Syst. Biol. 69(5):863–883, 2020

DOI:10.1093/sysbio/syaa003
Advance Access publication January 27, 2020

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, UK; and* [2] *Department of Evolutionary Biology, EBC, Uppsala University, Norbyägen 18D, 75236 Uppsala, Sweden*
[*] *Correspondence to be sent to: School of Biological Sciences, University of Manchester, Michael Smith Building, Oxford Road, Manchester M13 9PT, UK; E-mail: cpk@manchester.ac.uk*

Received 19 July 2019; reviews returned 3 January 2020; accepted 17 January 2020
Associate Editor: Matt Friedman

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 have 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, 1998; Swiderski et al. 1998; Bogdanowicz et al. 2005; Clouse et al. 2011) or principal component (PC) scores (MacLeod 2002; González-José et al. 2008, 2011; Aguilar-Medrano et al. 2011; Brocklehurst et al. 2016). These proposals, however, have been criticized for various reasons, especially the decomposition of phenotypic spaces into distinct characters (Bookstein 1994; Naylor 1996; Adams and Rosenberg 1998; Rohlf 1998; Monteiro 2000; Adams et al. 2011; Zelditch et al. 2012). Some authors have advocated methods that use landmarks as characters in cladistic analysis (Catalano et al. 2010, 2015; Goloboff and Catalano 2011; Catalano and Goloboff 2012; Perrard et al. 2016; Catalano and Torres 2017; Dehon et al. 2017; Ospina-Garcés and de Luna 2017; Ascarrunz et al. 2019; Palci and Lee 2019). An alternative is to use methods that avoid dividing the phenotypic variation into characters, but infer trees from distances among taxa using clustering techniques such as neighbor-joining (Polly 2001; Lockwood et al. 2004; Couette et al. 2005; Macholán 2006; Cardini and Elton 2008; Bjarnason et al. 2011, 2015, 2017; Cruz et al. 2012; Galland and Friess 2016; Galland et al. 2016; Schroeder et al. 2017; Ascarrunz et al. 2019), UPGMA (Marcus et al. 2000; Polly 2001; Cardini 2003; Cardini and O'Higgins 2004; Cardini and Elton 2008; Piras et al. 2010; Watanabe and Slice 2014; Koehl and Hass 2015; Pečnerová et al. 2015; Karanovic et al. 2016; Gabelaia et al. 2017; Zelditch et al. 2017), or other clustering methods (Cannon and Manos 2001; Polly 2001; Bjarnason et al. 2011). Other studies have estimated phylogenies from morphometric data using statistical approaches such as maximum likelihood (Cannon and Manos 2001; Polly 2003a,b; Caumul and Polly 2005; González-José et al. 2008; Ascarrunz et al. 2019) or Bayesian methods (Parins-Fukuchi 2018a, b; Álvarez-Carretero et al. 2019). Theoretical studies and computer simulations have demonstrated, however, that random evolutionary processes such as Brownian motion frequently produce convergence, so that phenotypic distance may not be a good indicator of time because divergence and the resulting estimates of phylogenies thus may be unreliable (Lynch 1989; Stayton 2008). A large empirical comparison of a range of methods in 41 morphometric data sets found that different methods tend to produce similar and fairly poor results (Catalano and Torres 2017).

These debates raise the question of how the quality of estimated trees can be assessed. So far, the majority of such assessments have compared trees obtained from morphometric data to reference trees obtained from other evidence, most often from molecular data (Cole et al. 2002; Lockwood et al. 2004; Cardini and Elton 2008; González-José et al. 2008; Klingenberg and Gidaszewski 2010; Catalano and Goloboff 2012; Perrard et al. 2016; Catalano and Torres 2017; Gabelaia et al. 2017; Ascarrunz et al. 2019). This type of comparison, however, can be problematic. First, it is often unclear whether the reference tree accurately represents the phylogeny of the taxa (e.g., because of differences between gene trees and species trees; Maddison 1997). Second, many of these studies produced partial agreement in the trees, so that the results are inherently ambiguous: adherents of a particular method can emphasize that the trees are partly correct, critics can point out that other aspects are wrong. For instance, Smith and Hendricks (2013, p. 377) “consider it impressive” that morphometric characters were able to allocate 33–45% of taxa successfully to their positions in a phylogenetic tree, whereas skeptics might argue that this implies a clear majority of failures. A way to avoid this ambiguity is to use computer simulations of evolution, where the true tree is known with certainty, and to use simple phylogenetic trees, so that there is no equivocation whether an estimated tree is right or wrong. This approach has been used for testing methods to infer phylogenies from molecular data (Huelsenbeck and Hillis 1993; Hillis et al. 1994; Huelsenbeck 1995). Simulations have been used in the context of geometric morphometrics to explore the consequences on phylogenetic inference (Polly 2004; Perrard et al. 2016; Parins-Fukuchi 2018a, b; Álvarez-Carretero et al. 2019). However, the simulations were conducted only under restricted sets of parameters (e.g., dimensionality, patterns of trait integration, branch lengths) and results are, therefore, difficult to generalize.

This study uses several sets of simulations to analyze how accurately phylogenies can be estimated using quantitative multidimensional data and what factors influence the quality of the resulting estimates. We use the four-taxon case as the simplest situation where different unrooted trees are possible (Felsenstein 1978a; Huelsenbeck and Hillis 1993). Because there are just three possible trees, there is no ambiguity whether estimated trees are partly correct or partly incorrect. This approach makes it possible to compare different methods for estimating phylogenies and to examine systematically the effects of different combinations of branch lengths in the phylogeny (Felsenstein 1978a; Huelsenbeck and Hillis 1993). Perhaps more importantly, we implement several models that make different assumptions of how phenotypic traits evolve. Because dimensionality is a fundamental characteristic of multivariate traits and is likely to affect the reliability of phylogeny estimation (Felsenstein 2002), we conduct simulations for different numbers of dimensions. A related concept is phenotypic integration, which reflects how different traits are related to each other and how variation is spread across the dimensions of the phenotypic space (Klingenberg 2008; Goswami et al. 2014). To examine its effect on the reliability of phylogenetic estimates, we conduct simulations with different patterns of integration. Because stabilizing selection has been shown to be an important factor in macroevolution (Estes and Arnold 2007), we include simulations that examine its effect on phylogenetic reliability. Finally, to assess how these results apply to analyses with more than four taxa, we conduct a further series of simulations with up to eight taxa. Together, these simulations assess how reliably phylogenies can be inferred from multidimensional data under a wide range of conditions. By examining the potential and limitations of the methods and of the data, the simulations provide new and decisive information to the debate about the role of multidimensional quantitative data in phylogenetics.

Materials and Methods

Simulation Strategy

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

Our strategy consists of repeatedly running evolutionary simulations for four taxa in a phenotypic space (Fig. 1) and estimating the unrooted tree from the resulting multidimensional phenotypes. The proportion of simulations in which these estimates match the tree topology used in the simulation, the proportion of

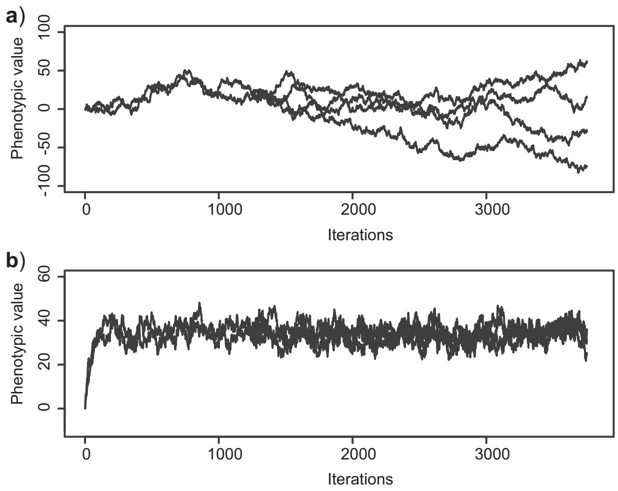

FIGURE 1. Examples of the two different evolutionary models used in the study. a) Brownian motion model. At each iteration,the phenotypic values change randomly. b) Stabilizing selection. At each iteration, the phenotypic values are attracted toward the phenotypic optimum (a phenotypic value of 35 in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

correct estimates, is a natural measure of reliability of the phylogeny reconstruction. Because there are only three possible trees (Fig. 2a), it is feasible to evaluate all three possible trees for each simulation and the analyses are, therefore, guaranteed to find the optimal tree in each simulation. Most importantly, however, it is completely clear that one tree is correct and the other two are incorrect. Therefore, there is none of the ambiguity about whether a reconstructed tree is "mostly correct" or "incorrect in some fundamental features," as it occurs almost inevitably in discussions of empirical examples involving more taxa. A separate set of simulations (Experiment 5, below) explores how the findings from the four-taxon trees extend to analyses with more taxa and also uses methods to quantify how much the true and estimated trees differ.

Evolutionary models.—Our simulations use evolutionary models that are variants of Brownian motion. Brownian motion has been of fundamental importance as an evolutionary model in discussions about phylogenies and quantitative traits (Cavalli-Sforza and Edwards 1967; Felsenstein 1973, 2002; Lynch 1989; Stayton 2008). This model assumes that the phenotype of each lineage evolves by a random change in each short time interval, that this change is equally likely in every direction of the phenotypic space, and that the change is additive over longer time spans. The resulting evolutionary trajectory is a random walk through the phenotypic space (Fig. 1a). Under a Brownian motion model, there is an association between the time since the splitting of two lineages and the expected distance between the corresponding phenotypes, providing a possible basis for estimating phylogeny. This association is not deterministic, however, but has a substantial stochastic component of variation, such that estimating the phylogeny from the distances between the phenotypes of the terminal nodes is inevitably fraught with a degree of uncertainty (Lynch 1989).

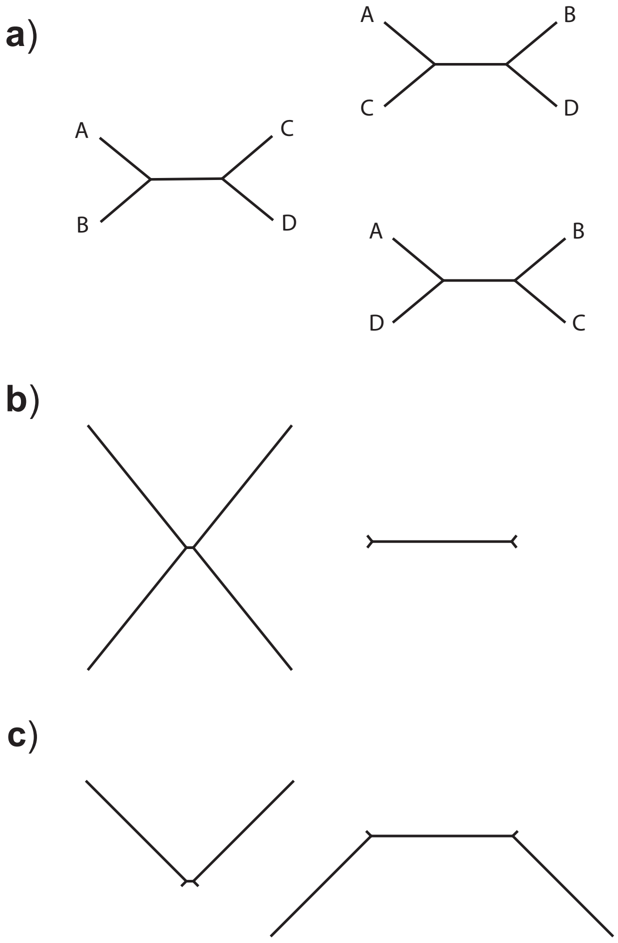

FIGURE 2. The three possible unrooted trees and two scenarios for varying branch lengths in the simulations. a) True phylogenetic tree simulated (left) and the two other possible tree topologies (right). b) Variation in branch lengths contrasting terminal-versus-internal branches. All the terminal branches 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.

To conduct simulations under a Brownian motion model, random walks of lineages through the phenotypic space can be implemented explicitly (Fig. 1a). It is more efficient, however, to obtain changes along the branches in the phylogeny directly as random vectors drawn from multivariate normal distributions with variances proportional to the respective branch lengths and zero covariances among variables (this follows from the multivariate version of the central limit theorem; e.g., Mardia et al. 1979). The phenotypes for the four terminal nodes can then be obtained by combining these changes in accordance with the true phylogenetic tree (tree 1; Fig. 2a). All the simulations were implemented using the R 2.10 statistical package (R Core Team 2013).

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

In the first case, one group contains the four terminal branches and second group consists of just the internal branch (Fig. 2b). Reconstructing the phylogeny should be easier when the internal branch is much longer than the terminal branches, because this situation provides ample opportunity for the two internal nodes to diverge, whereas each of them is likely to remain close to its two adjoining terminal nodes. Conversely, if the internal branch is much shorter than the terminal branches, such that the tree approaches a polytomy, all four taxa are expected to be roughly equidistant to one another and which tree fits the data best is substantially a matter of chance. If the internal branch actually has length zero (i.e., if there is a polytomy), the three possible unrooted trees represent the true tree equally well; in this case, evaluating the phylogenetic reconstruction does not make sense. Whereas these 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 characters. For multidimensional phenotypes, however, the difference matters. Computations for Wagner parsimony minimize the total amount of change for each variable separately, then adding up the resulting amounts across all variables, which corresponds to minimizing the total amount of change over the tree measured as Manhattan distance (Farris 1970; Swofford and Maddison 1987). By contrast, Euclidean parsimony minimizes the sum of changes over all the branches of the tree as Euclidean distances, using the Pythagorean theorem to combine changes in different variables. The task of finding such a tree is known in computer science as the Euclidean Steiner tree problem (Smith 1992; Prömel and Steger 2002; Brazil et al. 2008; Fampa et al. 2016). In the context of phylogenetic analyses of landmark data, some recent studies have used a hybrid approach, called "phylogenetic morphometrics," which combines features of both Wagner and Euclidean parsimony (Catalano et al. 2010; Goloboff and Catalano 2011; Catalano and Goloboff 2012).

To demonstrate the difference between methods, two 4-taxon phylogenies were used: a tree with a short internal branch and long terminal branches (Fig. 2b) and a second tree with two long terminal branches at either end of the internal branch and short remaining branches (Fig. 2c). We ran simulations for two sets of branch lengths, with the short branches at 10% and 30% of the length of the long branches, for which preliminary simulations had shown that they represented challenging conditions for phylogeny estimation. For each set of branch lengths, 1000 simulations of isotropic Brownian motion in 10 dimensions and another 1000 simulations in 50 dimensions were conducted.

For inferring the phylogeny from the phenotypes of the terminal nodes using squared-change parsimony, we used the algorithm of McArdle and Rodrigo (1994) to reconstruct the phenotypes for the internal nodes. Tree length was computed as the total of squared changes, summed over all branches and all variables, and the shortest tree for each simulation was accepted as the estimated tree. The maximum likelihood estimate, under a model of isotropic Brownian motion, was obtained using the *contml* program of the Phylip package (Felsenstein 2013). Euclidean parsimony was implemented using the optimization algorithm of Smith (1992), whereas Wagner parsimony was based on the algorithm by Farris (1970). Neighbor-joining and UPGMA trees were obtained from the matrix of Euclidean distances among phenotypes of the four taxa in each simulation, using the *neighbor* program in Phylip (Felsenstein 2013) with the appropriate settings.

Experiment 2: Detailed Analysis for the Isotropic Brownian Motion Model

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

We conducted separate sets of simulations, one contrasting the internal branch to all four terminal branches (Fig. 2b) and the other contrasting two terminal branches at either end of the internal branch to the other three branches (Fig. 2c). For Brownian motion, the absolute magnitude of the branch lengths affects only the overall scale of distances between taxa, but has no effect on how taxa are arranged relative to one another in phenotype space. This is different from molecular evolution, where there 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 5000 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 5000 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

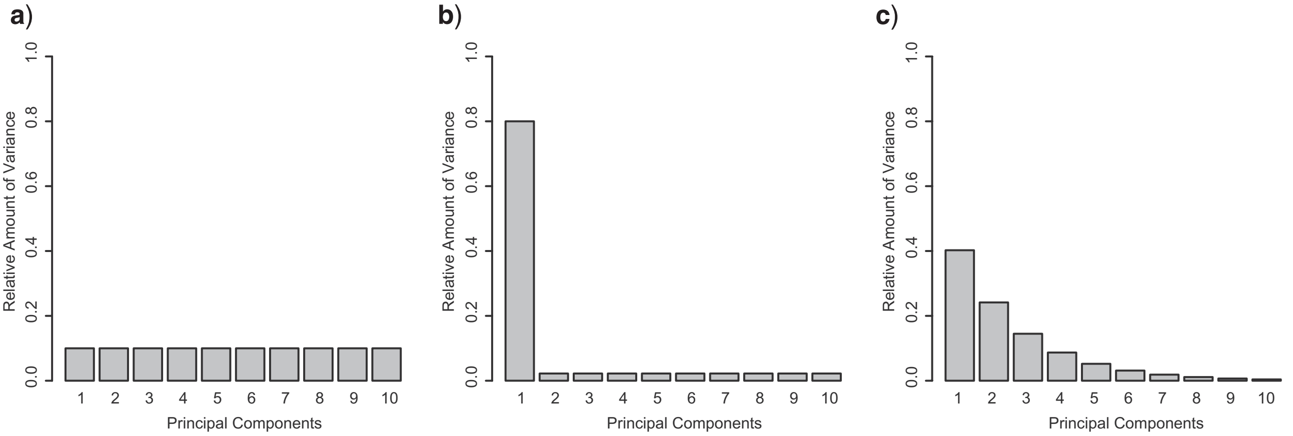

FIGURE 3. Examples of the models of integration used in the study (shown here for 10 dimensions). a) The model of isotropic Brownian motion, with no integration: all dimensions have the same amount of variation. b) The model of high integration, where a single dimension accounts for 80% of the total 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.

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 (Fig. 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 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). Mahalanobis distances are based on a transformation of the phenotypic space that, if the assumptions are met, produces a modified space where variation is isotropic. To achieve this, the transformation relatively shrinks those axes of the phenotypic space that account for much of the total variation and relatively stretches those axes that account for little variation. Usually, this transformation is applied to the variation within groups (Mardia et al. 1979; Klingenberg and Monteiro 2005), but in the present context, the phenotypic space is transformed so that evolutionary variation becomes isotropic. In this modified space, therefore, the effect of evolutionary integration has been removed. This transformation, however, comes with other potentially fundamental changes in the scaling of different dimensions and in the relative arrangement of taxon averages.

If the evolutionary covariance matrix were known, therefore, the phenotypic space could be scaled by the inverse of this matrix, transforming the space to a new space of Mahalanobis distances, in which the isotropic Brownian motion model for evolutionary change would apply. In practice, however, the evolutionary covariance matrix usually is not known, but must be estimated from the available data, which is exceedingly difficult if the phylogeny itself is also unknown (Felsenstein 1973, 1988, 2002). In principle, the phylogeny and evolutionary covariance matrix could be estimated simultaneously, but stringent limits on the relative number of taxa and dimensions of the phenotypic space apply (Felsenstein 2002).

For the purpose of this study, we made a series of assumptions that should be very favorable for phylogeny estimation, even though unrealistic for most clades of actual organisms: evolution is by pure drift, the phenotypic, additive genetic and mutational covariance matrices are proportional, and these covariance matrices are constant across the phylogeny. If these assumptions are met, the within-population covariance matrix can be used as a substitute for the evolutionary covariance matrix to obtain the transformed phenotypic space. Even though these assumptions are unlikely to be met by biological data, we use them in our simulations, as did a previous study (Álvarez-Carretero et al. 2019). We carried out separate simulations using the sample covariance matrix and a shrinkage estimator of the covariance matrix for computing Mahalanobis distances (Ledoit and Wolf 2004; Álvarez-Carretero et al. 2019). For further details, see Supplementary Appendix SA1, available on Dryad at https://doi.org/10.5061/dryad.sk244r4).

Experiment 4: Stabilizing Selection Model

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

The simulations of evolution under stabilizing selection were conducted as explicit random walks, starting from a root of the phylogeny located at the midpoint of the internal branch (Fig. 1b). At each interval from time t to $t+1$, each population changes its position from $\mathbf{x}_t$ to $\mathbf{x}_{t+1}$ following the equation $\mathbf{x}_{t+1}= \mathbf{x}_t + \alpha(\boldsymbol{\theta} - \mathbf{x}_t) + \boldsymbol{\sigma}$, where α is a coefficient indicating the strength of stabilizing selection, $\boldsymbol{\theta}$ is the position of the adaptive peak, and $\boldsymbol{\sigma}$ is an isotropic random deviation, drawn from a 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 are determined by the branch lengths, which were varied in steps of six 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, $\boldsymbol{\theta} = (0, \ldots ,0)$. 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 $\boldsymbol{\theta} = (35, 0, \ldots, 0)$ (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 2000 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, 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 Supplementary Appendix SA2, 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

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 neighbor-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 neighbor-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 neighbor-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 neighbor-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
Neighbor-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

Notes: Tabled values are the counts of how often particular combinations of trees were returned by the two methods in the comparison, for 1000 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.

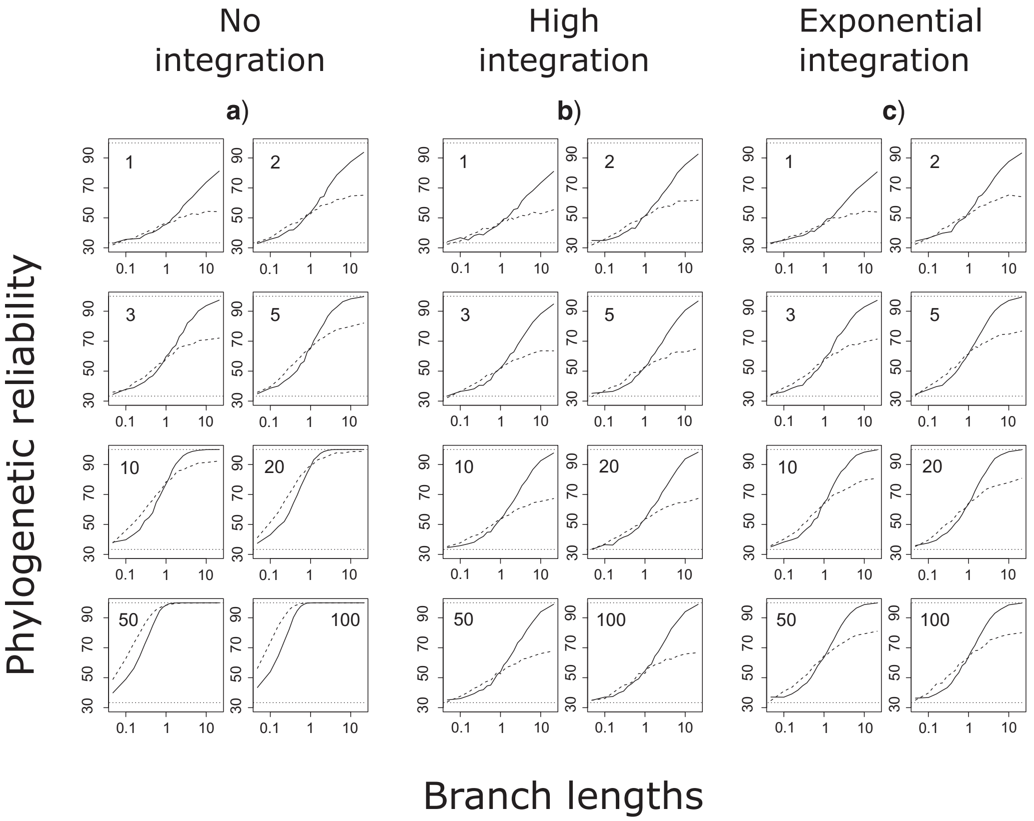

FIGURE 4. Phylogenetic reliability under Brownian motion models. a) Evolutionary model with isotropic Brownian motion (Experiment 2). b) Model of Brownian motion with high integration (Experiment 3, see Fig. 3b). c) Model of Brownian motion with exponential integration (Experiment 3, see Fig. 3c). The 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.

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 becomes more accentuated with increasing dimensionality. From lower to higher dimensionality of the phenotype, the region of high or perfect phylogenetic reliability expands toward shorter relative lengths of the internal branch.

At low dimensionality, the 2-versus-3 branch scenario (dashed lines in Fig. 4a) appears more challenging than the situation where the internal branch is contrasted to the four terminal branches (solid lines in Fig. 4a).

For very high dimensionality, however, the phylogenetic reliability is good even for simulations with a moderate degree of long-branch attraction, where two terminal branches at opposite ends of the internal branch are longer than the remaining three branches (left side of the diagrams in Fig. 4a, dashed lines).

Experiment 3: The Effect of Phenotypic Integration

Phenotypic integration has a strong adverse effect on the accuracy of phylogenetic estimates (Fig. 4b,c). For the model with high integration, where one dimension contains 80% of the total variation (Fig. 3b), there is a change in the relation between branch length ratios and reliability from one to two dimensions, but then this relation remains nearly the same for all simulations with greater dimensionality (Fig. 4b). In contrast to the simulations with isotropic variation (Fig. 4a), where phylogenetic reliability improves with increasing dimensionality, it appears that this improvement ceases after two dimensions for the high-integration model (Fig. 4b). Similarly, for the exponential integration model (Fig. 3c), the benefit of higher dimensionality extends to about five dimensions, but including dimensions beyond that provides no further improvement (Fig. 4c). This loss of the improved phylogenetic reliability with higher-dimensional phenotypes affects both the internal-versus-terminal and the 2-versus-3 branch scenarios (solid and dashed lines in Fig. 4b,c). Under either model of integration, high phylogenetic reliability is only achieved under very special conditions, if the internal branch is extremely long relative to the terminal branches (Fig. 4b,c).

The separate series of simulations using Mahalanobis distance in phylogenetic estimation showed that this approach ameliorated the effects of phenotypic integration partly but not completely. These simulations identified the sample size used to estimate covariance structure as a further complicating factor, and the shrinkage estimate performed somewhat better than sample covariance matrices (for further details, see Supplementary Appendix SA1, available on Dryad).

Experiment 4: The Effect of Stabilizing Selection

Phylogenetic reliability under an evolutionary process with stabilizing selection, for most combinations of branch lengths, is not much better than drawing trees randomly (Fig. 5). If stabilizing selection is weak, the accuracy of the estimates is better where the terminal branches are much shorter than the internal branch (at the bottom of the diagrams in Fig. 5a,b), especially when the dimensionality is high. For the 2-versus-3 branch simulations, reliability is best if all branches are short and more or less equal (lower-left corners of the diagrams in Fig. 5d,e; note that this situation, with all branches short, is similar to the lower-left corners of the diagrams in Fig. 5a,b). A particular situation occurs for the simulation with weak stabilizing selection with an initial phenotype at some distance from the optimum and with strong long-branch attraction. In this situation, only the lineages of the two long terminal branches have time to approach the optimum. Consequently, the incorrect tree ((A,D),(B,C)) tends to be shorter than the correct tree ((A,B),(C,D)). And phylogenetic reliability is systematically worse than drawing trees randomly (left edges of the diagrams in Fig. 5e).

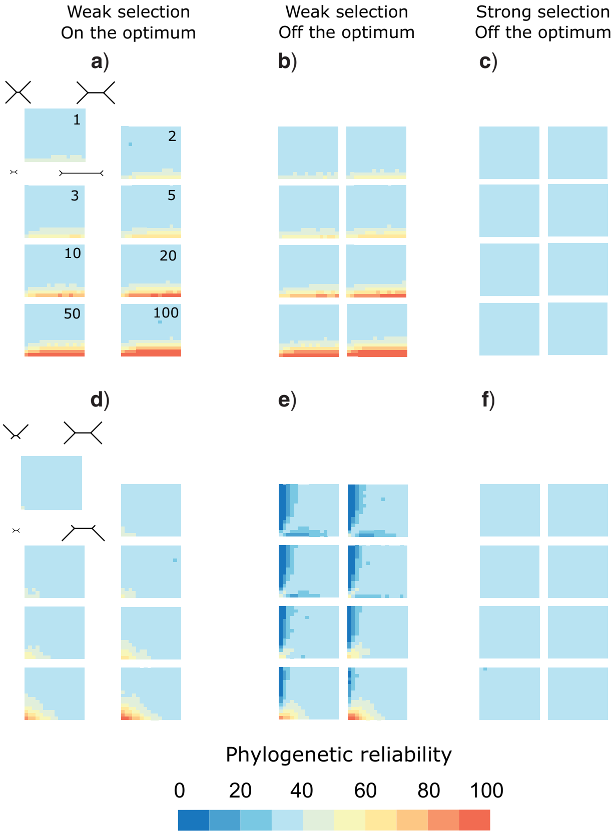

FIGURE 5. Phylogenetic reliability in the simulations using evolutionary models with stabilizing selection (Experiment 4). Phylogenetic reliability is indicated by color (see color scale at the bottom) as a function of the branch-length combination and dimensionality. For simulations with the internal-versus-terminal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

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 dimensionality, in that high-dimensional phenotypes yield lower relative tree distances than low-dimensional phenotypes (Fig. 6c). If there is phenotypic integration, this benefit does not extend beyond approximately five dimensions (Fig. 6d). Simulations with low branch length ratios produce no discernible change of relative tree distances with taxon number. For high branch length ratios, however, there is a gentle but clear trend for relative tree distances to rise with increasing numbers of taxa (Fig. 6c,d). Under the model with isotropic Brownian motion, high-dimensional phenotypes alleviate this trend (Fig. 6c), but when there is phenotypic integration, the trend is clearly apparent no matter how high the dimensionality of the phenotypic space (Fig. 6d).

To examine whether averaging over random tree topologies for any given number of taxa might obscure some relevant differences due to the topology of the tree used to simulate data, we conducted a set of simulation using specific topologies with extreme tree shapes. Phylogenetic reliability and the distributions of tree distances were similar, indicating that such differences are subtle (for details, see Supplementary Appendix SA2, available on Dryad).

Discussion

The simulations in this study have shown that the accuracy of phylogenetic estimates from multidimensional phenotypes depends on a number of factors: the relative branch lengths in the tree used to generate the data, the dimensionality of the phenotype under study, and the model of how phenotypes evolve. Two particularly important aspects of the evolutionary models are morphological integration and stabilizing selection. Here, we explore these results further and evaluate them in light of published evidence to assess their possible implications for the use of shape or other multivariate phenotypes for estimating phylogenies.

Comparison of Methods

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

There are marked differences in performance between the two variants of linear parsimony and between the two clustering methods, especially in the 2-versus-3-branch scenario. In this situation, Euclidean parsimony is nearly as accurate as squared-change parsimony and maximum likelihood (Table 1; but this does not hold for more extreme branch length ratios, Supplementary Table S2, available on Dryad), whereas Wagner parsimony performs clearly worse (under some circumstances worse than randomly picking trees; see particularly Table 1; Supplementary Table S2, available on Dryad). It seems plausible that this discrepancy relates to the difference in how the two methods combine changes across variables: Euclidean parsimony uses the Pythagorean theorem to combine changes across variables (which involves summing the squared changes on each branch of the tree), whereas Wagner parsimony minimizes changes in each variable separately and then sums them over all variables. Of

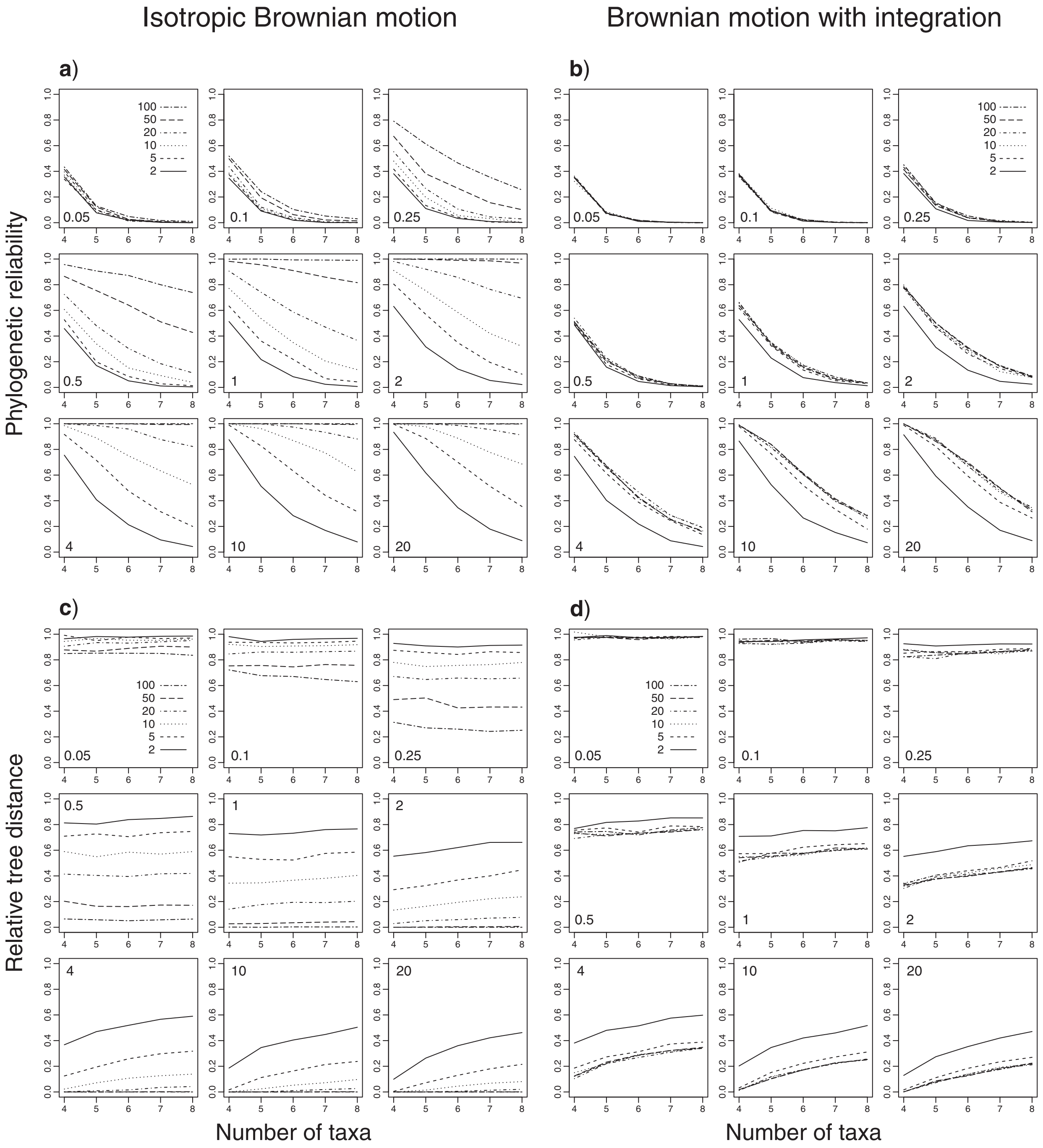

FIGURE 6. Simulations exploring the effect of the number of taxa (Experiment 5). For taxon numbers ranging from 4 to 8, true phylogenetic trees were drawn randomly from a uniform distribution of all unrooted trees with the respective number of taxa. These trees were used to generate phenotypic data with 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.

the two clustering methods, neighbor-joining performed consistently better than UPGMA. The difference is especially clear in the 2-versus-3-branch scenario: in three of the four series of simulations UPGMA was far worse than picking trees at random (Table 1; Supplementary Tables S1–S3, available on Dryad). It is well established that both Wagner parsimony and UPGMA can produce misleading results under long-branch attraction (Felsenstein 1978a; Huelsenbeck and Hillis 1993; Swofford et al. 1996; Felsenstein 2004).

In a methods comparison based on 41 morphometric data sets (Catalano and Torres 2017), Wagner parsimony, neighbor-joining, UPGMA, and the "phylogenetic morphometrics" method that combines Wagner and Euclidean parsimony (Catalano et al. 2010) all produced similar and fairly low degrees of congruence between estimated trees and reference phylogenies, whereas maximum likelihood and Wagner parsimony based on a subset of PC scores performed even slightly worse. Those results are quite different from the simulations in this study (Table 1; Supplementary Tables S1–S3, available on Dryad). It is conceivable that long-branch attraction may only have played a minor role in the 41 empirical data sets, so that the differences between methods were not as manifest as in our simulations. Another difference is that the data sets compiled by (Catalano and Torres, 2017) contained more than 4 taxa (range: 5–160 species), so that phylogeny estimation may have been inherently more challenging (Fig. 6). A further difficulty is that the reference phylogenies were estimated too, based on a variety of data, and that it is unclear how well they reflect the actual phylogenies of the respective clades.

The Effect of Dimensionality

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

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

The squared distance between the phenotypes at either end of a branch of the phylogeny, up to a scaling factor representing the expected magnitude of change along the branch, follows a chi-squared distribution with as many degrees of freedom as there are dimensions in the phenotypic space (this follows from the Pythagorean theorem and the definition of the chi-squared distribution with n degrees of freedom as the sum of squared values of n mutually independent random variates drawn from the standard normal distribution). The coefficient of variation for the chi-squared distribution is the square root of two divided by the square root of the degrees of freedom (Forbes et al. 2011). The relative variability of the phenotypic distances therefore diminishes with increasing degrees of freedom. Note, however, that a substantial improvement is only achieved with dimensionalities that are quite high: the coefficient of variation is 0.44 for 10 dimensions, 0.2 for 50 dimensions, 0.14 for 100 dimensions, and 200 dimensions are necessary for a coefficient of variation of 0.1. As a consequence, increasing dimensionality of a Brownian motion process causes phenotypic distances to become a more deterministic function of divergence times. With increasing dimensionality of the phenotype, the phenotypic distances are therefore expected to be a better reflection of the underlying branch lengths and it should become easier to infer phylogenies from phenotypic divergence.

The benefits of high dimensionality also can be understood intuitively by considering how probable it is for convergent evolution to occur, which is a form of homoplasy and may lead to erroneous phylogenetic inferences. There is always just one direction in which two lineages can converge toward each other in phenotypic space, but with increasing dimensionality, there are more and more directions in which the lineages can move away from each other. Convergence is quite likely in the 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, depend 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 two dimensions (Fig. 4b). This effect is more moderate for the exponential integration model (Fig. 3c), where the benefit stops at approximately five dimensions (Figs. 4c and 6b,d). 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 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 data sets 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 (Supplementary Appendix SA1, 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 in the phenotypic spaces will add up to the dimensionality of variation in the combined phenotypic space. The scenario of no integration at all is grossly unrealistic for actual morphological data, but how closely actual data can approximate the limiting scenario of complete integration is not clear. Although we are not aware of any examples of complete evolutionary integration, empirical studies show that associations among different structures are widespread and often strong (Gómez-Robles and Polly 2012; Hautier et al. 2012; Claverie and Patek 2013; Álvarez et al. 2015; Martín-Serra et al. 2015). Due to such evolutionary integration, combining data from multiple structures in phylogenetic analyses therefore is likely to provide only limited gains of phylogenetic reliability.

Stabilizing Selection

When the evolutionary model used in the simulations includes stabilizing selection, phylogenetic reliability drops and, for most simulations, is little better than for picking a tree at random (Fig. 5). For the simulations with strong stabilizing selection, this applies regardless of the dimensionality or branch length combinations used (Fig. 5c,f). In simulations with weak stabilizing selection, a combination of high dimensionality and a true phylogeny with a long internal branch and short terminal branches yielded a limited zone of better phylogenetic reliability (Fig. 5a,b). As soon as the terminal branches surpass a minimum length, however, even weak stabilizing selection is sufficient to eliminate the phylogenetic signal. In some of the simulations under the 2-versus-3-branch scenario with weak stabilizing selection, there was even a special set of circumstances where phylogenetic reliability was consistently worse than picking trees at random: if simulations started off the optimum and the set of three branches was sufficiently short, only the two lineages of the two long terminal branches tended to reach the optimal phenotype, and the analyses systematically returned the wrong tree (Fig. 5e). Overall, these simulations indicate clearly that stabilizing selection can have a severe detrimental effect on phylogenetic reliability. The reason for this is that stabilizing selection attracts every lineage to the optimum phenotype regardless of ancestry, and thereby erodes the phylogenetic signal. This general result is in agreement with findings from different simulations (Revell et al. 2008).

Because we used a model of stabilizing selection with a single adaptive peak, we need to ask whether using a model with two or more peaks might lead to different conclusions. The answer to this question depends on the processes that control transitions from one peak to another. It is possible to conceive of scenarios giving rise to strong phylogenetic signal, for instance, if clades are associated persistently with different adaptive peaks. Because the taxa within each of these clades would be under the same conditions as in a single-peak model, however, phylogenetic resolution within clades would also be poor. Alternatively, if switches between peaks are so frequent that closely related taxa are commonly associated with different peaks and remotely related taxa with the same peak, convergence will be rampant and phenotypic similarity will indicate association with adaptive peaks, not phylogenetic relatedness.

Whereas evolution under a model of Brownian motion, in principle at least, can continue without bounds, models of stabilizing selection ensure that phenotypes sooner or later converge toward the optimal phenotype. If stabilizing selection is sufficiently strong or the branches are sufficiently long, there is therefore no longer an association between the time of separation and the phenotypic distance between taxa. In other words, the phenotype loses the phylogenetic signal it may have had (see the upper-right regions of the diagrams in Fig. 5). This phenomenon is analogous to the problem of substitution saturation in molecular data, when the product of substitution rate and branch lengths is so large that each position is expected to have undergone multiple substitutions and therefore loses phylogenetic information. This is different from the other models used in this study, where no such phenomenon exists and phenotypic differences are expected to increase with time. In real organisms, however, there cannot be an indefinite amount of change. Simulations of Brownian motion can easily produce phenotypes that are clearly nonfunctional (Polly 2004), so that it seems best to view the models as restricted to a domain of phenotype space within which phenotypes are viable. If phenotypic variation extends to boundaries beyond which phenotypes are not functionally viable, evolving lineages are affected according to their phenotype and regardless of their ancestry, as for stabilizing selection. Therefore, the effect of such boundaries would probably be detrimental to phylogenetic reliability.

Studies of quantitative phenotypes such as morphological traits and gene expression have found extensive evolutionary conservation (e.g., Rifkin et al. 2003; Estes and Arnold 2007; Hunt

2007; Harmon et al. 2010; Kalinka et al. 2010; Gallego Romero et al. 2012) and many comparative analyses reported a good fit of Ornstein–Uhlenbeck models to morphometric data (e.g., Angielczyk et al. 2011; Monteiro and Nogueira 2011; Frédérich et al. 2013; Kimmel et al. 2017; Aristide et al. 2018). These findings support the view that stabilizing selection is widespread. It is therefore likely that many studies attempting to estimate phylogenies from multidimensional phenotypes will face problems similar to those in our simulations.

More Than Four Taxa

Because phylogenetic studies usually involve many more than four taxa, the question arises whether and how the results of our simulations extend to greater 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 picking trees. After this correction for the number of taxa, even though the Robinson–Foulds distance and quartet distance are known to differ in their properties (Steel and Penny 1993; Smith 2019a), they produced similar results in our simulations (Fig. 6c,d; Supplementary Fig. S1, available on Dryad). The main findings from the analyses of distances are consistent with the results on the four-taxon case: higher ratios of internal to terminal branch lengths produce estimated trees that tend to be closer to the true trees, and so does increased dimensionality, but phenotypic integration curtails the benefits of increased dimensionality.

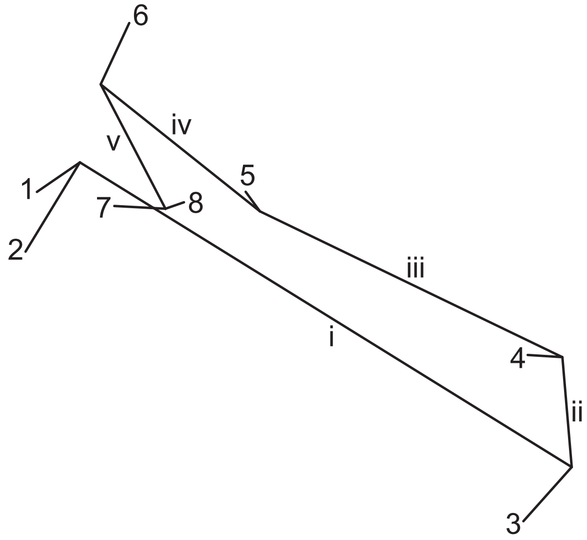

FIGURE 7. An example of convergence among internal nodes in a phylogeny with 8 taxa. Taxa are numbered 1–8 and internal branches i–v. The example was simulated under a Brownian motion model in a two-dimensional phenotype space (the plane of the graph). Taxa 1 and 2 are phylogenetically as remote from taxa 7 and 8 as it is possible on an 8-taxon tree, yet the two pairs are phenotypically quite close. Similarly, the sharp angle between branches iv and v brings taxa 7 and 8 very near to taxon 5, clearly closer than any of them are to taxon 6.

There is an additional result, however, which is not just extending the findings from the simulations with four taxa: when the branch length ratio is high and when dimensionality is low or there is integration, increasing the number of taxa yields a clear rise in the relative tree distances (Fig. 6c,d; Supplementary Fig. S1, available on Dryad). This suggests that, even with long internal branches, an increasing number of taxa poses an additional difficulty that is not present with fewer taxa. The likely reason is that, with more than four taxa, the internal branches of the tree can “fold over” so that taxa that are separated by two or more internal branches in the tree might end up being relatively close to each other in phenotypic space (Fig. 7). This effect depends on the dimensionality of the phenotypic space, because convergence among internal nodes is less likely when dimensionality is high. When dimensionality is low or integration confines variation to just a few dimensions of the phenotypic space, increasing the number of taxa enhances the probability of such convergence. Whereas, for just four taxa, a tree with an internal branch that is much longer than the terminal branches consistently yields accurate estimates of phylogeny under a Brownian motion model even when dimensionality is low (Fig. 4), it is surprisingly 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 article 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 smaller number of more complex trees under a more restricted set of conditions (Perrard et al. 2016; Parins-Fukuchi 2018b). Due to the small number of taxa, there is no ambiguity in the results whether estimated trees are correct or not, and the range of models permitted us not only to determine whether or not multidimensional phenotypic traits are reliable for estimating phylogenies but also to understand why.

The simulations identified the following three key factors: the dimensionality of the trait space, phenotypic integration, and stabilizing selection. Under Brownian motion, high dimensionality is crucial for estimating phylogenetic trees reliably (Fig. 4). Phenotypic integration is detrimental to phylogenetic reliability because variation is limited to just a few of the available dimensions (Fig. 4b,c). Integration is near ubiquitous in morphological structures (Klingenberg 2008; Goswami et al. 2014), suggesting that it imposes widespread limitations on phylogenetic reliability. Because there is no quantitative survey of the strength of evolutionary integration across a broad range of taxa and traits, it is currently impossible to judge how severe these limitations are. Stabilizing selection erodes phylogenetic signal from phenotypic data, and therefore is highly detrimental for estimating phylogenetic trees (Fig. 5). It is a widespread phenomenon (Estes and Arnold 2007), and therefore expected to have adverse effects on phylogenetic analyses using phenotypic data in many clades. Together, these factors conspire so that phylogenetic inference from morphometric data, or other high-dimensional phenotypic data in general, must be expected to be unreliable.

We understand that these results are frustrating to some investigators, particularly to paleontologists, because morphometric data may be the only or at least most easily available data for many fossil and even some extant taxa (MacLeod 2002; Smith and Hendricks 2013; Dehon et al. 2017; Parins-Fukuchi 2018a). Where possible, other data such as genomic sequence information can be used instead, which suffers from these difficulties to a lesser extent and where vast amounts of information are available (Rannala and Yang 2008). Even where such alternatives are not available, however, we think it is preferable to recognize the limitations of phylogenetic inference from such data, rather than to use approaches that may provide unreliable results.

Supplementary Material

Data available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.sk244r4

Acknowledgments

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

Funding

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

References

Adams D.C., Cardini A., Monteiro L.R., O'Higgins P., Rohlf F.J. 2011. Morphometrics and phylogenetics: principal components of shape from cranial modules are neither appropriate nor effective cladistic characters. J. Hum. Evol. 60:240–243.

Adams D.C., Rosenberg M.S. 1998. Partial warps, phylogeny, and ontogeny: a comment on (Fink and Zelditch, 1995). Syst. Biol. 47:168–173.

Aguilar-Medrano R., Frédérich B., de Luna E., Balart E.F. 2011. Patterns of morphological evolution of the cephalic region in damselfishes (Perciformes: Pomacentridae) of the Eastern Pacific. Biol. J. Linn. Soc. 102:593–613.

Álvarez A., Perez S.I., Verzi D.H. 2015. The role of evolutionary integration in the morphological evolution of the skull of caviomorph rodents (Rodentia: Hystricomorpha). Evol. Biol. 42:312–327.

Álvarez-Carretero S., Goswami A., Yang Z., dos Reis M. 2019. Bayesian estimation of species divergence times using correlated quantitative characters. Syst. Biol. 68:967–986.

Angielczyk K.D., Feldman C.R., Miller G.R. 2011. Adaptive evolution of plastron shape in emydine turtles. Evolution. 65:377–394.

Aristide L., Bastide P., dos Reis S.F., Pires dos Santos T.M., Lopes R.T., Perez S.I. 2018. Multiple factors behind early diversification of skull morphology in the continental radiation of New World monkeys. Evolution. 72:2697–2711.

Ascarrunz E., Claude J., Joyce W.G. 2019. Estimating the phylogeny of geoemydid turtles (Cryptodira) from landmark data: an assessment of different methods. PeerJ. 7:e7476.

Baab K.L., Perry J.M.G., Rohlf F.J., Jungers W.L. 2014. Phylogenetic, ecological, and allometric correlates of cranial shape in Malagasy lemuriforms. Evolution. 68:1450–1468.

Bergsten J. 2005. A review of long-branch attraction. Cladistics. 21:163–193.

Bjarnason A., Chamberlain A.T., Lockwood C.A. 2011. A methodological investigation of hominoid craniodental morphology and phylogenetics. J. Hum. Evol. 60:47–57.

Bjarnason A., Soligo C., Elton S. 2015. Phylogeny, ecology, and morphological evolution in the atelid cranium. Int. J. Primatol. 36:513–529.

Bjarnason A., Soligo C., Elton S. 2017. Phylogeny, phylogemetic inference, and cranial evolution in pitheciids and *Aotus*. Am. J. Primatol. 79:e22621.

Blanke A. 2018. Analysis of modularity and integration suggests evolution of dragonfly wing venation mainly in response to functional demands. J. R. Soc. Interface, 15:20180277.

Bogdanowicz W., Juste J., Owen R.D., Sztencel A. 2005. Geometric morphometrics and cladistics: testing evolutionary relationships in mega- and microbats. Acta Chiropt. 7:39–49.

Bookstein F. 1994. Can biometrical shape be a homologous character? In: Hall B.K. editor. Homology: the hierarchial basis of comparative biology. New York: Academic Press. p. 197–227.

Brawand D., Soumillon M., Necsulea A., Julien P., Csárdi G., Harrigan P., Weier M., Liechti A., Aximu-Petri A., Kircher M., Albert F.W., Zeller U., Khaitovich P., Grützner F., Bergmann S., Nielsen R., Pääbo S., Kaessmann H. 2011. The evolution of gene expression levels in mammalian organs. Nature. 478:343–350.

Brazil M., Thomas D.A., Nielsen B.K., Winter P., Wulff-Nilsen C., Zachariasen M. 2008. A novel approach to phylogenetic trees: *d*-dimensional geometric Steiner trees. Networks. 53:104–111.

Brocklehurst N., Romano M., Fröbisch J. 2016. Principal component analysis as an alternative treatment for morphometric characters: phylogeny of caseids as a case study. Palaeontology. 59:877–886.

Brusatte S.L., Sakamoto M., Montanari S., Harcourt Smith W.E.H. 2012. The evolution of cranial form and function in theropod dinosaurs: insights from geometric morphometrics. J. Evol. Biol. 25:365–377.

Cannon C.H., Manos P.S. 2001. Combining and comparing morphometric shape descriptors with a molecular phylogeny: the case of fruit type evolution in Bornean *Lithocarpus* (Fagaceae). Syst. Biol. 50:860–880.

Cardini A. 2003. The geometry of the marmot (Rodentia: Sciuridae) mandible: phylogeny and patterns of morphological evolution. Syst. Biol. 52:186–205.

Cardini A., Elton S. 2008. Does the skull carry a phylogenetic signal? Evolution and modularity in the guenons. Biol. J. Linn. Soc. 93:813–834.

Cardini A., O'Higgins P. 2004. Patterns of morphological evolution in *Marmota* (Rodentia, Sciuridae): geometric morphometrics of the cranium in the context of marmot phylogeny, ecology and conservation. Biol. J. Linn. Soc. 82:385–407.

Catalano S.A., Coloboff P.A., Giannini N.P. 2010. Phylogenetic morphometrics (I): the use of landmark data in a phylogenetic framework. Cladistics. 26:539–549.

Catalano S.A., Ercoli M.D., Prevosti F.J. 2015. The more, the better: the use of multiple landmark configurations to solve the phylogenetic relationships in musteloids. Syst. Biol. 64:294–306.

Catalano S.A., Goloboff P.A. 2012. Simultaneously mapping and superimposing landmark configurations with parsimony as optimality criterion. Syst. Biol. 61:392–400.

Catalano S.A., Torres A. 2017. Phylogenetic inference based on landmark data in 41 empirical data sets. Zool. Scr. 46:1–11.

Caumul R., Polly P.D. 2005. Phylogenetic and environmental components of morphological variation: skull, mandible, and molar shape in marmots (*Marmota*, Rodentia). Evolution. 59:2460–2472.

Cavalli-Sforza L.L., Edwards A.W.F. 1967. Phylogenetic analysis: models and estimation procedures. Evolution. 21:550–570.

Chamero B., Buscalioni Á.D., Marugán-Lobón J. 2013. Pectoral girdle and forelimb variation in extant Crocodylia: the coracoid–humerus pair as an evolutionary module. Biol. J. Linn. Soc. 108:600–618.

Cheverud J.M. 1996. Developmental integration and the evolution of pleiotropy. Amer. Zool. 36:44–50.

Claverie T., Patek S.N. 2013. Modularity and rates of evolutionary change in a power-amplified prey capture system. Evolution. 67:3191–3207.

Clouse R.M., de Bivort B.L., Giribet G. 2011. Phylogenetic signal in morphometric data. Cladistics. 27:337–340.

Cole T.M. III, Lele S.R., Richtsmeier J.T. 2002. A parametric bootstrap approach to the detection of phylogenetic signals in landmark data. In: MacLeod N., Forey P.L. editors. Morphology, shape and phylogeny. London: Taylor & Francis. p. 194–219.

Couette S., Escarguel G., Montuire S. 2005. Constructing, bootstrapping, and comparing morphometric and phylogenetic trees: a case study of New World monkeys (Platyrrhini, Primates). J. Mammal. 86:773–781.

Cruz R.A.L., Pante M.J.R., Rohlf F.J. 2012. Geometric morphometric analysis of shell shape variation in *Conus* (Gastropoda: Conidae). Zool. J. Linn. Soc. 165:296–310.

De Esteban-Trivigno S. 2011a. Buscando patrones ecomorfológicos comunes entre ungulados actuales y xenartros extintos. Ameghiniana. 48:189–209.

De Esteban-Trivigno S. 2011b. Ecomorfología de xenartros extintos: análisis de la mandíbula con métodos de morfometría geométrica. Ameghiniana. 48:381–398.

Degtjareva G.V., Valiejo-Roman C.M., Samigullin T.H., Guara-Requena M., Sokoloff D.D. 2012. Phylogenetics of *Anthyllis* (Leguminosae: Papilionoideae: Loteae): partial incongruence between nuclear and plastid markers, a long branch problem and implications for morphological evolution. Mol. Phylogenet. Evol. 62:693–707.

Dehon M., Perrard A., Engel M.S., Nel A., Michez D. 2017. Antiquity of cleptoparasitism among bees revealed by morphometric and phylogenetic analysis of a Paleocene fossil nomadine (Hymenoptera: Apidae). Syst. Entomol. 42:543–554.

Dryden I.L., Mardia K.V. 1998. Statistical shape analysis. New York: John Wiley & Sons.

Enard W., Khaitovich P., Klose J., Zöllner S., Heissig F., Giavalisco P., Nieselt-Struwe K., Muchmore E., Varki A., Ravid R., Doxiadis G.M., Bontrop R.E., Pääbo S. 2002. Intra- and interspecific variation in primate gene expression patterns. Science. 296:340–343.

Estabrook G.F., McMorris F.R., Meacham C.A. 1985. Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. Syst. Zool. 34:193–200.

Estes S., Arnold S.J. 2007. Resolving the paradox of stasis: models with stabilizing selection explain evolutionary divergence on all timescales. Am. Nat. 169:227–244.

Fampa M., Lee J., Maculan N. 2016. An overview of exact algorithms for the Euclidean Steiner tree problem in *n*-space. Intl. Trans. Op. Res. 23:861–874.

Farris J.S. 1970. Methods for computing Wagner trees. Syst. Zool. 19:83–92.

Felsenstein J. 1973. Maximum-likelihood estimation of evolutionary trees from continuous characters. Am. J. Hum. Genet. 25:471–492.

Felsenstein J. 1978a. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Biol. 27:401–410.

Felsenstein J. 1978b. The number of evolutionary trees. Syst. Zool. 27:27–33.

Felsenstein J. 1981. Evolutionary trees from gene frequencies and quantitative characters: finding maximum likelihood estimates. Evolution. 35:1229–1242.

Felsenstein J. 1988. Phylogenies and quantitative characters. Annu. Rev. Ecol. Syst. 19:455–471.

Felsenstein J. 2002. Quantitative characters, phylogenies, and morphometrics. In: MacLeod N., Forey P.L editors. Morphology, shape & phylogeny. London: Taylor & Francis. p. 27–44.

Felsenstein J. 2004. Inferring phylogenies. Sunderland (MA): Sinauer Associates.

Felsenstein J. 2013. PHYLIP (Phylogeny Inference Package). Seattle (WA): Department of Genome Sciences, University of Washington.

Figueirido B., Serrano-Alarcón F.J., Slater G.J., Palmqvist P. 2010. Shape at the cross-roads: homoplasy and history in the evolution of the carnivoran skull towards herbivory. J. Evol. Biol. 23:2579–2594.

Fink W.L., Zelditch M.L. 1995. Phylogenetic analysis of ontogenetic shape transformations: a reassessment of the piranha genus *Pygocentrus* (Teleostei). Syst. Biol. 44:343–360.

Forbes C., Evans M., Hastings N., Peacock B. 2011. Statistical distributions. 4th ed. Hoboken (NJ): Wiley.

Frédérich B., Sorenson L., Santini F., Slater G.J., Alfaro M.E. 2013. Iterative ecological radiation and convergence during the evolutionary history of damselfishes (Pomacentridae). Am. Nat. 181:94–113.

Friedman M. 2010. Explosive morphological diversification of spiny-finned teleost fishes in the aftermath of the end-Cretaceous extinction. Proc. R. Soc. Lond. Ser. B-Biol. Sci. 277:1675–1683.

Gabelaia M., Adriaens D., Tarkhnishvili D. 2017. Phylogenetic signals in scale shape in Caucasian rock lizards (*Darevskia* species). Zool. Anz. 268:32–40.

Galland M., Friess M. 2016. A three-dimensional geometric morphometrics view of the cranial shape variation and population history in the New World. Am. J. Hum. Biol. 28:646–661.

Galland M., Van Gerven D.P., von Cramon-Taubadel N., Pinhasi R. 2016. 11,000 years of craniofacial and mandibular variation in Lower Nubia. Sci. Rep. 6:31040.

Gallego Romero I., Ruvinsky I., Gilad Y. 2012. Comparative studies of gene expression and the evolution of gene regulation. Nat. Rev. Genet. 13:505–516.

Goloboff P.A., Catalano S.A. 2011. Phylogenetic morphometrics (II): algorithms for landmark optimization. Cladistics. 27:42–51.

Goloboff P.A., Mattoni C.I., Quinteros A.S. 2006. Continuous characters analyzed as such. Cladistics. 22:589–601.

Gómez-Robles A., Polly P.D. 2012. Morphological integration in the hominin dentition: evolutionary, developmental, and functional factors. Evolution. 66:1024–1043.

González-José R., Escapa I., Neves W.A., Cúneo R., Pucciarelli H.M. 2008. Cladistic analysis of continuous modularized traits provides phylogenetic signals in *Homo* evolution. Nature. 453:775–778.

González-José R., Escapa I., Neves W.A., Cúneo R., Pucciarelli H.M. 2011. Morphometric variables can be analyzed using cladistic methods: a reply to Adams et al. J. Hum. Evol. 60:244–245.

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

Hansen T.F. 1997. Stabilizing selection and the comparative analysis of adaptation. Evolution. 51:1341–1351.

Harmon L.J., Losos J.B., Davies T.J., Gillespie R.G., Gittleman J.L., Jennings W.B., Kozak K.H., Schluter D., Schulte J.A., II, Seehausen O., Sidlauskas B.L., Torres-Carvajal O., Weir J.T., Mooers A.Ø. 2010. Early bursts of body size and shape evolution are rare in comparative data. Evolution. 64:2385–2396.

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

Hillis D.M., Huelsenbeck J.P., Cunningham C.W. 1994. Application and accuracy of molecular phylogenies. Science. 264:671–677.

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

Huelsenbeck J.P., Hillis D.M. 1993. Success of phylogenetic methods in the four-taxon case. Syst. Biol. 42:247–264.

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

Huey R.B., Bennett A.F. 1987. Phylogenetic studies of coadaptation: preferred temperatures versus optimal performance temperatures of lizards. Evolution. 41:1098–1115.

Hunt G. 2007. The relative importance of directional change, random walks, and stasis in the evolution of fossil lineages. Proc. Natl. Acad. Sci. USA 104:18404–18408.

Kalinka A.T., Varga K.M., Gerrard D.T., Preibisch S., Corcoran D.L., Jarrells J., Ohler U., Bergman C.M., Tomancak P. 2010. Gene expression divergence recapitulates the developmental hourglass model. Nature. 468:811–814.

Karanovic T., Djurakic M., Eberhard S.M. 2016. Cryptic species or inadequate taxonomy? Implementation of 2D geometric morphometrics based on integumental organs as landmarks for delimitation and description of copepod taxa. Syst. Biol. 65:304–327.

Kendall D.G., Barden D., Carne T.K., Le H. 1999. Shape and shape theory. Chichester: Wiley.

Kimmel C.B., Small C.M., Knope M.L. 2017. A rich diversity of opercle bone shape among teleost fishes. PLoS One 12:e0188888.

Kitching I.J., Forey P.L., Humphries C.J., Williams D.M. 1998. Cladistics: the theory and practice of parsimony analysis. 2nd ed. Oxford: Oxford University Press.

Klingenberg C.P. 2008. Morphological integration and developmental modularity. Annu. Rev. Ecol. Evol. Syst. 39:115–132.

Klingenberg C.P. 2013. Cranial integration and modularity: insights into evolution and development from morphometric data. Hystrix. 24:43–58.

Klingenberg C.P. 2015. Analyzing fluctuating asymmetry with geometric morphometrics: concepts, methods, and applications. Symmetry. 7:843–934.

Klingenberg C.P., Barluenga M., Meyer A. 2002. Shape analysis of symmetric structures: quantifying variation among individuals and asymmetry. Evolution. 56:1909–1920.

Klingenberg C.P., Duttke S., Whelan S., Kim M. 2012. Developmental plasticity, morphological variation and evolvability: a multilevel analysis of morphometric integration in the shape of compound leaves. J. Evol. Biol. 25:115–129.

Klingenberg C.P., Gidaszewski N.A. 2010. Testing and quantifying phylogenetic signals and homoplasy in morphometric data. Syst. Biol. 59:245–261.

Klingenberg C.P., Marugán-Lobón J. 2013. Evolutionary covariation in geometric morphometric data: analyzing integration, modularity and allometry in a phylogenetic context. Syst. Biol. 62:591–610.

Klingenberg C.P., Monteiro L.R. 2005. Distances and directions in multidimensional shape spaces: implications for morphometric applications. Syst. Biol. 54:678–688.

Koehl P., Hass J. 2015. Landmark-free geometric methods in biological shape analysis. J. R. Soc. Interface. 12:20150795.

Ledoit O., Wolf M. 2004. A weel-conditioned estimator for large-dimensional covariance matrices. J. Multivariate Anal. 88:365–411.

Lockwood C.A., Kimbel W.H., Lynch J.M. 2004. Morphometrics and hominoid phylogeny: support for a chimpanzee-human clade and differentiation among great ape subspecies. Proc. Natl. Acad. Sci. USA 101:4356–4360.

Lynch M. 1989. Phylogenetic hypotheses under the assumption of neutral quantitative-genetic variation. Evolution. 43:1–17.

Macholán M. 2006. A geometric morphometric analysis of the shape of the first upper molar in mice of the genus *Mus* (Muridae, Rodentia). J. Zool. (Lond.), 270:672–681.

MacLeod N. 2002. Phylogenetic signals in morphometric data. In: MacLeod N., Forey P.L. editors. Morphology, shape and phylogeny. London: Taylor & Francis. p. 100–138.

Maddison W.P. 1991. Squared-change parsimony reconstructions of ancenstral states for continuous-valued characters on a phylogenetic tree. Syst. Zool. 40:304–314.

Maddison W.P. 1997. Gene trees in species trees. Syst. Biol. 46:523–536.

Marcus L.F., Hingst-Zaher E., Zaher H. 2000. Application of landmark morphometrics to skulls representing the orders of living mammals. Hystrix. 11:27–47.

Mardia K.V., Kent J.T., Bibby J.M. 1979. Multivariate analysis. London: Academic Press.

Martín-Serra A., Figueirido B., Palmqvist P. 2014. A three-dimensional analysis of morphological evolution and locomotor performance of the carnivoran forelimb. PLoS One 9:e85574.

Martín-Serra A., Figueirido B., Pérez-Claros J.A., Palmqvist P. 2015. Patterns of morphological integration in the appendicular skeleton of mammalian carnivores. Evolution. 69:321–340.

Martins E.P. 1999. Estimation of ancestral states of continuous characters: a computer simulation study. Syst. Biol. 48:642–650.

McArdle B., Rodrigo A.G. 1994. Estimating the ancestral states of a continuous-valued character using squared-change parsimony: an analytical solution. Syst. Biol. 43:573–578.

Monteiro L.R. 2000. Why morphometrics is special: the problem with using partial warps as characters for phylogenetic inference. Syst. Biol. 49:796–800.

Monteiro L.R., Bonato V., dos Reis S.F. 2005. Evolutionary integration and morphological diversification in complex morphological structures: mandible shape divergence in spiny rats (Rodentia, Echimyidae). Evol. Dev. 7:429–439.

Monteiro L.R., Nogueira M.R. 2011. Evolutionary patterns and processes in the radiation of phyllostomid bats. BMC Evol. Biol. 11:137.

Naylor G.J.P. 1996. Can partial warp scores be used as cladistic characters? In: Marcus L.F., Corti M., Loy A., Naylor G.J.P., Slice D.E. editors. Advances in morphometrics. New York: Plenum Press. p. 519–530.

Olson E.C., Miller R.L. 1958. Morphological integration. Chicago: University of Chicago Press.

Ospina-Garcés S.M., de Luna E. 2017. Phylogenetic analysis of landmark data and the morphological evolution of cranial shape and diets in species of *Myotis* (Chiroptera: Vespertilionidae). Zoomorphology (Berl.). 136:251–265.

Palci A., Lee M.S.Y. 2019. Geometric morphometrics, homology and cladistics: review and recommendations. Cladistics. 35:230–242.

Paradis E., Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics (Oxf.). 35:526–528.

Parins-Fukuchi C. 2018a. Bayesian placement of fossils on phylogenies using quantitative morphometric data. Evolution. 72:1801–1814.

Parins-Fukuchi C. 2018b. Use of continuous traits can improve morphological phylogenetics. Syst. Biol. 67:328–339.

Pavlicev M., Cheverud J.M., Wagner G.P. 2009. Measuring morphological integration using eigenvalue variance. Evol. Biol. 36:157–170.

Pečnerová P., Moravec J.C., Martínková N. 2015. A skull might lie: modeling ancestral ranges and diet from genes and shape of tree squirrels. Syst. Biol. 64:1074–1088.

Perrard A., Lopez-Osorio F., Carpenter J.M. 2016. Phylogeny, landmark analysis and the use of wing venation to study the evolution of social wasps (Hymenoptera: Vespidae: Vespinae). Cladistics. 32:406–425.

Philippe H., Zhou Y., Brinkmann H., Rodrigue N., Delsuc F. 2005. Heterotachy and long-branch attraction in phylogenetics. BMC Evol. Biol. 5:50.

Piras P., Colangelo P., Adams D.C., Buscalioni A., Cubo J., Kotsakis T., Meloro C., Raia P. 2010. The *Gavialis–Tomistoma* debate: the contribution of skull ontogenetic allometry and growth trajectories to the study of crocodylian relationships. Evol. Dev. 12:568–579.

Polly P.D. 2001. On morphological clocks and paleophylogeography: towards a timescale for *Sorex* hybrid zones. Genetica. 112–113:339–357.

Polly P.D. 2003a. Palaeophylogeography: the tempo and mode of geographic differentiation in marmots (*Marmota*). J. Mammal. 84:369–384.

Polly P.D. 2003b. Paleophylogeography of *Sorex araneus* (Insectivora, Soricidae): molar shape as a morphological marker for fossil shrews. Mammalia. 68:233–243.

Polly P.D. 2004. On the simulation of the evolution of morphological shape: multivariate shape under selection and drift. Palaeontol. Electron. 7:7A.

Prömel H.J., Steger A. 2002. The Steiner tree problem: a tour through graphs, algorithms, and complexity. Braunschweig: Vieweg.

R Core Team. 2013. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Rannala B., Yang Z. 2008. Phylogenetic inference using whole genomes. Annu. Rev. Genomics Hum. Genet. 9:217–231.

Revell L.J., Harmon L.J., Collar D.C. 2008. Phylogenetic signal, evolutionary process, and rate. Syst. Biol. 57:591–601.

Rifkin S.A., Kim J., White K.P. 2003. Evolution of gene expression in the *Drosophila melanogaster* subgroup. Nat. Genet. 33:138–144.

Robinson D.F., Foulds L.R. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131–147.

Rohlf F.J. 1998. On applications of geometric morphometrics to studies of ontogeny and phylogeny. Syst. Biol. 47:147–158.

Saitou N., Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4:406–425.

Santana S.E., Lofgren S.E. 2013. Does nasal echolocation influence the modularity of the mammal skull? J. Evol. Biol. 26:2520–2526.

Schluter D., Price T., Mooers A.Ø., Ludwig D. 1997. Likelihood of ancestor states in adaptive radiation. Evolution. 51:1699–1711.

Schroeder L., Scott J.E., Garvin H.M., Laird M.F., Dembo M., Radovèiæ D., Berger L.R., de Ruiter D.J., Ackermann R.R. 2017. Skull diversity in the *Homo* lineage and the relative position of *Homo naledi*. J. Hum. Evol. 104:124–135.

Sherratt E., Gower D.J., Klingenberg C.P., Wilkinson M. 2014. Evolution of cranial shape in caecilians (Amphibia: Gymnophiona). Evol. Biol. 41:528–545.

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

Smith M.R. 2019a. Bayesian and parsimony approaches reconstruct informative trees from simulated morphological datasets. Biol. Lett. 15:20180632.

Smith M.R. 2019b. Quartet: comparison of phylogenetic trees using quartet and bipartition measures. doi:10.5281/zenodo.2536318.

Smith U.E., Hendricks J.R. 2013. Geometric morphometrics character suites as phylogenetic data: extracting phylogenetic signal from gastropod shells. Syst. Biol. 62:366–385.

Smith W.D. 1992. How to find Steiner minimal trees in Euclidean *d*-space. Algorithmica. 7:137–177.

Sneath P.H.A., Sokal R.R. 1973. Numerical taxonomy: the principles and practice of numerical classification. San Francisco: W. H. Freeman.

Stayton C.T. 2008. Is convergence surprising? An examination of the frequency of convergence in simulated datasets. J. Theor. Biol. 252:1–14.

Steel M.A., Penny D. 1993. Distributions of tree comparison metrics—some new results. Syst. Biol. 42:126–141.

Swiderski D.L., Zelditch M.L., Fink W.L. 1998. Why morphometrics is not special: coding quantitative data for phylogenetic analysis. Syst. Biol. 47:508–519.

Swofford D.L., Maddison W.P. 1987. Reconstructing ancestral character states under Wagner parsimony. Math. Biosci. 87:199–229.

Swofford D.L., Olsen G.J., Waddell P.J., Hillis D.M. 1996. Phylogenetic inference. In: Hillis D.M., Moritz C., Mable B.K. editors. Molecular systematics. Sunderland (MA): Sinauer. p. 407–514.

Thompson E.A. 1973. The method of minimum evolution. Ann. Hum. Genet. 36:333–340.

Uddin M., Wildman D.E., Liu G., Xu W., Johnson R.M., Hof P.R., Kapatos G., Grossman L.I., Goodman M. 2004. Sister grouping of chimpanzees and humans as revealed by genome-wide phylogenetic analysis of brain gene expression profiles. Proc. Natl. Acad. Sci. USA 101:2957–2962.

Wägele J.W., Mayer C. 2007. Visualizing differences in phylogenetic information content of alignments and distinction of three classes of long-branch effects. BMC Evol. Biol. 7:147.

Wagner G.P. 1984. On the eigenvalue distribution of genetic and phenotypic dispersion matrices: evidence for a nonrandom organization of quantitative character variation. J. Math. Biol. 21:77–95.

Wagner G.P., Pavlicev M., Cheverud J.M. 2007. The road to modularity. Nat. Rev. Genet. 8:921–931.

Watanabe A., Slice D.E. 2014. The unitility of cranial ontogeny for phylogenetic inference: a case study in crocodylians using geometric morphometrics. J. Evol. Biol. 27:1078–1092.

Wiens J.J., Hollingsworth B.D. 2000. War of the iguanas: conflicting molecular and moprhological phylogenies and long-branch attraction in iguanid lizards. Syst. Biol. 49:143–159.

Zelditch M.L., Fink W.L., Swiderski D.L. 1995. Morphometrics, homology, and phylognetics: quantified characters as synapomorphies. Syst. Biol. 44:179–189.

Zelditch M.L., Fink W.L., Swiderski D.L., Lundrigan B.L. 1998. On applications of geometric morphometrics to studies of ontogeny and phylogeny: a reply to Rohlf. Syst. Biol. 47:159–167.

Zelditch M.L., Swiderski D.L., Sheets H.D. 2012. Geometric morphometrics for biologists: a primer. 2nd ed. Amsterdam: Elsevier.

Zelditch M.L., Ye J., Mitchell J.S., Swiderski D.L. 2017. Rare ecomorphological convergence on a complex adaptive landscape: body size and diet mediate evolution of jaw shape in squirrels (Sciuridae). Evolution. 71:633–649.