WILEY

Cladistics

Cladistics 38 (2022) 392–401

10.1111/cla.12496

# An empirical test of the relationship between the bootstrap and likelihood ratio support in maximum likelihood phylogenetic analysis

Denis Jacob Machado[a,b*], Fernando Portella de Luna Marques[c], Larry Jiménez-Ferbans[d] and Taran Grant[c]

[a]*Programa Inter-unidades de Pós-graduação em Bioinformática, Universidade de São Paulo, Rua do Matão 1010 São Paulo, SP 05508-090, Brazil;* [b]*Department of Bioinformatics and Genomics, College of Computing and Informatics, University of North Carolina at Charlotte, 9331 Robert D. Snyder Rd, Charlotte, NC 28223, USA;* [c]*Departamento de Zoologia, Instituto de Biociências, Universidade de São Paulo, Tv. 14, 101 - Butantã, São Paulo, SP 05508-090, Brazil;* [d]*Facultad de Ciencias Básicas, Universidad del Magdalena, Carrera 32 No 22-08, Santa Marta D.T.C.H., Magdalena 470004, Colombia*

Accepted 15 November 2021

**Abstract**

In maximum likelihood (ML), the support for a clade can be calculated directly as the likelihood ratio (LR) or log-likelihood difference (*S*, LLD) of the best trees with and without the clade of interest. However, bootstrap (BS) clade frequencies are more pervasive in ML phylogenetics and are almost universally interpreted as measuring support. In addition to theoretical arguments against that interpretation, BS has several undesirable attributes for a support measure. For example, it does not vary in proportion to optimality or identify clades that are rejected by the evidence and can be overestimated due to missing data. Nevertheless, if BS is a reliable predictor of *S*, then it might be an efficient indirect method of measuring support—an attractive possibility, given the speed of many BS implementations. To assess the relationship between *S* and BS, we analyzed 106 empirical datasets retrieved from TreeBASE. Also, to evaluate the degree to which *S* and BS are affected by the number of replicates during suboptimal tree searches for *S* and pseudoreplicates during BS estimation, we randomly selected 5 of the 106 datasets and analyzed them using variable numbers of replicates and pseudoreplicates, respectively. The correlation between *S* and BS was extremely weak in the datasets we analyzed. Increasing the number of replicates during tree search decreased the estimated values of *S* for most clades, but the magnitude of change was small. In contrast, although increasing pseudoreplicates affected BS values for only approximately 40% of clades, values both increased and decreased, and they did so at much greater magnitudes. Increasing replicates/pseudoreplicates affected the rank order of clades in each tree for both *S* and BS. Our findings show decisively that BS is not an efficient indirect method of measuring support and suggest that even quite superficial searches to calculate *S* provide better estimates of support.

## Introduction

The discussion of support, or strength of evidence, in maximum likelihood (ML) analysis is not new. Hacking (1965) related likelihood and support through the “law of likelihood” and proposed the likelihood ratio (LR) as a measure of support: “If *h* and *i* are simple joint propositions included in the joint proposition *e*, then *e* supports *h* better than *i* if the likelihood ratio of *h* to *i* exceeds 1” (Hacking, 1965:63). Although LR can be calculated for any pair of hypotheses, it can be most useful for comparing optimal and next-best hypotheses. As such, the evidential support for any clade in the optimal cladogram is obtained by dividing the likelihood of the maximum likelihood

*Corresponding author:
*E-mail address:* dmachado@uncc.edu

hypothesis by the likelihood of the best hypothesis that lacks that clade, or, equivalently, subtracting the log-likelihood of the best contradictory hypothesis from the log-likelihood of the optimal hypothesis (see Grant & Kluge, 2008 for a more detailed explanation). The likelihood ratio does not necessarily correlate with accuracy or the expectation that the clade in the optimal cladogram is present in the true phylogeny, as these questions would require prior (unavailable) knowledge about the conditional probability of the clade, given the cladogram (see Royall, 2004). Therefore, LR does not answer how confident we should be that a particular clade is true, either in the context of belief or error probabilities. Instead, LR serves as a measure of the strength of the evidence for competing hypotheses based only on the evidence and the models involved (Taper & Lele, 2004), which is consistent with the philosophical underpinnings of ML. As summarized by Hacking (1965), confidence intervals are for before-trial betting, whereas LR is for after-trial evaluation.

Despite the increasing focus on evidential statistics as an alternative to both standard frequentist and Bayesian paradigms, as well as the concomitant recognition of LR as an evidence function used to measure support in ML outside phylogenetics (e.g., Cahusac, 2020; Taper & Ponciano, 2016), LR has received less attention from phylogeneticists than bootstrap (BS) frequencies or proportions. Although it is not hard to find examples of authors using LR (e.g., Meireles et al., 1999; Lee & Hugall, 2003; Padial et al., 2014; Marques & Caira, 2016; Trevisan et al., 2017; see also Grant & Kluge, 2008), BS values are the standard measure of clade support in ML. Bootstrap frequencies are computed by non-parametric bootstrap resampling (Efron, 1979) of characters from the original alignment or character matrix to create pseudoreplicate matrices of the same size as the original dataset and conducting a tree search for each pseudoreplicate matrix (Felsenstein, 1985). The BS frequency of each clade is equal to the frequency of that clade among the trees obtained from all the pseudoreplicates.

Although BS is commonly used as a method for measuring support (e.g., Bardin et al., 2016; Boyd et al., 2017; Linkem et al., 2016; Simon, 2020; Yuan et al., 2016), that is not its purpose. Indeed, as with the use of *P*-values as support generally, “this suggestion is always informal, and no theory is ever put forward for what properties a measure of support or evidence should have” (Schervish, 1996:204). The bootstrap was proposed in phylogenetics as a means of placing confidence intervals (Felsenstein, 1985). As such, it is a measure of uncertainty or imprecision, which can be empirically related to, but is logically different from, support, or strength of evidence (e.g., Goodman & Royall, 1988; Hacking, 1965). Despite the decades-old conflation of confidence and support in phylogenetics, their distinction is easily demonstrated empirically by the well-known fact that clades that are present in the optimal tree—and are therefore supported by the evidence—can have lower BS frequencies than clades that are absent from the optimal tree—and are therefore unsupported by the evidence—as well as the fact that BS values do not vary in direct relation to optimality (Grant & Kluge, 2008). An additional concern is the increasing problem of highly significant *P*-values being observed for contrasting phylogenetic hypotheses when analyzing large datasets (Kumar et al., 2012).

Nevertheless, BS continues to be interpreted as a measure of support. A review of recent phylogenetic publications gives an idea of just how prevalent this use of BS is. We reviewed 233 papers from recent volumes of Cladistics (volumes 33–36), Molecular Phylogenetics and Evolution (142–147), and Systematic Biology (66–69) that included phylogenetic analyses. In total, 177 of those 233 papers (~76%) used the ML criterion. All of the papers that used the ML criterion used BS to measure support (Table S1). Furthermore, not only is BS most widely used, but there continues to be an influx of new algorithms to improve BS analysis (e.g., Chang et al., 2019; Chatzou et al., 2018; Klötzl & Haubold, 2016; Lemoine et al., 2018; Wang & Liu, 2020; Wang et al., 2016, 2020).

Although BS does not technically measure support, given the speed of many BS implementations, the support interpretation could be salvaged if BS proves to be an efficient indirect measure of support. That is, if BS frequencies are a reliable predictor of LR, it might be possible to use BS frequencies obtained via fast BS analysis as a proxy for the direct support values obtained from slower LR analyses. Of course, this assumes that BS is faster than LR, but the speed of a given BS or LR analysis depends on the specified search parameters. Large numbers of pseudoreplicates and/or a thorough search of each pseudoreplicate in a BS analysis could be slower than a superficial search for the best trees lacking each clade of the ML tree. To date, the empirical relationship between BS and direct measures of support and the effects of variations in the number of replicates or pseudoreplicates have been poorly explored.

Our objective is to test if BS frequencies are an efficient indirect method of measuring support by systematically analyzing the relationship between them and LR. Specifically, we ask: First, how well does BS correlate with LR? And second, how sensitive are BS and LR to variations in the number of pseudoreplicates and replicates, respectively? By answering these questions, we will provide empirical comparisons of these measures of support and add light to the discussion regarding what they are measuring, how their values should be interpreted, and how they should be used in practice.

## Materials and methods

### Empirical data and tree search

Empirical data used in the current study comprised 106 matrices of DNA sequences downloaded from TreeBASE (https://treebase.org). These 106 datasets were obtained with the help of a Bash script that modifies download links according to different accession numbers. We attempted to download all datasets between M15000 and M35000 from TreeBASE. We ignored all instances in which the download failed, or the dataset format was not standardized and would not allow retrieving the original data without editing the original file. Accession numbers of the 106 selected datasets, as well as their number of taxa and characters, are listed in Table S2. ML analysis was performed using Garli version 2.01 (Zwickl, 2006). To analyze the 106 data matrices retrieved from TreeBASE, we used 1000 random addition sequence replicates to find optimal trees, 1000 pseudoreplicates for BS, and 100 replicates for each search constrained to find the best tree lacking the nodes in the optimal tree, one node at a time. We used original Python scripts to facilitate Garli parallelization and support calculation. Software is freely available under the GNU General Public License version 3.0 (GPL-3.0) at https://grant.ib.usp.br/anfibios/researchSoftware.html, https://gitlab.com/MachadoDJ/sudoparallelgarli, and https://gitlab.com/MachadoDJ/s4ml. The software includes the configuration file templates used for all tree search experiments.

### Computational resources

All in silico procedures were executed using “ACE” (https://grant.ib.usp.br/anfibios/researchHPC.html), an SGI rackable computer cluster housed in the Museum of Zoology of the University of São Paulo. Selected servers had four 2.3 GHz Operon CPUs with 16 cores each and 256 or 516 GB of memory. The software environment in ACE consisted of a SUSE Linux Enterprise Server with SGI Performance Suite, SGI Management Center, and PBS Pro Job Scheduler.

### Likelihood ratios and support

The likelihood ratio (LR) for any node, $n$, in the maximum likelihood tree is the likelihood of that tree ($L_1$) divided by the likelihood of the best tree lacking that node ($L_2$):

$$\mathrm{LR}_n = \frac{L_1}{L_2}$$

Although the theoretical basis for support in ML is LR (e.g., Hacking, 1965), support ($S$) is defined technically as the natural logarithm of the likelihood ratio (Edwards, 1972), which converts the ratio of likelihood scores into the log-likelihood difference (LLD of Grant & Kluge, 2008; see also Marques & Caira, 2016; Trevisan et al., 2017):

$$S = \ln\mathrm{LR} = \ln L_1 - \ln L_2$$

Thus defined, $S$ is symmetrically distributed around 0 (which obtains when both hypotheses are equally supported by the evidence) and unbounded, with positive values indicating that $L_1$ is more strongly supported and negative values indicating that $L_2$ is more strongly supported. Although $S$ is not associated with particular cutoffs, by convention a value of 1 (LR = 2.7) is considered weak, 2 (LR = 7.4) moderate, 3 (LR = 20) strong, and 4 (LR = 55) extremely strong (Goodman & Royall, 1988). Conveniently, Garli likelihood scores ($s$) are reported as the natural logarithm of the likelihood ($\ln L$) of the tree. Consequently, for any node ($n$) in the optimal tree, $S$ can be calculated as:

$$S_n = s_1 - s_2$$

where $s_1$ is the log-likelihood of the optimal tree and $s_2$ is the log-likelihood of the best tree lacking node $n$, and negative values for $S_n$ indicate that a better tree was found during the support search. Note that this would indicate that the original search was inadequate and lead researchers to perform more exhaustive analyses before calculating support and, therefore, it is unlikely to be reported in empirical studies unless authors are interested in measuring strength of contradiction. To facilitate comparison with BS frequencies, we also compare them using normalized $S$ values,

$$\hat{S} = 1 - \left(\mathrm{e}^{-S}\right)$$

Accordingly, an $\hat{S}$ of 0.7 would indicate that $s_2$ is 30% lower than $s_1$.

### Correlation and sensitivity to the number of replicates and pseudoreplicates

If a method is an efficient indirect measure of support, we would expect an increase in its value to be matched, on average, by a proportionate increase in the value of the direct measure of support, $S$. That is, we would expect the $S$ of a clade with a BS of 80% to be, on average, twice that of a clade with a BS of 40%. To understand the relationship between direct and indirect measures of support, we extracted $S$, $\hat{S}$, and BS values for each internal node to examine their distributions and test the correlation between $\hat{S}$ and BS using linear and non-linear regression analysis and both parametric and non-parametric strategies in R version 4.0.4 (http://www.R-project.org). The dataset and detailed analysis protocol in R are given in Table S3 and Appendix S1, respectively.

We conducted a nonlinear dependence test between BS and $S$ values using the R package “canova” (Wang et al., 2015) with various bin sizes (2, 5, 10, 20, 30, 40, 50, 100, 198, and 200) and permutation numbers (1000, 2000, 5000, and 10 000). In addition, we also computed the “multiinformation” (also called “total correlation”) between BS and $S$ using the R package “infotheo” (Cover, 1999; Meyer, 2008). Together, these two analyses indicate if BS values lead to similar $S$ values, although the correlation may not be strong enough to find that BS is a good predictor of $S$.

We used ordinary least squares (OLS) estimation for linear models to evaluate the proportionality between BS and $S$. In this context, even if BS does not correlate linearly with $S$, we can use linear regressions to analyze relationships that are not inherently linear after data transformation. For BS values, we tested the logit, natural logarithm, quadratic, and square-root transformations.

For $S$, we performed transformations based on $\hat{S}$ described above. This transformation is based on the function $f(S) = 1 - \mathrm{e}^{(-S/b)}$, with $S \geq 0$ and $b > 0$, which is the common cumulative distribution function. Every probability distribution based on the real numbers, discrete or “mixed” as well as continuous, is uniquely identified by a continuous monotonic increasing cumulative distribution function $f$ with limits of 0–1 (which allows intuitive comparison between transformed $S$ values and BS values, which range from 0 to 100). We used $b = 1$ to make the curve more pronounced, resembling the scatterplot of raw BS and $S$ values.

Additionally, we tested the Box Cox transformation of $S$ values. In short, the Box Cox transformation transforms the data to closely resemble a normal distribution. Thus, the Box Cox transformation is a strategy to normalize errors and improve the predictive power of our model.

After computing the OLS regression models for the raw and transformed data, we compared the resulting *P*-values, adjusted $R^2$, and residual statistics. For every computation, we evaluated the effect of heteroscedasticity, deviation of residuals from normality, and outliers.

We also used the raw and transformed data to perform non-parametric linear correlation tests using Pearson's and Spearman's coefficients. Moreover, we used non-parametric quantile regressions (using R's "quantreg" package; Koenker, 2005) with calculations of the Nagelkerke's pseudo-$R^2$ (with the "nagelkerke" function in R's "rcompanion" package; Nagelkerke, 1991).

In addition to the linear models, we implemented nonlinear correlation estimations using an adaptive local linear correlation computation available via the R's package "nlcor" (Ranjan & Najari, 2020; the "nlcor" package is available https://github.com/ProcessMiner/nlcor, accessed on July 6, 2021). Furthermore, we tested nonlinear models using the raw and transformations data comparing the root mean square error (RMSE) and $R^2$ of polynomial, spline, and generalized additive model (GAM) regressions to select the best nonlinear model. Finally, we estimated how much BS would over- or underestimate *S* under a 95% confidence interval given the best linear and nonlinear models, and we examined the frequency with which BS under- and overestimated support relative to *S* by making predictions based on the models that showed strongest correlation between BS and *S* and classifying the results using a 95% confidence interval.

Although both *S* and BS analyses require multiple search parameters to be set (e.g., tree searching algorithms, number of trees held during each search), the simplest and most commonly adjusted parameters are the number of random addition sequence replicates per node for *S* and the number of pseudoreplicates for BS. As such, to evaluate the degree to which *S* and BS are affected by the number of replicates and pseudoreplicates, respectively, during the computations of these indices, we randomly selected 5 of the 106 datasets and analyzed them using variable numbers of replicates/pseudoreplicates. We obtained *S* for each node by searching for the best tree lacking each clade of the optimal tree using 10, 100, and 1000 random addition sequence replicates and the BS frequency for each node using 10, 100, and 1000 pseudoreplicates, repeating each analysis three times and comparing the means, for a total of 1224 data points. The tree searches used the same parameters described above. For each clade shared across search parameters, we determined (i) if the values increased, decreased, or were constant; (ii) the magnitude of increase or decrease; and (iii) the rank order of clades in each tree.

To evaluate whether variable numbers of replicates/pseudoreplicates changed *S* and BS values uniformly, we ranked all clades for each dataset according to those values as ordered lists. Next, we converted those lists into strings of characters and submitted them to pairwise alignment to calculate differences between the order of ranked clades. Scores for each pairwise alignment employed a cost of 1 for matches and −1 for mismatches, gap opening, and gap extensions using the "pairwise2.align.globalms" function of Biopython version 1.79 (available from https://biopython.org/). We measured the distance between two ordered lists that contain the same clades as the percentage of mismatches and gaps in their pairwise alignment. This strategy allowed us to use alignment distances as a proxy to changes in the rank order of clades according to their BS or *S* values.

## Results

### *Correlation of direct and indirect measures of support*

The nucleotide matrices downloaded from TreeBASE resulted in 106 trees with 3959 nodes, among which the mean *S* and $\hat{S}$ were 8.676 ± 13.000 and 0.799 ± 0.304, respectively, and the mean BS was 67.277 ± 29.120. The final dataset used for correlation analyses is available in Table S3. All of the results of the statistical analyses are available in Appendix S1.

Due to the heteroscedasticity observed in the residual statistics, parametric linear regressions are not ideal to indicate the correlation between BS and *S*. Comparison between Pearson's ($R = 0.56$, $R^2 = 0.31$) and Spearman's ($R = 0.84$, $R^2 = 0.71$) coefficients also suggest that the putative relationship between BS and *S* is non-parametric.

Comparisons between non-parametric linear correlations using quantile regressions showed that we maximized pseudo-$R^2$ when applying the Box Cox transformation of *S* values compared to the raw data and other transformations of BS (logit, natural logarithm, quadratic, and square-root) and *S* values (using $\hat{S}$). The quantile regression of BS and Box Cox transformations of *S* values resulted in a *P*-value of 0 and a pseudo-$R^2$ of 0.67 (Figure 1a). Furthermore, the changes in quartile coefficients and confidence intervals for the BS variable show that the quantile slope estimates differ from the least-squares estimate (Figure 1b). However, despite the significant value of *P* and relatively high pseudo-$R^2$ value, this model overestimated and underestimated support in the test dataset (composed of 20% of the raw data) for 47.67% and 47.79% of the predictions, respectively (Figure 1a). This model's application to each matrix shows that predictions are often significantly different from the observations, with an overall greater tendency for BS to overestimate the support value in the observed trees (Figure 1c).

Non-uniform piecewise linear correlations with R's "nlcor" package did not indicate that non-linear models would significantly improve the correlation between BS and *S*, since the regression returned a single straight line cutting through the entire scatter plot. Nevertheless, we tested polynomial, spline, and GAM non-linear regressions, which resulted in RMSE values of 6.83, 10.30, and 13.24 and $R^2$ values of 0.68, 0.27, and 0.37, respectively. Given that the polynomial regression produced the best results among non-linear models and the corresponding $R^2$ values were greater than the pseudo-$R^2$ produced by the quantile regression using the Box Cox transformation of *S* values, we trained a polynomial regression model with 80% of the data and estimated the frequency of overestimated and underestimated predictions (Figure 1d). Among all predicted values in the test dataset, the model had 30.21% overestimated and 39.06% underestimated results. Thus, only 30.73% of the results were within the 95% confidence interval.

### *Sensitivity to variation in the number of replicates/pseudoreplicates*

The mean BS frequencies for increasing the number of pseudoreplicates increased in 18.32% of the nodes

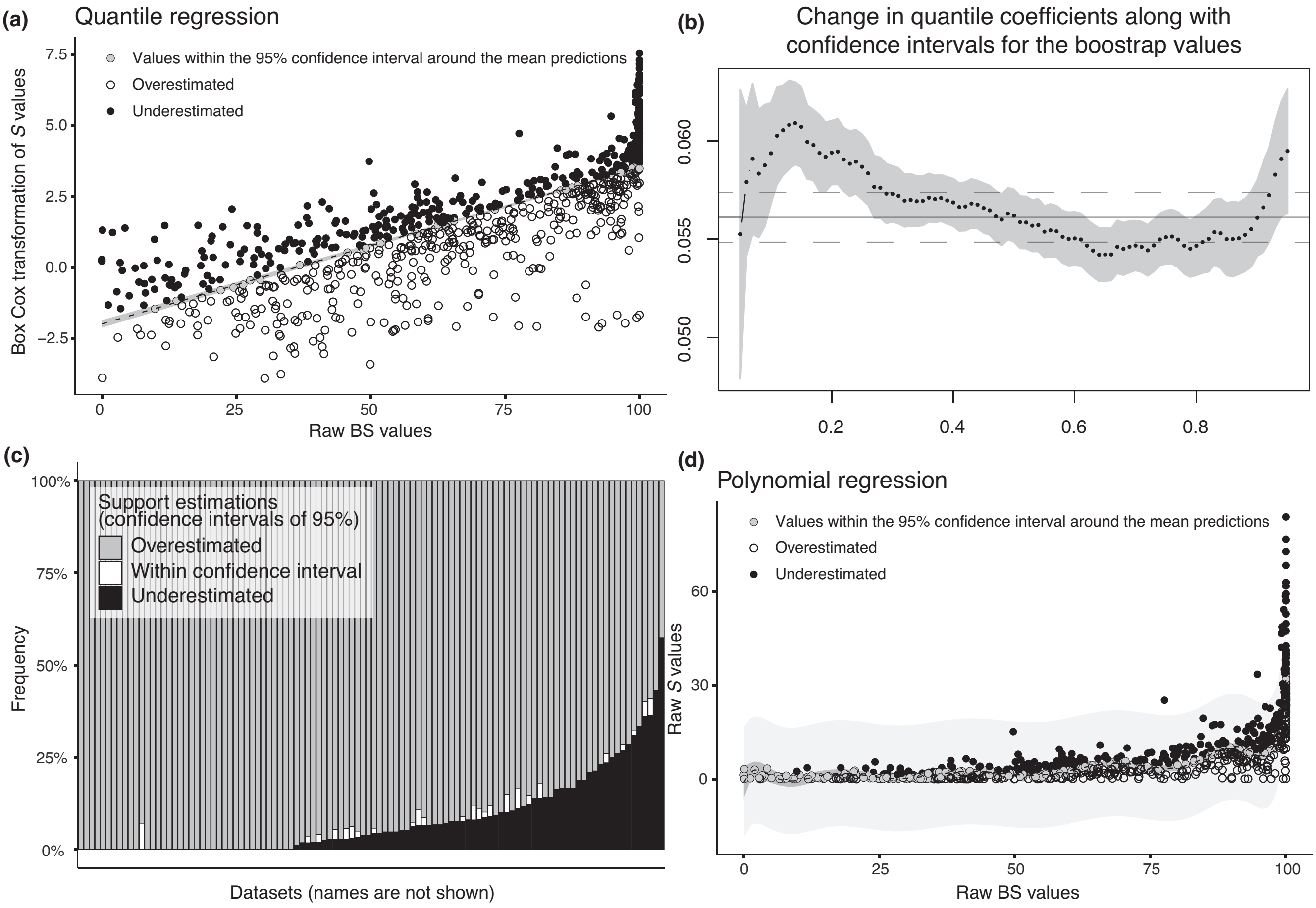

Fig. 1. This figure shows the main plots from regression analyses, using linear and nonlinear models. (a) Quantile regression analysis with *P*-value of 0 and pseudo-$R^2$ of 0.67. In this plot, 47.67% and 47.79% of the predictions underestimated or overestimated support, respectively, considering the 95% confidence interval (indicated by the dashed line and surrounding shaded area). (b) Visualization of the change in quantile coefficients along with confidence intervals for the BS variable. The horizontal lines are the least-squares estimate (continuous line) and its confidence interval (dashed lines). (c) Stacked bar plot showing that applying the quantile regression model using the Box Cox transformation of *S* values results in predictions that frequently underestimated or overestimated support for each matrix analyzed. (d) Polynomial (11 terms) regression analysis with a root mean square error (RMSE) of 6.832121 and $R^2$ of 0.68. The 95% confidence interval is indicated by the dashed line and surrounding grey shaded area. The wider, light grey shaded area indicates the prediction interval. In this plot, 30.21% and 39.06% of the predictions underestimated or overestimated support considering the 95% confidence interval.

and decreased in 21.34% of the nodes. However, the minimum and maximum mean BS frequencies for different numbers of pseudoreplicates overlapped in 95.32% of the nodes (i.e., although the mean BS frequencies varied, the range of that variation was not exclusive to different numbers of pseudoreplicates). The mean $\hat{S}$ values for increasing numbers of replicates decreased in 92.22% of the nodes and never increased. The minimum and maximum mean $\hat{S}$ values overlapped in 63.63% of the nodes. See Table 1 below and Table S4 for details.

Although increasing the number of replicates during the computation of this index generally resulted in lower values of *S*, among the internal nodes whose *S* values decreased, $\hat{S}$ changed only −0.38% on average (Figure 2). In contrast, the variation among the internal nodes whose BS values changed when the number of pseudoreplicates increased was much more pronounced, changing −4.41% and 5.26% on average, indicating that BS is significantly more sensitive to variation in the number of replicates/pseudoreplicates than $\hat{S}$ is to variation in the number of replicates.

The data in Figure 2 and on Table S4 show that both $\hat{S}$ and BS varied more within the number of replicates/pseudoreplicates during their calculation than between different numbers of replicates/pseudoreplicates. This is expected, since both $\hat{S}$ and BS are prone to variation due to the heuristic nature of phylogenetic analysis. However, BS values are much more sensitive to variation in the number of pseudoreplicates than $\hat{S}$ is to the variation in the number of replicates.

The rank order of clades varied across almost all analyses for both *S* and BS (Table S4). The single

Table 1
Observed variation in bootstrap frequencies (BS) and support ($\hat{S}$) among 136 internal nodes of five datasets

| Matrix | Internal nodes (total nodes) | BS increase (%) | BS decrease (%) | $\hat{S}$ increase (%) | $\hat{S}$ decrease (%) | Overlaps in BS (%) | Overlaps in $\hat{S}$ (%) |
|---|---|---|---|---|---|---|---|
| M16688 | 37 (333) | 5.41 | 21.62 | 0 | 91.89 | 94.59 | 45.95 |
| M19355 | 40 (360) | 20 | 30 | 0 | 100 | 90 | 75 |
| M21278 | 27 (243) | 29.63 | 18.52 | 0 | 85.19 | 100 | 70.37 |
| M21595 | 7 (63) | 28.57 | 28.57 | 0 | 100 | 100 | 42.86 |
| M21596 | 25 (225) | 8 | 8 | 0 | 84 | 92 | 84 |
| TOTAL | 136 (1224) | 18.32 | 21.34 | 0 | 92.22 | 95.32 | 63.63 |

Bootstrap and $\hat{S}$ were calculated for 10, 100, and 1000 replicates/pseudo-pseudoreplicates, summing 1224 data points. The percentages indicate the number of nodes affected.

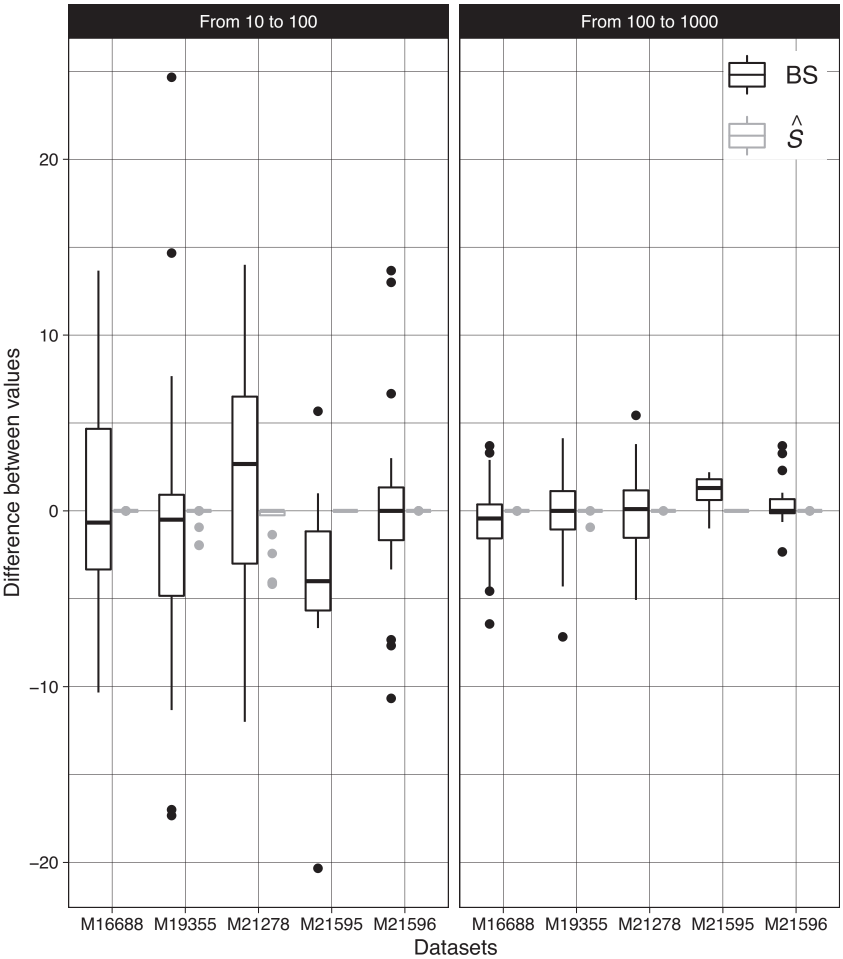

Fig. 2. Boxplots summarizing the differences between BS and $\hat{S}$ values calculated by subtracting the value corresponding to the smallest number of replicates/pseudoreplicates from the value corresponding to the greatest number of replicates/pseudoreplicates for all internal nodes in each dataset. The upper and lower boxplot hinges correspond to the first and third quartiles. Internal bars on each boxplot correspond to the median. The top and bottom whiskers extend to 150% of the respective interquartile range.

exception was dataset M21595, for which the rank order of clades remained constant across all analyses for $S$ but not BS. We estimate that variation in BS and $S$ affected 20.73% and 5.54% of clade ranking order, on average, respectively.

## Discussion

Common usage in phylogenetics aside, BS is a measure of confidence or uncertainty, not support. This is not merely a matter of semantics, as these concepts have different meanings and derive from vastly different statistical traditions, the former from the frequentist Neyman–Pearson approach concerned with Type I and Type II errors and the latter from the evidential approach concerned with the strength of evidence in favour of one hypothesis over another (Royall, 2004). It is widely known that BS does not vary in proportion to optimality, a fundamental desideratum of any measure of support, and that BS frequencies can be higher for clades that are unsupported by the evidence (i.e., absent in the optimal tree) than clades that are supported by the evidence (Grant & Kluge, 2008). Here, we have shown that BS is also not an efficient indirect measure of support, being an extremely unreliable predictor of the direct measure of support, *S*.

Although we see no justification for interpreting BS as either a direct or indirect measure of support in ML, this does not rule out its usefulness for other purposes. The discreteness of trees and violation of the assumptions that data were randomly sampled and are independent and identically distributed make rigorous statistical interpretations questionable (e.g., Anisimova & Gascuel, 2006; Carpenter, 1992; Farris, 1983; Galtier, 2004; Kluge & Wolf, 1993). Farris et al. (1996: p. 109) rejected Felsenstein's (1985) view of the BS as representing the statistical confidence level for clades and instead proposed the less ambitious interpretation of BS "simply as a way of discovering ambiguities in the data," although in practice they used BS frequencies as indicating group support. Alternatively, Wheeler (2010) interpreted BS as measuring "expected support," or the expectation, given a particular probability distribution on trees in tree space, of the occurrence of a clade over all trees. That is, it is not a measure of the degree of support a clade is expected to have, but rather the degree to which a clade is expected to be supported, which is equivalent to the repeatability interpretation described by Yang and Rannala (2005). Since expected support eschews any reference to either strength of evidence or estimation of the true tree and is explicitly descriptive only of the data at hand (Wheeler, 2010: p. 662), it is unclear what other inferences, if any, can be drawn from it.

Insofar as *S* is dependent only on the relationship between evidence and hypotheses given the background knowledge, it is a function of explanatory power and directly proportional to optimality. In this regard, although they employ different philosophical foundations, *S* is the ML analogue of other direct measures of support like Goodman-Bremer (GB) and ratio of explanatory power (REP) values in parsimony analysis (Grant & Kluge, 2007, 2010) and posterior odds ratio in Bayesian statistics (Bolstad, 2007; Grant & Kluge, 2008). As such, it also enjoys the advantageous properties of all those measures that have a clear and direct relationship to optimality, in contrast to clade credibility values (often referred to as posterior probabilities) in Bayesian phylogenetics and BS in parsimony, neither of which assesses the strength of evidence for competing hypotheses (Grant & Kluge, 2008; see also Goloboff et al., 2003; Goloboff et al., 2021).

A number of monophyly tests (internal branch tests) use *S* as their test statistic (e.g., Anisimova & Gascuel, 2006; Anisimova et al., 2011; Kishino & Hasegawa, 1989; Shimodaira & Hasegawa, 1999). However, those tests focus entirely on significance levels and error probabilities (i.e., Neyman–Pearson statistics) and not *S* itself. As such, two nodes considered equally supported because they have the same Shimodaira–Hasegawa-aLRT value, for example, need not have the same underlying *S*. Consequently, these methods are more closely related to BS than support.

Also, in order to decrease computation time these methods use heuristics expected to result in inflated values of *S*, including searching for the best trees lacking each node in the optimal tree using nearest-neighbour interchange (NNI) and reusing log-likelihood scores calculated for individual sites in the original analysis to avoid performing complete ML optimizations (i.e., resampling estimated log-likelihoods; Kishino et al., 1990). Although the precise extent to which those heuristics affect *S* remains to be determined, we confirmed that increasing the number of random addition sequence replicates from 10 to 100 and 1000 resulted in decreasing values of *S* (although values decreased very little, on average) and that the rank order of nodes varied among analyses. In contrast to the monotonic decreasing relationship between the number of replicates during the computation of *S*, we found that increasing the number of pseudoreplicates can both increase and decrease BS frequencies (see also Simmons & Freudenstein, 2011) and that the magnitude of the differences was much greater, further complicating interpretation of BS frequencies obtained under different search criteria.

Other, non-algorithmic alternatives to decrease the computation time required to calculate *S* also exist. For example, although it is customary to report support values for all nodes, systematists are usually interested in a far smaller number of nodes that are most relevant to specific taxonomic, evolutionary, or methodological questions. By targeting nodes of interest, the total number of searches can be greatly reduced, allowing computation time to be decreased without sacrificing the thoroughness of searches used to calculate *S* (Padial et al., 2014).

Nevertheless, given that finding the optimal tree for a given alignment is NP-complete (Garey & Johnson, 1977), both likelihood scores and support values should always be interpreted at best as upper bounds. As such, conventions regarding values that indicate weak, moderate, strong, and extremely strong support should be interpreted cautiously, taking dataset size and the number of replicates into account. Indeed, the naïve application of arbitrary thresholds and cut-offs for support values leads to some of the same problems inherent in common usage of *P*-values (Goodman, 2008). All nodes present in the optimal tree are supported by the evidence, and collapsing branches that have low *S* values entails the rejection of nodes that are supported by the evidence. More important than the absolute support value of a node is its value relative to other nodes, as this provides a means to heuristically guide future research cycles (Grant & Kluge, 2003). That is, regardless of their absolute values, less contradictory evidence is required to overturn a node with a lower *S* than another node with a higher *S*.

## Conclusions

The log-likelihood difference (*S*) is a direct measure of support in ML. It is simple to interpret and easy to calculate, with numerous algorithmic and non-algorithmic approaches to decrease computation time. Increasing the number of replicates during tree searches usually decreases *S* values, although the effect was small in the datasets we analyzed. In contrast, BS is neither a direct measure of support nor an efficient indirect measure of support, and increasing the number of pseudoreplicates both increases and decreases BS values, with the magnitude of differences being much greater than the magnitude of changes in *S* caused by increasing the number of random addition sequences replicates during tree searchers.

Nevertheless, the fact that BS fails to measure support is not a criticism of the method per se, as it was not originally intended to measure support. Just as *P*-values and support have been conflated in statistics generally (Schervish, 1996), a host of different concepts have been conflated with support in phylogenetics, including reliability, repeatability, confidence, and significance. One may argue that any and all of these different measures can be called “support,” but that semantic argument misses the point that they all measure something different, and using the same term for all has been a lasting source of confusion. In a recent example, Evangelista et al. (2018) examined a plethora of conceptually and methodologically different “support” measures, concluding that “assessing the strength of a phylogenetic hypothesis [i.e., the strength of the evidence in favour of a hypothesis] requires integration of tests and metrics” (p. 120). We believe they arrived at that pluralist conclusion because not one of the tests and metrics they examined actually assessed the strength of the hypothesis. Certainly, they all measured *something* and provide potentially useful information, but the strength of evidence in favour of one hypothesis over another (i.e., support) is unequivocally measured in ML by either LR or *S*, which were not included in their study. The precise use of terminology is a first step to understanding the utility of the different approaches and their relationships to each other, both conceptually and empirically.

Although BS is not an efficient indirect measure of support, other methods might be (for a current list of options of support measures in RAxML, Garli, and IQ-Tree, see Table S5). For example, RAxML (Stamatakis, 2014) uses fast heuristics to measure internode certainty (IC) and IC All (ICA) to quantify the log magnitude of the difference between the two most prevalent conflicting bipartitions or all most prevalent conflicting bipartitions, respectively. Both IC and ICA aim to calculate the degree of incongruence for a given internode and are similar to tree certainty (TC) or TC All (TCA) measures for the entire tree (Salichos et al., 2014). Similarly, IQ-Tree (Minh et al., 2020b) uses fast heuristics to calculate the gene concordance factor (gCF) and the novel site concordance factor (sCF). The gCF describes the proportion of inferred single-locus trees that contain a given branch, and the sCF represents the percentage of decisive alignment sites with character-state transformations on a branch (Minh et al., 2020a). If these or other measures are faster to calculate and also strongly correlated with *S*, then they would be efficient indirect measures of support; if not, then they are still potentially useful as descriptors of different characteristics of the data, but they would not be justifiably interpreted as support measures.

## Acknowledgements

We thank Gabriel Massami Izumi de Freitas for the helpful discussion and assistance during the literature review. We also thank Dr. Paulo Inácio K. L. Prado for the useful discussion regarding our statistical analyses. Finally, we acknowledge our three reviewers for their constructive criticism. This work was made possible thanks to the financial support from the São Paulo Research Foundation (FAPESP; grants 2012/10000-5, 2013/05958-8, 2015/18654-2, 2018/15425-0) and Brazilian National Council for Scientific and Technological Development (CNPq; grant 306823/2017-9).

## Conflict of interest

None declared.

## Data availability statement

Articles used in our literature review are listed in Table S1. Empirical data used in the current study comprised 106 matrices of DNA sequences downloaded from TreeBASE (https://treebase.org). Information about the TreeBASE datasets is given in Table S2. The original software is freely available under the GNU General Public License version 3.0 (GPL-3.0) at https://grant.ib.usp.br/anfibios/researchSoftware.html, https://gitlab.com/MachadoDJ/sudoparallelgarli, and https://gitlab.com/MachadoDJ/s4ml. Additional data and statistical analyses are provided as supplementary materials.

## References

Anisimova, M. & Gascuel, O. (2006) Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. Systematic Biology, 55, 539–552. Available from: https://doi.org/10.1080/10635150600755453

Anisimova, M., Gil, M., Dufayard, J.F., Dessimoz, C. & Gascuel, O. (2011) Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. Systematic Biology, 60, 685–699. Available from: https://doi.org/10.1093/sysbio/syr041

Bardin, J., Rouget, I. & Cecca, F. (2016) Ontogenetic data analyzed as such in phylogenies. Systematic Biology, 66(1), 23–37. Available from: https://doi.org/10.1093/sysbio/syw052

Bolstad, W.M. (2007) Introduction to Bayesian statistics, 2nd edition. Hoboken, NJ: John Wiley & Sons.

Boyd, B.M., Allen, J.M., Nguyen, N., Sweet, A.D., Warnow, T., Shapiro, M.D. et al. (2017) Phylogenomics using target-restricted assembly resolves intra-generic relationships of parasitic lice (Phthiraptera: *Columbicola*). Systematic Biology, 66(6), syx027. Available from: https://doi.org/10.1093/sysbio/syx027

Cahusac, P. (2020) Data as evidence. Experimental Physiology, 105 (7), 1071–1080. Available from: https://doi.org/10.1113/EP088664

Carpenter, J.M. (1992) Random cladistics. Cladistics, 8(2), 147–153.

Chang, J.-M., Floden, E.W., Herrero, J., Gascuel, O., Di Tommaso, P. & Notredame, C. (2019) Incorporating alignment uncertainty into Felsenstein's phylogenetic bootstrap to improve its reliability. Bioinformatics, 37, 1506–1514. Available from: https://doi.org/10.1093/bioinformatics/btz082

Chatzou, M., Floden, E.W., Di Tommaso, P., Gascuel, O. & Notredame, C. (2018) Generalized bootstrap supports for phylogenetic analyses of protein sequences incorporating alignment uncertainty. Systematic Biology, 67(6), 997–1009. Available from: https://doi.org/10.1093/sysbio/syx096

Cover, T.M. (1999) Elements of information theory, 1st edition. Hoboken, NJ: John Wiley & Sons.

Edwards, A. (1972) Likelihood. Cambridge: Cambridge University Press.

Efron, B. (1979) Computers and the theory of statistics: thinking the unthinkable. SIAM Review, 21(4), 460–480.

Evangelista, D., Thouzé, F., Kohli, M.K., Lopez, P. & Legendre, F. (2018) Topological support and data quality can only be assessed through multiple tests in reviewing Blattodea phylogeny. Molecular Phylogenetics and Evolution, 128, 112–122. Available from: https://doi.org/10.1016/j.ympev.2018.05.007

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

Farris, J.S., Albert, V.A., Källersjö, M., Lipscomb, D. & Kluge, A.G. (1996) Parsimony jackknifing outperforms neighbor-joining. Cladistics, 12(2), 99–124. Available from: https://doi.org/10.1006/clad.1996.0008

Felsenstein, J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution, 39(4), 783–791. Available from: https://doi.org/10.1111/j.1558-5646.1985.tb00420.x

Galtier, N. (2004) Sampling properties of the bootstrap support in molecular phylogeny: influence of nonindependence among sites. Systematic Biology, 53(1), 38–46.

Garey, M.R. & Johnson, D.S. (1977) The rectilinear Steiner tree problem is NP-complete. SIAM Journal on Applied Mathematics, 32(4), 826–834.

Goloboff, P.A., Catalano, S.A. & Torres, A. (2021) Parsimony analysis of phylogenomic datasets (II): evaluation of PAUP*, MEGA and MPBoot. Cladistics. Available from: http://dx.doi.org/10.1111/cla.12476

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

Goodman, S. (2008) A dirty dozen: twelve p-value misconceptions. Seminars in Hematology, 45(3), 135–140. Available from: https://doi.org/10.1053/j.seminhematol.2008.04.003

Goodman, S.N. & Royall, R. (1988) Evidence and scientific research. American Journal of Public Health, 12, 1568–1574. Available from: https://doi.org/10.2105/ajph.78.12.1568

Grant, T. & Kluge, A.G. (2003) Data exploration in phylogenetic inference: scientific, heuristic, or neither. Cladistics, 19(5), 379–418. Available from: https://doi.org/10.1111/j.1096-0031.2003.tb00311.x

Grant, T. & Kluge, A.G. (2007) Ratio of explanatory power (REP): a new measure of group support. Molecular Phylogenetics and Evolution, 44(1), 483–487.

Grant, T. & Kluge, A.G. (2008) Clade support measures and their adequacy. Cladistics, 24(6), 1051–1064. Available from: https://doi.org/10.1111/j.1096-0031.2008.00231.x

Grant, T. & Kluge, A.G. (2010) REP provides meaningful measurement of support across datasets. Molecular Phylogenetics and Evolution, 55, 340–342. Available from: https://doi.org/10.1016/j.ympev.2009.10.028

Hacking, I. (1965) The logic of statistical inference. Cambridge, UK: Cambridge University Press.

Kishino, H. & Hasegawa, M. (1989) Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. Journal of Molecular Evolution, 29(2), 170–179.

Kishino, H., Miyata, T. & Hasegawa, M. (1990) Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. Journal of Molecular Evolution, 31, 151–160. Available from: https://doi.org/10.1007/BF02109483

Klötzl, F. & Haubold, B. (2016) Support values for genome phylogenies. Life, 6(1), 11. Available from: https://doi.org/10.3390/life6010011

Kluge, A.G. & Wolf, A.J. (1993) Cladistics: what's in a word? Cladistics, 9(2), 183–199.

Koenker, R. (2005) Quantile Regression (Econometric Society Monographs). New York, NY: Cambridge University Press.

Kumar, S., Filipski, A.J., Battistuzzi, F.U., Kosakovsky Pond, S.L. & Tamura, K. (2012) Statistics and truth in phylogenomics. Molecular Biology and Evolution, 29(2), 457–472. Available from: https://doi.org/10.1093/molbev/msr202

Lee, M.S.Y. & Hugall, A.F. (2003) Partitioned likelihood support and the evaluation of data set conflict. Systematic Biology, 52, 15–22.

Lemoine, F., Domelevo Entfellner, J.-B., Wilkinson, E., Correia, D., Dávila Felipe, M., de Oliveira, T. et al. (2018) Renewing Felsenstein's phylogenetic bootstrap in the era of big data. Nature, 556(7702), 452–456. Available from: https://doi.org/10.1038/s41586-018-0043-0

Linkem, C.W., Minin, V.N. & Leaché, A.D. (2016) Detecting the anomaly zone in species trees and evidence for a misleading signal in higher-level skink phylogeny (Squamata: Scincidae). Systematic Biology, 65(3), 465–477. Available from: https://doi.org/10.1093/sysbio/syw001

Marques, F.P.L. & Caira, J.N. (2016) *Pararhinebothroides* – neither the sister-taxon of *Rhinebothroides* nor a valid genus. Journal of Parasitology, 102(2), 249–259. Available from: https://doi.org/10.1645/15-894

Meireles, C.M., Czelusniak, J., Schneider, M.P.C., Muniz, J.A.P.C., Brigido, M.C., Ferreira, H.S. et al. (1999) Molecular phylogeny of ateline new world monkeys (Platyrrhini, Atelinae) based on γ-globin gene sequences: evidence that *Brachyteles* is the sister group of *Lagothrix*. Molecular Phylogenetics and Evolution, 12 (1), 10–30. Available from: https://doi.org/10.1006/mpev.1998.0574

Meyer, P.E. (2008) Information-theoretic variable selection and network inference from microarray data. Ph.D. thesis of the Universite Libre de Bruxelles.

Minh, B.Q., Hahn, M.W. & Lanfear, R. (2020a) New methods to calculate concordance factors for phylogenomic datasets. Molecular Biology and Evolution, 37(9), 2727–2733. Available from: https://doi.org/10.1093/molbev/msaa106

Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., Von Haeseler, A. et al. (2020b) IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Molecular Biology and Evolution, 37(5), 1530–1534. Available from: https://doi.org/10.1093/molbev/msaa015

Nagelkerke, N.J.D. (1991) A note on the general definition of the coefficient of determination. Biometrika, 78(3), 691–692.

Padial, J.M., Grant, T. & Frost, D.R. (2014) Molecular systematics of terraranas (Anura: Brachycephaloidea) with an assessment of the effects of alignment and optimality criteria. Zootaxa, 3825(1), 1–132.

Ranjan, C. & Najari, V. (2020) Package "nlcor:" compute nonlinear correlations. Research Gate. Available from: https://doi.org/10.13140/RG.2.2.33716.68480

Royall, R.M. (2004) The likelihood paradigm for statistical evidence. In: Taper, M.L. & Lele, S.R. (Eds.) The nature of scientific evidence: empirical, statistical, and philosophical considerations. Chicago and London: University of Chicago, pp. 119–152. Available from: https://doi.org/10.7208/chicago/9780226789583.003.0005

Salichos, L., Stamatakis, A. & Rokas, A. (2014) Novel information theory-based measures for quantifying incongruence among phylogenetic trees. Molecular Biology and Evolution, 31(5), 1261–1271. Available from: https://doi.org/10.1093/molbev/msu061

Schervish, M.J. (1996) *P* values: what they are and what they are not. The American Statistician, 50(3), 203–206.

Shimodaira, H. & Hasegawa, M. (1999) Likelihood-based tests of topologies in phylogenetics. Molecular Biology and Evolution, 16 (8), 1114–1116.

Simmons, M.P. & Freudenstein, J.V. (2011) Spurious 99% bootstrap and jackknife support for unsupported clades. Molecular Phylogenetics and Evolution, 61(1), 177–191. Available from: https://doi.org/10.1016/j.ympev.2011.06.003

Simon, C. (2020) An evolving view of phylogenetic support. Systematic Biology, syaa068. Available from: https://doi.org/10.1093/sysbio/syaa068

Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics, 30(9), 1312–1313.

Taper, M.L. & Lele, S.R. (2004) The nature of scientific evidence: a forward-looking synthesis. In: Taper, M.L. & Lele, S.R. (Eds.) The nature of scientific evidence: empirical, statistical, and philosophical considerations. Chicago and London: University of Chicago, pp. 527–551. Available from: https://doi.org/10.7208/chicago/9780226789583.003.0016

Taper, M.L. & Ponciano, J.M. (2016) Evidential statistics as a statistical modern synthesis to support 21st century science. Population Ecology, 58(1), 9–29.

Trevisan, B., Primon, J.F. & Marques, F.P.L. (2017) Systematics and diversification of *Anindobothrium* Marques, Brooks & Lasso, 2001 (Eucestoda: Rhinebothriidea). PLoS One, 12(9), e0184632. Available from: https://doi.org/10.1371/journal.pone.0184632

Wang, H.-C., Susko, E. & Roger, A.J. (2016) Split-specific bootstrap measures for quantifying phylogenetic stability and the influence of taxon selection. Molecular Phylogenetics and Evolution, 105, 114–125. Available from: https://doi.org/10.1016/j.ympev.2016.08.017

Wang, W. & Liu, K.J. (2020) Build a better bootstrap and the RAWR shall beat a random path to your door: phylogenetic support estimation revisited. BioRxiv. Available from: https://doi.org/10.1101/2020.02.02.931063

Wang, W., Smith, J., Hejase, H.A. & Liu, K.J. (2020) Non-parametric and semi-parametric support estimation using SEquential RESampling random walks on biomolecular sequences. Algorithm. Mol. Biol, 15(1), 1–15. Available from: https://doi.org/10.1186/s13015-020-00167-0

Wang, Y., Li, Y., Cao, H., Xiong, M., Shugart, Y.Y. & Jin, L. (2015) Efficient test for nonlinear dependence of two continuous variables. BMC Bioinformatics, 16(1), 1–8.

Wheeler, W.C. (2010) Distinctions between optimal and expected support. Cladistics, 26(6), 657–663. Available from: https://doi.org/10.1111/j.1096-0031.2010.00308.x

Yang, Z. & Rannala, B. (2005) Branch-length prior influences Bayesian posterior probability of phylogeny. Systematic Biology, 54(3), 455–470.

Yuan, Z.-Y., Zhou, W.-W., Chen, X., Poyarkov, N.A., Chen, H.-M., Jang-Liaw, N.-H. et al. (2016) Spatiotemporal diversification of the true frogs (genus *Rana*): a historical framework for a widely studied group of model organisms. Systematic Biology, 65 (5), 824–842. Available from: https://doi.org/10.1093/sysbio/syw055

Zwickl, D.J. (2006) Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. [Doctoral dissertation, The University of Texas at Austin]. Available at: http://hdl.handle.net/2152/2666

## Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

**Table S1**. Literature review of the number of papers using bootstrap to measure clade support in maximum likelihood phylogenetic analysis.

**Table S2**. This table contains TreeBASE accession numbers and the number of taxa and characters per dataset.

**Table S3**. Values used for testing the proportionality between BS and *S*.

**Table S4**. Sensitivity of BS and *S* to the increasing number of replicates/pseudoreplicates.

**Table S5**. Options of support measures in RAxML, Garli, and IQ-Tree.

**Appendix S1**. R Markdown document with detailed statistical analysis exploring the potential proportionality between BS and *S*.