Palaeontology
The Palaeontological Association
www.palass.org

[Palaeontology, 2018, pp. 1–17]

SYMPOSIUM

PROBABILISTIC METHODS OUTPERFORM PARSIMONY IN THE PHYLOGENETIC ANALYSIS OF DATA SIMULATED WITHOUT A PROBABILISTIC MODEL

by MARK N. PUTTICK[1,2,3,*], JOSEPH E. O'REILLY[1,2,*], DAVIDE PISANI[1,4] *and* PHILIP C. J. DONOGHUE[1]

[1]School of Earth Sciences, University of Bristol, Life Sciences Building, Tyndall Avenue, Bristol, BS8 1TQ, UK; mark.puttick@bristol.ac.uk, joe.oreilly@bristol.ac.uk davide.pisani@bristol.ac.uk, phil.donoghue@bristol.ac.uk
[2]Department of Life Sciences, The Natural History Museum, Cromwell Road, London, SW7 5BD, UK
[3]School of Biochemistry & Biological Sciences, University of Bath, Bath, UK
[4]School of Biological Sciences, University of Bristol, Life Sciences Building, Tyndall Avenue, Bristol, BS8 1TQ, UK

Typescript received 19 March 2018; accepted in revised form 10 June 2018

Abstract: To understand patterns and processes of the diversification of life, we require an accurate understanding of taxon interrelationships. Recent studies have suggested that analyses of morphological character data using the Bayesian and maximum likelihood Mk model provide phylogenies of higher accuracy compared to parsimony methods. This has proved controversial, particularly studies simulating morphology-data under Markov models that assume shared branch lengths for characters, as it is claimed this leads to bias favouring the Bayesian or maximum likelihood Mk model over parsimony models which do not explicitly make this assumption. We avoid these potential issues by employing a simulation protocol in which character states are randomly assigned to tips, but datasets are constrained to an empirically realistic distribution of homoplasy as measured by the consistency index. Datasets were analysed with equal weights and implied weights parsimony, and the maximum likelihood and Bayesian Mk model. We find that consistent (low homoplasy) datasets render method choice largely irrelevant, as all methods perform well with high consistency (low homoplasy) datasets, but the largest discrepancies in accuracy occur with low consistency datasets (high homoplasy). In such cases, the Bayesian Mk model is significantly more accurate than alternative models and implied weights parsimony never significantly outperforms the Bayesian Mk model. When poorly supported branches are collapsed, the Bayesian Mk model recovers trees with higher resolution compared to other methods. As it is not possible to assess homoplasy independently of a tree estimate, the Bayesian Mk model emerges as the most reliable approach for categorical morphological analyses.

Key words: phylogenetics, parsimony, likelihood, Bayesian, morphology, simulation.

MORPHOLOGY is integral to restoring fossil species to their rightful place among their living relatives within the tree of life, which is prerequisite to inferring their evolutionary significance. It is tempting to conclude that the hegemony of parsimony is the consequence of an absence of competing phylogenetic methods, yet parsimony methods have undergone modest diversification (Goloboff 1997) and a simple Markov model of character change has been available for more than a decade (Lewis 2001). Rather, it is perhaps the development and enthusiastic adoption of phylogenetic comparative methods by palaeontologists which has led to renewed interest in the relative performance of morphology-based phylogenetic methods. Indeed, it has become conventional to undertake parallel analyses of morphological datasets using the gamut of phylogenetic methods (e.g. O'Leary *et al.* 2013; Parry *et al.* 2016), but it is not possible to determine which method yields the most accurate estimate when the true phylogeny is unknown. Hence, a number of studies have resorted to simulations, testing between competing phylogenetic methods based on morphology-like datasets generated

*These authors contributed equally to the manuscript.

doi: 10.1111/pala.12388

from known phylogenies (Wright & Hillis 2014; Congreve & Lamsdell 2016; O'Reilly *et al.* 2016, 2017; Brown *et al.* 2017; Puttick *et al.* 2017*a*, *b*; Goloboff *et al.* 2017). A number of these studies have relied on continuous time Markov models to simulate morphology-like data (Wright & Hillis 2014; O'Reilly *et al.* 2016, 2017; Brown *et al.* 2017; Puttick *et al.* 2017*a*, *b*) and, while some have attempted to generate data that violate a number of the assumptions of the Mk model used in statistical phylogenetic methods (Brown *et al.* 2017), it has been argued that they remain biased against parsimony methods (Goloboff *et al.* 2017, 2018). Specifically, Goloboff *et al.* (2017, 2018) argued that the framework employed in previous simulations stretches and compresses all the branches of a tree by the same factor when altering the underlying rate at which character state changes occur in individual characters. Goloboff *et al.* (2017, 2018) stated this approach to simulation has given an advantage to maximum likelihood and Bayesian analyses using the Mk model, which make this assumption about character evolution, over parsimony-based methods in these simulation-based benchmarking analyses of phylogenetic methods.

Here, we address concerns with previous simulation approaches, using a protocol for generating approximately random data on a known tree, and establishing the empirical realism of simulated data by ensuring that they meet the expectations of real morphological datasets, based on an analysis of the distribution of character consistency in empirical data. As such, the ensuing simulated datasets violate the assumptions of probabilistic evolutionary models and, if anything, are likely to favour parsimony-based methods, including equal weights and implied weights parsimony due to the implicit assumptions of these approaches (Tuffley & Steel 1997). In particular, our simulation protocol generates datasets in which all character state changes are observed and there is no assumption of equality or proportionality in the branch lengths among different characters. The results of our analyses follow previous studies in demonstrating that the accuracy and precision of all phylogenetic methods increase with the scale of the available data. The Bayesian implementation of the Mk model performs best in the analysis of datasets exhibiting very high levels of homoplasy, reaching significantly higher levels of accuracy. In contrast, at levels of homoplasy in which implied weights achieves higher accuracy, there is only a minimal (nonsignificant) improvement compared to the other methods. Since the consistency of characters and datasets is never known in isolation from the generating tree, we conclude that the Bayesian implementation of the Mk model should be preferred for the phylogenetic analysis of empirical categorical morphological data.

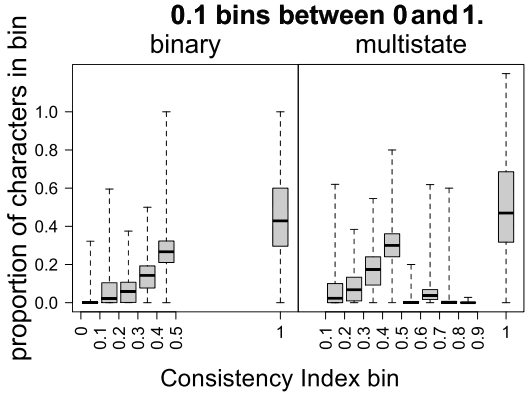
METHOD

With few exceptions (e.g. experimental viral strains: Hillis *et al.* 1992; Cunningham *et al.* 1998) no phylogeny estimated from empirical data can be demonstrated to represent the true relationships of its constituent taxa. Thus, any meaningful test of phylogenetic method efficacy requires a known tree upon which data are simulated. However, in contrast to molecular models of evolution, there is no unifying empirical or theoretical model to describe the process by which changes in morphological character states accumulate through time or between lineages. Previous approaches to simulating morphological data have pragmatically co-opted models of molecular evolution to produce approximately realistic datasets (Wright & Hillis 2014; O'Reilly *et al.* 2016, 2017; Puttick *et al.* 2017*a*), but this approach faces two main criticisms. First, if we do not understand how morphological cladistic data are expected to evolve, it is impossible to apply a model to simulate realistic data (Goloboff *et al.* 2017, 2018). Second, data simulated in a framework where characters share branch lengths or proportional branch lengths under a model of between character rate heterogeneity (Yang 1994) are biased in favour of an inference framework that explicitly makes the same assumption (e.g. maximum likelihood and Bayesian implementations of the Mk model) over a nonprobabilistic framework where these assumptions are not made (e.g. equal weights or implied weights parsimony) (Goloboff *et al.* 2017, 2018).

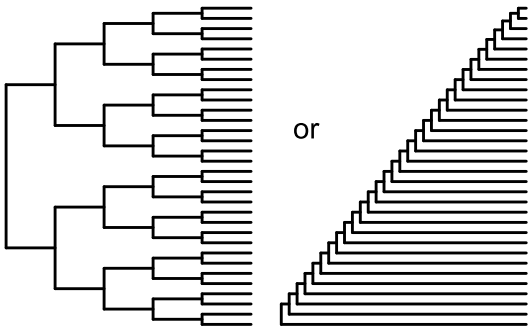
We cannot address the first criticism since, if we had an accurate model of morphological evolution, there would be no debate; it could be implemented in phylogenetic analysis. In the absence of a realistic model of morphological evolution, we attend to the second criticism employing a procedure that generates datasets with a realistic distribution of homoplasy across multiple simulated matrices but does not require the assumption of shared branch lengths among characters. In this

FIG. 1. Schematic of workflow followed in simulating and analysing the data. A, variance in the consistency index (CI) of characters was assayed based on a large compilation of empirical datasets. Data were simulated on two strictly bifurcating trees, one symmetrical and another maximally asymmetrical (B), characters were randomly asserted to tips (C), with the ensuing characters allocated to ten bins according to their CI. Matrices were assembled by drawing characters from these ten bins until the desired matrix size (100, 350, 1000 characters) and CI was achieved (D). The matrices were then analysed using equal weights and implied weights parsimony in TNT, under the maximum likelihood implementation of the Mk model in IQ-Tree, and using Bayesian implementation of the Mk model in MrBayes (E). Colour online.

A Estimate empirical distribution of per-character Consistency Index (CI) and place in 0.1 bins between 0 and 1.

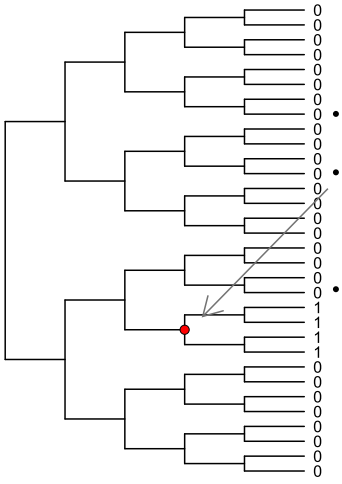


B True trees. Fully symmetric and fully asymmetric with no edge lengths



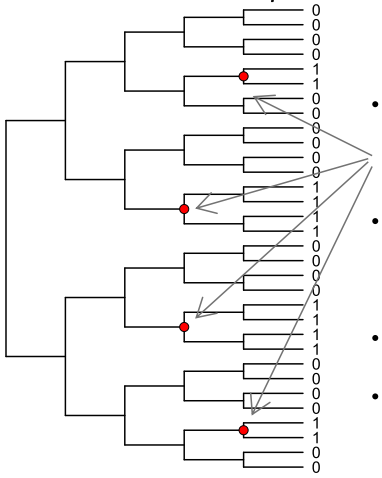
C Simulate datasets that have binary or a mixture of binary and multistate characters (max states = 4). A character is either homologous or homoplastic with a probability taken from the empirical survey (from A).

Homologous



- One internal node chosen at random
- All nodes (except root) are equally likely to be chosen
- No memory of previously selected nodes

Homoplastic

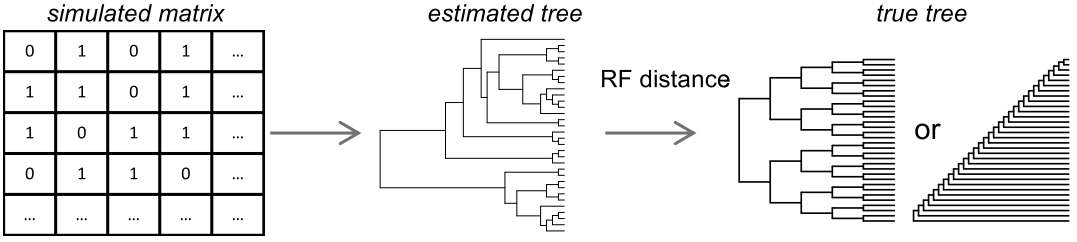


- Random *n* nodes chosen to define paraphyletic clades
- All nodes equally likely to be chosen in any combination
- Number of nodes (*n*) is random
- No memory of previously selected nodes

D Repeat C until reaching a set number of characters (*100, 350,* or *1000* characters) and when the matrix-wide CI value falls in the specified CI bin. Repeat so each CI bin on each tree has 100 independent datasets

Matrix CI value		0.0-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1.0
Number of datasets	binary	100	100	100	100	100	100	100	100	100	100
	mixture		100	100	100	100	100	100	100	100	100

E Analyse datasets using Parsimony, Implied Weights (k=2, k=10, k=20), Maximum likelihood, and Bayesian phylogenetics and compare each inferred tree with relevant generating tree



simulation framework (Fig. 1), we generated data on two trees comprised of 32 tips: (1) fully symmetrical; and (2) maximally asymmetrical. Tips of the known trees were randomly assigned character states. Tip states were designated using a procedure that considered the topology only with no branch lengths or continuous time model.

Characters were assigned to the tree by first selecting if a character was homologous (CI = 1) or homoplastic (0 < CI < 1). The probability of a character being homologous or homoplastic was informed by a survey of empirical data, and the overall proportion of homologous: homoplastic characters in each matrix varied in each of the ten consistency index (CI) bins (Fig. 1). If a character was selected as homologous, a monophyletic group was selected at random and all its tips were given a shared unique character. If a character was selected to be homoplastic, characters were assigned to tips that did not form a monophyletic group. All internal nodes and tips were equally likely to be sampled in the simulation (apart from the root node), and the selection was memory-less so the same node(s) or tip(s) could be selected multiple times. This process was repeated until (1) the set number of characters was simulated, and (2) the matrix-wide CI value fell within the desired bin. All simulated characters were variable.

The simulation procedure used here with the random assortments of character state data ensures that no simulating proportional branch lengths (measured as the expected number of changes per character) are incorporated. We do not consider this to be a realistic model of morphological evolution; it was designed as a procedure for simulating morphology-like datasets that is independent of the Markov models used to estimate tree topologies and, therefore, favours neither likelihood-based nor parsimony methods. The simulation procedure yields empirically realistic datasets, as assayed by the distribution of character consistency exhibited by the generated datasets. Importantly, the procedure does not impose equal or proportional branch lengths, describing evolutionary distances between taxa, among characters, nor does it allow for unobserved changes; both features of previous model-based simulations. As such, the procedure does not favour model-based phylogenetic methods and, indeed, diminishes their benefits, such as in accounting for unobserved changes.

For each character in the simulated datasets, character states were assigned to leaves following the procedure described above. However, a nonrandom filtering strategy was then applied to ensure that the datasets used in our analyses matched the characteristics observed in empirical datasets. Specifically, we only retained datasets that matched the levels of character congruence exhibited by real datasets as measured using the consistency index (CI), a scaled (0–1) measure of the congruence of a character to a tree (Kluge & Farris 1969). CI can be measured character-by-character or averaged over all the characters of a matrix; both of these measures are used in our simulation procedure. Here, we use a framework of 10 bins of CI (0–1 increasing in increments of 0.1). Overall, we simulated datasets that possessed whole-matrix CI values (based on the true tree) that fell into each of those bins, respectively (100 simulations per bin). For each of these datasets, the per-character CI values were based upon empirical estimates of per-character CIs in each of the 10 bins.

Empirical estimates of per-character CI were measured using 1544 empirical datasets spanning all kingdoms of life (Wright *et al.* 2016). For each dataset, the CI of each character was calculated using a single most parsimonious tree estimated from the empirical data. For each matrix, we calculated the overall proportion of characters that fall into each of the 10 CI bins (Fig. 2). We then pooled the distributions of the proportion of characters that fall into each CI bin from all the empirical datasets to get a global distribution of the proportion of characters that fall into each CI bin. These distributions were then used as a target distribution to be approximated when simulating data. For example, if across all empirical datasets 40–50% of characters possess a CI value between 0.5 and 0.6, we constrain the simulations such that 40–50% of per-character CI values for the resulting simulated data fell within bin 0.5–0.6.

Using these per-matrix constraints of CI values, we simulated 100 datasets that possessed per-matrix CI values within each of the 10 bins, resulting in 1000 simulated matrices. Two unique cases of this general simulation procedure were produced, with matrices consisting of either exclusively binary [0, 1] or a mixture of binary and multistate characters with a maximum of four states: [0, 1, 2, 3]. Data were either designated as fully congruent with the tree (CI = 1) or incongruent (0 < CI < 1). We repeated this procedure to create datasets composed of 100, 350 and 1000 characters on both the fully symmetrical and the fully asymmetrical 32 tip generating trees. For the lower CI bins (0–0.1 binary characters, 0.1–0.2 multistate characters), the constraint of the per-character CI values was violated to obtain the correct per-matrix CI value. The lowest CI bin (0.0–0.1) was only used for the binary data, not multistate data. R scripts to simulate data are available in Dryad (Puttick *et al.* 2018).

Tree estimation

Trees were estimated from the simulated matrices using both equal weights and implied weights parsimony in

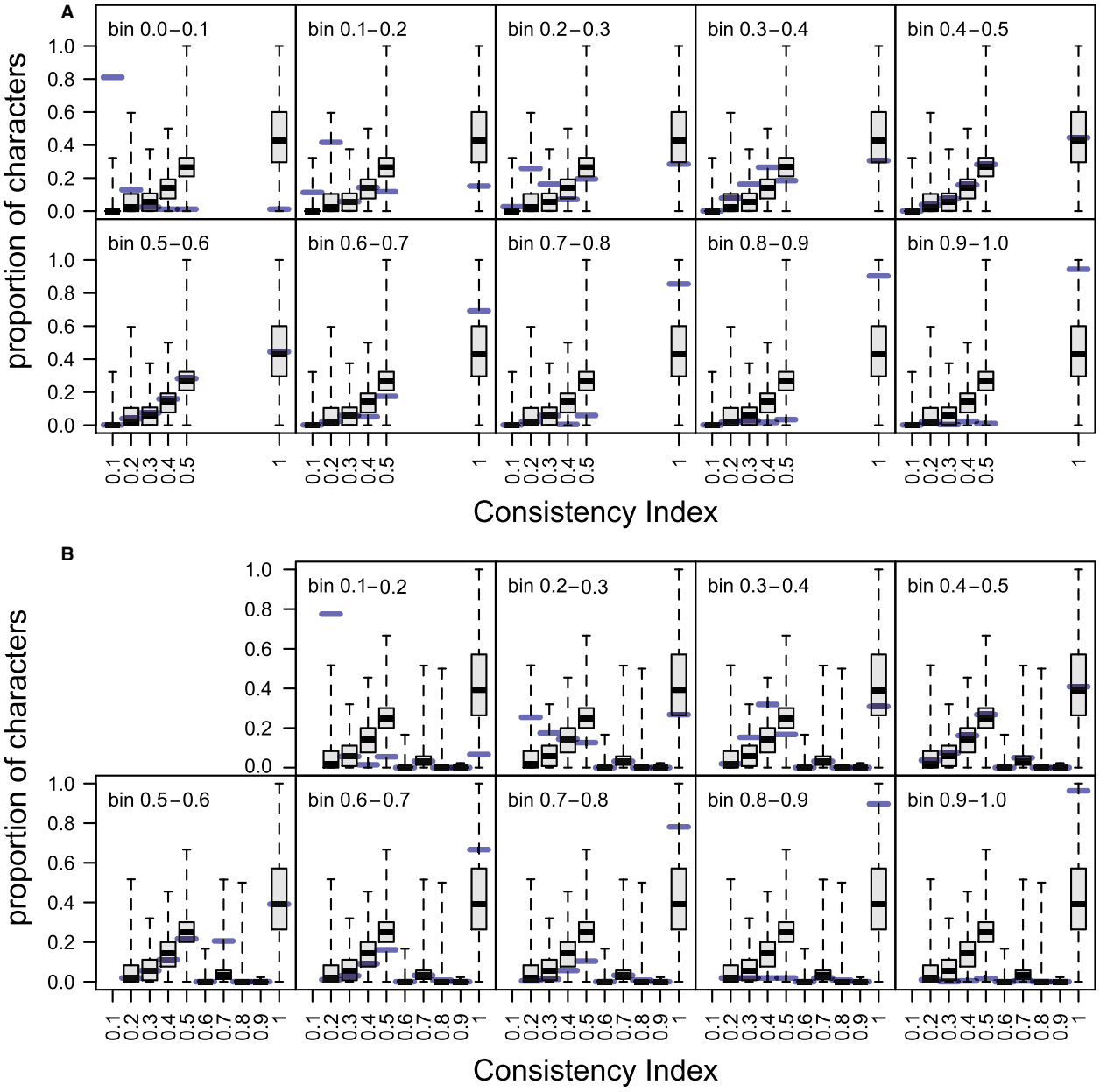


FIG. 2. Empirical distribution of the proportion of characters in each bin from empirical datasets for: A, binary; B, multistate characters. The value that was used to simulate datasets with whole-matrix CI value in each bin is shown in blue. The bin of whole-matrix values of 0.0–0.1 was not used for multistate data. Note that binary characters cannot achieve values of CI > 0.5 < 1.0.

TNT (Goloboff *et al.* 2003*a*, 2008; Goloboff & Catalona 2016), maximum likelihood in IQ-Tree (Minh *et al.* 2013) and with Bayesian inference using MrBayes 3.2.6 (Ronquist *et al.* 2012). For maximum likelihood and Bayesian estimation of trees, the Mkv model of morphological evolution was applied with rate heterogeneity modelled with a discretized gamma distribution with four categories. For Bayesian estimation of trees, 1 000 000 MCMC generations were performed for each replicate, with every 100th sample retained to produce a final posterior sample of 15 000 trees, over two runs of four metropolis-coupled chains after a 25% burn-in. A stop rule was also applied so that, if the standard deviation of split frequencies dropped below 0.01, then the analysis would automatically terminate as the posterior distribution had been judged to be adequately sampled. Three values of the concavity constant (k = 2, 10, 20) were tested for the implied weights parsimony analyses, and the results from analyses using different k-values were pooled together.

Output trees

We considered three consensus-tree types constructed from the output of the different inference frameworks: *standard output, split support > 0.5* and *split support > 0.95*. Standard output trees are outputs from each analysis: 50% majority-rule trees of the post burn-in MCMC sample obtained during Bayesian analysis, the fully bifurcating maximum likelihood estimate of topology, and the 50% majority-rule consensus constructed from the set of most parsimonious trees from equal weights or implied weights analyses.

As well as the standard output trees, we incorporated uncertainty in clade support into our analyses by collapsing splits into soft polytomies if their associated support value fell below a specified value. Bayesian estimation with MrBayes produces a 50% majority-rule consensus from the posterior sample by default; these trees were used in the subsequent analyses. For parsimony methods and maximum likelihood, we incorporated uncertainty via nonparametric bootstrapping. For maximum likelihood, we employed the ultrafast bootstrapping algorithm with 1000 replicates (Minh *et al.* 2013), and for equal weights and implied weights parsimony, we used nonparametric bootstrapping to measure the proportion of replicates containing the relevant split (Felsenstein 1985; Goloboff *et al.* 2003*b*). Using these proportions, we subsequently collapsed branches if they had lower than 0.5 bootstrap support (split support ≥ 0.5) or 0.95 support (split support ≥ 0.95).

Assessing the accuracy of topology estimates

We assessed topological accuracy by comparing the estimated tree topologies to the generating trees using the Robinson–Foulds distance between these two trees (Robinson & Foulds 1981). The Robinson–Foulds metric is equal to the sum of splits found in one tree but not the other: a value of zero indicates two trees that are either identical or that one tree is fully unresolved; higher values indicate increasing topological discordance. As this measure does not discriminate whether topological concordance is achieved because the estimated tree is similar and well resolved, or because it is poorly resolved, we also compared the distribution of Robinson–Foulds values with the number of resolved nodes in each estimated tree.

The proportion of accurate and inaccurate nodes

For each estimated tree, we identified the number of accurate and inaccurate nodes: those present and absent, respectively, in the generating tree. We also examined whether nodes that were deeper in the true phylogenies were more or less likely or to be accurately resolved than nodes closer to the tips of the tree.

RESULTS

All phylogenetic methods show increasing topological accuracy (lower Robinson–Foulds distance) with an increase in the number of analysed characters and/or an increase in character congruence with the simulation tree (greater CI values) (Figs 3–7; Puttick *et al.* 2018, figs S1–S11). The Bayesian implementation of the Mk model was able to recover the highest numbers of nodes when branches with less than 0.95 support are collapsed (Table 1; Puttick *et al.* 2018, tables S1, S3, S5, S7 and S9). All methods are generally more accurate when estimating topology using data simulated on the symmetrical tree than when analysing data simulated on the asymmetrical tree (Table 2; Puttick *et al.* 2018, tables S2, S4, S6, S8 and S10). Overall, the Bayesian implementation of the Mk model was able to recover the greatest number of correct nodes; this trend is most pronounced when only splits with ≥ 0.95 support are presented in the estimated topology.

Ability to resolve nodes

For the standard output trees, the Bayesian implementation of the Mk model recovers the fewest correct nodes in the lowest CI bins and the maximum likelihood implementation of the Mk model recovers the most nodes (the topology is strictly bifurcating) (Figs 5–7). The majority-rule consensus topologies of the most parsimonious trees from the equal weights and implied weights analyses tend to recover a large number of nodes, albeit with high variance (Figs 5–7; Table 1).

All methods, apart from the maximum likelihood implementation of the Mk model, struggle to recover nodes at ≥ 0.5 support, when analysing the smallest character matrices comprised of characters with the lowest consistency (Figs 5–7). In the 0.1 CI bin, the median resolution is for a single node for all methods apart from maximum likelihood (Figs 5–7). For both parsimony methods, only one node is resolved in the 0.1 bin with 350 and 1000 characters (Puttick *et al.* 2018, tables S1, S3). For CI bins 0.2–0.5 (inclusive), Bayesian and maximum likelihood implementations of the Mk model resolve a higher number of nodes than either parsimony method with both data types (binary and multistate). At CI bins of > 0.7, for 350 and 1000 character datasets, all methods achieve full resolution with almost no variance (Figs 5–7).

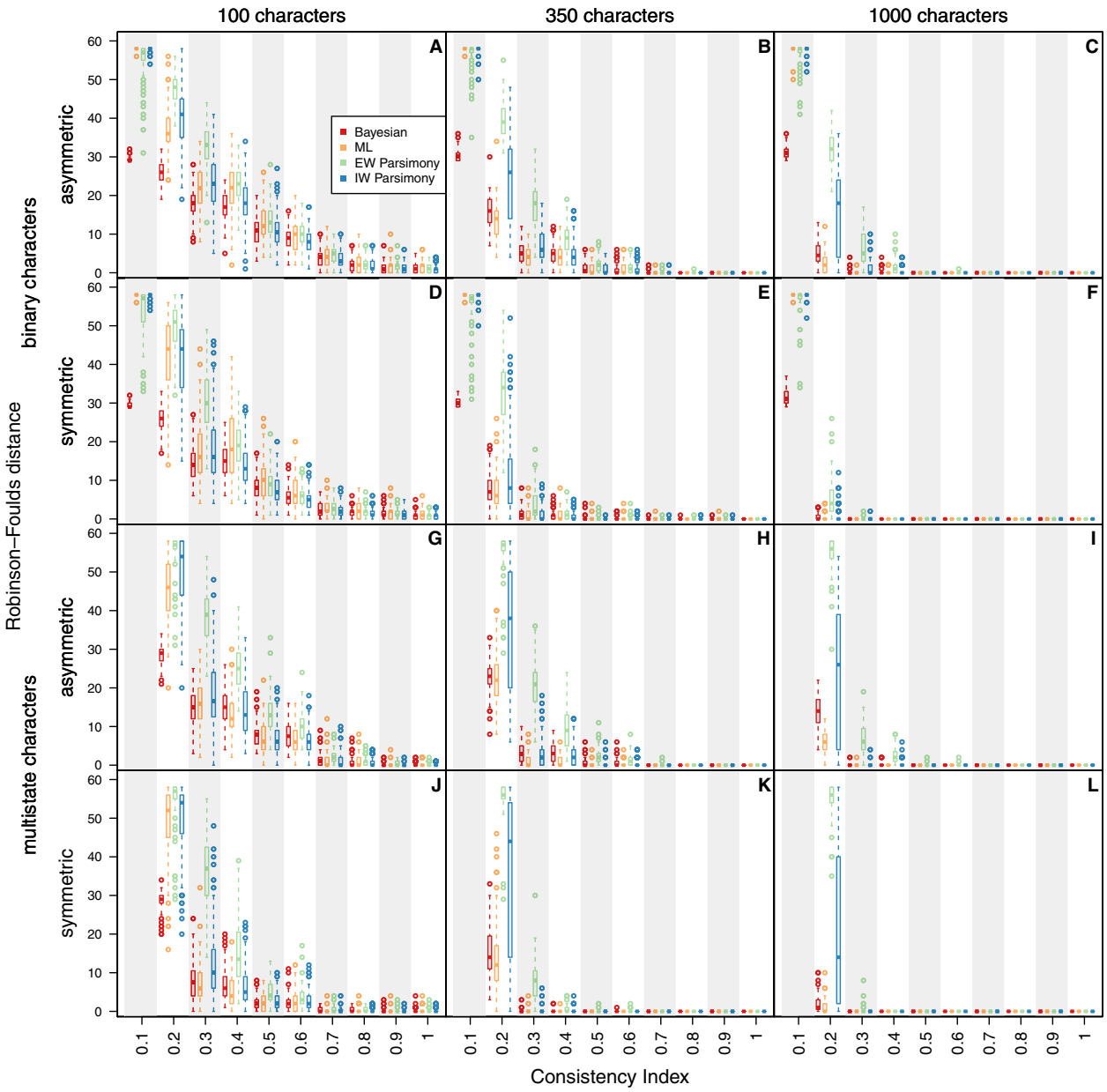


FIG. 3. Robinson–Foulds distances for standard output trees: 50% majority-rule Bayesian, equal weights (EW) parsimony, implied weights (IW) parsimony, and maximum likelihood (ML) trees recovered from analysis of the binary (A–F) and multistate (G–L) character datasets, generated from asymmetric (A–C, G–I) and symmetric (D–F, J–L) trees. All methods converge in the analysis of large datasets of very consistent characters. With small and low consistency datasets, Bayesian exhibits greatest accuracy, followed in order by IW parsimony, EW parsimony and ML.

For *0.95 support* trees, all methods estimate trees with few nodes when analysing 100 characters from the lower CI bins. However, Bayesian inference and maximum likelihood methods resolve a larger number of nodes when compared to either parsimony method; the Bayesian implementation of the Mk model recovers a greater number of nodes than the maximum likelihood implementation (Figs 5–7). These trends also hold at 350 and 1000 analysed characters, but only in the lower CI bins (Figs 5–7).

Tree accuracy

Different trends in accuracy are seen with different output types, but only at lower CI bins. Above a CI value of

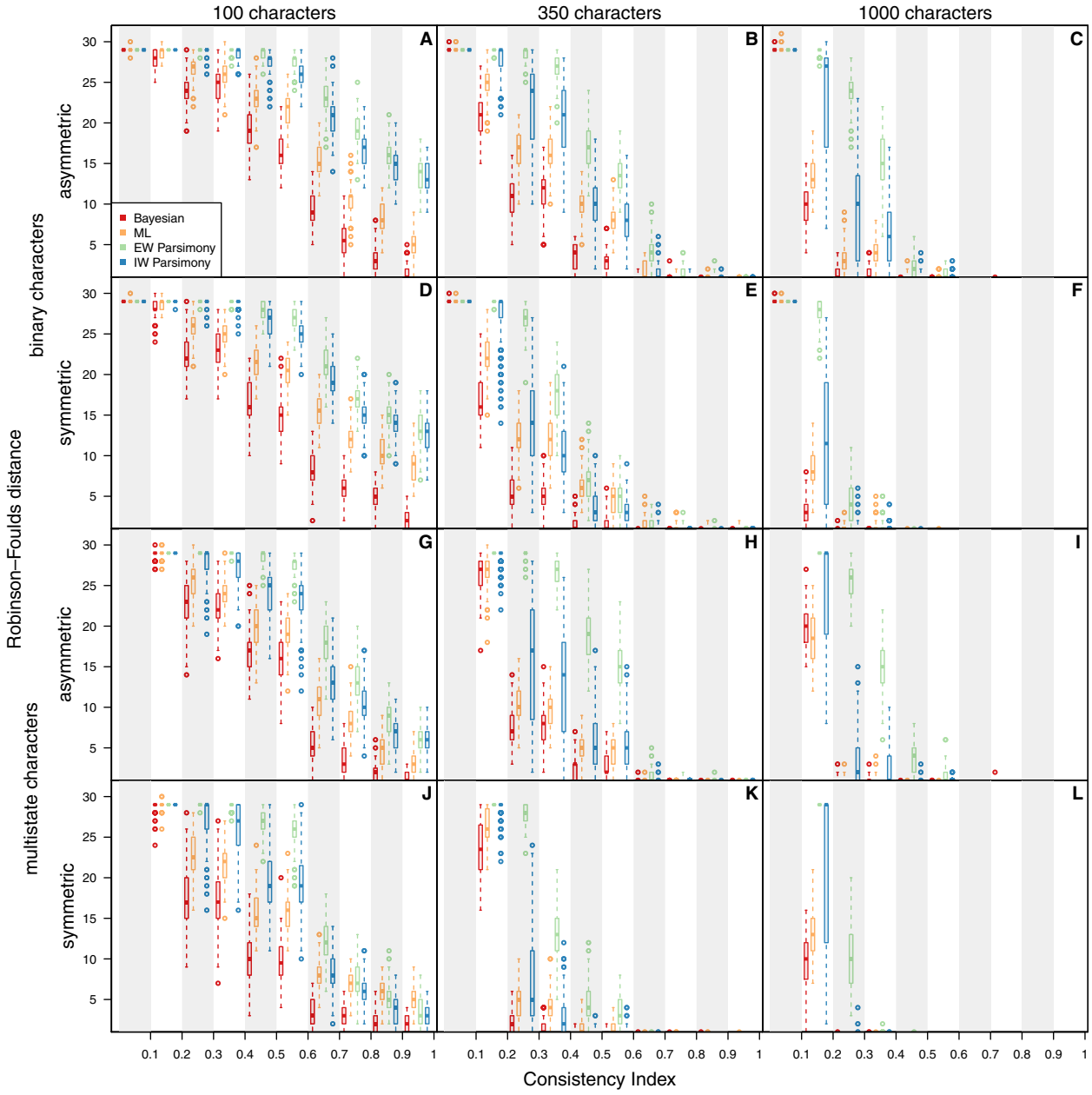


FIG. 4. Robinson–Foulds distances for 95% support trees from Bayesian, equal weights (EW) parsimony, implied weights (IW) parsimony and maximum likelihood (ML) analyses of binary data generated from asymmetric (A–C, G–I) and symmetric (D–F, J–L) trees. All methods converge in the analysis of large datasets of very consistent characters. With small and low consistency datasets, Bayesian exhibits greatest accuracy, followed in order by ML and/or IW parsimony and EW parsimony.

around 0.7, all methods recover the generating tree with little imprecision (Table 2; Fig. 3). At lower CI bins, Bayesian inference outperforms all other methods with the *standard output* trees.

With the 0.5 support trees, the Bayesian Mk model generally has the highest median performance, and the relative performance of implied weights improves considerably with increasing CI values. The accuracy of implied weights is sometimes nonsignificantly higher than other methods in bins with moderate CI values (i.e. ~0.5). Of the other methods, maximum likelihood tends to infer trees that are slightly more accurate than trees inferred under equal weights parsimony (Fig. 3).

When splits with only ≥ 0.95 support are considered, all methods perform equally poorly in the lowest CI bins (0.0–0.1 for binary data, 0.1–0.2 for multistate). In the remaining bins with low CI values (0.1–0.5 binary, 0.2–0.5 multistate), Bayesian inference under the Mk model is

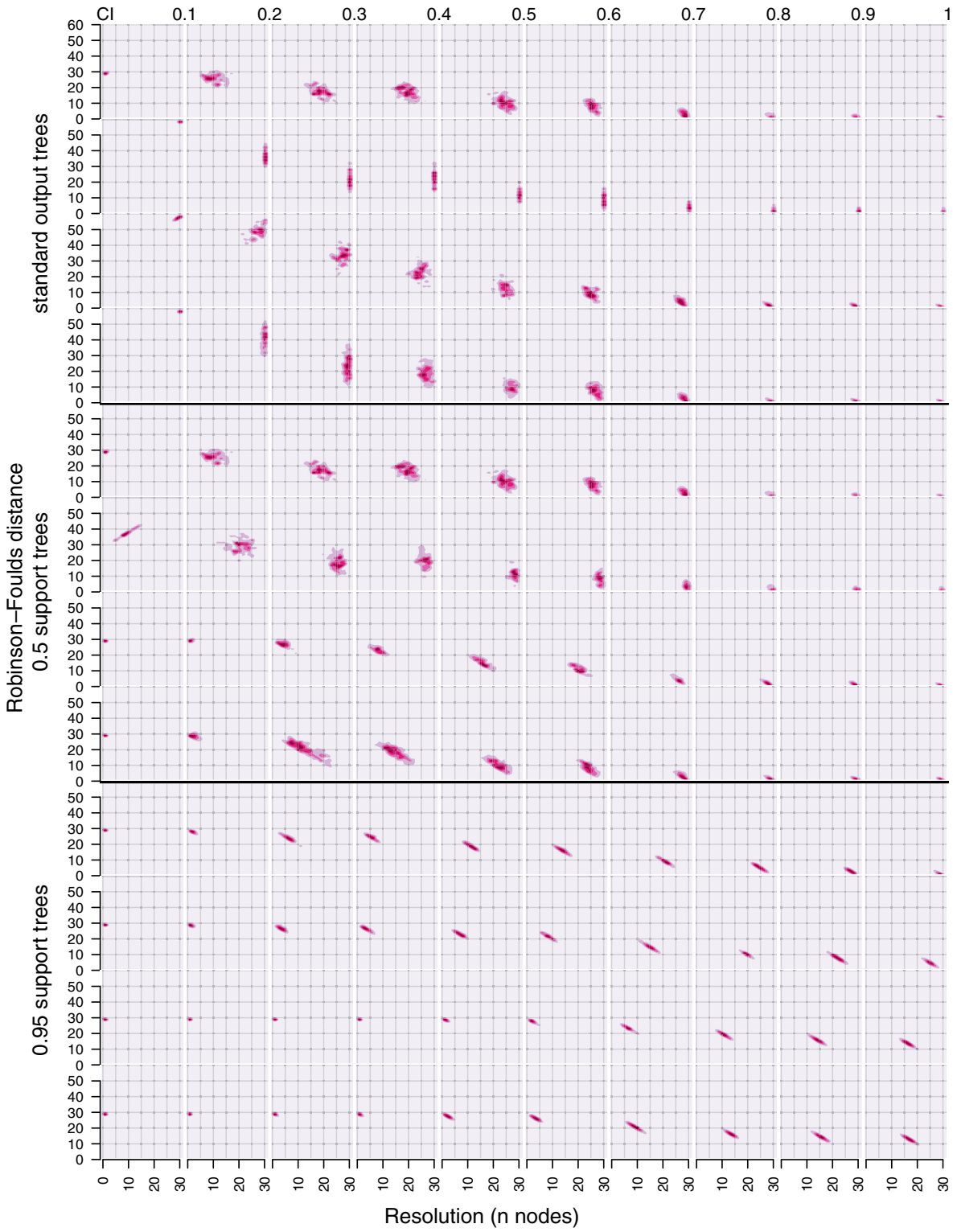


FIG. 5. Density plot showing the resolution and Robinson–Foulds distance for the standard output, 50% majority-rule support trees, and 95% support trees, based on 100 binary character datasets generated from an asymmetrical tree. All methods exhibit accuracy (low Robinson–Foulds distances) and precision (high resolution) in analysing datasets with a high consistency index (CI). Bayesian analysis yields the most accurate standard output trees (A) based on analysis of datasets with a low CI, but achieves this at low precision. This contrast diminishes when considering (B) 50% and (C) 95% support trees, but Bayesian analysis continues to perform best in recovering trees that are more resolved, while maintaining accuracy. Colour online.

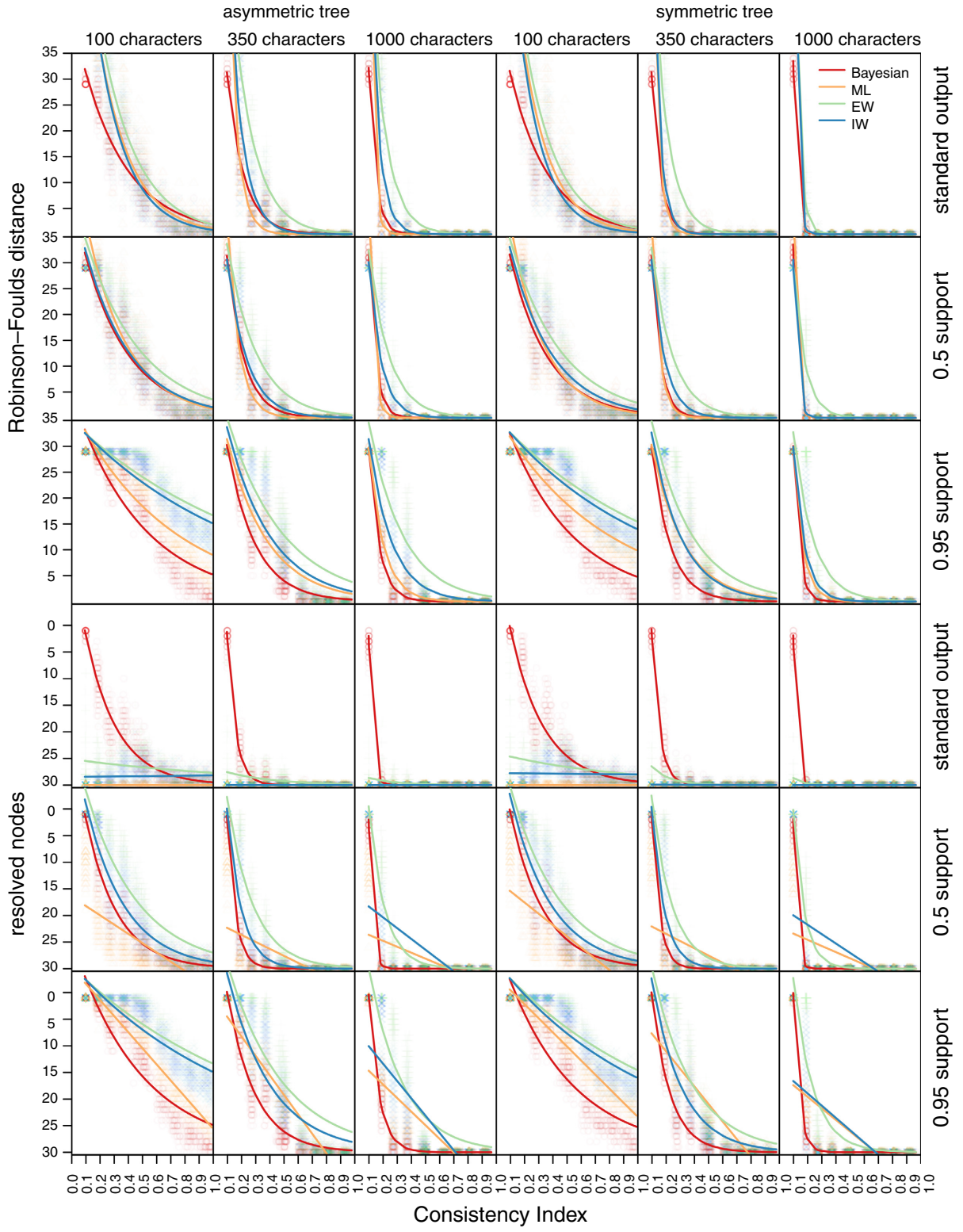


FIG. 6. Exponential model fit to Robinson–Foulds distance and resolution with increasing consistency index (CI) values based on the standard output, 50% and 95% support trees, based on datasets composed of binary characters generated from asymmetric and symmetric trees. The performance of all methods converge with increasing CI and with increased data; however, Bayesian analysis achieves greater accuracy (lower Robinson–Foulds distances) and precision (higher resolution) than the other methods in analysis of datasets with low CI and small numbers of characters.

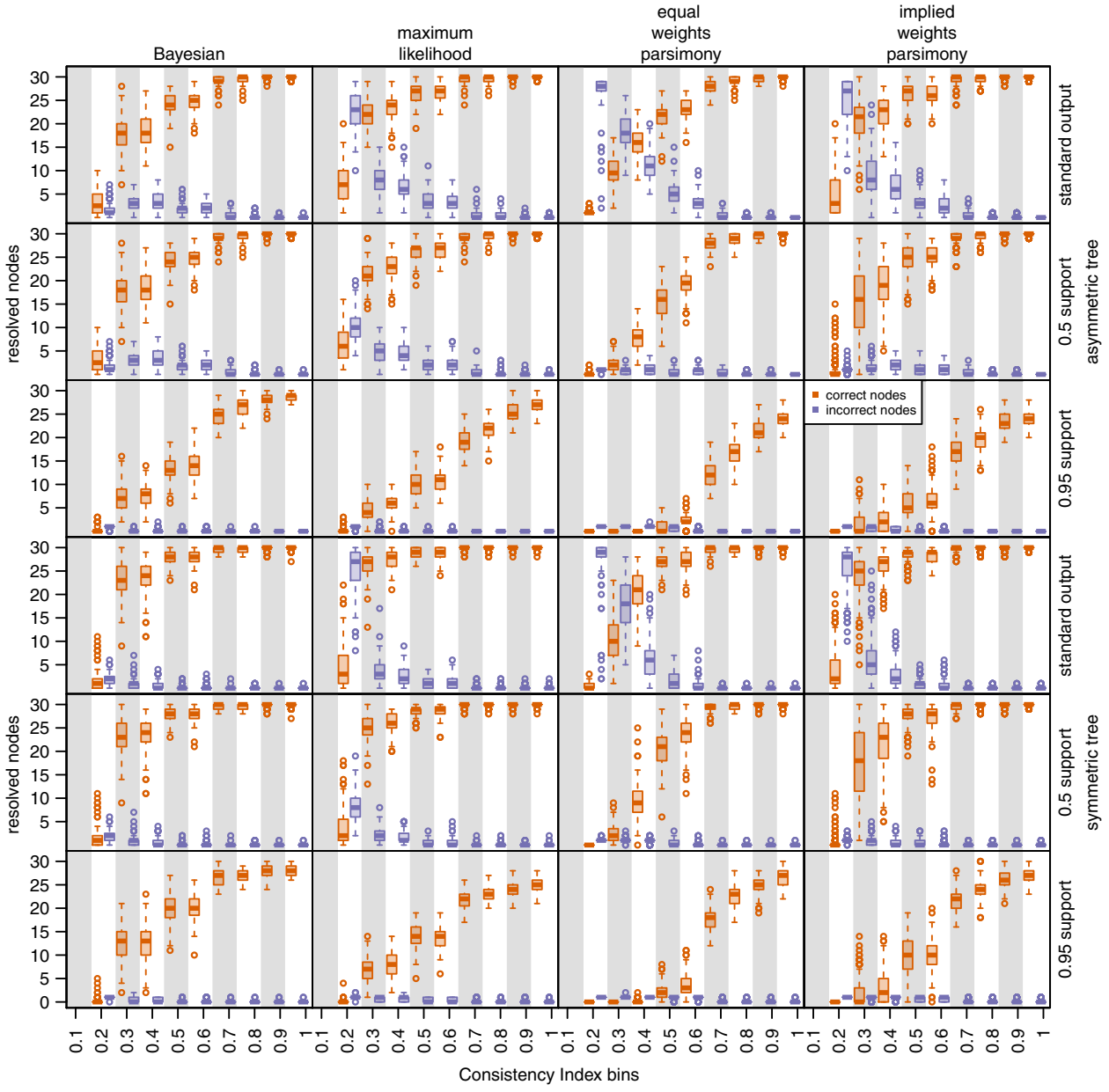


FIG. 7. Number of correct and incorrect nodes recovered using the different phylogenetic methods based on datasets comprised of 100 binary characters. The performance of all methods converges with increasing CI and increased stringency in resolving nodes based on their levels of support. Bayesian analysis consistently recovers the fewest incorrect nodes in analysing datasets with low overall CI, recovering comparable numbers of correct nodes to the other phylogenetic methods.

the most accurate method. Bayesian inference outperforms all other methods when branches with less than 0.95 support are collapsed, with maximum likelihood being the second most accurate method (Fig. 4).

Proportion of accurate and inaccurate nodes

For the standard output trees, Bayesian inference presents the fewest incorrect nodes compared to the other methods (Figs 5, 6). The 0.5 support trees recovered from analysis of datasets from the lowest CI bins show a dramatic decrease in the number of correct and incorrect nodes from maximum likelihood, equal weights and implied weights. Overall, Bayesian inference has a higher median number of correct nodes, and fewer incorrect nodes compared to all other methods (Fig. 7, figs S12–S16). The median number of correct nodes in the 0.95 support trees is higher for Bayesian analysis than for all other methods in analysis of datasets from CI bins 0.0–

TABLE 1. Median and range of the number of resolved nodes for all methods based on the 100 binary character dataset.

	CI Bin	Asymmetric tree				Symmetric tree			
		Bayesian	ML	EW	IW	Bayesian	ML	EW	IW
Standard output	0.1	1 (1–4)	30 (30–30)	29 (3–30)	30 (29–30)	1 (1–4)	30 (30–30)	29 (5–30)	30 (29–30)
	0.2	11 (3–19)	30 (30–30)	27 (18–30)	30 (24–30)	7 (2–17)	30 (30–30)	28 (8–30)	29 (25–30)
	0.3	19 (12–27)	30 (30–30)	27 (18–30)	29 (25–30)	19 (8–25)	30 (30–30)	26 (17–30)	28 (23–30)
	0.4	20 (10–27)	30 (30–30)	25 (18–30)	26.5 (22–30)	17 (10–26)	30 (30–30)	23 (18–29)	25 (20–29)
	0.5	24 (15–29)	30 (30–30)	25 (15–30)	27 (20–30)	24 (15–30)	30 (30–30)	24 (20–30)	26 (20–30)
	0.6	25 (19–30)	30 (30–30)	25 (21–29)	27 (22–30)	25 (21–30)	30 (30–30)	25 (20–30)	26 (22–30)
	0.7	28 (25–30)	30 (30–30)	27 (24–30)	28 (24–30)	28 (25–30)	30 (30–30)	28 (25–30)	28 (25–30)
	0.8	29 (25–30)	30 (30–30)	28 (25–30)	29 (25–30)	29 (26–30)	30 (30–30)	28 (25–30)	29 (26–30)
	0.9	29 (27–30)	30 (30–30)	29 (25–30)	29 (26–30)	29 (26–30)	30 (30–30)	29 (25–30)	29 (25–30)
	1.0	29 (26–30)	30 (30–30)	29 (26–30)	29 (26–30)	29 (27–30)	30 (30–30)	29 (27–30)	29 (27–30)
0.5 support	0.1	1 (1–4)	9 (2–19)	1 (1–3)	1 (1–3)	1 (1–4)	8 (1–18)	1 (1–2)	1 (1–2)
	0.2	11 (3–19)	20 (11–26)	1 (1–6)	3 (1–18)	7 (2–17)	13 (6–24)	1 (1–4)	2 (1–14)
	0.3	19 (12–27)	25 (19–29)	5 (2–10)	13 (3–29)	19 (8–25)	24.5 (15–29)	4 (1–9)	10 (1–26)
	0.4	20 (10–27)	26 (21–29)	8 (3–14)	15 (7–26)	17 (10–26)	24 (14–29)	7 (2–14)	12 (4–25)
	0.5	24 (15–29)	28 (22–30)	16 (8–24)	22 (11–29)	24 (15–30)	27 (19–30)	17 (7–26)	23 (9–29)
	0.6	25 (19–30)	28 (24–30)	20 (14–26)	24 (18–29)	25 (21–30)	28 (24–30)	21 (14–28)	24 (16–29)
	0.7	28 (25–30)	29 (26–30)	26 (21–30)	27 (24–30)	28 (25–30)	29 (26–30)	27 (21–30)	28 (24–30)
	0.8	29 (25–30)	30 (26–30)	28 (22–30)	29 (25–30)	29 (26–30)	29 (26–30)	28 (24–30)	29 (25–30)
	0.9	29 (27–30)	30 (28–30)	28 (25–30)	29 (26–30)	29 (26–30)	29 (26–30)	29 (25–30)	29 (25–30)
	1.0	29 (26–30)	30 (26–30)	29 (26–30)	29 (26–30)	29 (27–30)	29 (28–30)	29 (27–30)	29 (27–30)
0.95 support	0.1	1 (1–1)	1 (1–2)	1 (1–1)	1 (1–1)	1 (1–1)	1 (1–2)	1 (1–1)	1 (1–1)
	0.2	2 (1–5)	1 (1–7)	1 (1–1)	1 (1–1)	2 (1–6)	1 (1–4)	1 (1–1)	1 (1–2)
	0.3	6 (2–11)	4 (1–9)	1 (1–2)	1 (1–4)	8 (1–13)	4 (1–9)	1 (1–2)	1 (1–4)
	0.4	6 (2–11)	4 (2–10)	1 (1–3)	1 (1–4)	7 (2–13)	5 (2–10)	1 (1–2)	1 (1–5)
	0.5	11 (5–17)	7 (2–13)	1 (1–4)	2 (1–8)	14 (8–20)	9 (4–13)	2 (1–5)	3 (1–9)
	0.6	14 (8–18)	8 (4–13)	2 (1–6)	4 (1–8)	15 (8–21)	9.5 (6–15)	3 (1–7)	5 (1–10)
	0.7	21 (16–25)	15 (10–19)	7 (2–13)	9 (2–16)	22 (17–28)	14.5 (10–19)	9 (3–14)	11 (5–16)
	0.8	25 (20–30)	19 (14–25)	11 (5–17)	13 (8–18)	24 (20–28)	18 (13–22)	13 (8–17)	15 (10–20)
	0.9	27 (22–30)	22 (18–26)	14 (9–18)	15 (10–20)	25 (22–29)	20 (15–24)	15 (10–20)	16 (11–21)
	1.0	29 (25–30)	25 (21–29)	16 (12–21)	17 (13–21)	28 (25–30)	21 (18–25)	17 (12–23)	17 (12–23)

Different levels of resolution are achieved when different support values are used to collapse branches.

0.5 (Fig. 7); all other methods apart from maximum likelihood tend to recover an unresolved, star-tree.

Location of correct nodes

For the asymmetrical tree, there is no correlation between the length of descendent terminal branches and the ability of different methods to resolve nodes correctly in datasets from any CI bin (Fig. 8). For datasets comprised of binary characters, only two datasets out of 120 show a significant correlation between the distance separating the node from its descendent tip and the ability to accurately reconstruct a tip value (Spearman's rank) for all methods. Of the multistate datasets, 22% show a significant correlation between the distance of a node from the tips and accuracy.

For the symmetrical generating tree, all methods demonstrate a greater ability to resolve nodes separating the two largest clades (Fig. 8). The trend is only significantly different for Bayesian inference and maximum likelihood, but this is probably due to the smaller sample size in the parsimony analyses.

DISCUSSION

The performance of competing phylogenetic methods is very similar when analysing cladistic matrices that exhibit a high proportion of consistent characters. If the matrix-wide CI value for a dataset is ~0.5 and above, all methods tend to estimate the correct topology with minimal error (Table 1; Figs 3, 5, 6). Thus, it could be argued that any method could be applied to morphological data matrices to recover the true tree when data are of high quality and have been generated by random state assignment followed by screening on their numbers of steps (Goloboff &

TABLE 2. Median and range of Robinson–Foulds distances for all methods based on the 100 binary character dataset.

	CI Bin	Asymmetric tree				Symmetric tree			
		Bayesian	ML	EW	IW	Bayesian	ML	EW	IW
Standard output	0.1	29 (29–32)	58 (56–58)	57 (31–58)	58 (54–58)	29 (29–32)	58 (56–58)	57 (33–58)	58 (54–58)
	0.2	26 (19–32)	36 (24–56)	48 (38–56)	35 (19–58)	26 (17–33)	44 (14–56)	51 (32–58)	35 (15–56)
	0.3	18 (8–28)	22 (8–34)	33 (13–44)	20 (5–37)	14 (6–27)	16 (4–44)	30 (13–49)	13 (4–40)
	0.4	17 (5–24)	22 (2–36)	23 (11–33)	17 (1–28)	15 (6–25)	18 (4–42)	19 (5–33)	12 (4–28)
	0.5	11 (3–20)	12 (4–26)	13 (4–28)	10 (2–21)	8 (1–17)	10 (0–26)	9 (1–22)	7 (1–20)
	0.6	9 (2–16)	10 (2–20)	10 (2–18)	8 (1–17)	5.5 (1–14)	6 (0–20)	6 (1–13)	5 (1–14)
	0.7	4 (0–10)	4 (0–12)	5 (0–10)	3 (0–10)	2 (0–7)	2 (0–10)	2.5 (0–8)	2 (0–6)
	0.8	2 (0–7)	2 (0–10)	2 (0–7)	1 (0–7)	2 (0–6)	2 (0–8)	2 (0–7)	1 (0–6)
	0.9	1 (0–7)	2 (0–10)	2 (0–7)	1 (0–6)	1 (0–6)	2 (0–8)	1 (0–5)	1 (0–5)
	1.0	1 (0–5)	2 (0–6)	1 (0–4)	1 (0–4)	1 (0–5)	1 (0–6)	1 (0–3)	1 (0–3)
0.5 support	0.1	29 (29–32)	37 (30–47)	29 (29–31)	29 (29–31)	29 (29–32)	36 (29–46)	29 (29–30)	29 (29–30)
	0.2	26 (19–32)	30 (20–42)	29 (27–32)	28 (16–33)	26 (17–33)	31 (14–40)	29 (27–30)	28 (16–30)
	0.3	18 (8–28)	19 (7–29)	27 (21–31)	20 (5–29)	14 (6–27)	14 (6–31)	27 (22–30)	21 (5–29)
	0.4	17 (5–24)	20 (3–32)	23 (16–31)	18 (4–26)	15 (6–25)	15 (5–31)	23.5 (17–30)	18 (9–27)
	0.5	11 (3–20)	11 (4–24)	15 (7–26)	11 (3–24)	8 (1–17)	8 (2–18)	13 (4–23)	8 (1–21)
	0.6	9 (2–16)	9 (2–18)	11 (4–19)	8.5 (1–16)	5.5 (1–14)	6 (1–18)	9.5 (3–16)	6 (1–14)
	0.7	4 (0–10)	4 (0–11)	5 (0–10)	3 (0–10)	2 (0–7)	3 (0–8)	3 (0–11)	2 (0–7)
	0.8	2 (0–7)	2 (0–8)	3 (0–8)	2 (0–7)	2 (0–6)	2 (0–6)	2 (0–6)	1 (0–6)
	0.9	1 (0–7)	1.5 (0–9)	2 (0–7)	1 (0–6)	1 (0–6)	1 (0–7)	1 (0–5)	1 (0–5)
	1.0	1 (0–5)	1 (0–6)	1 (0–4)	1 (0–4)	1 (0–5)	1 (0–5)	1 (0–3)	1 (0–3)
0.95 support	0.1	29 (29–29)	29 (28–30)	29 (29–29)	29 (29–29)	29 (29–29)	29 (29–30)	29 (29–29)	29 (29–29)
	0.2	28 (25–29)	29 (27–30)	29 (29–29)	29 (29–29)	28 (24–30)	29 (27–30)	29 (29–29)	29 (28–29)
	0.3	24 (19–29)	27 (22–29)	29 (28–29)	29 (26–29)	22 (17–29)	26 (21–29)	29 (28–29)	29 (26–29)
	0.4	25 (19–29)	26 (21–30)	29 (27–29)	29 (26–29)	23 (17–28)	25 (20–28)	29 (28–29)	29 (25–29)
	0.5	19 (13–26)	23 (17–28)	29 (26–29)	28 (22–29)	16 (10–22)	21.5 (17–26)	28 (25–29)	27 (21–29)
	0.6	16 (12–22)	22 (17–26)	28 (24–29)	26 (22–29)	15 (9–22)	20.5 (15–24)	27 (23–29)	25 (20–29)
	0.7	9 (5–14)	15 (11–20)	23 (17–28)	21 (14–28)	8 (2–13)	15.5 (11–20)	21 (16–27)	19 (14–25)
	0.8	5.5 (0–11)	11 (5–16)	19 (13–25)	17 (12–22)	6 (2–10)	12 (8–17)	17 (13–22)	15 (10–20)
	0.9	3 (0–8)	8 (4–12)	16 (12–21)	15 (10–20)	5 (1–8)	10 (6–15)	15 (10–20)	14 (9–19)
	1.0	1 (0–5)	5 (1–9)	14 (9–18)	13 (9–17)	2 (0–5)	9 (5–14)	13 (7–18)	13 (7–18)

Different levels of resolution are achieved when different support values are used to collapse branches.

Wilkinson 2018). However, when there are high levels of homoplasy in the data, the relative performance of competing phylogenetic methods varies (Table 2; Fig. 3). Unfortunately, the proportion of homoplastic characters in a dataset can only be evaluated with reference to a phylogenetic hypothesis and, for empirical datasets, there can be no knowledge of the levels of homoplasy before estimating a phylogeny. Thus, it is the relative performance of phylogenetic methods in analysis of datasets dominated by homoplasy that is most informative in designing phylogenetic analyses of empirical datasets. Bayesian inference with the Mk model achieves the highest accuracy in analyses of datasets exhibiting the highest levels of homoplasy (Table 2; Fig. 3). However, multistate datasets from CI bins 0.3–0.6 (100 and 350 characters) are an exception; for these, implied weights parsimony has a median value of 2 units of Robinson–Foulds distance lower than Bayesian inference (Puttick *et al.* 2018, tables S6, S8). However, the only examples in which there is no overlap in accuracy between different methods are the analyses in which Bayesian inference is the best-performing method (Table 2). In analyses of datasets comprised of more consistent characters, there is always substantial overlap between the performance of methods, with low variation between all methods.

Tree accuracy

Binary data. The standard output and 0.5 support trees show greatest discrepancy between phylogenetic methods for datasets with a CI between 0.0 and 0.4 (Table 2; Figs 3, 6). In the first bin, CI 0.0–0.1, all methods perform poorly. However, the Bayesian implementation of the Mk model is much more accurate than the next-best method, implied weights parsimony (Fig. 3). The median

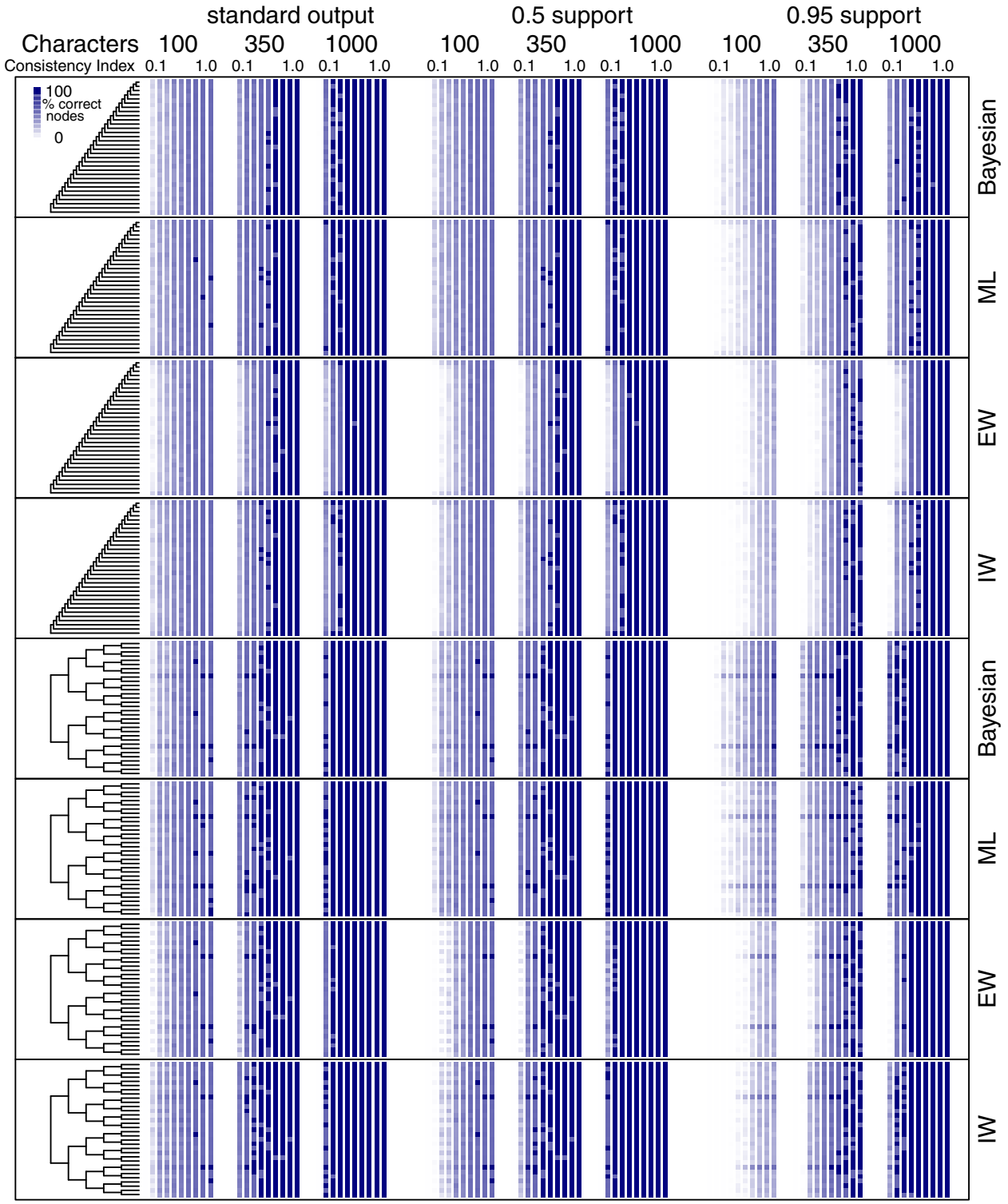


FIG. 8. Location of correct nodes resolved on each tree for all methods using the binary dataset. None of the phylogenetic methods exhibit obvious trends in the relationship between node accuracy and topology in analysis of data generated from the asymmetric tree; all methods show the same trend in topological accuracy in analysis of data generated from the symmetric tree. Colour online.

Robinson–Foulds distance for Bayesian inference is 25 units lower than implied weights parsimony for both the symmetrical and the asymmetrical topologies. These results are also reflected in the difference between datasets from the upper and lower range of CI values, as Bayesian inference exhibits a smaller upper range. Before support is taken into account, equal weights parsimony and maximum likelihood perform poorly in comparison with other methods, with equal weights parsimony generally being the most inaccurate method (O'Reilly *et al.* 2017; Puttick *et al.* 2017*a*).

At this extremely high level of homoplasy (CI 0.0–0.1), Bayesian inference outperforms other methods as it recovers trees that lack resolution and, thus, achieves accuracy when competing methods resolve only incorrect nodes (Table 1; Fig. 7). In the next two bins (CI 0.1–0.3), the differences between methods decrease, but Bayesian inference has the lower median and upper-range Robinson–Foulds values. In CI bins above 0.5, the variance between methods is generally only 1 Robinson–Foulds unit and implied weights parsimony generally outperforms Bayesian inference, albeit marginally (Table 2). In a comparison to Bayesian inference, implied weights parsimony achieves a lower value for the upper range of Robinson–Foulds distances in only 12 out of 60 combinations of analysis.

Multistate data. The trends exhibited by analyses of binary character datasets are similar, but not identical to results of analyses of the multistate character datasets. Bayesian inference is still the best-performing method when there is large variation in the data (Robinson–Foulds distances > 3). Unlike the binary character datasets, implied weights parsimony is the best-performing method in CI bins 0.3–0.6. Yet in some comparable bins, the upper range of Robinson–Foulds values is equal or superior for Bayesian inference compared to implied weights.

Impact of support

The general trends in the relative performance of methods at low CI values continue when nodes with less than 0.5 support are collapsed. However, differences in the performance of Bayesian inference and competing phylogenetic methods are diminished relative to the standard output trees (Brown *et al.* 2017; O'Reilly *et al.* 2017). The largest increase in accuracy achieved by incorporating this measure of support is seen in trees recovered by implied weights parsimony (Fig. 3). However, Bayesian inference is still the most accurate method in analysis of datasets with low CI.

When only nodes with the highest (≥ 0.95) support are considered, Bayesian inference outperforms all methods (Fig. 7). When only nodes with high levels of support are considered (Figs 4–8), all methods, bar Bayesian inference, exhibit relatively high levels of inaccuracy. At ≥ 0.95 support, only Bayesian inference is able to achieve both high levels of accuracy and precision. Thus, if the goal of researchers is to achieve high accuracy with confidence, this ≥ 0.95 support threshold should be applied to avoid the inclusion of erroneous clades. These results contradict previous findings, that Bayesian methods achieved higher accuracy at the expense of low precision, even when accuracy and precision are measured in the same way (O'Reilly *et al.* 2016, 2017; Puttick *et al.* 2017*a*).

Simulation procedure

Previous analyses comparing the relative efficacy of different methods have used continuous time Markov chain models of evolution in which branch lengths are shared across characters in models similar to those used in molecular analyses (Wright & Hillis 2014; O'Reilly *et al.* 2016, 2017; Puttick *et al.* 2017*a*, *b*; Brown *et al.* 2017). This approach has been criticized for its potential to generate datasets that are biased towards model-based approaches where changes are proportional to branch lengths, and long-branch attraction is not a known problem (Felsenstein 1978; Siddall 1998; Philippe *et al.* 2005). In this vein, Goloboff *et al.* (2017) simulated data using a model in which there is no assumption of shared branch lengths amongst characters, concluding that implied weights parsimony is the most accurate phylogenetic method for the analysis of categorical morphological data. Responding to work by O'Reilly and colleagues (O'Reilly *et al.* 2016, 2017; Puttick *et al.* 2017*a*), Goloboff *et al.* (2017) attempted to account for the empirical realism of their simulated matrices in a different manner, not by screening simulated matrices for realism, but by incorporating empirical characteristics of homoplasy into the simulation procedure itself. Goloboff *et al.* (2017) pooled all characters from 158 empirical datasets into one large homoplasy distribution that appeared to be approximately exponentially distributed; the rate parameter of an exponential fitted to this distribution was then used to guide their simulation procedure. This procedure produced an overall distribution of homoplasy that resembled the empirical survey. However, data sets simulated using their code exhibit a per-character homoplasy distribution within datasets that have a proportionally greater number of consistent characters than do empirical datasets (O'Reilly *et al.* 2018). Thus, in the light of our results, the superior accuracy of implied weights parsimony in analysis of the datasets simulated by Goloboff *et al.* (2017) is equally unsurprising and unrealistic, as the per-matrix proportion of highly consistent characters used by Goloboff *et al.* (2017) falls in a narrow

simulation area in which all methods do well, but implied weights performs best (Table 2; Figs 3, 5, 6).

There is substantial overlap in the performance of all methods in analysis of the datasets in which implied weights parsimony performs best (Goloboff *et al.* 2017). Indeed, all methods perform extremely well on datasets with a high proportion of consistent characters; at lower levels of consistency, the Bayesian implementation of the Mk model tends to outperform the other phylogenetic methods. Implied weights parsimony performs well when there are large numbers of consistent characters because it can up-weight the large number of consistent characters, and down-weight the small number of inconsistent characters in datasets. However, implied weights parsimony is not generally able to correctly assign weights to consistent characters when homoplasy is high (CI bins 0.0–0.4), suggesting that in these circumstances implied weights gives higher weight to homoplastic characters, rather than consistent characters (Congreve & Lamsdell 2016).

Our simulation procedure addresses fully the concerns and criticisms raised by Goloboff *et al.* (2017, 2018). There is no reliance on Markov models used for phylogenetic inference, and there is no expectation of shared branch lengths between characters. Encouragingly, these assumptions of the simulation procedure are evident in the simulated data. For example, there is evidence that all nodes are equally likely to be resolved (Fig. 8). Furthermore, the entire range of possible CI values is simulated in our datasets and we differentiate the performance of the competing phylogenetic methods in analysis of datasets exhibiting different overall CI. Given that our simulation procedure did not allow for unobserved character changes, it is perhaps surprising that the model-based phylogenetic methods performed so well relative to parsimony methods.

As we have shown here, the method of simulation is perhaps less significant than the empirical realism of the data simulated (O'Reilly *et al.* 2018). Our simulation procedure could not be accused of faithfully reflecting the process of morphological evolution; it was formulated to complement previous simulation studies (Wright & Hillis 2014; O'Reilly *et al.* 2016, 2017; Puttick *et al.* 2017*a*). These also demonstrated the superiority of model-based phylogenetic methods in analysing morphology-like data but, while many of them were designed to violate assumptions underlying the Mk model, it could be argued that they biased against parsimony-based phylogenetic methods (Goloboff *et al.* 2017, 2018). If anything, by violating assumptions of model-based methods, the simulations should favour parsimony over other methods. The results of all of these simulation-based tests demonstrate that when datasets are large and/or comprised of principally consistent characters, competing phylogenetic methods recover similar estimates. However, when there are large differences between the estimates from competing phylogenetic methods, Bayesian inference generally recovers the most accurate estimate. Hence, we conclude that the Bayesian implementation of the Mk model should be preferred for the phylogenetic analysis of categorical morphological data.

Acknowledgements. We thank the Palaeontological Association for inviting this submission. We also thank April Wright, Graeme Lloyd, Thomas Guillerme and members of the Bristol Palaeobiology research group for discussion, and Pablo Goloboff and an anonymous reviewer for comments that improved this manuscript. We acknowledge the Willi Hennig Society for the freely available version of TNT. Finally, we acknowledge funding from the Royal Commission for the Exhibition of 1851 to MNP, Biotechnology & Biological Sciences Research Council (BB/N000919/1) to PCJD, and the Natural Environment Research Council (NE/P013678/1; NE/N002067/1) to PCJD and DP.

DATA ARCHIVING STATEMENT

Data and supplementary information for this study are available in the Dryad digital depository: https://doi.org/10.5061/dryad.h8r2629

Editor. Imran Rahman

REFERENCES

BROWN, J. W., PARINS-FUKUCHI, C., STULL, G. W., VARGAS, O. M. and SMITH, S. 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*, **284**, 20170986.

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

CUNNINGHAM, C. W., ZHU, H. and HILLIS, D. M. 1998. Best-fit maximum-likelihood models for phylogenetic inference: empirical tests with known phylogenies. *Evolution*, **52**, 978–987.

FELSENSTEIN, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology*, **27**, 401–410.

—— 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, **39**, 783–791.

GOLOBOFF, P. A. 1997. Self-weighted optimization: tree searches and character state reconstructions under implied transformation costs. *Cladistics*, **13**, 225–245.

—— and CATALONA, S. A. 2016. TNT version 1.5, including a full implementation of phylogenetic morphometrics. *Cladistics*, **32**, 221–238.

—— and WILKINSON, M. 2018. On defining a unique phylogenetic tree with homoplastic characters. *Molecular Phylogenetics & Evolution*, **122**, 95–101.

—— FARRIS, S. and NIXON, K. 2003*a*. TNT (Tree analysis using New Technology). http://www.lillo.org.ar/phylogeny/tnt

—— FARRIS, J. S., KÄLLERSJÖ, M., OXELMAN, B., RAMÍREZ, M. J. and SZUMIK, C. A. 2003*b*. Improvements to resampling measures of group support. *Cladistics*, **19**, 324–332.

—— —— and NIXON K. C. 2008. TNT, a free program for phylogenetic analysis. *Cladistics*, **24**, 774–786.

—— TORRES, A. and ARIAS, J. S. 2017. Weighted parsimony outperforms other methods of phylogenetic inference under models appropriate for morphology. *Cladistics*, **34**, 407–437.

—— —— —— 2018. Parsimony and model-based phylogenetic methods for morphological data: a comment on O'Reilly *et al*. *Palaeontology*, **61**, 625–630.

HILLIS, D. M., BULL, J. J., WHITE, M. E., BADGETT, M. R. and MOLINEUX, I. J. 1992. Experimental phylogenetics: generation of a known phylogeny. *Science*, **255**, 589–592.

KLUGE, A. G. and FARRIS, J. S. 1969. Quantitative phyletics and the evolution of anurans. *Systematic Zoology*, **18**, 1–32.

LEWIS, P. O. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology*, **50**, 913–925.

MINH, B. Q., NGUYEN, M. A. and HAESELER, A. VON 2013. Ultrafast approximation for phylogenetic bootstrap. *Molecular Biology & Evolution*, **30**, 1188–1195.

O'LEARY, M. A., BLOCH, J. I., FLYNN, J. J., GAUDIN, T. J., GIALLOMBARDO, A., GIANNINI, N. P., GOLDBERG, S. L., KRAATZ, B. P., LUO, Z.-X., MENG, J., NI, X., NOVACEK, M. J., PERINI, F. A., RANDALL, Z. S., ROUGIER, G. W., SARGIS, E. J., SILCOX, M. T., SIMMONS, N. B., SPAULDING, M., VELAZCO, P. M., WEKSLER, M., WIBLE, J. R. and CIRRANELLO, A. L. 2013. The placental mammal ancestor and the post–K-Pg radiation of placentals. *Science*, **339**, 662–667.

O'REILLY, J. E., PUTTICK, M. N., PARRY, L. A., TANNER, A. R., TARVER, J. E., FLEMING, J., PISANI, D. and DONOGHUE, P. C. J. 2016. Bayesian methods outperform parsimony but at the expense of precision in the estimation of phylogeny from discrete morphological data. *Biology Letters*, **12**, 20160081.

—— —— PISANI, D. and DONOGHUE, P. C. J. 2017. Probabilistic methods surpass parsimony when assessing clade support in phylogenetic analyses of discrete morphological data. *Palaeontology*, **61**, 105–118.

—— —— —— —— 2018. 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.

PARRY, L. A., EDGECOMBE, G. D., EIBYE-JACOBSEN, D. and VINTHER, J. 2016. The impact of fossil data on annelid phylogeny inferred from discrete morphological characters. *Proceedings of the Royal Society B*, **283**, 20161378.

PHILIPPE, H., DELSUC, F., BRINKMANN, H. and LARTILLOT, N. 2005. Phylogenomics. *Annual Review of Ecology, Evolution, & Systematics*, **36**, 541–562.

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. and DONOGHUE, P. C. 2017*a*. Uncertain-tree: discriminating among competing approaches to the phylogenetic analysis of phenotype data. *Proceedings of the Royal Society B*, **284**, 20162290.

—— —— PISANI, D. and DONOGHUE, P. C. J. 2018. Data from: Probabilistic methods outperform parsimony in the phylogenetic analysis of data simulated without an underlying probabilistic model. *Dryad Digital Repository*. https://doi.org/10.5061/dryad.h8r2629

—— —— OAKLEY, D., TANNER, A. R., FLEMING, J. F., CLARK, J., HOLLOWAY, L., LOZANO-FERNANDEZ, J., PARRY, L. A., TARVER, J. E., PISANI, D. and DONOGHUE, P. C. J. 2017*b*. Parsimony and maximum-likelihood phylogenetic analyses of morphology do not generally integrate uncertainty in inferring evolutionary history: a response to Brown et al. *Proceedings of the Royal Society B*, **284**, 20171636.

ROBINSON, D. F. and FOULDS, L. R. 1981. Comparison of phylogenetic trees. *Mathematical Biosciences*, **53**, 131–147.

RONQUIST, F., TESLENKO, M., MARK, P. VAN DER, AYRES, D. L., DARLING, A., HÖHNA, S., LARGET, B., LIU, L., SUCHARD, M. A. and HUELSENBECK, J. P. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, **61**, 539–542.

SIDDALL, M. E. 1998. Success of parsimony in the four-taxon case: long-branch repulsion by likelihood in the Farris Zone. *Cladistics*, **14**, 209–220.

TUFFLEY, C. and STEEL, M. 1997. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. *Bulletin of Mathematical Biology*, **59**, 581–607.

WRIGHT, A. M. and 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.

—— LLOYD, G. T. and HILLIS, D. M. 2016. Modeling character change heterogeneity in phylogenetic analyses of morphology through the use of priors. *Systematic Biology*, **65**, 602–611.

YANG, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution*, **39**, 306–314.