Syst. Biol. 69(5):897–912, 2020

DOI:10.1093/sysbio/syaa012
Advance Access publication February 19, 2020

Morphological Phylogenetics Evaluated Using Novel Evolutionary Simulations

JOSEPH N. KEATING[1,2], ROBERT S. SANSOM[1]*, MARK D. SUTTON[3], CHRISTOPHER G. KNIGHT[1] AND RUSSELL J. GARWOOD[1,4,*]

[1]*Department of Earth and Environmental Sciences, University of Manchester, Williamson Building, Oxford Road, Manchester M13 9PL, UK;* [2]*School of Earth Sciences, University of Bristol, Life Sciences Building, Tyndall Avenue, Bristol BS8 1TQ, UK;* [3]*Department of Earth Science and Engineering, South Kensington Campus, Imperial College London, London SW7 2AZ, UK; and* [4]*Earth Sciences Department, Natural History Museum, Cromwell Rd, South Kensington, London SW7 5BD, UK*

**Correspondence to be sent to: Department of Earth and Environmental Sciences, University of Manchester, Manchester M13 9PL, UK; E-mail: russell.garwood@manchester.ac.uk; robert.sansom@manchester.ac.uk.*

Received 15 July 2019; reviews returned 31 January 2020; accepted 7 February 2020
Associate Editor: Jeanne Serb

Abstract.—Evolutionary inferences require reliable phylogenies. Morphological data have traditionally been analyzed using maximum parsimony, but recent simulation studies have suggested that Bayesian analyses yield more accurate trees. This debate is ongoing, in part, because of ambiguity over modes of morphological evolution and a lack of appropriate models. Here, we investigate phylogenetic methods using two novel simulation models—one in which morphological characters evolve stochastically along lineages and another in which individuals undergo selection. Both models generate character data and lineage splitting simultaneously: the resulting trees are an emergent property, rather than a fixed parameter. Standard consensus methods for Bayesian searches (Mki) yield fewer incorrect nodes and quartets than the standard consensus trees recovered using equal weighting and implied weighting parsimony searches. Distances between the pool of derived trees (most parsimonious or posterior distribution) and the true trees—measured using Robinson-Foulds (RF), subtree prune and regraft (SPR), and tree bisection reconnection (TBR) metrics—demonstrate that this is related to the search strategy and consensus method of each technique. The amount and structure of homoplasy in character data differ between models. Morphological coherence, which has previously not been considered in this context, proves to be a more important factor for phylogenetic accuracy than homoplasy. Selection-based models exhibit relatively lower homoplasy, lower morphological coherence, and higher inaccuracy in inferred trees. Selection is a dominant driver of morphological evolution, but we demonstrate that it has a confounding effect on numerous character properties which are fundamental to phylogenetic inference. We suggest that the current debate should move beyond considerations of parsimony versus Bayesian, toward identifying modes of morphological evolution and using these to build models for probabilistic search methods. [Bayesian; evolution; morphology; parsimony; phylogenetics; selection; simulation.]

Phylogenetic trees provide a vital framework for evolutionary inferences. Consequently, the accuracy of phylogenetic estimates built using empirical characters underpins our understanding of evolutionary history. Morphology was fundamental to the conception and development of phylogenetic methods (Hennig 1950, 1965). However, in the genomic age, sequence data have replaced morphology as both the dominant source of phylogenetic information for estimating tree topology and the basis of numerical phylogenetic method development (Lee and Palci 2015; Lartillot et al. 2016). Molecular characters are more numerous than morphological ones and their evolution can be modeled based on empirical observations (Kimura 1980; Felsenstein 1981; Hasegawa et al. 1985). Nevertheless, morphology still plays a fundamental role. It is the only form of data by which we can incorporate fossils, and thus a deep-time perspective, in phylogenies. Fossil taxa allow the calibration of molecular clocks (Donoghue and Yang 2016); offer an independent test of evolutionary developmental hypotheses (Raff 2007); break long branches, and thus clarify otherwise intractable relationships (Donoghue et al. 1989; Wiens and Soltis 2005; Legg et al. 2013); and provide the only means of understanding diversity and evolution in deep time (Raup and Sepkoski 1982). For these reasons, integrating fossils in phylogenies is necessary in order to derive accurate phylogenies and reconstructions of character and clade evolution.

Morphological data have conventionally been analyzed using maximum parsimony (Kitching et al. 1998) in which trees that necessitate the fewest character changes are considered optimal. Characters are either treated as weighted equally or rescaled in relation to their homoplasy, for example, using implied weighting (IW; Goloboff 2013). Likelihood-based models have also been used to analyze morphological data, primarily through Bayesian analysis using the Mk model of character evolution (Lewis 2001). The Mk model is a k-parameter model, where k is the number of possible unordered states for a discrete morphological character (e.g., in an M2 model, characters could have $k = 2$ states). The model assumes that character state changes follow a Markov process, and thus the likelihood of changing from one state to another is determined only by the current state. The basic Mk model assumes that all state changes are equally likely and occur at the same rate, although these assumptions are not always true (Lewis 2001). Some characters might be gained or lost much faster or slower than others; as such numerous refinements have been proposed to account for asymmetrical evolutionary rates. For instance, the symmetrical (SYM) and all-rates-different (ARD) models (Paradis et al. 2004) are two extensions of the Mk model that can relax this assumption.

Recently, a suite of simulation studies have assessed the relative performance of Bayesian and parsimony phylogenetic inference built on categorical data (Wright and Hillis 2014; O'Reilly et al. 2016; Congreve and Lamsdell 2016; Puttick et al. 2017b, 2019; O'Reilly et al. 2018a; Smith 2019a). These simulation studies have followed one of two general approaches: they have utilized either a random model of character evolution (Goloboff et al. 2018; Puttick et al. 2019) or likelihood-based Markov models of character evolution (Wright and Hillis 2014; O'Reilly et al. 2016; Brown et al. 2017; Puttick et al. 2017b). Both approaches simulate stochastic state changes upon a fixed tree, but they differ in their genesis of characters. Markov models of character evolution use branch lengths; for each character, the probability of a state change occurring on a given branch is proportional to the length of that branch. In contrast, simulations employing random character evolution do not require branch lengths and consequently, there is no underlying probability of state change per branch shared between characters (Puttick et al. 2019). Comparison of results between previous simulation studies is made more challenging by the variety of metrics that have been used to estimate accuracy (e.g., by measuring tree distances using different consensus methods). The results, and subsequent recommendation, vary between studies (see Table 1).

Here, we present two custom-built evolutionary models that simulate lineage splitting and character evolution simultaneously. In contrast to previous studies, the tree is, therefore, an emergent property of the simulation, rather than a predefined parameter. The models vary with respect to both the level at which evolution occurs (one is lineage-based and one operates at the level of the individual), and the underlying mode of evolution (stochastic in one model, and via selection in the other). We characterize these models using a wide variety of tree and data metrics, then use data simulated by each to evaluate the relative performance of topology estimation with different phylogenetic reconstruction techniques (parsimony methods and Bayesian implementation of the Mk model). Previous studies have quantified accuracy using Robinson-Foulds (RF) distance between the simulated tree and a single form of consensus tree. In this study, we employ measures of phylogenetic distance to the consensus type typically employed for each inference method. Distance metrics include not just RF, but also subtree prune and regraft (SPR), tree bisection reconnection (TBR), and quartets distances. We also consider the larger pool of trees from which the consensus trees are derived. Finally, we assess the relationship between different tree/data attributes, and the accuracy of phylogenetic estimation. By doing so, we (i) assess the accuracy of phylogenetic reconstruction techniques using evolutionary simulations; (ii) broaden the range of models available to simulate discrete morphological character data; (iii) assess the performance of phylogenetics methods in light of different modes of discrete character evolution and data set properties; and (iv) characterize the relationship between modes of morphological evolution and resulting data properties.

Materials and Methods

Phylogenetic data are derived from two evolutionary simulations designed for this study. One is a stochastic, lineage-based model (MBL2017); the other includes selection and operates at the level of the individual (TREvoSim). The models thus encompass significant diversity in evolutionary simulations: the presence or absence of natural selection (Huneman 2014) and the level (individuals vs. taxa) at which evolution is simulated to occur. Both generate character data concurrently with trees in which branch lengths represent time. In both packages, apomorphies are accrued within lineages via random mutation, but in MBL2017 both these, and speciation, are stochastic: it is a neutral birth–death model. In contrast, TREvoSim simulates natural selection: mutations which increase fitness are selected for and drive evolution. We created three data sets of 1000 replicates (128, 512, and 1024 parsimony-informative characters) using each software package. All trees comprise 32 terminals. All analysis scripts, software code, exemplar outputs for both models, and redistributables are available in the Supplementary material, available on Dryad at https://doi.org/10.5061/dryad.4b8gtht8h, hosted in the SI Zenodo repository associated with this article (DOI:10.5281/zenodo.3609738). TREvoSim and MBL2017 are also available—both code and distributable binaries—in GitHub repositories (https://github.com/palaeoware), which will be updated with future versions. The versions employed herein are TREvoSim v1.0.0 (https://github.com/palaeoware/trevosim/; doi:10.5281/zenodo.3619356) and MBL2017 v2.0.0 (https://github.com/palaeoware/MBL2017; doi:10.5281/zenodo.3614075). All code is published under a GNU General Public License v3.0. Supplementary Figures, available on Dryad.

Stochastic Data Simulation—MBL2017

The MBL model has a rich history, having been developed and first applied in the 1970s (Raup et al. 1973; Raup and Gould 1974; Gould et al. 1977), and used in a number of studies since (Sepkoski 1978; Uhen 1996; Foote 1999; Sigwart et al. 2018). MBL provides a stochastic null-hypothesis for evolutionary modeling within paleobiology (Raup et al. 1973), and has resulted in further discussion regarding scale and the impact of chance factors in macroevolution (e.g. Stanley et al. 1981). Overviews and history are provided by Huss (2009) and Sepkoski (2012).

To generate binary morphological character data using species-level lineages, we modified the MBL2017 program described by (Sigwart et al., 2018). MBL2017 generates birth-death trees: the simulation starts with a

TABLE 1. An overview of recent studies that have assessed the relative performance of different phylogenetic inference methods using simulated data

Study	Data simulation	Data type	Phylogenetic methods	Trees compared	Recommended method
Wright et al. (2014)	Mk	Binary	Bayesian Mk EW	Bayesian Mk (50% MRC) EW (mean distance from all MPTs)	Bayesian Mk
(Congreve and Lamsdell, 2016)	Mk	Binary	EW IW (k = 1, 3, 5, 10)	EW (SC) IW (SC)	EW
O'Reilly et al. (2016)	HKY(Characters fit a predefined homoplasy distribution)	Binary	Bayesian Mk EW IW (k = 2, 3, 5, 20, 200)	Bayesian Mk (50% MRC) EW (50% MRC) IW (50% MRC)	Bayesian Mk
Brown et al. (2017)	HKY (Characters fit a predefined homoplasy distribution)	Binary	Bayesian Mk ML Mk	Bayesian Mk (50% MRC) ML Mk (MLT, 50% bootstrap NST)	Bayesian Mk/ML bootstrap
Puttick et al. (2017a)	HKY (Characters fit a predefined homoplasy distribution)	Binary + multistate	Bayesian Mk EW IW (k = 2)	Bayesian Mk (50% MRC) EW (50% MRC) IW (50% MRC)	Bayesian Mk
Goloboff et al. (2018)	Random (Characters fit a predefined homoplasy distribution)	multistate	Bayesian Mk EW IW (k = 2:200) ML Mk	Bayesian Mk (50% MRC) EW (SC) IW (SC)	IW
O'Reilly et al. (2018a)	HKY (Characters fit a predefined homoplasy distribution)	Binary + multistate	Bayesian Mk EW ML Mk	Bayesian Mk (50% MRC) EW (50% MRC, 50% bootstrap SS) ML Mk (50% MRC, 50% bootstrap NST)	Bayesian Mk/ML bootstrap
Puttick et al. (2019)	Random (Characters fit a predefined homoplasy distribution)	Binary, Binary + multistate	Bayesian Mk EW IW (k = 2, 10, 20)	Bayesian Mk (50, 95% MRC) EW (50, 95% bootstrap SS)	Bayesian Mk
Smith (2019a)	Mk, HKY (Characters fit a predefined homoplasy distribution)	Binary	Bayesian Mk EW IW (k = 1, 2, 3, 5, 10, 20, 200)	Bayesian (50, 55, 60... 95 MRC) IW & EW (0, 2, 4... 100 bootstrap / jackknife NST, -100, -95...95, 100 Bremer NST)	Bayesian Mk/IW bootstrap

Key: Mk. Lewis Model 2001; HKY. Hasegawa et al. 1985 model; EW. equal weighting parsimony; IW. implied weighting parsimony; ML. Maximum likelihood; MLT. Maximum likelihood tree; MRC. majority-rule consensus; SC. strict consensus; k. concavity constant; NST. node support tree.

single lineage, which iterates through a fixed number of discrete time intervals. At each time interval, a lineage has a fixed chance to speciate (terminate and be replaced by two daughter lineages), and a separate fixed chance to go extinct (terminate without replacement). Speciation and extinction in the same interval are not allowed. The code has been modified for the current article to add phylogenetic information to lineages in the form of a user-specified number of binary characters (constrained to a multiple of 32; here 128, 512, or 1024). Each character of the initial lineage has a random state (either 0 or 1). At each time interval, each binary character of each lineage has a fixed probability of mutating (i.e. flipping 0 to 1, or 1 to 0). Daughter lineages inherit the characters of their parent lineage. Mutations occur before extinction or speciation in each time interval. Apomorphies hence accrue within lineages via random mutation and represent pure "drift"; there is no selection as characters do not influence the extinction and speciation mechanism.

To simulate data for this study, we used the following parameters: speciation probability 0.055 (λ); extinction probability 0.045 (μ); mutation probability per character per time interval 0.02; and time intervals per simulation 400. These settings follow those of Sigwart et al. (2018) and employ the median speciation/extinction value pair of that study; mutation probability was assessed experimentally and a value was selected that balanced the need to avoid identical taxa, but minimized saturation. For each resulting tree, extinct taxa were removed as in Sigwart et al. (2018), data sets with one or more uninformative characters were removed, and

taxa with identical character scores recorded (mean 3.5, representing the last speciations in sister lineages). Constraint to 32 taxa was achieved by selecting a clade from each simulation run with precisely 32 terminals; trees in which no such clade existed were discarded.

Data Simulation under Selection—TREvoSim

We also generated binary morphological character data using an individual- or agent-based evolutionary simulation, called TREvoSim. This model represents nonstochastic evolution as it incorporates natural selection. It derives some concepts from the package REvoSim (Garwood et al. 2019)—for example, the fitness algorithm—but has a focus on the simulation of trees and associated character data. It does not incorporate concepts of space or sexual reproduction. TREvoSim employs digital organisms comprising binary strings, which provide characters for both phylogenetic inference and for selection within the simulation. Organisms compete, replicate, and mutate, and the simulation incorporates a species concept. Speciation is emergent in the simulation, allowing the software to output a phylogenetic tree showing the species relationships, with associated character data.

The principles of the model are as follows: a user-defined number of organisms are alive at any given point during a simulation (we use 128). These are held in a list, the *playing field* (*pf*); this population can include members of different species, as well as multiple organisms from the same species. The binary string of an organism is ultimately the character data within this study, and the number of characters present (n) is user defined (here 128, 512, or 1024 characters). The fitness of the organisms alive at any given time is calculated by comparison with the *environment* (all organisms in the playing field are competing, and thus this might represent a niche). The environment is formed of five random numbers (*masks*) of size n, where n equals the length of the character binary string of each organism. The fitness of every organism in the playing field is calculated following the approach described by Garwood et al. (2019). In brief, this employs an exclusive OR ($\oplus$) operation to sum the Hamming distance (*hd*) of the organism to each of the five masks. Where the input bits are the same the exclusive OR returns a zero, otherwise it returns a one. Thus by comparing every bit of the binary string to the equivalent bit in each mask and summing the results, this provides a value between 0 and $5n$. The fitness (f) is an integer calculated as the distance from a target value; for this release, that is defined as halfway between zero and $5n$, that is:

$$f = |\sum^{hd}[m1..m5] - \frac{5n}{2}|$$

Those organisms best-suited to their environment (the fittest) thus have a distance of zero, and the worst $2.5n$. The advantages of this approach are (i) it is relatively computationally efficient; (ii) small environment changes result in small changes to an organism's fitness; and (iii) mutations within organisms will also result in relatively small fitness changes. Multiple character strings that allow an organism to be optimally fit for any environment exist (i.e. there are numerous fitness peaks; Supplementary Fig. 1, available on Dryad). The algorithm used for each TREvoSim iteration is described in full below. Key points are: members of the playing field compete, and their fitness is linked to reproductive success; and species within TREvoSim are defined based on Hamming distance (character distance) to past organisms within an evolving lineage (user defined; species difference, *sd*). A simulation is initiated by filling the masks with random binary numbers and the *playing field* with multiple identical organisms (species zero). Initializing with a single organism is a necessary simplification to allow all organisms in the simulation to belong to the same phylogeny. The chosen organism is within the top 10% of possible fitnesses for the starting masks (thus preventing the simulation being a single lineage adapting to one fitness peak). A simulation then runs until the desired number of species is achieved by repeating the following steps (this is also summarized in Fig. 1):

1) Organisms within the *playing field* are sorted by fitness, with the fittest organisms at the top of the list. If a number of organisms have the same fitness (e.g., at initialization), these are randomly ordered.

2) An organism is picked to be duplicated via a sequential coin toss (with a 50% chance of selecting the first in the list, then if that is not chosen, a 50% chance of selecting the second, and so on). If the simulation reaches the end of the *playing field* without selecting one, it starts from the beginning again.

3) The organism selected for duplication has a user-defined chance of mutation (defined as mutations per hundred characters per iteration; 1.5 for these data sets). The user can select whether deleterious mutations are accepted (they were discarded here).

4) If the duplicated organism, after mutation, is sufficiently different (*sd*, species difference) to its character string at origination it is defined as a new species (if this is not the first speciation in the lineage, *sd* bits from the last species to diverge is used as benchmark). Comparison to last speciation (if one has occurred) rather than the original genome prevents bursts of speciation from closely related organisms sharing a common parent, but still allows cladogenesis within a species.

5) The duplicated, mutated organism is then returned to the *playing field*, overwriting the least fit organism in the playing field (or randomly selecting one of the least fit if multiple least fit organisms exist).

6) Organisms in the playing field typically represent multiple species once a simulation is running. Thus the playing field is checked each iteration, and any species that have become extinct are identified. On extinction, the characters of the last surviving organism are appended to the character matrix (this is optional, but on by default, and ensures that if a single lineage has

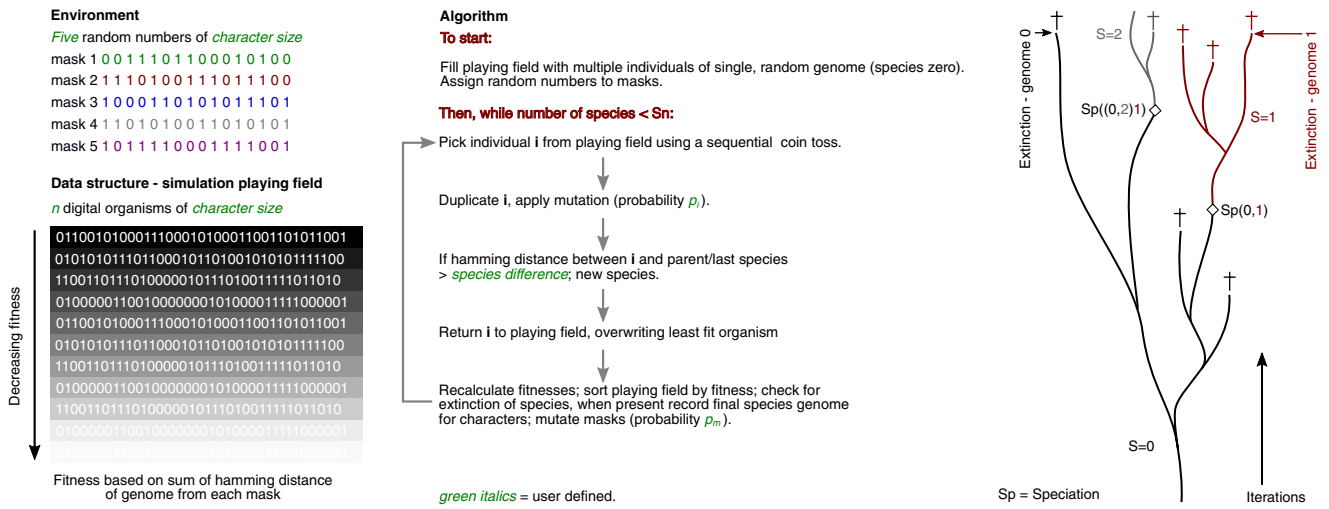

FIGURE 1. A summary flowchart showing the algorithm of the TREvoSim model/software used to simulate data under selection for this study.

given birth to multiple species, the recorded character set is closest to the most recently branching terminal).

7) Masks are then mutated at the end of each iteration (there is a user-defined chance of mutation per hundred bits per iteration, here 1.0), providing environmental change throughout the simulation.

Once the requested number of species has been achieved, the simulation finishes, the character data of all extant taxa are appended to the character matrix (the fittest organism, or one of these, is selected if multiple organisms within a species are surviving). The final character matrix contains all extinct and extant species. If stripping of uninformative characters is requested, the number of characters, and species difference, are increased at the start of a run (using an empirically calculated factor based on the requested settings), and then informative characters are randomly subsampled at this stage to achieve the requested number of characters. A check for identical taxa is then conducted (the data are discarded and simulation repeated if the number of identical terminals is above a user-defined cutoff). The tree and character matrix are output through a customized logging system, which allows, e.g. standard (nexus/TNT) formats. See Supplementary Material, available on Zenodo, for examples of the output strings used in the current study.

To simulate data for this study, the simulation ran until it reached 32 species, the cutoff above which runs were discarded was five identical terminals, and uninformative characters were stripped. For creating data sets of 128 characters, the species difference was set to 12, for 512 it was set to 50, and for 1024 it was 100.

Estimating of Properties of Simulated Data

We have characterized the properties of data generated under the two different models using eight tree and data attributes. To capture the distribution of homoplastic characters within each data set, we measured ensemble Consistency Index (CI; Kluge and Farris 1969), ensemble Retention Index (RI; Farris 1989) and the mean number of excess steps of the character data mapped onto the true tree using the R package phangorn (Schliep 2011). To estimate the phylogenetic signal of the simulated characters, we calculated the character dispersion (*D*) metric of Fritz and Purvis (2010) *D* estimates the phylogenetic signal of binary characters using sum of sister-clade differences in a given phylogeny, and as such, it is a measure of how "clumped" characters are, independent of tree size, shape, and character prevalence (given a minimum of 25 taxa). We estimated mean *D* for each data set using the phylo.d function in the R package caper (Orme et al. 2012). We recorded the Colless Index of treeshape (1982) for each simulated tree using the R package apTreeshape (Bortolussi et al. 2006). In order to further describe tree shape, we also calculated the "stemminess" metric outlined in Fiala and Sokal (1985). This is defined as the proportion of the sum of branch lengths of a clade, including the branch subtending the clade, that is accounted for by the branch subtending the clade. The stemminess of a tree we report herein is the mean stemminess value for each clade of the tree. Finally, we characterized the relationship between morphological similarity and recency of common ancestry, termed "morphological coherence" by Raup and Gould (1974). For each taxon pair within each data set, we measured the phylogenetic distance (i.e. the branch length from Taxon A to the common ancestor of Taxon A and B + the branch length from Taxon B to the common ancestor of Taxon A and B) and the character difference (i.e. the number of different characters—or hamming distance—between Taxon A and Taxon B) using the R packages phytools (Revell 2012) and phylobase (Hackathon et al. 2011). Plotting phylogenetic distance against character difference allows visualization of

morphological coherence for each simulated data set. We quantified morphological coherence using two methods. Firstly, we measured the Spearman's Rank correlation of all taxon pairs (termed "raw morphological coherence" henceforth). This metric can potentially be biased by long-branch artifacts, thus we also measured the Spearman Rank correlation for just taxon pairs where phylogenetic distance was less than or equal to half the maximum possible phylogenetic distance (termed "adjusted morphological coherence" henceforth).

Phylogenetic Analyses

For Bayesian estimation, we analyzed batch nexus files created by the simulation software in MrBayes version 3.2 (Ronquist et al. 2012) using the Mk + informative model (which accounts for only parsimony-informative characters having been scored, and ascertainment bias) with gamma-shaped rate variation (Mki + Γ). Extracts from the batch files are provided in the Supplementary Material, available on Zenodo. We chose to explicitly exclude uninformative characters from simulations to better reflect the attributes of the vast majority of empirical cladistic data sets (Brazeau 2011). We used 2 runs of 4 chains and sampled 10,000 trees each run, of which 30% were discarded as burn-in. To confirm that independent runs had reached convergence, we examined the ".pstat" output files. Convergence was accepted for data sets that had average ESS values >200 and PSRF values between 0.9 and 1.1 for all parameters. We ensured that at least 50% of each batch (>500 replicates) had achieved convergence, and 500 of those converged analyses were randomly selected for subsequent analysis. The number of generations required to achieve 50% convergence varied with respect to both the number of characters and the model used to simulate the data (see Supplementary Table S1, available on Dryad).

For parsimony, we analyzed batch files in TNT version 1.5 (Goloboff and Catalano 2016 made available with the sponsorship of the Willi Hennig Society). We used "new technology" with tree-drifting, tree-fusing, and sectorial searches (*xmult: level 10*) and subsequent branch breaking (*bbreak*) retaining a maximum of 100,000 MPTs for each matrix. We used equal weighting (EW) and IW searches. IW (Goloboff 2013) is an extension of maximum parsimony in which homoplastic characters (i.e. those with additional steps) are down-weighted according to a concavity constant, k. We used $k = 3$, which enforces strong down-weighting of homoplastic characters, is widely used, and is the default in TNT.

Tree Distance Calculations

Consensus trees.—Bayesian and parsimony searches take very different approaches to tree sampling which can confound direct comparison of precision (the number of nodes resolved) and accuracy (how many of these are correct). Parsimony analyses typically record the optimal (most parsimonious) trees, which are summarized using a strict consensus approach. In contrast, Bayesian inference generates a posterior sample of thousands of trees, sampled relative to their posterior probability. This posterior distribution is typically summarized using a 50% majority-rule consensus (MRC) tree, which contains all bipartitions recovered in greater than 50% of the posterior trees. Alternatively, the posterior distribution may be summarized using a maximum clade credibility (MCC) tree, which is a single tree within the posterior distribution containing the maximum sum or product of posterior probabilities across each clade. The MCC tree is analogous to a most parsimonious tree in that it represents an optimal point estimate. Previous studies assessing the relative performance of these methods have done so by comparing consensus trees and collapsing poorly supported nodes at different thresholds (Brown et al. 2017; O'Reilly et al. 2018a; Puttick et al. 2019; Smith 2019a). Collapsing branches under a certain threshold of support is not currently standard practice for parsimony studies (Puttick et al. 2017a). Here, we compare the most commonly used outputs for the respective methods: namely strict consensus (SC) trees for parsimony analyses, and the 50% MRC tree for Bayesian inference. We also compare the mean distance from the larger pool of binary trees from which the consensuses are drawn. Mean distances for parsimony estimation were calculated using all MPTs. For computational efficiency, mean distance for Bayesian estimation was calculated using one of the two post-burnin runs of 7000 trees, reflecting the stationary distribution of the MCMC analysis. We also used the Bayesian MCC tree.

Distance measures, accuracy and precision.—Tree distances between the simulated trees and the derived parsimony and Bayesian phylogenies were conducted on unrooted trees. We used the package phangorn (Schliep 2011) to calculate RF's distances, but in view of this metric's sensitivity to wildcard taxa (Kuhner and Yamato 2015), we also report SPR distances. SPR within phangorn employs a heuristic search, and is thus an approximation. As such, we additionally provide true TBR distances which have been calculated using the software USPR (Whidden and Matsen 2018) modified to automate batch comparisons (code is included in the SI). We also computed quartets distance using the R package quartet (Smith 2019b).

All tree distance metrics, and in particular RF distance, risk conflating precision and accuracy. Consequently, the distance between the estimated and true tree is not always proportional to the accuracy of the estimated tree. A fully unresolved estimated tree with a single node must be 100% accurate (all nodes that occur in the estimated tree also occur in the true tree). However, the estimated tree contains no bipartitions and thus will be 50% of maximum possible RF distance from the true tree. Here, we have explicitly distinguished between measures of tree distance (e.g. RF, SPR,

TBR) and measures of accuracy and inaccuracy: here absolute accuracy (i.e. the number of nodes or quartets shared between the true and estimated tree), absolute inaccuracy (i.e. the number of nodes or quartets in the estimated tree that are not present in the true tree) and percentage accuracy (i.e. the number of correct nodes or quartets as a percentage of the number of resolved nodes or quartets). We achieved these using custom R functions (see Supplementary Material, available on Zenodo). Results were plotted using the R-package "ggplot2" (Wickham 2016).

Results

Simulated Data Properties

The models differ in the proportion of homoplastic characters found in their data sets. TREvoSim produces trees and associated data matrices with higher CI and RI values (i.e. less homoplasy) than those simulated under MBL2017 (Fig. 2A,B). Similarly, TREvoSim characters have fewer excess steps on the true tree than those characters simulated under MBL2017 (Fig. 2C, Supplementary Figs. 2 and 3, available on Dryad). The excess steps of individual TREvoSim characters for each data set show a distribution resembling that derived by Goloboff et al. (2018) from mapping excess steps for empirical data sets against a most parsimonious tree (Supplementary Fig. 3, available on Dryad). In contrast, the distribution of extra steps for MBL2017 data is normally distributed with a mode of approximately 6 (Supplementary Fig. 2, available on Dryad): characters simulated in MBL2017 are highly homoplastic.

MBL2017 and TREvoSim also show very different patterns of character dispersion. Values of D are standardized from 0 (clumping consistent with Brownian motion) to 1 (no clumping, random character distribution) whilst negative values indicate extreme clumping. The mean D values observed for MBL2017 data sets are positive, between 0 and 0.6, whilst for TREvoSim the distribution is broader and more negative with mean D values of between -2.5 and -2 (Fig. 2D). This fits with patterns expected given the stochastic and selective nature of the models, respectively (see Discussion section). MBL2017 and TREvoSim produce very different tree shapes (Fig. 2E). MBL2017 trees tend to be highly symmetrical (Colless Indices of between 0 and 150) whereas TREvoSim trees have a broader distribution of tree shapes ranging from moderately symmetrical to very asymmetrical (Colless Indices of between 50 and 400). The trees from each model also differ in their stemminess (Fig. 2F). MBL2017 trees have stemminess values of between 0.2 and 0.5, indicating that the component clades tend to consist of relatively short branches, and are subtended by relatively long branches. In contrast, TREvoSim simulated trees have stemminess values of between 0 and 0.2, indicating that clades tend to consist of long branches subtended by proportionally shorter branches.

The models also differ in their morphological coherence i.e. the relationship between the phylogenetic distance between taxon pairs and the accumulated character difference between those pairs (Fig. 2G,H). MBL2017 data sets show a positive nonlinear relationship between taxon distances and taxon character differences (Supplementary Fig. 4, available on Dryad), which suggests that the rate of character state changes is initially high, but decreases with increasing phylogenetic distance between taxon pairs. The relationship is heteroscedastic, indicating variance increases with phylogenetic distance and character difference. Both these observations are compatible with saturation, i.e. multiple state changes to the same characters result in the apparent character difference being less than would be expected based on the actual phylogenetic distance. In contrast, TREvoSim data sets show no evidence of saturation; however, the relationship between taxon phylogenetic distances and taxon character differences is less constrained (Supplementary Fig. 5, available on Dryad). Importantly, TREvoSim data sets show taxon pairs with very high phylogenetic distance and very low character difference; something that is not seen in MBL2017 data sets. This is compatible with convergence or parallel evolution. Spearman's Rank correlations of all taxon pairs (raw morphological coherence, Fig. 2G) suggest that both MBL2017 and TREvoSim data sets show similar morphological coherence, with TREvoSim data showing marginally more. However, if we remove the most distantly related taxon pairs to account for long-branch artifacts such as saturation (adjusted morphological coherence, Fig. 2H), we find that MBL2017 has much stronger correlation between phylogenetic distance and character difference for the most recently diverged taxon pairs.

Phylogenetic Analyses

Bayesian estimation yields the most accurate standard consensus trees for both morphological simulations: Bayesian majority-rule consensus (50% MRC) trees are, on average, closer to the true tree in RF distance (Fig. 3). Parsimony trees have higher absolute accuracy (i.e. they contain more correct nodes and quartets, Supplementary Fig. 6, available on Dryad); however, they also have higher absolute inaccuracy (i.e. they contain more incorrect nodes/quartets; Fig. 4, Supplementary Fig. 7, available on Dryad) and consequently, they have lower percentage accuracy (Supplementary Fig. 8, available on Dryad). The EW parsimony SC tree is marginally closer to the true tree than the IW parsimony SC in terms of RF distance (Fig. 3); however, both trees are largely equivalent in terms of absolute accuracy, absolute inaccuracy, and percentage accuracy (Supplementary Figs. 6–8, available on Dryad). Bayesian MRC trees contain, on average, the fewest resolved nodes; Bayesian estimation yields the least precise standard consensus tree. IW SC trees contain the most resolved

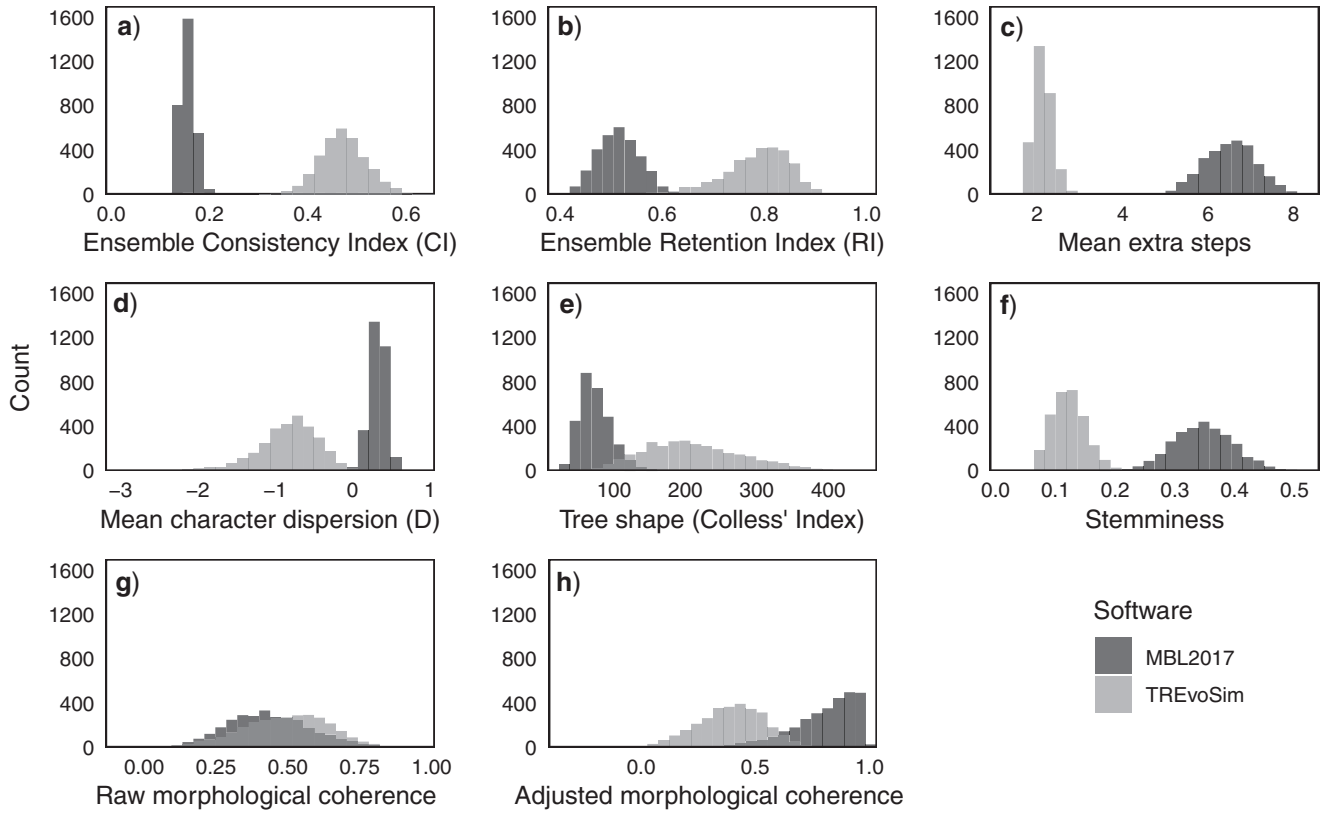

FIGURE 2. Tree and data properties for simulations under MBL2017 (stochastic) and TREvoSim (selection) models. Data comprise 1000 replicates of 128, 512, and 1024 character runs. a–c) Homoplasy measurements; (d) the level of clumping of characters across terminals; (e, f) tree shape properties; and (g, h) the relationship between branch length and character difference across all taxon pairs (see Estimating of Properties of Simulated Data section under Materials and methods section for full description).

nodes (Fig. 4; Supplementary Fig. 15, available on Dryad).

The performance of all methods varies with respect to the model under which the data were simulated, as well as the metric used to assess accuracy. All methods recover a higher percentage of correct nodes in standard consensus trees for data simulated using the stochastic (MBL2017) model (Supplementary Fig. 8, available on Dryad). Tree estimates from TREvoSim data have a low (mean <50%) percentage of correct nodes. In contrast, there is little difference in the percentage of correct quartets between models: all methods attain a high percentage (mean >75%) of correct quartets irrespective of how the data were generated. The impact of increasing character number varies with respect to the model. As characters are added in MBL2017, the performance of all methods improves. The absolute inaccuracy of Bayesian inference remains consistent with the addition of more characters; however, its precision increases (Fig. 4, Supplementary Fig. 15, available on Dryad). EW parsimony exhibits a decrease in absolute inaccuracy and an increase in precision with additional characters. IW parsimony is consistently precise, but absolute inaccuracy decreases with additional characters. Phylogenetic analyses of TREvoSim data sets, in contrast, show a slight increase in precision together with a slight increase in absolute inaccuracy with higher character numbers (Fig. 4, Supplementary Fig. 15, available on Dryad).

When using the pool of all derived trees rather than consensus trees (i.e. posterior distribution trees and all most parsimonious trees), parsimony searches and Bayesian searches are equivalent for TREvoSim data (Fig. 5, Supplementary Figs. 9 and 10, available on Dryad). All methods produce pools of trees that are distant from the true tree. However, for data created with our stochastic model (MBL2017), we find that posterior trees are, on average (when considered separately for all trees), further from the true tree than most parsimonious trees. This reflects the fact that the posterior distribution includes trees that are suboptimal under a likelihood framework, whereas most parsimonious trees are, by definition, optimal under a parsimony framework. A fairer comparison can be made using the MCC tree drawn from the posterior distribution of each search. If we compare the MCC tree, rather than all posterior trees, we find there is little difference between Bayesian and Parsimony estimation (Fig. 5, Supplementary Figs. 9 and 10, available on Dryad).

Data Attributes and Phylogenetic Accuracy

We also compare attributes of the data (tree shape, homoplasy, etc.) with the RF distances between

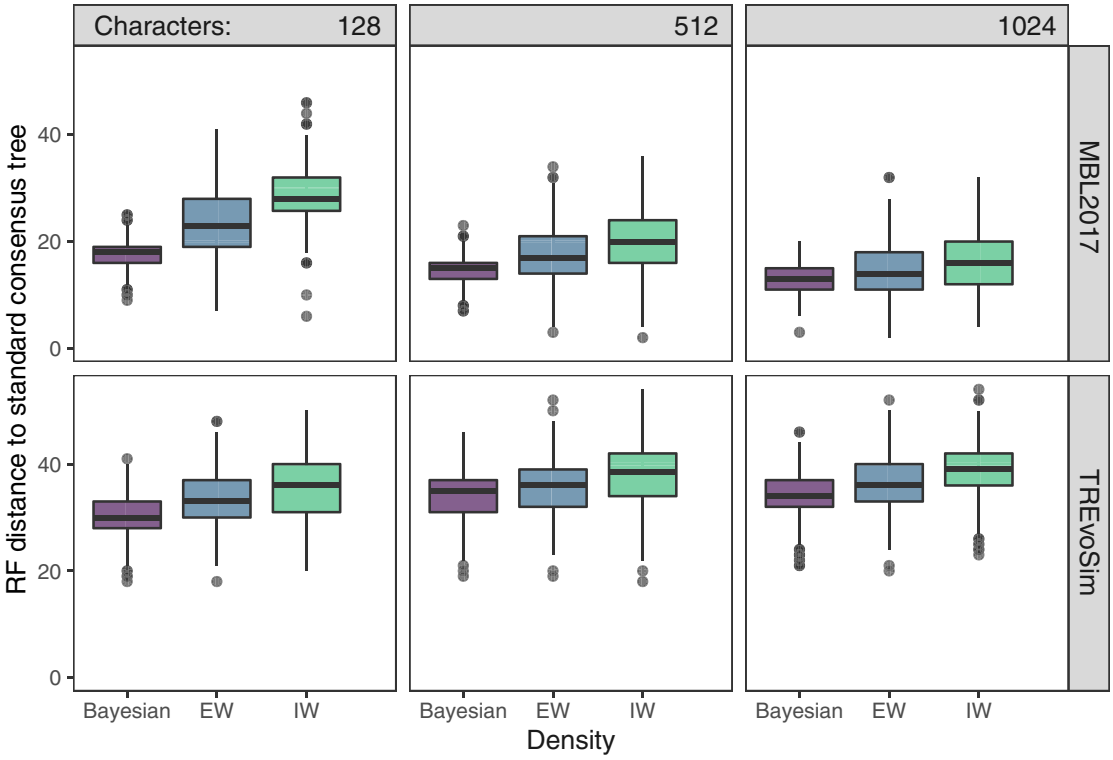

FIGURE 3. Box plot of RF distances between the simulated tree and the standard consensus trees for Bayesian inference (Bayesian), EW parsimony, and IW parsimony (500 replicates; 128, 512, and 1024 characters). Bayesian MRC trees are, on average, closer to the true tree than SC trees of EW or IW parsimony.

derived and true trees (Figs. 6 and 7, Supplementary Figs. 11–14, available on Dryad). Homoplasy measures (ensemble CI/RI, mean excess steps) are correlated with phylogenetic accuracy: within each model, more homoplastic characters yield less accurate trees. Between models, the opposite pattern is apparent (more homoplastic characters yields more accurate trees). This results from differences between the models (see *Characterizing Models of Morphological Evolution* section); the data exhibit Simpson's paradox (Simpson 1951). Focusing on within model patterns, tree shape is unrelated to phylogenetic accuracy. This contrasts with some previous studies which have found that asymmetric tree tends to be recovered with less accuracy (Rohlf et al. 1990; Puttick et al. 2017b). Mean character dispersion (D) is positively correlated with RF distance for both models. Thus, under both stochastic and selection-based data, as characters are more randomly distributed relative to the tree tips (D of 1), phylogenetic inference becomes more inaccurate.

The degree of stemminess has no correlation with the accuracy of equal or IW parsimony estimation (Supplementary Figs. 11–14, available on Dryad). However, Bayesian estimation of MBL2017 data shows a weak correlation, suggesting that the stemminess of the true tree is related to the distance of the Bayesian estimate from the true tree. Morphological coherence shows different patterns for MBL2017 and TREvoSim. For MBL2017 data there is a negative relationship between both raw and adjusted morphological coherence and RF distance, whereas for TREvoSim data there is no correlation.

DISCUSSION

Bayesian Versus Parsimony

Bayesian Inference is more accurate than parsimony methods at estimating phylogeny from discrete morphological data when using standard consensus methods. Our results suggest that this is true irrespective of the data attributes or means by which they were generated. Bayesian MRC trees are on average, closer to the true tree (RF distance, Fig. 3), contain fewer incorrect nodes/quartets (Supplementary Fig. 7, available on Dryad), and have a higher percentage accuracy (Supplementary Fig. 8, available on Dryad) than SC trees of both equal weight and implied weight parsimony searches. We note, however, that Bayesian MRC trees do not contain more correct nodes/quartets than parsimony SC trees (Supplementary Fig. 6, available on Dryad). For data generated under the selection (TREvoSim) model, parsimony and Bayesian Inference resolve a similar number of correct nodes/quartets, whereas for data generated under the stochastic (MBL2017) model, parsimony SC trees contain more correct nodes/quartets than Bayesian MRC trees.

If we consider the pool of trees from which standard consensus trees are drawn, Bayesian posterior trees tend

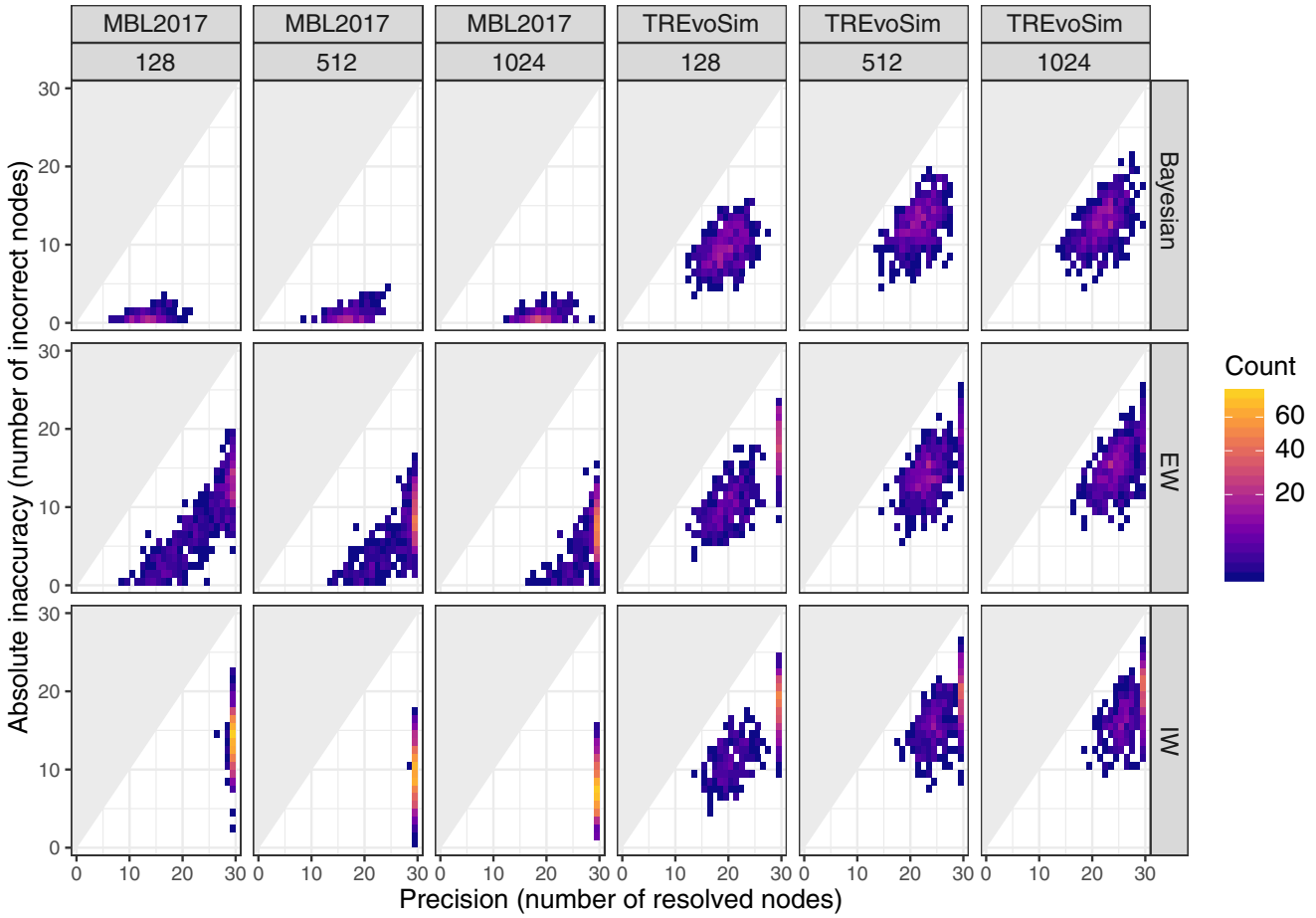

FIGURE 4. Heatmap of the number of resolved nodes against the number of incorrect nodes for each standard consensus tree for 128, 512 and 1024 character datasets (500 replicates). Bayesian estimates are, on average, less inaccurate and less precise than EW or IW parsimony estimates. All methods provide less inaccurate estimates for data generated using the stochastic model (MBL2017).

to be further from the true tree than most parsimonious trees (Fig. 5, Supplementary Figs. 9 and 10, available on Dryad). Despite this, Bayesian inference outperforms parsimony because incorrect nodes in posterior trees tend to have low posterior probability, and are typically collapsed in the Bayesian MRC tree, whereas these nodes are frequently retained within the parsimony SC tree. As a result, Bayesian MRC trees have higher percentage accuracy, but lower resolution than SC parsimony trees. We concur with O'Reilly et al. (2016) that this is preferable to precision without accuracy, present in parsimony searches on these data (most notably with IW).

As such, it is important to consider resolution in addition to percentage accuracy. In cases where percentage accuracy of two methods is equivalent, we should favor methods that also provide resolution, because these methods provide more phylogenetic information. Smith (2019a) demonstrates that if poorly supported nodes are collapsed using bootstrapping, IW parsimony estimates are comparable with Bayesian estimates in both accuracy and precision. We suggest it is thus premature to reject all parsimony estimates. Our results, and those of Smith (2019a), suggest the SC method provides a suboptimal summary of parsimony searches.

Characterizing Models of Morphological Evolution

The character dispersion (D) of data generated under each model aligns with the expectations given their differing modes of evolution. TREvoSim characters are generally extremely clumped (negative average D), concording with modes of natural selection and constraint, whilst MBL2017 characters exhibit less clumped distributions, consistent with stochastic evolution (Brownian motion and drift) through to more random character distributions that potentially indicate saturation.

Our results demonstrate that parsimony and Bayesian phylogenetic analyses estimate a higher percentage of correct nodes when applied to data generated under a stochastic evolutionary model (MBL2017) than under a selection-based one (TREvoSim). This is unexpected given the relatively elevated levels of homoplasy in data generated in our stochastic model (Fig. 2A–C).

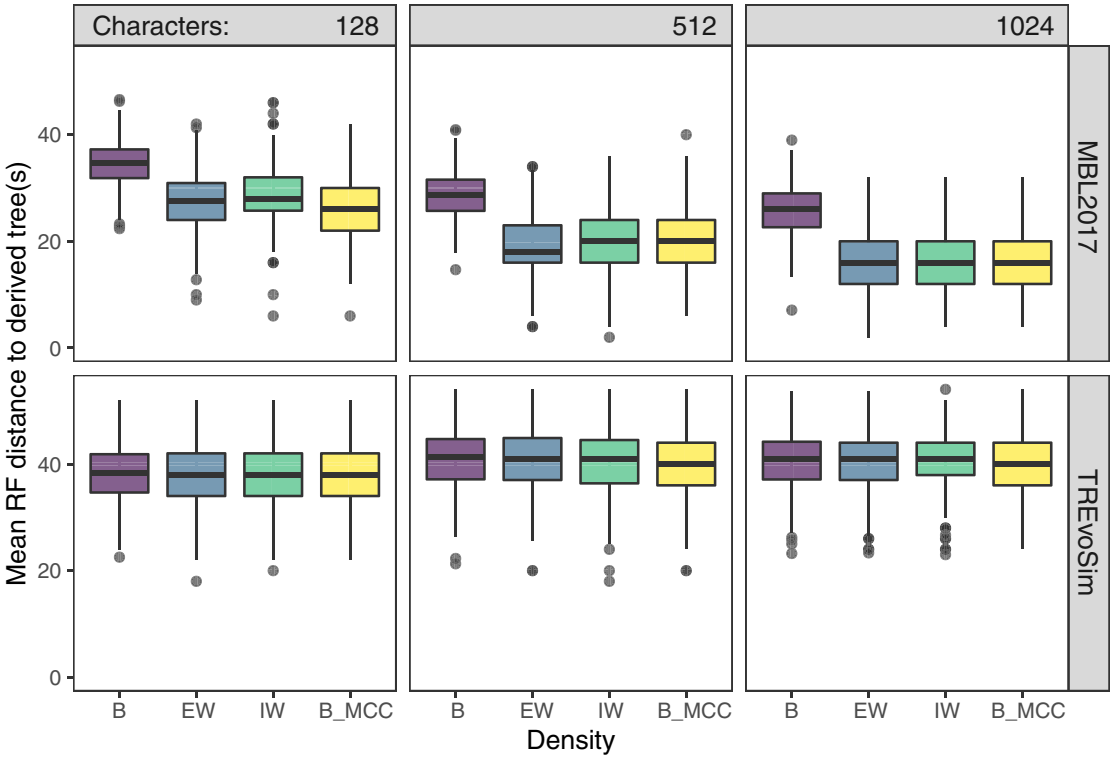

FIGURE 5. Box plot of mean RF distance between the true tree generated using TREvoSim and MBL2017 simulated data, and every tree derived for each inference method: Bayesian stationary distribution (B); Bayesian maximum clade credibility tree (B_MCC); EW most parsimonious trees; and IW most parsimonious trees (500 replicates for 128, 512, and 1024 characters).

This result is dependent on distance metric: both stochastic- and selection-generated data are estimated with a high percentage of correct quartets. This probably reflects the sensitivity of the correct nodes metric to wildcard taxa. For data from our stochastic-based model, quartet and node differences are broadly equivalent, indicating that wildcard taxa are either absent or rare, thus having limited impact. In the selection model, a smaller percentage of correct nodes than quartets likely results from wildcard taxa, which will impact a higher proportion of bipartitions than quartets. For example, a single taxon being recovered outside the correct clade will have a large impact on the number of correct nodes, but will compromise fewer quartets. Hence, selection results in more wildcards, and these taxa are routinely reconstructed incorrectly by both Bayesian and parsimony estimation. The cause of this is likely to be convergent or parallel evolution within the model: we explore this further when considering the impact of character number.

The effect of adding characters also varies between models. When estimating trees from data generated stochastically, increasing the number of characters improves performance of all phylogenetic methods via decreasing absolute inaccuracy and/or increasing precision. In contrast, additional characters generated under selection slightly increase both precision and absolute inaccuracy (Fig. 4, Supplementary Fig. 15, available on Dryad). This probably relates to differing levels of morphological coherence in characters generated under stochastic and selection models. Under a stochastic evolutionary model with no among-lineage rate heterogeneity, state changes are random. The graphs of between-taxon branch length distance and between-taxon character difference for MBL2017 (morphological coherence, Supplementary Fig. 4, available on Dryad) show that lineages rapidly accumulate character differences following initial divergence as each lineage accrues independent random mutations. As diverging lineages become more distant, the rate at which character differences accumulate slows, and eventually stops, as character state changes become increasingly homoplasious (i.e. saturation occurs). Consequently, the most recently diverged lineages will always show strong morphological coherence, even if the rate of character change is high (Fig. 2H). By increasing the character number, but maintaining relative character rates, the point at which two lineages stop accumulating character differences due to saturation is delayed. Additional characters thus improve morphological coherence.

In contrast, under selection (e.g. TREvoSim), state changes may be highly nonrandom. Homoplastic state changes can be concerted amongst particular lineages due to convergence or parallel evolution. As a result, distantly related taxa can show few character differences (Supplementary Fig. 5, available on Dryad), violating the assumption of morphological coherence, an important precept of phylogenetic analysis. Under such

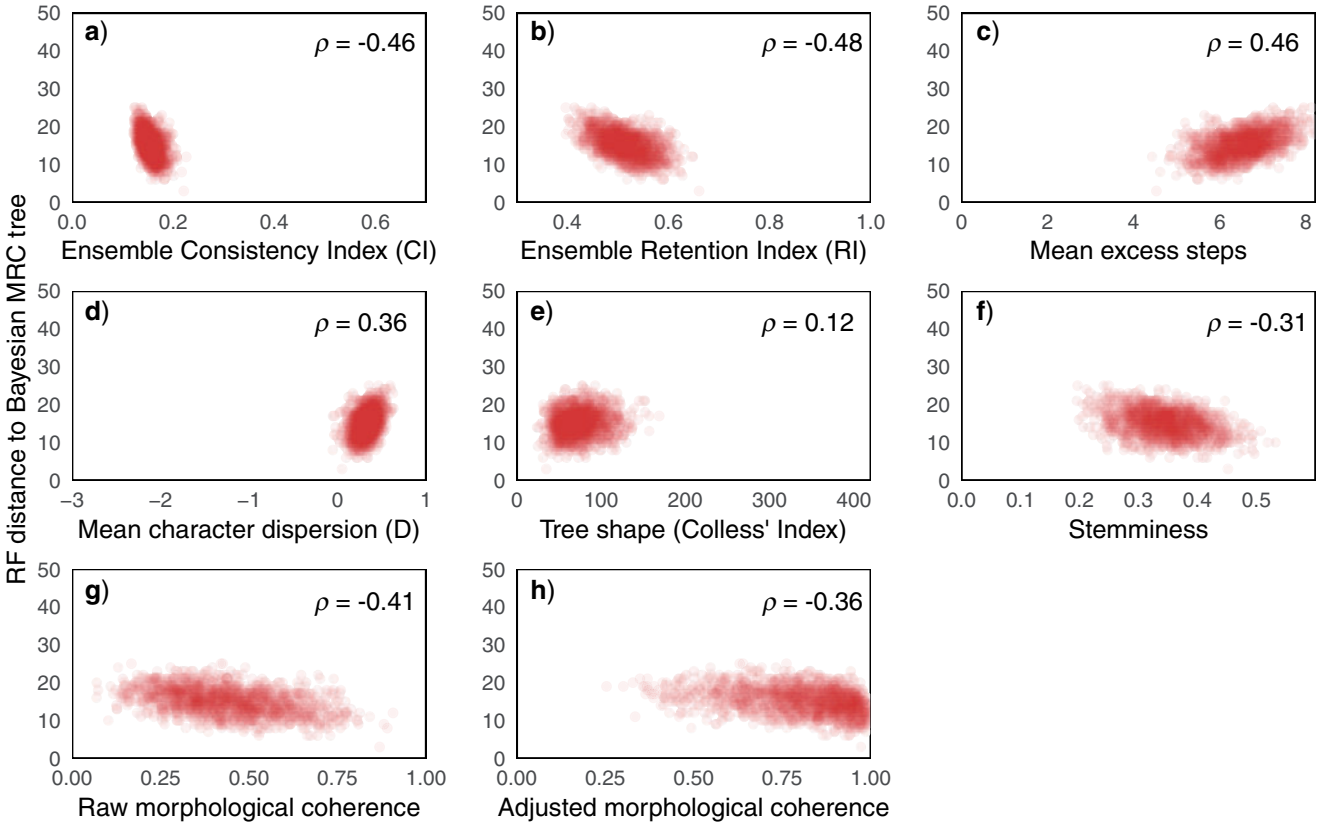

FIGURE 6. Scatter plots of correlation of eight measured tree and data attributes against RF distance between the true tree and the Bayesian MRC tree for MBL2017 datasets of 128, 512, and 1024 simulated characters (500 replicates). ρ = Spearman's Rank correlation.

circumstances, these taxa will likely behave as wildcards in phylogenetic estimation under Bayesian or parsimony. Additional characters will not alleviate this problem: they have evolved under the same selection pressures and thus contain the same concerted false signals. A similar bias will affect models of stochastic evolution with among-lineage rate heterogeneity. Here, lineage specific rates may result in concerted homoplasy on particular branches. This bias will affect all characters equally, and additional characters will not necessarily improve phylogenetic estimation.

Realism of Simulation Models

Previous studies that make inferences about the performance of different methods of morphological phylogenetic inference have all advocated the need for morphological evolutionary simulations to have empirical realism (O'Reilly et al. 2016, 2018b; Puttick et al. 2017b; Goloboff et al. 2018). These studies have used the minimum amount of homoplasy present in empirical parsimony-based estimates of tree topology as a benchmark for realism. As such, these measures of homoplasy are inherently linked to the derived tree topology. To what extent the minimum amount of homoplasy expected within empirical data relates to the real amount of homoplasy is open to debate. Furthermore, the focus for such considerations has been on the amount of homoplasy present, but it is clear from our results that this has much less impact on the accuracy of phylogenetic methods than might be expected. Rather, we find that the distribution of homoplasy among lineages has a significant impact on the accuracy of phylogenetic estimation.

Two important evolutionary mechanisms can introduce nonrandom homoplasy among lineages: natural selection and among-lineage rate heterogeneity. Consequently, to simulate "realistic" morphological data, we must consider the extent to which these mechanisms drive morphological evolution in the real world. Empirical studies have been equivocal on the relative importance of selection within morphological evolution (Lande 1976; Lynch 1990; Ho et al. 2017), and it seems likely that the extent to which this process drives evolution varies at differing taxonomic levels and timescales. On microevolutionary scales, stochastic evolution of morphology via genetic drift may be expected (Ackermann and Cheverud 2004; Marroig and Cheverud 2004). On a deeper scale, including cladogenesis from the origin and diversification of phyla through to genera, we expect selection to dominate (Rieseberg et al. 2002; Ho et al. 2017).

Among-lineage rate heterogeneity has been detected in morphological data for a wide variety of different clades over different evolutionary scales. It seems likely

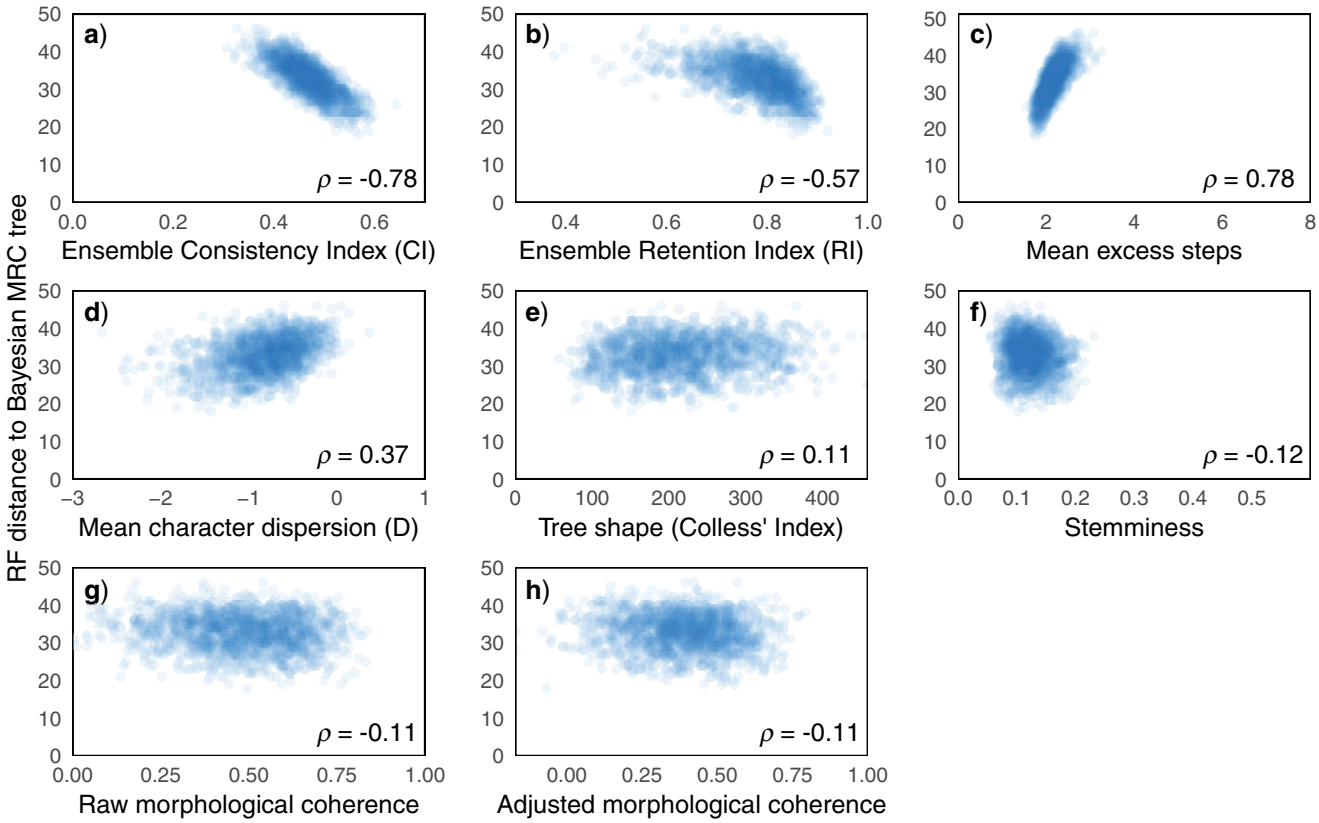

FIGURE 7. Scatter plots of correlation of eight measured tree and data attributes against RF distance between the true tree and the Bayesian MRC tree for TREvoSim datasets of 128, 512, and 1024 simulated characters (500 replicates). ρ = Spearman's Rank correlation.

that this is a general feature of morphological evolution (Lovette et al. 2002; Lloyd et al. 2012; Lee et al. 2013; Rabosky et al. 2013; Beck and Lee 2014; Puttick et al. 2014; Close et al. 2015; Wang and Lloyd 2016; Castiglione et al. 2018). It is also clear that among-lineage rate heterogeneity is not independent of selection. Elevated evolutionary rates have been frequently linked to adaptive radiations (e.g. Lee et al. 2013; Beck and Lee 2014; Close et al. 2015). Furthermore, traits that are probably subject to strong selection, such as body size, correlate with inferred evolutionary rate (Rabosky et al. 2013). As such, we would expect nonrandom distributions of homoplasy among lineages in real morphological data, particularly for data spanning macroevolutionary scales and/or adaptive radiations. These are exactly the scenarios for which morphology, especially fossil data, is a key source of information and is most frequently applied. Unfortunately, it is also under these circumstances that phylogenetic analyses—parsimony and Bayesian alike—struggle to recover accurate phylogenetic estimates.

Fortunately, more sophisticated Bayesian models offer potential solutions to this problem. For example, in contrast to standard Bayesian Mk analyses, those which employ relaxed morphological clocks do not assume constant evolutionary rate per lineage. Such models can better accommodate data containing nonrandom distributions of homoplasy. Clock models can have a profound effect on topology estimates for morphological data (King et al. 2017), but further simulation studies using models that include rate heterogeneity are required to determine if phylogenetic inference using a relaxed morphological clock is more accurate than a standard Mk analysis. Alternatives to the Mk model that incorporate directional character evolution may be more appropriate for data that have evolved under selection. For continuous morphological data, there are numerous well-established Gaussian models, including the Ornstein–Uhlenbeck model (Beaulieu et al. 2012) and the Lévy process (Landis et al. 2013). A recent simulation study by Parins-Fukuchi (2018) suggests that continuous characters perform at least as well as discrete characters in phylogenetic estimation. As such, directional continuous character models provide an intriguing alternative to standard discrete morphological models. Alternatively, Klopfstein et al. (2015) provide a nonstationary Markov model for directional evolution of discrete characters, although this has yet to be used to estimate tree topology. Assessing the efficacy of these different approaches will require comparison of their performance with simulated data containing nonrandom distributions of homoplasy among lineages, coupled with careful consideration of the prevalence of nonrandom homoplasy distributions among lineages within empirical data.

Conclusions

Here, we provide two new evolutionary models that derive trees and phylogenetic character data simultaneously: one in which lineages evolve stochastically, and the other at the level of individuals undergoing natural selection. We demonstrate that Bayesian searches are more accurate than parsimony searches using their respective standard consensus methods: Bayesian MRC trees have a higher percentage of correct nodes/quartets than parsimony SC trees under both equal and IW. Through in-depth characterization of the properties of the data using a variety of metrics, we find that homoplasy and character dispersion are related to phylogenetic accuracy. Tree estimation using selection-generated data is generally less accurate than stochastically generated data, despite having less homoplasy. We interpret this as resulting from the lower morphological coherencey of selection data: There is a weaker relationship between the phylogenetic and character distance between taxon pairs, which lead to a prevalence of wildcard taxa. Our results indicate that the inclusion of selection into models of character evolution potentially violates some important tenets of phylogenetic estimation and impacts our ability to resolve the correct tree. This could be problematic given the important role that selection plays in morphological evolution, despite having not previously been accounted for in phylogenetic simulations of this type. As such, rather than focusing on the relative merits of Bayesian or parsimony analyses of morphological data, future analyses might be better directed at identifying modes and patterns of morphological evolution that can ultimately be incorporated into more nuanced models for phylogenetic inference.

Supplementary Material

Data available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.4b8gtht8h Data analysis Zenodo repository associated with this article (DOI:10.5281/zenodo.3609738). TREvoSim v1.0.0 (https://github.com/palaeoware/trevosim/; doi: 10.5281/zenodo.3619356) MBL2017 v2.0.0 (https://github.com/palaeoware/MBL2017; doi:10.5281/zenodo.3614075).

Acknowledgements

We thank Rachel Warnock, Julia Sigwart, Roger Benson, James McInerney, and Martin Smith for discussion; Mark Puttick and Dominic Bennett for assistance in R; and Chris Whidden and Frederick Matsen for coding USPR and sharing it under a GNU General Public License. R.J.G would like to acknowledge the assistance of the Software Sustainability Institute. The work carried out by the Software Sustainability Institute is supported by the UK Engineering and Physical Sciences Research Council (EPSRC) through grant EP/H043160/1 and EPSRC, BBSRC, and ESRC Grant EP/N006410/1. We would like to thank our reviewers—three anonymous reviewers, David Polly, and Mark Wilkinson—in addition to editors Robert Thomson and Jeanne M. Serb, for suggestions that significantly strengthened the manuscript.

Funding

This work was supported by BBSRC award BB/N015827/1 to R.S.S. and R.J.G., and NERC award NE/T000813/1 to R.J.G. and R.S.S.

References

Ackermann R.R., Cheverud J.M. 2004. Detecting genetic drift versus selection in human evolution. Proc. Natl. Acad. Sci. USA 101:17946–17951.

Beaulieu J.M., Jhwueng D.-C., Boettiger C., O'Meara B.C. 2012. Modeling stabilizing selection: expanding the Ornstein–Uhlenbeck model of adaptive evolution. Evolution. 66:2369–2383.

Beck R.M.D., Lee M.S.Y. 2014. Ancient dates or accelerated rates? Morphological clocks and the antiquity of placental mammals. Proc. Biol. Sci. 281:20141278.

Bortolussi N., Durand E., Blum M., François O. 2006. apTreeshape: statistical analysis of phylogenetic tree shape. Bioinformatics. 22:363–364.

Brazeau M.D. 2011. Problematic character coding methods in morphology and their effects. Biol. J. Linn. Soc. 104:489–498.

Brown J.W., Parins-Fukuchi C., Stull G.W., Vargas O.M., Smith S.A. 2017. Bayesian and likelihood phylogenetic reconstructions of morphological traits are not discordant when taking uncertainty into consideration: a comment on Puttick et al. Proceedings of the Royal Society B: Biological Sciences, 284:20170986.

Castiglione S., Tesone G., Piccolo M., Melchionna M., Mondanaro A., Serio C., Febbraro M.D., Raia P. 2018. A new method for testing evolutionary rate variation and shifts in phenotypic evolution. Methods Ecol. Evol. 9:974–983.

Close R.A., Friedman M., Lloyd G.T., Benson R.B.J. 2015. Evidence for a mid-jurassic adaptive radiation in mammals. Curr. Biol. 25:2137–2142.

Colless D.H. 1982. Review of phylogenetics: the theory and practice of phylogenetic systematics. Syst. Zool. 31:100–104.

Congreve C.R., Lamsdell J.C. 2016. Implied weighting and its utility in palaeontological datasets: a study using modelled phylogenetic matrices. Palaeontology. 59:447–462.

Donoghue M.J., Doyle J.A., Gauthier J., Kluge A.G., Rowe T.B. 1989. The importance of fossils in phylogeny reconstruction. Annu. Rev. Ecol. Syst. 20:431–460.

Donoghue P.C.J., Yang Z. 2016. The evolution of methods for establishing evolutionary timescales. Phil. Trans. R. Soc. B. 371:20160020.

Farris J. 1989. The retention index and rescaled consistency index. Cladistics. 5:417–419.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17:368–376.

Fiala K.L., Sokal R.R. 1985. Factors determining the accuracy of cladogram estimation: evaluation using computer simulation. Evolution. 39:609–622.

Foote M. 1999. Evolutionary and preservational constraints on origins of biologic groups: divergence times of eutherian mammals. Science. 283:1310–1314.

Fritz S.A., Purvis A. 2010. Selectivity in mammalian extinction risk and threat types: a new measure of phylogenetic signal strength in binary traits. Conserv. Biol. 24:1042–1051.

Garwood R.J., Spencer A.R.T., Sutton M.D. 2019. REvoSim: organism-level simulation of macro and microevolution. Palaeontology. 62:339–355.

Goloboff P.A. 2013. Extended implied weighting. Cladistics. 30:260–272.

Goloboff P.A., Catalano S.A. 2016. TNT version 1.5, including a full implementation of phylogenetic morphometrics. Cladistics. 32:221–238.
Goloboff P.A., Torres A., Arias J.S. 2018. Weighted parsimony outperforms other methods of phylogenetic inference under models appropriate for morphology. Cladistics. 34:407–437.
Gould S.J., Raup D.M., Sepkoski J.J., Schopf T.J.M., Simberloff D.S. 1977. The shape of evolution: a comparison of real and random clades. Paleobiology. 3:23–40.
Hackathon R., Bolker B., Butler M., Cowan P., Vienne D.D., Eddelbuettel D. 2011. phylobase: base package for phylogenetic structures and comparative data. R package version 0.6. 3. Available from: https://cran.r-project.org/web/packages/phylobase/index.html.
Hasegawa M., Kishino H., Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22:160–174.
Hennig D.W. 1950. Grundzüge einer Theorie der phylogenetischen Systematik. Berlin: Deutscher Zentralverlag.
Hennig W. 1965. Phylogenetic systematics. Annu. Rev. Entomol. 10:97–116.
Ho W.-C., Ohya Y., Zhang J. 2017. Testing the neutral hypothesis of phenotypic evolution. Proc. Natl. Acad. Sci. USA 114: 12219–12224.
Huneman P. 2014. Mapping an expanding territory: computer simulations in evolutionary biology. Hist. Philos. Life Sci. 36:60–89.
Huss J. 2009. The shape of evolution: the MBL model and clade shape. In: Sepkoski D., Ruse M., editors. The paleobiological revolution: essays on the growth of modern paleontology. Chicago: University of Chicago Press. p. 326–345.
Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16:111–120.
King B., Qiao T., Lee M.S.Y., Zhu M., Long J.A. 2017. Bayesian morphological clock methods resurrect placoderm monophyly and reveal rapid early evolution in jawed vertebrates. Syst. Biol. 66:499–516.
Kitching I.J., Forey P.L., Humphries C.J., Williams D.M. 1998. Cladistics—second edition—the theory and practice of parsimony analysis. Oxford: Oxford University Press.
Klopfstein S., Vilhelmsen L., Ronquist F. 2015. A nonstationary markov model detects directional evolution in hymenopteran morphology. Syst. Biol. 64:1089–1103.
Kluge A.G., Farris J.S. 1969. Quantitative phyletics and the evolution of anurans. Syst. Zool. 18:1–32.
Kuhner M.K., Yamato J. 2015. Practical performance of tree comparison metrics. Syst. Biol. 64:205–214.
Lande R. 1976. Natural selection and random genetic drift in phenotypic evolution. Evolution. 30:314–334.
Landis M.J., Schraiber J.G., Liang M. 2013. Phylogenetic analysis using Lévy processes: finding jumps in the evolution of continuous traits. Syst. Biol. 62:193–204.
Lartillot N., Phillips M.J., Ronquist F. 2016. A mixed relaxed clock model. Philos. Trans. R. Soc. Lond. B Biol. Sci. 371:20150132.
Lee M.S.Y., Palci A. 2015. Morphological phylogenetics in the genomic age. Curr. Biol. 25:R922–R929.
Lee M.S.Y., Soubrier J., Edgecombe G.D. 2013. Rates of phenotypic and genomic evolution during the Cambrian explosion. Curr. Biol. 23:1889–1895.
Legg D.A., Sutton M.D., Edgecombe G.D. 2013. Arthropod fossil data increase congruence of morphological and molecular phylogenies. Nat. Commun. 4:2485.
Lewis P.O. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. Syst. Biol. 50:913–925.
Lloyd G.T., Wang S.C., Brusatte S.L. 2012. Identifying heterogeneity in rates of morphological evolution: discrete character change in the evolution of lungfish (Sarcopterygii; Dipnoi). Evolution. 66:330–348.
Lovette I.J., Bermingham E., Ricklefs R.E. 2002. Clade-specific morphological diversification and adaptive radiation in Hawaiian songbirds. Proc. R. Soc. Lond. B Biol. Sci. 269:37–42.
Lynch M. 1990. The rate of morphological evolution in mammals from the standpoint of the neutral expectation. Am. Nat. 136:727–741.
Marroig G., Cheverud J.M. 2004. Did natural selection or genetic drift produce the cranial diversification of neotropical monkeys? Am. Nat. 163:417–428.
O'Reilly J.E., Puttick M.N., Parry L., Tanner A.R., Tarver J.E., Fleming J., Pisani D., Donoghue P.C.J. 2016. Bayesian methods outperform parsimony but at the expense of precision in the estimation of phylogeny from discrete morphological data. Biol. Lett. 12:20160081.
O'Reilly J.E., Puttick M.N., Pisani D., Donoghue P.C.J. 2018a. Probabilistic methods surpass parsimony when assessing clade support in phylogenetic analyses of discrete morphological data. Palaeontology. 61:105–118.
O'Reilly J.E., Puttick M.N., Pisani D., Donoghue P.C.J. 2018b. Empirical realism of simulated data is more important than the model used to generate it: a reply to Goloboff *et al*. Palaeontology. 61:631–635.
Orme D., Freckleton R., Thomas G., Petzoldt T., Fritz S., Isaac N., Pearse W. 2012. Caper: comparative analyses of phylogenetics and evolution in R. Available from: https://CRAN.R-project.org/package=caper/.
Paradis, E., J. Claude, and K. Strimmer. 2004. APE: Analyses of phylogenetics and evolution in R language. Bioinformatics. 20:289–290.
Parins-Fukuchi C. 2018. Bayesian placement of fossils on phylogenies using quantitative morphometric data. Evolution. 72:1801–1814.
Puttick M.N., O'Reilly J.E., Oakley D., Tanner A.R., Fleming J.F., Clark J., Holloway L., Lozano-Fernandez J., Parry L.A., Tarver J.E., Pisani D., Donoghue P.C.J. 2017a. Parsimony and maximum-likelihood phylogenetic analyses of morphology do not generally integrate uncertainty in inferring evolutionary history: a response to Brown *et al.* Proc. R. Soc. B. 284:20171636.
Puttick M.N., O'Reilly J.E., Pisani D., Donoghue P.C.J. 2019. Probabilistic methods outperform parsimony in the phylogenetic analysis of data simulated without a probabilistic model. Palaeontology. 62:1–17.
Puttick M.N., O'Reilly J.E., Tanner A.R., Fleming J.F., Clark J., Holloway L., Lozano-Fernandez J., Parry L.A., Tarver J.E., Pisani D., Donoghue P.C.J. 2017b. Uncertain-tree: discriminating among competing approaches to the phylogenetic analysis of phenotype data. Proc. R. Soc. B. 284:20162290.
Puttick M.N., Thomas G.H., Benton M.J. 2014. High rates of evolution preceded the origin of birds. Evolution. 68:1497–1510.
Rabosky D.L., Santini F., Eastman J., Smith S.A., Sidlauskas B., Chang J., Alfaro M.E. 2013. Rates of speciation and morphological evolution are correlated across the largest vertebrate radiation. Nat. Commun. 4:1958.
Raff R.A. 2007. Written in stone: fossils, genes and evo–devo. Nat. Rev. Genet. 8:911–920.
Raup D.M., Gould S.J. 1974. Stochastic simulation and evolution of morphology-towards a nomothetic paleontology. Syst. Zool. 23:305–322.
Raup D.M., Gould S.J., Schopf T.J.M., Simberloff D.S. 1973. Stochastic models of phylogeny and the evolution of diversity. J. Geol. 81:525–542.
Raup D.M., Sepkoski J.J. 1982. Mass extinctions in the marine fossil record. Science. 215:1501–1503.
Revell L.J. 2012. phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol. Evol. 3:217–223.
Rieseberg L.H., Widmer A., Arntz A.M., Burke J.M. 2002. Directional selection is the primary cause of phenotypic diversification. Proc. Natl. Acad. Sci. USA 99:12242–12245.
Rohlf F.J., Chang W.S., Sokal R.R., Kim J. 1990. Accuracy of estimated phylogenies: effects of tree topology and evolutionary model. Evolution. 44:1671–1684.
Ronquist F., Teslenko M., van der Mark P., Ayres D.L., Darling A., Höhna S., Larget B., Liu L., Suchard M.A., Huelsenbeck J.P. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61:539–542.
Schliep K.P. 2011. phangorn: phylogenetic analysis in R. Bioinformatics. 27:592–593.
Sepkoski D. 2012. Towards a nomothetic paleontology: the MBL model and stochastic paleontology. In: Sepkoski D, editor. Rereading the fossil record: the growth of paleobiology as an evolutionary discipline. Chicago: University of Chicago Press. p. 215–270.

Sepkoski J.J. 1978. A kinetic model of phanerozoic taxonomic diversity I. Analysis of marine orders. Paleobiology. 4:223–251.
Sigwart J.D., Sutton M.D., Bennett K.D. 2018. How big is a genus? Towards a nomothetic systematics. Zool. J. Linn. Soc. 183: 237–252.
Simpson E.H. 1951. The interpretation of interaction in contingency tables. J. R. Stat. Soc. B. 13:238–241.
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. Available from: https://cran.r-project.org/web/packages/Quartet/index.html.
Stanley S.M., Signor P.W., Lidgard S., Karr A.F. 1981. Natural clades differ from "random" clades: simulations and analyses. Paleobiology. 7:115–127.
Uhen M.D. 1996. An evaluation of clade-shape statistics using simulations and extinct families of mammals. Paleobiology. 22:8–22.
Wang M., Lloyd G.T. 2016. Rates of morphological evolution are heterogeneous in early cretaceous birds. Proc. R. Soc. B Biol. Sci. 283:20160214.
Whidden C., Matsen F. 2018. Calculating the unrooted subtree prune-and-regraft distance. IEEE/ACM Trans. Comput. Biol. Bioinf. 16:898–911.
Wickham H. 2016. ggplot2: elegant graphics for data analysis. 2nd ed. New York: Hadley Wickham Springer-Verlag.
Wiens J.J., Soltis P. 2005. Can incomplete taxa rescue phylogenetic analyses from long-branch attraction? Syst. Biol. 54:731–742.
Wright A.M., Hillis D.M. 2014. Bayesian analysis using a simple likelihood model outperforms parsimony for estimation of phylogeny from discrete morphological data. PLoS One. 9:e109210.