum amount of homoplasy possible) is an arbitrary value
113 somewhere within that range, an artefact arising from the theoretical maximum tree length
114 (L_{max}; Fig. 1, Table 1) used when calculating RI (Archie, 1989; Farris, 1989; Goloboff 1991b;
115 Archie, 1996).

116 To mitigate this scaling issue, additional methods incorporating a randomisation step
117 such as the homoplasy excess ratio (HER) (Archie, 1989) and the homoplasy slope ratio (HSR)
118 (Meier *et al.*, 1991) were developed to measure relative homoplasy that can be compared
119 across datasets. However, these methods are limited due to long computation times, and
120 that the method of estimating the maximum value of homoplasy complicates interpretations

121 of relative homoplasy values with the use of particular datasets and topologies (Fig. 1, Table
122 1). Although HER and HSR have been applied in several studies (e.g., Wills, 1999; Hughes,
123 2013; Hoyal Cuthill, 2015b; Oyston *et al.*, 2015, 2022; Mendler *et al.*, 2019; Murphy *et al.*,
124 2021), they are generally challenging to implement and require several time-consuming
125 steps to calculate.

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.	Kluge and Farris (1969)
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.	Farris (1989)
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.	Archie (1989)
Homoplasy slope ratio	$HSR = \frac{L - L_{min}}{{L_{max}}^{r} - L_{min}}$	${L_{max}}^{r}$ = 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.	Meier *et al.* (1991)
Modified homoplasy slope ratio	$HSRm = \frac{L - L_{min}}{{L_{max}}^{r} - L_{min}}$	See HSR formula.	$0 \leq mHSR \leq 1$	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 ≤ *DD* ≤ 1	Negative	L_{all} is not strictly an estimate of maximum possible homoplasy, therefore *DD* does not measure relative homoplasy.	Goloboff (1991a)
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 ≤ *RHI* ≤ 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

126 **Table 1 Summary of homoplasy indices.**

Towards improving the rigour of systematic investigations of patterns of homoplasy, we present a method with a novel randomisation 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 randomisation 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 summarised below, CI is the only index that measures an absolute

151 value of homoplasy (i.e., observed homoplasy that is not compared to an estimated
152 maximum), whereas the other indices presented here measure relative homoplasy (except
153 DD, but see below).
154
155 *Consistency and retention indices*
156 The methodology of the consistency (CI) and retention (RI) indices (Kluge and Farris, 1969;
157 Farris, 1989) are summarised in Supporting Information, and their formulas and terms are
158 summarised in Table 1. See Archie (1996) for further discussion.
159 CI and RI can be expressed with respect to each other:
160
161 $$CI = \frac{L_{min}}{L_{max} - RI(L_{max} - L_{min})}$$
162
163 $$RI = \frac{(CI * L_{max}) - L_{min}}{CI * (L_{max} - L_{min})}$$
164
165 RI is equal to 0 when CI = L_{min}/L_{max}, i.e., when $L = L_{max}$ (Farris 1989). It should be noted that
166 the complement of CI is referred to as the homoplasy index in Kluge and Farris (1969), and
167 the complement of RI is referred to as the distortion index in Farris (1989).
168
169 *Homoplasy excess ratio*
170 The formula format for the homoplasy excess ratio (HER) is identical to RI, but L_{max} is
171 replaced by L_{max}^{p}, which represents the mean tree length of minimum length trees inferred
172 through maximum parsimony for a set of *n* matrices that have been randomised through
173 permutations of the character states within each character for all taxa (Archie, 1989).

174 $$HER = 1 - \frac{L - L_{min}}{{L_{max}}^{p} - L_{min}} = \frac{{L_{max}}^{p} - L}{{L_{max}}^{p} - L_{min}}$$

175

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

185

186 *Data decisiveness*

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

196 $$DD = 1 - \frac{L - L_{min}}{L_{all} - L_{min}} = \frac{L_{all} - L}{L_{all} - L_{min}}$$

where L_{all} is the mean tree length for all possible trees for the dataset (denoted by $\overline{S}$ in Goloboff, 1991a). In theory, L_{all} can be calculated analytically using Theorem 1 from Carter *et al.* (1990) (see 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_{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_{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_{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:

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

where the HS of a set of random matrices is calculated and averaged to obtain HS_{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.* 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:

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

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

241 $$HSR = \frac{L - L_{min}}{L_{random} - L_{min}} = \frac{L - L_{min}}{{L_{max}}^{r} - L_{min}}$$

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

246

247 *Modified homoplasy slope ratio*

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

256

257 *Relative homoplasy index*

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

260 $$RHI = \frac{L - L_{min}}{L_{null} - L_{min}}$$

261 Where L_{null} (null tree length) represents the tree length for a given matrix when no
262 characters are phylogenetically informative with respect to the topology except those that
263 are informative due to chance. L_{null} is calculated by taking the median tree length for the

empirical matrix of a set of *n* randomised trees. In these randomised (null) trees, the topology of the empirical tree (i.e., tree shape, including the root) is maintained, but the tip names are randomised, including outgroups. In this way, observed homoplasy ($L - L_{min}$) is compared to an estimated maximum value of homoplasy ($L_{null} - L_{min}$) 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 (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_{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 randomisation procedure to calculate RHI is unique among homoplasy indices described to date and does not rely on any randomisation of the empirical matrix, unlike the HSR or the HER (Fig. 1), thus avoiding the computationally expensive tree inference step necessitated by each matrix randomisation for HER or time-consuming matrix permutations for both HSR and HER. Further, by only randomising 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 phylogenetic signal of individual characters

288 and provides an efficient way to generate a distribution of trees that have no phylogenetic
289 information.

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

297 It is possible for the value of RHI to exceed 1 under certain conditions (for example, if
298 the empirical tree has a much higher tree length than the minimum length tree). This would
299 happen if the observed homoplasy was higher than that of the estimated maximum
300 homoplasy given the empirical tree shape and empirical matrix, and therefore the matrix
301 and tree combination contains more homoplasy than a matrix and tree with no phylogenetic
302 signal with respect to each other. This is generally unlikely to occur in real phylogenetic
303 datasets, and the aim of introducing RHI is to provide an alternative method to calculating
304 relative homoplasy in empirical datasets. It remains to be shown how limiting this feature of
305 RHI is, and would be best investigated with a combination of theoretical simulated datasets
306 and a meta-analysis of empirical datasets in future work. In datasets that do exceed unity,
307 RHI would not be a useful measure of relative homoplasy and an alternative method, such as
308 the retention index, may be more appropriate. Throughout the analyses in this study, RHI
309 does not exceed unity even at the highest levels of homoplasy in simulated datasets with
310 realistic parameters, whereas HER and HSRm do exceed unity (Table S2). Due to randomising
311 the tips as opposed to the matrix, RHI is unique among homoplasy indices in that its value

312 can be different for two minimum-length trees with different tree shapes inferred from the
313 same matrix. Therefore, one tree shape may have a higher estimated maximum homoplasy
314 value than an alternative tree shape of the same length.

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

321 We provide novel functions in the R programming language (R Core Team, 2021) to
322 carry out all calculations and analyses (see Supporting Information for all R code associated
323 with this work). Our methodology relies on the *phangorn* and *APE* packages (Paradis *et al.*,
324 2004; Schliep, 2011) and includes functions to make morphological matrices compatible for
325 use with *phangorn*. RHI can be calculated for whole matrices, for individual characters, when
326 incorporating specific characters as ordered, or for a distribution of trees (e.g., a set of most
327 parsimonious trees).

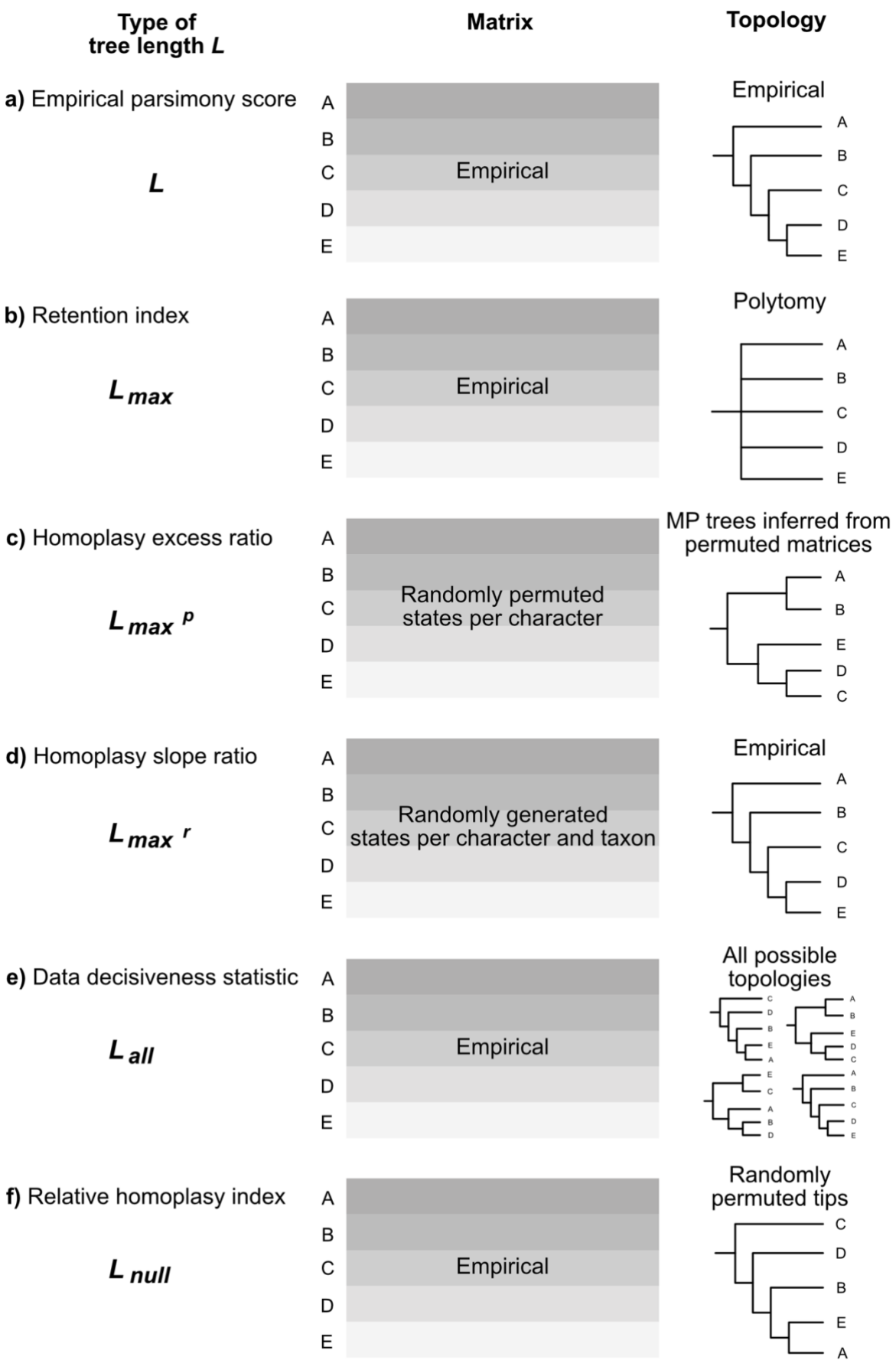

Type of tree length L
Matrix
Topology
a) Empirical parsimony score
L
A
B
C
D
E
Empirical
Empirical
A
B
C
D
E
b) Retention index
L max
A
B
C
D
E
Empirical
Polytomy
A
B
C
D
E
c) Homoplasy excess ratio
L max p
A
B
C
D
E
Randomly permuted states per character
MP trees inferred from permuted matrices
A
B
E
D
C
d) Homoplasy slope ratio
L max r
A
B
C
D
E
Randomly generated states per character and taxon
Empirical
A
B
C
D
E
e) Data decisiveness statistic
L all
A
B
C
D
E
Empirical
All possible topologies
f) Relative homoplasy index
L null
A
B
C
D
E
Empirical
Randomly permuted tips
C
D
B
E
A

Figure 1: Summary of tree length calculations per method. Schematic to summarise randomisation 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).

Materials 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.* and Steell *et al.* were formatted so that two different treatments of the matrix and/or empirical tree could be analysed. For Benito *et al.*, 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.*, two datasets were analysed with the same characters but 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, utilising empirical matrices, randomly generated matrices and matrices simulated under an equal rates model of character state evolution. All analyses were carried out in R v4.3.2 (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).

Analysis	Independent variable (X)	X values	Rate categories (μ)	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 (μ)	-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
		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
4. Variable number of characters	Number of characters	Avialae (a): 10:280 by 30	-6, -4, -2	1	100	10	1000
5. Principal coordinates in phylomorphospace	Mean log (μ)	-6:0 by 2	N/A	1	1	4	4

Table 2: Summary of analyses.

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 randomisations (*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* (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 'pratchet' 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 homoplasy in worst-case scenario, high-homoplastic matrices with no missing data or polymorphisms.

Run time (seconds)

Dataset	N taxa	N characters	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

405 **Table 3: Summary of run times of each index per empirical dataset.** For methods involving
406 a randomisation step, *n* = 100 unless stated otherwise.

407

408 *Simulated matrices (Analyses 2-4)*

409 Analyses 2–4 investigated patterns of homoplasy in simulated matrices for topologies of
410 Telluraves, Passeriformes *a*, Neornithes and Avialae *a*; see Supporting Information for details
411 on the simulation protocol. Due to the substantial run times for the HER and HSR (Table 3),
412 we decided not to include these indices in our simulations due to the expected
413 computational time being orders of magnitude longer than when calculating CI, RI and RHI
414 for the 800–2,100 simulated matrices per topology per analysis (Table 2). We also believe
415 Analysis 1 is sufficient to determine the efficacy of each method at detecting homoplasy in a
416 range of datasets, and instead Analyses 2–4 focus on the behaviour of homoplasy in discrete
417 matrices with changing variables.

418

419 *Character state transition rates (Analysis 2)*

420 We investigated homoplasy across different character state transition rates by varying mean
421 log (μ) values when generating simulated datasets. We used the rate of character state
422 change (hereon referred to as transition rate) as a proxy for homoplasy in simulated
423 matrices. Higher transition rates lead to increased homoplasy because a greater number of
424 character state transitions may result in independent acquisitions of or reversals to the same
425 state (Simmons *et al.*, 2004; Harrison and Larsson, 2015). However, the efficacy of increasing
426 transition rates as a proxy for homoplasy will decrease with increased numbers of character

427 states present in a matrix (i.e., more multistate characters with more states per character)
428 and may produce unexpected effects on homoplasy detection in datasets with many
429 multistate characters (Hoyal Cuthill, 2015b). Moreover, the assumption of homoplasy
430 increasing in tandem with transition rates may not be valid if transitions between states do
431 not occur with equal probability. Therefore, all our simulations assume a conservative 9:1
432 ratio of binary to three-state characters and equal probabilities of character state change
433 directionality. We varied transition rates by employing a range of values for the mean log (μ)
434 transition rate and simulated 100 matrices per rate category (Table 2). The range of values (-
435 8 to 2) for μ was selected based on trialling multiple scenarios; the selected range was
436 appropriate for the empirical datasets investigated here, as it covers very low transition rates
437 up to very high rates. The median, 5% and 95% quantiles of RHI, 1 – CI and 1 – RI were
438 plotted against mean log transition rate for the Telluraves, Passeriformes *a*, Neornithes and
439 Avialae *a* topologies. We also normalised median treelength (*L*) between 0 and 1 with the
440 following formula:

441 $$\frac{L - min(L)}{max(L) - min(L)}$$

442 where min(*L*) and max(*L*) represent the minimum and maximum treelengths from the
443 simulation.

444

445 *Number of taxa (Analysis 3)*

446 Homoplasy across datasets with varying numbers of taxa was investigated by sampling tips
447 from each topology without replacement in decreasing increments (Table 2). We sampled
448 100 reduced-taxa trees per taxon-number increment and simulated one matrix per reduced-
449 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: μ = -6 (slow), μ = -4 (medium) and μ = -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 (μ = -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 Table S2 (Supporting Information).

Results

We quantified homoplasy in empirical and randomly generated datasets across a range of methods (Analysis 1, Table 2). Results are summarised in Table 4 and Figure 2 (note that Fig.

2 shows CI, RI and HER values as one 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 are identical to two decimal points or near identical across all datasets.

Dataset	Input tree	'?' (%)	Tree length	1 - CI	1 - RI	1 - HER	HSR	HSR m	RHI
Empirical									
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
Random									
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

Table 4: Homoplasy index results from Analysis 1 (homoplasy across empirical and random matrices).

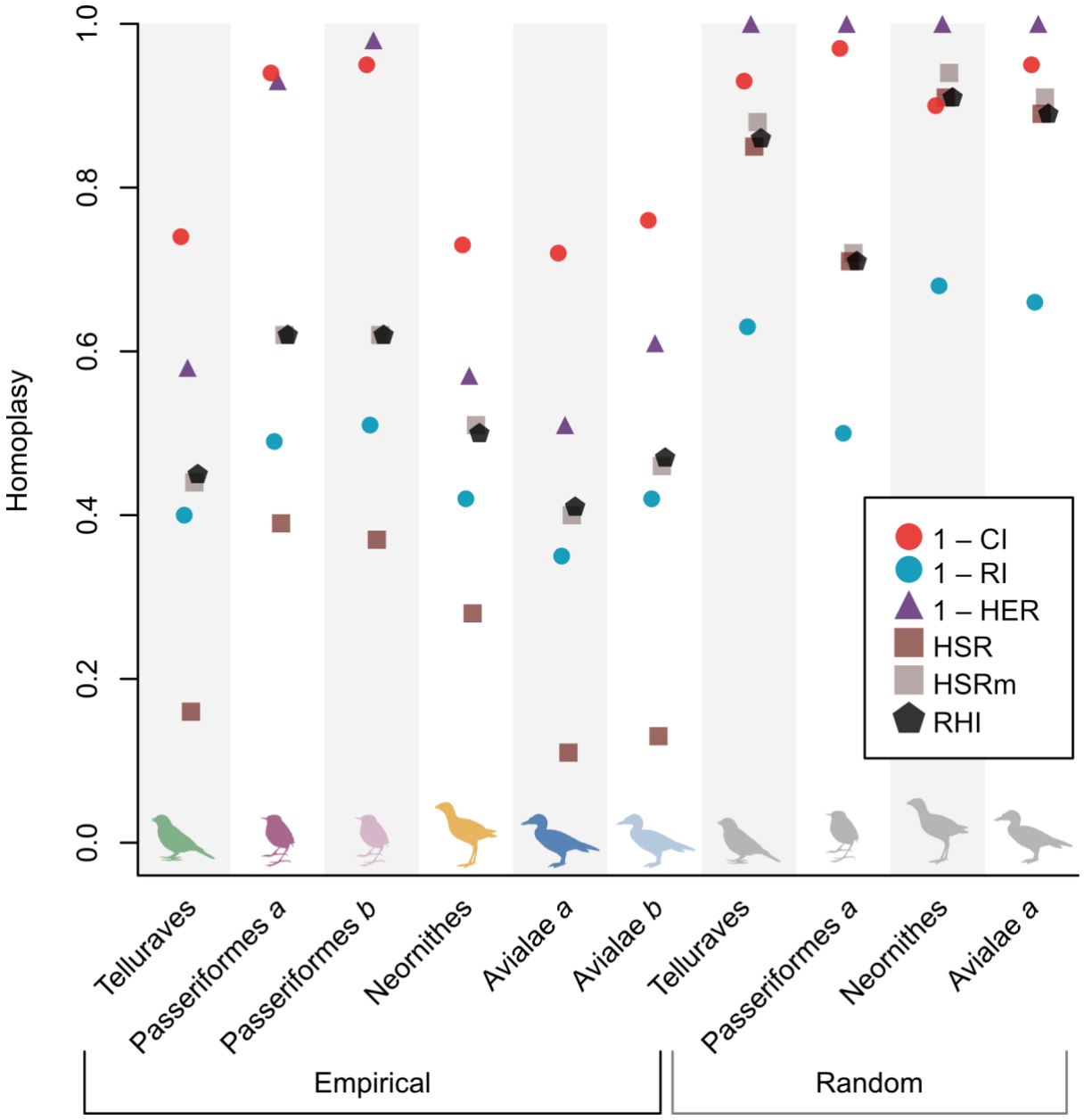
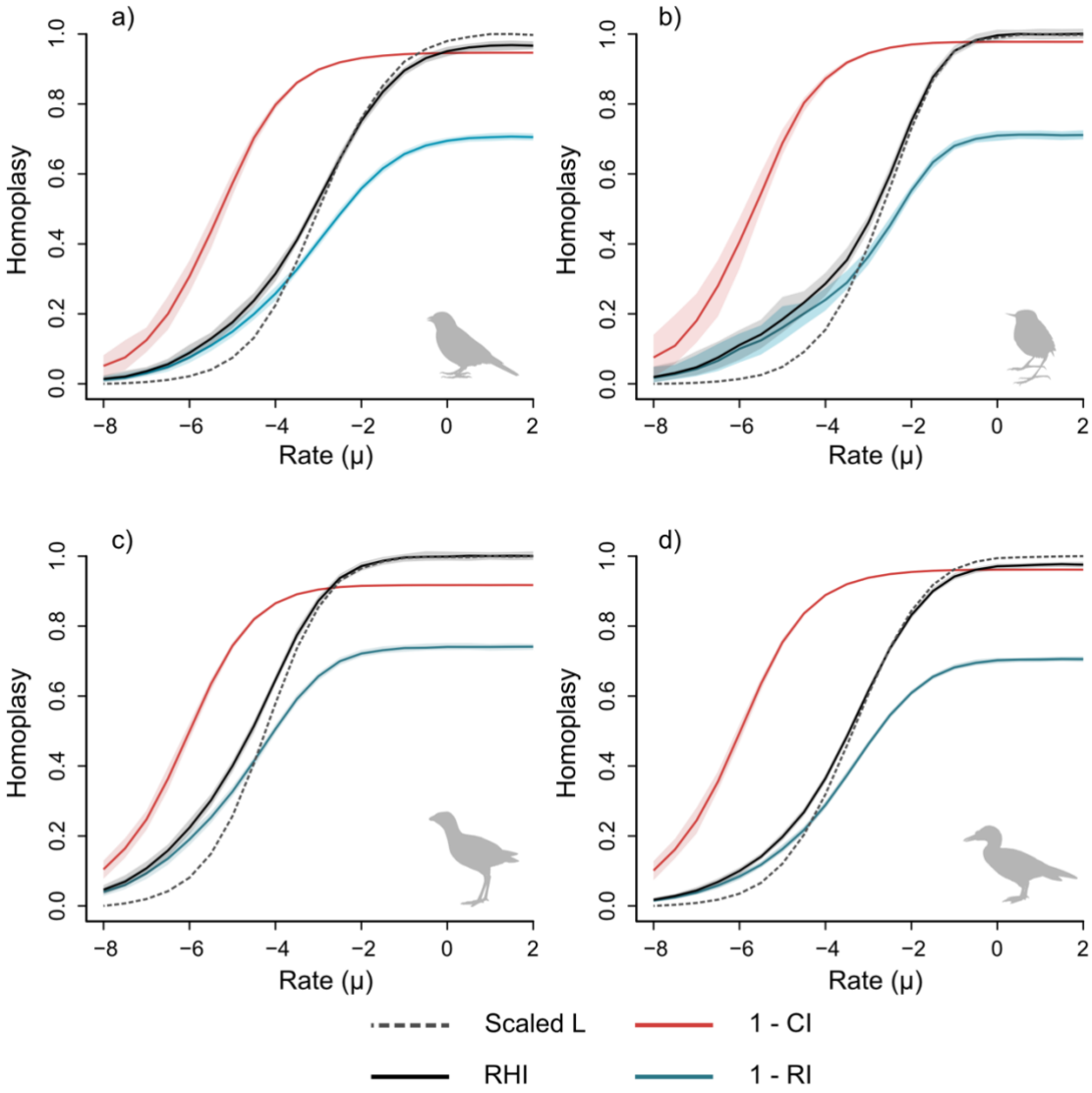


Figure 2: Homoplasy indices for empirical and random matrices (Analysis 1). Homoplasy values (Table 4) for consistency index (CI), retention index (RI), homoplasy excess ratio (HER), homoplasy slope ratio (HSR), modified homoplasy slope ratio (HSRm) and relative homoplasy index (RHI) are plotted for 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 homoplasy excess ratio are subtracted

494 from 1 for ease of comparison to other methods. A value of 1 indicates the maximum
495 amount of detectable homoplasy for each method.
496
497 We varied the rate of character state change as a proxy for varying levels of
498 homoplasy in simulated datasets in Analysis 2. Results for CI, RI and RHI calculations as well
499 as normalised tree length (*L**) across these simulations are summarised in Figure 3. The
500 results indicate that as transition rate increases (by increasing mean log μ), homoplasy
501 increases (Fig. 3), although the extent of homoplasy detected is dependent on the index
502 applied. Normalised tree length for each dataset shows that at approximately μ = -2, tree
503 length reaches its maximum value for each dataset (Fig. 3). The CI rapidly indicates higher
504 levels of absolute homoplasy with a small increase in transition rate, and subsequently
505 asymptotes at a consistently lower μ value than the other indices. Ranges of Y-axis values are
506 reported in Table 5, as well as area under the curve values, the difference between areas
507 under the *L** curve and areas under curves for 1 – CI, 1 – RI and RHI. RHI consistently yields
508 the greatest Y-axis range across each dataset, and RI consistently yields the smallest range of
509 homoplasy values (i.e., 1 – RI always asymptotes at a homoplasy value lower than one).
510 Normalised tree length, 1 – RI and RHI all asymptote at approximately the same μ values for
511 each dataset. The area difference between RHI and *L** for each dataset is the smallest across
512 the three methods (Table 5), which can be seen in Fig. 3 in which the RHI curve closely tracks
513 the *L** curve. The Y-axis range difference for each index is reported in Table S1 (Supporting
514 Information).

515



516

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 normalised between 0 and 1; grey dotted line), when the mean log for character state transition rate (μ) 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.

517

518

	Y-axis range				Area under curve				Area difference from L*		
Topology	***L*†***	**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

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

520 † *L** denotes normalised tree length.

521

522 We tested the effects of varying numbers of taxa and characters using simulated
523 matrices in Analyses 3 and 4, respectively. Results for Analysis 3 are summarised in Figure 4
524 and show that as taxa are added to datasets (i.e., as tips are added to a phylogeny), the
525 levels of homoplasy detected by CI, RI and RHI change. CI detects more absolute homoplasy
526 as number of taxa is increased, and RI and RHI both detect less relative homoplasy as
527 number of taxa is increased. These effects are also variable depending on the underlying
528 transition rate of character state change (Fig. 4, Table S1); a greater difference in detected
529 homoplasy is observed for a mid to high transition rate (μ = -4) than for either the low- or
530 high-rate categories (μ = -6 and μ = -2, respectively). Results for Analysis 4 are summarised in
531 Figure 5 and show that as characters are added to a matrix, no change in levels of
532 homoplasy is detected across all methods tested (Table S1). We find that in small datasets
533 (either low numbers of taxa or low numbers of characters), the variance in detected
534 homoplasy is greater (Figs. 4 and 5).

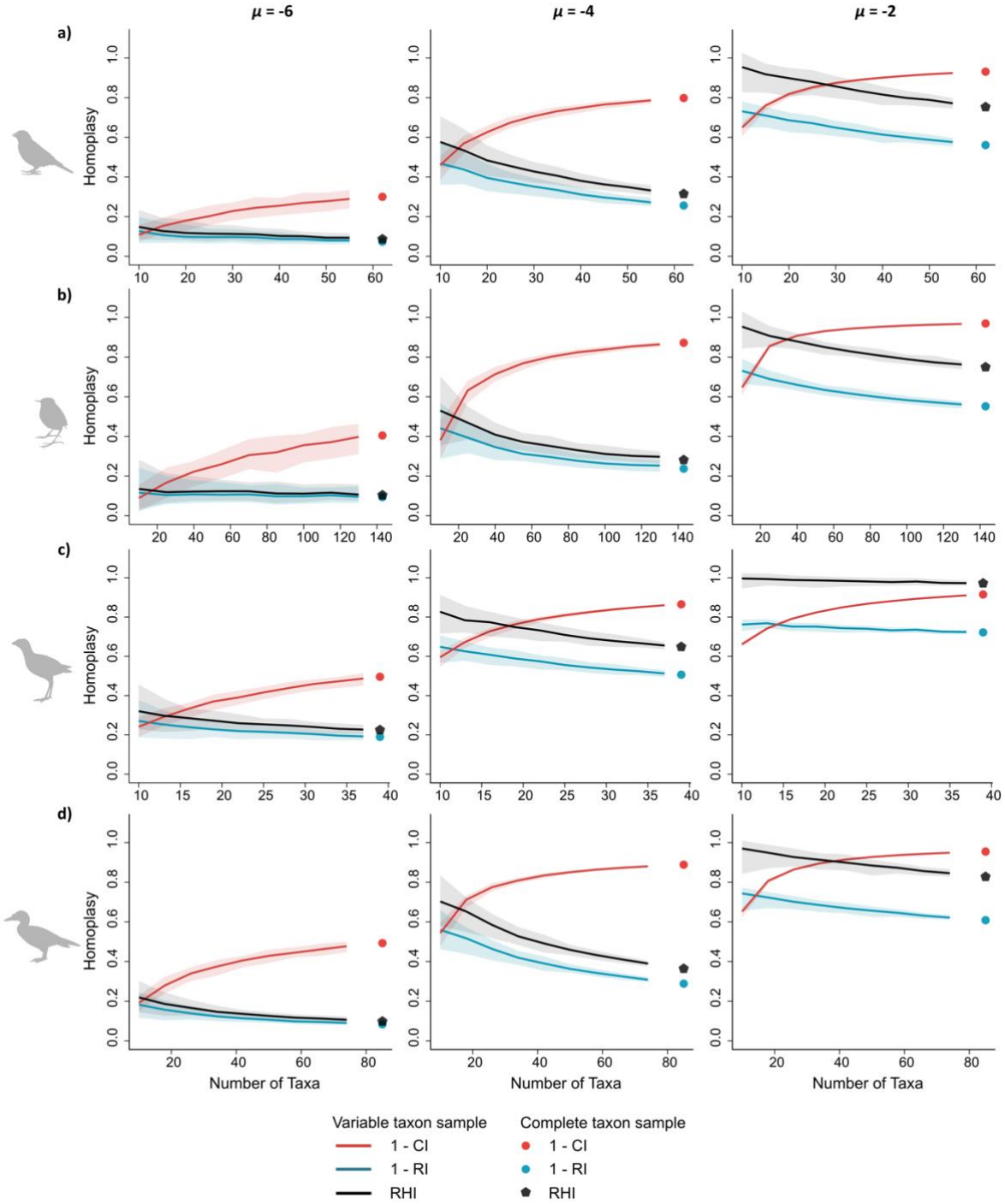


Figure 4: Levels of homoplasy when varying numbers of taxa in simulated matrices (Analysis 3). Results for varying number of taxa in simulated matrices at low (μ = -6), medium (μ = -4) and high (μ = -2) rates of character state change for the Telluraves (A),

Passeriformes *α* (B), Neornithes (C) and Avialae *α* (D) topologies. 1 – CI in red, 1 – RI in blue and RHI in black with 95% and 5% quantile ranges as envelopes. Points represent index values with the full set of tips. Y axis ranges are reported in Table S1.

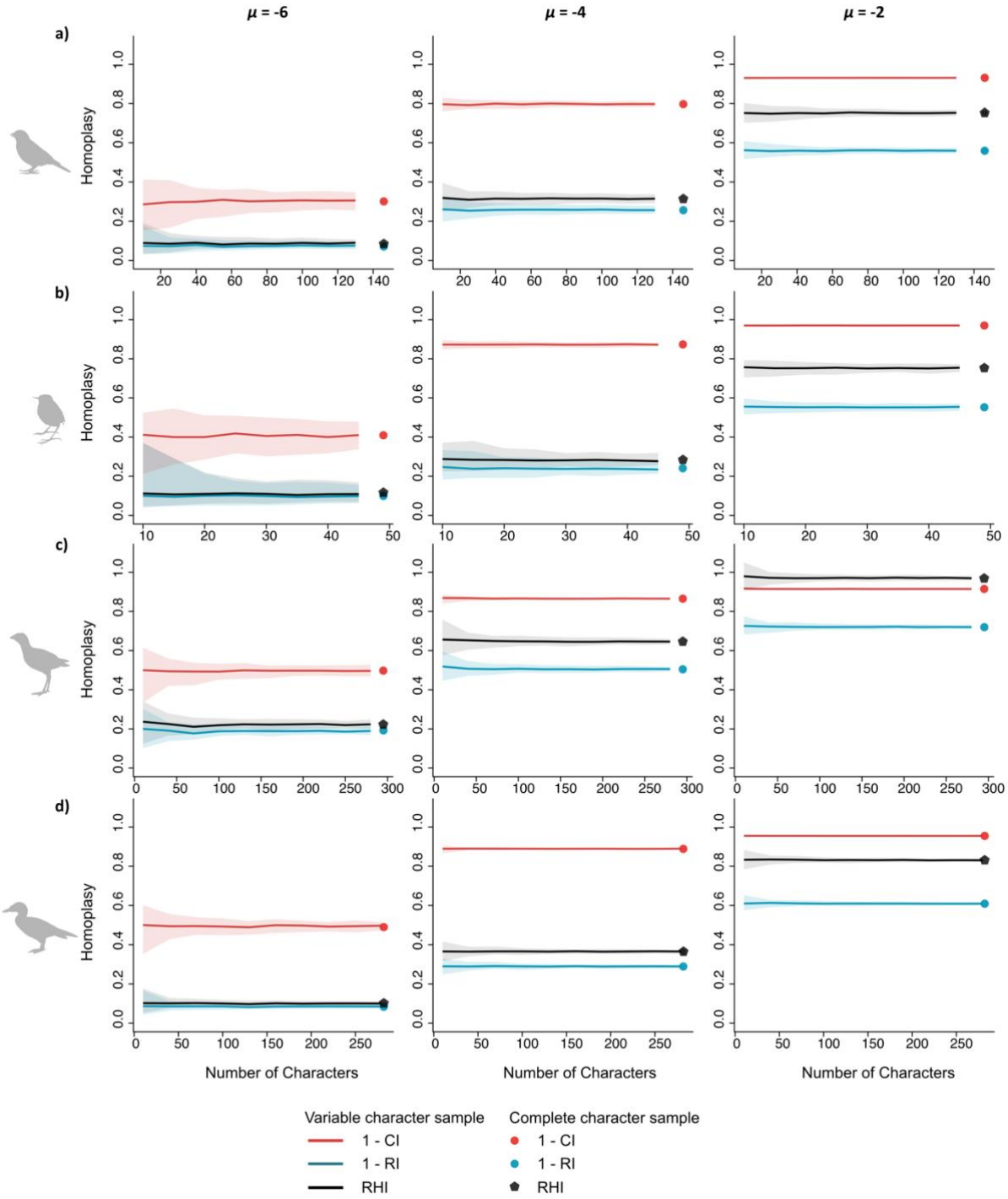


Figure 5: Levels of homoplasy when varying numbers of characters in simulated matrices (Analysis 4). Results for varying number of characters in simulated matrices at low (μ = -6), medium (μ = -4) and high (μ = -2) rates of character state change for the Telluraves (A), Passeriformes *a* (B), Neornithes (C) and Avialae *a* (D) topologies. 1 – CI in red, 1 – RI in blue and RHI in black with 95% and 5% quantile ranges as envelopes. Points represent index values with the full set of characters. Y axis ranges are reported in Table S1.

Finally, we generated phylomorphospace plots (Sidlauskas, 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 show distinct morphospace partitioning between clades, with a trend towards more overlap between clades as rates are increased. Visually, internal branches at low transition rates are longer than terminal branches, whereas at high transition rates (and therefore high homoplasy), branch lengths across the whole tree appear equivalent, or with longer terminal branches. The amount of partitioning and overlap in morphospace depends to some extent on properties of the tree and character matrix. For instance, the dataset with the most taxa (Passeriformes *a*) shows less partitioning of morphospace even at very low levels of homoplasy (RHI = 0.08; Fig. 6b, Table S2) compared to Telluraves or Avialae *a* at equivalent homoplasy levels, whereas the dataset with the fewest taxa (Neornithes) shows substantial partitioning even at more moderate levels of homoplasy (RHI = 0.25; Fig. 6c, Table S2). There is 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, Table S2).

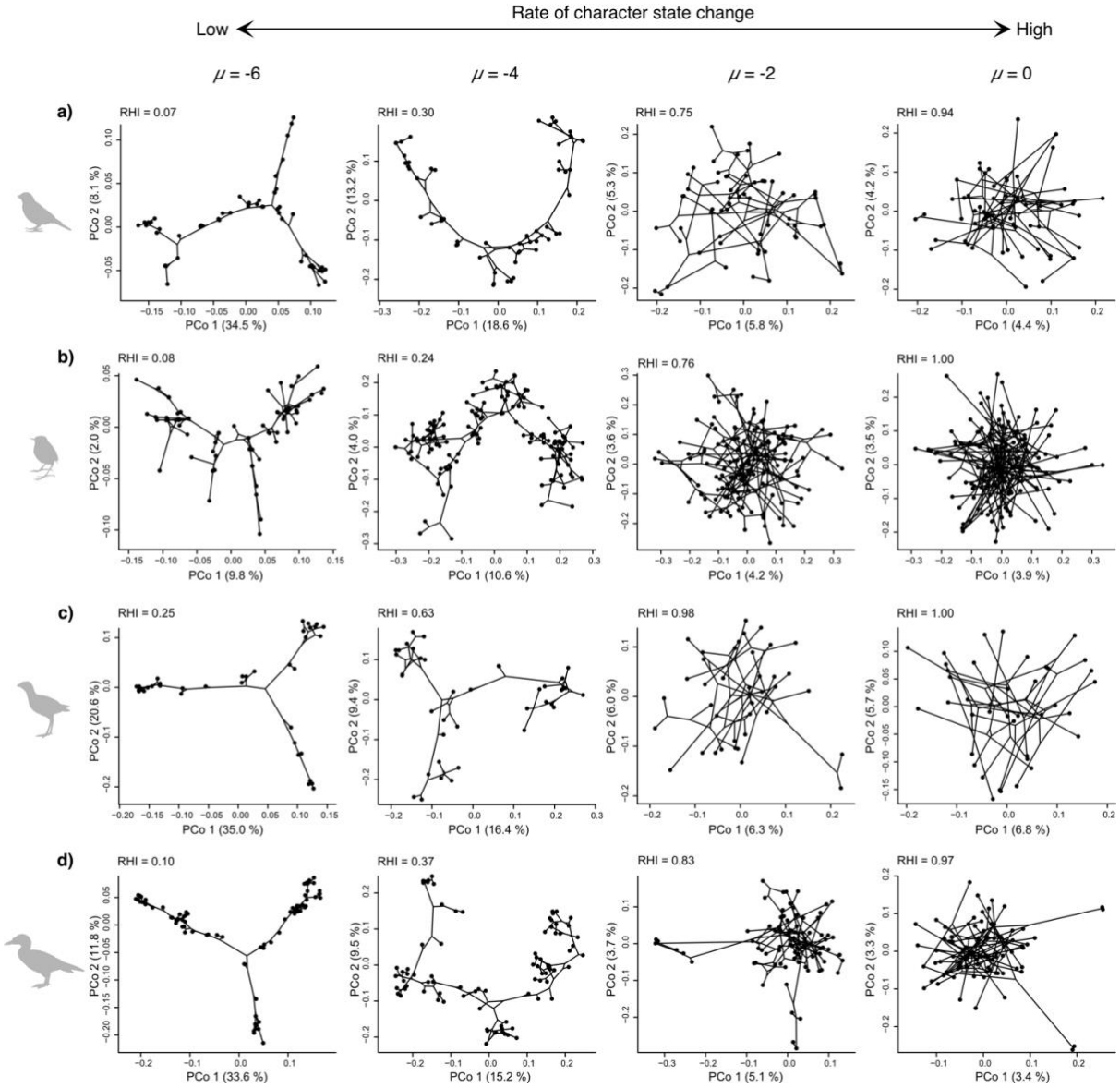


Figure 6: Phylomorphospace plots of simulated matrices at different character state transition rates from Analysis 5. Axes 1 and 2 from each Principal Coordinate (PCo) analysis for simulated matrices based on the empirical topologies for Telluraves (A), Passeriformes *a* (B), Neornithes (C) and Avialae *a* (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.

Discussion

We investigated the behaviour of homoplasy in discrete matrices by testing different methods with empirical, random, and simulated matrices (Tables 1 and 2). In general, levels of relative homoplasy 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 (Fig. 2, Table 4), hampering meaningful comparisons among different datasets. The homoplasy excess ratio (HER), first developed to overcome the scaling issue with CI (Archie 1989; Archie 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 homoplasy slope ratio (HSR) is also not appropriate with datasets that contain missing data or other non-state scorings due to its randomisation step (Fig. 1). The modified homoplasy slope ratio (HSRm, described here) proves to be more appropriate for empirical datasets. The relative homoplasy index (RHI), our new method with a novel randomisation step (Fig. 1), overcomes these issues by being scaled between zero and one for realistic datasets and maintaining the characteristics unique to the input matrix when estimating maximum homoplasy, 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 randomisation 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 ~16-48x.

Comparison of methods

The major limitation of the RI is that it underestimates the proportional extent of homoplasy 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 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 (2015) 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 (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 (μ = 0), suggesting that

623 tree length for the input tree and simulated matrix far exceeds the average tree length for
624 permuted matrices in the HER randomisation step.
625 We found that HSR greatly underestimates relative homoplasy in empirical matrices
626 due to the randomly generated matrices not being an appropriate null model to estimate
627 maximum homoplasy, but that HSRm does appear to be a suitable alternative when all
628 character scorings are taken into consideration. Although not identical, the matrix
629 randomisation protocol in HSRm (Fig. 1) is almost numerically equivalent to the tip
630 randomisation in RHI, and the two indices consistently yield very similar values (Table 4,
631 Table S2). However, generating randomised matrices is more computationally demanding
632 than randomising tips on a tree, so we recommend the use of the more efficient RHI method
633 (Table 3).
634
635 *Homoplasy in discrete matrices*
636 Analysis 3 shows that, with the addition of taxa, relative homoplasy detected by RHI and RI
637 decreases, whereas absolute homoplasy detected by CI increases considerably (Fig. 4). This
638 pattern in CI mirrors early work investigating homoplasy in discrete matrices, which
639 concluded that adding more taxa into a dataset would increase estimated levels of
640 homoplasy (Sanderson and Donoghue 1989). This longstanding hypothesis has been
641 corroborated by numerous subsequent studies (Meier *et al.*, 1991; Hauser and Boyajian,
642 1997; Murphy *et al.*, 2021), and is explained by the increased detection of individual
643 homoplastic character state changes with the addition of taxa and/or characters, thereby
644 resulting in more absolute homoplasy (Sanderson and Donoghue, 1989; Hauser and
645 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)

669 due to the exponentially increasing number of possible arrangements of tips as taxa are
670 added (Felsenstein, 1978) and the increasing number of possible character state
671 combinations available in character state space (Foote, 1994; Hoyal Cuthill, 2015a).
672 However, provided the rate distribution of character state transitions for the matrix is
673 maintained, there is no reason to expect more relative homoplasy in a dataset with more
674 taxa. On the contrary, observed homoplasy values will be proportionally lower compared to
675 the increasing potential homoplasy of the dataset as a taxon sample increases.
676 We find that CI, RI and RHI do not vary with increasing numbers of characters (Fig. 5).
677 For CI, provided the rate distribution of character state transitions is the same regardless of
678 number of characters, L and L_{min} (and therefore H) increase proportionally as characters are
679 added. For RI and RHI, again assuming transition rates are maintained, L and L_{max} and L and
680 L_{null}, respectively, also increase proportionally as characters are added.
681 Previous studies investigating patterns of detected homoplasy with CI, RI and
682 additional parsimony indices such as the HER (Archie, 1989) and the HSR (Meier *et al.*, 1991)
683 report conflicting results. Archie (1989) also observed that RI (therein referred to as the
684 'homoplasy excess ratio maximum') detects less relative homoplasy in subsets of datasets
685 with more taxa, although a slight increase in relative homoplasy with the addition of
686 characters is observed (Archie, 1989). Other studies using multiple non-subsetted empirical
687 datasets report mixed results regarding RI. For example, Hauser and Boyajian (1997) found a
688 slight decrease in relative homoplasy with increasing taxa in their meta-analysis of 600
689 datasets, but Murphy *et al.* (2021) showed the opposite trend in an analysis of 364 datasets.
690 These variable patterns are likely explained by idiosyncrasies specific to the datasets
691 analysed; each dataset exhibits a particular amount of estimated homoplasy dependent on
692 factors such as tree shape, character state space (Simmons *et al.*, 2004) and biological

693 drivers of homoplasy. As our results show, larger datasets do not show more relative
694 homoplasy based on their dimensions alone.
695 Tree shape is expected to impact patterns of homoplasy estimation (Simmons *et al.*,
696 2004), which is why an advantage of the RHI is the novel tip randomisation process. We
697 randomised tips to retain the overall tree shape of each empirical tree, rather than sampling
698 randomly generated trees that may come from a distribution of variable tree shapes with
699 unequal frequencies (Felsenstein, 2004). Simmons *et al.* (2004) explored the effect of tree
700 shape on RI and CI with simulated datasets and found that with increasing transition rate,
701 homoplasy detected by RI is affected substantially by tree shape. At low transition rates,
702 levels of homoplasy increase significantly more for a completely asymmetrical, unequal
703 branch-length tree than they do for a completely symmetrical, equal branch-length tree
704 (Simmons *et al.*, 2004). We expect that tree shape affects relative homoplasy measured by
705 RHI in cases with multiple most-parsimonious trees for the same dataset. Trees of different
706 shapes but the same length may have different RHI values, as we expect that some tree
707 shapes are associated with a greater amount of possible homoplasy than others, which
708 warrants further exploration in future work. Investigating homoplasy patterns with random
709 trees generated by different methods, for example those implemented in *APE* (Paradis *et al.*,
710 2004), was beyond the scope of the current study but could be useful for the development
711 of new, non-parsimony-based methods for homoplasy quantification.
712
713 *Using homoplasy in macroevolutionary studies*
714 Estimating relative homoplasy for discrete matrices is not only useful for estimating the
715 overall phylogenetic informativeness of characters but has wider implications when
716 considering macroevolutionary patterns of morphological disparity and evolution. Although

717 some phylogenetic matrices may not be an ideal source of data for disparity analysis (Lloyd,
718 2016), particularly at very broad taxonomic scales, many datasets provide sufficient
719 information to explore morphological evolution with discrete data (Gerber, 2019).
720 Quantifying homoplasy in discrete datasets may therefore become more commonplace for
721 evaluating macroevolutionary patterns across different clades, particularly if this is
722 combined with additional analyses investigating disparity and character state space
723 occupation (Oyston *et al.*, 2015; Lloyd, 2016, 2018; Gerber, 2019).
724 Our phylomorphospace results from simulated matrices show decreasing
725 morphospace partitioning and longer terminal branch lengths for principal coordinate (PCo)
726 axes 1 and 2 as character state transition rates (and therefore absolute and relative
727 homoplasy) increase (Fig. 6, Table S2). Our results appear to be analogous to snapshots of
728 clade evolution through time; high morphospace partitioning and low homoplasy are
729 consistent with the idea that clades explore morphospace early in their evolutionary history
730 and, as a result, disparity rapidly increases as taxa occupy new combinations of character
731 states (Wagner, 2000; Harmon *et al.*, 2003; Hughes *et al.*, 2013; Oyston *et al.*, 2015).
732 Similarly, higher levels of homoplasy corresponding with weak morphospace partitioning are
733 consistent with character state exhaustion that is expected to occur later in the evolutionary
734 history of a clade when biological constraints are acting on morphological evolution,
735 resulting in convergent character state combinations (Wagner, 2000; Oyston *et al.*, 2015;
736 Brocklehurst and Benson, 2021). We also see that increasing relative homoplasy is
737 associated with a general trend of decreasing amounts of variation explained per principal
738 coordinate axis (Fig. 6, Table S2). This is potentially explained by varying levels of integration
739 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 multidimensional character state space (Gerber, 2019). Therefore, it is increasingly recommended to exercise caution when using these techniques, particularly as the tools to visualise 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.

References

Archie J.W. Homoplasy excess ratios: new indices for measuring levels of homoplasy in phylogenetic systematics and a critique of the consistency index. *Systematic Zoology* 1989; **38:** 253–269.

Archie J.W. Measures of homoplasy. In: Sanderson M.J., Hufford L., eds. *Homoplasy: the recurrence of similarity in evolution*. San Diego, California: Academic Press, 1996; 153–188.

Benito J., Kuo P.C., Widrig K.E., *et al.* Cretaceous ornithurine supports a neognathous crown bird ancestor. *Nature* 2022; **612:** 100–105.

763 Brocklehurst N., Benson R.J. Multiple paths to morphological diversification during the origin
764 of amniotes. *Nature Ecology and Evolution* 2021; **5:** 1243–1249.
765 Brocklehurst N., Panciroli E., Benevento G.L., *et al.* Mammaliaform extinctions as a driver of
766 the morphological radiation of Cenozoic mammals. *Current Biology* 2021; **31:** 2955-2963.
767 Carter M., Hendy M., Penny D., *et al.* On the distribution of lengths of evolutionary trees.
768 *SIAM Journal on Discrete Mathematics* 1990; **3:** 38-47.
769 Davis J.I., Simmons M.P., Stevenson D.W., *et al.* Data decisiveness, data quality, and
770 incongruence in phylogenetic analysis: an example from the monocotyledons using
771 mitochondrial atp A sequences. *Systematic Biology* 1998; **47:** 282-310.
772 Farris J.S. The retention index and the rescaled consistency index. *Cladistics* 1989; **5:** 417–
773 419.
774 Farris J.S. Excess homoplasy ratios. *Cladistics* 1991; **7:** 81–91.
775 Felsenstein J. The number of evolutionary trees. *Systematic Zoology* 1978; **27:** 27–33.
776 Felsenstein J. *Inferring Phylogenies*. Sunderland, Massachusetts: Sinauer Associates, 2004.
777 Field D.J., Benito J., Chen A., *et al.* Late Cretaceous neornithine from Europe illuminates the
778 origins of crown birds. *Nature* 2020; **579:** 397–401.
779 Foote M. Morphological disparity in Ordovician-Devonian crinoids and the early saturation
780 of morphological space. *Paleobiology* 1994; **20:** 320–344.
781 Gerber S. On the relationship between the macroevolutionary trajectories of morphological
782 integration and morphological disparity. *PLoS ONE* 2013; **8:** e63913.
783 Gerber S. Use and misuse of discrete character data for morphospace and disparity analyses.
784 *Palaeontology* 2019; **62:** 305–319.
785 Goloboff P.A., Farris J.S., Nixon K.C. TNT, a free program for phylogenetic analysis. *Cladistics*
786 2008; **24:** 774–786.

787 Goloboff P.A. Homoplasy and the choice among cladograms. *Cladistics* 1991a; **7:** 215-232.
788 Goloboff P.A. Random data, homoplasy and information. *Cladistics* 1991b; **7:** 395-406.
789 Guillerme T. dispRity: a modular R package for measuring disparity. *Methods in Ecology and*
790 *Evolution* 2018; **9:** 1755–1763.
791 Harmon L.J., Schulte, J.A., Larson, A., *et al.* Tempo and mode of evolutionary radiation in
792 iguanian lizards. *Science* 2003; **301:** 961–965.
793 Harrison L.B., Larsson H.C.E. Among-character rate variation distributions in phylogenetic
794 analysis of discrete morphological characters. *Systematic Biology* 2015; **64:** 307–324.
795 Harvey M.G., Bravo G.A., Claramunt S., *et al.* The evolution of a tropical biodiversity hotspot.
796 *Science* 2020; **370:** 1343–1348.
797 Hauser D.L., Boyajian G. Proportional change and patterns of homoplasy: Sanderson and
798 Donoghue revisited. *Cladistics* 1997; **13:** 97–100.
799 Hillis, D.M., Huelsenbeck, J.P. Signal, noise, and reliability in molecular phylogenetic analyses.
800 *Journal of Heredity* 1992; **83:** 189-195.
801 Hoyal Cuthill J. The morphological state space revisited: what do phylogenetic patterns in
802 homoplasy tell us about the number of possible character states? *Interface Focus* 2015a; **5:**
803 20150049.
804 Hoyal Cuthill J. The size of the character state space affects the occurrence and detection of
805 homoplasy: modelling the probability of incompatibility for unordered phylogenetic
806 characters. *Journal of Theoretical Biology* 2015b; **366:** 24–32.
807 Hoyal Cuthill J.F., Lloyd G.T. Measuring homoplasy I: comprehensive measures of maximum
808 and minimum cost under parsimony across discrete cost matrix character types. *Cladistics*
809 2024; **0:** 1–27.

Huelsenbeck, J.P. Tree-length distribution skewness: an indicator of phylogenetic information. *Systematic Zoology* 1991: **40:** 257–270.

Hughes M. Major evolutionary trends. Unpublished thesis, University of Bath 2013.

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

Izrailev S. tictoc: functions for timing R scripts, as well as implementations of 'Stack' and 'Stacklist' structures. R Package version 1.2 2024.

Klassen G.J., Mooi R.D., Locke A. Consistency indices and random data. *Systematic Biology,* 1991; **40:** 446–457.

Kluge A.G., Farris J.S. Quantitative phyletics and the evolution of anurans. *Systematic Biology* 1969; **18:** 1–32.

Ksepka D.T., Grande L., Mayr G. Oldest Finch-Beaked Birds Reveal Parallel Ecological Radiations in the Earliest Evolution of Passerines. *Current Biology* 2019; **29:** 657-663.

Lankester E.R. 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.

Laurin, M. The evolution of body size, Cope's rule and the origin of amniotes. *Systematic Biology* 2004; **53:** 594–622.

Le Quesne, W.J. Frequency distributions of lengths of possible networks from a data matrix. *Cladistics* 1989; **5:** 395–407.

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

834 Lloyd G.T. Journeys through discrete-character morphospace: synthesizing phylogeny, tempo,
835 and disparity. *Palaeontology* 2018; **61:** 637–645.
836 Meier R., Kores P., Darwin S. Homoplasy slope ratio: a better measurement of observed
837 homoplasy in cladistic analyses. *Systematic Zoology* 1991; **40:** 74–88.
838 Mendler K., Chen H., Parks D.H., *et al.* Annotree: visualization and exploration of a
839 functionally annotated microbial tree of life. *Nucleic Acids Research* 2019; **47:** 4442–4448.
840 Murphy J.L., Puttick M.N., O'Reilly J.E., *et al.* Empirical distributions of homoplasy in
841 morphological data. *Palaeontology* 2021; **64:** 505–518.
842 Naylor, G. and Kraus, F., The relationship between *s* and *m* and the retention index.
843 *Systematic biology* 1995; **44:** 559–562.
844 Oliveros C.H., Field D.J., Ksepka D.T., *et al.* Earth history and the passerine superradiation.
845 *Proceedings of the National Academy of Sciences of the United States of America* 2019; **116:**
846 7916–7925.
847 Oyston J.W., Hughes M., Wagner P.J., *et al.* What limits the morphological disparity of
848 clades? *Interface Focus* 2015; **5:** 20150042.
849 Oyston J.W., Wilkinson M., Ruta M., *et al.* Molecular phylogenies map to biogeography
850 better than morphological ones. *Communications Biology* 2022; **5:** 521.
851 Paradis E., Claude J, Strimmer K. APE: analyses of phylogenetics and evolution in R language.
852 *Bioinformatics* 2004; **20:** 289–290.
853 Prum R.O., Berv J.S., Dornburg A., *et al*. A comprehensive phylogeny of birds (Aves) using
854 targeted next-generation DNA sequencing. *Nature* 2015; **526:** 569–573.
855 R Core Team. R: a language and environment for statistical computing. *Foundation for*
856 *Statistical Computing, Vienna, Austria* 2021.

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

Sanderson M.J., Donoghue M.J. Patterns of Variation in Levels of Homoplasy. *Evolution* 1989; **43:** 1781–1795.

Schliep K.P. phangorn: phylogenetic analysis in R. *Bioinformatics* 2011; **27:** 592–593.

Sidlauskas B. Continuous and arrested morphological diversification in sister clades of characiform fishes: a phylomorphospace approach. *Evolution* 2008; **62:** 3135–3156.

Simmons M.P., Reeves A., Davis J.I. Character-state space versus rate of evolution in phylogenetic inference. *Cladistics* 2004; **20:** 191–204.

Sookias R.B. 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:** 670-699.

Steell E.M., Nguyen J.M.T., Benson R.B.J., *et al.* Comparative anatomy of the passerine carpometacarpus helps illuminate the early fossil record of crown Passeriformes. *Journal of Anatomy* 2023; **242:** 495–509.

Wagner P.J. Exhaustion of morphologic character states among fossil taxa. *Evolution* 2000; **54:** 365–386.

Wilkinson M. Homoplasy and parsimony analysis. *Systematic Zoology* 1991 **40:** 105–9.

Wills M.A. Congruence between phylogeny and stratigraphy: randomization tests and the gap excess ratio. *Systematic Biology* 1999; **48:** 559–580.

Figure legends

Figure 1: Summary of tree length calculations per method. Schematic to summarise randomisation 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).

Figure 2: Homoplasy indices for empirical and random matrices (Analysis 1). Homoplasy values (Table 4) for consistency index (CI), retention index (RI), homoplasy excess ratio (HER), homoplasy slope ratio (HSR), modified homoplasy slope ratio (HSRm) and relative homoplasy 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 homoplasy excess ratio are subtracted from 1 for ease of comparison to other methods. A value of 1 indicates the maximum amount of detectable homoplasy for each method.

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 normalised between 0 and 1; grey dotted line), when the mean log for character state transition rate (μ) 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.

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 (μ = -6),

medium (μ = -4) and high (μ = -2) rates of character state change for the Telluraves (A), Passeriformes *a* (B), Neornithes (C) and Avialae *a* (D) topologies. 1 – CI in red, 1 – RI in blue and RHI in black with 95% and 5% quantile ranges as envelopes. Points represent index values with the full set of tips. Y axis ranges are reported in Table S1.

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 (μ = -6), medium (μ = -4) and high (μ = -2) rates of character state change for the Telluraves (A), Passeriformes *a* (B), Neornithes (C) and Avialae *a* (D) topologies. 1 – CI in red, 1 – RI in blue and RHI in black with 95% and 5% quantile ranges as envelopes. Points represent index values with the full set of characters. Y axis ranges are reported in Table S1.

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 *a* (B), Neornithes (C) and Avialae *a* (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.

Supporting Information

'Supporting_Information.docx' document containing lists of data and R script files, extended methodology, supplementary results and custom R function usage notes.

'R_scripts.zip' zipped folder containing all R files to carry out analyses.

928 ‘Data.zip’ zipped folder containing all NEXUS and PHYLIP data files (topologies and matrices)

929 to carry out the analyses.

930