WILEY-BLACKWELL

Cladistics (2017) 1–31

Cladistics

10.1111/cla.12205

Weighted parsimony outperforms other methods of phylogenetic inference under models appropriate for morphology

Pablo A. Goloboff*, Ambrosio Torres and J. Salvador Arias

Unidad Ejecutora Lillo, Fundación Miguel Lillo, CONICET, Miguel Lillo 251, 4000 San Miguel de Tucumán, Argentina

Accepted 22 March 2017

Abstract

One of the lasting controversies in phylogenetic inference is the degree to which specific evolutionary models should influence the choice of methods. Model-based approaches to phylogenetic inference (likelihood, Bayesian) are defended on the premise that without explicit statistical models there is no science, and parsimony is defended on the grounds that it provides the best rationalization of the data, while refraining from assigning specific probabilities to trees or character-state reconstructions. Authors who favour model-based approaches often focus on the statistical properties of the methods and models themselves, but this is of only limited use in deciding the best method for phylogenetic inference—such decision also requires considering the conditions of evolution that prevail in nature. Another approach is to compare the performance of parsimony and model-based methods in simulations, which traditionally have been used to defend the use of models of evolution for DNA sequences. Some recent papers, however, have promoted the use of model-based approaches to phylogenetic inference for discrete morphological data as well. These papers simulated data under models already known to be unfavourable to parsimony, and modelled morphological evolution as if it evolved just like DNA, with probabilities of change for all characters changing in concert along tree branches. The present paper discusses these issues, showing that under reasonable and less restrictive models of evolution for discrete characters, equally weighted parsimony performs as well or better than model-based methods, and that parsimony under implied weights clearly outperforms all other methods.

Introduction

In recent years, a number of papers have used simulations to examine different methods for phylogenetic analysis of discrete morphological data. Wright and Hillis (2014), with the aim of encouraging palaeontologists to "adopt model-based approaches", used their simulations to conclude that MrBayes (Ronquist et al., 2012) with the so-called "Mk model" of Lewis (2001), produces better trees than parsimony. Wright and Hillis (2014) used a fixed model tree, taken from an earlier empirical study on amphibians, and generated their data with Lewis' model. In another paper, O'Reilly et al. (2016) used the same model tree as Wright and Hillis (2014) but a slightly different evolutionary model to generate their data, comparing Bayesian with parsimony analyses under both equal and implied weighting (Goloboff, 1993). O'Reilly et al. (2016) concluded that Bayesian analysis produces better trees than equal-weights parsimony, and implied weighting performs the worst of all three methods. Congreve and Lamsdell (2016) concurred with the previous authors that the wider use of parsimony instead of model-based methods to analyse palaeontological data is only because of a "legacy issue" (p. 447) instead of appropriateness. However, they provided no comparison between parsimony and model-based methods; they only compared parsimony with equal and implied weights. Like O'Reilly et al. (2016), Congreve and Lamsdell (2016) concluded that parsimony with equal weights is preferable to implied weighting. The last study in the series of simulation papers, by Puttick et al. (2017), is very similar in design to the one by O'Reilly et al. (2016); the authors of Puttick et al. (2017) include all the eight authors in O'Reilly

*Corresponding author.
E-mail address: pablogolo@yahoo.com.ar

et al. (2016) plus another three. The main difference is that Puttick et al. (2017) explore the influence of factors that (in their view) could have biased previous comparisons (i.e. resolution of the model tree, impact of the probabilistic model, tree shape and presence of multistate characters), to conclude that a "consensus shows that the Bayesian Mk model is the most accurate method of phylogenetic reconstruction" for morphological data (p. 8).

In the present paper, we discuss both the use of models based on notions of molecular evolution [such as the models of Neyman (1971) and Felsenstein (1978, 1981), from which Lewis' (2001) Mk model is a derivative] to analyse morphology, and the simulations for which implied weighting produced worse trees than equal-weights analyses. When the data are generated with alternative models (i.e. less restrictive, with branch lengths not fixed, using distributions of homoplasy approaching that observed in empirical data sets) and the trees inferred from each method of analysis are more carefully compared, the advantage of model-based methods over equal-weights parsimony vanishes, and implied weighting outperforms the other methods examined.

Realism

Felsenstein (1978) proposed a model of evolution under which parsimony (then the most widely used criterion for phylogenetic inference) would be statistically inconsistent, and thus "positively" misleading. The main reason why parsimony becomes inconsistent under Felsenstein's (1978, 1981) models is that there is a uniform branch length for all sites (with inconsistency becoming more likely for lower numbers of possible character states, as noted by Farris, 1983, p. 14). Felsenstein's model assumed that all characters evolve at the same rate. Later work on likelihood (e.g. Jin and Nei, 1990; Yang, 1994) incorporated the idea that some characters may evolve at a faster rate than others, but the relative branch lengths still determine the relative probabilities of change along each branch for all characters (or all the characters in a partition, in the case of unlinked models). Thus, if change in a slow-evolving character is more likely on one tree branch than on another, then the same is also true for fast-evolving characters. In other words, the probability of change for all characters, fast and slow, decreases or increases in concert at each branch–in which case, despite the rate-heterogeneity model, parsimony is also prone to inconsistency. But in the specific case of morphology, it is well known that some character systems may well change mostly along some branches, whereas others do so on other branches. For example, in the branches interconnecting some species of insects, thoracic wings evolve more than vertebrae; the opposite is true for the branches interconnecting some species of vertebrates (Farris, 1983 made this exact same point, pp. 14–15, using tooth counts and paired appendages as example). On a given branch of the tree, the evolution of some characters may be sped up while the evolution of others is slowed down; this is not how different rates are normally treated in model-based analyses, which preserve the relative branch lengths across all characters. Evolutionary models with uniform branch lengths, which forbid these differences, may be defensible in the case of molecular sequences, but seem implausible for morphology.[1] This implausibility, more than any "legacy issue" (Congreve and Lamsdell, 2016, p. 447) or "consequence of tradition" (O'Reilly et al., 2016, p. 1), is the main reason why many phylogeneticists still prefer parsimony for analysing morphological data.

The almost universal use of models with common relative branch lengths may create the impression that those are the most natural models, or even the only ones possible. This universality also may have been fostered by the availability of numerous programs (e.g. Seq-Gen, Rambaut and Grassly, 1997; Paml, Yang, 2007; Geiger, Harmon et al., 2008) that facilitate generating data sets under that model, and by most cladists apparently lacking interest in simulations and thus not proposing alternatives.

The studies of Wright and Hillis (2014), O'Reilly et al. (2016) and Puttick et al. (2017; but see below for symmetrical trees) are no different in this regard, because they also assume proportional branch lengths across all simulated characters. These studies generate their data sets using the Mk model and use a tree (taken from Pyron, 2011) with very unequal branch lengths. Thus, their finding that model-based inference outperforms parsimony is hardly shocking—phylogeneticists have known about long-branch attraction for almost 40 years. Implied weighting is conceived to correct for differences in the reliability of different characters, but it is affected by long-branch attraction in exactly the same manner as parsimony under equal weights (if this seems counterintuitive, consider the 4-taxon case, the prime example of inconsistency for parsimony, for which implied and equal weighting always produce the same result). O'Reilly et al. (2016) stated that their simulation would not favour Bayesian inference because they used data violating the Mk

[1] We are aware, of course, that a likelihood model without such restriction could be constructed (e.g. taking ideas from covarion models; Fitch and Markowitz, 1970), but not having been implemented in any of the major programs or used in studies assessing the performance of parsimony, that hardly matters in practical terms. Likewise, that different partitions can have independent branch lengths, as in so-called unlinked models, makes no difference to our present argument.

model (they generated their data using HKY85 instead of the Mk model, and subsequently recoded the data in binary form, as purines/pyrimidines). It is true that the data patterns so produced are not exactly the ones expected under the Mk model, but that hardly means that the recoding will diminish long-branch attraction. The recoding might instead increase long-branch attraction, because it is reducing the number of states (thus making independent parallel derivations more likely), or at least leave long-branch attraction unaffected (much like the case of supersites or "k-tuples"; see Steel and Penny, 2000, p. 845). O'Reilly et al. (2016) are correct that their data slightly violate the model, but there is no real basis for their statement that the resulting data are then not "in favour of either method of phylogenetic inference" (p. 2).

O'Reilly et al. (2016, p. 1) worried that Wright and Hillis's (2014) study "did not consider whether the simulated data exhibited realistic levels of homoplasy", and set out to correct this problem. They filtered their simulated data sets, eliminating those with an ensemble consistency index (CI; Kluge and Farris, 1969) below 0.26, using as reference the values found by Sanderson and Donoghue (1989). Puttick et al. (2017) used exactly the same type of filtering for their simulated matrices. The problem is that the ensemble consistency index is an overall measure of homoplasy. Being an overall measure, the same averages may well be obtained by very different distributions of the homoplasy in the characters. For example, a binary data set with 500 characters with no extra steps and 500 characters with four extra steps produces (on the most-parsimonious tree) a CI = 0.333, exactly the same as another binary data set with, 1000 characters with two extra steps each (and, yes, two data sets like these can exist). Yet the first data set comprises some characters that are much more reliable than others, whereas the reliability of all the characters in the second data set is exactly the same (thus making a method such as implied weighting simply irrelevant). The use of the ensemble CI to determine the realism of the homoplasy distribution is thus overly simplistic.

Congreve and Lamsdell (2016) used a gamma distribution which, in their view, "allows for every character within the matrix to have a variable rate of change, which results in more naturalistic data sets by accounting for the observed patterns of mosaic[2] evolution...and allows for the overall levels of homoplasy within each data set to be highly variable." Unlike O'Reilly et al. (2016), Congreve and Lamsdell (2016) did not use any quantitative comparison to back their claim that the levels of homoplasy in their data sets are realistic—but they are not.

Although the CI may be a crude approximation of the distribution of homoplasy among characters, a much better estimation can be gained by looking at precisely that—the distribution of homoplasy among characters in morphological data sets.

Data sets and methods

We used the 70 data sets of Goloboff et al. (2008a), as well as 88 other empirical data sets (with 50–149 taxa, and 22–1844 characters; average number of taxa is 80.7, average number of characters is 214.3). A few of the data sets are elaborations of others (e.g. the improvement of a previous matrix), but most of the data sets are independent. The details of the 158 morphological data sets are in Appendix 1, and the data sets themselves are included as Supplementary Material (at http://www.lillo.org.ar/phylogeny/published/).

A rough estimation of a most-parsimonious tree was performed (with the ***mult 10 = hold 1*** command of TNT; see Goloboff et al., 2008b) for each of these 158 empirical data sets. A single tree was held at the end of the search. The frequencies of characters with different numbers of extra steps were then calculated and pooled over the 158 data sets (continuous and uninformative characters, if present, were excluded from these counts). The results are shown in Fig. 1a. The curve (black dots) is a very regular one, with an almost perfect fit to an exponential distribution ($R^2 = 0.985$). Note that the exponential distribution is typical of Poisson and Markov processes; the assumption of a Markov process in phylogeny is essentially sound, as lineages are independent after cladogenesis. Standard likelihood models are much more restrictive in that the Markov process is supposed to be homogeneous, making use of branch lengths common to all characters.

The same procedure was repeated for the 100 data sets of Congreve and Lamsdell (2016), and the data sets for 100 characters of O'Reilly et al. (2016) for which CI $\geq$ 0.26 (these are 172 out of the 1000[3]). The shape of the distribution of homoplasy in the data sets of O'Reilly et al. (2016) resembles that observed in empirical data, approaching an exponential distribution, although the proportion of characters with no homoplasy is considerably larger than the observed

[2] Note that using the term "mosaic evolution" to refer to differences in rate of change of the different characters is highly unusual; mosaic evolution normally refers to the fact that, on a given branch of the tree, only some characters change to the derived state, whereas others remain in their plesiomorphic condition, so that taxonomic groups are a "mosaic" of primitive and derived features. This will happen even for completely uniform rates of change, so that there is no logical connection between mosaicism and rates of change.

[3] O'Reilly et al. (2016) report that only 128 of their 1000 data sets had a CI $\geq$ 0.26. They probably measured homoplasy on the model tree, whereas we measured it on a most-parsimonious tree.

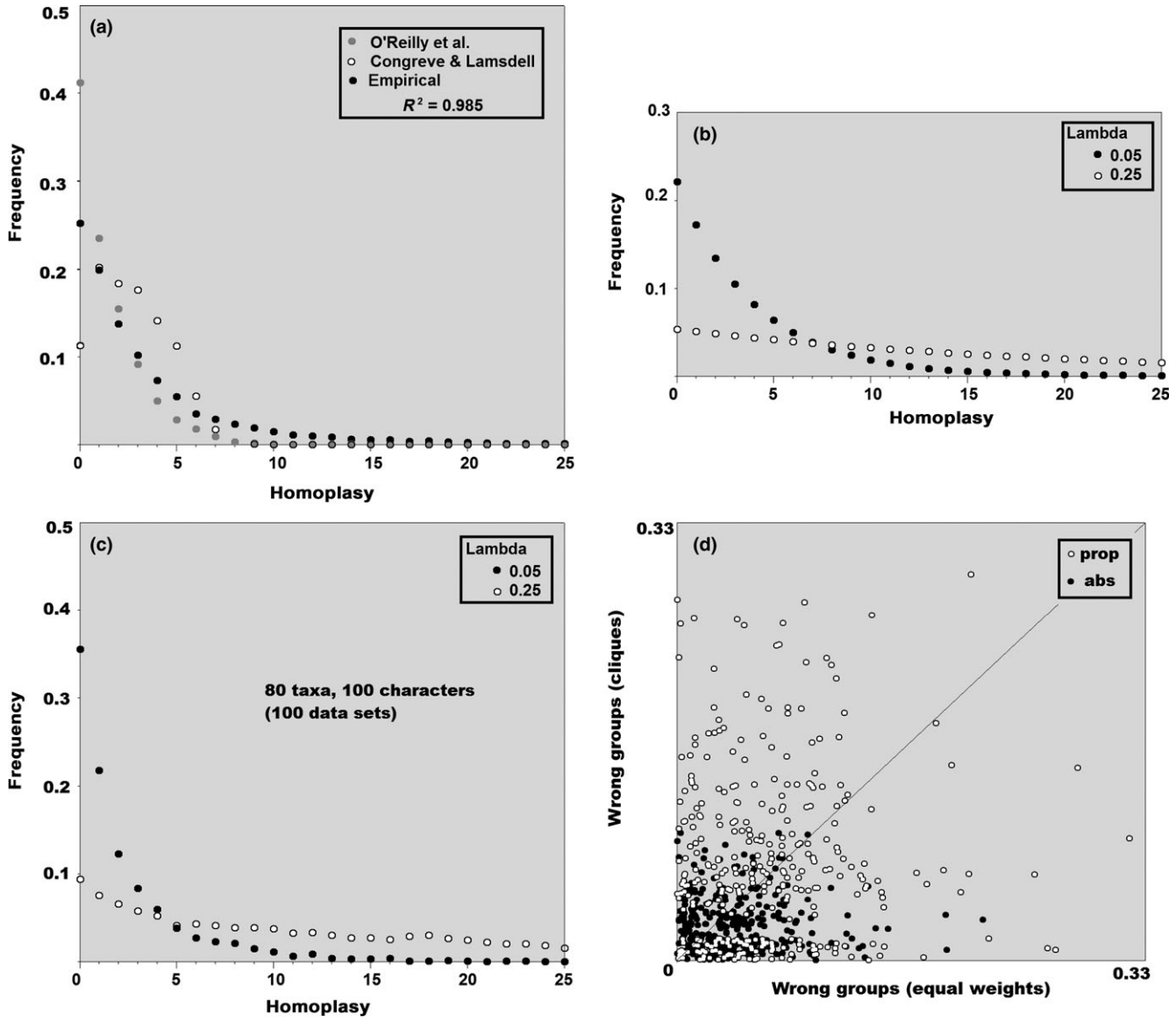

Fig. 1. (a) Homoplasy distribution in different data sets: our 158 empirical data sets (black), O'Reilly et al. (2016) (grey), and Congreve and Lamsdell (2016) (white). (b) Probabilities of numbers of changes, for $\lambda = 0.05$ and $\lambda = 0.25$; probabilities are normalized so that they sum to unity at 50 extra steps. (c) Probabilities of different numbers of extra steps in generated data sets (this is not exactly the same as in (b) because some changes can appear on sister or consecutive branches). (d) Absolute (black) and proportional (white) number of incorrect groups for clique analysis and equal weights.

(0.41212 vs. 0.25556; Fig. 1a, grey dots). Congreve and Lamsdell's data sets (Fig. 1a, white dots) depart more significantly from empirical data sets, because the frequency of characters with no extra steps (0.11255) is much less than the frequency of characters with one (0.20127) to three extra steps (0.18291). Implied weighting works by identifying the more reliable characters, but these are rarer in Congreve and Lamsdell's simulated data sets than in empirical data sets. Because there are basically no reliable characters to be identified in their matrices, Congreve and Lamsdell's (2016) assertion that "in incorporating rate heterogeneity our simulations were designed to favour implied weighting over equal weights parsimony" (p. 457) is unjustified.

Given that the exponential distribution ($\lambda e^{-\lambda x}$) almost perfectly fits the observed pattern of homoplasy and that it is a simpler distribution than the gamma distribution (for which there is no reason other than convenience; see, e.g., Felsenstein, 2004, p. 219), we used the exponential distribution to simulate our data sets. The exponential function provides point values that were converted into cumulative probabilities (so that the probabilities sum to 1 at 50 extra steps).

A different "model"

In addition to using the frequencies observed in real data sets to determine more realistically the probabilities of homoplasy in the simulated data sets, the mode of generation of data used here is less restrictive than the Lewis model in that it does not assume that there is a branch-length parameter common to all characters. To evolve a character with n transformations on the model tree (where the number n is decided with the exponential function), n branches of the tree are chosen at random and marked as points of change. Then, the character always starts as state 0 at the root, changing to a different state (randomly picked from each of three other alternatives) every time a branch marked as a point of change is encountered. We hesitate to call this a "model," because it is very unrestrictive, but it does evolve the characters on a given tree (the "model" tree) according to a fixed set of rules.

Incidentally, this way to evolve data produces changes distributed on the branches of the tree just as in Lewis's (2001) model with the parameters used by Congreve and Lamsdell (2016). These authors used (2016, p. 452) a model tree for which all branches have the same length, and different characters have different rates. Thus, in a given character, a change has the same chances of occurring at any of the branches of the tree, just as in our simulations. The distribution of changes across the tree branches is also ("paradoxically", as noted by Steel, 2011, p. 104) the same one that results if branches are assigned different lengths for each character independently, and the characters are then evolved in the standard way, with suitable distributions of branch lengths and character rates; the models of extreme uniformity and extreme heterogeneity thus converge to similar outcomes. As a consequence, there are significant similarities between our "model" and standard Markov models, and the simulations are not expected to be specially unfavourable to methods based on maximum likelihood—other models, with mixes of different rates (e.g. as in Kolaczkowski and Thornton, 2004) are likely to produce worse results for likelihood.

The performance of parsimony under equal and implied weights (with TNT v.1.5; Goloboff and Catalano, 2016), Bayesian analysis (with MrBayes v.3.2.6, using the default Lewis model, summarizing the results of the Markov chain with the standard procedure of using frequency of groups in the tree sample as the posterior probability) and maximum likelihood (with RAxML; Stamatakis et al., 2005; v.7.7.2, with default settings) was compared. For this, we generated 200 combinations of different numbers of taxa, characters, and values of λ (with the number of taxa randomly chosen between 40 and 80, number of characters one and half times the number of taxa, and λ randomly chosen between 0.05 and 0.25). The resulting probabilities of the different numbers of changes are shown in Fig. 1b. For each of those 200 combinations, we generated 10 data sets, and analysed them with the different methods (the 10 data sets per combination of parameters is intended to reduce dispersion). Thus, a total of 2000 data sets was generated; the model tree was generated at random (all trees equiprobable, with a different random seed) for each of the 2000 data sets. TNT scripts were used to generate the data, denoted MrBayes and RAxML, import the trees produced by those programs and produce all of the comparisons. Parsimony analyses considered all characters as nonadditive, and eliminated unsupported groups with TBR-collapsing. Note that parallel changes in sister branches are not corrected by the script when it generates the data; thus, the resulting homoplasy need not be exactly the same as in the exponential function. The actual homoplasy distribution in the resulting data sets was examined in 100 data sets with 80 taxa and 1000 characters (Fig. 1c). The values of λ used here produced data sets that completely include the distribution of homoplasy observed in the empirical data sets. The scripts and individual results of the simulations are included in the Supplementary Material (at http://www.lillo.org.ar/phylogeny/published/).

Measures of comparison

The comparisons of Wright and Hillis (2014), O'Reilly et al. (2016) and Puttick et al. (2017) were based exclusively on the Robinson–Foulds distances (RF; Robinson and Foulds, 1981). Although there is no reason to think that this measure could be biased in favour of any method of phylogenetic analysis, the measure is not without problems, because just one or a few taxa moving to a faraway location in one of the trees will strongly increase the distance without significantly altering the tree. In the present paper, other measures that are less affected by such floating taxa are used in addition to RF. The second measure used is a modification of the distortion coefficient (DC) of Farris (1973), the complement of $(Ga + Gb - Sab - Sba)/(Ga + Gb - Ma - Mb)$, where Ga is the maximum possible steps of the matrix representing tree a, Sab is the number of steps of the matrix representing tree a when mapped onto tree b, and Ma is the minimum possible number of steps of the matrix representing tree a. By using the reciprocal sums, the DC is made symmetrical (the original formulation is an asymmetric measure; see Goloboff, 2005). This measure is implemented in the TNT command ***tcomp***. The third measure used is the SPR distance (SPRd), but with moves weighted by distance (Goloboff, 2007), so that longer moves, which change the tree more, are more "costly" than shorter ones

(with the ***sprdiff*** command of TNT). Finally, the fourth measure is (the complement of) the average group similarity between the model tree and the inferred tree, as measured in terms of group composition (as in Goloboff et al., 2009, p. 215; Goloboff and Catalano, 2012, p. 509, implemented in the ***tcomp*** command of TNT).

The previous four measures provide an evaluation of the similarity between model and inferred tree. Another aspect that is of interest is the number of groups of the model tree that are recovered by an inference method, which is the number of correct groups retrieved, normalized by the number of taxa minus two (because model trees are binary), and the proportion of groups in the inferred tree that are incorrect (the number of incorrect groups divided by the total number of groups in the inferred consensus tree). Congreve and Lamsdell (2016) used the *absolute* number of incorrect groups, instead of the *proportion*, but (as they noted themselves, p. 457) this will favour unresolved trees. What matters in this context is the probability that a group concluded from an analysis is incorrect, and this measure must consider the ratio between the number of incorrect groups and the total number of groups in the inferred tree. Another reason to not consider the absolute number of wrong groups is that such a measure is in fact improved by cliques, an extreme form of weighting (where characters have weights either unity or zero). Clique analysis (Fig. 1d) recovers a lower absolute number of incorrect groups than equal weights (average 0.01777 instead of 0.03196, black dots), but a larger proportion (0.05468 instead of 0.05005, white dots). We do not consider clique analysis a defensible method for phylogenetic analysis (Farris, 1983), and presumably neither would Congreve and Lamsdell, so we conclude that a decreasing absolute number of groups is not a reliable indication that a method performs worse.

A possible improvement in calculating the proportion of wrong groups would be by taking into account that missing a group because of a single terminal has been misplaced is not equivalent to missing it because no group in the inferred tree resembles the reference group from the model tree. This proportional difference can be calculated as (the complement of) the sum of the similarities of each group in the inferred tree to the closest group in the model tree (the individual maximum for each group is one), divided by the number of nodes in the inferred tree. In this way, all the groups that are made incorrect by a long move of one single terminal are still counted as relatively close to the expected groups. Only those groups for which no similarity is detected are counted as completely wrong. This function was implemented with the following TNT code:

```
ttag =;
rfreq ['inferred']'model';
set f 0;
loop = danod (root + 1) nnodes['inferred']
  set f + = $ttag #danod;
  stop
  set propdiff 1 - ('f'/(tnodes['inferred'] * 100));
ttag -;
```

In the figures comparing methods, these measures are plotted with one method on the *X*-axis and another method on the *Y*-axis, with both axes having the same scale. In this way, the diagonal indicates a perfect tie between the methods in question. For most measures, the points above the diagonal indicate a poorer performance of the method on the *Y*-axis; the only exception is the number of retrieved groups, which would indicate better performance (more correct groups recovered) for the method on the *Y*-axis when the dots are above the diagonal.

Simulation results

Equal-weights parsimony and Bayesian analysis

For three of the measures, the distance between the model and the inferred tree was similar, although there is a lot of dispersion. The average RF between model and inferred tree for equal-weights parsimony (0.09421) was slightly less than that distance for Bayesian analysis (0.09642), as was the average group similarity (0.05332 for equal weights, 0.05391 for Bayesian). The DC indicated a slightly larger difference between the model and the parsimony tree (0.01696) than between the model and the Bayesian tree (0.01510). The SPRd indicated a much larger difference between the model and the Bayesian tree (0.09546, instead of 0.07756 for equal-weights parsimony), but this may be an artifact produced by counting the number of moves needed to resolve the Bayesian tree (which often has polytomies); the algorithm used by TNT to count the number of resolving moves is a rough approximation (and parsimony uses the single best SPRd to one of the equally optimal trees found by the search, not the consensus). The measures of resemblance to the model tree, therefore, do not indicate any clear advantage of using either equal-weights parsimony or Bayesian analysis. The plot for the individual values is shown in Fig. 2a.

The proportions of correct and incorrect groups found by each method do not provide a clear choice either. Bayesian analysis finds a higher proportion of wrong groups, but also recovers more correct groups than equal-weights parsimony, and these seem in principle to balance with each other. Figure 2b shows the values on a common 0–1 scale in order to make the

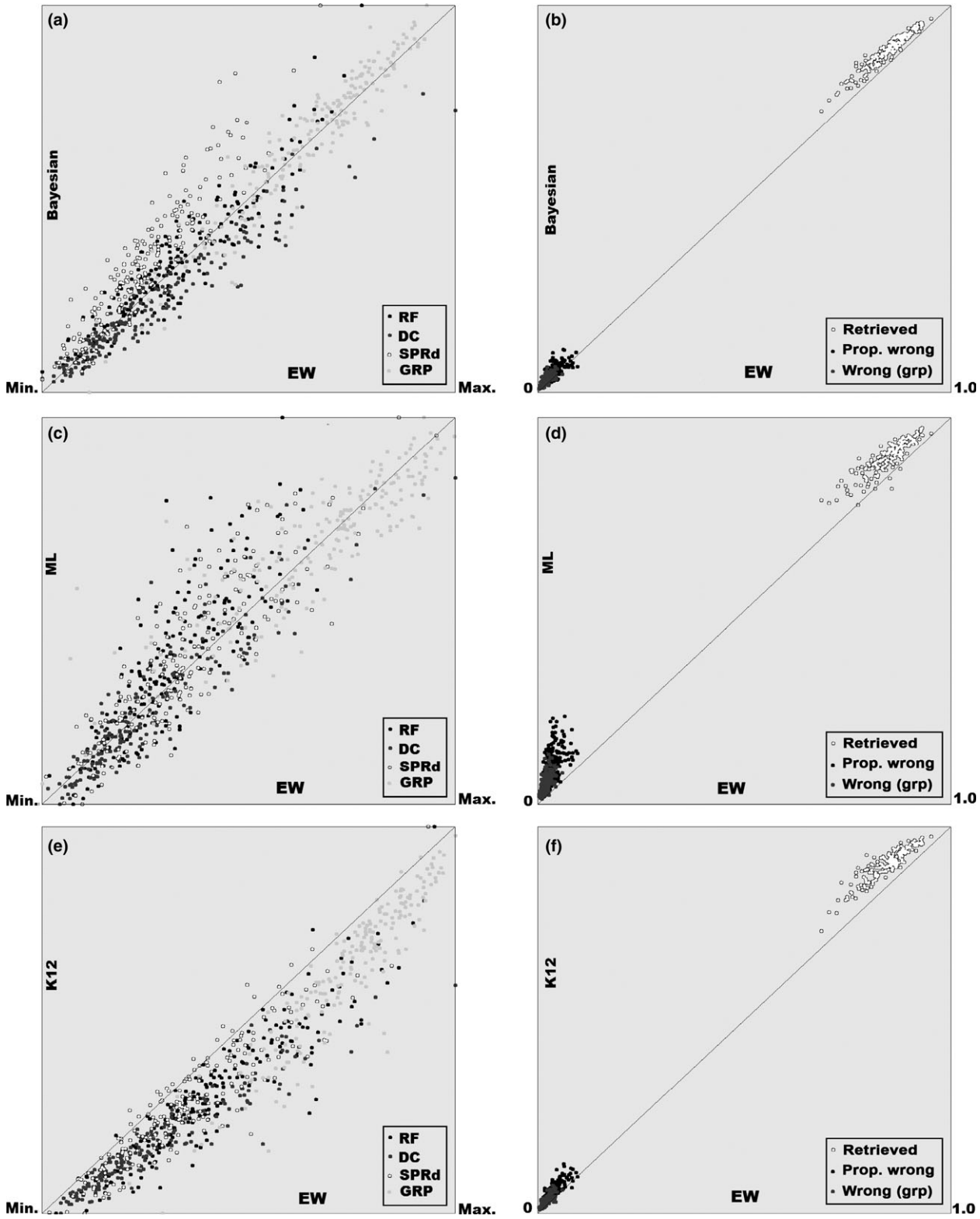

Fig. 2. (a) Distances between model and inferred tree, for Bayesian inference. (b) Retrieved and incorrect groups for Bayesian analysis. (c) Distances between model and inferred tree for maximum likelihood (ML). (d) Retrieved and incorrect groups for ML. (e) Distances between model and inferred tree, for implied weights. (f) Retrieved and incorrect groups for implied weights. In all cases, values plotted against equal weights (EW). Data sets generated with an exponential distribution of homoplasy (as explained in the text), without missing entries.

relative differences between incorrect and retrieved groups more evident. An exact comparison is difficult because the number of incorrect groups is normalized by the number of nodes in the inferred tree, which may vary among the different data sets. See below for alternative comparisons.

Equal-weights parsimony and maximum likelihood

The comparison of equal weights and maximum likelihood (Fig. 2c,d) is somewhat clearer than the comparison between equal weights and Bayesian analysis. The differences between the model and the inferred likelihood or parsimony tree are about the same when measured with DC (average 0.016599 for parsimony and 0.016495 for likelihood) or with group similarity (0.053320 for parsimony and 0.05408 for likelihood). The inferred parsimony trees are clearly more similar to the model tree than the inferred likelihood trees, when measured with the RF (0.09421 for parsimony and 0.105194 for likelihood) or with SPRd (0.077562 for parsimony and 0.08244 for likelihood). In this case, the possibility that SPRd is biased against the model-based method does not exist, because the comparison was done with the optimal tree returned by RAxML, which is binary.

The number of correct groups retrieved by maximum likelihood is larger than the number of correct groups retrieved by parsimony (Fig. 2d). The problem with maximum likelihood, however, is the very large proportion of incorrect groups (Fig. 2d). Some of those incorrect groups might have been found by rearranging relatively few terminals (as the grey dots in Fig. 2d are much closer to the diagonal than the black dots), but are still farther from having been recovered exactly than with parsimony.

Equal-weights parsimony and implied weighting

Although equal-weights parsimony and model-based methods seem to perform similarly, implied weighting outperformed equal weights for almost every measure. The plots show the results for implied weighting under $K = 12$. Goloboff (1993) introduced the method with a strong concavity ($K = 3$, which is still retained as the default in TNT), but it has since become evident that, particularly for larger data sets, better results are obtained with larger values of K (e.g. Goloboff et al., 2008a, p. 765), so that the method weights more mildly against homoplastic characters. The results for $K = 6$ and $K = 8$ were also calculated, but (even if superior to those of equal-weights and model-based methods) they were normally inferior to those of $K = 12$.

The tree inferred by implied weighting is (for almost every combination of taxa, characters and exponential function) more similar to the model tree than the tree inferred using equal weights (Fig. 2e). This difference is overwhelmingly in favour of implied weighting. The number of correct groups retrieved by implied weighting is, for all numbers of taxa, characters, and λ, higher than for equal weights[4] (Fig. 2f). The only aspect in which implied weighting performed worse than equal weights is in finding a larger proportion of incorrect groups.

Model-based methods and implied weighting

Although the difference between model-based methods and equal-weights parsimony is subtle to nonexistent, the case is different with regard to implied weights. The tree inferred by implied weighting is much more similar to the model tree than the tree inferred by MrBayes (Fig. 3a) or maximum likelihood (Fig. 3c), for all measures of tree distance. The only comparisons for which implied weighting does not strongly outperform model-based methods are with the number of correct nodes retrieved by maximum likelihood (the average of which, 0.89672 is only slightly worse than the average for implied weights 0.90446), and the proportion of wrong groups found by Bayesian analysis (0.02944, somewhat better than the 0.05794 in implied weights).

When all three methods (Bayesian, likelihood and implied weights) are compared against equal weights (Fig. 4a–d) the same picture emerges. The white dots (representing implied weighting) are situated lower than the black (likelihood) and grey (Bayesian) dots for the measures that evaluate the degree of difference between inferred and model tree (RF and DC).

Effect of different values of lambda

With different values of lambda, the distribution of homoplasy changes (Fig. 1c). As the data set contains (Fig. 5) more characters with no or very little homoplasy (higher values of lambda), the results of implied weights become more similar to those of equal weights (white dots). When there is greater homoplasy, implied weighting produces results more different from those of equal weights (black dots), with inferred trees more similar to model tree (e.g. Fig. 5a, b), and a larger advantage in number of correct groups recovered (Fig. 5c); in contrast, the proportion of incorrect groups inferred by implied weights also exceeds that for equal weights by a larger factor (Fig. 5d).

[4]Note that each dot in the graphs represents the average of 10 values, corresponding to a combination of taxa/characters/lambda. It is then possible that in some of the individual data sets the number of correct groups recovered was larger for equal weights, although this clearly cannot be the norm.

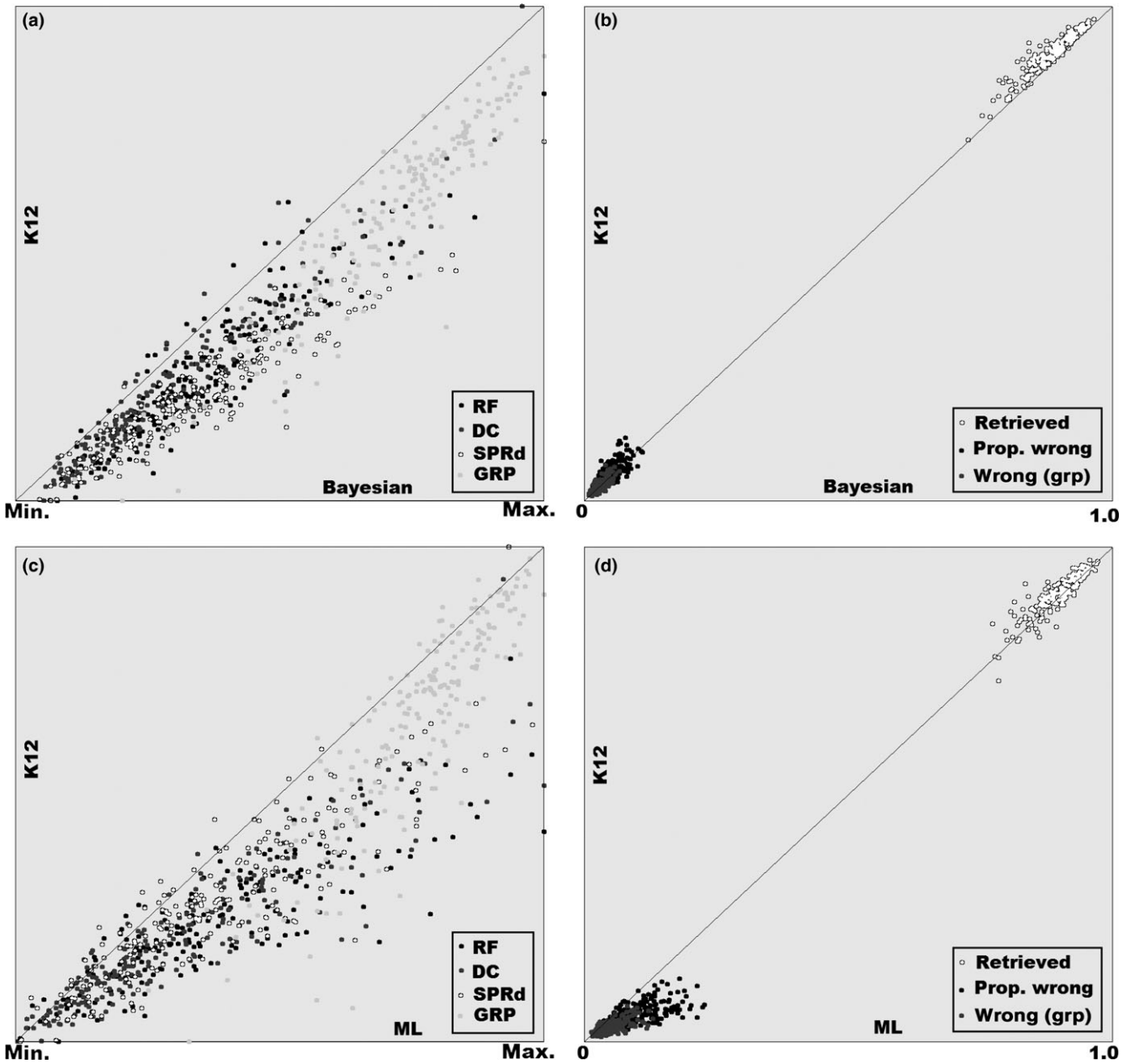

Fig. 3. (a) Distances between model and inferred tree, for implied weights, plotted against Bayesian inference. (b) Retrieved and incorrect groups for implied weights, plotted against Bayesian analysis. (c) Distances between model and inferred tree, for implied weights, plotted against maximum likelihood (ML). (d) Retrieved and incorrect groups for implied weights, plotted against ML. Data sets generated as for Fig. 2.

Empirically derived frequencies of homoplasy

The results shown in Figs 2–5 correspond to data sets generated such that the probability of different numbers of steps of homoplasy is determined by an exponential distribution. The script used has also been adapted to use the exact frequencies observed in a most-parsimonious tree under equal weights for a given empirical data set. For each of the 158 empirical data sets, 10 simulated data sets with frequencies of homoplasy matching the observed frequencies were generated and analysed as before (for a total of 1580 simulated data sets). The data were evolved using a random tree as model. The results are similar to those already discussed, and shown in Figs 6–7. As in the data sets generated with the exponential function, likelihood retrieves a high number of correct groups, slightly outperforming implied weights in this regard (0.89269 instead of 0.87910; Fig. 6d), but the difference in the proportion of incorrect groups is more disadvantageous for likelihood than in the simulated data. For parsimony, the proportion of incorrect

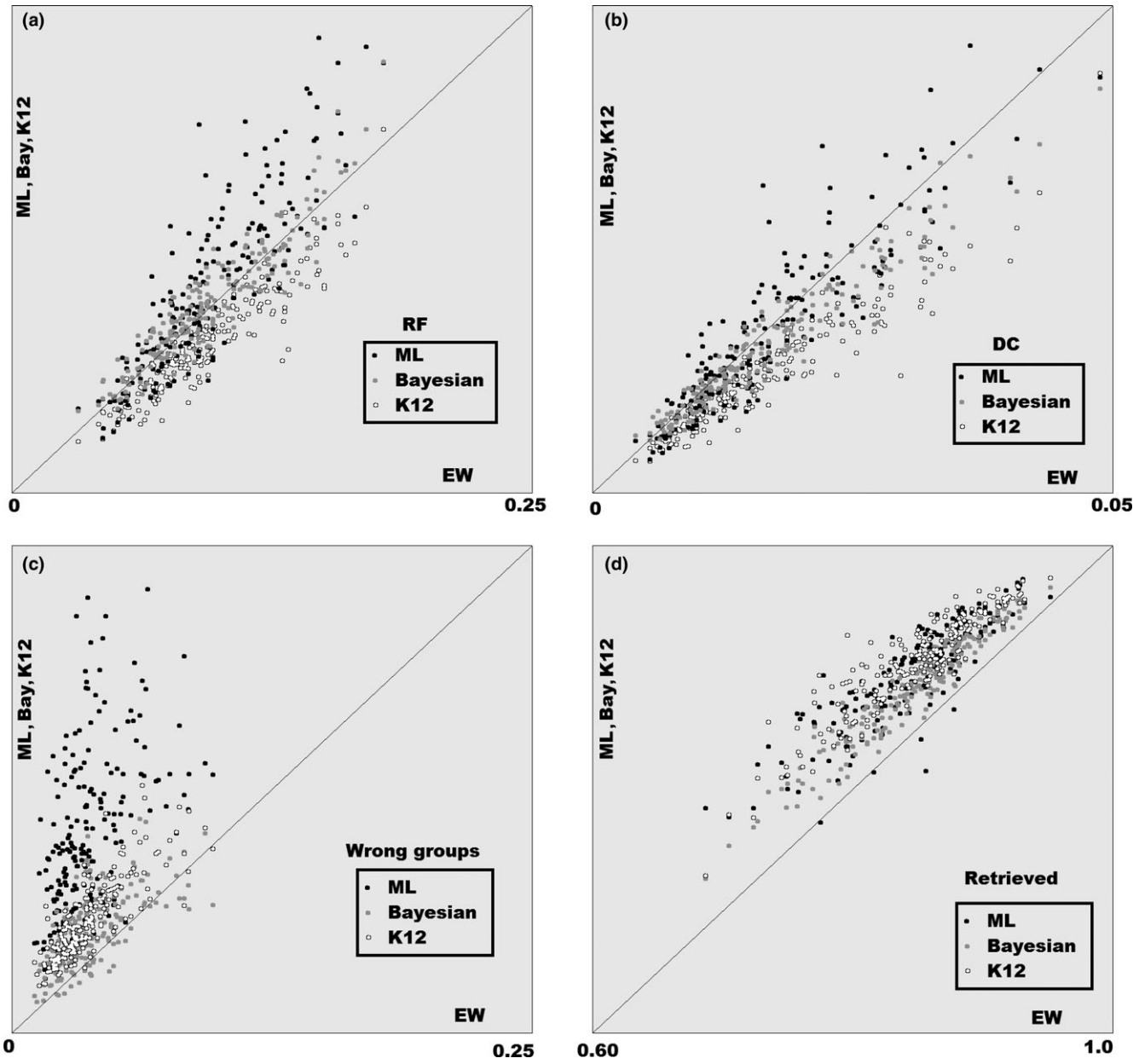

Fig. 4. (a) Robinson–Foulds (RF) distance to the model tree, for maximum likelihood (ML), Bayesian inference and implied weights (K12), plotted against equal-weights parsimony. (b) Same, for symmetric distortion coefficient (DC). (c) Same, for proportion of wrong groups. (d) Same, for retrieved (=correct) groups. Data sets generated as for Fig. 2.

groups decreases for empirically derived frequencies (0.03593 in implied weights, 0.024531 in equal weights) relative to the exponential function (0.05794 in implied weights, 0.03702 in equal weights), whereas in likelihood this proportion remains almost the same in frequencies derived empirically (0.10462) relative to the exponential function (0.10316); as a consequence, the difference in proportions of incorrect groups between likelihood and parsimony is more pronounced for empirically derived frequencies of homoplasy.

Convenience and "Philosophy"

The papers by Wright and Hillis (2014), O'Reilly et al. (2016) and Puttick et al. (2017) are quite succinct, and their conclusions presented as if entirely factual. Yet several important decisions in those studies are taken without any discussion or rationale, even when those decisions are crucial to their conclusions. The main decision is the choice of a model like Lewis's (2001); all those authors simply write as if no possible alternatives existed. But possible

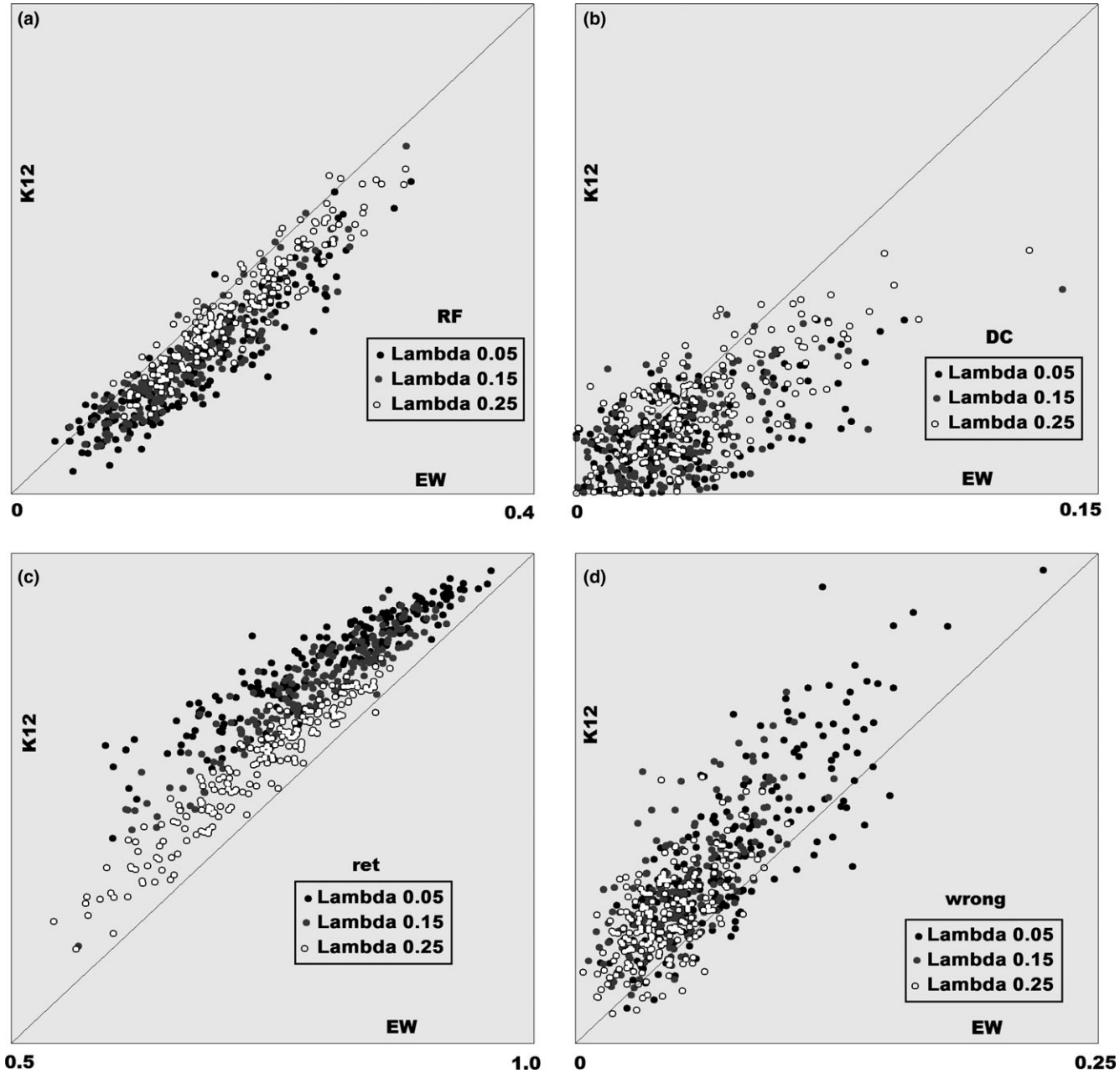

Fig. 5. (a) Robinson–Foulds distance to the model tree, for implied weights, plotted against equal weights (EW), for different distributions of homoplasy. (b) Same, for the symmetric distortion coefficient. (c) Same, for the retrieved (=correct) groups. (d) Same, for proportion of wrong groups. Data sets generated as those of Fig. 2.

alternatives do exist, and the results favouring Bayesian methods need not hold unless one is willing to accept (with Lewis, 2001) that all characters get speeded up or slowed down on exactly the same branches of the tree. If that requirement of simultaneous acceleration and deceleration is dropped, the papers' conclusions do not hold, as shown in the previous section. The problem then becomes one of how morphological characters are more likely to evolve in the real world, a problem that neither Wright and Hillis (2014), O'Reilly et al. (2016) nor Puttick et al. (2017) address.

In addition to the choice of model, there also other aspects in which those authors reach conclusions that are far from obvious. The choice of $K = 2$, which O'Reilly et al. (2016) use for analysing most of the results under implied weights, is one of the most egregious examples. O'Reilly et al. (2016) examined several different values of K in their preliminary assessments; they found that different values of K produce trees

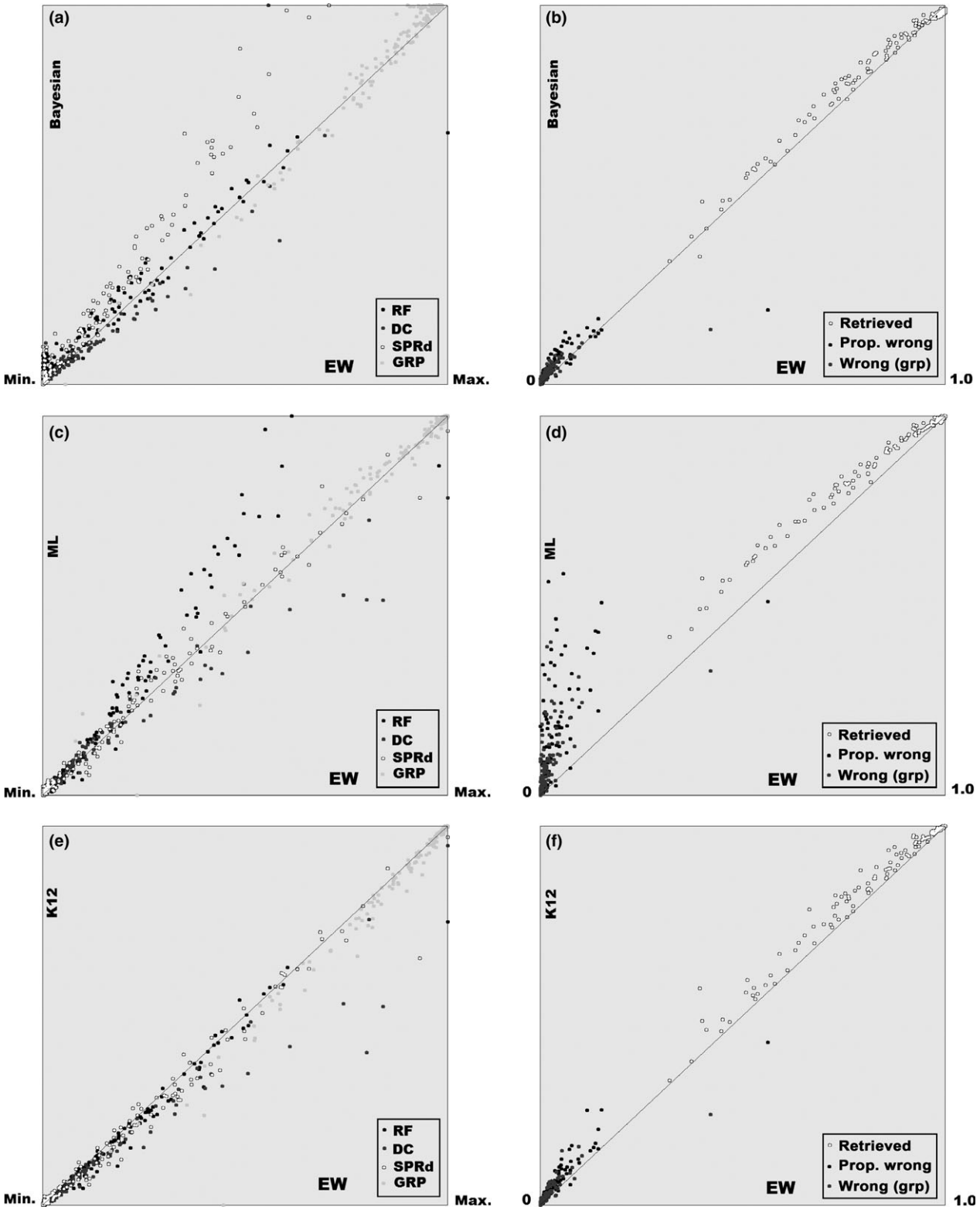

Fig. 6. (a) Distances between model and inferred tree, for Bayesian inference. (b) Retrieved and incorrect groups for Bayesian analysis. (c) Distances between model and inferred tree, for maximum likelihood. (d) Retrieved and incorrect groups for maximum likelihood. (e) Distances between model and inferred tree, for implied weights. (f) Retrieved and incorrect groups for implied weights. In all cases, values plotted against equal weights (EW). Data sets generated so that numbers of taxa and characters, and homoplasy distributions, correspond exactly to each of the 158 empirical data sets (as explained in the text); no missing entries.

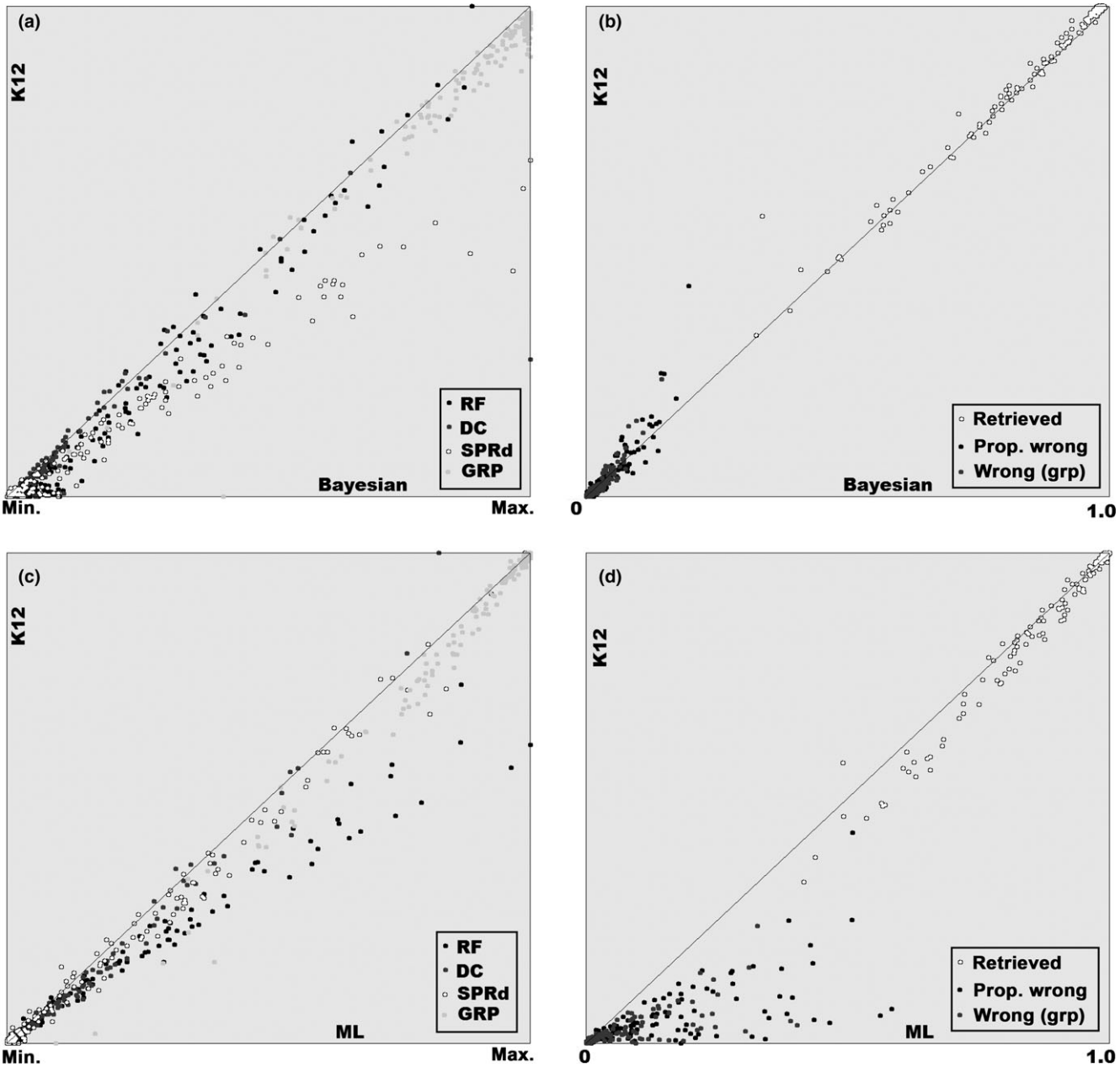

Fig. 7. Distances between model and inferred tree, for implied weights, plotted against Bayesian inference. (b) Retrieved and incorrect groups for implied weights, plotted against Bayesian analysis. (c) Distances between model and inferred tree, for implied weights, plotted against maximum likelihood. (d) Retrieved and incorrect groups for implied weights, plotted against maximum likelihood; note likelihood retrieves a slightly larger number of correct groups (average 0.89530) than implied weights (0.88117), but a much larger number of incorrect groups. Data sets generated as those of Fig. 6.

with different distances to the model tree (their table 1, values plotted here in Fig. 8), and chose precisely the value of K that makes implied weights perform the worst—the only reason offered for that choice being "convenience" (p. 2). As evident in Fig. 8, the values of K that produced the closest results to the model tree were $K = 10$ and $K = 20$; $K = 2$ instead performed worst. O'Reilly et al. (2016) chose the worst K value and criticized implied weights for performing poorly, instead of reaching the almost self-evident conclusion: do not use values of K that weight too strongly against homoplasy, as had already been suggested long ago by Goloboff (1995, pp. 99–100). The same authors, in Puttick et al. (2017), subsequently use only $K = 2$, with neither exploration of alternative concavities nor even a hint that different values of concavity of the weighting function may produce better values—that fact being apparent only from their previous paper

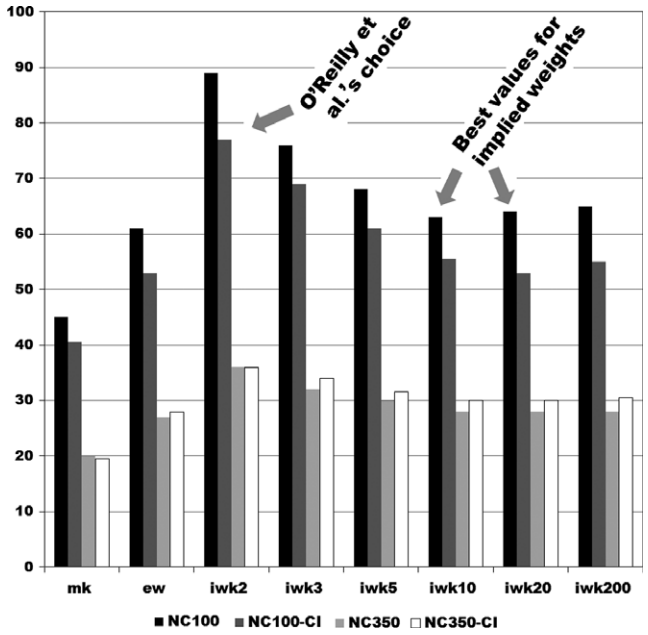

Fig. 8. Median values for Robinson–Foulds distances in O'Reilly et al.'s (2016) study. Mk, Bayesian analysis; ew, equal weights; iwk2 to iwk200, implied weights with $K = 2$–200. See text for discussion.

(O'Reilly et al., 2016), and not explained in the 2017 contribution.

Puttick et al. (2017) differs from O'Reilly et al. (2016) in comparing phylogenetic methods on two alternative topologies, one fully symmetrical, the other fully pectinate (asymmetrical). They explored the two topologies because one of the coauthors had earlier found (Holton et al., 2014) that, on real data sets, different reconstruction methods may produce trees with different degrees of balance, suggesting that at least "some of the methods are biased with respect to tree shape" (p. 436). The surprising finding of Puttick et al. (2017) is that deeper nodes in asymmetric trees are harder to reconstruct, for all phylogenetic methods—whereas all nodes are reconstructed more accurately for symmetrical trees. Puttick et al. (2017) conclude that "the impact of tree topology is of particular concern because empirical phylogenetic trees are invariably asymmetric, and trees of fossil species are infamous for their asymmetry" (p. 6). They never discuss, however, the true cause of the difference between deep and shallow nodes in their asymmetric trees; as all the methods examined by Puttick et al. (2017) are time-reversible (i.e. transitions back and forth between states equally costly or likely), their finding goes directly against theory: the location of the root should be immaterial. The real reason for the difference is evident only in their supplementary R script (available online): the model trees in Puttick et al. (2017) are ultrametric (i.e. perfectly clocklike). The successive splits branching off terminals in their asymmetric trees have decreasing lengths (1, 0.968, 0.936, decreasing by 0.032 every time), whereas the branch leading to the group of the remaining taxa has a constant length (0.032). The symmetric model trees that they used, instead, have all their branches with exactly the same length (0.2). Incredibly, Puttick et al. (2017) never discuss this difference between their two model trees, or the crucial implication that ultrametricity on an asymmetric tree forces a mixture of very long and very short branches. Very long branches are difficult to place for any phylogenetic method, moving around with only minor differences in likelihood or parsimony; this property has been known for decades (e.g. Wheeler, 1990), and Goloboff and Pol (2005, pp. 153–154) used very long branches to emulate the behaviour of missing entries. Because the basalmost branch will move around its location with minor differences in the fit to the data (possibly even suffering long-branch attraction from the almost equally long branch subtending the next terminal split, separated by a short intermediate branch), this will naturally decrease the rate of correct recovery for the group of all taxa but the first split in the tree. The difficulty Puttick et al. (2017) found in recovering some nodes therefore has nothing to do with asymmetry *per se*, but is instead the result of the long branches that ultrametricity forces on an asymmetric tree—the right conclusion would then be that the hardest nodes to recover are those around very long branches, not the deep ones. Several authors have been sceptical about the "morphological clock" (e.g. Beck and Lee, 2014; Puttick et al., 2016), but even if a perfect morphological clock really held in nature, that would still not mean that asymmetric trees including fossils as the earliest splits conform to the pattern simulated by Puttick et al. (2017)—the clock assumes that change along time is a constant, but the earliest splits fossilized millions of years ago, not having had much time to evolve, thus connecting to the tree by short branches. So, not only is Puttick et al.'s (2017) difficulty in recovering some groups the result of very long branches instead of tree shape, it is also irrelevant for the situation they claim it affects the most: fossil phylogenies.

Puttick et al. (2017) also consider very important that "[a]s parsimony methods appear to be less effective than probabilistic approaches, it may be necessary to alter data collection practices by moving away from choosing a selection of characters that undergo few changes, and moving towards scoring all possible characters from the available taxa irrespective of their expected homoplasy" (p. 8). The notion that cladists select for their matrices only those characters thought to have low homoplasy is an absurd caricature—as any sample of cladistic analyses published in the last decades immediately shows. Indeed, if standard cladistic practice consisted of selecting only characters with

low homoplasy, then why would Goloboff (1993) have bothered to propose a method to allow homoplastic characters to be included in the matrix (and given less influence as a result of the analysis, instead of an assumption)? But there is even more irony in Puttick et al.'s (2017) notion that homoplastic characters must be included because "[u]nder the likelihood model, branch length, informed by the number of character changes, contributes to topology estimation" (p. 7). Not only is branch length determined by highly variable characters, but it is also determined by invariant characters, which should be included (or assumed, as in the "Mkv" variant of the Lewis, 2001; model) in vast numbers. Including large numbers of invariant characters leads (other things being equal[5]) to the same trees being optimal under standard likelihood models and parsimony (Tuffley and Steel, 1997). Taking to heart Puttick et al.'s (2017) recommendation would thus always lead to most-parsimonious trees!

Puttick et al. (2017), in addition to their simulations, also discuss four empirical cases. Their discussion of these examples shows, more than the advantages of the Bayesian approach, the depth of their prejudice against parsimony methods. When Bayesian analysis produces an unresolved tree (in the reanalysis of the *Kulindroplax* and the seed–plant data sets), they take this to be the correct result—for no reason other than their conviction that previous conflicts in phylogenetic hypotheses for these groups were "largely an artefact of the false resolution of parsimony methods" (p. 6) and that Bayesian analysis must be correct. When previous authors had postulated possible alternative conclusions on the position of some taxon (as in the case of *Nyasasaurus*, in Nesbitt et al., 2013), Puttick et al. (2017, p. 6) emphasized the agreement of the Bayesian analysis with only one of the alternatives offered (that of Nesbitt et al., 2013), so as to make their results in apparent agreement with firmly established conclusions. The selective presentation of Puttick et al. (2017) is similar to the one used in Lee and Worthy's (2012) claim that model-based methods are superior because they correctly place *Archaeopteryx*, whereas equal-weights parsimony places it in an unusual position. Spencer and Wilberg (2013, p. 666) and Xu and Pol (2014, p. 325) noted that Lee and Worthy (2012) conveniently avoided mentioning that model-based analysis made the Tyrannosauroidea (a widely accepted group) paraphyletic, or that implied weighting simultaneously recovered a monophyletic Tyrannosauroidea and placed *Archaeopteryx* in the standard position.

Congreve and Lamsdell's (2016) paper differs from the three others examined here in presenting more extensive discussions in addition to the simulations. Their arguments, however, are less than compelling. What they call "philosophical" discussion of implied weighting has very little to do with philosophy, and consists instead of repeating criticisms (already addressed by Goloboff et al., 2008a, p. 760) based on loose biological arguments, such as the notion that

> One major flaw of implied weighting is that it assumes homoplastic characters have an equal likelihood of homoplasy across the entire tree … Furthermore, directed homoplasy (homoplasy due to adaptive convergence) can exhibit strong phylogenetic signal, and there is always the risk that, by maximizing character fit, implied weighting could converge on an erroneous topology through maximizing homoplasy. This concern is, in our opinion, the most crucial. No study has been performed to see whether implied weighting is actually consistently minimizing homoplasy or maximizing fit in an ad hoc manner to the extent of retrieving a well-resolved but erroneous topology. (p. 451)

Indeed, implied weighting may be negatively affected by characters that do not "have an equal likelihood of homoplasy across the entire tree", but Congreve and Lamsdell (2016) do not mention that the same is equally true of equal weighting. They do not mention, either, that Goloboff et al. (2008a, p. 760) had already noted that both equal and implied weighting are equally affected by this problem, or that Goloboff et al. (2009) implemented a method with weights that vary on the tree and found it to be incompatible with parsimony. Also, "directed homoplasy due to adaptive convergence" (one presumes, in several characters at the same time) may well mislead implied weighting, but Congreve and Lamsdell do not explain why such convergence would be any less detrimental to equal-weights[6] or model-based methods.

Congreve and Lamsdell (2016, p. 449) state that implied weighting "utilizes a model (based on character fitness, f) to successively downweight characters that the model deems to be homopastic against every possible generated topology". The term "model" has several meanings; implied weighting is certainly not a "model" in the same sense as (say) a JC69 model: it is not a model in the sense of being a statement about how characters are thought to evolve. And implied weighting does not downweight characters "successively"—it was, in fact, proposed as an alternative to successive weighting (Farris, 1969). But perhaps the most enigmatic of Congreve and Lamsdell's (2016)

[5] That is, leaving aside problems in summarizing the results of the Monte Carlo Markov chain discussed by Goloboff and Pol (2005), which Puttick et al. (2017) do not consider. It also leaves aside the problem of different numbers of possible alternative states in the different characters, which can also cause differences.

[6] Obviously, "adaptive convergence" may mislead implied weighting, but not because the method will "maximize homoplasy", as Congreve and Lamsdell (2016) state.

notions is that it is the "model" in implied weights which "deems" a character to be homoplastic. What "deems" a character to be homoplastic is not any "model", but instead a tree; for example, if current notions of insect phylogeny are approximately correct, then there is no "model" which could "deem" the evolution of wings to lack homoplasy.

Congreve and Lamsdell (2016) also seriously misrepresent Goloboff (1993, 1995) when they say (pp. 450–451), that "Goloboff (1993, 1995) suggested that *k* values need to be tweaked for each data set so that the model gives an appropriate response". Goloboff (1993, 1995) never talked about "tweaking" or "appropiate responses"; he said instead that the proper values of *K* remained to be investigated, and that these might vary with numbers of taxa (Goloboff, 1993, p. 89, 1995, p. 100). Goloboff et al. (2008a) provide some hints as to how this might be done, and also stress the need to evaluate more than a single concavity (so as to make results more conservative), given that the hope of determining a unique, "optimal" value of *K* is misguided.

Congreve and Lamsdell (2016) make additional criticisms of implied weighting, but replying to all of those is difficult, for in many cases it is hard to understand what they really mean. For example, they say that although implied weighting "has been demonstrated to increase internal consistency within real data sets (Goloboff et al., 2008a), this pattern is not in and of itself a true test of the efficacy of the methodology because the same could be said of any form of character weighting" (p. 448) and that in the end "the argument rests on implied weighting being preferred because it produces results consistent with itself" (p. 451). Goloboff et al. (2008a) never claimed that the method should be preferred on the grounds that it produces a result identical to itself, as Congreve and Lamsdell (2016) incorrectly state; that would certainly have been silly on the part of Goloboff et al. (2008a), as no method could produce results "inconsistent with itself". Likewise, Goloboff et al. (2008a) never made any reference to an "increase in internal consistency". If by "internal consistency" Congreve and Lamsdell (2016) mean the increase in both jackknife frequencies and measures of stability to addition of characters and taxa (which is what Goloboff et al., 2008a; showed implied weighting improves), then it is not true that "the same could be said of any form of character weighting": Goloboff et al. (2008a) also showed that randomly chosen weights worsen such aspects. And if that is indeed what they mean by "internal consistency", then Congreve and Lamsdell (2016) contradict themselves when they reject jackknife frequencies as evidence, for they consider those same values relevant when they (Congreve and Lamsdell, 2016, p. 451) approvingly cite Källersjö et al.'s (1999) finding that eliminating third positions decreases jackknife frequencies, interpreting this as indication that homoplastic characters should not be downweighted (an idea discussed in depth by Goloboff et al., 2008a, pp. 761–762, 765–767).

Errors

Despite their extensive discussion of abstract principles, Congreve and Lamsdell (2016) think that the only "true test of the utility of this method is to compare how well implied weighting converges on a known tree" (p. 448), which can only be done "using simulated data, for which the tree topology is known" (p. 450). But, interestingly, the simulations that they perform still do not solve the problem automatically—additional considerations are needed. The dilemma in their simulations is that implied weighting, compared to equal weights, recovers a larger number of incorrect groups, but also a larger number of correct groups—thus, one aspect of the results is worsened, whereas another is improved. They then "turn to statistical hypothesis testing" (p. 452), and flatly state (as if this was an obvious imperative of statistical theory) that the only important concern is whether a method minimizes the number of incorrect groups (which they analogize with errors of Type I). They give absolutely no reason for their preference; any serious method of inference must consider also the degree to which correct groups are recovered by the method, and it is well known that equal weights does poorly in this regard, by virtue of being very conservative.

It is also far from evident that the absolute number of erroneous groups (as measured by Congreve and Lamsdell, 2016) can be analogized with Type I errors. A much better candidate for measuring Type I errors is the *proportion* of incorrect groups, relative to the number of groups inferred by the method in question. That is, the probability of a group picked from the (consensus of) optimal trees being correct, and this is measured by the proportion (relative to the number of nodes in the consensus), not by the absolute number. More simply: a method that indicates 100 groups, only three of which are wrong, seems generally preferable to another method which indicates three groups, two of which are wrong. With Congreve and Lamsdell's (2016) measure (absolute number of incorrect groups), a complete bush is a "perfect" tree. They admit this much themselves:

> given how we have scored this data [sic], an analysis in which the equal weights consensus was entirely a polytomy would by definition be considered superior to an implied weights tree with only one error. Such an example would not be a fair treatment of the accuracy of implied weights (p. 457)

Given that problem, they compare the slopes for the difference in numbers of correct groups between equal and implied weights (their fig. 5, graphs on the left), with the slopes for the difference in numbers of incorrect groups (their fig. 5, graphs on the right), in both cases as a function of the number of polytomies. Congreve and Lamsdell (2016) find that "the slope of the regression line for the incorrect vs unknown plot was substantially steeper than the correct vs unknown plots". They conclude from this (p. 457) that "using implied weights to resolve polytomies found in equal weighted analyses... shows a far stronger trend towards incorrectly resolving data rather than correctly resolving data" and that "implied weights simply 'picks' a solution to these conflicts seemingly at random, and more often than not it tends to be incorrect." That conclusion hardly follows from the mere fact that one of the regression lines has a steeper slope: line positions are determined by both slope and *intercept*. Our Fig. 9 (redrawn from Congreve and Lamsdell's fig. 5, superimposing the lines for correct and incorrect groups in the same graph, with different colours) shows that, even if the line for the difference in correct groups is steeper, for all practical purposes (i.e. the maximum possible differences in correct/incorrect nodes, bounded by the numbers of taxa) the difference in number of incorrect groups (black line) is always going to be less than or equal to the difference in number of correct groups (lighter line). Congreve and Lamsdell's (2016) comparison is just poor statistics.

As already discussed, rather than the absolute numbers of wrong groups, it seems more reasonable to consider proportions of wrong groups. The proportion of incorrect groups, however, is higher in implied

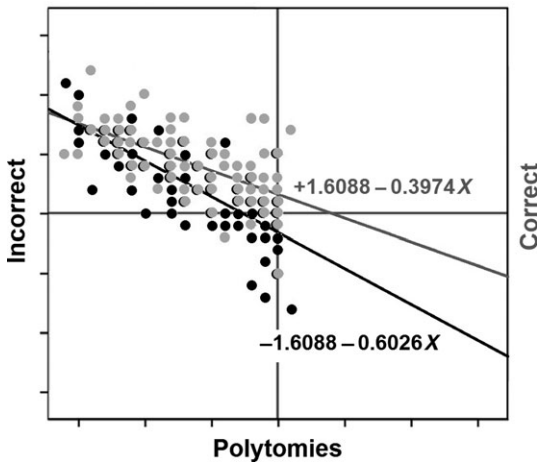

Fig. 9. Example showing that consideration of slopes of regression lines for correct and incorrect groups (relative to unresolved groups) in implied groups is insufficient to establish a proper comparison; consideration of intercepts is also required. Redrawn from Congreve and Lamsdell (fig. 5). See text for discussion.

weights (unless tree-collapsing is taken into account; see next section); the number of correct groups retrieved is also higher. The problem then becomes one of how to trade-off the additional correct groups, relative to the larger proportion of incorrect groups. In this case, comparing differences in slopes (and intercepts!) is complicated by the fact that the number of groups supported by each method and data set (used to rescale one of the two variables but not the other) can be different.

A different approach to the problem must look at the proportion of groups present in one method but not the other that are also correct (i.e. present in the model tree). Congreve and Lamsdell (2016) write as if only implied weights could find some groups not present under equal weights; as a matter of fact, equal weights may also produce consensus trees (or well supported groups) that are not found under implied weights. What truly matters here is whether a group that is present in implied weights is more likely to be correct than wrong, and vice versa. If a group that is present in implied but not equal weights is more likely to be correct than wrong, then it is preferable to add those groups to the results of equal weights. This probability is simply calculated as the number of groups present under implied but not equal weights that are correct, divided by the number of groups present under implied but not equal weights. As long as this probability is above 0.5, then adding the group to the results of equal weight is preferable to not adding it. When comparing two methods, the proportion of groups present in one but not the other being correct may be above 0.5 for both methods, given that either method may produce trees with polytomies. Consider the case of model tree (A(BC)(DE)), method 1 producing the tree (A(BC)DE) and method 2 producing (ABC(DE)). The proportion of groups in method 1 but not method 2 that are correct is 1.0 (only one group, BC, which is correct), and the proportion of groups in method 2 but not method 1 that are correct is *also* 1.0 (only one group, DE, which is correct). A full comparison of these values is included in the Supplementary Material; for each combination of taxa, characters and homoplasy distributions, a total of 100 data sets was simulated, for a total of 3600 simulated data sets.

Figure 10a shows the results of such a comparison between equal and implied weights. For the majority of cases, the groups found by implied weighting and not equal weights are more likely to be correct than wrong (mean 0.66257, in data sets with no missing entries), whereas the groups found by equal weights are more likely to be wrong than correct (0.48179). In 98.89% of cases implied weights had some groups not present in equal weights (an average number of 7.18155 groups), whereas in 41.89% of cases equal

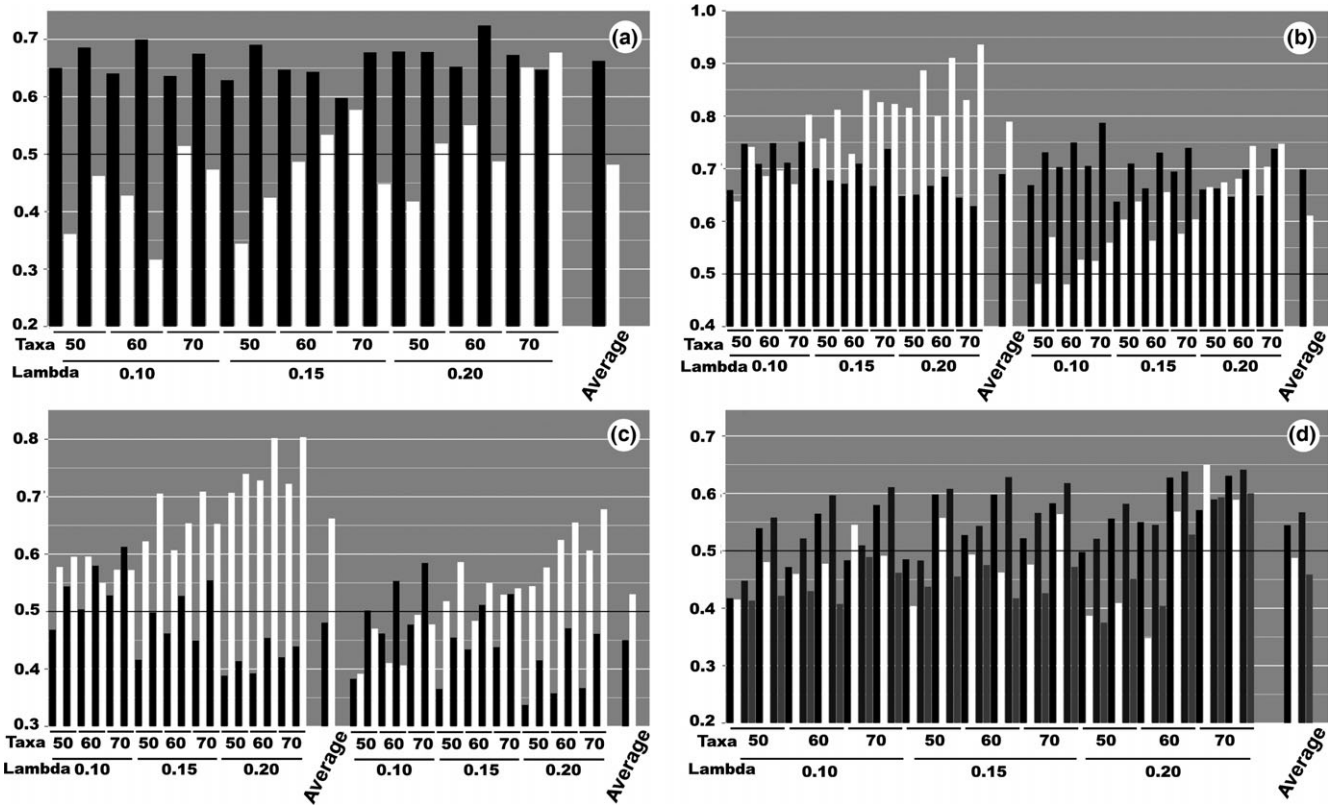

Fig. 10. (a) Proportions of groups found in by equal-weights parsimony but not implied weights (white), and groups found by implied weights but not equal-weights parsimony (black) that are correct. (b) Same, for equal-weights parsimony (black) and Bayesian analysis (white); the set of values on the right side of the graph corresponds to data sets with 25% of missing entries, as described in the text. (c) Same, for equal-weights parsimony (white) and maximum likelihood (black). (d) Proportions of groups found in by equal-weights parsimony but not implied weights (white), groups found by implied weights but not equal-weights parsimony (black), proportion of groups found by extended implied weighting but not equal-weights parsimony (dark grey), proportion of groups found by equal-weights parsimony and not extended implied weighting (light grey) that are correct, for data sets with 25% missing entries (as described in the text). See text for additional discussion.

weights had some groups not present in implied weights (an average number of 2.58630 groups).

The comparison between Bayesian analysis and equal weights (Fig. 10b) produces the hypothetical situation described above, where the unique groups found by both methods have good chances of being correct. Bayesian analysis had groups not found by equal weights in 99.11% of cases (an average number of 6.35743 groups), whereas equal weights had them in 100% (average of 2.52778 groups). The proportion of groups found by equal weights and not Bayesian analysis that are correct is 0.68972, and the proportion of groups found by Bayesian analysis and not equal weights that are correct is 0.78944. Thus, the best result would come—according to this comparison—from combining the groups (if combinable) of Bayesian analysis and equal weights; either of these two methods alone is likely to miss some supported groups.

Maximum-likelihood produced groups (Fig. 10c) not present in equal weights in 100% of the cases (an average of 12.94611 groups), and equal weights produced groups not present in maximum likelihood in 86.17% of cases (an average of 3.75057 groups). Of the groups in maximum likelihood but not equal weights, the proportion of correct groups is 0.48057. Of the groups in equal weights but not maximum likelihood, the proportion of correct groups is 0.66211. Then, any group present in equal weights but not maximum likelihood should be included in the result, whereas a group present in maximum likelihood but not equal weights should not. The combination of both results (maximum likelihood and equal weights) should thus yield exactly the equal weights tree.

This approach to the problem solves the dilemma posed by Congreve and Lamsdell (2016), of how to consider the differences in errors without unduly favouring methods that produce completely unresolved trees. Their conclusions regarding the relative merits of equal and implied weights is due, more than to the model used in the simulations, to a poor analysis of the trees resulting from their simulations. Contrary to Congreve and Lamsdell (2016), implied weights does not select solutions to character conflict at random.

The present comparisons show that a group found by implied but not equal weights is more likely to be correct than wrong, and a group found by equal but not implied weights is (slightly) more likely to be wrong than correct. Therefore, the best final conclusion from considering both the tree produced by implied weights and the tree produced by equal weights is exactly the same as ... the implied weights tree.

Collapsing poorly supported groups

The documentation of TNT states that "it is very important that you evaluate group support when doing implied weighting. Because exact ties are very unlikely, this criterion may produce overresolved trees if poorly supported groups are not collapsed." However, neither O'Reilly et al. (2016) nor Congreve and Lamsdell (2016) considered the effect of tree-collapsing. In the case of Bayesian analysis, the tree produced should already lack poorly supported groups. RAxML produces a single optimal tree, without attempting to find multiple equally likely trees (groups with no or poor support can be eliminated only by considering their bootstrap frequencies). Note that these differences do not actually correspond to differences implicit in the optimality criteria themselves, but instead simply to the specific way in which these programs have been designed. For that reason, no explicit comparison is provided here between model-based and parsimony methods for poorly supported groups eliminated; only the parsimony methods (equal and implied weighting) are properly compared in this regard.

The elimination of weakly supported groups to reduce the number of incorrect groups retrieved could be done either by calculating Bremer supports (absolute, Bremer, 1994; relative, Farris et al., 1996; or a combination of both, Goloboff, 2014b, p. 278), or resampling (bootstrapping, Felsenstein, 1985; or parsimony jackknifing, Farris et al., 1996). Puttick et al. (2017) offered a strange reason to avoid bootstrapping in morphological data sets: "phenotypic data does not meet the assumption that phylogenetic signal is distributed randomly among characters" (p. 2). As a result, their study used "parsimony methods which recover resolved trees" (p. 2), effectively not giving likelihood or parsimony the chance of improving the RF distance to model tree by eliminating poorly supported groups. The same point was made by Brown et al. (2017). The idea that the bootstrap requires that "phylogenetic signal" is uniformly distributed among characters would make the bootstrap incompatible with standard analysis of DNA sequence data: multiple evolutionary rates are premised on the notion that such "signal" is not equally strong in all characters. And Felsenstein himself considered that both morphological data and multiple rates are compatible with the bootstrap: "If we could consider successive characters as independently drawn, having a mix of rates of evolution, or a mix of body regions, would not endanger the bootstrap" (Felsenstein, 2004, p. 344). Felsenstein (2004) was discussing the statistical interpretation of the bootstrap, and even fewer assumptions are needed when resampling is used only as a means to eliminate weakly supported groups (Farris et al., 1996, p. 109; Farris, 2002, p. 352). When the standard bootstrap is used to eliminate weakly supported groups, the problems created by characters with different weights can be avoided with modifications proposed by Goloboff et al. (2003). And, finally, even if Puttick et al. (2017) were right in the argument against bootstrapping morphological data sets, they might as well have used Bremer supports to eliminate weakly supported groups.

In this section, poorly supported groups are eliminated by TBR-collapsing, accepting rearrangements with an absolute score difference equivalent to a step (or a step in a character with no homoplasy, for implied weighting) and a relative fit difference (Goloboff and Farris, 2001) of 0.25 to 0.50. When collapsing poorly supported groups, a lower proportion of incorrect groups is found (Fig. 11a,b), both for equal weights and implied weights, although the decrease in the proportion of incorrect groups is more significant for implied weights. This happens regardless of missing entries. Unsurprisingly, the number of correct groups retrieved (Fig. 11c) also decreases; obviously, some inferred groups are correct, but supported only weakly. The trade-off between avoiding incorrect groups and missing correct ones will necessarily come as the method is made more or less restrictive. Because collapsing improves some aspect of the results (not finding incorrect groups) and worsens another (finding fewer correct groups), the overall effect of collapsing on the closeness to the model tree is minimal for most measures of tree distance (Fig. 11d).

Figure 11 compares the results with and without collapsing for the same method of inference. Comparing the results for equal and implied weighting (Fig. 12a, b), it is seen more clearly that the proportion of incorrect groups for implied weighting approaches the proportion for equal weighting, as poorly supported groups are collapsed. Of course, if the goal was only to eliminate incorrect groups from the inferred tree, then collapsing even more stringently (Fig. 12c) will accomplish this goal—but this would have the undesirable side effect of retrieving fewer correct groups (Fig. 12d).

Collapsing poorly supported groups may improve results considerably, in terms of reducing the incorrect groups retrieved. For example, in Congreve and Lamsdell's (2016) own 100 data sets, the probability of a

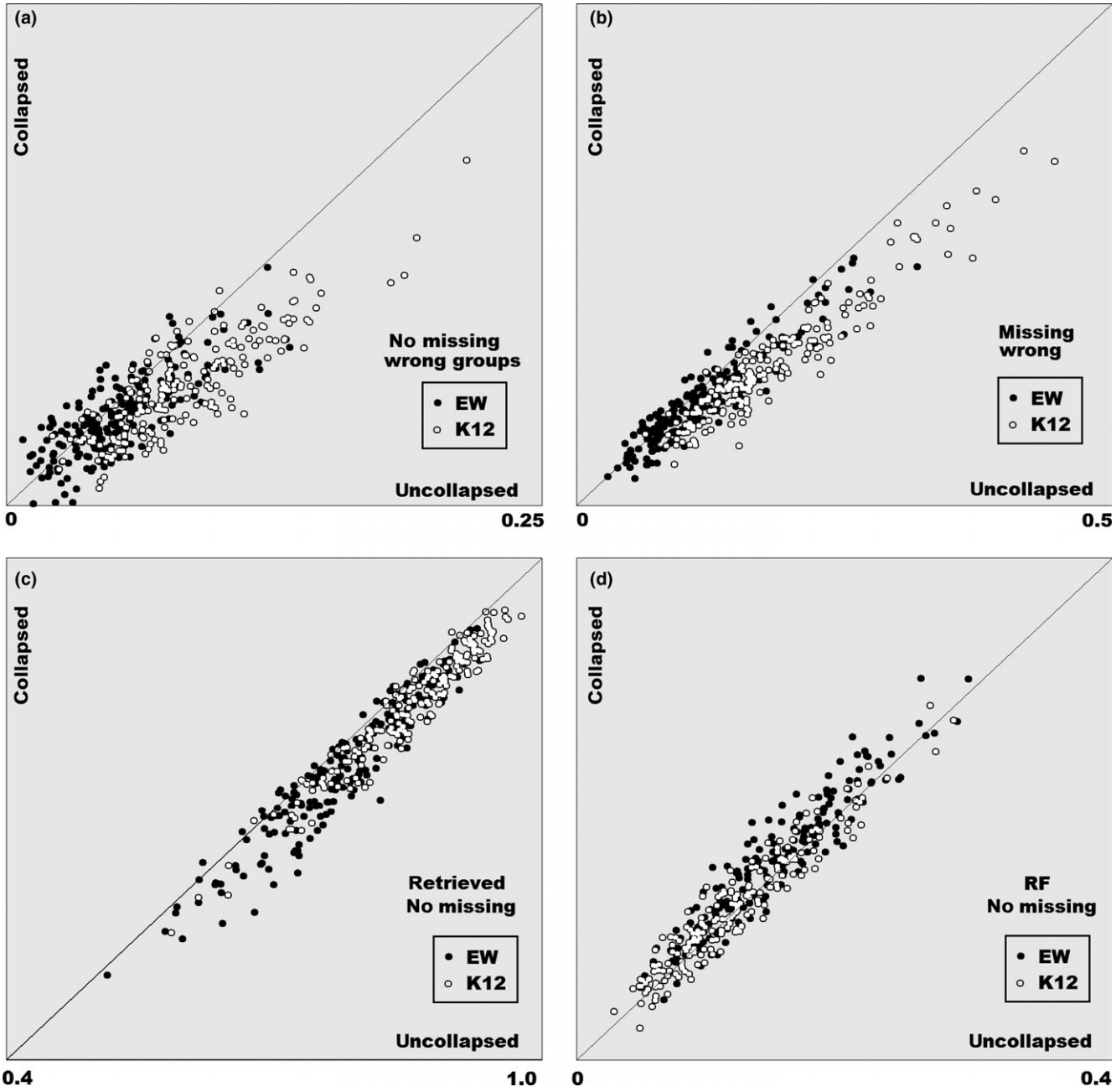

Fig. 11. (a) Proportion of wrong groups found by collapsing poorly supported groups, plotted against the proportion of wrong groups found by not collapsing them, for both equal-weights parsimony (black) and implied weights (white), for data sets with no missing entries. (b) Same as (a) for data sets with 25% of missing entries. (c) Retrieved (=correct) groups found by collapsing poorly supported groups, plotted against the retrieved groups found by not collapsing them, for equal-weights parsimony (black) and implied weights, in data sets with no missing entries. (d) Distance to the model tree when collapsing poorly supported groups, plotted against the distance when not collapsing them, for both equal-weights parsimony (black) and implied weights (white), for data sets with no missing entries.

group found by one method and not the other to be correct is below 0.5, for both implied ($K = 12$) and equal weights. This is in part because their data sets have only 22 taxa (our simulations for the previous section considered 50 to 70 taxa). But if poorly supported groups are collapsed for Congreve and Lamsdell's (2016) data sets (for both methods, with an absolute score difference of one step, or one step in a homoplasy-free character, and a relative fit difference of 0.25), the probability of a group found by equal but not implied weights (which occur in 58% of the data sets, with an average number of 2.94828 groups) being

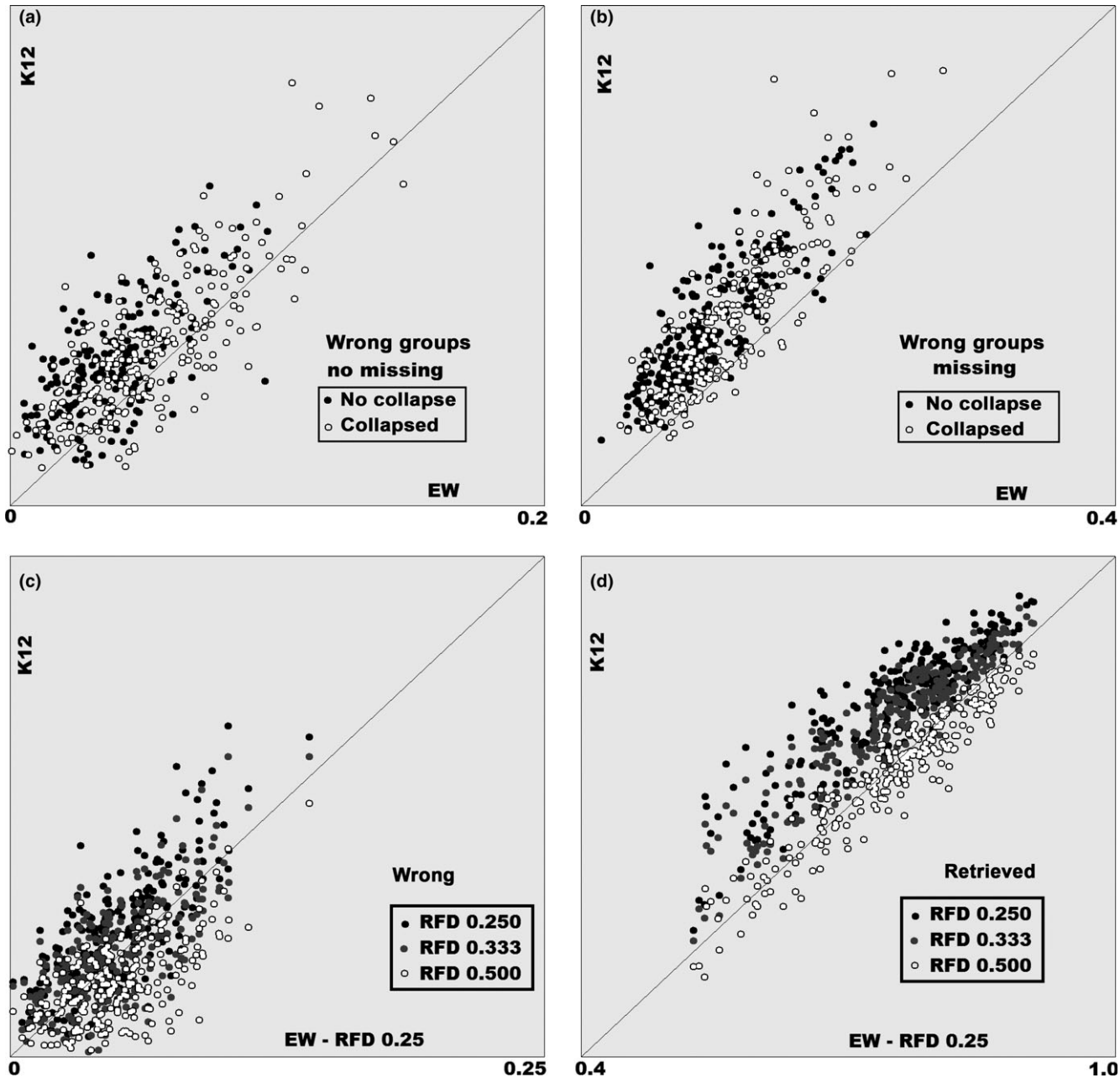

Fig. 12. (a) Proportion of wrong groups for equal and implied weighting, when trees are (white dots) or are not collapsed (black dots), for data sets with no missing entries. (b) Same as for (a) for data sets with 25% of missing entries. (c) Same as (a), when groups are collapsed with varying relative fit differences (RFD). (d) Proportion of retrieved groups for equal and implied weighting, when groups are collapsed with varying relative fit differences.

correct is 0.37671, whereas the probability of a group found by implied but not equal weights (which occur in 83% of the data sets, with an average number of 3.60241 groups) being correct is 0.64679. Thus, consideration of Congreve and Lamsdell's (2016) own matrices, when poorly supported groups are collapsed and the proper comparison of errors is done, is clearly in favour of implied weights.

Missing entries

Of the four studies discussed in this paper, only Wright and Hillis (2014) included missing entries in their simulations, and they did not consider implied weights. Although Congreve and Lamsdell (2016) were of the opinion that the only "true test" for a method is "using simulated data, for which the tree topology is

known" (p. 450), and their own simulated data sets had no missing entries, they still criticized the treatment of missing entries proposed by Goloboff (2014a) and implemented in TNT, from what they consider to be first principles.

Goloboff (2014a), noting that missing entries will usually increase the weight of characters (as some homoplasy may then be unobserved), proposed a method to extrapolate the homoplasy detected among observed entries to missing entries. The contribution of Goloboff (2014a) is not the proposal that missing entries may conceal homoplasy,[7] but instead proposing a practical way in which such influence can be logically considered. A requirement—Goloboff (2014a, p. 265) was explicit about this—is that missing entries do not lower weights by themselves, and that in the absence of homoplasy all characters are equally influential, regardless of missing entries. To accomplish this, the weight of the characters with more missing entries must decrease more rapidly with homoplasy; this can be done by using weighting functions of different strengths for the different characters, with strength depending on proportion of missing entries (rescaled so that the cost of the first step of homoplasy is the same for all characters). The weight in a character with numerous missing entries may well be higher than for a character with none; this depends on the differences in homoplasy, and to the degree to which the user extrapolates observed homoplasy to missing entries (which can be only a small fraction).

Congreve and Lamsdell (2016) confuse this, thinking either that missing entries per se lower character weights or that characters with missing entries can get weights of zero. They state (p. 458) that "assuming homoplasy in missing data under implied weighting effectively results in characters that exhibit missing data being unfairly discounted during the analysis and, by extension, dismisses the signal which such characters have been shown to impart", but provide no example. Congreve and Lamsdell (2016) incorrectly describe the method to consider missing entries as if it could lead to ignore characters, and as if it could fail to identify groups that are in truth defined by synapomorphies in homoplastic characters with missing entries. Their comments (p. 451) on reversible characters go along the same lines, stating that implied weights may fail to recognize groups defined by reversals. But the numerical results of Goloboff's (2014a) method are simply not those which Congreve and Lamsdell (2016) conjecture: in implied weights (extended or otherwise), no character is ever "discounted"; no matter how low its weight, the additional homoplasy in a character supporting a possible group in the tree will always increase the score contribution of that character, and a tree lacking that group could be preferred if and only if there is another character (or several) with higher weights supporting an alternative grouping. Thinking otherwise, as Congreve and Lamsdell (2016) do, reflects only a lack of understanding of how parsimony works as an optimality criterion to select trees.

Congreve and Lamsdell (2016) are also mistaken on the treatment of inapplicable characters, which they think pose a fatal threat to the method. They note that "treating inapplicables as missing data ... is the preferred protocol" (p. 458) and therefore the method of Goloboff (2014a) is "wholly inappropriate to use with inapplicable characters as it will result in assuming homoplasy where no homoplasy can exist, in turn penalizing characters through downweighting simply because they are inapplicable for some taxa." This would indeed be a serious flaw of the method, only if it were true that one is forced to treat inapplicable characters and missing entries in the same way. Goloboff (2014a, pp. 265–266) showed that the concavity ratios for two characters should be proportional to the ratios between observed entries, and also made it clear (p. 268) that the concavities can be manually adjusted by the user. All the user has to do is not count inapplicables as missing entries and set the concavities manually. On top of that, versions of TNT implementing extended implied weights include the option to code inapplicable states differently (as an asterisk), so that those characters are properly weighted automatically.[8]

Congreve and Lamsdell's (2016) criticism of extended implied weighting is not only based on misunderstanding the method. Their idea that taking into account missing entries will produce worse trees is also rejected by what they consider the best means to settle preference for phylogenetic methods:

[7] Goloboff (2014a) attributed this notion to "conversation with colleagues". In fact, the earliest version of Piwe (written in late, 1991 and early, 1992, available at http://www.lillo.org.ar/phylogeny/), before publication of Goloboff (1993), attempted to correct for the fact that missing entries were likely to have some unobserved homoplasy. If Goloboff's memory serves, Kevin Nixon (who was at that time in Goloboff's supervising committee) was one of the colleagues who insisted that missing entries could create problems for standard implied weighting. That early attempt of Goloboff to take into account missing entries, however, did have the undesirable property (which Congreve and Lamsdell erroneously attribute to the method of Goloboff, 2014a) of downweighting even nonhomoplastic characters characters in direct proportion to the number of missing entries. Therefore, that original formula was never published, and subsequently changed to the simpler one in Goloboff (1993), which makes no attempt to consider missing entries. This historical digression serves to show that the issue of missing entries and unobserved homoplasy had already been considered by cladists long ago.

[8] In fact, the 1992 version of Piwe referred to in the previous footnote *also* made this distinction between inapplicable and unobserved entries (with the *set* command, explained in the documentation), thus making Congreve and Lamsdell's (2016) criticism doubly invalid.

simulations. For this, missing entries in 25% of the matrix cells were simulated, and the results of equal, standard and extended implied weights were compared (Fig. 13). The missing entries were simulated by randomly choosing half the characters in the matrix, and for each chosen character replacing the state in half the taxa by a missing entry (the sets of taxa chosen randomly for each character). In the presence of missing entries, standard implied weights continues outperforming equal weights, both in terms of closeness to the model tree (Fig. 13a) and number of correct groups retrieved (Fig. 13b, white dots). It does more poorly in the proportion of groups that are incorrect (Fig. 13b, grey and black dots), but it continues producing groups not found for equal weights that are more likely to be correct than incorrect (Fig. 10d) for 13 of 18 combinations of taxa/characters/lambda (average probability of being correct 0.54454), when equal weights produces groups not found by implied weighting that are more likely to be correct in only 6 of 18 combinations (average probability of being correct 0.48766). When the homoplasy in the observed entries is extrapolated to the missing entries (assuming that missing entries will have 0.75 as much homoplasy as the observed entries), these statistics are slightly improved (Fig. 13c,d); the comparison with implied weighting (Fig. 13e,f) shows this more clearly, with the tree inferred by extended implied weighting generally closer to the model tree, as well as a trend to recover more correct and fewer incorrect groups. Extended implied weighting produces groups not found by equal weights (Fig. 10d) that are now even more likely to be correct than incorrect (in 16 of 18 combinations, with an average probability of being correct of 0.56707). Given that extended implied weighting misses fewer correct groups, equal weighting produces groups not found by extended implied weighting with a somewhat smaller average probability of being correct, 0.45878 (now above 0.5 in only 3 of 18 combinations). Thus, the average difference in probability of being correct for a group not found by the other method, of 0.05689 between standard implied weights and equal weights, is almost doubled (0.10829) when missing entries are taken into account with extended implied weights.

Discussion

The controversy between proponents of parsimony and model-based methods has been standing at least since the early 80s (with Farris and Felsenstein as the main defendants of each position at the time). It is clear that, today, model-based inference is more ubiquitous than parsimony. A clear symptom is that, in the almost 3 years since Bayesian methods were first compared against parsimony for morphological data (Wright and Hillis 2014), and as papers based on simulations for morphological data continued appearing, no response was offered—and within 2 months of the first comparison proclaiming Bayesian methods to perform better than maximum likelihood (Puttick et al., 2017), there is already a reply to that paper (Brown et al., 2017)! But although Brown et al. (2017) are laudably open-minded regarding the relative merits of parsimony vs. model-based methods, they cover just a few aspects of Puttick et al. (2017)—only the claims on maximum likelihood—with no comments on the use of models of DNA evolution for the simulation and analysis of morphological data.

Despite the widespread belief that model-based methods are superior, the existing literature still provides no obvious scientific reasons for such clear-cut preference. A brief summary of the current state of affairs is provided below. Although this summary reflects our views on the subject, it provides no new arguments and borrows heavily from many published discussions. The views from both sides of the dispute are far from monolithic (often using different approaches and arguments to justify the same methods or conclusions), so we feel that the summary below reflects the attitude and rationales that lead many cladists to consider parsimony as a method of analysis justifiable under a wide set of circumstances (regardless of how they choose to articulate those rationales when pressed for justification).

Back to square one

Simulations have been used in the past both to produce results in favour of parsimony (Siddall, 1998; Pol and Siddall, 2001; Kolaczkowski and Thornton, 2004, 2009) or against (Huelsenbeck, 1995; Yang, 1996; Swofford et al., 2001). The results presented in this paper clearly show that, in terms of closeness to the model tree (and when rates of evolution are not assumed to uniformly increase or decrease for all characters along tree branches), the trees produced by equal-weights parsimony are preferable to the trees produced by maximum likelihood, and the trees produced by Bayesian analysis are about as good as those found by equal-weights parsimony. For differences in homoplasy in the different characters corresponding to real data sets, the use of implied weights produces trees that are closer to the model tree than any of the alternative methods, and although implied weights increases the proportion of incorrect groups relative to equal weights, the additional groups produced are more likely to be correct than wrong.

The very fact that the relative merits of alternative phylogenetic methods for simulations under the present model are so different from those under models

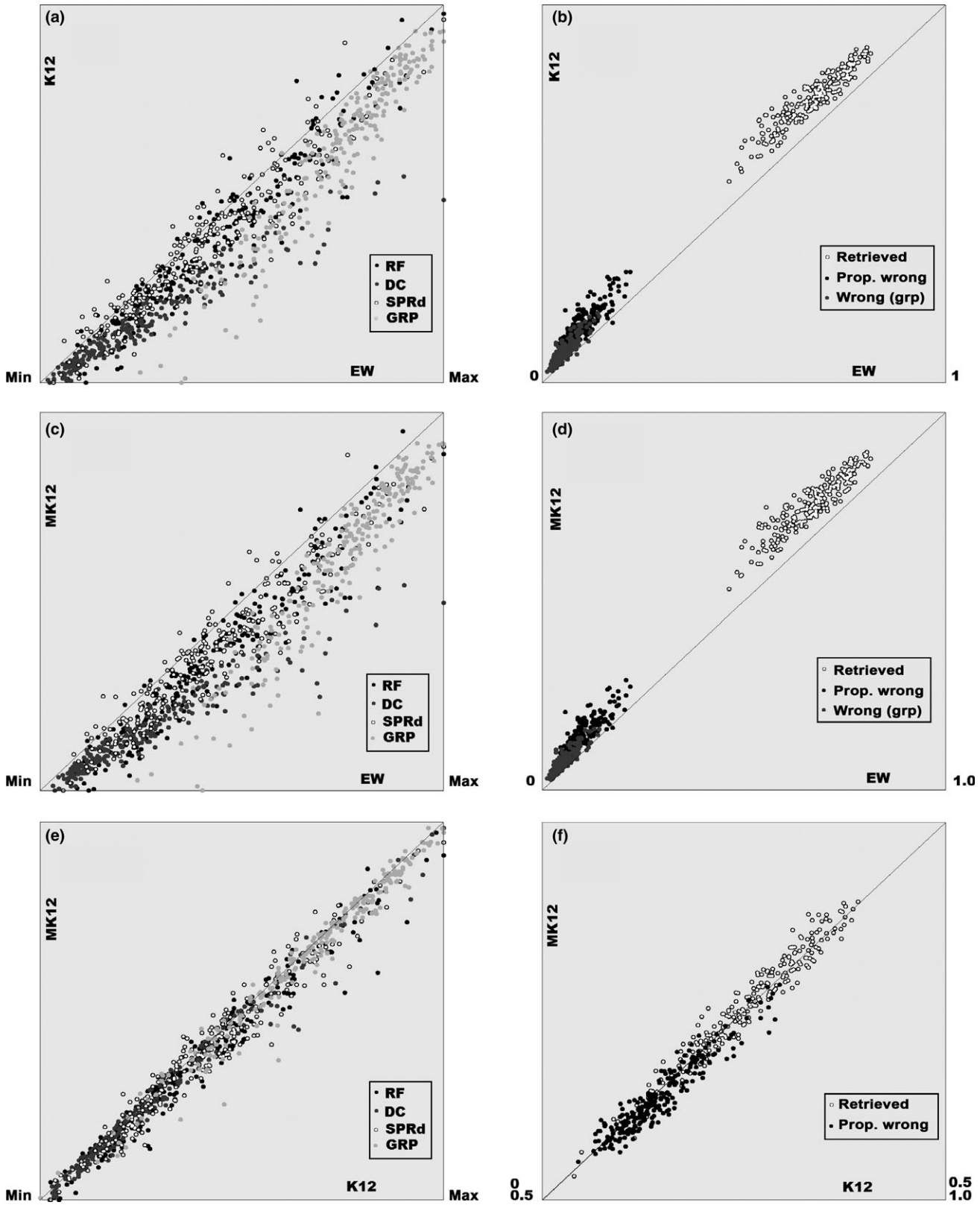

Fig. 13. (a) Distance to model tree, for equal and standard implied weights, for data sets with 25% of missing entries. (b) Retrieved and incorrect groups, for equal and standard implied weights, for data sets with 25% of missing entries. (c, d) Same as (a, b), comparing equal and extended implied weights (MK12). (e, f) Same as (a, b), comparing standard and extended implied weighting.

like the one used by Wright and Hillis (2014), O'Reilly et al. (2016) and Puttick et al. (2017), is enough to highlight the influence of the model used to evolve the data—hardly a surprising result. The model used here is much less restrictive than the models derived from ideas on evolution of DNA sequences normally used by proponents of likelihood. In the case of DNA sequences, the models may or may not be accurate in every detail, but at least well-established theories make those models plausible; however, no such justification can be invoked in the case of morphology. The prime reason for preferring likelihood methods—the statistical inconsistency of parsimony under certain conditions—does not necessarily apply to morphology with the same force. The inconsistency of parsimony results from parallel independent derivations of the same state in nonsister branches, and although this may occur when only four states are the possible conditions that characters can take, it is less likely to occur by chance in morphology, where the number of states a character can have is potentially very large (Farris, 1983, pp. 14–15; Bergsten, 2005, p. 165). This is not to say that parsimony will always produce consistent results for morphological data sets, only that incorrect results are unlikely to result from long-branch attraction. In the case of sequences, for data sets of one or a few genes, parsimony does sometimes produce some questionable groups that are absent from model-based analyses (although it is unclear whether this is the norm; according to Rindal and Brower (2011) it is not); it seems more likely that this results more from the general way in which rates of change are seen as correlated along a given branch of the tree, than because of the finer details of the model (GTR matrix, exact values of the gamma distribution or proportion of invariants, etc.). And even so, in most cases, the only reason to consider those groups preferred by parsimony as questionable is that they conflict with widely accepted morphological synapomorphies, which—however informal—is a parsimony argument: the reason why the bearers of those apomorphic conditions are thought to form a monophyletic group is that separating them would require independent derivations (Farris, 1983). If those cases are to be seen as a justification for likelihood-based analyses of DNA sequences, then the congruence used to justify the preference for likelihood is—somewhat paradoxically—with parsimony for morphology.

Thus, proposals proclaiming the superiority of methods based on models of DNA evolution for analyzing morphological characters lack serious justification—there is no justification other than the fact those models are so commonly used for DNA. All of the assumptions that more or less reasonably can be made for DNA (e.g. all characters increasing or decreasing their probabilities of change at the same branches in concert, fixed substitution rates, base frequencies at equilibirum resulting from the substitution rates, selective regimes constant through time, a maximum of four states) are certainly inapplicable in the case of morphology. Evidently, preferring the results of one method over those of the others amounts to a statement that evolution for that kind of character is more likely to proceed in one or the other manner. In appropriate contexts simulations may well provide important insights on the behaviour of methods that cannot easily be analysed formally, but using them as a way to validate methods is a more delicate question—validation cannot occur unless the model used to evolve the data is itself validated independently. Farris (1983) noted that the empirical consequences of simulations would necessarily be quite limited, precisely for this reason.

Although they generated their data sets with models specifically chosen to make Bayesian methods perform better than parsimony, Wright and Hillis (2014), O'Reilly et al. (2016) and Puttick et al. (2017) asserted, with typical grandiloquence, that Bayesian methods are superior to parsimony in general. That superiority, however, vanishes when data are generated under different models, because simulations alone cannot lead to preference for one method over another unless there is empirical evidence in favour of the model used to run the simulations. In that sense, the results presented here take us back to square one: despite claims by Wright and Hillis (2014), O'Reilly et al. (2016) and Puttick et al. (2017), there is still no empirical evidence that application of probabilistic methods of phylogenetic inference to morphological data is advantageous.

General statistical principles?

Instead of predicating model-based methods on how evolution may actually proceed, a different approach is to defend them by invoking general statistical principles. It is interesting that one of the criticisms of parsimony often involves rejecting any justification that is not strictly framed in probabilistic terms—but considering that methods can be justified only with statistical principles is itself a "philosophical" position. Farris's (1983) argument that *ad hoc* hypotheses (=homoplasies) must be minimized because they serve only to dispose of contradicting evidence appeals to the specifics of phylogenetic inference and common sense, rather than to a particular philosopher. In this sense, both parsimony and model-based methods are justified by recourse to first principles, despite likelihoodists and Bayesians pretending otherwise (Goloboff, 2003, p. 93). But even if the point of view that only explicitly statistical principles can guide the selection of methods is accepted, the choice continues being far from clear.

Although consistency used to be considered a fundamental property by likelihoodists (e.g. Felsenstein, 1978, 1981; Swofford et al., 1996), it now figures less prominently among defences of model-based methods (with Sanderson and Kim, 2000; probably having marked the turning point). The diminished emphasis on consistency is both because the infinite quantities of data on which consistency is premised will never be available in practice, and because consistency is guaranteed only when the data actually evolve under certain conditions (see Steel, 2011), conditions which must be incorporated into the model used to estimate the phylogeny for consistency to occur. Thus, other considerations besides consistency must come into play as well: "although statistical consistency is desirable, it should not override all other considerations—for example, a powerful method that is consistent in most regions of parameter space would generally be preferred over a statistically consistent method that requires huge amounts of data to converge" (Steel, 2011, p. 105). It is ironic that, whereas the lack of consistency was in the past supposed to be the main argument to abandon parsimony, the development of many new models is now made with no concern whatsoever for whether the model results in statistically consistent estimations (e.g. Klopfstein et al., 2015; Wright et al., 2016; Pyron, 2017). And (with the exception of Yang, 2006, p. 176) no Bayesian has expressed any concern in print with examples showing that the standard form of Bayesian analysis (i.e. summarizing the results of the Monte Carlo Markov chain using the frequency of groups from the sample of trees) can easily lead to concluding trees different from the model tree (Goloboff and Pol, 2005); Bayesians have cited this argument (e.g. Brandley et al., 2006; Velasco, 2008; Wright and Hillis, 2014) only in passing, mentioning no details, and without offering any counterargument—thus showing little concern for inconsistency, when suspected to occur in Bayesian analysis. Likewise, the finding of Goloboff (2003) that calculation of likelihoods integrating branch lengths (as estimated in Bayesian phylogenetic programs) may lead to inconsistency (Goloboff, 2003, p. 97) has received no attempted refutation from Bayesians or likelihoodists. But if consistency wasn't so important after all, why was it that parsimony had to be abandoned in the first place? Is there any reason, other than *esprit de corps*, to explain why, in many journals, it became almost impossible to publish a paper using parsimony as the only method?

Another common theme is that parsimony amounts to overparameterization, in that it "estimates" the branch lengths for every character separately (e.g. Huelsenbeck et al., 2008; following Tuffley and Steel, 1997), or that it "estimates" the individual ancestral conditions for every character at every node (following Felsenstein, 1978; and Goldman, 1990; *contra* Goloboff, 2003, pp. 99–101); the alternative is that parsimony can be viewed as a simpler model, in that it is justifiable when probabilities of change for a given character are the same across all tree branches (e.g. Yang, 1996; Goloboff, 2003). That discussion centres solely on whether the models assumed for the inference of phylogeny have many parameters (thus "fitting elephants"; Steel, 2005; borrowing von Neumann's famous quote), while dedicating very little attention to the problem of how morphology might actually evolve. Steel (2005, p. 308) claims that separate rates for different characters are not to be expected because it is difficult to imagine biochemical mechanisms that would act separately on different characters, but the argument is not too compelling. Evolution is a *biological*, not a *biochemical*, phenomenon: when differences in biology, ecology or behaviour are brought into the picture, there can be plenty of biological reasons to expect such differences, even for DNA sequences, but more obviously so for morphology. As shown by the difference in performance between parsimony and model-based methods in this and other simulations, how morphological characters actually evolve does make a difference, and so the problem is not just one of the properties of the inference, but of realism as well.

When it comes to realism, likelihoodists and Bayesians admit that "all models are wrong" (another catchphrase, this one by statistician George Box), in the sense that models are not expected to faithfully describe every detail of evolution. In the case of DNA, standard models of sequence evolution do not consider indels (let alone more complex transformations such as gene rearrangements!); thus, it is far from obvious how to decide which is preferable: a more accurate description of substitions excluding indels, or a more simplistic analysis of substitutions and indels together (as in Poy, Wheeler et al., 2015). That likelihoodists may have the hopes that substitutions alone (when gaps in appropriate alignments are considered as missing data) will suffice to guarantee consistency (Truszkowski and Goldman, 2016) is besides the point, especially when considering that, in practice, different programs for sequence alignment produce very different results, and that phylogeneticists are indeed interested in understanding the evolution of sequences, not just substitutions. Even for substitutions alone, deciding whether current models provide a sufficient approximation to reality is very difficult to evaluate. As with discussions of the problems with either too many or too few parameters, there is the also the subliminal idea that, just as a side effect of mentioning these problems, the methods preferred by likelihoodists and Bayesians will automatically be the right answer to this balance. Nonetheless, as attested by the history of phenetics, just expressing concern with some general

statistical principle does not automatically provide methods fulfilling that principle.

In the end, regardless of how well they articulate their reasons for being sceptical towards model-based methods, the general attitude of cladists is that, based on the necessarily low quantities of data used in phylogenetic analyses,[9] any hopes of estimating with any precision something as complex as a phylogeny (and all its concomitant parameters) are only wishful thinking. If anything, the significant differences in results of different model-based programs for the same data (usually in the order of the difference between parsimony and model-based methods), or by the analysis of different genes with model-based methods,[10] clearly point to statistical phylogenetic inference—in practice—being far from perfectly accurate. Ironically, although parsimony is often criticized for overparameterization, the clear trend in model-based inference is towards increasingly complex models, touted as being more realistic. Note that this "realism" is not exactly the same one that proponents of parsimony expect: not making unrealistic assumptions (Farris, 1983, 2008). The "realism" sought by proponents of model-based methods, instead, assumes that the parameters for ever more specific details can be estimated (e.g. Klopfstein et al., 2015; Wright et al., 2016; Pyron, 2017). The great fanfare with which the ever-increasing precision and level of detail of model-based methods are announced is, obviously, a good selling point, but some taxonomists prefer applying a method that refrains from attributing specific probabilities to phylogenies—not because inference can be established with certainty (as proponents of parsimony are sometimes caricaturized), but because uncertainty occurs at such a basic level. It is hard to argue with Yang's (2006) charitable characterization of parsimony, that "perhaps one should be content to consider parsimony as a heuristic method of tree reconstruction that often works well under simple conditions, rather than seeking a rigorous statistical justification for it", but most cladists would also add that exactly the same reasoning applies to maximum-likelihood and Bayesian analysis. Contrary to what is stated by some statistical phylogeneticists (e.g. Felsenstein, 1987, p. 208, 2001, pp. 466–467), this position does not result from adopting abstract, bizarre philosophical positions, but rather, from common sense considerations of what is known and *un*known about the process of evolution.

[9] Phylogenomic studies are no exception to this, as the trend is to analyse multi-gene data sets with unlinked models, thus estimating the parameters for each partition from the low quantities of data available from a single gene. Interestingly, the more partitions, the more similar the results should be between likelihood and parsimony (from Tuffley and Steel, 1997).

[10] Another trend here is to consider that the discordance in phylogenetic trees for different genes does reflect reality, distinguishing "gene" from "species" trees and blaming all of the incongruence on incomplete lineage sorting. This amounts to keeping faith in the results of model-based methods even when they produce conflicting answers—i.e. with an ever-increasing detachment from evidence (see discussion in Gatesy and Springer, 2014; Simmons and Gatesy, 2015; Springer and Gatesy, 2016).

Acknowledgements

We appreciate comments on the manuscript made by Santiago Catalano, Mark Simmons and Claudia Szumik. Three reviewers (two anonymous, and Ward Wheeler), as well as the Associate Editor, provided useful comments on the original submission which helped improve the paper. The financial support of CONICET (PIP, 0687, PUE-UEL, to PAG) as well as NSF, NASA and FAPESP (via a grant to J. Cracraft and L. Lohman, entitled "Assembly and evolution of the Amazonian biota and its environment: an integrated approach", with participation from PAG) is deeply appreciated.

References

Beck, R., Lee, M., 2014. Ancient dates or accelerated rates? Morphological clocks and the antiquity of placental mammals. Proc. R. Soc. B 281, 20141278.

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

Brandley, M., Leache, A., Warren, D., Mcguire, J., 2006. Are unequal clade priors problematic for Bayesian phylogenetics? Syst. Biol. 55, 138–146.

Bremer, K., 1994. Branch support and tree stability. Cladistics 10, 295–304.

Brown, J., Parins-Fukuchi, C., Stull, G., Vargas, O., Smith, S., 2017. Missing the point (estimate): Bayesian and likelihood phylogenetic reconstructions of morphological characters produce generally concordant inferences. A comment on Puttick et al. bioRxiv 114793. https://doi.org/10.1101/114793.

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

Farris, J., 1969. A successive approximmations approach to character weighting. Syst. Zool. 19, 374–385.

Farris, J., 1973. On comparing the shapes of taxonomic trees. Syst. Zool. 22, 50–54.

Farris, J., 1983. The logical basis of phylogenetic analysis. In: Platnick, N., Funk, V. (Eds.), Advances in Cladistics II. Columbia University Press, New York, NY, pp. 7–36.

Farris, J., 2002. RASA attributes highly significant structure to randomized data. Cladistics 18, 334–353.

Farris, J., 2008. Parsimony and explanatory power. Cladistics 24, 825–847.

Farris, J., Albert, V., Källersjö, M., Lipscomb, D., Kluge, A., 1996. Parsimony jackknifing outperforms neighbor-joining. Cladistics 12, 99–124.

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

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

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

Felsenstein, J., 1987. Comment [on Statistical analysis of hominoid molecular evolution]. Stat. Sci. 2, 208–209.
Felsenstein, J., 2001. The troubled growth of statistical phylogenetics. Syst. Biol. 50, 465–467.
Felsenstein, J., 2004. Inferring Phylogenies. Sinauer Associates, Sunderland, MA.
Fitch, W., Markowitz, E., 1970. An improved method for determining codon variability in a gene and its applications to the rate of fixation of mutations in evolution. Biochem. Genet. 4, 579–593.
Gatesy, J., Springer, M., 2014. Phylogenetic analysis at deep timescales: unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. Mol. Phylogenet. Evol. 80, 231–266.
Goldman, N., 1990. Maximum likelihood inference of phylogenetic trees, with special reference to a Poisson process model of DNA substitution and to parsimony analyses. Syst. Zool. 39, 345–361.
Goloboff, P., 1993. Estimating character weights during tree search. Cladistics 9, 83–91.
Goloboff, P., 1995. Parsimony and weighting: a reply to Turner and Zandee. Cladistics 11, 91–104.
Goloboff, P., 2003. Parsimony, likelihood, and simplicity. Cladistics 19, 91–103.
Goloboff, P., 2005. Minority rule supertrees? MRP, compatibility, and Minimum Flip supertrees may display the least frequent groups. Cladistics 21, 282–294.
Goloboff, P., 2007. Calculating SPR-distances between trees. Cladistics 24, 591–597.
Goloboff, P., 2014a. Extended implied weighting. Cladistics 30, 260–272.
Goloboff, P., 2014b. Oblong, a program to analyse phylogenomic data sets with millions of characters, requiring negligible amounts of RAM. Cladistics 30, 273–281.
Goloboff, P., Catalano, S., 2012. GB-to-TNT: facilitating creation of matrices from GenBank and diagnosis of results in TNT. Cladistics 28, 503–513.
Goloboff, P., Catalano, S., 2016. TNT version 1.5, including a full implementation of phylogenetic morphometrics. Cladistics 32, 221–238.
Goloboff, P., Farris, J., 2001. Methods for quick consensus estimation. Cladistics 17, S26–S34.
Goloboff, P., Pol, D., 2005. Parsimony and Bayesian phylogenetics. In: Albert, V. (Ed.), Parsimony, Phylogeny, and Genomics. Oxford University Press, London, pp. 148–159
Goloboff, P., Farris, J., Källersjö, M., Oxelmann, B., Ramírez, M., Szumik, C., 2003. Improvements to resampling measures of group support. Cladistics 19, 324–332.
Goloboff, P., Carpenter, J., Arias, J., Miranda-Esquivel, D., 2008a. Weighting against homoplasy improves phylogenetic analysis of morphological data sets. Cladistics 24, 1–16.
Goloboff, P., Farris, J., Nixon, K., 2008b. TNT, a free program for phylogenetic analysis. Cladistics 24, 774–786.
Goloboff, P., Mirande, M., Arias, J., 2009. On weighting characters differently in different parts of the cladogram. In Szumik, C., Goloboff, P. (Eds.), A summit of cladistics: abstracts of the 27th Annual Meeting of the Willi Hennig Society and VIII Reunión Argentina de Cladística y Biogeografía. Cladistics 26, pp. 217–218.
Harmon, L., Weir, J., Brock, C., Glor, R., Challenger, W., 2008. GEIGER: investigating evolutionary radiations. Bioinformatics 24, 129–131.
Holton, T., Wilkinson, M., Pisani, D., 2014. The shape of modern tree reconstruction methods. Syst. Biol. 63, 436–441.
Huelsenbeck, J., 1995. Performance of phylogenetic methods in simulation. Syst. Biol. 44, 17–48.
Huelsenbeck, J., Anné, C., Larget, B., Ronquist, F., 2008. A Bayesian perspective on a non-parsimonious parsimony model. Syst. Biol. 57, 406–419.
Jin, L., Nei, M., 1990. Limitations of the evolutionary parsimony method of phylogenetic analysis [erratum in Mol. Biol. Evol., 1990 7. 201]. Mol. Biol. Evol. 7, 82–102.
Källersjö, M., Albert, V., Farris, J., 1999. Homoplasy increases phylogenetic structure. Cladistics 15, 91–93.
Klopfstein, S., Vilhelmsen, L., Ronquist, F., 2015. A nonstationary Markov model detects directional evolution in hymenopteran morphology. Syst. Biol. 64, 1089–1103.
Kluge, A., Farris, J., 1969. Quantitative phyletics and the evolution of anurans. Syst. Zool. 18, 1–32.
Kolaczkowski, B., Thornton, J., 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature 431, 980–984.
Kolaczkowski, B., Thornton, J., 2009. Long-branch attraction bias and inconsistency in Bayesian phylogenetics. PLoS ONE 4, e7891.
Lee, M., Worthy, T., 2012. Likelihood reinstates *Archaeopteryx* as a primitive bird. Biol. Lett. 8, 299–303.
Lewis, P., 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. Syst. Biol. 50, 913–925.
Nesbitt, S., Barrett, P., Werning, S., Sidor, C., Charig, A., 2013. The oldest dinosaur? A Middle Triassic dinosauriform from Tanzania. Biol. Lett. 9, 20120949. https://doi.org/10.1098/rsbl.2012.0949
Neyman, J., 1971. Molecular studies of evolution: a source of novel statistical problems. In: Gupta, S., Yackel, J. (Eds.), Statistical Decision Theory and Related Topics. Academic Press, New York, NY, pp. 1–17.
O'Reilly, J., Puttick, M., Parry, L., Tanner, A., Tarver, J., Fleming, J., Pisani, D., Donoghue, P., 2016. Bayesian methods outperform parsimony but at the expense of precision in the estimation of phylogeny from discrete morphological data. Biol. Lett. 12, 20160081, 1–5. https://doi.org/10.1098/rsbl.2016.0081.
Pol, D., Siddall, M., 2001. Biases in maximum likelihood and parsimony: a simulation approach to a 10-taxon case. Cladistics 17, 266–281.
Puttick, M., Thomas, G., Benton, M., 2016. Dating Placentalia: morphological clocks fail to close the molecular fossil gap. Evolution 70, 873–886.
Puttick, M., O'Reilly, J., Tanner, A., Fleming, J., Clark, J., Holloway, L., Lozano-Fernandez, J., Parry, L., Tarver, J., Pisani, D., et al., 2017. Uncertain-tree: discriminating among competing approaches to the phylogenetic analysis of phenotype data. Proc. R. Soc. B 284, 20162290.
Pyron, R.A., 2011. Divergence time estimation using fossils as terminal taxa and the origins of Lissamphibia. Syst. Biol. 60, 466–481.
Pyron, R.A., 2017. Novel approaches for phylogenetic inference from morphological data and total-evidence dating in squamate reptiles (lizards, snakes, and amphisbaenians). Syst. Biol. 66, 38–56.
Rambaut, A., Grassly, N., 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13, 235–238.
Rindal, E., Brower, A., 2011. Do model-based phylogenetic analyses perform better than parsimony? A test with empirical data. Cladistics 27, 331–334.
Robinson, D., Foulds, L., 1981. Comparison of phylogenetic trees. Math. Biosci. 53, 131–147.
Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M., Huelsenbeck, J., 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61, 539–542.
Sanderson, M., Donoghue, M., 1989. Patterns of variation in levels of homoplasy. Evolution 43, 1781–1795.
Sanderson, M., Kim, J., 2000. Parametric phylogenetics? Syst. Biol. 49, 817–829.
Siddall, M., 1998. Success of parsimony in the four-taxon case: long-branch repulsion by likelihood in the Farris zone. Cladistics 14, 209–220.
Simmons, M., Gatesy, J., 2015. Coalescence vs. concatenation: Sophisticated analyses vs. first principles applied to rooting the angiosperms. Mol. Phylogenet. Evol. 91, 98–122.
Spencer, M., Wilberg, E., 2013. Efficacy or convenience? Model-based approaches to phylogeny estimation using morphological data. Cladistics 29, 663–671.

Springer, M., Gatesy, J., 2016. The gene tree delusion. Mol. Phylogenet. Evol. 94, 1–33.
Stamatakis, A., Ludwig, T., Meier, M., 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21, 456–463.
Steel, M., 2005. Should phylogenetic models be trying to "fit an elephant"? Trends Genet. 21, 307–309.
Steel, M., 2011. Can we avoid "sin" in the house of "no common mechanism"? Syst. Biol. 60, 96–109.
Steel, M., Penny, D., 2000. Parsimony, likelihood, and the role of models in molecular phylogenetics. Mol. Biol. Evol. 17, 839–850.
Swofford, D., Olsen, G., Waddell, P., Hillis, D., 1996. Phylogenetic inference. In: Hillis, D., Moritz, C., Mable, B. (Eds.), Molecular Systematics, second ed. Sinauer, Sunderland, MA, pp. 407–514.
Swofford, D., Waddell, P., Huelsenbeck, J., Foster, P., Lewis, P., Rogers, J., 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. Syst. Biol. 50, 525–539.
Truszkowski, J., Goldman, N., 2016. Maximum likelihood phylogenetic inference is consistent on multiple sequence alignments, with or without gaps. Syst. Biol. 65, 328–333.
Tuffley, C., Steel, M., 1997. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. Bull. Math. Biol. 59, 581–607.
Velasco, J., 2008. The prior probabilities of phylogenetic trees. Biol. Philos. 23, 455–473.
Wheeler, W., 1990. Nucleic acid sequence and random outgroups. Cladistics 3, 363–367.
Wheeler, W., Lucaroni, N., Hong, L., Crowley, L., Varón, A., 2015. POY version 5: phylogenetic analysis using dynamic homologies under multiple optimality criteria. Cladistics 31, 189–196.
Wright, A., Hillis, D., 2014. Bayesian analysis using a simple likelihood model outperforms parsimony for estimation of phylogeny from discrete morphological data. PLoS ONE 9, e109210. https://doi.org/10.1371/journal.pone.0109210.
Wright, A., Lloyd, G., Hillis, D., 2016. Modeling character change heterogeneity in phylogenetic analyses of morphology through the use of priors. Syst. Biol. 65, 602–611.
Xu, X., Pol, D., 2014. *Archaeopteryx*, paravian phylogenetic analyses, and the use of probability-based methods for palaeontological datasets. J. Syst. Paleontol. 12, 323–334.
Yang, Z., 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39, 306–314.
Yang, Z., 1996. Phylogenetic analysis using parsimony and likelihood methods. J. Mol. Evol. 42, 294–307.
Yang, Z., 2006. Computational Molecular Evolution. Oxford Series in Ecology and Evolution, University College, London, pp 357.
Yang, Z., 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24, 1586–1591.

Appendix 1

List of empirical data sets used for this study. Data sets 1–70 are from Goloboff et al. (2008a). For each data set, the numbers of characters and then taxa are indicated first, followed by the source. The taxonomic groups are indicated only for unpublished matrices. The data sets themselves are included as Supplementary Material (in the case of unpublished matrices, with taxon names randomized).

1. agonum, 138 x 150: Carabid beetles, genus *Agonum*. Liebherr & Schmidt., 2004. Dtsch. Entomol. Z. 51, 151–206. **2.** amerem, 64 x 56: American Eumeninae (Vespidae). Unpublished (J. Carpenter). **3.** amphi, 156 x 85: Anuran amphibians. Unpublished (a version of the matrix in A. Haas., 2003. Cladistics 19, 2389; J. Faivovich, pers. comm.). **4.** anyph, 200 x 93: Anyphaenid spiders. Ramírez, M. et al., 2004. Zootaxa 668, 18. **5.** apoidea, 139 x 54: Bees and sphecid wasps (Apoidea). Melo, G., 1999. Sci. Pap. Nat. Hist. Mus. Univ. Kansas 14, 155. **6.** araneo, 302 x 83: Araneoid spiders. Agnarsson, I., 2003. Inv. Syst. 17, 719–734. **7.** astr, 36 x 103: *Astragalus* legumes. Camp, P., in Platnick, N., 1989. Cladistics 5, 145–161. **8.** ausmat, 96 x 72: Thynnine wasps (Tiphiidae). Kimsey, L., 2000. J. Hymen. Res. 9, 18–28. **9.** bats, 250 x 75: Bats. Unpublished (modified from Giannini & Simmons, 2005. Cladistics 21, 411–437; N. Giannini, pers. comm.). **10.** bemby, 163 x 53: Carabids (genus *Bembidion* et al.). Maddison, D.R., 1993. Bull. Mus. Comp. Zool. 153, 143–299. **11.** bertetal, 59 x 54: Tinamou species. Bertelli et al., 2002. Syst. Biol. 51, 959–979. **12.** bivalvia, 183 x 76: Bivalves (Mollusca). Giribet & Wheeler., 2002. Invert. Biol. 121, 271–324. **13.** bomb, 44 x 50: Bees (genus *Bombus*). Williams, P., 1994. Syst. Entomol. 19, 327–344. **14.** bracon, 89 x 126: Braconid wasps. (D. Quicke, pers. comm.). **15.** brochu, 164 x 62: Gavialids (Crocodilia). C. Brochu., 1997. Syst. Biol. 46, 479–522. **16.** bryo, 43 x 56: Polytrichales (Bryophita). J. Hyvonen et al., 2004. Mol. Phyl. Evol. 31, 915–928. **17.** camilo, 110 x 50: Scorpions (genus *Bothriurus*). Unpublished (Camilo Mattoni, Ph.D. Thesis). **18.** caronieto, 110 x 64: Mayflies. Unpublished (C. Nieto, Ph.D. Thesis). **19.** centip, 222 x 80: Centipedes. Edgecombe & Giribet., 2004. J. Zool. Syst. Evol. Res. 42, 89–134. **20.** cephalo, 101 x 78: Cephalopods. Lindgren et al., 2004. Cladistics 20, 454–486. **21.** cocos, 268 x 53: Crocodyles. A version from Pol & Apesteguia., 2005. Am. Mus. Novit., 3490, 1–38. **22.** corydo, 83 x 68: Corydoradine fishes. Britto, M., 2003. Proc. Acad. Nat. Sci. Philadelphia 153, 119154. **23.** cristian2, 128 x 64: *Liolaemus* lizards. Unpublished (C. Abdala, Ph.D. Thesis). **24.** crust, 352 x 68: Crustaceans and other arthropods. Giribet et al., 2005. Crustacean Issues 16, 307–52. **25.** das, 60 x 85: *Dasybasis* (Tabanidae). Unpublished (González et al.). **26.** dinos, 276 x 50: Prosauropods. Unpublished (D. Pol, Ph.D. Thesis). **27.** diony, 381 x 145: Dionychan spiders. Unpublished (M. Ramírez). **28.** dro, 217 x 159: Drosophilid flies. Grimaldi, 1990. Bull. Am. Mus. Nat. Hist. 197, 1–139. **29.** embia, 186 x 157: Embiopterans. Unpublished (C. Szumik). **30.** entelo, 247 x 55: Entelogyne spiders. Griswold et al., 2005. Proc. Calif. Acad. Sci. 4th Ser. 56, Suppl. II:1–324. **31.** erigo, 176 x 82: Erigonid spiders. Miller, J., Hormiga, G., 2004. Cladistics 20, 385–442. **32.** ethe, 51 x 58: Iguanid lizards. Etheridge & de Queiroz., 1988. Stanford University Press. **33.** fannia, 157 x 83: Muscoid flies (genus *Fannia*). Unpublished (C. Domínguez, Ph.D. Thesis). **34.** firefly, 100 x 96: Branham, M.A., Wenzel, J.W, 2003. Cladistics 19, 1–22. **35.** gig_nw, 71 x 66: Genus *Gidantodax* (Simuliidae, Diptera). Pinto Sanchez et al., 2005. Insect Syst. Evol. 36, 219–240. **36.** gui_m, 76 x 55: Tingid heteropterans. Guilbert, E., 2001. Zool. Scr. 30, 313–324. **37.** holmorph, 176 x 85: Holometabolous insects. Whiting, M. et al., 1997. Syst. Biol. 46, 1–68. **38.** hymen, 169 x 77: Hymenopteran families. Ronquist, F. et al., 1999. Zool. Scr. 28, 13–50. **39.** kearney, 162 x 80: Amphisbaenians. Kearney, M., 2003. Herpet. Monogr. 17, 1–74. **40.** liebherr, 206 x 170: Platynine carabids. Liebherr & Zimmerman., 1998. Syst. Entomol. 23, 137–172. **41.** lobo3, 45 x 76: *Liolaemus* lizards. Unpublished (F. Lobo). **42.** lorica, 215 x 128: Loricariid fishes. Armbruster, J., 2004. Zool. J. Linn. Soc. 141, 1–80. **43.** ltbees, 131 x 83: Long-tongued bees. RoigAlsina, A., Michener, C.D., 1993. Univ. Kansas Sci. Bull. 55, 123–162. **44.** lucena, 119 x 66 Characid fishes, unpublished (Carlos A.S. Lucena, Ph.D. Thesis). **45.** lucho3, 139 x 83: *Trichomycterus* fishes and related genera. Unpublished (Luis Fernandez, pers. comm.). **46.** lycos, 147 x 98: Ctenid spiders and relatives. Unpublished (an earlier version of Silva–Dávila, D., 2003. Bull. Am. Mus. Nat. Hist. 274, 1–86). **47.** mammals, 319 x 90: Tetrapods. Ruta et al., 2003. Biol. Rev. 78, 251–345. **48.** marcos2, 370 x 91: Characid fishes. Unpublished (M. Mirande). **49.** mischo, 60 x 73: *Mischocyttarus* wasps. Unpublished (O. Silveira, Ph.D. Thesis). **50.** mitt, 159 x 78: Chrysomelid beetles. From Platnick, N., 1989, Cladistics 5, 145–161. **51.** molina, 123 x 73: Leptohyphid mayflies. Unpublished (Molineri, Ph.D. Thesis, an earlier version of Molineri, C., 2006. Syst. Entomol. 31, 711). **52.** morph, 252 x 117: Hexapod orders. Wheeler, W. et al., 2001. Cladistics 17, 113–169. **53.** nixseed, 103 x 49: Seed plants. Nixon et al., 1994. Ann. Mo. Bot. Gard. 81, 484–533. **54.** norell, 222 x 56: Troodontid dinosaurs. Xu & Norell., 2004. Nature 431, 838–841. **55.** nsfmorph, 31 x 51: Polistes wasps. Unpublished

(Pickett et al.). **56.** odonata, 132 x 121: Dragonflies. An enlarged version of Rehn, A., 2003. Syst. Entomol. **57.** parambly, 132 x 92: *Paramblynotus* wasps. Liu, Z. et al. in press, Bull. Am. Mus. Nat. Hist. **58.** pilo, 149 x 113: Pilophorine hemipterans: Schuh, R., 1991. Cladistics 7, 157–189. **59.** po, 95 x 68: Polistine wasps: Arevalo, E. et al., 2004. BioMed Central Evol. Biol. 4, 8. **60.** prendi, 115 x 71: Scorpion genera: Unpublished (L. Prendini, with duplicates removed). **61.** pulawski, 74 x 135: Species of *Tachysphex* (Sphecidae). Unpublished (W. Pulawski, with duplicates removed). **62.** realdata, 124 x 90: Vespid wasps. Unpublished (Carpenter et al.). **63.** ropa, 95 x 106: *Ropalidia* wasps. Unpublished (Kojima and Carpenter). **64.** sch, 75 x 76: Phylinae bugs (Hemiptera). Schuh, R., 1984. Bull. Am. Mus. Nat. Hist. 177, 1–476. **65.** tab_m, 96 x 65: Tabanids (Diptera). Unpublished (Coscarón and Miranda–Esquivel.). **66.** tenu, 262 x 56: Tenuipalpid mites. QuirozGonzales (Ph.D. Thesis), in Platnick, N., 1989. Cladistics 5, 145–161. **67.** tetrao, 219 x 58: Tetraodontiform fishes. Santini & Tyler, 1999. Am. Zool. 39, 10. **68.** total, 104 x 84: Nemesiid spiders. Goloboff, P., 1995. Bull. Am. Mus. Nat. Hist. 224, 1–189. **69.** virg7, 93 x 75: Lizards (muscles). Unpublished (V. Abdala). **70.** west, 73 x 66: Legumes. Crisp & Weston, 1987. Adv. Legume Syst., Part 3, R. Bot. Gard. Kew, table 4. **71.** Agnolin_2011, 368 x 88: Agnolin, F.L., Novas, F.E., 2011. An. Acad. Bras. Cienc. 83, 117–162. **72.** Agnolin_2012, 423 x 103: Agnolin, F.L., Powell, J.E., Novas, F.E., Kundrát, M., 2012. Cretac. Res. 1–24. **73.** Andres_2013, 154 x 109: Andres, B.B., Myers, T.S., 2013. Earth Environ. Sci. Trans. R. Soc. Edinb. 103, 383–398. **74.** Apaldetti_2011, 361 x 54: Apaldetti, C., 2011. PLoS One. 6, 1–19. **75.** Arbo_2009, 48 x 99: Arbo, M., Espert, S., 2009. Taxon 58, 457–467. **76.** Averianov_2010, 353 x 62: Averianov, A.O., Krasnolutskii, S.A., Ivantsov, S.V., 2010. Proc. Zool. Inst. Russ. Akad. Sci. 314, 42–57. **77.** Barret_2011, 137 x 62: Barrett, P.M., Butler, R.J., 2011. Pap. Paleontol. 86, 131–163. **78.** Bittencourt_2009, 224 x 54: Bittencourt, J.S., Kellner, A.W., 2009. Zootaxa, 2079, 1–56. **79.** Brower_2014, 353 x 88: Brower, A.V., Willmott, K.R., Silva–brandão, K.L., Garzón–orduña, I.J., Freitas, A.V., 2014. Syst. Biodivers. 12, 133–147. **80.** Brusatte_2010, 187 x 55: Brusatte, S.L., Benton, M.J., Desojo, J.B., Langer, M.C., 2010. J. Syst. Palaeontol. 8, 3–47. **81.** Buenaventura_2015, 115 x 93: Buenaventura, E., Pape, T., 2015. Org. Divers. Evol. 15, 1–31. **82.** Butler_2012, 227 x 50: Butler, R.J., Galton, P.M., Porro, L.B., Chiappe, L.M., Henderson, D.M., Erickson, G.M., 2012. Proc. R. Soc. B 277, 375–381. **83.** Butler_2014, 413 x 80: Butler, R., Sullivan, C., Ezcurra, M.D., Liu, J., Lecuona, A., Sookias, R., 2014. BMC Evol. Biol. 14, 128. **84.** Cardoso_2012, 55 x 104: Cardoso, D.B., Cavalcante, de lima H., Schütz, rodrigues R., Queiroz, L.P., Pennington R., Lavin, M., 2012. Taxon. 61, 1057–1073. **85.** Carrano_2012, 351 x 61: Carrano, M.T., Benson, R.B., Sampson, S.D., 2012. J. Syst. Palaeontol. 10, 211–300. **86.** Chemisquy_2010, 52 x 50: Chemisquy, M., Giussani, L., Scataglini, M., Kellogg, E., Morrone, O., 2010. Ann. Bot. 106, 107–130. **87.** Choiniere_2010, 421 x 99: Choiniere, J.N., Xu, X., Clark, J.M., Forster, C.A., Guo, Y., Han, F., 2010. Sci. 327, 571–574. **88.** Choiniere_2010_2, 472 x 60: Choiniere, J.N., Clark, J.M., Forster, C.A., Xu, X., 2010. J. Vertebr. Paleontol. 30, 1773–1796. **89.** Clement_2009, 81 x 94: Clement, W., Weiblen, G., Weiblen, G., 2009. Syst. Bot. 34, 530–552. **90.** Cohen_2011, 22 x 67: Cohen, J.I., 2011. Cladistics. 27, 559–580. **91.** Csiki_2010, 251 x 70: Csiki, Z., Vremir, M., Brusatte, S.L., Norell, M.A., 2010. Can. J. Earth. Sci. 47, 1507–1517. **92.** Damgaard_2012, 64 x 76: Damgaard, J., 2008. Insect. Syst. Evol. 39, 431–460. **93.** Daqing_2010, 296 x 72: Daqing, L., Norell, M.A., Gao, K., Smith, N.D., Makovicky, P.J., 2010. Proc. R. Soc. B 277, 183–190. **94.** Davalos_2012, 150 x 63: Davalos, L.M., Cirranello, A.L., Geisler, J., Simmons, N., 2012. Biol. Rev. Camb. Philos. Soc. 87, 991–1024. **95.** Davalos_2012_2, 220 x 80: Davalos, L.M., Cirranello, A.L., Geisler, J., Simmons, N., 2012. Biol. Rev. Camb. Philos. Soc. 87, 991–1024. **96.** Davalos_2014, 278 x 113: Davalos, L.M., Velazco, P.M., Warsi, O., Smits, P.D., Simmons, N., 2014. Syst. Biol. 63, 582–601. **97.** DeWit_2011, 54 x 80: De wit, P., Rota, E., Erseus, C., 2011. Zool. Scr. 40, 509–519. **98.** Dikow_2009_2, 220 x 88: Dikow, T., 2009. Bull. Am. Mus. Nat. Hist. 319, 1–175. **99.** Dongyu_2010, 363 x 87: Dongyu, H., Lianhai, H., Lijun, Z., Xing, X., 2009. Nat. 461, 640–643. **100.** Ezcurra_2010, 378 x 50: Ezcurra, M.D., 2010. J. Syst. Palaeontol. 8, 371–425. **101.** Freire_2014, 36 x 58: Freire, S.E., Chemisquy, M.A., Anderberg, A., Beck, S., Meneses, R.I., Loeuille, B.F., Urtubey, E., 2014. Plant. Syst. Evol. 301, 1227–1248. **102.** Frick_2010, 176 x 111: Frick, H., Nentwig, W., Kropf, C., 2010. Org. Divers. Evol. 10, 297–310. **103.** Gernandt_2010, 54 x 50: Gernandt, D.S., León–Gomez, C., Hernández–León, S., Olson, M.E., 2011. Syst. Bot. 36, 583–594. **104.** Gonzalez_2011, 453 x 104: Gonzalez–Marquez, M.E., Ramirez, M.J., 2011. Zootaxa., 3201, 1–26. **105.** Gonzalez_2013, 97 x 105: Gonzalez, V.H., Griswold, T.L., 2013. Zool. J. Linn. Soc. 168, 221–425. **106.** Huang_2011, 50 x 78: Huang, D., Fitzhugh, K., Rouse, G., 2011. Cladistics. 27, 356–379. **107.** James_2009, 242 x 79: James, F.C., Pourtless, J.A., 2009. Ornithol. Monogr. 66, 1–78. **108.** Jansa_2008, 129 x 51: Voss, R., Jansa, S., 2009. Bull. Am. Mus. Nat. Hist. 322, 1–178. **109.** Ksepka_2012, 245 x 71: Ksepka, D., Fordyce, R., Ando, Y., Jones, C.M., 2012. J. Vertebr. Paleontol. 32, 235–254. **110.** Laborda_2013, 447 x 113: Laborda, A., Ramírez, M.J., Pizarro–Araya, J., 2013. Zootaxa 3731, 133–152. **111.** Lambkin_2011, 207 x 78: Lambkin, C.L., Bartlett, J.S., 2011. Zookeys 150, 231–280. **112.** Makovicky_2010, 296 x 72: Makovicky, P.J., Li, D., Gao, K., Lewin, M., Erickson, G.M., Norell, M.A., 2010. Proc. R. Soc. B. 277, 191–198. **113.** Makovicky_2011, 227 x 52: Makovicky, P.J., Kilbourne, B.M., Sadlier, R.W., Norell, M.A., 2011. J. Vertebr. Paleontol. 31, 626–640. **114.** Mao_2012, 53 x 51: Mao, K., Milne, R., Zhang, L., Peng, Y., Liu, J., Thomas, P., Mill, R.R., Renner, S.S., 2012. Proc. Natl. Acad. Sci. U.S.A. 109, 7793–7798. **115.** McDonald_2010, 131 x 62: McDonald, A.T., Barrett, P.M., Chapman, S.D., 2010. Zootaxa, 2569, 1–43. **116.** McDonald_2011, 131 x 61: Mcdonald A.T., 2011. Zootaxa, 2783, 52–68. **117.** McKenna_2010, 56 x 58: McKenna, M.J., Simmons, M.P., Bacon, C.D., Lombardi, J., 2011. Syst. Bot. 36, 922–932. **118.** Naish_2012, 1025 x 72: Naish, D., Dyke, G., Cau, A., Escuillié, F., 2012. Biol. Lett. 8, 97–100. **119.** Nesbitt_2011, 412 x 82: Nesbitt, S.J., 2011. Bull. Am. Mus. Nat. Hist. 352, 1–292. **120.** Nesbitt_2011_2, 256 x 71: Nesbitt, S.J., Clarke, J.A., Turner, A.H., Norell, M.A., 2011. J. Vertebr. Paleontol. 31, 144–153. **121.** Ni_2013, 1844 x 108: Ni, X., Gebo, D.L., Dagosto, M., Meng, J., Tafforeau, P., Flynn, J.J., Beard, K.C., 2013. Nature 498, 60–64. **122.** Novas_2011, 380 x 56: Novas, F.E., Ezcurra, M.D., Chatterjee, S., Kutty, T.S., 2011. Earth Environ. Sci. Trans. R. Soc. Edinb. 101, 333–349. **123.** O'connor_2011, 245 x 54: O'connor, J.K., Chiappe, L.M., Bell, A., 2011. Wiley–Blackwell, Oxford, 39–114. **124.** Pol_2011, 277 x 50: Pol, D., Garrido, A.C., Cerda, I., 2011. PLoS One 6, 1–24. **125.** Pol_2011_2, 230 x 51: Pol, D., Rauhut, O.W., Becerra, M., 2011. Naturwissenschaften 98, 369–379. **126.** Price_2013, 131 x 119: Price, S., Powell, S., Kronauer, D.J., Tran, L.A., Pierce, N., Wayne, R., 2014. J. Evol. Biol. 27, 242–258. **127.** Price_2013_2, 131 x 82: Price, S., Powell, S., Kronauer, D.J., Tran, L.A., Pierce, N., Wayne, R., 2013. J. Evol. Biol. 27, 242–258. **128.** Prieto_2009, 299 x 52: Prieto–Márquez, A., Wagner, J.R., 2009. Cret. Res. 30, 1238–1246. **129.** Prieto_2010, 370 x 53: Prieto–Márquez A., 2010. Zootaxa 2452, 1–17. **130.** Prieto_2011, 289 x 51: Prieto–Márquez A., Wagner, J.R., 2013. Acta. Palaeontol. Pol. 58, 255–268. **131.** Pyron_2011, 161 x 76: Pyron, R.A., 2011. Syst. Biol. 60, 466–481. **132.** Ramirez_2014_2, 82 x 52: Ramírez, M.J., Grismado, C.J., Labarque, F.M., Izquierdo, M.A., Ledford, J., Miller, J., Haddad, C.R., Griswold, C., 2014. Zool. Anzeig. 253, 382–393. **133.** Ronquist_2012, 353 x 113: Ronquist, F., Klopfstein, S., Lars, V., Schulmeister, S., Murray, D., Rasnitsyn, A.P., 2012. Syst. Biol. 61, 973–999. **134.** Rowe_2011, 361 x 51: Rowe, T.B., Sues, H., Reisz, R., 2011. Proc. R. Soc. B 278, 1044–1053. **135.** Scataglini_2013, 57 x 132: Scataglini, M., Zuloaga, F., 2013. Syst. Bot. 38, 1076–1086. **136.** Senter_2010, 364 x 89: Senter, P., 2010. J. Evol. Biol. 23, 1732–1743. **137.** Senter_2012, 394 x 109:

Senter, P., Kirkland, J.I., Deblieux, D.D., Madsen, S., Toth, N., 2012. PLoS One 7, 1–20. **138.** Sertich_2010, 361 x 50: Sertich, J.J., Loewen, M.A., 2010. PLoS One 5, 1–17. **139.** Sharkey_2011, 392 x 105: Sharkey, M., Carpenter, J., Vilhelmsen, L., Heraty, J., Hawks, D., Dowling, A.P., Schulmeister, S., Murray, D., Deans, A., Ronquist, F., Krogmann, L., Wheeler, W., 2011. Cladistics 28, 80–112. **140.** Soto_2012, 445 x 108: Soto, E.M., Ramírez, M.J., 2012. Zootaxa 3443, 1–65. **141.** Sundue_2010, 109 x 129: Sundue, M.A., 2010. Syst. Bot. 35, 716–729. **142.** Syme_2011, 69 x 149: Syme, A.E., Oakley, T.H., 2011. Syst. Biol. 61, 314–336. **143.** Thode_2013, 209 x 97: Thode, V.A., O'Leary, N., Olmstead, R., Freitas, L.B., 2013. Syst. Bota. 38, 805–817. **144.** Thompson_2011, 170 x 51: Thompson, R.S., Parish, J.C., Maidment, S.C., Barrett, P.M., 2012. J. Syst. Palaeontol. 10, 301–312. **145.** Thulin_2013, 55 x 77: Thulin, M., Phillipson, P., Lavin, M., 2013. Adansonia, ser. 3. 35, 61–71. **146.** Tippery_2011, 28 x 54: Tippery, N., Les, D., 2011. Syst. Bot. 36, 1101–1113. **147.** Werenkraut_2009, 444 x 101: Werenkraut, V., Ramírez, M., 2009. Zootaxa., 2212, 1–40. **148.** Whitmore_2013, 84 x 88: Whitmore, D., Pape, T., Cerretti, P., 2013. Zool. J. Linn. Soc. 169, 604–639. **149.** Wiens_2010, 363 x 64: Wiens, J.J., Kuczynski, C.A., Townsend, T., Reeder, T., Mulcahy, D., Sites, J.W., 2010. Syst. Biol. 59, 674–688. **150.** Wilson_2015, 423 x 68: Wilson, G.D., Shaik, S., Ranga–Reddy Y., 2015. J. Crustac. Biol. 35, 216–240. **151.** Wolfe_2013, 146 x 148: Wolfe, J.M., Hegna, T.A., 2013. Cladistics 30, 366–390. **152.** Xing_2012, 334 x 53: Xing, H., Prieto–Márquez, A., Gu, W., Yu T., 2012. Vertebr. PalAsiatica 50, 160–169. **153.** Xu_2011, 474 x 92: Xu, X., You, H., Du, K., Han, F., 2011. Nat. 475, 465–470. **154.** Xu_2012, 363 x 88: Xu, X., Sullivan, C., Tan, Q., Sander, M., Ma, Q., 2012. Vertebrata. PalAsiatica 50. 140–150. **155.** Yi_2013, 435 x 86: Yi, H., Norell, M.A., 2013. Am. Mus. Novit., 3767, 1–31. **156.** Zanno_2010, 348 x 76: Zanno, L.E., 2010. J. Syst. Palaeontol. 8, 37–41. **157.** Zanol_2012, 230 x 82: Zanol, J., Fauchald, K., Paiva, P.C., 2007. Zool. J. Linn. Soc. 150, 413–434. **158.** Zanol_2013, 213 x 82: Zanol, J., Halanych, K., Fauchald, K., 2013. Zool. Scr. 43, 79–100.