Cladistics 26 (2010) 86–97

Cladistics

10.1111/j.1096-0031.2009.00269.x

Behaviour of resampling methods under different weighting schemes, measures and variable resampling strengths

Cecilia Kopuchian[1]* and Martín J. Ramírez[2]

[1]División Ornitología and [2]Division Aracnología, CONICET, Museo Argentino de Ciencias Naturales "Bernardino Rivadavia", Av. Ángel Gallardo 470, C1405DJR, Buenos Aires, Argentina

Accepted 14 May 2009

Abstract

We compared general behaviour trends of resampling methods (bootstrap, bootstrap with Poisson distribution, jackknife, and jackknife with symmetric resampling) and different ways to summarize the results for resampling (absolute frequency, F, and frequency difference, GC′) for real data sets under variable resampling strengths in three weighting schemes. We propose an equivalence between bootstrap and jackknife in order to make bootstrap variable across different resampling strengths. Specifically, for each method we evaluated the number of spurious groups (groups not present in the strict consensus of the unaltered data set), of real groups, and of inconsistencies in ranking of groups under variable resampling strengths. We found that GC′ always generated more spurious groups and recovered more groups than F. Bootstrap methods generated more spurious groups than jackknife methods; and jackknife is the method that recovered more real groups. We consistently obtained a higher proportion of spurious groups for GC′ than for F; and for bootstrap than for jackknife. Finally, we evaluated the ranking of groups under variable resampling strengths qualitatively in the trajectories of "support" against resampling strength, and quantitatively with Kendall coefficient values. We found fewer ranking inconsistencies for GC′ than for F, and for bootstrap than for jackknife.
© The Willi Hennig Society 2009.

Background: resampling methods and measures

When we make a phylogenetic reconstruction, we want to know how well supported the groups we have obtained are. For this purpose, there are methods that assess the support of the groups obtained. Two resampling methods used in assessing group support in a cladogram are jackknife and bootstrap. Both are based in matrix data perturbation by means of random resampling of all or part of the characters, replicated a number of times. The frequency with which a given group is found in the trees obtained from these resampled matrices is used as a measure of its support.

The bootstrap method (BT) was proposed by Efron (1979) as a general-purpose statistical tool, and was first applied to phylogenetics by Felsenstein (1985) to place confidence intervals on phylogenies. Nowadays, BT is widely used by systematists to test the level of support of individual nodes in a phylogenetic tree. As noted by Goloboff et al. (2003a), several interpretations of BT have been advanced by different authors (Efron, 1979; Felsenstein, 1985; Berry and Gascuel, 1996). A common interpretation is that BT measures the probability of recovering a given group if a data set for the same organisms is to be sampled again from scratch. In this way, BT is interpreted as a measure of stability under specific circumstances. However, in this paper we have used this method as a measure of support instead of a measure of stability (for a discussion see Goloboff et al., 2003a; Grant and Kluge, 2003; Ramírez, 2005).

The bootstrap method consists of obtaining new data sets from the original one by means of random resampling with replacement. The final size of the resampled data set is set to be the same as in the original data set, but some characters will not be represented and others will be represented more than once. The most parsimonious cladograms are then

*Corresponding author:
E-mail address: ckopuchian@macn.gov.ar

searched for each pseudoreplicate, and represented in each case by their strict consensus (as implemented in TNT) (De Laet et al., 2004). The process is repeated many times (1000, for example) and the frequency with which a group is found in the trees obtained from the pseudoreplicates is interpreted as a measure of the confidence level (Felsenstein, 1985) or support (Farris et al., 1996; Goloboff et al., 2003a; Ramírez, 2005) of that group. Well supported groups (favoured by many characters) are frequently recovered, and poorly supported groups are represented few times.

Bootstrap frequencies for a group decrease with the addition of informative characters that are compatible but not informative for that group (Faith and Cranston, 1991), autapomorphies (Carpenter, 1992), or invariant characters (Kluge and Wolf, 1993; Harshman, 1994). The latter author argued that bootstrapping is only slightly affected by the inclusion of irrelevant characters (see also Felsenstein, 2004, p. 344); however, he proposed that anyone concerned about the variation in BT values due to irrelevant characters can equalize the effect of irrelevant characters on all nodes by adding a large number of invariant characters to the matrix. This correction is equivalent to using a character weight distribution with a Poisson distribution of mean one (BP) (Farris, 1999; Goloboff et al., 2003a) and the problem of having non-informative characters disappears.

In jackknife resampling (JD), each character in the original data set has a probability of being deleted (without replacement), hence the pseudoreplicate data sets are smaller than the original. Farris et al. (1996) proposed a probability of deletion of $e^{-1} \approx 0.36$ to produce JD group frequencies comparable to those obtained under BT.

In order to avoid the distortion of frequencies under BT or JD (either under- or overestimations of the actual group support) that may occur under heterogeneous prior weights or state transformation costs, Goloboff et al. (2003a) have proposed symmetric resampling (JS) where each character has a probability $2P$ to be changed, and if changed, it can be duplicated or deleted with equal probability. JS has the advantage that it can be applied to any weighting scheme: successive weighting (Farris, 1969), implied weighting (Goloboff, 1993), or weighting of state transformations, including asymmetries in transformation costs.

All the methods mentioned above are expected to obtain meaningful frequencies above 50% which are summarized in a majority rule consensus tree. Below this threshold, frequencies may not be correlated with support, as in the example given in Goloboff et al. (2003a, fig. 7), where a supported group has lower frequency than a contradicted one. For this study we have computed group frequencies even below 50%, as shown in Figs 1–3.

Goloboff et al. (2003a) proposed that what actually measures the support is not the absolute frequency (F), but the difference in frequency between a group and its most frequent contradictory group (GC, for "group present/contradicted"). GC values of −1, 0 and 1 indicate maximum contradiction, indifference, and maximum support, respectively. GC is useful to measure strength of contradiction and to obtain support values for groups with positive but low support, which are otherwise not reported by methods using absolute frequencies (real groups with frequencies below 50% that are not retained in the majority consensus tree).

It is known that resampling methods may produce spurious results, especially in weakly supported groups. For example, a measure should never indicate a group contradicted by the data as "better supported" than a group supported by the data (Goloboff et al., 2003a; implicit in Farris et al., 1996), but this situation can occur in real data sets (for example, see the thicker broken line in Figs 2 and 3, and Appendix S1: Fig. S1). In this contribution we use the terms "real" for those groups present in the strict consensus of the unaltered data set, and "spurious" for the opposite situation. The most evident way to detect artefacts with support methods is by comparison with the strict consensus tree, looking for spurious groups appearing as supported, and real groups that appear to lack support.

In addition, the trajectories of group frequencies against resampling strength are also informative for detecting problems with the resampling methods (Goloboff et al., 2003a). If a group is better supported than another under a given resampling probability P, but the situation is reversed under other probability P', then the method is internally inconsistent (Ramírez, 2005). If the frequency is taken as an indication of support, the trajectories should not cross each other in the range of reasonable probabilities of up- or down-weighting (e.g. $0 < P < 0.5$). When two trajectories intersect, it means that the ranking of the two groups is inverted before and after the point of intersection (for example see Fig. 2 and the three thicker profiles in Appendix S1: Fig. S2a,b), which means that the method produces different rankings at different resampling strengths (Ramírez, 2005). The crossing of trajectories occurs across all the examined range of resampling strengths, which suggests that the problem described here is intrinsic to the methods, rather than to a given resampling strength.

Objectives

Our aim was to compare the performance of resampling methods in real data sets, in terms of internal consistency. In the following experiments, negative indicators for a given method or measure are (i) spurious

(a)

```
  Dataset without conflict                                            Dataset with conflict
A 0000000000 000000000 00000000 0000000 000000 00000 0000 000 00 0   A 000000000 000000000 000000000 000000000
B 0000000000 000000000 00000000 0000000 000000 00000 0000 000 00 0   B 000000001 000000000 000000000 000000000
C 1111111111 000000000 00000000 0000000 000000 00000 0000 000 00 0   C 111111111 000000000 000000000 000000000
D 1111111111 111111111 00000000 0000000 000000 00000 0000 000 00 0   D 111111110 000000000 000000000 000000000
E 1111111111 111111111 11111111 0000000 000000 00000 0000 000 00 0   E 000000000 000000011 000000000 000000000
F 1111111111 111111111 11111111 1111111 000000 00000 0000 000 00 0   F 000000000 111111111 000000000 000000000
G 1111111111 111111111 11111111 1111111 111111 00000 0000 000 00 0   G 000000000 111111100 000000000 000000000
H 1111111111 111111111 11111111 1111111 111111 11111 0000 000 00 0   H 000000000 000000000 000000111 000000000
I 1111111111 111111111 11111111 1111111 111111 11111 1111 000 00 0   I 000000000 000000000 111111111 000000000
J 1111111111 111111111 11111111 1111111 111111 11111 1111 111 00 0   J 000000000 000000000 111111000 000000000
K 1111111111 111111111 11111111 1111111 111111 11111 1111 111 11 0   K 000000000 000000000 000000000 000001111
L 1111111111 111111111 11111111 1111111 111111 11111 1111 111 11 1   L 000000000 000000000 000000000 111111111
M 1111111111 111111111 11111111 1111111 111111 11111 1111 111 11 1   M 000000000 000000000 000000000 111110000
```

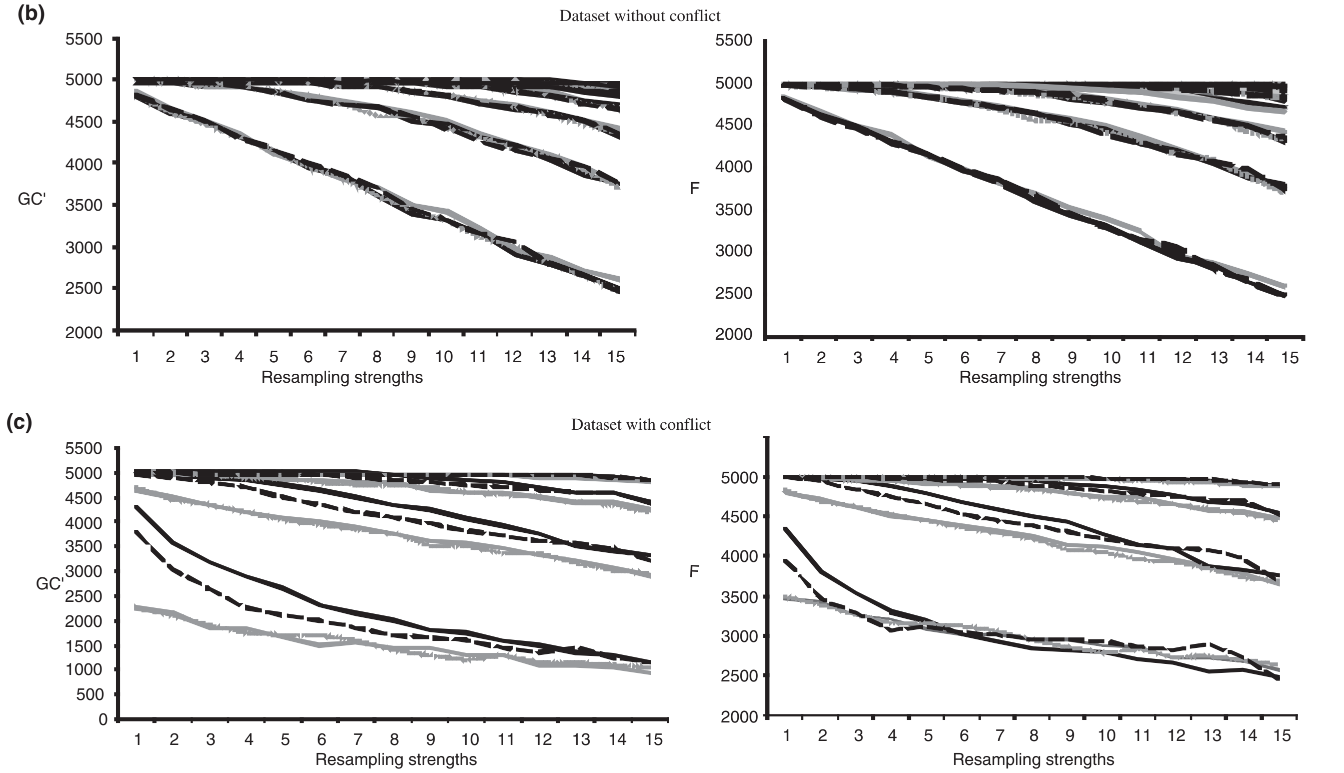

Fig. 1. (a) Hypothetical data sets without (left) and with (right) conflict. (b) Comparison between methods: bootstrap (grey line), bootstrap with Poisson distribution (grey broken line), jackknife (black line), and symmetric resampling (black broken line), using frequency trajectories against resampling strengths, in case of data without conflict. (c) Same as (b), for data with conflict.

groups reported as supported; (ii) inversion of rankings of group support under variable resampling strengths; and (iii) real groups reported as unsupported or contradicted. We evaluated these parameters by tracing group frequencies over a range of resampling strengths.

Because JS was proposed to correct biases under differential costs schemes, we also compared the performance of methods and measures under implied weights and differential transformation costs between states.

In this work we focus on the groups present in the strict consensus of the unaltered data set, the reference against which the spurious results are defined. We do not address here other interesting aspects related to support measures, such as the detection of secondary signals (e.g. Baker and DeSalle, 1997; Gatesy et al., 1999), or comparisons between resampling methods and Bremer support (Ramírez, 2005).

Methods

We analysed one hypothetical and 10 molecular data sets deposited in TREEBASE (http://www.treebase.org) or obtained from the authors (Lijtmaer et al., 2004; Ramírez, 2005) (see Appendix 1).

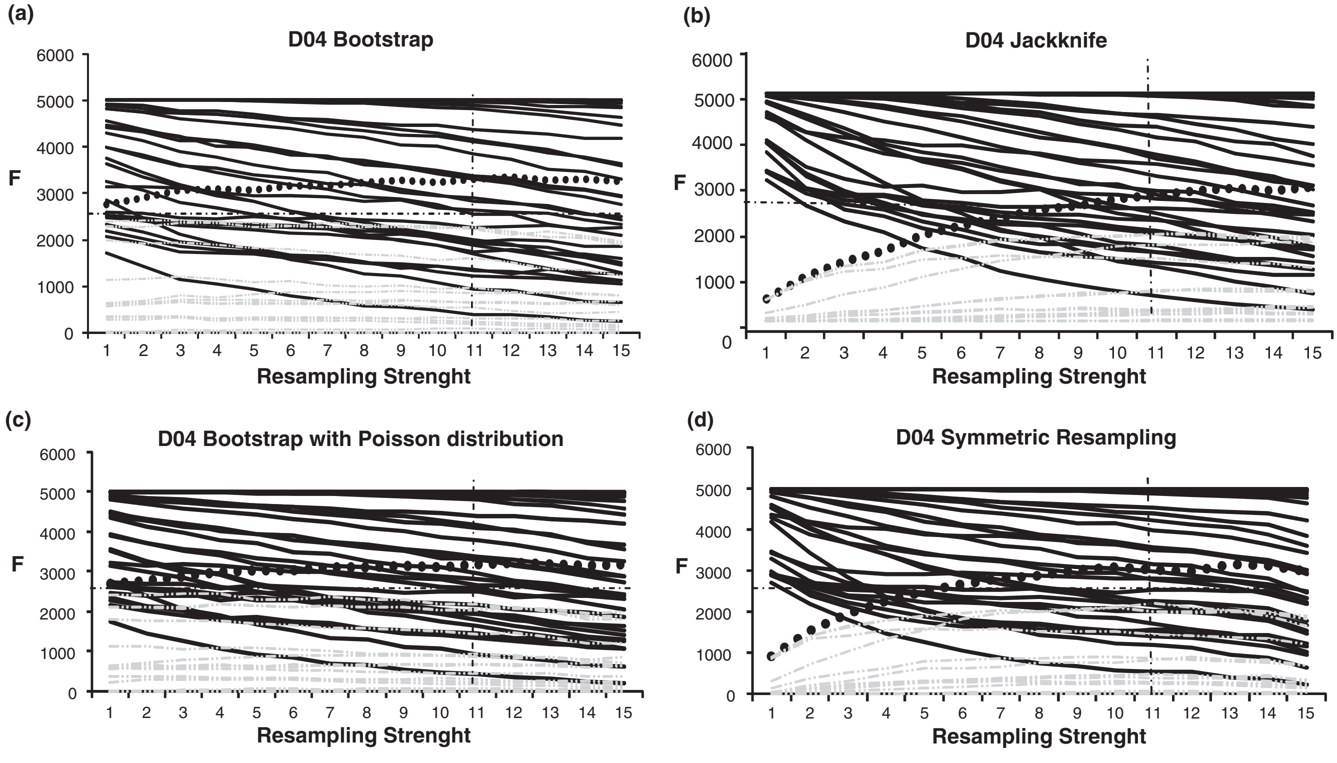

Fig. 2. Profiles of absolute frequency values as a function of different resampling strengths ($0 \le P \le 0.5$) with variable bootstrap (a), jackknife (b), variable bootstrap with Poisson distribution (c), and symmetric resampling (d). Black continuous lines represent real groups; grey broken lines represent spurious groups. The thicker black broken line is an example of a spurious group with $F > 50\%$ under some resampling strengths. Some groups contradicted by the data are indicated as "better supported" than groups with positive support under some resampling strengths. Horizontal broken line indicates the 50% frequency value; vertical broken line resampling strength 11, the one commonly used (see text for details). Data set D04, equal weights.

In order to examine a wide spectrum of cases, we have chosen data sets with variable numbers of taxa (from nine species in D01 and D02 to 85 in D06) and characters (from 24 characters in D03 to 1785 in D07). The data sets come from very different groups: birds (*Sporophila* and Stercorariidae), primates, fungi (*Peziza* and *Ophiocera*), insects (Holometabola and cynipid gall wasps), spiders (Ctenidae), plants (*Robinsonia*), and hypothetical data. Molecular data sets include both coding and non-coding regions from ribosomal, nuclear, and mitochondrial markers. We have also chosen data sets with different degrees of conflict. For example, D01 and D02 have the same number of taxa and a similar number of characters, but there is more character conflict in D02 than in D01.

We analysed the data sets under equal weights (EqWe), implied weights (ImWe) and, in molecular data sets, different state transformation costs, with transition:transversion cost 1:10 (T_i:T_v). For ImWe we used a constant of concavity $k = 8$, which produces a mild weighting function. We were not able to analyse the largest data set D06 for differential costs or ImWe because the search is much more time-consuming under these conditions.

We calculated both absolute frequency (F) and GC′, which is the difference between the frequencies obtained for a group and its most frequent contradictory group among the strict consensus trees of the pseudoreplicate analyses (rather than the most frequent contradictory group considering all the most parsimonious trees of the resampled data set, as in the actual GC) (Goloboff et al., 2003a). GC′ is an estimation of the actual GC value, and it was used here because it is much more easily obtained and is already implemented in TNT. Supplementary data for this paper can be found in Appendix S1.

Variable bootstrap

In JD and JS analyses, we used deletion probabilities (P) between 0.033 and 0.5 (Table 1). $P = 0$ is the unaltered data set.

In data sets without conflict, the frequency for a group G with a character deletion probability P in JD and JS, is:

$$\mathrm{JD}(\mathrm{G}, P) = 1 - P^r$$

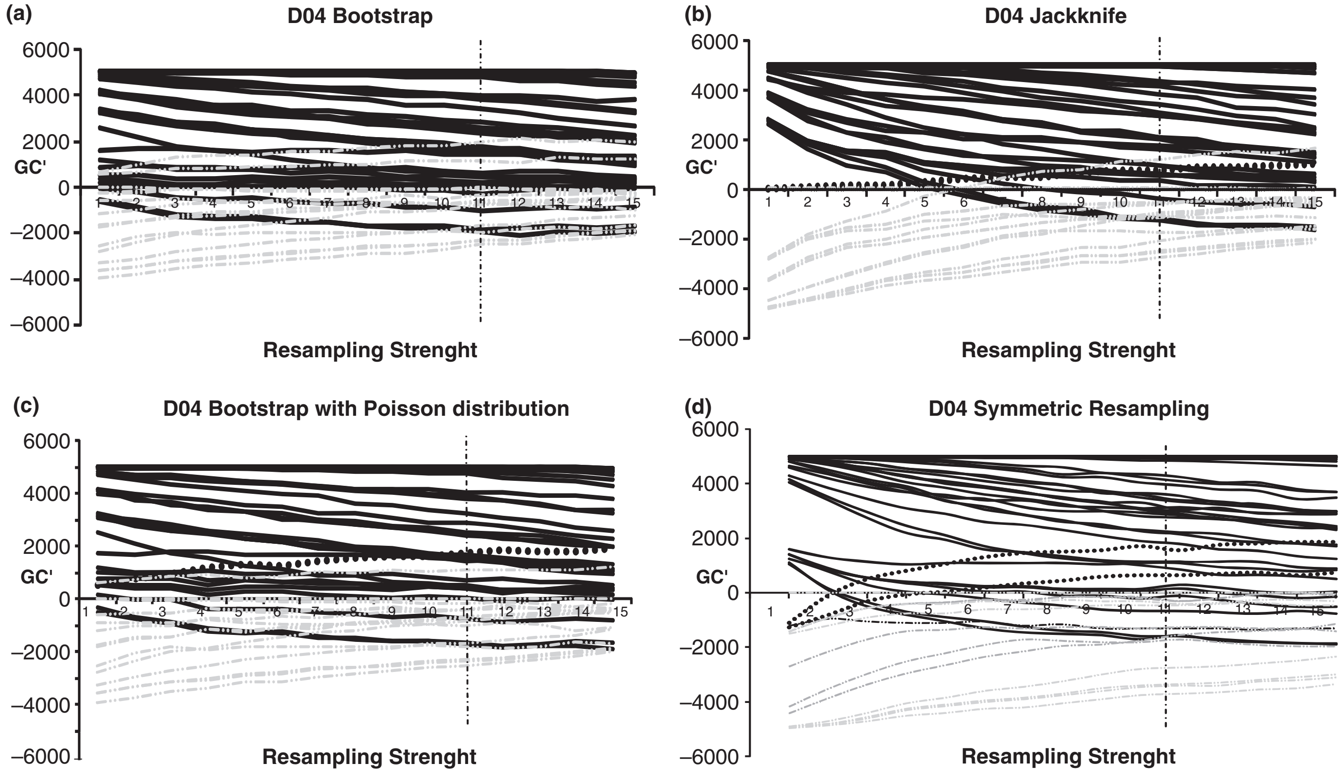

Fig. 3. Profiles of GC′ values as a function of different resampling strengths (0 ≤ P ≤ 0.5) with variable bootstrap (a), jackknife (b), variable bootstrap with Poisson distribution (c), and symmetric resampling (d). Black continuous lines represent groups with actual support; grey broken lines represent spurious groups. The thicker broken line is an example of a spurious group with a positive GC′ value under some resampling strengths. Data set D04, equal weights. See Fig. 2 for conventions.

where r is the number of uncontradicted characters supporting group G (Farris et al., 1996).

In BT, the probability of resampling a given character is $1/n$, where n is the total number of characters in the data set. If there were r characters supporting a given group G, the probability of resampling one of those characters is r/n. The probability of resampling a character not relevant for the group is $1 - r/n$. For the group G being absent in the tree, neither of the r characters must be resampled, which occurs with a probability $(1 - r/n)^n$.

Then the BT frequency of group G will be:

$$\mathrm{BT(G)} = 1 - (1 - r/n)^n \quad \text{(Harshman, 1994)}.$$

In order to compare BT with JD and JS through a range of resampling strengths, we calculated a variable bootstrap (BT*), which consists of a BT with variable final data set size. In this method, the BT frequency is:

$$\mathrm{BT^*(G,}m) = 1 - (1 - r/n)^{mn} \qquad (1)$$

where mn is the final size of the resampled data set. As n increases, BT and BT* very quickly approach their limits, as:

$$\mathrm{BT} = 1 - e^{-r}$$

$$\mathrm{BT^*} = 1 - e^{-mr} \qquad (2)$$

We used this limit (Equation 2), which is simpler to compute than Equation 1. For $n = 30$ the differences in BT values between Equations 2 and 1 are in the third decimal.

Table 1
Jackknife–bootstrap resampling strength equivalencies

Resampling strength	1	2	3	4	5	6	7	8	9	10	**11**	12	13	14	15
P	0.033	0.066	0.1	0.13	0.16	0.2	0.23	0.26	0.3	0.33	**0.36**	0.4	0.43	0.46	0.5
m	3.40	2.71	2.30	2.01	1.79	1.61	1.45	1.32	1.20	1.09	**1.00**	0.91	0.83	0.76	0.69

Resampling strength 11 (bold) corresponds to the values commonly used in jackknife and bootstrap.

P, Character deletion probability in jackknife; m, factor defining the “size” of the resampled data set. See text for details.

To make JD comparable with BT in the case of data sets without conflict:

$$\mathrm{BT}^* = \mathrm{JD}$$

$$1 - e^{-mr} = 1 - P^r$$

$$e^{-mr} = P^r$$

$$e^{-m} = P$$

$$\ln e^{-m} = \ln P$$

$$m = -\ln P \tag{3}$$

By using Equation 3 we calculated the values of m for the different resampling strengths under BT*, being in this way comparable with the resampled strength used in JD and JS (Table 1).

Note that our 11th resampled strength corresponds to $m = 1$, thus $mn = n$, as in traditional BT, where the final size of the resampled data set is the same as the size of the original data set (n) (Table 1). Similarly for JD, our 11th strength is $P = 0.3667$, close to the value established by Farris et al. (1996) ($P = 0.36$) to make it comparable with the traditional BT.

To obtain a variable bootstrap using a Poisson distribution (BP*) we simply assigned weights with a Poisson distribution with mean m (Appendix 2).

As BT* and BP* are novel ways to evaluate support by BT, we analysed two hypothetical data sets, one with conflict and the other without conflict (Fig. 1a) in order to check the equivalences between the four resampling methods. We obtained good correspondences between F and GC′ values in the case without conflict (Fig. 1b), but the equivalences are only approximate in the presence of some conflict (Fig. 1c).

This behaviour occurs in real data sets as well (Appendix S1: Fig. S3). Data sets D01 and D02 both had nine taxa and almost the same number of characters (486 and 446 characters, respectively). As there is low conflict in data set D01, the F and GC′ values are equivalent across the resampling methods analysed (Appendix S1: Fig. S3a), but this is not true for data set D02 (Appendix S1: Fig. S3b), where extensive conflict does exist in the data.

Analyses

For each resampling method (BT*, BP*, JD and JS) and weighting scheme considered (EqWe, $T_{\mathrm{i}}{:}T_{\mathrm{v}}$, and ImWe with $k = 8$), we used 15 resampling strengths. For each strength we made 5000 pseudoreplicates (except for large data sets D04 and D05 under $T_{\mathrm{i}}{:}T_{\mathrm{v}}$, only 1000) and assessed F and GC′ values (Figs 1–3).

For each pseudoreplicate we:

1. analysed one to five replicates of random addition sequence Wagner trees (RAS) followed by TBR swapping, with more replicates for larger data sets (Appendix 1)—in this way every pseudoreplicate obtained trees that were close to the optimum;
2. retained up to five different trees (different as collapsed with TBR) for each RAS + TBR obtained in (1) (we retained only the shortest trees); and
3. saved the strict consensus of each pseudoreplicate.

Upon completion of the 5000 pseudoreplicates for a given combination of measure, resampling method and weighting scheme, we:

4. saved the majority rule consensus tree and the frequency difference tree.

After all the combinations of resampling methods and weighting schemes were evaluated, we:

5. considered all the candidate groups appearing in at least one of the majority rule consensus trees of step (4), and for each group assigned a unique identifier number; and,
6. computed the support measures (GC′ and F) for each of the sets of 5000 trees saved in step (3).

All these analyses were made with TNT (Goloboff et al., 2003b) using custom scripts (available in Supporting information).

To evaluate differences in frequency rankings, we computed the Kendall coefficient (*Ke*) (Daniel, 1978) for each combination of data set, method, measure, and weighting scheme. Kendall coefficient values were calculated using SPSS 12.0 (http://www.spss.com/spss).

Sources of error

We identified several sources of error that might bias the results presented here, and present a description of how far we went in controlling them.

Aggressiveness of search routines

If the search algorithms applied to each pseudoreplicate are not exhaustive enough, the optimal trees will not be found, and the results may be biased in complex ways. Also, data sets producing large polytomies may be resolved in arbitrary ways if the tree buffer is not large enough. We have used search routines that very often reached optimal trees, but a small tree buffer that may over-resolve the consensus. For certain sensitive results we have increased the tree buffer size (see below) to check against such bias. For the most complex data sets, we invested about 50 h in completing 5000 pseudoreplicates using a personal computer.

Precision

We have used 5000 pseudoreplicates, but frequency functions built in TNT reports truncate percentages. For F we used our scripts to count exact frequencies. For GC′, which is much more complex to calculate, we relied on the truncate percentages built in TNT.

Sampling error

The standard deviations of frequencies follow a binomial distribution, with a standard deviation $s_P = (PQ/(n-1))^{0.5}$, where n is the number of pseudoreplicates, and P and Q are the frequency and its complement, respectively. For 5000 replicates, the worst-case scenario is $s_{50\%} = 0.7\%$. (For the two large data sets where we used 1000 replicates under T_i:T_v, $s_{50\%} = 1.6\%$.) Such a small error, combined with the fact that our conclusions are drawn not from one point estimation, but from a sequence of resampling strengths each with 5000 pseudoreplicates, implies that sampling error is adequately controlled in our experiments.

Estimation of frequency differences

The measure GC′ is an approximation of the frequency difference GC, using the consensus of each pseudoreplicate instead of all the possible resolutions (Goloboff et al., 2003a,b). The computation of GC is not available in current software. In summary, we have controlled several sources of error, except those from computational power for larger data sets (‘‘Aggressiveness of search routines’’ above) and implementation of algorithms in current software [‘‘Precision’’ (in part) and ‘‘Estimation of frequency differences’’]. It might happen that some of our results come from biased estimations, but even in that case we believe our results are useful. First, they can be used as an indication of methodological problems and so may lead to the development of more refined algorithms or implementations. Second, they show the performance of current algorithms and implementations, which are the only tools available for empirical systematists.

Recovered groups

As expected, GC′ recovered more groups than F, because the former recovers groups with low but actual support that are not recovered by computing F only.

We have analysed which method recovered more groups and which method recovered fewer groups in each of the weighting schemes. JD is the method that recovered more groups across all cases. BT*, and to a lesser degree BP*, are the methods that recovered fewer groups across all the resampling strengths analysed (Table 2). In general, JD recovered more groups under mild resampling strengths, switching to more groups for JS under stronger resampling strengths, but there is no clear association between the low resolution of BT* or BP* in recovering groups, and resampling strength. For details in different weighting schemes analysed see Table 2 and text in Supporting information.

Spurious groups

We calculated the number of spurious groups generated by each method/measure. GC′ always generated more spurious groups than *F* (Table 3) and the general tendency is that methods derived from bootstrap (B methods: BT* and BP*) generated more spurious groups than those derived from jackknife (J methods: JD and JS). In particular, JD is the method that generated fewest spurious groups (Table 3). These tendencies still occur (but to a lower degree) when only the groups with support above 5% are considered (results not shown).

We report the first examples of spurious groups with absolute frequencies above 50%, even under symmetrical resampling (Fig. 2). In this example we have increased the tree buffer of each pseudoreplicate and obtained similar results, which suggests that this is a real phenomenon; those spurious groups also have a positive GC′ value (Fig. 3). Unsupported groups with positive GC have been already reported by Goloboff et al. (2003a, fig. 11).

We expected that JS would generate fewer spurious groups than JD, because it was proposed as an improvement of JD (Goloboff et al., 2003a). However, this tendency was not observed here. On the other hand, the number of spurious groups did not differ consistently among the weighting schemes analysed.

Spurious groups are (as shown in Fig. 2 for J methods) easy to identify in curves of frequency against resampling strengths, because they have a positive slope at weak resampling strengths (at P = 0, F = 0), while real groups have a negative slope (at P = 0, F = 1) (Goloboff et al., 2003a,b). However, we should be careful with these generalizations because frequency slopes can change drastically at very low resampling strengths (Ramírez, 2005).

We have also computed an estimation of the error rate as $E = S/R$, where S is the number of spurious groups, and R the number of real groups recovered. Then we expected low error rates for methods (BT*, BP*, JD, JS) and measures (F, GC′) that did not generate many spurious groups and are efficient in recovering real groups. In general, we have shown that error rates for GC′ were higher than or equal to those

Table 2
Methods recovering more (+) and fewer (−) groups according to *F* and GC′ measures, under equal weights (EqWe), transition:transversion cost 1:10, and implied weights

Resampling strength	*F*						G					
	EqWe		T_i:T_v		ImWe		EqWe		T_i:T_v		ImWe	
	+	−	+	−	+	−	+	−	+	−	+	−
1	JD	BT	JD	BT	JD	BT BP	JD JS	BT	JD	BT BP	JD	BT BP
2	JD	BT	JD JS	BP	JD	BT	JD	BT BP	JD	BT BP	JD	BP
3	JD	BT	JD	BT	JD	BT	JD	BT	JD	BP	JD	BT
4	JD	BT	JD	BT	JD	BT	JD	BT	JD	BT	JD	BT
5	JD	BT	JD JS	BT	JD	BT	JD	BT BP	JD	BT	JD	BT
6	JD	BT	JD	BT	JD	BT	JD	BT BP	JD JS	BT	JD	BT
7	JD	BT BP	JD	BT	JD	BT BP	JD	BT	JD	BT	JD	BT
8	JD	BT	JD	BT	JD	BT	JD	BT	JD	BT BP	JD	BT
9	JD	BP	JD	BT	JD	BT	JD	BT	JD	BT BP	JD	BT
10	JD	BT	JD	BT	JD	BT	JD	BT	JD	BT	JD	BT
11	**JD JS**	**BP**	**JD**	**BT**	**JD**	**BT**	**JD JS**	**BT**	**JD**	**BT**	**JD JS**	**BT BP**
12	JS	BP	JD JS	BT BP	JD	BP	JD	BT BP	JD	BT	JD JS	BP
13	JS	BT BP	JD	BT BP	JD	BT	JD	BT	JD	BT	JS	BT
14	JS	BT BP	JD	BT	JD	BT BP	JD JS	BT	JD JS	BT	JD JS	BP
15	JS	BP	JS	BP	JS	BT BP	JD	BT BP	JD JS	BT	JS	BT

For resampling strengths values see Table 1. Resampling strength 11 (bold) corresponds to the values commonly used in jackknife and bootstrap.

Abbreviations: BT, variable bootstrap; JD, jackknife; BP, variable bootstrap with Poisson distribution; JS, symmetric resampling.

obtained for F (Table 4; Fig. 4), and that BT* had the highest error rates, followed by BP*, JS and JD, in that order (Table 4). All weighting schemes presented the highest error rate for B* methods (Appendix S1: Table S1). See also tendencies considering only the 11th resampling strength at Appendix S1 text and Table S2.

In general, the spurious groups have low support values (Fig. 4; see also Figs 2 and 3 and Supporting Information). Counting over all resampling strengths and treatments, the mean F of spurious groups was 57%, and the mean GC′ was 11% (see Supporting information). These values are nearly the same when only the 11th strength is considered. For the 11th strength, real groups had mean F = 65%, and mean GC′ = 52%, although the frequency distribution is not symmetrical (Fig. 4).

Contradictions in "support" rankings between different resampling strengths

We have plotted group frequencies against resampling strengths (like those in Figs 1–3) for all the data sets, methods, measures and resampling strengths (available in Supporting information). These plots were inspected visually for qualitative differences in rankings of group frequencies in different analysis situations.

In order to evaluate differences in frequency rankings that we assessed qualitatively in the graphics, we calculated and compared the Kendall coefficient (*Ke*) for the ranking of groups across resampling strengths. Comparisons are limited to a given data set because *Ke* is dependent of the number of groups and, as different data sets had different number of groups, they are not comparable.

We evaluated whether the GC measure reduces the crossing trajectories as resampling strengths vary. We found a corroborating tendency, with higher Kendall coefficient values for GC′ (for all groups, $Ke_{GC'}$ higher in 68, Ke_F higher in 48 cases; for not spurious groups, $Ke_{GC'}$ higher in 62, Ke_F higher in 54 cases; see Supporting information), and this situation is reflected in the fewer crossing trajectories for GC′ than for F in the graphics.

When we compared methods, we found that BT* had a greater proportion of higher *Ke* values (fewer crossing trajectories in the graphics) in all weighting schemes, followed by BP*, JS, and JD, in that order (Figs 2, 3 and 5). This was true both for F and GC′ measures, considering all groups or not spurious groups only.

When we checked which method had a higher proportion of low values of *Ke* (more crossing trajectories in the graphics), we found that it was clearly JD (Fig. 5; cf. Fig. 2b with 2a,c,d; cf. Fig. 3b with 3a,c,d). The same tendency occurs within each weighting scheme (Table 5), except in the T_i:T_v weighting scheme (GC′ measure), where JS also had low values.

Because JS is a modified version of JD, and BP* is a modified version of BT*, we compared values of *Ke* for JD against JS, and for BT* against BP*, separately. We have found that JS had consistently higher values than JD, and BP* had more higher values than BT*, but the tendency is not so clear as when comparing JD with JS (for details see Appendix S1: Table S3).

Table 3
Number of spurious groups generated by each resampling method with F and GC′ measures, for the three weighting schemes analysed

	F				GC′			
Data set	Bootstrap	Jackknife	Bootstrap with Poisson distribution	Symmetric resampling	Bootstrap	Jackknife	Bootstrap with Poisson distribution	Symmetric resampling
Equal weights								
D01	0	0	0	0	0	0	0	0
D02	0	0	0	0	**1**	(0)	**1**	(0)
D03	0	0	0	0	0	0	0	0
D04	1	1	1	1	**5**	(3)	(3)	(3)
D05	**3**	(1)	**3**	(1)	**10**	(8)	**10**	(8)
D06	**5**	(2)	4	(2)	(24)	**28**	(24)	25
D07	0	0	0	0	5	**8**	(4)	5
D08	0	0	0	0	0	0	0	0
D09	**2**	(0)	**2**	(0)	(8)	**9**	(8)	(8)
D10	0	0	0	0	3	3	3	3
D11	0	0	0	0	2	2	2	2
Transition:transversion 1:10								
D01	0	0	0	0	0	0	0	0
D02	0	0	0	0	2	2	2	2
D04	1	1	1	1	6	6	6	6
D05	2	2	2	2	8	(5)	**9**	(5)
D07	(0)	**1**	**1**	**1**	(5)	(5)	(5)	**6**
D08	0	0	0	0	0	0	0	0
D09	0	0	0	0	**5**	(4)	(4)	(4)
D10	**2**	(0)	**2**	1	(7)	**9**	(7)	8
D11	0	0	0	0	0	0	0	0
Implied weights								
D01	0	0	0	0	0	0	0	0
D02	0	0	0	0	2	2	2	2
D03	0	0	0	0	0	0	0	0
D04	**4**	(1)	2	(1)	**6**	(2)	4	(2)
D05	2	2	2	2	**8**	(5)	7	7
D07	0	0	0	0	4	(3)	**5**	(3)
D08	0	0	0	0	0	0	0	0
D09	(0)	**1**	**1**	**1**	3	3	3	3
D10	0	0	0	0	2	2	2	2
D11	0	0	0	0	0	0	0	0
Number of maximum	**5**	2	**5**	2	**6**	4	4	1
Number of minimum	2	(5)	0	4	4	(9)	7	8

For each data set, cells containing the highest numbers of spurious groups generated are in bold; cells containing the lowest values are in parentheses. The two last rows summarize the number of cases in which we found the highest and the lowest number of spurious groups for each method and for each measure.

Conclusions

In this study we have analysed general tendencies of resampling support values using (i) 15 resampling strengths, (ii) three weighting schemes, (iii) four resampling methods, and (iv) two measures. We compared the methods traditionally used (BT and JD) and their modified versions (BP and JS). We have also evaluated different measures of frequency values: F, the one traditionally used, and GC, which is experimental and was proposed recently (Goloboff et al., 2003a). Also, in order to make B methods comparable with J methods, we have proposed a variable bootstrap (BT*), which has a variable final data set size when varying the resampling strength. All the methods and measures examined produced spurious groups, although usually with low "support" values.

We have found a clear tendency showing that J methods recovered more real groups, while producing fewer spurious groups, thus having the lowest error rate. However, JD is the method that produced more inconsistencies in ranking of frequencies with variable resampling strengths (more crossing trajectories in frequency against resampling strength).

Goloboff et al. (2003a) have proposed JS to correct the distortion of frequencies under BT or JD (either under- or overestimations of the actual group support) that may happen when character weights or state transformation costs are heterogeneous. Hence we might have expected that JS would recover more real

Table 4
Number of maximum error rate values per method for all data set, considering all the resampling strengths

Weighting scheme/measure	Total of maximum error rate value for each method			
	BT	JD	BP	JS
EqWe *F*	36	7	28	4
EqWe GC′	43	11	37	23
Iv1:10 *F*	31	7	31	8
Iv1:10 GC′	65	7	51	14
P_w *F*	27	0	13	3
P_w GC′	47	9	35	5
Total of máximum S/R index value for each method	249	41	195	56

Abbreviations—for weighting schemes: EqWe, equal weights; Iv1:10, transition:transversion cost 1:10; P_w, implied weights; for methods: BT, variable bootstrap; JD, jackknife; BP, variable bootstrap with Poisson distribution; JS, symmetric resampling. For more details see Appendix S1: Table S1.

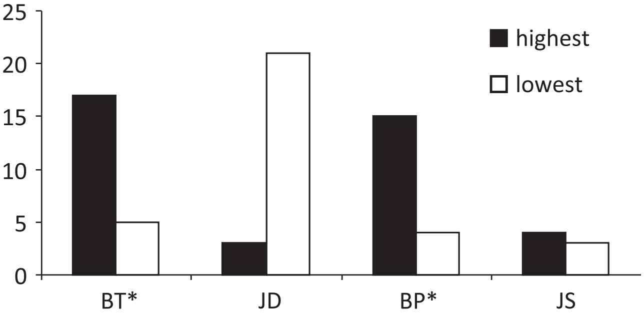

Fig. 5. Total number of matrices in which each method had the highest and lowest value of Kendall coefficient (*Ke*) using absolute frequency measure and considering all groups. Similar results were obtained with GC′, or considering not spurious groups only (data not shown). References: BT*, variable bootstrap; JD, jackknife; BP*, variable bootstrap with Poisson distribution; JS, symmetric resampling.

groups, generate fewer spurious groups, and exhibit fewer crossing trajectories in frequency against resampling strength, under T_i:T_v and ImWe weighting schemes. We have not obtained a clear tendency verifying this expectation.

JS always generated fewer group ranking differences under different resampling strengths than JD, so we can consider JS as more consistent than JD. However, JS did not completely solve the problem of inconsistent rankings between different resampling strengths. We also present examples of spurious groups with frequencies above 50% or positive GC′ values under JS, indicating that symmetric resampling is not enough to correct against spurious groups.

B* methods produced fewer ranking inconsistencies with variable resampling strength, but generated more spurious groups and recovered fewer real groups (higher error rate) than J methods, especially under weak resampling strengths. This is reflected in flatter trajectories in frequency against resampling strength graphics (Figs 2 and 3).

We also wanted to assess whether BP* generated fewer spurious groups than BT*, and whether JS generated fewer spurious groups than JD, because they were proposed as improvements of these resampling methods. We could not find a clear tendency in our experiments. We noticed, however, that JD and BT* had lower *Ke* values than their improved versions JS and BP*, as expected. There is a clear tendency showing that JS produced fewer crossings than JD in the trajectories of frequency against resampling strengths. However, the differences were deeper for J methods than between B* methods.

We have also found that GC′, designed to recover more real groups, does so at the expense of also recovering more spurious groups, and having a higher error rate.

The tendencies in this study suggest that F values may be better indicators of support than GC′ values. Our results also suggest that J methods produce better indicators of support than B* methods, because they show lower error rates. Contrary to our expectations, we have not found clear interactions between weighting schemes and resampling methods.

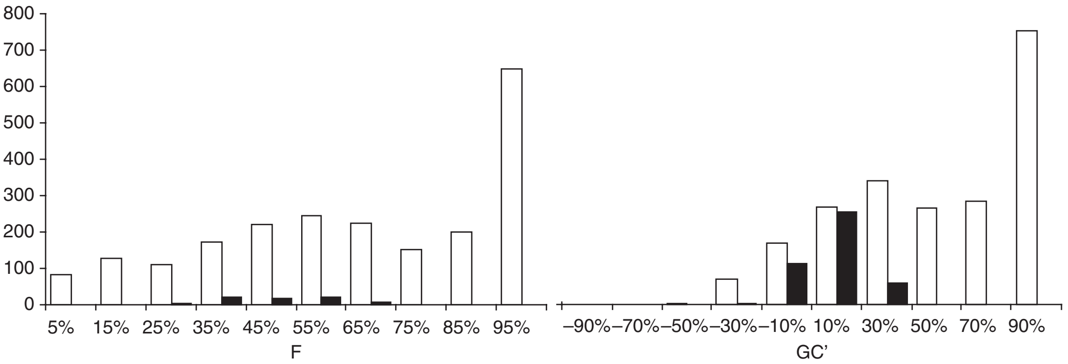

Fig. 4. Count of real (white) and spurious (black) groups for all treatments under resampling strength 11, classified according to their absolute frequency (F) and frequency difference (GC′).

Table 5
Total number of matrices in which each method had the highest/lowest value of Kendall coefficient, for each frequency measure, considering weighting schemes separately (highest proportion of lowest Kendall coefficient in bold)

Methods		Bootstrap		Jackknife		Bootstrap with Poisson distribution		Symmetric resampling	
Measures	Weighting schemes	All groups	Not spurious groups	All groups	Not spurious groups	All groups	Not spurious groups	All groups	Not spurious groups
F	Equal weights	7/2	5/1	1/**8**	1/**9**	7/0	7/0	1/1	2/1
	Transition:transversion 1:10	4/3	4/0	1/**5**	1/**5**	3/3	3/2	2/1	1/2
	Implied weights	6/0	5/0	1/**8**	1/**7**	5/1	6/1	1/1	0/2
	Total	17/5	14/1	3/**21**	3/**21**	15/4	16/3	4/3	3/5
GC′	Equal weights	6/1	8/1	1/**10**	0/**10**	5/0	4/0	1/0	2/0
	Transition:transversion 1:10	5/0	5/0	1/**5**	3/3	4/0	2/1	2/**5**	0/**5**
	Implied weights	6/0	6/0	0/**7**	0/**7**	5/1	6/0	2/2	1/3
	Total	17/1	19/1	2/**22**	3/**20**	14/1	12/1	5/7	3/8

Acknowledgements

We thank Santiago Catalano, Marcos Mirande, Pablo Goloboff and an anonymous reviewer for helpful comments on the manuscript and Leonardo Campagna for help with the figures. This study was supported by PIP 0032 and 6502 from CONICET, EAR-0228699 from NSF, and PICT 14092 from ANPCyT.

References

Baker, R.H., DeSalle, R., 1997. Multiple sources of character information and the phylogeny of Hawaiian drosophilids. Syst. Biol. 46, 654–673.

Berry, V., Gascuel, O., 1996. On the interpretation of bootstrap trees: appropriate threshold of clade selection and induced gain. Mol. Biol. Evol. 13, 999–1011.

Carpenter, J.M., 1992. Random cladistics. Cladistics 8, 147–153.

Chen, W., Shearer, C.A., Crane, L.J., 1999. Phylogeny of *Ophioceras* spp. based on morphological and molecular data. Mycologia 21, 84–94.

Cohen, B.L., Baker, A.J., Blechschmidt, K., Dittmann, D.L., Furness, R.W., Gerwin, J.A., Helbig, A.J., De Korte, J., Marshall, H.D., Palma, R.L., Peter, H.U., Ramli, R., Siebold, I., Willcox, M.S., Wilson, R.H., Zink, R.M., 1997. Enigmatic phylogeny of skuas (Aves: Stercorariidae). Proc. R. Soc. Lond. B 264 (1379), 181–190.

Daniel, W.W., 1978. Applied Nonparametric Statistics. Houghton Mifflin, Boston, MA, USA.

De Laet, J., Farris, J.S., Goloboff, P.A., 2004. Treatment of multiple trees in resampling analyses. In: Grandcolas, P. Abstract of the 23rd Annual Meeting of the Willi Henning Society. ''Phylogenetics and Evolutionary Biology''. Cladistics 20, 590.

Efron, B., 1979. Bootstrap methods: another look at the jackknife. Ann. Stat. 7, 1–26.

Faith, D.P., Cranston, P.S., 1991. Could a cladogram this short have arisen by chance alone? On permutation test for cladistic structure. Cladistics 7, 1–28.

Farris, J., 1969. A successive approximations approach to character weighting. Syst. Zool. 18, 374–385.

Farris, J.S., 1999. In: Horovitz, I. (Eds.), A report on ''One day Symposium on Numerical Cladistics''. Cladistics 15, 177–182.

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

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

Felsenstein, J., 2004. Inferring Phylogenies. Springer-Verlag, New York, NY, USA.

Gatesy, J., O'Grady, P., Baker, R.H., 1999. Corroboration among data sets in simultaneous analysis: hidden support for phylogenetic relationships among higher level artiodactyl taxa. Cladistics 15, 271–313.

Goloboff, P., 1993. Estimating character weights during tree search. Cladistics 9, 83–91.

Goloboff, P.A., Farris, J.S., Källersjö, M., Oxelman, B., Ramírez, M.J., Szumik, C.A., 2003a. Improvements to resampling measures of group support. Cladistics 19, 324–332.

Goloboff, P.A., Farris, J.S., Nixon, K. 2003b. TNT: Tree Analysis Using New Technology. Program and documentation available from the authors: http://www.zmuc.dk/public/phylogeny.

Grant, T., Kluge, A.G., 2003. Data exploration in phylogenetic inference: scientific, heuristic or neither. Cladistics 19, 379–418.

Harshman, J., 1994. The effect of irrelevant characters on bootstrap values. Syst. Biol. 43, 419–424.

Hayasaka, K., Gojobori, T., Horai, S., 1988. Molecular phylogeny and evolution of primate mitochondrial DNA. Mol. Biol. Evol. 5, 626–644.

Huber, K.C., Haider, T.S., Müller, M.W., Huber, B.A., Schweyen, R.J., Barth, F.G., 1993. DNA sequence data indicates the polyphyly of the family Ctenidae (Araneae). J. Arachnol. 21, 194–201.

Kluge, A.G., Wolf, A.J., 1993. Cladistics: what's in a word? Cladistics 9, 183–199.

Lijtmaer, D.A., Sharp, N.M.M., Tubaro, P.L., Lougheed, S.C., 2004. Molecular phylogenetics and diversification of the genus *Sporophila* (Aves: Passeriformes). Mol. Phylogenet. Evol. 33, 562–579.

Norman, J.E., Egger, K.N., 1999. Molecular phylogenetic analysis of *Peziza* and related genera. Mycologia 91, 820–829.

Nylander, J.A.A., Ronquist, F., Huelsenbeck, J.P., Nieves-Aldrey, J.L., 2003. Bayesian phylogenetic inference of combined data. Syst. Biol. 53, 47–67.

Ramírez, M., 2005. Resampling measures of group support: a reply to Grant and Kluge. Cladistics 21, 83–89.

Sang, T., Crawford, D.J., Stuessy, T.F., Silva-O, M., 1995. ITS sequences and the phylogeny of the genus *Robinsonia* (Asteraceae). Syst. Bot. 20, 55–64.

Whiting, M.F., Carpenter, J.C., Wheeler, Q.D., Wheeler, W.C., 1997. The Strepsiptera problem: phylogeny of the holometabolous insect orders inferred from 18S and 28S ribosomal DNA sequences and morphology. Syst. Biol. 46, 1–68.

Supporting Information

Additional Supporting information may be found in the online version of this article:

Appendix S1. Background: resampling methods and measures.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

Appendix 1

Sources, characteristics and accession numbers of the matrices used in this study

Data set	Number of taxa	Number of characters	Number of replicates	Source	Treebase study accession number	Treebase matrix accession number	Type of character
D01	9	486	1	Sang et al. (1995)	S11×6×95c10c06c51	M57c11×6×95c10c09c35	DNA
D02	9	446	1	Huber et al. (1993)	S11×16×96c22c02c22	M161c11×16×96c22c10c59	DNA
D03	13	24	1	Ramírez (2005)	–	–	Hypothetical
D04	42	1070	2	Chen et al. (1999)	S259	M299	DNA
D05	52	398	3	Whiting et al. (1997)	S325	M419	DNA
D06	85	1016	5	Whiting et al. (1997)	S325	M420	DNA
D07	24	1785	1	Norman and Egger (1999)	S380	M526	DNA
D08	10	1020	1	Cohen et al. (1997)	S428	M626	DNA
D09	22	381	1	Nylander et al. (2003)	S970	M1611	DNA
D10	33	498	2	Lijtmaer et al. (2004)	–	–	DNA
D11	12	898	1	Hayasaka et al. (1988)	–	–	DNA

Appendix 2

Values of $(1 - P) \times 1000$ for bootstrap with Poisson distribution, where P = probability of receiving a certain character weight (mean character weight = m as calculated for variable bootstrap)

Strength	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
m	3.40	2.70	2.30	2.01	1.79	1.61	1.46	1.32	1.20	1.10	1.00	0.92	0.84	0.76	0.69
Character weight															
0	333	667	1000	1333	1667	2000	2333	2667	3000	3333	3667	4000	4333	4667	5000
1	1467	2472	3303	4020	4653	5219	5729	6191	6612	6995	7345	7665	7957	8223	8466
2	3395	4917	5954	6726	7328	7809	8200	8521	8786	9007	9191	9344	9472	9579	9667
3	5581	7123	7988	8544	8926	9199	9398	9547	9659	9744	9808	9857	9895	9923	9944
4	7440	8617	9159	9460	9642	9758	9835	9886	9922	9946	9963	9975	9983	9989	9992
5	8704	9426	9699	9829	9898	9938	9961	9976	9985	9990	9994	9996	9998	9999	9999
6	9421	9791	9906	9953	9975	9986	9992	9996	9997	9999	9999	10000	10000	10000	10000
7	9769	9933	9974	9989	9995	9997	9999	9999	10000	10000	10000	10000	10000	10000	10000
8	9917	9980	9994	9997	9999	10000	10000	10000	10000	10000	10000	10000	10000	10000	10000
9	9973	9995	9999	10000	10000	10000	10000	10000	10000	10000	10000	10000	10000	10000	10000
10	9992	9999	10000	10000	10000	10000	10000	10000	10000	10000	10000	10000	10000	10000	10000