Estimating Phylogenies from Shape and Similar Multidimensional Data: Why It Is Not Reliable

DOI:
10.1093/sysbio/syaa003

Document Version
Accepted author manuscript

Citation for published version (APA):
Varón González, C., Whelan, S., & Klingenberg, C. (2020). Estimating Phylogenies from Shape and Similar Multidimensional Data: Why It Is Not Reliable. *Systematic Biology*, *69*(5), 863–883. https://doi.org/10.1093/sysbio/syaa003

Published in:
Systematic Biology

Running head: Reliability of multidimensional data

Estimating Phylogenies from Shape and Similar Multidimensional Data: Why It Is Not Reliable

CEFERINO VARÓN-GONZÁLEZ [1], SIMON WHELAN[1,2]

AND CHRISTIAN PETER KLINGENBERG[1*]

[1]*School of Biological Sciences, University of Manchester, Michael Smith Building, Oxford Road, Manchester M13 9PT, United Kingdom;*

[2]*Dept. of Evolutionary Biology, EBC, Uppsala University, Norbyägen 18D, 75236 Uppsala, Sweden*

**Corresponding author:*

Christian Peter Klingenberg, School of Biological Sciences, University of Manchester, Michael Smith Building, Oxford Road, Manchester M13 9PT, United Kingdom;

Phone: +44 161 2753899

E-mail: cpk@manchester.ac.uk

Abstract.—In recent years, there has been controversy whether multidimensional data such as geometric morphometric data or information on gene expression can be used for estimating phylogenies. This study uses simulations of evolution in multidimensional phenotype spaces to address this question and to identify specific factors that are important for answering it. Most of the simulations use phylogenies with four taxa, so that there are just three possible unrooted trees and the effect of different combinations of branch lengths can be studied systematically. In a comparison of methods, squared-change parsimony performed similarly well as maximum likelihood, and both methods outperformed Wagner and Euclidean parsimony, neighbor-joining and UPGMA. Under an evolutionary model of isotropic Brownian motion, phylogeny can be estimated reliably if dimensionality is high, even with relatively unfavorable combinations of branch lengths. By contrast, if there is phenotypic integration such that most variation is concentrated in one or a few dimensions, the reliability of phylogenetic estimates is severely reduced. Evolutionary models with stabilizing selection also produce highly unreliable estimates, which are little better than picking a phylogenetic tree at random. To examine how these results apply to phylogenies with more than four taxa, we conducted further simulations with up to eight taxa, which indicated that the effects of dimensionality and phenotypic integration extend to more than four taxa, and that convergence among internal nodes may produce additional complications specifically for greater numbers of taxa. Overall, the simulations suggest that multidimensional data, under evolutionary models that are plausible for biological data, do not produce reliable estimates of phylogeny. [Brownian motion; gene expression data; geometric morphometrics; morphological integration; squared-change parsimony; phylogeny; shape; stabilizing selection.]

Whether quantitative data should be used for estimating phylogenies has long been debated (Kitching et al. 1998; Felsenstein 2002). Much of these discussions has concerned scalar traits such as single length measurements or ratios between two measurements. In recent years, the debate has shifted mostly to multidimensional characters, where a number of quantities jointly characterize complex features of organisms or populations. Some early studies that pioneered phylogenetics were based on considerations of multidimensional spaces of allele frequencies for multiple loci (Cavalli-Sforza and Edwards 1967) and several more recent studies have estimated phylogenetic trees from data on gene expression (Enard et al. 2002; Rifkin et al. 2003; Uddin et al. 2004; Brawand et al. 2011), but most such analyses have used morphometric data on the shapes of organisms or their parts (e.g., Lockwood et al. 2004; González-José et al. 2008; Aguilar-Medrano et al. 2011; Smith and Hendricks 2013; Watanabe and Slice 2014; Catalano et al. 2015; Brocklehurst et al. 2016; Perrard et al. 2016; Bjarnason et al. 2017; Catalano and Torres 2017; Schroeder et al. 2017; Parins-Fukuchi 2018b; Álvarez-Carretero et al. 2019). It remains contentious, however, whether the phylogenies estimated from quantitative multidimensional variables are reliable.

During the last two decades, several proposals for estimating phylogenies from morphometric data have been discussed contentiously. Some authors have suggested phylogenetic analyses based on cladistic characters derived from partial warp scores (Fink and Zelditch 1995; Zelditch et al. 1995; Swiderski et al. 1998; Zelditch et al. 1998; Bogdanowicz et al. 2005; Clouse et al. 2011) or principal component scores (MacLeod 2002; González-José et al. 2008; Aguilar-Medrano et al. 2011; González-José et al. 2011; Brocklehurst et al. 2016). These proposals, however, have been criticized for various reasons, especially the decomposition of phenotypic

68 spaces into distinct characters (Bookstein 1994; Naylor 1996; Adams and Rosenberg
69 1998; Rohlf 1998; Monteiro 2000; Adams et al. 2011; Zelditch et al. 2012). Some
70 authors have advocated methods that use landmarks as characters in cladistic analysis
71 (Catalano et al. 2010; Goloboff and Catalano 2011; Catalano and Goloboff 2012;
72 Catalano et al. 2015; Perrard et al. 2016; Catalano and Torres 2017; Dehon et al.
73 2017; Ospina-Garcés and de Luna 2017; Ascarrunz et al. 2019; Palci and Lee 2019).
74 An alternative is to use methods that avoid dividing the phenotypic variation into
75 characters, but infer trees from distances among taxa using clustering techniques such
76 as neighbor joining (Polly 2001; Lockwood et al. 2004; Couette et al. 2005; Macholán
77 2006; Cardini and Elton 2008; Bjarnason et al. 2011; Cruz et al. 2012; Bjarnason et al.
78 2015; Galland and Friess 2016; Galland et al. 2016; Bjarnason et al. 2017; Schroeder
79 et al. 2017; Ascarrunz et al. 2019), UPGMA (Marcus et al. 2000; Polly 2001; Cardini
80 2003; Cardini and O'Higgins 2004; Cardini and Elton 2008; Piras et al. 2010;
81 Watanabe and Slice 2014; Koehl and Hass 2015; Pečnerová et al. 2015; Karanovic et
82 al. 2016; Gabelaia et al. 2017; Zelditch et al. 2017), or other clustering methods
83 (Cannon and Manos 2001; Polly 2001; Bjarnason et al. 2011). Other studies have
84 estimated phylogenies from morphometric data using statistical approaches such as
85 maximum likelihood (Cannon and Manos 2001; Polly 2003a, b; Caumul and Polly
86 2005; González-José et al. 2008; Ascarrunz et al. 2019) or Bayesian methods (Parins-
87 Fukuchi 2018a, b; Álvarez-Carretero et al. 2019). Theoretical studies and computer
88 simulations have demonstrated, however, that random evolutionary processes such as
89 Brownian motion frequently produce convergence, so that phenotypic distance may
90 not be a good indicator of time since divergence and the resulting estimates of
91 phylogenies thus may be unreliable (Lynch 1989; Stayton 2008). A large empirical

92 comparison of a range of methods in 41 morphometric datasets found that different
93 methods tend to produce similar and fairly poor results (Catalano and Torres 2017).
94 These debates raise the question of how the quality of estimated trees can be
95 assessed. So far, the majority of such assessments have compared trees obtained from
96 morphometric data to reference trees obtained from other evidence, most often from
97 molecular data (Cole et al. 2002; Lockwood et al. 2004; Cardini and Elton 2008;
98 González-José et al. 2008; Klingenberg and Gidaszewski 2010; Catalano and
99 Goloboff 2012; Perrard et al. 2016; Catalano and Torres 2017; Gabelaia et al. 2017;
100 Ascarrunz et al. 2019). This type of comparison, however, can be problematic. First, it
101 is often unclear whether the reference tree accurately represents the phylogeny of the
102 taxa (e.g., because of differences between gene trees and species trees; Maddison
103 1997). Second, many of these studies produced partial agreement in the trees, so that
104 the results are inherently ambiguous: adherents of a particular method can emphasize
105 that the trees are partly correct, critics can point out that other aspects are wrong. For
106 instance, Smith and Hendricks (2013, p. 377) "consider it impressive" that
107 morphometric characters were able to allocate 33–45% of taxa successfully to their
108 positions in a phylogenetic tree, whereas skeptics might argue that this implies a clear
109 majority of failures. A way to avoid this ambiguity is to use computer simulations of
110 evolution, where the true tree is known with certainty, and to use simple phylogenetic
111 trees, so that there is no equivocation whether an estimated tree is right or wrong. This
112 approach has been used for testing methods to infer phylogenies from molecular data
113 (Huelsenbeck and Hillis 1993; Hillis et al. 1994; Huelsenbeck 1995). Simulations
114 have been used in the context of geometric morphometrics to explore the
115 consequences on phylogenetic inference (Polly 2004; Perrard et al. 2016; Parins-
116 Fukuchi 2018a, b; Álvarez-Carretero et al. 2019). However, the simulations were

117 conducted only under restricted sets of parameters (e.g., dimensionality, patterns of
118 trait integration, branch lengths) and results are therefore difficult to generalize.
119 This study uses several sets of simulations to analyze how accurately
120 phylogenies can be estimated using quantitative multidimensional data and what
121 factors influence the quality of the resulting estimates. We use the four-taxon case as
122 the simplest situation where different unrooted trees are possible (Felsenstein 1978a;
123 Huelsenbeck and Hillis 1993). Because there are just three possible trees, there is no
124 ambiguity whether estimated trees are partly correct or partly incorrect. This approach
125 makes it possible to compare different methods for estimating phylogenies and to
126 examine systematically the effects of different combinations of branch lengths in the
127 phylogeny (Felsenstein 1978a; Huelsenbeck and Hillis 1993). Perhaps more
128 importantly, we implement several models that make different assumptions of how
129 phenotypic traits evolve. Because dimensionality is a fundamental characteristic of
130 multivariate traits and is likely to affect the reliability of phylogeny estimation
131 (Felsenstein 2002), we conduct simulations for different numbers of dimensions. A
132 related concept is phenotypic integration, which reflects how different traits are
133 related to each other and how variation is spread across the dimensions of the
134 phenotypic space (Klingenberg 2008; Goswami et al. 2014). To examine its effect on
135 the reliability of phylogenetic estimates, we conduct simulations with different
136 patterns of integration. Because stabilizing selection has been shown to be an
137 important factor in macroevolution (Estes and Arnold 2007), we include simulations
138 that examine its effect on phylogenetic reliability. Finally, to assess how these results
139 apply to analyses with more than four taxa, we conduct a further series of simulations
140 with up to eight taxa. Together, these simulations assess how reliably phylogenies can
141 be inferred from multidimensional data under a wide range of conditions. By

142 examining the potential and limitations of the methods and of the data, the simulations
143 provide new and decisive information to the debate about the role of multidimensional
144 quantitative data in phylogenetics.

145 MATERIALS AND METHODS

146 *Simulation Strategy*

147 Complex phenotypes can be represented in multidimensional spaces, in which
148 evolving populations appear as points in locations corresponding to their average
149 phenotypes. Examples of such multidimensional spaces are the space of gene
150 expression (e.g., Brawand et al. 2011) and shape tangent spaces (Dryden and Mardia
151 1998; Kendall et al. 1999) or, for structures with object symmetry, the subspace of the
152 shape tangent space containing the symmetric component of variation (Klingenberg et
153 al. 2002; Klingenberg 2015). Evolution of the mean phenotype in a population
154 corresponds to movement of the respective point through the phenotypic space.

155 Our strategy consists of repeatedly running evolutionary simulations for four
156 taxa in a phenotypic space (Fig. 1) and estimating the unrooted tree from the resulting
157 multidimensional phenotypes. The proportion of simulations in which these estimates
158 match the tree topology used in the simulation, the proportion of correct estimates, is
159 a natural measure of reliability of the phylogeny reconstruction. Because there are
160 only three possible trees (Fig. 2a), it is feasible to evaluate all three possible trees for
161 each simulation and the analyses are therefore guaranteed to find the optimal tree in
162 each simulation. Most importantly, however, it is completely clear that one tree is
163 correct and the other two are incorrect. Therefore, there is none of the ambiguity
164 about whether a reconstructed tree is “mostly correct” or “incorrect in some
165 fundamental features”, as it occurs almost inevitably in discussions of empirical

166 examples involving more taxa. A separate set of simulations (Experiment 5, below)
167 explores how the findings from the four-taxon trees extend to analyses with more taxa
168 and also uses methods to quantify how much the true and estimated trees differ.
169 *Evolutionary models*.—Our simulations use evolutionary models that are
170 variants of Brownian motion. Brownian motion has been of fundamental importance
171 as an evolutionary model in discussions about phylogenies and quantitative traits
172 (Cavalli-Sforza and Edwards 1967; Felsenstein 1973; Lynch 1989; Felsenstein 2002;
173 Stayton 2008). This model assumes that the phenotype of each lineage evolves by a
174 random change in each short time interval, that this change is equally likely in every
175 direction of the phenotypic space, and that the change is additive over longer time
176 spans. The resulting evolutionary trajectory is a random walk through the phenotypic
177 space (Fig. 1a). Under a Brownian motion model, there is an association between the
178 time since the splitting of two lineages and the expected distance between the
179 corresponding phenotypes, providing a possible basis for estimating phylogeny. This
180 association is not deterministic, however, but has a substantial stochastic component
181 of variation, such that estimating the phylogeny from the distances between the
182 phenotypes of the terminal nodes is inevitably fraught with a degree of uncertainty
183 (Lynch 1989).
184 To conduct simulations under a Brownian motion model, random walks of
185 lineages through the phenotypic space can be implemented explicitly (Figure 1a). It is
186 more efficient, however, to obtain changes along the branches in the phylogeny
187 directly as random vectors drawn from multivariate normal distributions with
188 variances proportional to the respective branch lengths and zero covariances among
189 variables (this follows from the multivariate version of the central limit theorem; e.g.,
190 Mardia et al. 1979). The phenotypes for the four terminal nodes can then be obtained

191 by combining these changes in accordance with the true phylogenetic tree (tree 1; Fig.
192 2a). All the simulations were implemented using the R 2.10 statistical package (R
193 Core Team 2013).

194 *Variation in branch lengths*.— Branch lengths reflect the opportunity for
195 evolutionary change along the branches of a phylogeny, and result jointly from the
196 rate of evolutionary change and the time interval corresponding to the respective
197 branch of the phylogeny. To examine the effects of variation in branch lengths, we
198 systematically explore different combinations of branch lengths, as in the simulation
199 study of Huelsenbeck and Hillis (1993). We conduct two different sets of simulations,
200 one to analyze the effects of the relative lengths of internal versus terminal branches
201 (Figure 2b) and another set to study the effect of long-branch attraction and related
202 difficulties for phylogeny reconstruction (Fig. 2c). In both cases, we divide the five
203 branches into two groups, within which all the branches have the same length.

204 In the first case, one group contains the four terminal branches and second
205 group consists of just the internal branch (Fig. 2b). Reconstructing the phylogeny
206 should be easier when the internal branch is much longer than the terminal branches,
207 because this situation provides ample opportunity for the two internal nodes to
208 diverge, while each of them is likely to remain close to its two adjoining terminal
209 nodes. Conversely, if the internal branch is much shorter than the terminal branches,
210 such that the tree approaches a polytomy, all four taxa are expected to be roughly
211 equidistant to one another and which tree fits the data best is substantially a matter of
212 chance. If the internal branch actually has length zero (i.e., if there is a polytomy), the
213 three possible unrooted trees represent the true tree equally well; in this case,
214 evaluating the phylogenetic reconstruction does not make sense. Whereas these
215 expectations are fairly straightforward, it is not clear to what extent intermediate

combinations of branch lengths provide reliable estimates of phylogeny. Our simulations aim to establish this under several evolutionary models.

In the second type of simulations, the internal branch and one terminal branch at either end of it have one branch length and the other two terminal branches have another branch length (Fig 2c). This arrangement of relative branch lengths has been shown to pose potential challenges to phylogenetic methods (Felsenstein 1978a; Huelsenbeck and Hillis 1993; Huelsenbeck 1995). Some methods may erroneously group together terminal nodes that are linked to the rest of the tree by long branches. This situation has long been known as long-branch attraction or heterotachy, where the rate of evolutionary changes differs among lineages in the phylogeny, and has been widely studied in molecular phylogenetics (Wiens and Hollingsworth 2000; Bergsten 2005; Philippe et al. 2005; Wägele and Mayer 2007; Degtjareva et al. 2012). It is less clear, however, whether this problem has similarly serious effects on phylogeny estimation from multidimensional phenotypes.

Experiment 1: Comparison of Estimation Methods

To examine the effect of different methods on phylogenetic reliability, we conducted a series of simulations using squared change parsimony (Huey and Bennett 1987; Maddison 1991), maximum likelihood (Felsenstein 1973, 1981), neighbor joining (Saitou and Nei 1987), UPGMA clustering (Sneath and Sokal 1973), as well as two variants of linear parsimony: Wagner parsimony (Farris 1970; Swofford and Maddison 1987; Goloboff et al. 2006) and Euclidean parsimony (first introduced under the name "minimum evolution" by Cavalli-Sforza and Edwards 1967; Thompson 1973; new name suggested by Klingenberg and Gidaszewski 2010). These variants have previously not been clearly distinguished in the phylogenetics literature, possibly because both methods reduce to the same minimization criterion for scalar

241 characters. For multidimensional phenotypes, however, the difference matters.
242 Computations for Wagner parsimony minimize the total amount of change for each
243 variable separately, then adding up the resulting amounts across all variables, which
244 amounts to minimizing the total amount of change over the tree measured as
245 Manhattan distance (Farris 1970; Swofford and Maddison 1987). By contrast,
246 Euclidean parsimony minimizes the sum of changes over all the branches of the tree
247 as Euclidean distances, using the Pythagorean theorem to combine changes in
248 different variables. The task of finding such a tree is known in computer science as
249 the Euclidean Steiner tree problem (Smith 1992; Prömel and Steger 2002; Brazil et al.
250 2008; Fampa et al. 2016). In the context of phylogenetic analyses of landmark data,
251 some recent studies have used a hybrid approach, called "phylogenetic
252 morphometrics", which combines features of both Wagner and Euclidean parsimony
253 (Catalano et al. 2010; Goloboff and Catalano 2011; Catalano and Goloboff 2012).
254 To demonstrate the difference between methods, two four-taxon phylogenies
255 were used: a tree with a short internal branch and long terminal branches (Fig. 2b) and
256 a second tree with two long terminal branches at either end of the internal branch and
257 short remaining branches (Fig. 2c). We ran simulations for two sets of branch lengths,
258 with the short branches at 10% and 30% of the length of the long branches, for which
259 preliminary simulations had shown that they represented challenging conditions for
260 phylogeny estimation. For each set of branch lengths, 1,000 simulations of isotropic
261 Brownian motion in 10 dimensions and another 1,000 simulations in 50 dimensions
262 were conducted.
263 For inferring the phylogeny from the phenotypes of the terminal nodes using
264 squared-change parsimony, we used the algorithm of McArdle and Rodrigo (1994) to
265 reconstruct the phenotypes for the internal nodes. Tree length was computed as the

266 total of squared changes, summed over all branches and all variables, and the shortest
267 tree for each simulation was accepted as the estimated tree. The maximum likelihood
268 estimate, under a model of isotropic Brownian motion, was obtained using the *contml*
269 program of the Phylip package (Felsenstein 2013). Euclidean parsimony was
270 implemented using the optimization algorithm of Smith (1992), whereas Wagner
271 parsimony was based on the algorithm by Farris (1970). Neighbor joining and
272 UPGMA trees were obtained from the matrix of Euclidean distances among
273 phenotypes of the four taxa in each simulation, using the *neighbor* program in Phylip
274 (Felsenstein 2013) with the appropriate settings.

275 *Experiment 2: Detailed Analysis for the Isotropic Brownian Motion Model*

276 To assess the effect of different combinations of dimensionality and of branch
277 lengths on phylogenetic reliability in more detail, we conducted further simulations of
278 evolution by Brownian motion. Dimensionality is a key aspect of multivariate data,
279 because more phenotypic attributes (e.g., more landmarks in morphometric studies)
280 can potentially carry more information, and therefore might plausibly improve the
281 quality of phylogenetic estimates. To examine the effects of dimensionality, we
282 conducted the simulations using Brownian motion models with 1, 2, 3, 5, 10, 20, 50
283 and 100 dimensions.

284 We conducted separate sets of simulations, one contrasting the internal branch
285 to all four terminal branches (Fig. 2b) and the other contrasting two terminal branches
286 at either end of the internal branch to the other three branches (Fig. 2c). For Brownian
287 motion, the absolute magnitude of the branch lengths affects only the overall scale of
288 distances between taxa, but has no effect on how taxa are arranged relative to one
289 another in phenotype space. This is different from molecular evolution, where there
290 are saturation effects if the product of time and substitution rate becomes very large,

because there are only four possible nucleotides (or 20 amino acids). Therefore, simulations only need to vary the ratio of branch lengths in the two groups of branches, but not the absolute branch length. In both sets of simulations, the ratio of branch lengths ranged from 1:20 to 20:1.

The phenotypic variation in these simulations was isotropic, with variances that were proportional to branch lengths and the same for all dimensions, and variation was independent among dimensions. For each number of dimensions and combination of branch lengths, phenotypes were obtained from 5,000 simulations. To reconstruct phylogenies, we used squared change parsimony for this set of simulations (and all subsequent ones), because the comparisons in Experiment 1 showed that this method performs well and because it is computationally efficient. Phylogenetic reliability was quantified as the percentage of the 5,000 simulations in which squared change parsimony returned the correct tree (tree 1).

Experiment 3: Brownian Motion with Phenotypic Integration

The model of isotropic variation, implying independent evolution of all phenotypic traits at equal rates and an equal amount of variation in all dimensions of the phenotypic space (Fig. 3a), is not a realistic representation of biological data, where integration among traits is virtually ubiquitous (Olson and Miller 1958; Cheverud 1996; Wagner et al. 2007; Klingenberg 2008, 2013). Integration means that traits are correlated with each other and that, as a result, variation is concentrated in certain directions in phenotypic space (Wagner 1984; Klingenberg 2008; Pavlicev et al. 2009). Integration may be detrimental for phylogeny estimation because multiple traits may convey the same information, rather than each trait adding new information, or because the variation may not occupy the entire dimensionality available in the phenotypic space.

We include two sets of simulations to investigate the effects of integration on estimation of phylogeny from multidimensional traits (Figure 3). One model simulates very strong integration, in which a single dimension accounts for 80% of the total variation and all the other ones take up the remaining 20% of variation in equal amounts (Fig. 3b). In another model, the relative amount of variance decreases in an exponential manner from one dimension to the next, so that the variance in each dimension is 60% of the variance in the preceding dimension (Fig. 3c). For comparison with empirical data, these variances are equivalent to the eigenvalues obtained from a principal component analysis (PCA) of the evolutionary covariance matrix in the data.

For this set of simulations, tree length was computed using squared-change parsimony, which treats changes in every direction of phenotypic space in the same way. Because this method for estimating phylogeny is based on the relative arrangement of phenotypes of the different taxa in a multidimensional space, the orientation of the coordinate system does not influence the results. Because of this invariance to orientation, we can choose any coordinate system without loss of generality. Accordingly, we use the principal components (PCs) of the evolutionary covariance matrix as the coordinate system for our simulations, so that evolutionary changes in the resulting coordinates are uncorrelated with one another. We can therefore simulate the evolutionary change on each branch by independently drawing random deviations from normal distributions with variances as described above (Fig. 3), multiplied with the respective branch length.

Compensating for integration.—In principle, it is possible to address the problem of integration among traits by using Mahalanobis distances for estimating phylogenies (Felsenstein 1973, 1981, 1988; Álvarez-Carretero et al. 2019).

341 Mahalanobis distances are based on a transformation of the phenotypic space that, if
342 the assumptions are met, produces a modified space where variation is isotropic. To
343 achieve this, the transformation relatively shrinks those axes of the phenotypic space
344 that account for much of the total variation are and relatively stretches those axes that
345 account for little variation. Usually, this transformation is applied to the variation
346 within groups (Mardia et al. 1979; Klingenberg and Monteiro 2005), but in the
347 present context, the phenotypic space is transformed so that evolutionary variation
348 becomes isotropic. In this modified space, therefore, the effect of evolutionary
349 integration has been removed. This transformation, however, comes with other
350 potentially fundamental changes in the scaling of different dimensions and in the
351 relative arrangement of taxon averages.
352 If the evolutionary covariance matrix were known, therefore, the phenotypic
353 space could be scaled by the inverse of this matrix, transforming the space to a new
354 space of Mahalanobis distances, in which the isotropic Brownian motion model for
355 evolutionary change would apply. In practice, however, the evolutionary covariance
356 matrix usually is not known, but must be estimated from the available data, which is
357 exceedingly difficult if the phylogeny itself is also unknown (Felsenstein 1973, 1988,
358 2002). In principle, the phylogeny and evolutionary covariance matrix could be
359 estimated simultaneously, but stringent limits on the relative number of taxa and
360 dimensions of the phenotypic space apply (Felsenstein 2002).
361 For the purpose of this study, we made a series of assumptions that should be
362 very favorable for phylogeny estimation, even though unrealistic for most clades of
363 actual organisms: evolution is by pure drift, the phenotypic, additive genetic and
364 mutational covariance matrices are proportional, and these covariance matrices are
365 constant across the phylogeny. If these assumptions are met, the within-population

366 covariance matrix can be used as a substitute for the evolutionary covariance matrix
367 to obtain the transformed phenotypic space. Even though these assumptions are
368 unlikely to be met by biological data, we use them in our simulations, as did a
369 previous study (Álvarez-Carretero et al. 2019). We carried out separate simulations
370 using the sample covariance matrix and a shrinkage estimator of the covariance
371 matrix for computing Mahalanobis distances (Ledoit and Wolf 2004; Álvarez-
372 Carretero et al. 2019). For further details, see Online Appendix 1 (available on Dryad,
373 doi:10.5061/dryad.sk244r4).

374 *Experiment 4: Stabilizing Selection Model*

375 Stabilizing selection appears to be widespread (e.g., Estes and Arnold 2007)
376 and it can potentially have serious effects on estimates of phylogeny from the traits it
377 affects (Polly 2004). We simulated stabilizing selection using an Ornstein–Uhlenbeck
378 model with a single adaptive peak (Hansen 1997). With more than one adaptive peak,
379 the behavior of the model would be dominated by the assumptions about the
380 processes of switching between peaks. Because little is known about these processes
381 and implementation is problematic for small numbers of taxa, we limited the
382 simulations to a single adaptive peak.

383 The simulations of evolution under stabilizing selection were conducted as
384 explicit random walks, starting from a root of the phylogeny located at the midpoint
385 of the internal branch (Fig. 1b). At each interval from time t to $t + 1$, each population
386 changes its position from $\mathbf{x}_t$ to $\mathbf{x}_{t+1}$ following the equation $\mathbf{x}_{t+1} = \mathbf{x}_t + \alpha(\theta - \mathbf{x}_t) + \sigma$,
387 where α is a coefficient indicating the strength of stabilizing selection, θ is the
388 position of the adaptive peak, and σ is an isotropic random deviation, drawn from a
389 multivariate normal distribution with zero mean and an identity matrix as the

covariance matrix. The coefficient α can take values from zero (in this case, the model will be the same as isotropic Brownian motion) to unity (in that case, the phenotype will be returned exactly to the optimum at each iteration, and will only deviate by the random effect newly added in that round).

Each simulation consisted of a number of iterations that is determined by the branch lengths, which were varied in steps of 6 iterations from 10 to 100 iterations, as required for the simulation (Fig. 2b, c). We conducted separate simulations with weak and strong stabilizing selection, which use values of $\alpha = 0.05$ and $\alpha = 0.3$ respectively. The simulations started with two populations at the root of the phylogeny, midway on the internal branch of the unrooted tree, both with initial phenotypes $\mathbf{x}_0 = (0, \ldots,0)$. To test for the effect of the initial conditions, we conducted separate simulations where the starting point coincides with the optimal phenotype, $\theta = (0, \ldots,0)$. A separate set of simulations was conducted for the situation where the starting point is at a distance to the optimum, which was set to $\theta = (35, 0, \ldots, 0)$ (Fig. 1b). This is equivalent to a model that initially contains a component of directional selection, which then diminishes as each lineage approaches the optimum phenotype.

For each set of branch lengths, dimensionality, strength of stabilizing selection, and location of the optimum, we conducted 2,000 simulations. Squared-change parsimony was used to estimate phylogenies.

Experiment 5: Simulations with More Than Four Taxa

To examine how the results for trees with four taxa extend to a greater number of taxa, we ran additional simulations using up to eight taxa. The main difference to four-taxon simulations is that there are many more possible tree topologies (e.g., for 8 taxa, there are 10,395 unrooted bifurcating trees; Felsenstein 1978b; Felsenstein

2004). This rise in the number of possible trees entails some further complications. First, the computational effort required increases rapidly with the number of taxa. We chose the limit of eight taxa because this is the maximum for which it is feasible to conduct exhaustive searches in order to identify shortest trees with certainty. Second, there is the question of how the topology for the true phylogenetic tree to be used in the simulations should be chosen.

To obtain an insight into the overall effect of taxon number, we used trees randomly drawn from a uniform distribution over all unrooted bifurcating tree topologies with the appropriate number of taxa. Branch lengths were chosen so that they were the same for all internal branches and for all terminal branches, with ratios of internal to terminal branch lengths of 0.05, 0.1, 0.25, 0.5, 1, 2, 4, 10 and 20 in different simulation runs. Phenotypic evolution was simulated as isotropic Brownian motion or as Brownian motion according to the exponential integration model (see above and Fig. 3c). These simulations were run for phenotypes with 2, 5, 10, 20, 50 and 100 dimensions. For each combination of branch length ratio, evolutionary model, dimensionality and number of taxa, 1000 simulations were run. Squared-change parsimony was used as the method for estimating phylogeny.

To assess the performance of phylogenetic estimation, we scored the results for phylogenetic reliability as the proportion of simulation runs where the correct tree was returned. For more than four taxa, there is also the question how close an estimated tree is to the true one, even if it is not exactly correct. To address this question, we computed distances between the true and estimated trees using two topological distance measures: the Robinson–Foulds distance (Robinson and Foulds 1981) and the quartet distance (Estabrook et al. 1985). Both are metrics, but they differ somewhat in their properties (Steel and Penny 1993; Smith 2019a). The

Robinson–Foulds distance was computed with the ape package in R (Paradis and Schliep 2019) and quartet distance with the Quartet library (Smith 2019b). Because both these distance measures depend on the number of taxa in the trees being compared, we standardized the distances. To do so, we divided the distances by the expected distance between pairs of random trees with the appropriate number of taxa. For up to seven taxa, those expected distances were computed by full enumeration and are therefore exact; for eight taxa, the average was taken over a sample of 1 million pairs of random trees.

To explore whether the topology of the tree used to generate has an effect on the reliability of phylogenetic estimation, we conducted a series of simulations using extreme tree shapes. For details, see Online Appendix 2 (available on Dryad).

RESULTS

Experiment 1: Comparison of Methods

The differences in performance among methods depend on the branch-length scenarios. For simulations with a true tree in which the internal branch is 30% as long as the terminal branches (Fig. 2b) and a 50-dimensional phenotype, all methods did similarly well: squared-change parsimony found the correct tree in 70.5% of simulations, maximum likelihood in 70.4%, Euclidean parsimony in 70.6%, Wagner parsimony in 67.9%, neighbor joining in 70.0%, and UPGMA in 63.0% (Table 1). In the 2-versus-3-branch scenario (Fig. 2c), there were marked differences among methods: squared-change parsimony found the correct tree in 84.3% of simulations, maximum likelihood in 84.7%, Euclidean parsimony in 81.9%, Wagner parsimony in 74.8%, neighbor joining in 78.5%, and UPGMA in 14.8% of the simulations (Table 1). In the vast majority of simulations, squared-change parsimony and maximum

likelihood yielded the same trees, regardless whether correct or incorrect (99.7% for internal vs. terminal branches, 99.5% for 2 vs. 3 branches; Table 1). In corresponding simulations with Brownian motion in 10 instead of 50 dimensions, the results were similar, but all methods performed somewhat worse and the differences between them were slightly less accentuated (Supplementary Table S1, available on Dryad).

Success rates were lower overall in simulations where the shorter branch lengths were 10% of those of the longer branches, instead of 30%. In particular, in the 2-versus-3-branch scenario with a 50-dimensional phenotype, some pronounced differences between methods emerged: squared-change parsimony found the correct tree in 60.2% of simulations, maximum likelihood in 66.3%, Euclidean parsimony in 44.8%, Wagner parsimony in 23.6%, neighbor joining in 21.7%, whereas UPGMA never produced the correct tree (Supplementary Table S2, available on Dryad). In this series of simulations, squared-change parsimony and maximum likelihood returned the same tree in 92.0% of simulations, confirming the close relation between the two methods (Supplementary Table S2, available on Dryad). Because squared-change parsimony consistently performed best or close to best (in those series where maximum likelihood performed better), and because of its computational efficiency, we exclusively use squared-change parsimony for the remaining simulations, which focus on the reliability of estimated phylogenies in response to properties of the data.

Experiment 2: Detailed Analysis for the Isotropic Brownian Motion Model

The more detailed simulations using isotropic Brownian motion show that two major determinants of phylogenetic reliability are the relative branch lengths and dimensionality of the phenotype (Fig. 4a). Phylogenetic reliability improves consistently, and may reach 100%, as the ratio of internal to terminal branch lengths increases (Fig. 4a, solid lines, from left to right in the diagrams). This improvement

488 becomes more accentuated with increasing dimensionality. From lower to higher
489 dimensionality of the phenotype, the region of high or perfect phylogenetic reliability
490 expands toward shorter relative lengths of the internal branch.
491 At low dimensionality, the 2-versus-3 branch scenario (dashed lines in Fig. 4a)
492 appears more challenging than the situation where the internal branch is contrasted to
493 the four terminal branches (solid lines in Fig. 4a). For very high dimensionality,
494 however, the phylogenetic reliability is good even for simulations with a moderate
495 degree of long-branch attraction, where two terminal branches at opposite ends of the
496 internal branch are longer than the remaining three branches (left side of the diagrams
497 in Fig. 4a, dashed lines).

498 *Experiment 3: The Effect of Phenotypic Integration*

499 Phenotypic integration has a strong adverse effect on the accuracy of
500 phylogenetic estimates (Fig 4b, c). For the model with high integration, where one
501 dimension contains 80% of the total variation (Fig. 3b), there is a change in the
502 relation between branch length ratios and reliability from one to two dimensions, but
503 then this relation remains nearly the same for all simulations with greater
504 dimensionality (Fig. 4b). In contrast to the simulations with isotropic variation (Fig.
505 4a), where phylogenetic reliability improves with increasing dimensionality, it
506 appears that this improvement ceases after two dimensions for the high-integration
507 model (Fig. 4b). Similarly, for the exponential integration model (Fig. 3c), the benefit
508 of higher dimensionality extends to about 5 dimensions, but including dimensions
509 beyond that provides no further improvement (Fig. 4c). This loss of the improved
510 phylogenetic reliability with higher-dimensional phenotypes affects both the internal-
511 versus-terminal and the 2-versus-3 branch scenarios (solid and dashed lines in Figs.
512 4b, c). Under either model of integration, high phylogenetic reliability is only

513 achieved under very special conditions, if the internal branch is extremely long
514 relative to the terminal branches (Fig. 4b,c).
515 The separate series of simulations using Mahalanobis distance in phylogenetic
516 estimation, showed that this approach ameliorated the effects of phenotypic
517 integration partly, but not completely. These simulations identified the sample size
518 used to estimate covariance structure as a further complicating factor, and the
519 shrinkage estimate performed somewhat better than sample covariance matrices (for
520 further details, see Online Appendix 1, available on Dryad).

521 *Experiment 4: The Effect of Stabilizing Selection*

522 Phylogenetic reliability under an evolutionary process with stabilizing
523 selection, for most combinations of branch lengths, is not much better than drawing
524 trees randomly (Fig. 5). If stabilizing selection is weak, the accuracy of the estimates
525 is better where the terminal branches are much shorter than the internal branch (at the
526 bottom of the diagrams in Fig. 5a, b), especially when the dimensionality is high. For
527 the 2-versus-3 branch simulations, reliability is best if all branches are short and more
528 or less equal (lower-left corners of the diagrams in Fig. 5d, e; note that this situation,
529 with all branches short, is similar to the lower-left corners of the diagrams in Fig. 5a,
530 b). A particular situation occurs for the simulation with weak stabilizing selection
531 with an initial phenotype at some distance from the optimum and with strong long-
532 branch attraction. In this situation, only the lineages of the two long terminal branches
533 have time to approach the optimum. Consequently, the incorrect tree ((A,D),(B,C))
534 tends to be shorter than the correct tree ((A,B),(C,D)). And phylogenetic reliability is
535 systematically worse than drawing trees randomly (left edges of the diagrams in Fig.
536 5e).
537 With strong stabilizing selection, there is no combination of branch lengths

where the phylogenetic reliability for estimating phylogenies from the phenotypic data is perceptibly better than for drawing phylogenies at random. This is true regardless of dimensionality, and it makes no noticeable difference whether the simulations start with the optimal phenotype or at a distance from it (Fig. 5c, f; simulations starting at the optimal phenotypes not shown because the graphs look the same).

Experiment 5: Trees with More than Four Taxa

Phylogenetic reliability tends to decrease with increasing number of taxa (Fig 6a, b). For more favorable branch length ratios (internal branches long relative to terminal branches), the reliability is higher for four taxa and the decrease more gradual than for unfavorable branch ratios (internal branches short relative to terminal branches). The decrease of reliability with increasing number of taxa is more accentuated for low- than high-dimensional phenotypes under isotropic Brownian motion (Fig 6a), but with phenotypic integration, the benefit of increasing dimensionality beyond about five dimensions vanishes (Fig. 6b).

To assess whether estimated trees, even if they did not match the true trees exactly, were at least a reasonable approximation, we examined the distances between true and estimated trees, relative to the distances expected for pairs of random trees with the corresponding numbers of taxa. The results for both the Robinson–Foulds metric (Fig. 6c, d) and for the quartet metric (Supplementary Fig. S1, available on Dryad) are very similar. For low or medium branch length ratios, the average relative tree distances between the true and estimated trees are essentially constant regardless of the number of taxa, indicating that division by the expected distance between random trees appears to be an effective correction for the dependence of tree distances on the number of taxa. For isotropic Brownian motion, there is a clear benefit of

563 dimensionality, in that high-dimensional phenotypes yield lower relative tree
564 distances than low-dimensional phenotypes (Fig. 6c). If there is phenotypic
565 integration, this benefit does not extend beyond approximately five dimensions (Fig.
566 6d). Simulations with low branch length ratios produce no discernible change of
567 relative tree distances with taxon number. For high branch length ratios, however,
568 there is a gentle but clear trend for relative tree distances to rise with increasing
569 numbers of taxa (Fig. 6c, d). Under the model with isotropic Brownian motion, high-
570 dimensional phenotypes alleviate this trend (Fig. 6c), but when there is phenotypic
571 integration, the trend is clearly apparent no matter how high the dimensionality of the
572 phenotypic space (Fig. 6d).
573 To examine whether averaging over random tree topologies for any given
574 number of taxa might obscure some relevant differences due to the topology of the
575 tree used to simulate data, we conducted a set of simulation using specific topologies
576 with extreme tree shapes. Phylogenetic reliability and the distributions of tree
577 distances were similar, indicating that such differences are subtle (for details, see
578 Online Appendix 2, available on Dryad).

579 ## DISCUSSION

580 The simulations in this study have shown that the accuracy of phylogenetic
581 estimates from multidimensional phenotypes depends on a number of factors: the
582 relative branch lengths in the tree used to generate the data, the dimensionality of the
583 phenotype under study, and the model of how phenotypes evolve. Two particularly
584 important aspects of the evolutionary models are morphological integration and
585 stabilizing selection. Here, we explore these results further and evaluate them in light

586 of published evidence to assess their possible implications for the use of shape or
587 other multivariate phenotypes for estimating phylogenies.

588 *Comparison of Methods*

589 The comparison of methods is largely consistent with earlier results that
590 focused on molecular data (Huelsenbeck and Hillis 1993; Huelsenbeck 1995;
591 Swofford et al. 1996; Felsenstein 2004), but the choice of methods covered here
592 reflects those used in studies with morphometric data. Squared-change parsimony and
593 maximum likelihood performed similarly well, in the vast majority of simulation runs
594 returning the same trees, regardless of whether those were correct or incorrect (Table
595 1; Supplementary Tables S1–S3, available on Dryad). The close relation between
596 squared-change parsimony and maximum likelihood is well established (Maddison
597 1991; Schluter et al. 1997; Martins 1999; Felsenstein 2004). The difference is that
598 maximum likelihood includes a weighting by branch lengths (Felsenstein 1981); the
599 calculations therefore also include estimation of the branch lengths and the weighting
600 as extra steps that are not carried out for squared-change parsimony. Note also that,
601 for a uniform prior distribution, the maximum likelihood tree is also the tree with the
602 highest posterior probability and therefore a Bayesian point estimate of the phylogeny
603 (Huelsenbeck et al. 2001). Because squared-change parsimony performed nearly as
604 well as maximum likelihood, but is faster computationally, it is a reasonable choice
605 for the remainder of the simulations in this study: based on the comparisons, it is very
606 unlikely that a different method would produce substantially better results.

607 There are marked differences in performance between the two variants of
608 linear parsimony and between the two clustering methods, especially in the in the 2-
609 versus-3-branch scenario. In this situation, Euclidean parsimony is nearly as accurate
610 as squared-change parsimony and maximum likelihood (Table 1; but this does not

611 hold for more extreme branch length ratios, Supplementary Table S2, available on
612 Dryad), whereas Wagner parsimony performs clearly worse (under some
613 circumstances worse than randomly picking trees; see particularly Table 1;
614 Supplementary Table S2, available on Dryad). It seems plausible that this discrepancy
615 relates to the difference in how the two methods combine changes across variables:
616 Euclidean parsimony uses the Pythagorean theorem to combine changes across
617 variables (which involves summing the squared changes on each branch of the tree),
618 whereas Wagner parsimony minimizes changes in each variable separately and then
619 sums them over all variables. Of the two clustering methods, neighbor-joining
620 performed consistently better than UPGMA. The difference is especially clear in the
621 2-versus-3-branch scenario: in three of the four series of simulations UPGMA was far
622 worse than picking trees at random (Table 1; Supplementary Tables S1–S3, available
623 on Dryad). It is well established that both Wagner parsimony and UPGMA can
624 produce misleading results under long-branch attraction (Felsenstein 1978a;
625 Huelsenbeck and Hillis 1993; Swofford et al. 1996; Felsenstein 2004).
626 In a methods comparison based on 41 morphometric datasets (Catalano and
627 Torres 2017), Wagner parsimony, neighbor joining, UPGMA and the "phylogenetic
628 morphometrics" method that combines Wagner and Euclidean parsimony (Catalano et
629 al. 2010) all produced similar and fairly low degrees of congruence between estimated
630 trees and reference phylogenies, whereas maximum likelihood and Wagner parsimony
631 based on a subset of PC scores performed even slightly worse. Those results are quite
632 different from the simulations in this study (Table 1; Supplementary Tables S1–S3,
633 available on Dryad). It is conceivable that long-branch attraction may only have
634 played a minor role in the 41 empirical datasets, so that the differences between
635 methods were not as manifest as in our simulations. Another difference is that the

636 datasets compiled by Catalano and Torres (2017) contained more than 4 taxa (range:
637 5–160 species), so that phylogeny estimation may have been inherently more
638 challenging (Fig. 6). A further difficulty is that the reference phylogenies were
639 estimated too, based on a variety of data, and that it is unclear how well they reflect
640 the actual phylogenies of the respective clades.

641 *The Effect of Dimensionality*

642 Increasing dimensionality has a favorable effect on phylogenetic accuracy
643 (Figs. 4, 6). This finding is in agreement with previous observations that using more
644 landmarks or variables in simulations produces better agreement between the
645 estimated trees and the true trees used to generate the data (Perrard et al. 2016; Parins-
646 Fukuchi 2018b). It also agrees with the basic intuition that using more information
647 should lead to a better estimate of phylogeny.

648 For fully understanding this result, it is important to consider the evolutionary
649 models used in the simulations and how the dimensionality of the phenotype affects
650 them. Brownian motion has been widely used as a model for the evolution of
651 phenotypic traits in one- or multidimensional settings (Cavalli-Sforza and Edwards
652 1967; Felsenstein 1973; Lynch 1989; Polly 2004; Stayton 2008; Perrard et al. 2016).
653 It is an evolutionary model that is favorable for estimating phylogeny because the
654 expected distance between taxa increases monotonically with the time of separation
655 (Lynch 1989). Yet, a difficulty is that this distance also has a high variability (a
656 coefficient of variation of 1.4 for one-dimensional Brownian motion), which may
657 often lead to convergence, reversals, and parallel evolution that may produce
658 erroneous phylogenetic estimates (Lynch 1989; Stayton 2008; Klingenberg and
659 Gidaszewski 2010).

660 The squared distance between the phenotypes at either end of a branch of the
661 phylogeny, up to a scaling factor representing the expected magnitude of change
662 along the branch, follows a chi-squared distribution with as many degrees of freedom
663 as there are dimensions in the phenotypic space (this follows from the Pythagorean
664 theorem and the definition of the chi-squared distribution with *n* degrees of freedom
665 as the sum of squared values of *n* mutually independent random variates drawn from
666 the standard normal distribution). The coefficient of variation for the chi-squared
667 distribution is the square root of two divided by the square root of the degrees of
668 freedom (Forbes et al. 2011). The relative variability of the phenotypic distances
669 therefore diminishes with increasing degrees of freedom. Note, however, that a
670 substantial improvement is only achieved with dimensionalities that are quite high:
671 the coefficient of variation is 0.44 for 10 dimensions, 0.2 for 50 dimensions, 0.14 for
672 100 dimensions, and 200 dimensions are necessary for a coefficient of variation of
673 0.1. As a consequence, increasing dimensionality of a Brownian motion process
674 causes phenotypic distances to become a more deterministic function of divergence
675 times. With increasing dimensionality of the phenotype, the phenotypic distances are
676 therefore expected to be a better reflection of the underlying branch lengths and it
677 should become easier to infer phylogenies from phenotypic divergence.
678 The benefits of high dimensionality also can be understood intuitively by
679 considering how probable it is for convergent evolution to occur, which is a form of
680 homoplasy and may lead to erroneous phylogenetic inferences. There is always just
681 one direction in which two lineages can converge toward each other in phenotypic
682 space, but with increasing dimensionality, there are more and more directions in
683 which the lineages can move away from each other. Convergence is quite likely in the
684 univariate case, as shown in previous studies (Lynch 1989), but it becomes less

probable as more dimensions are added (Stayton 2008), thus improving phylogenetic reliability, as can be seen in our simulation results (Fig. 4a).

Because high dimensionality reduces stochastic effects, it also can alleviate the problems of long-branch attraction and differences in evolutionary rates among branches in the phylogeny (Fig 4a, dashed lines). Yet for methods that are sensitive to long-branch attraction, such as UPGMA or Wagner parsimony, high dimensionality can exacerbate such problems (cf. Table 1 vs. Supplementary Table S1; Supplementary Tables S2 vs. S3, available on Dryad). In general, the weaker stochastic effects in simulations using high dimensionality tend to make the differences in performance among methods more apparent.

Above all, the benefit of high dimensionality has implications for the data used in phylogenetic analyses. Using methods such as PCA to reduce the dimensionality of phenotypic data before phylogenetic analyses would definitely be ill-advised. In the comparison of Catalano and Torres (2017), methods including a dimension reduction via PCA performed slightly worse than methods using the full dimensionality of the data, and it is possible that this poorer performance was due to the reduced dimensionality. Studying phenotypes with high dimensionality has been proposed as one way of increasing phylogenetic reliability (Felsenstein 1973; Polly 2004; González-José et al. 2008; Stayton 2008). Similarly, the suggestion to combine morphometric data from multiple structures (Catalano et al. 2015; Perrard et al. 2016; Catalano and Torres 2017) also can be viewed as a strategy to increase the dimensionality of the phenotypic space used for inferring phylogenies. Whether such strategies are effective, however, depends not only on the dimensionality of the data space, but also on how closely the phenotypic traits are integrated.

Phenotypic Integration

In the simulations of Brownian motion with integration, the benefit of increasing dimensionality ceases at some intermediate level—beyond that dimensionality, phylogenetic reliabilities seem to be constant and always worse than the corresponding simulations with isotropic variation (cf. Fig. 4b,c vs. 4a). The effect resulting from phenotypic integration is similar to that of a reduction of the dimensionality to a level that is less than the actual dimensionality of the phenotypic space. It is worst in the model of extreme integration (Fig. 3b), where reliability does not increase beyond a level comparable to isotropic motion in 2 dimensions (Fig. 4b). This effect is more moderate for the exponential integration model (Fig. 3c), where the benefit stops at approximately 5 dimensions (Fig. 4c, 6b, 6d). For both models, high phylogenetic reliability only results if the internal branch is very long relative to the terminal branches (Fig. 4b,c), a condition that is very unlikely to be met for most empirical data. In both these models, the point where phylogenetic reliability ceases to benefit from higher dimensionality relates to the distribution of variation across the phenotypic space: because most variation is concentrated within just a few dimensions and this distribution remains essentially the same no matter how many additional dimensions are included, the overall dimensionality of the phenotypic space is immaterial for phylogenetic reliability. Including additional dimensions adds directions that are mostly devoid of variation and therefore have little or no effect on phylogenetic reliability.

It appears from these simulations that integration is a serious problem for phylogenetic reconstruction. This raises the question whether the simulations of integration are realistic at all. In actual biological data, integration is ubiquitous—the variation in the data does not "fill" the entire dimensionality of the phenotypic space,

but is concentrated mostly in a few of the available dimensions because of integration (Olson and Miller 1958; Cheverud 1996; Klingenberg 2008, 2013; Goswami et al. 2014). The scenario of high integration, in which 80% of the variation is contained in the a single dimension (Fig. 3b), was designed to be extreme and probably exceeds the level of integration in real data, although some examples come quite close (e.g., analyses where the first PC accounts for more than 60% of variation among species; Klingenberg et al. 2012). The exponential model of integration (Fig. 3c) is more realistic, as numerous examples show comparable or greater strengths of interspecific integration in geometric morphometric data (e.g., Monteiro et al. 2005; Sidlauskas 2008; Friedman 2010; De Esteban-Trivigno 2011a, b; Monteiro and Nogueira 2011; Brusatte et al. 2012; Santana and Lofgren 2013; Baab et al. 2014; Martín-Serra et al. 2014; Watanabe and Slice 2014; Blanke 2018), although some other studies found somewhat weaker integration, albeit still with most variation concentrated in just a few dimensions (Figueirido et al. 2010; Chamero et al. 2013; Klingenberg and Marugán-Lobón 2013; Sherratt et al. 2014). Altogether, by comparison with empirical data, the exponential model of integration used in the simulation seems to be fairly realistic. Accordingly, those simulations are likely to represent evolutionary integration in actual biological datasets realistically, and the levels of phylogenetic reliability obtained in our simulations under the exponential model of integration represent what usually should be expected in empirical data.

In principle, the adverse effects of phenotypic integration can be mitigated by using Mahalanobis distance in the process of estimating phylogeny (Felsenstein 1973, 2002; Álvarez-Carretero et al. 2019). If the correct evolutionary covariance matrix is used to compute Mahalanobis distances, this eliminates the effects of integration and phylogenetic reliability therefore should be the same as for Brownian motion with no

integration. Our simulations show some improvements of phylogenetic reliability, especially when the shrinkage estimator of the covariance matrix is used (Ledoit and Wolf 2004). This is similar to the results of Álvarez-Carretero et al. (2019). Nevertheless, phylogenetic reliability is not restored completely to the levels for Brownian motion without integration and sampling errors may produce inaccuracies (Online Appendix 1, available on Dryad). It is important to note that the approach of using within-taxon phenotypic variation to estimate evolutionary covariance structure makes a number of key assumptions: evolution is by random drift, and the phenotypic, additive genetic and mutational covariance matrices are proportional and constant across the whole phylogeny. All these assumptions are at best questionable, and probably unrealistic for most clades and traits. Therefore, even though it is theoretically possible (Felsenstein 2002), the difficulties involved in estimating the evolutionary covariance matrix without knowing the phylogeny are likely to render this approach unworkable. Accordingly, no remedy against the effects of phenotypic integration exists that is practically viable for empirical studies.

Phenotypic integration is also of key importance when considering the suggestion to combine morphometric data from multiple structures (Catalano et al. 2015; Perrard et al. 2016; Catalano and Torres 2017). Whether, or to what extent, combining data from different structures results in a dimensionality of the combined phenotypic space that is higher than the dimensionality of the phenotypic spaces of the individual structures depends on the strength of integration among structures. The possible outcomes are on a spectrum limited by two extremes: complete integration, for which combining different structures will not have any effect at all (the phenotypic space of each structure contains the complete information about variation in any other structure), or no integration at all, where the dimensionalities of variation

784 in the phenotypic spaces will add up to the dimensionality of variation in the
785 combined phenotypic space. The scenario of no integration at all is grossly unrealistic
786 for actual morphological data, but how closely actual data can approximate the
787 limiting scenario of complete integration is not clear. Although we are not aware of
788 any examples of complete evolutionary integration, empirical studies show that
789 associations among different structures are widespread and often strong (Gómez-
790 Robles and Polly 2012; Hautier et al. 2012; Claverie and Patek 2013; Álvarez et al.
791 2015; Martín-Serra et al. 2015). Due to such evolutionary integration, combining data
792 from multiple structures in phylogenetic analyses therefore is likely to provide only
793 limited gains of phylogenetic reliability.

794 *Stabilizing Selection*

795 When the evolutionary model used in the simulations includes stabilizing
796 selection, phylogenetic reliability drops and, for most simulations, is little better than
797 for picking a tree at random (Fig. 5). For the simulations with strong stabilizing
798 selection, this applies regardless of the dimensionality or branch length combinations
799 used (Fig. 5c,f). In simulations with weak stabilizing selection, a combination of high
800 dimensionality and a true phylogeny with a long internal branch and short terminal
801 branches yielded a limited zone of better phylogenetic reliability (Fig. 5a,b). As soon
802 as the terminal branches surpass a minimum length, however, even weak stabilizing
803 selection is sufficient to eliminate the phylogenetic signal. In some of the simulations
804 under the 2-versus-3-branch scenario with weak stabilizing selection, there was even a
805 special set of circumstances where phylogenetic reliability was consistently worse
806 than picking trees at random: if simulations started off the optimum and the set of 3
807 branches was sufficiently short, only the two lineages of the 2 long terminal branches
808 tended to reach the optimal phenotype, and the analyses systematically returned the

809 wrong tree (Fig. 5e). Overall, these simulations indicate clearly that stabilizing
810 selection can have a severe detrimental effect on phylogenetic reliability. The reason
811 for this is that stabilizing selection attracts every lineage to the optimum phenotype
812 regardless of ancestry, and thereby erodes the phylogenetic signal. This general result
813 is in agreement with findings from different simulations (Revell et al. 2008).
814 Because we used a model stabilizing selection with a single adaptive peak, we
815 need to ask whether using a model with two or more peaks might lead to different
816 conclusions. The answer to this question depends on the processes that control
817 transitions from one peak to another. It is possible to conceive of scenarios giving rise
818 to strong phylogenetic signal, for instance, if clades are associated persistently with
819 different adaptive peaks. Because the taxa within each of these clades would be under
820 the same conditions as in a single-peak model, however, phylogenetic resolution
821 within clades would also be poor. Alternatively, if switches between peaks are so
822 frequent that closely related taxa are commonly associated with different peaks and
823 remotely related taxa with the same peak, convergence will be rampant and
824 phenotypic similarity will indicate association with adaptive peaks, not phylogenetic
825 relatedness.
826 Whereas evolution under a model of Brownian motion, in principle at least,
827 can continue without bounds, models of stabilizing selection ensure that phenotypes
828 sooner or later converge toward the optimal phenotype. If stabilizing selection is
829 sufficiently strong or the branches are sufficiently long, there is therefore no longer an
830 association between the time of separation and the phenotypic distance between taxa.
831 In other words, the phenotype loses the phylogenetic signal it may have had (see the
832 upper-right regions of the diagrams in Fig. 5). This phenomenon is analogous to the
833 problem of substitution saturation in molecular data, when the product of substitution

834 rate and branch lengths is so large that each position is expected to have undergone
835 multiple substitutions and therefore loses phylogenetic information. This is different
836 from the other models used in this study, where no such phenomenon exists and
837 phenotypic differences are expected to increase with time. In real organisms,
838 however, there cannot be an indefinite amount of change. Simulations of Brownian
839 motion can easily produce phenotypes that are clearly non-functional (Polly 2004), so
840 that it seems best to view the models as restricted to a domain of phenotype space
841 within which phenotypes are viable. If phenotypic variation extends to boundaries
842 beyond which phenotypes are not functionally viable, evolving lineages are affected
843 according to their phenotype and regardless of their ancestry, as for stabilizing
844 selection. Therefore, the effect of such boundaries would probably be detrimental to
845 phylogenetic reliability.

846 Studies of quantitative phenotypes such as morphological traits and gene
847 expression have found extensive evolutionary conservation (e.g., Rifkin et al. 2003;
848 Estes and Arnold 2007; Hunt 2007; Harmon et al. 2010; Kalinka et al. 2010; Gallego
849 Romero et al. 2012) and many comparative analyses reported a good fit of Ornstein–
850 Uhlenbeck models to morphometric data (e.g., Angielczyk et al. 2011; Monteiro and
851 Nogueira 2011; Frédérich et al. 2013; Kimmel et al. 2017; Aristide et al. 2018). These
852 findings support the view that stabilizing selection is widespread. It is therefore likely
853 that many studies attempting to estimate phylogenies from multidimensional
854 phenotypes will face problems similar to those in our simulations.

855 *More Than Four Taxa*

856 Because phylogenetic studies usually involve many more than four taxa, the
857 question arises whether and how the results of our simulations extend to greater
858 numbers of taxa. Our simulations with more than four taxa give some indications

about this (Fig. 6). First, there is a clear continuity from the results of simulations with four taxa to those with more taxa. Second, dimensionality and integration, two of the main factors accounting for the results in simulations with four taxa, can also explain the findings about trees with more taxa in the same manner.

Phylogenetic reliability tends to drop with increasing numbers of taxa. This reflects the fact that the number of possible tree topologies rises sharply with increasing number of taxa (Felsenstein 1978b, 2004). If random variation plays any substantial role, an increasing number of taxa means that one is picking at random (to some extent, at least) from a much greater number of trees, and consequently the chance of success drops. For the model of isotropic Brownian motion, and provided that internal branches are sufficiently long, high dimensionality of the phenotype can alleviate this effect (Fig. 6a). In the presence of integration, however, this favorable effect does not extend beyond approximately five dimensions (Fig. 6b). This is the same limitation from phenotypic integration that we discussed above for four taxa (Experiment 3). In the current context, the consequence of integration is that high-dimensional phenotypes provide no escape from the trend of falling phylogenetic reliability with increasing number of taxa.

With more than four taxa, it makes sense not just to ask whether an estimated tree is the same as the true phylogeny, but also to quantify how similar or how different they are. The rationale of this is that, even though the estimated trees might not match the true phylogeny perfectly, they might be sufficiently close to provide a reasonable approximation. Because tree distances depend on the number of taxa, we applied a correction by scaling distances in relation to the expected distance between random trees with the corresponding number of taxa—the scaled distances therefore indicate how much closer estimated trees are to the true phylogeny than randomly

884 picking trees. After this correction for the number of taxa, even though the Robinson–
885 Foulds distance and quartet distance are known to differ in their properties (Steel and
886 Penny 1993; Smith 2019a), they produced similar results in our simulations (Fig.
887 6c,d; Supplementary Fig. S1, available on Dryad). The main findings from the
888 analyses of distances are consistent with the results on the four-taxon case: higher
889 ratios of internal to terminal branch lengths produce estimated trees that tend to be
890 closer to the true trees, and so does increased dimensionality, but phenotypic
891 integration curtails the benefits of increased dimensionality.
892 There is an additional result, however, which is not just extending the findings
893 from the simulations with four taxa: when the branch length ratio is high and when
894 dimensionality is low or there is integration, increasing the number of taxa yields a
895 clear rise in the relative tree distances (Fig. 6c,d; Supplementary Fig. S1, available on
896 Dryad). This suggests that, even with long internal branches, an increasing number of
897 taxa poses an additional difficulty that is not present with fewer taxa. The likely
898 reason is that, with more than four taxa, the internal branches of the tree can "fold
899 over" so that taxa that are separated by two or more internal branches in the tree might
900 end up being relatively close to each other in phenotypic space (Fig. 7). This effect
901 depends on the dimensionality of the phenotypic space, because convergence among
902 internal nodes is less likely when dimensionality is high. When dimensionality is low
903 or integration confines variation to just a few dimensions of the phenotypic space,
904 increasing the number of taxa enhances the probability of such convergence.
905 Whereas, for just four taxa, a tree with an internal branch that is much longer than the
906 terminal branches consistently yields accurate estimates of phylogeny under a
907 Brownian motion model even when dimensionality is low (Fig. 4), it is surprisingly
908 difficult, under the same conditions, to conceive a similarly favorable scenario for

greater numbers of taxa because there is no way to avoid convergence among internal nodes. As a consequence, in those conditions, convergence among internal nodes is an extra source of error for phylogenetic inference. Because a greater number of taxa provides more opportunity for this problem to occur, its effect rises with increasing number of taxa (Fig. 6c,d; Supplementary Fig. S1, available on Dryad).

The majority of phylogenetic analyses include more than eight taxa, raising the question how the results of these simulations extend to greater numbers of taxa. The trends over the range of four to eight taxa indicate that the adverse effects of low dimensionality and phenotypic integration apply to the simulations similarly or, for the convergence among internal nodes, that increasing numbers of taxa even exacerbate the difficulties in estimating phylogenies (Fig. 6; Supplementary Fig. S1, available on Dryad). Our simulations with more than four taxa did not include stabilizing selection, but there is no apparent reason why the attraction of separate lineages to a common optimum would erode phylogenetic signal any less for greater numbers of taxa than it does for four taxa (Fig. 5).

CONCLUSIONS

The approach in this paper differs from previous studies in several ways and provides new insights. First, we used simulations rather than empirical examples, such as comparisons of phylogenetic trees estimated from morphometric data and reference trees (e.g., Cole et al. 2002; Lockwood et al. 2004; Klingenberg and Gidaszewski 2010; Catalano and Torres 2017). For this reason, there is certainty about the true phylogeny and the model of the evolutionary processes. Second, our simulations used simple trees, so that it was possible for simulations to explore systematically factors such as relative branch lengths and different evolutionary models, rather than a

933 smaller number of more complex trees under a more restricted set of conditions
934 (Perrard et al. 2016; Parins-Fukuchi 2018b). Due to the small number of taxa, there is
935 no ambiguity in the results whether estimated trees are correct or not, and the range of
936 models permitted us not only to determine whether or not multidimensional
937 phenotypic traits are reliable for estimating phylogenies, but also to understand why.
938 The simulations identified three key factors: the dimensionality of the trait
939 space, phenotypic integration, and stabilizing selection. Under Brownian motion, high
940 dimensionality is crucial for estimating phylogenetic trees reliably (Fig. 4).
941 Phenotypic integration is detrimental to phylogenetic reliability because variation is
942 limited to just a few of the available dimensions (Fig. 4b,c). Integration is near
943 ubiquitous in morphological structures (Klingenberg 2008; Goswami et al. 2014),
944 suggesting that it imposes widespread limitations on phylogenetic reliability. Because
945 there is no quantitative survey of the strength of evolutionary integration across a
946 broad range of taxa and traits, it is currently impossible to judge how severe these
947 limitations are. Stabilizing selection erodes phylogenetic signal from phenotypic data,
948 and therefore is highly detrimental for estimating phylogenetic trees (Fig. 5). It is a
949 widespread phenomenon (Estes and Arnold 2007), and therefore expected to have
950 adverse effects on phylogenetic analyses using phenotypic data in many clades.
951 Together, these factors conspire so that phylogenetic inference from morphometric
952 data, or other high-dimensional phenotypic data in general, must be expected to be
953 unreliable.
954 We understand that these results are frustrating to some investigators,
955 particularly to paleontologists, because morphometric data may be the only or at least
956 most easily available data for many fossil and even some extant taxa (MacLeod 2002;
957 Smith and Hendricks 2013; Dehon et al. 2017; Parins-Fukuchi 2018a). Where

958 possible, other data such as genomic sequence information can be used instead, which
959 suffers from these difficulties to a lesser extent and where vast amounts of
960 information are available (Rannala and Yang 2008). Even where such alternatives are
961 not available, however, we think it is preferable to recognize the limitations of
962 phylogenetic inference from such data, rather than to use approaches that may provide
963 unreliable results.
964

965 ## SUPPLEMENTARY MATERIAL

966 Supplementary material, including online-only appendices and R scripts
967 used for the simulations, can be found in the Dryad repository at
968 http://datadryad.org, doi:10.5061/dryad.sk244r4.

969 ## FUNDING

970 We gratefully acknowledge the assistance given by Research IT and the use of
971 the Computational Shared Facility at The University of Manchester.

972 ## ACKNOWLEDGMENTS

973 We thank Leandro Monteiro, Rob Sansom, Nicolas Navarro, and Sylvain
974 Gerber for helpful discussions, and Associate Editor Matt Friedman, Brian
975 Sidlauskas, and the anonymous reviewers for helpful comments on earlier versions of
976 this manuscript.

977 REFERENCES

978 Adams DC, Cardini A, Monteiro LR, O'Higgins P, Rohlf FJ. 2011. Morphometrics
979 and phylogenetics: Principal components of shape from cranial modules are neither
980 appropriate nor effective cladistic characters. J. Hum. Evol., 60:240-243.
981 Adams DC, Rosenberg MS. 1998. Partial warps, phylogeny, and ontogeny: A
982 comment on Fink and Zelditch (1995). Syst. Biol., 47:168-173.
983 Aguilar-Medrano R, Frédérich B, de Luna E, Balart EF. 2011. Patterns of
984 morphological evolution of the cephalic region in damselfishes (Perciformes:
985 Pomacentridae) of the Eastern Pacific. Biol. J. Linn. Soc., 102:593–613.
986 Álvarez A, Perez SI, Verzi DH. 2015. The role of evolutionary integration in the
987 morphological evolution of the skull of caviomorph rodents (Rodentia:
988 Hystricomorpha). Evol. Biol., 42:312–327.
989 Álvarez-Carretero S, Goswami A, Yang Z, dos Reis M. 2019. Bayesian estimation of
990 species divergence times using correlated quantitative characters. Syst. Biol., 68:967–
991 986.
992 Angielczyk KD, Feldman CR, Miller GR. 2011. Adaptive evolution of plastron shape
993 in emydine turtles. Evolution, 65:377–394.
994 Aristide L, Bastide P, dos Reis SF, Pires dos Santos TM, Lopes RT, Perez SI. 2018.
995 Multiple factors behind early diversification of skull morphology in the continental
996 radiation of New World monkeys. Evolution, 72:2697–2711.
997 Ascarrunz E, Claude J, Joyce WG. 2019. Estimating the phylogeny of geoemydid
998 turtles (Cryptodira) from landmark data: an assessment of different methods. PeerJ,
999 7:e7476.

1000 Baab KL, Perry JMG, Rohlf FJ, Jungers WL. 2014. Phylogenetic, ecological, and
1001 allometric correlates of cranial shape in Malagasy lemuriforms. Evolution, 68:1450–
1002 1468.
1003 Bergsten J. 2005. A review of long-branch attraction. Cladistics, 21:163–193.
1004 Bjarnason A, Chamberlain AT, Lockwood CA. 2011. A methodological investigation
1005 of hominoid craniodental morphology and phylogenetics. J. Hum. Evol., 60:47–57.
1006 Bjarnason A, Soligo C, Elton S. 2015. Phylogeny, ecology, and morphological
1007 evolution in the atelid cranium. Int. J. Primatol., 36:513–529.
1008 Bjarnason A, Soligo C, Elton S. 2017. Phylogeny, phylogemetic inference, and
1009 cranial evolution in pitheciids and *Aotus*. Am. J. Primatol., 79:e22621.
1010 Blanke A. 2018. Analysis of modularity and integration suggests evolution of
1011 dragonfly wing venation mainly in response to functional demands. J. R. Soc.
1012 Interface, 15:20180277.
1013 Bogdanowicz W, Juste J, Owen RD, Sztencel A. 2005. Geometric morphometrics and
1014 cladistics: Testing evolutionary relationships in mega- and microbats. Acta Chiropt.,
1015 7:39–49.
1016 Bookstein F. 1994. Can biometrical shape be a homologous character? In: Hall BK
1017 editor. Homology: The hierarchial basis of comparative biology. New York,
1018 Academic Press, p. 197-227.
1019 Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, Weier M,
1020 Liechti A, Aximu-Petri A, Kircher M, Albert FW, Zeller U, Khaitovich P, Grützner F,
1021 Bergmann S, Nielsen R, Pääbo S, Kaessmann H. 2011. The evolution of gene
1022 expression levels in mammalian organs. Nature, 478:343-350.

1023 Brazil M, Thomas DA, Nielsen BK, Winter P, Wulff-Nilsen C, Zachariasen M. 2008.
1024 A novel approach to phylogenetic trees: *d*-dimensional geometric Steiner trees.
1025 Networks, 53:104–111.
1026 Brocklehurst N, Romano M, Fröbisch J. 2016. Principal component analysis as an
1027 alternative treatment for morphometric characters: phylogeny of caseids as a case
1028 study. Palaeontology, 59:877–886.
1029 Brusatte SL, Sakamoto M, Montanari S, Harcourt Smith WEH. 2012. The evolution
1030 of cranial form and function in theropod dinosaurs: insights from geometric
1031 morphometrics. J. Evol. Biol., 25:365–377.
1032 Cannon CH, Manos PS. 2001. Combining and comparing morphometric shape
1033 descriptors with a molecular phylogeny: the case of fruit type evolution in Bornean
1034 *Lithocarpus* (Fagaceae). Syst. Biol., 50:860–880.
1035 Cardini A. 2003. The geometry of the marmot (Rodentia: Sciuridae) mandible:
1036 phylogeny and patterns of morphological evolution. Syst. Biol., 52:186–205.
1037 Cardini A, Elton S. 2008. Does the skull carry a phylogenetic signal? Evolution and
1038 modularity in the guenons. Biol. J. Linn. Soc., 93:813-834.
1039 Cardini A, O'Higgins P. 2004. Patterns of morphological evolution in *Marmota*
1040 (Rodentia, Sciuridae): geometric morphometrics of the cranium in the context of
1041 marmot phylogeny, ecology and conservation. Biol. J. Linn. Soc., 82:385–407.
1042 Catalano SA, Coloboff PA, Giannini NP. 2010. Phylogenetic morphometrics (I): the
1043 use of landmark data in a phylogenetic framework. Cladistics, 26:539–549.
1044 Catalano SA, Ercoli MD, Prevosti FJ. 2015. The more, the better: the use of multiple
1045 landmark configurations to solve the phylogenetic relationships in musteloids. Syst.
1046 Biol., 64:294–306.

1047 Catalano SA, Goloboff PA. 2012. Simultaneously mapping and superimposing
1048 landmark configurations with parsimony as optimality criterion. Syst. Biol., 61:392-
1049 400.
1050 Catalano SA, Torres A. 2017. Phylogenetic inference based on landmark data in 41
1051 empirical data sets. Zool. Scr., 46:1–11.
1052 Caumul R, Polly PD. 2005. Phylogenetic and environmental components of
1053 morphological variation: skull, mandible, and molar shape in marmots (*Marmota*,
1054 Rodentia). Evolution, 59:2460–2472.
1055 Cavalli-Sforza LL, Edwards AWF. 1967. Phylogenetic analysis: models and
1056 estimation procedures. Evolution, 21:550–570.
1057 Chamero B, Buscalioni ÁD, Marugán-Lobón J. 2013. Pectoral girdle and forelimb
1058 variation in extant Crocodylia: the coracoid–humerus pair as an evolutionary module.
1059 Biol. J. Linn. Soc., 108:600–618.
1060 Cheverud JM. 1996. Developmental integration and the evolution of pleiotropy.
1061 Amer. Zool., 36:44–50.
1062 Claverie T, Patek SN. 2013. Modularity and rates of evolutionary change in a power-
1063 amplified prey capture system. Evolution, 67:3191–3207.
1064 Clouse RM, de Bivort BL, Giribet G. 2011. Phylogenetic signal in morphometric
1065 data. Cladistics, 27:337–340.
1066 Cole TM, III., Lele SR, Richtsmeier JT. 2002. A parametric bootstrap approach to the
1067 detection of phylogenetic signals in landmark data. In: MacLeod N, Forey PL editors.
1068 Morphology, shape and phylogeny. London, Taylor & Francis, p. 194–219.
1069 Couette S, Escarguel G, Montuire S. 2005. Constructing, bootstrapping, and
1070 comparing morphometric and phylogenetic trees: A case study of New World
1071 monkeys (Platyrrhini, Primates). J. Mammal., 86:773–781.

1072 Cruz RAL, Pante MJR, Rohlf FJ. 2012. Geometric morphometric analysis of shell
1073 shape variation in *Conus* (Gastropoda: Conidae). Zool. J. Linn. Soc., 165:296–310.
1074 De Esteban-Trivigno S. 2011a. Buscando patrones ecomorfológicos comunes entre
1075 ungulados actuales y xenartros extintos. Ameghiniana, 48:189–209.
1076 De Esteban-Trivigno S. 2011b. Ecomorfología de xenartros extintos: análisis de la
1077 mandíbula con métodos de morfometría geométrica. Ameghiniana, 48:381–398.
1078 Degtjareva GV, Valiejo-Roman CM, Samigullin TH, Guara-Requena M, Sokoloff
1079 DD. 2012. Phylogenetics of *Anthyllis* (Leguminosae: Papilionoideae: Loteae): Partial
1080 incongruence between nuclear and plastid markers, a long branch problem and
1081 implications for morphological evolution. Mol. Phylogenet. Evol., 62:693-707.
1082 Dehon M, Perrard A, Engel MS, Nel A, Michez D. 2017. Antiquity of
1083 cleptoparasitism among bees revealed by morphometric and phylogenetic analysis of
1084 a Paleocene fossil nomadine (Hymenoptera: Apidae). Syst. Entomol., 42:543–554.
1085 Dryden IL, Mardia KV. 1998. Statistical shape analysis. New York, John Wiley &
1086 Sons.
1087 Enard W, Khaitovich P, Klose J, Zöllner S, Heissig F, Giavalisco P, Nieselt-Struwe
1088 K, Muchmore E, Varki A, Ravid R, Doxiadis GM, Bontrop RE, Pääbo S. 2002. Intra-
1089 and interspecific variation in primate gene expression patterns. Science, 296:340-343.
1090 Estabrook GF, McMorris FR, Meacham CA. 1985. Comparison of undirected
1091 phylogenetic trees based on subtrees of four evolutionary units. Syst. Zool., 34:193–
1092 200.
1093 Estes S, Arnold SJ. 2007. Resolving the paradox of stasis: Models with stabilizing
1094 selection explain evolutionary divergence on all timescales. Am. Nat., 169:227-244.
1095 Fampa M, Lee J, Maculan N. 2016. An overview of exact algorithms for the
1096 Euclidean Steiner tree problem in *n*-space. Intl. Trans. Op. Res., 23:861–874.

1097 Farris JS. 1970. Methods for computing Wagner trees. Syst. Zool., 19:83–92.
1098 Felsenstein J. 1973. Maximum-likelihood estimation of evolutionary trees from
1099 continuous characters. Am. J. Hum. Genet., 25:471–492.
1100 Felsenstein J. 1978a. Cases in which parsimony or compatibility methods will be
1101 positively misleading. Syst. Biol., 27:401–410.
1102 Felsenstein J. 1978b. The number of evolutionary trees. Syst. Zool., 27:27–33.
1103 Felsenstein J. 1981. Evolutionary trees from gene frequencies and quantitative
1104 characters: finding maximum likelihood estimates. Evolution, 35:1229–1242.
1105 Felsenstein J. 1988. Phylogenies and quantitative characters. Annu. Rev. Ecol. Syst.,
1106 19:455–471.
1107 Felsenstein J. 2002. Quantitative characters, phylogenies, and morphometrics. In:
1108 MacLeod N, Forey PL editors. Morphology, Shape & Phylogeny. London, Taylor &
1109 Francis, p. 27-44.
1110 Felsenstein J. 2004. Inferring phylogenies. Sunderland, MA, Sinauer Associates.
1111 Felsenstein J. 2013. PHYLIP (Phylogeny Inference Package). Seattle, WA,
1112 Department of Genome Sciences, University of Washington.
1113 Figueirido B, Serrano-Alarcón FJ, Slater GJ, Palmqvist P. 2010. Shape at the cross-
1114 roads: homoplasy and history in the evolution of the carnivoran skull towards
1115 herbivory. J. Evol. Biol., 23:2579–2594.
1116 Fink WL, Zelditch ML. 1995. Phylogenetic analysis of ontogenetic shape
1117 transformations: a reassessment of the piranha genus *Pygocentrus* (Teleostei). Syst.
1118 Biol., 44:343–360.
1119 Forbes C, Evans M, Hastings N, Peacock B. 2011. Statistical distributions. 4th ed.
1120 Hoboken, NJ, Wiley.

1121 Frédérich B, Sorenson L, Santini F, Slater GJ, Alfaro ME. 2013. Iterative ecological
1122 radiation and convergence during the evolutionary history of damselfishes
1123 (Pomacentridae). Am. Nat., 181:94–113.
1124 Friedman M. 2010. Explosive morphological diversification of spiny-finned teleost
1125 fishes in the aftermath of the end-Cretaceous extinction. Proc. R. Soc. Lond. Ser. B-
1126 Biol. Sci., 277:1675–1683.
1127 Gabelaia M, Adriaens D, Tarkhnishvili D. 2017. Phylogenetic signals in scale shape
1128 in Caucasian rock lizards (*Darevskia* species). Zool. Anz., 268:32–40.
1129 Galland M, Friess M. 2016. A three-dimensional geometric morphometrics view of
1130 the cranial shape variation and population history in the New World. Am. J. Hum.
1131 Biol., 28:646–661.
1132 Galland M, Van Gerven DP, von Cramon-Taubadel N, Pinhasi R. 2016. 11,000 years
1133 of craniofacial and mandibular variation in Lower Nubia. Sci. Rep., 6:31040.
1134 Gallego Romero I, Ruvinsky I, Gilad Y. 2012. Comparative studies of gene
1135 expression and the evolution of gene regulation. Nat. Rev. Genet., 13:505–516.
1136 Goloboff PA, Catalano SA. 2011. Phylogenetic morphometrics (II): Algorithms for
1137 landmark optimization. Cladistics, 27:42-51.
1138 Goloboff PA, Mattoni CI, Quinteros AS. 2006. Continuous characters analyzed as
1139 such. Cladistics, 22:589–601.
1140 Gómez-Robles A, Polly PD. 2012. Morphological integration in the hominin
1141 dentition: evolutionary, developmental, and functional factors. Evolution, 66:1024–
1142 1043.
1143 González-José R, Escapa I, Neves WA, Cúneo R, Pucciarelli HM. 2008. Cladistic
1144 analysis of continuous modularized traits provides phylogenetic signals in *Homo*
1145 evolution. Nature, 453:775–778.

González-José R, Escapa I, Neves WA, Cúneo R, Pucciarelli HM. 2011. Morphometric variables can be analyzed using cladistic methods: A Reply to Adams et al. J. Hum. Evol., 60:244-245.

Goswami A, Smaers JB, Soligo C, Polly PD. 2014. The macroevolutionary consequences of phenotypic integration: from development to deep time. Philos. Trans. R. Soc. Lond. B Biol. Sci., 369:20130254.

Hansen TF. 1997. Stabilizing selection and the comparative analysis of adaptation. Evolution, 51:1341-1351.

Harmon LJ, Losos JB, Davies TJ, Gillespie RG, Gittleman JL, Jennings WB, Kozak KH, Schluter D, Schulte JA, II, Seehausen O, Sidlauskas BL, Torres-Carvajal O, Weir JT, Mooers AØ. 2010. Early bursts of body size and shape evolution are rare in comparative data. Evolution, 64:2385–2396.

Hautier L, Lebrun R, Cox PG. 2012. Patterns of covariation in the masticatory apparatus of hystricognathous rodents: implications for evolution and diversification. J. Morphol., 273:1319–1337.

Hillis DM, Huelsenbeck JP, Cunningham CW. 1994. Application and accuracy of molecular phylogenies. Science, 264:671–677.

Huelsenbeck JP. 1995. Performance of phylogenetic methods in simulation. Syst. Biol., 44:17–48.

Huelsenbeck JP, Hillis DM. 1993. Success of phylogenetic methods in the four-taxon case. Syst. Biol., 42:247-264.

Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science, 294:2310–2314.

1169 Huey RB, Bennett AF. 1987. Phylogenetic studies of coadaptation: preferred
1170 temperatures versus optimal performance temperatures of lizards. Evolution,
1171 41:1098–1115.
1172 Hunt G. 2007. The relative importance of directional change, random walks, and
1173 stasis in the evolution of fossil lineages. Proc. Natl. Acad. Sci. USA, 104:18404–
1174 18408.
1175 Kalinka AT, Varga KM, Gerrard DT, Preibisch S, Corcoran DL, Jarrells J, Ohler U,
1176 Bergman CM, Tomancak P. 2010. Gene expression divergence recapitulates the
1177 developmental hourglass model. Nature, 468:811–814.
1178 Karanovic T, Djurakic M, Eberhard SM. 2016. Cryptic species or inadequate
1179 taxonomy? Implementation of 2D geometric morphometrics based on integumental
1180 organs as landmarks for delimitation and description of copepod taxa. Syst. Biol.,
1181 65:304–327.
1182 Kendall DG, Barden D, Carne TK, Le H. 1999. Shape and shape theory. Chichester,
1183 Wiley.
1184 Kimmel CB, Small CM, Knope ML. 2017. A rich diversity of opercle bone shape
1185 among teleost fishes. PLOS ONE, 12:e0188888.
1186 Kitching IJ, Forey PL, Humphries CJ, Williams DM. 1998. Cladistics: the theory and
1187 practice of parsimony analysis. 2nd ed. Oxford, Oxford University Press.
1188 Klingenberg CP. 2008. Morphological integration and developmental modularity.
1189 Annu. Rev. Ecol. Evol. Syst., 39:115-132.
1190 Klingenberg CP. 2013. Cranial integration and modularity: insights into evolution and
1191 development from morphometric data. Hystrix, 24:43–58.
1192 Klingenberg CP. 2015. Analyzing fluctuating asymmetry with geometric
1193 morphometrics: concepts, methods, and applications. Symmetry, 7:843–934.

1194 Klingenberg CP, Barluenga M, Meyer A. 2002. Shape analysis of symmetric
1195 structures: quantifying variation among individuals and asymmetry. Evolution,
1196 56:1909–1920.
1197 Klingenberg CP, Duttke S, Whelan S, Kim M. 2012. Developmental plasticity,
1198 morphological variation and evolvability: a multilevel analysis of morphometric
1199 integration in the shape of compound leaves. J. Evol. Biol., 25:115–129.
1200 Klingenberg CP, Gidaszewski NA. 2010. Testing and quantifying phylogenetic
1201 signals and homoplasy in morphometric data. Syst. Biol., 59:245-261.
1202 Klingenberg CP, Marugán-Lobón J. 2013. Evolutionary covariation in geometric
1203 morphometric data: analyzing integration, modularity and allometry in a phylogenetic
1204 context. Syst. Biol., 62:591–610.
1205 Klingenberg CP, Monteiro LR. 2005. Distances and directions in multidimensional
1206 shape spaces: Implications for morphometric applications. Syst. Biol., 54:678-688.
1207 Koehl P, Hass J. 2015. Landmark-free geometric methods in biological shape
1208 analysis. J. R. Soc. Interface, 12:20150795.
1209 Ledoit O, Wolf M. 2004. A weel-conditioned estimator for large-dimensional
1210 covariance matrices. J. Multivariate Anal., 88:365–411.
1211 Lockwood CA, Kimbel WH, Lynch JM. 2004. Morphometrics and hominoid
1212 phylogeny: Support for a chimpanzee-human clade and differentiation among great
1213 ape subspecies. Proc. Natl. Acad. Sci. USA, 101:4356-4360.
1214 Lynch M. 1989. Phylogenetic hypotheses under the assumption of neutral
1215 quantitative-genetic variation. Evolution, 43:1-17.
1216 Macholán M. 2006. A geometric morphometric analysis of the shape of the first upper
1217 molar in mice of the genus *Mus* (Muridae, Rodentia). J. Zool. (Lond.), 270:672–681.

1218 MacLeod N. 2002. Phylogenetic signals in morphometric data. In: MacLeod N, Forey
1219 PL editors. Morphology, shape and phylogeny. London, Taylor & Francis, p. 100–
1220 138.
1221 Maddison WP. 1991. Squared-change parsimony reconstructions of ancenstral states
1222 for continuous-valued characters on a phylogenetic tree. Syst. Zool., 40:304-314.
1223 Maddison WP. 1997. Gene trees in species trees. Syst. Biol., 46:523–536.
1224 Marcus LF, Hingst-Zaher E, Zaher H. 2000. Application of landmark morphometrics
1225 to skulls representing the orders of living mammals. Hystrix, 11:27–47.
1226 Mardia KV, Kent JT, Bibby JM. 1979. Multivariate analysis. London, Academic
1227 Press.
1228 Martín-Serra A, Figueirido B, Palmqvist P. 2014. A three-dimensional analysis of
1229 morphological evolution and locomotor performance of the carnivoran forelimb.
1230 PLoS ONE, 9:e85574.
1231 Martín-Serra A, Figueirido B, Pérez-Claros JA, Palmqvist P. 2015. Patterns of
1232 morphological integration in the appendicular skeleton of mammalian carnivores.
1233 Evolution, 69:321–340.
1234 Martins EP. 1999. Estimation of ancestral states of continuous characters: a computer
1235 simulation study. Syst. Biol., 48:642–650.
1236 McArdle B, Rodrigo AG. 1994. Estimating the ancestral states of a continuous-valued
1237 character using squared-change parsimony: An analytical solution. Syst. Biol.,
1238 43:573-578.
1239 Monteiro LR. 2000. Why morphometrics is special: the problem with using partial
1240 warps as characters for phylogenetic inference. Syst. Biol., 49:796–800.

1241 Monteiro LR, Bonato V, dos Reis SF. 2005. Evolutionary integration and
1242 morphological diversification in complex morphological structures: Mandible shape
1243 divergence in spiny rats (Rodentia, Echimyidae). Evol. Dev., 7:429–439.
1244 Monteiro LR, Nogueira MR. 2011. Evolutionary patterns and processes in the
1245 radiation of phyllostomid bats. BMC Evol. Biol., 11:137.
1246 Naylor GJP. 1996. Can partial warp scores be used as cladistic characters? In: Marcus
1247 LF, Corti M, Loy A, Naylor GJP, Slice DE editors. Advances in morphometrics. New
1248 York, Plenum Press, p. 519–530.
1249 Olson EC, Miller RL. 1958. Morphological integration. Chicago, University of
1250 Chicago Press.
1251 Ospina-Garcés SM, de Luna E. 2017. Phylogenetic analysis of landmark data and the
1252 morphological evolution of cranial shape and diets in species of *Myotis* (Chiroptera:
1253 Vespertilionidae). Zoomorphology (Berl.), 136:251–265.
1254 Palci A, Lee MSY. 2019. Geometric morphometrics, homology and cladistics: review
1255 and recommendations. Cladistics, 35:230–242.
1256 Paradis E, Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and
1257 evolutionary analyses in R. Bioinformatics (Oxf.), 35:526–528.
1258 Parins-Fukuchi C. 2018a. Bayesian placement of fossils on phylogenies using
1259 quantitative morphometric data. Evolution, 72:1801–1814.
1260 Parins-Fukuchi C. 2018b. Use of continuous traits can improve morphological
1261 phylogenetics. Syst. Biol., 67:328–339.
1262 Pavlicev M, Cheverud JM, Wagner GP. 2009. Measuring morphological integration
1263 using eigenvalue variance. Evol. Biol., 36:157–170.
1264 Pečnerová P, Moravec JC, Martínková N. 2015. A skull might lie: modeling ancestral
1265 ranges and diet from genes and shape of tree squirrels. Syst. Biol., 64:1074–1088.

1266 Perrard A, Lopez-Osorio F, Carpenter JM. 2016. Phylogeny, landmark analysis and
1267 the use of wing venation to study the evolution of social wasps (Hymenoptera:
1268 Vespidae: Vespinae). Cladistics, 32:406–425.
1269 Philippe H, Zhou Y, Brinkmann H, Rodrigue N, Delsuc F. 2005. Heterotachy and
1270 long-branch attraction in phylogenetics. BMC Evol. Biol., 5:50.
1271 Piras P, Colangelo P, Adams DC, Buscalioni A, Cubo J, Kotsakis T, Meloro C, Raia
1272 P. 2010. The *Gavialis–Tomistoma* debate: the contribution of skull ontogenetic
1273 allometry and growth trajectories to the study of crocodylian relationships. Evol.
1274 Dev., 12:568–579.
1275 Polly PD. 2001. On morphological clocks and paleophylogeography: towards a
1276 timescale for *Sorex* hybrid zones. Genetica, 112–113:339–357.
1277 Polly PD. 2003a. Palaeophylogeography: the tempo and mode of geographic
1278 differentiation in martmots (*Marmota*). J. Mammal., 84:369–384.
1279 Polly PD. 2003b. Paleophylogeography of *Sorex araneus* (Insectivora, Soricidae):
1280 molar shape as a morphological marker for fossil shrews. Mammalia, 68:233–243.
1281 Polly PD. 2004. On the simulation of the evolution of morphological shape:
1282 multivariate shape under selection and drift. Palaeontol. Electron., 7:7A.
1283 Prömel HJ, Steger A. 2002. The Steiner tree problem: a tour through graphs,
1284 algorithms, and complexity. Braunschweig, Vieweg.
1285 R Core Team. 2013. R: A language and environment for statistical computing.
1286 Vienna, Austria, R Foundation for Statistical Computing.
1287 Rannala B, Yang Z. 2008. Phylogenetic inference using whole genomes. Annu. Rev.
1288 Genomics Hum. Genet., 9:217–231.
1289 Revell LJ, Harmon LJ, Collar DC. 2008. Phylogenetic signal, evolutionary process,
1290 and rate. Syst. Biol., 57:591–601.

1291 Rifkin SA, Kim J, White KP. 2003. Evolution of gene expression in the *Drosophila*
1292 *melanogaster* subgroup. Nat. Genet., 33:138-144.
1293 Robinson DF, Foulds LR. 1981. Comparison of phylogenetic trees. Math. Biosci.,
1294 53:131–147.
1295 Rohlf FJ. 1998. On applications of geometric morphometrics to studies of ontogeny
1296 and phylogeny. Syst. Biol., 47:147-158.
1297 Saitou N, Nei M. 1987. The neighbor-joining method: a new method for
1298 reconstructing phylogenetic trees. Mol. Biol. Evol., 4:406–425.
1299 Santana SE, Lofgren SE. 2013. Does nasal echolocation influence the modularity of
1300 the mammal skull? J. Evol. Biol., 26:2520–2526.
1301 Schluter D, Price T, Mooers AØ, Ludwig D. 1997. Likelihood of ancestor states in
1302 adaptive radiation. Evolution, 51:1699–1711.
1303 Schroeder L, Scott JE, Garvin HM, Laird MF, Dembo M, Radovčić D, Berger LR, de
1304 Ruiter DJ, Ackermann RR. 2017. Skull diversity in the *Homo* lineage and the relative
1305 position of *Homo naledi*. J. Hum. Evol., 104:124–135.
1306 Sherratt E, Gower DJ, Klingenberg CP, Wilkinson M. 2014. Evolution of cranial
1307 shape in caecilians (Amphibia: Gymnophiona). Evol. Biol., 41:528–545.
1308 Sidlauskas B. 2008. Continuous and arrested morphological diversification in sister
1309 clades of characiform fishes: a phylomorphospace approach. Evolution, 62:3135–
1310 3156.
1311 Smith MR. 2019a. Bayesian and parsimony approaches reconstruct informative trees
1312 from simulated morphological datasets. Biol. Lett., 15:20180632.
1313 Smith MR. 2019b. Quartet: comparison of phylogenetic trees using quartet and
1314 bipartition measures. doi: doi:10.5281/zenodo.2536318.

1315 Smith UE, Hendricks JR. 2013. Geometric morphometrics character suites as
1316 phylogenetic data: extracting phylogenetic signal from gastropod shells. Syst. Biol.,
1317 62:366–385.
1318 Smith WD. 1992. How to find Steiner minimal trees in Euclidean *d*-space.
1319 Algorithmica, 7:137–177.
1320 Sneath PHA, Sokal RR. 1973. Numerical taxonomy: the principles and practice of
1321 numerical classification. San Francisco, W. H. Freeman.
1322 Stayton CT. 2008. Is convergence surprising? An examination of the frequency of
1323 convergence in simulated datasets. J Theor Biol, 252:1-14.
1324 Steel MA, Penny D. 1993. Distributions of tree comparison metrics—some new
1325 results. Syst. Biol., 42:126–141.
1326 Swiderski DL, Zelditch ML, Fink WL. 1998. Why morphometrics is not special:
1327 Coding quantitative data for phylogenetic analysis. Syst. Biol., 47:508-519.
1328 Swofford DL, Maddison WP. 1987. Reconstructing ancestral character states under
1329 Wagner parsimony. Math. Biosci., 87:199–229.
1330 Swofford DL, Olsen GJ, Waddell PJ, Hillis DM. 1996. Phylogenetic inference. In:
1331 Hillis DM, Moritz C, Mable BK editors. Molecular systematics. Sunderland, MA,
1332 Sinauer, p. 407–514.
1333 Thompson EA. 1973. The method of minimum evolution. Ann. Hum. Genet., 36:333–
1334 340.
1335 Uddin M, Wildman DE, Liu G, Xu W, Johnson RM, Hof PR, Kapatos G, Grossman
1336 LI, Goodman M. 2004. Sister grouping of chimpanzees and humans as revealed by
1337 genome-wide phylogenetic analysis of brain gene expression profiles. Proc. Natl.
1338 Acad. Sci. USA, 101:2957–2962.

1339 Wägele JW, Mayer C. 2007. Visualizing differences in phylogenetic information
1340 content of alignments and distinction of three classes of long-branch effects. BMC
1341 Evol. Biol., 7:147.
1342 Wagner GP. 1984. On the eigenvalue distribution of genetic and phenotypic
1343 dispersion matrices: evidence for a nonrandom organization of quantitative character
1344 variation. J. Math. Biol., 21:77–95.
1345 Wagner GP, Pavlicev M, Cheverud JM. 2007. The road to modularity. Nat. Rev.
1346 Genet., 8:921-931.
1347 Watanabe A, Slice DE. 2014. The unitility of cranial ontogeny for phylogenetic
1348 inference: a case study in crocodylians using geometric morphometrics. J. Evol. Biol.,
1349 27:1078–1092.
1350 Wiens JJ, Hollingsworth BD. 2000. War of the iguanas: conflicting molecular and
1351 moprhological phylogenies and long-branch attraction in iguanid lizards. Syst. Biol.,
1352 49:143–159.
1353 Zelditch ML, Fink WL, Swiderski DL. 1995. Morphometrics, homology, and
1354 phylognetics: quantified characters as synapomorphies. Syst. Biol., 44:179–189.
1355 Zelditch ML, Fink WL, Swiderski DL, Lundrigan BL. 1998. On applications of
1356 geometric morphometrics to studies of ontogeny and phylogeny: a reply to Rohlf.
1357 Syst. Biol., 47:159–167.
1358 Zelditch ML, Swiderski DL, Sheets HD. 2012. Geometric morphometrics for
1359 biologists: a primer. 2nd ed. Amsterdam, Elsevier.
1360 Zelditch ML, Ye J, Mitchell JS, Swiderski DL. 2017. Rare ecomorphological
1361 convergence on a complex adaptive landscape: body size and diet mediate evolution
1362 of jaw shape in squirrels (Sciuridae). Evolution, 71:633–649.
1363

FIGURE CAPTIONS

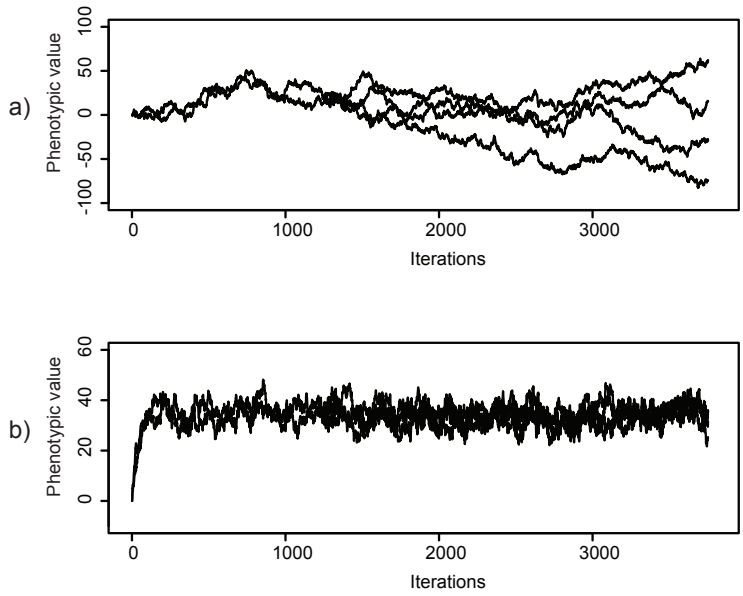

Figure 1. Examples of the two different evolutionary models used in the study. a) Brownian motion model. At each iteration the phenotypic values change randomly. b) Stabilizing selection. At each iteration, the phenotypic values are attracted towards the phenotypic optimum (a phenotypic value of 35 in this case) and also have a small amount of random movement.

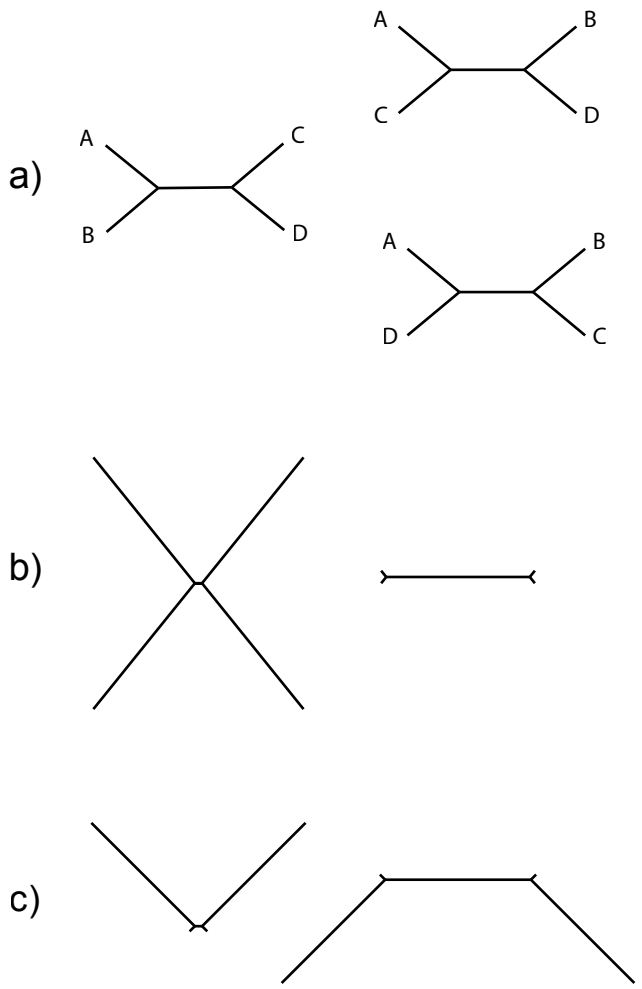

Figure 2. The three possible unrooted trees and two scenarios for varying branch lengths in the simulations. a) True phylogenetic tree simulated (left) and the two other possible tree topologies (right). b) Variation in branch lengths contrasting terminal versus internal branches. All the terminal branches share a length and the internal branch has a different length. The relative lengths of the two sets of branches are varied from 1:20 to 20:1. When the internal branch is very long relative to the terminal branches (right), it is expected that estimating the phylogeny should be reliable. c) Variation in branch lengths contrasting two terminal branches with the three remaining branches (2-versus-3-branch scenario). The situation at the left, where

two terminal branches at either end of the internal branch are much longer than the remaining three branches, is well known to be particularly challenging.

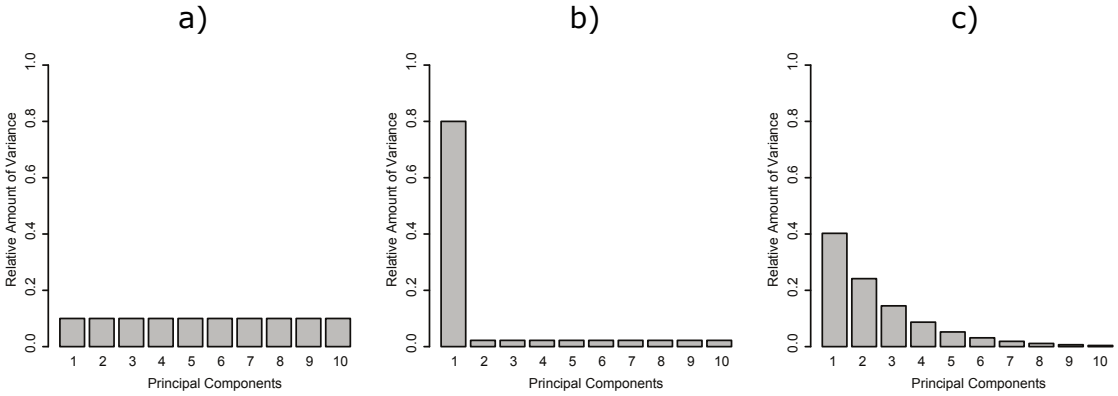

Figure 3. Examples of the models of integration used in the study (shown here for 10 dimensions). a) The model of isotropic Brownian motion, with no integration: all dimensions have the same amount of variation. b) The model of high integration, where the a single dimension accounts for 80% of the total variation and the other dimensions share the remaining 20%. c) The exponential integration model, where the distribution of variances across dimensions of the phenotypic space follows an exponential function, with each dimension accounting for 60% of the variance in the preceding dimension.

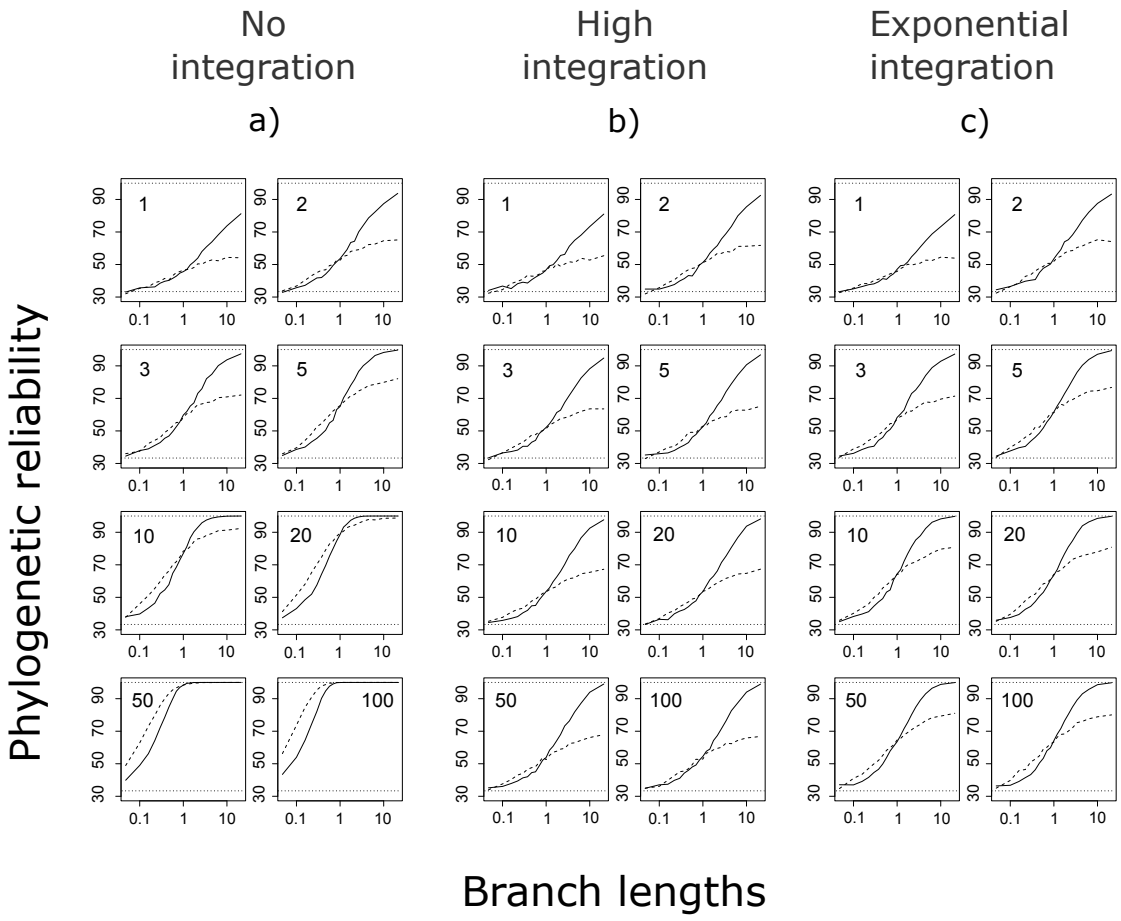

Figure 4. Phylogenetic reliability under Brownian motion models. a) Evolutionary model with isotropic Brownian motion (Experiment 2). b) Model of Brownian motion with high integration (Experiment 3, see Fig. 3b). c) Model of Brownian motion with exponential integration (Experiment 3, see Fig. 3c). The solid lines represent the simulations with the internal-versus-terminal branch scenario, the dashed lines those with the 2-versus-3 branch scenario. In each panel, phylogenetic reliability, as the percentage of correct phylogenetic estimates, is plotted on the vertical axis (dotted horizontal lines at 33.33% and 100%, for randomly chosen trees and perfect reliability) and the branch length ratios on the horizontal axis (logarithmic scaling; challenging scenarios with short internal branch or high long-branch attraction with low ratios, to the left; easier scenarios with higher ratios, to the right). The number at the top of each panel is the dimensionality of the phenotypic space used in the respective set of simulations.

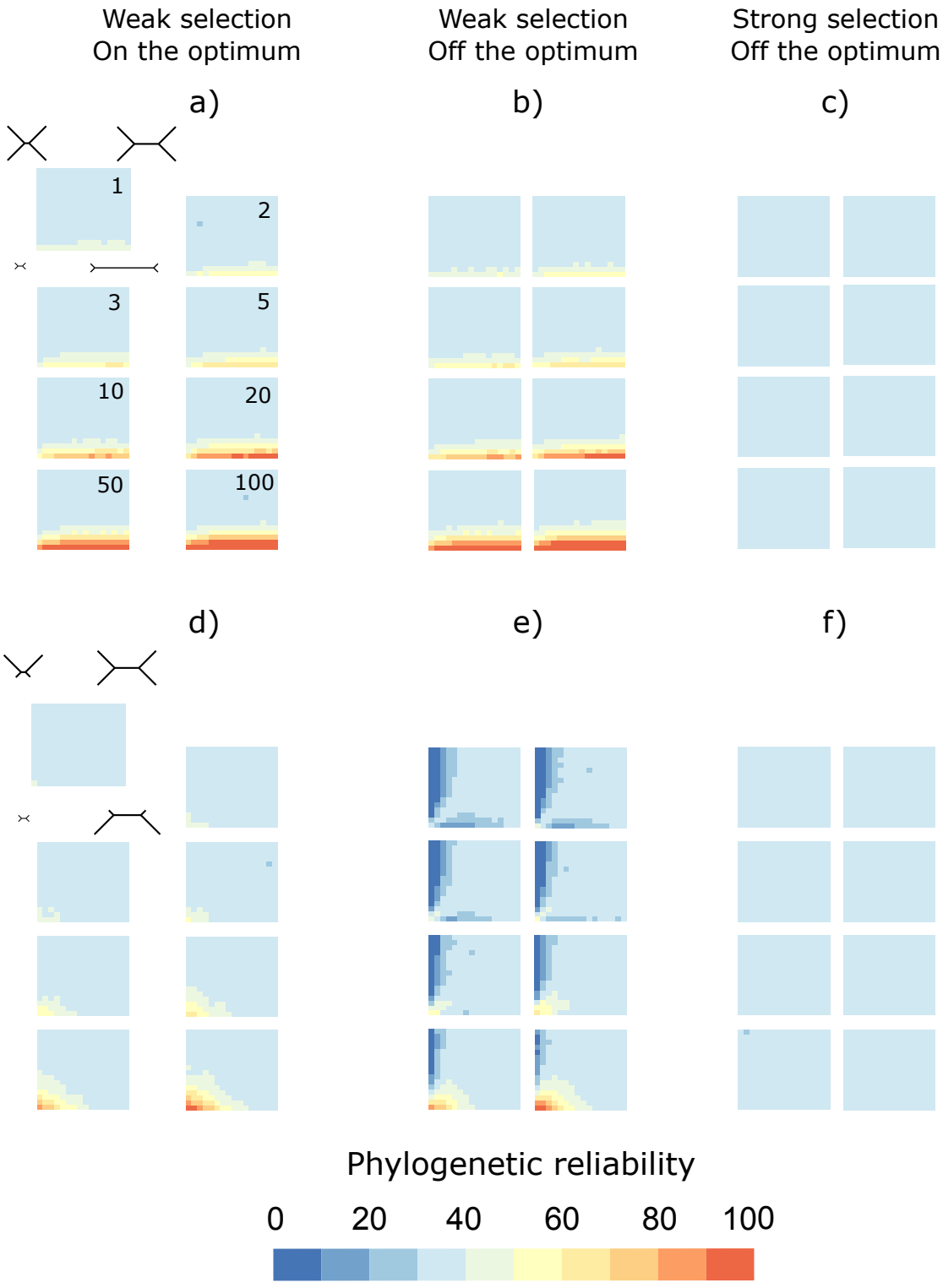

Figure 5. Phylogenetic reliability in the simulations using evolutionary models with stabilizing selection (Experiment 4). Phylogenetic reliability is indicated by color (see color scale at the bottom) as a function of the branch-length combination and dimensionality. For simulations with the internal-versus-terminal branch scenario (a–

c), the *x*-axis in each diagram represents the length of the internal branch and the y-axis the lengths of the terminal branches. For the simulations using the 2-versus-3 branch scenario (d–f), the *y*-axis represents the lengths of two branches at opposite ends of the internal branch and the *x*-axis the lengths of the remaining three branches (i.e., the strongest long-branch attraction occurs in the upper left corner of each panel). The numbers in the panels of part a) indicate the number of dimensions used in the simulations; the other parts use the same arrangement.

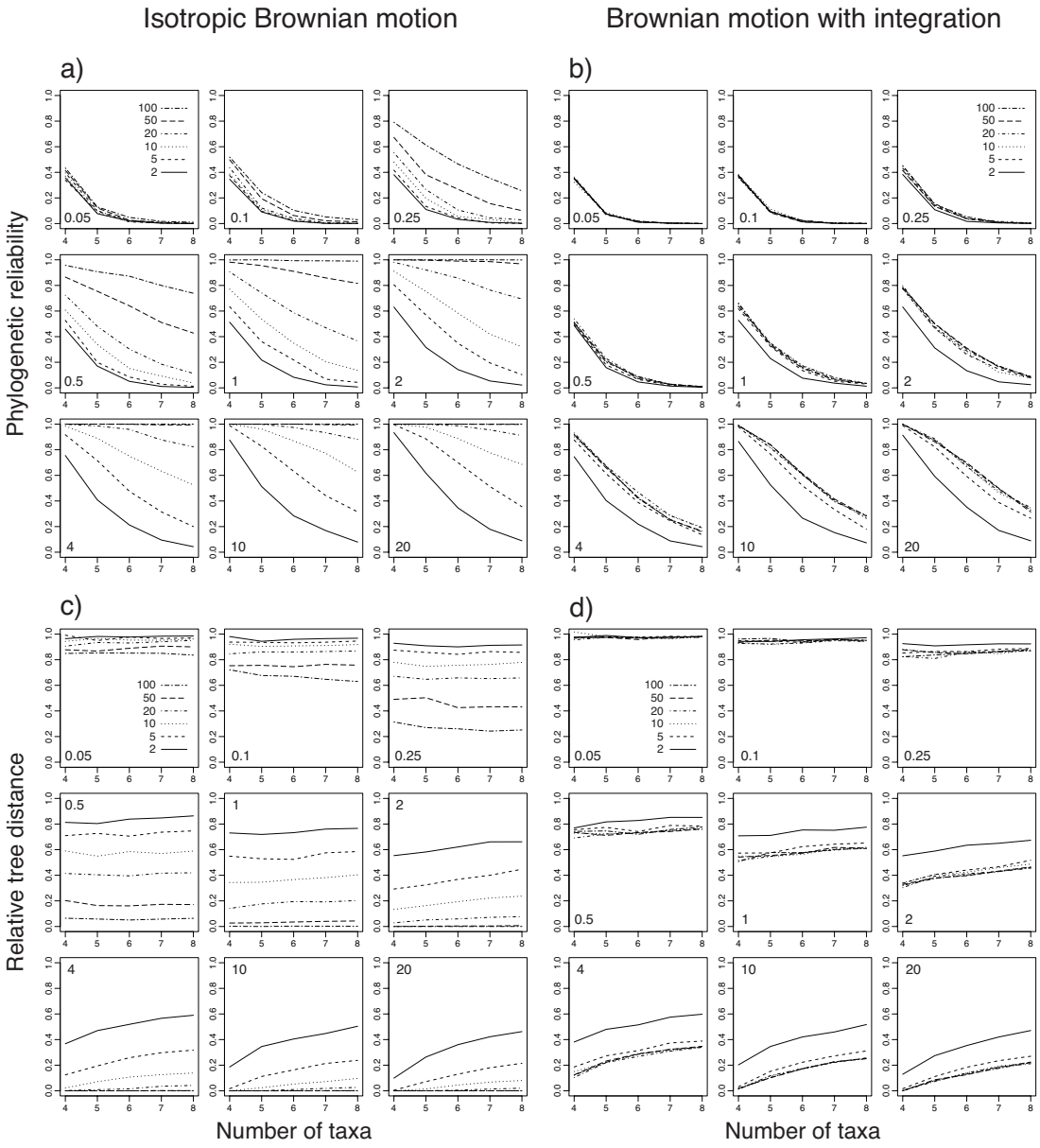

Figure 6. Simulations exploring the effect of the number of taxa (Experiment 5). For taxon numbers ranging from 4 to 8, true phylogenetic trees were drawn randomly from a uniform distribution of all unrooted trees with the respective number of taxa. These trees were used to generate phenotypic data with different dimensionalities, from which the trees then were estimated using squared-change parsimony. a) Phylogenetic reliability (as the proportion of phylogenetic trees estimated correctly) under an evolutionary model of isotropic Brownian motion (Fig. 3a). b) Phylogenetic reliability under the exponential integration model (Fig. 3c). c) Average relative tree

distances between true and estimated trees (scaled relative to the expected distance between two random trees for the respective number of taxa) for the simulations under the model of isotropic Brownian motion. d) Average relative tree distances between true and estimated trees for the simulations under the model of exponential integration. The number in the corner of each panel indicates the ratio of the lengths of the internal branches to the lengths of the terminal branches in the trees used in the respective set of simulations. Dimensionalities of phenotypes are distinguished by the types of lines. The relative tree distances (c,d) are the average of the Robinson–Foulds distances between the true and estimated trees in each set of simulations, divided by the expected Robinson–Foulds distance between pairs of random trees with the respective number of taxa.

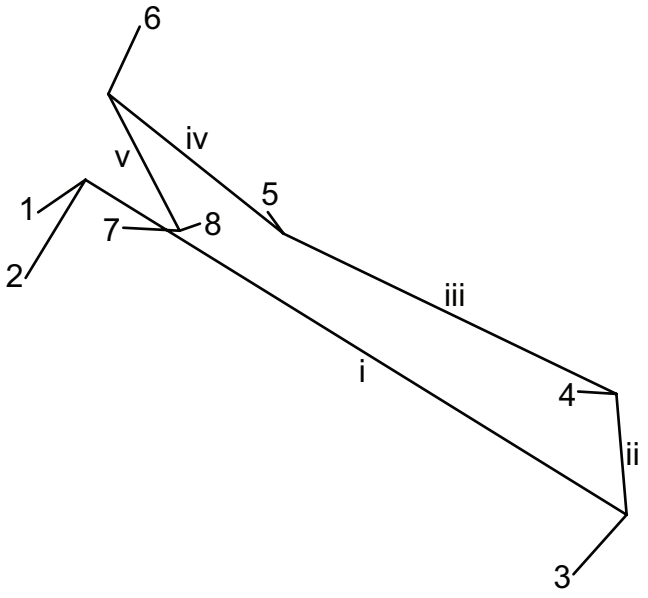

Figure 7. An example of convergence among internal nodes in a phylogeny with 8 taxa. Taxa are numbered 1–8 and internal branches i–v. The example was simulated under a Brownian motion model in a two-dimensional phenotype space (the plane of the graph). Taxa 1 and 2 are phylogenetically as remote from taxa 7 and 8 as it is possible on an 8-taxon tree, yet the two pairs are phenotypically quite close. Similarly,

1445 the sharp angle between branches iv and v brings taxa 7 and 8 very near to taxon 5,

1446 clearly closer than any of them are to taxon 6.

1447

1448 Table 1. Comparisons of different methods for estimating phylogenies.

	Internal versus terminal branches			2 versus 3 branches		
	Tree 1	Tree 2	Tree 3	Tree 1	Tree 2	Tree 3
Squared-change parsimony (rows) versus maximum likelihood (columns)						
Tree 1	703	1	1	843	0	0
Tree 2	0	154	0	4	78	1
Tree 3	1	0	140	0	0	74
Squared-change parsimony (rows) versus Euclidean parsimony (columns)						
Tree 1	703	2	0	819	22	2
Tree 2	2	150	2	0	83	0
Tree 3	1	1	139	0	7	67
Squared-change parsimony (rows) versus Wagner parsimony (columns)						
Tree 1	632	40	33	730	85	28
Tree 2	23	118	13	9	71	3
Tree 3	24	10	107	9	15	50
Squared-change parsimony (rows) versus neighbour joining (columns)						
Tree 1	695	8	2	782	59	2
Tree 2	3	148	3	0	83	0
Tree 3	2	1	138	3	16	55
Squared-change parsimony (rows) versus UPGMA (columns)						
Tree 1	558	82	65	144	689	10
Tree 2	40	96	18	4	79	0
Tree 3	32	24	85	0	68	6
Maximum likelihood (rows) versus Euclidean parsimony (columns)						
Tree 1	701	2	1	819	26	2

Tree 2	3	150	2	0	78	0
Tree 3	2	1	138	0	8	67
Maximum likelihood (rows) versus Wagner parsimony (columns)						
Tree 1	630	40	34	731	88	28
Tree 2	24	118	13	8	68	2
Tree 3	25	10	106	9	15	51
Maximum likelihood (rows) versus neighbour joining (columns)						
Tree 1	693	8	3	782	63	2
Tree 2	4	148	3	0	78	0
Tree 3	3	1	137	3	17	55
Maximum likelihood (rows) versus UPGMA (columns)						
Tree 1	556	82	66	144	693	10
Tree 2	41	96	18	4	74	0
Tree 3	33	24	84	0	69	6
Euclidean parsimony (rows) versus Wagner parsimony (columns)						
Tree 1	633	40	33	725	67	27
Tree 2	23	118	12	13	93	6
Tree 3	23	10	108	10	11	48
Euclidean parsimony (rows) versus neighbour joining (columns)						
Tree 1	698	6	2	782	37	0
Tree 2	1	151	1	0	112	0
Tree 3	1	0	140	3	9	57
Euclidean parsimony (rows) versus UPGMA (columns)						
Tree 1	561	80	65	144	665	10
Tree 2	38	99	16	4	108	0

Tree 3	31	23	87	0	63	6
Wagner parsimony (rows) versus neighbour joining (columns)						
Tree 1	633	23	23	710	28	10
Tree 2	35	123	10	47	119	5
Tree 3	32	11	110	28	11	42
Wagner parsimony (rows) versus UPGMA (columns)						
Tree 1	542	75	62	143	598	7
Tree 2	45	103	20	3	166	2
Tree 3	43	24	86	2	72	7
Neighbour joining (rows) versus UPGMA (columns)						
Tree 1	563	73	64	144	631	10
Tree 2	37	105	15	4	154	0
Tree 3	30	24	89	0	51	6

Tabled values are the counts of how often particular combinations of trees were returned by the two methods in the comparison, for 1,000 simulations per scenario. Two scenarios, corresponding to trees with different branch lengths, were used for simulations with Brownian motion in 50 dimensions: internal versus terminal branches (Fig. 2b), in which the internal branch had a length of 0.3 and the terminal branches had lengths of 1.0, and a scenario of 2 versus 3 branches (Fig. 2c), in which the internal branch and two terminal branches at either end of it had lengths of 0.3 and the two remaining terminal branches had lengths of 1.0. The correct tree in all simulations is tree 1.