Robustness of Felsenstein's Versus Transfer Bootstrap Supports With Respect to Taxon Sampling

Paul Zaharias, Frédéric Lemoine, Olivier Gascuel

Syst. Biol. XX(XX):1–16, 2023

https://doi.org/10.1093/sysbio/syad052
Advance Access Publication August 10, 2023

Robustness of Felsenstein's Versus Transfer Bootstrap Supports With Respect to Taxon Sampling

Paul Zaharias[1,*], Frédéric Lemoine[2,3], and Olivier Gascuel[1,*]

[1]*Institut de Systématique, Evolution, Biodiversité (ISYEB UMR7205—CNRS, Muséum National d'Histoire Naturelle, SU, EPHE, UA), 75005 Paris, France*
[2]*Institut Pasteur, Université Paris Cité, G5 Evolutionary Genomics of RNA Viruses, 75015, Paris, France*
[3]*Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, 75015, Paris, France*
**Correspondence to be sent to: Institut de Systématique, Evolution, Biodiversité (ISYEB UMR7205—CNRS, Muséum National d'Histoire Naturelle, SU, EPHE, UA), 75005 Paris, France; E-mail: paul.zaharias@mnhn.fr; olivier.gascuel@mnhn.fr.*

Received 24 February 2022; reviews returned 26 June 2023; accepted 09 August 2023
Associate Editor: Siavash Mirarab

Abstract.—The bootstrap method is based on resampling sequence alignments and re-estimating trees. Felsenstein's bootstrap proportions (FBP) are the most common approach to assess the reliability and robustness of sequence-based phylogenies. However, when increasing taxon sampling (i.e., the number of sequences) to hundreds or thousands of taxa, FBP tend to return low support for deep branches. The transfer bootstrap expectation (TBE) has been recently suggested as an alternative to FBP. TBE is measured using a continuous transfer index in [0,1] for each bootstrap tree, instead of the binary {0,1} index used in FBP to measure the presence/absence of the branch of interest. TBE has been shown to yield higher and more informative supports while inducing a very low number of falsely supported branches. Nonetheless, it has been argued that TBE must be used with care due to sampling issues, especially in datasets with a high number of closely related taxa. In this study, we conduct multiple experiments by varying taxon sampling and comparing FBP and TBE support values on different phylogenetic depths, using empirical datasets. Our results show that the main critique of TBE stands in extreme cases with shallow branches and highly unbalanced sampling among clades, but that TBE is still robust in most cases, while FBP is inescapably negatively impacted by high taxon sampling. We suggest guidelines and good practices in TBE (and FBP) computing and interpretation. [Felsenstein's bootstrap; phylogenetic trees; support robustness; taxon sampling; transfer bootstrap.]

Branch supports are essential for interpreting trees because they allow quantifying the degree of uncertainty in our phylogenetic hypotheses. For maximum-likelihood (ML) tree estimation (and other approaches, e.g., distance-based), the most popular branch support is undeniably the bootstrap method proposed by Felsenstein (1985), one of the most cited articles of all time (Van Noorden et al. 2014). The procedure relies on resampling with the replacement of the sites of a reference alignment until obtaining a pseudo- (or bootstrap) alignment of the same length. Then, pseudo- (or bootstrap) trees are estimated using the same inference method. Finally, the support for every branch on the reference tree is measured as the bootstrap proportion (BP) of pseudo-trees containing that branch.

The interpretation of bootstrap support values has led to great controversies in the 1990s (reviewed in Sanderson 1995; Soltis and Soltis 2003; Simon 2022). Originally, Felsenstein (1985) suggested interpreting it as a measure of repeatability, meaning the "probability that a specified group will be found in an analysis of an independent sample of characters" (Hillis and Bull 1993). However, this interpretation has been contrasted with another view that bootstrap could be interpreted as a confidence region of some kind in a null hypothesis framework (Hillis and Bull 1993), and was further discussed in the literature (Felsenstein and Kishino 1993; Sanderson 1995; Efron et al. 1996; Susko 2009) but did not find success among practitioners. Through simulations, Hillis and Bull (1993) suggested a threshold value of 70%, although this value was proposed under very specific conditions, namely "equal rates of change," "symmetric phylogenies," and "internodal change of <20% of the characters." Despite those important limitations, Soltis and Soltis (2003) note that "many systematists have adopted Hillis and Bull's '70%' value as an indication of support," an observation still true today, even though the "95%" and the more arbitrary "80%" cut-offs can often be seen in the literature.

Aside from conceptual matters, two main criticisms against FBP are recurring in the literature, especially for "large" datasets. By large datasets, we mean in this study matrices where the number of taxa is large, while the number of sites remains moderate, usually corresponding to a single gene or phylogenetic marker. The first limitation is technical: re-estimation (usually by ML) of pseudo-trees is computationally demanding on large datasets and can be impractical on very large datasets. A rapid bootstrap support (RBS; Stamatakis et al. 2008) was proposed and consists in using some of the trees found during the ML topological research as bootstrap trees. The RBS is fast and reliable, but it remains computationally intensive and tends to be more liberal than standard FBP (Anisimova et al. 2011). Even faster

is the Ultrafast Bootstrap approximation approach (UFBoot; Minh et al. 2013; Hoang et al. 2018), but recent works suggest that it is considerably more liberal than standard FBP (and RBS), questioning the comparability of UFBoot to standard bootstrap (Gascuel and Lemoine 2022). The second main criticism is FBP's sensitivity to rogue taxa (Wilkinson 1996), that is, taxa whose position varies from (pseudo-)trees to (pseudo-)trees. Indeed, if a single taxon is unstable (e.g., due to homoplasy or missing data) in the overall tree or in a particular region of it, then the FBP support values are expected to be considerably lowered in that region. Other criticisms discuss the problem of "large" datasets in the sense of "large number of sites" (e.g., Sharma and Kumar 2021), but this will not be covered here as it is not the subject of this study.

The commonly used alternatives to Felsenstein Bootstrap Support (FBP) can be divided into 2 main classes. The first one is the Posterior Probability (PP) of Bayesian phylogenetics (Rannala and Yang 1996), calculated as the proportion of all sampled trees in the MCMC chain (post burn-in) in which the branch of interest is found. Due to their parametric nature and implementation, PP is often considered liberal, as opposed to the more conservative FBP. Abundant literature exists on comparing BP and PP (e.g., Douady et al. 2003), but expanding on that matter goes beyond the frame of this article. Local supports are another class of alternative supports, responding to a need for computational speed in a context of ever-growing datasets. These supports are obtained by locally rearranging the tree topology around the branch of interest, using Nearest Neighbor Interchange. Some of the most popular ones include the approximate likelihood-ratio test (aLRT; Anisimova and Gascuel 2006) and the non-parametric Shimodaira–Hasegawa-like version (aLRT SH-like; Guindon et al. 2010). However, these approaches only provide a local view of the support. It is also important to note that none of these two classes explicitly addresses the problem or rogue taxa: Bayesian PP is expected to behave similarly to FBP with a tree that contains rogues; the local supports are little affected by the presence of a few rogues, but they are also unable to detect them and measure their impact on the overall tree due to their local nature.

Recently, an alternative to FBP has been proposed: the transfer bootstrap expectation (TBE; Lemoine et al. 2018). TBE in itself is fast to compute (Lutteropp et al. 2020) but is also based on resampling and re-estimating pseudo-trees, and thus is overall computationally heavy, although easily parallelizable and applicable with RBS and UFBoot. The difference lies in the comparison of the pseudo-trees to the reference tree. Rather than the binary presence/absence of a reference branch in the pseudo-trees, TBE uses a "transfer" distance that is measured using the number of taxa that must be transferred (or removed) to make two branches identical (Lemoine et al. 2018). Because of its continuous nature, TBE scores are always higher than FBP, except for cherries (i.e., a clade comprising only two taxa) and when FBP = 100%, where both supports are identical. TBE is also less affected by rogues and it has a natural interpretation: on a given reference branch, we can easily calculate the average number of taxa that have to be transferred to recover that branch in bootstrap trees. Finally, yet importantly, results with real and simulated data showed that TBE induces very few falsely supported branches when used with common thresholds. This does not mean that the supported branches are entirely correct, as implicitly assumed with FBP, but that they are nearly correct with a low level of conflict with the true branches. While their initial results indicate that 70% is a reasonable threshold for supporting branches that are at least 95% accurate (based on a quartet distance to the true tree), Lemoine et al. (2018) suggest that it is "better to interpret TBE values depending on the data and the phylogenetic question being addressed." TBE also has solid mathematical ground and is guaranteed to converge in probability to 0 when the size of the tree grows and there is no signal in the data (Dávila Felipe et al. 2019).

TBE was proposed in a context of the ever-growing number of sequences, in particular in the epidemiology field. It has become recurrent in the literature to find extremely large datasets (in terms of number of sequences): bacteria/archaea (10,575 tips; Zhu et al. 2019), mushrooms (5284 tips; Varga et al. 2019), fishes (31,526 tips; Rabosky et al. 2018), diatoms (19,197 tips; Lewitus et al. 2018), angiosperms (36,101 tips; Janssens et al. 2020), and especially viruses-like HIV (9147 tips, Lemoine et al. 2018) and of course SARS-CoV-2 (>15 million tips; Turakhia et al. 2022). In datasets of this size, it is expected that some of the taxa will behave like rogues, thus lowering FBP values. In fact, it is not rare to find FBP scores below 20% (or even at 0%), especially for deep branches, even when those have a strong phylogenetic signal. Yet these deep branches are usually the primary focus of large-scale studies. Lemoine et al. (2018) showed that TBE was in fact able to support deep branches that have strong phylogenetic signals without being affected by a few rogues. TBE is available on the BOOSTER platform (https://booster.pasteur.fr/) and in software programs like Gotree (Lemoine and Gascuel 2021), PhyML (Guindon et al. 2010), Seaview (Gouy et al. 2021), IQ-TREE 2 (Minh et al. 2020), and RAxML-NG (Kozlov et al. 2019).

Since its publication 5 years ago, TBE has been cited more than 400 times (Google Scholar, June 2023), but has generated little debate. However, a recent review of the history and development of branch support measures (Simon 2022) contained a stimulating critique communicated by Nick Goldman, that "TBE must be used with care due to sampling issues. For example, if many closely related taxa are added to the tree TBE values will increase across the entire tree. This is because the measure is based on counting the number of sequences sampled, not taking into account their variation." This critique raises a series of important questions related to sampling variation (or taxon sampling) and sampling disequilibrium in datasets with large numbers of taxa.

In this study, we will explore the impact on both FBP and TBE of taxon sampling, rogue taxa, and sampling disequilibrium in large datasets, using theoretical examples and empirical datasets. We provide a series of guidelines in the "Discussion" section for good practices and interpretation when estimating FBP and TBE support values on large datasets.

Theoretical results

In this section, we explore two theoretical cases with sampling variation and presence of rogues. The goal is mainly pedagogical, partly to address the critique published by Simon (2022), but it is unlikely that any of these examples could be found as such in real data, which we explore further in the next section. To facilitate the reader's comprehension, we remind here the basic formulae for FBP and TBE.

In FBP, the support of a branch (or bipartition) in a tree T is the proportion of bootstrap trees T^* in which the bipartition is present. If a bipartition b^* from T^* is found identical as b in T, then the support of b given T^* is equal to 1, else it is equal to 0. As opposed to the binary nature of FBP, TBE was proposed as a continuous version, using the transfer distance. The distance $\delta(b,b^*)$ is equal to the number of taxa that must be transferred (or removed) to make both bipartitions identical. The transfer index $\Phi(b,T^*)$ is the minimum of the transfer distances for all bipartitions b^* present in T^*: $\Phi(b,T^*) = Min_{b^*\in T^*}\{\delta(b,b^*)\}$. This index is then normalized in [0,1] and averaged over all bootstrap trees. The normalization relies on p, the size of the smallest of the two subsets of taxa defined by b. It is easily seen that $\Phi(b,T^*) \leq p-1$. The normalized version of the transfer index is thus equal to $1-\Phi(b,T^*)/(p-1)$, and the TBE support is defined by: $\mathrm{TBE}(b) = 1-\overline{\Phi(b,T^*)}/(p-1)$, where the bar denotes the average over all bootstrap trees. It follows that $\mathrm{TBE}(b) \geq \mathrm{FBP}(b)$ when using the same set of bootstrap trees. The difference between the two supports depends on p and thus the depth of b (as in Lemoine et al. 2018, we assume here and in the whole article that the depth of a branch is measured by p). If $p = 2$ (i.e., a cherry), then both supports are equal, because $\Phi(b,T^*)$ values can only be 0 or 1 (i.e., the cherry is recovered in T^* or absent). If b is a deep branch and p is large, there will often be a large difference between the two supports. This is what we will explore in this study, first through theoretical examples and then with real data.

Impact of Duplicated Sequences on TBE Scores

In his personal communication to Simon (2022), Nick Goldman suggests that, for a given phylogenetic signal, TBE supports will be increased if we add many closely related taxa to the tree. A model that closely matches this hypothesis is presented hereafter. Let $P|Q$ be a bipartition of reference tree T, where P and Q have sampling size p and q, respectively, $p \leq q$ and $p+q = n$, where n corresponds to the total number of taxa (p is again the depth of bipartition/branch $P|Q$). To keep the same phylogenetic signal while increasing the sampling, we simply add duplicated taxa (i.e., strictly identical sequences). The number of taxa will be multiplied by a factor k, where k represents the number of duplicated sequences, $k = 1$ corresponds to the original dataset, and p and q are transformed into kp and kq when duplicates are added. With any reasonable phylogenetic program, the reference and bootstrap trees will remain essentially the same, with all duplicated sequences grouped together into clusters that form the "tips" of the tree, while the rest of the tree and the internal branches are unchanged. In this model, the phylogenetic signal remains the same and the FBP support is unchanged regardless of the value of k.

Let us now describe the behavior of the TBE supports. In the absence of duplicates ($k = 1$), the TBE support of $P|Q$ (let us call it $\sigma(1)$) is equal to $1-\tau/(p-1)$, where τ is the average number of taxa that need to be transferred to retrieve $P|Q$ in the bootstrap trees. Then, we have $\tau = (1-\sigma(1))(p-1)$. Let us now assume that each sequence is duplicated k times. Then, it is easily seen that the support of the duplicated version of $P|Q$ is equal to:

$$\sigma(k) = 1-\frac{k\tau}{kp-1} = 1-\frac{k(1-\sigma(1))(p-1)}{kp-1}.$$

Using the derivative, we note that $\sigma(k)$ is an increasing function of k. When k is large, $\sigma(k)$ converges to $\sigma_{max} = \sigma(1)(1-1/p)+1/p$, and with large p we have $\sigma(1) \approx \sigma_{max} \approx \sigma(k)$ for any value of k. In other words, the support of large clades (deep branches) remains unchanged when adding duplicates. However, things are different when p is small (i.e., close to 2). Figure 1 shows the TBE supports for different values of p and k, and for $\sigma(1) = 0.3$, 0.7, and 0.9 (i.e., very low, moderate, and high phylogenetic signal). We see that sampling redundancy (i.e., the addition of duplicates) has a strong impact on cherries ($p = 2$), much less on small clades ($p = 3,5,10$), and very little when the clade is large ($p = 100$). The gap between FBP and TBE is maximized for cherries: with $p = 2$, FBP = TBE and FBP is not impacted by duplicates, whereas TBE is clearly impacted and increases with the number of duplicates, especially if the support is low.

This simple model allows us to capture some trends in TBE supports. In line with Goldman's hypothesis, the TBE supports of very small clades will indeed increase when duplicates are added to the tree, whereas in this model, the FBP supports remain the same. However, this is much less true for medium-sized clades (e.g., 10–30 taxa), and for large clades, the TBE supports are practically unchanged. Furthermore, the removal of duplicates before tree inference should be done systematically and is already considered good practice by most phylogeny software. We will confirm these results on TBE supports with our empirical data sets in the following sections. FBP is robust in this example, but

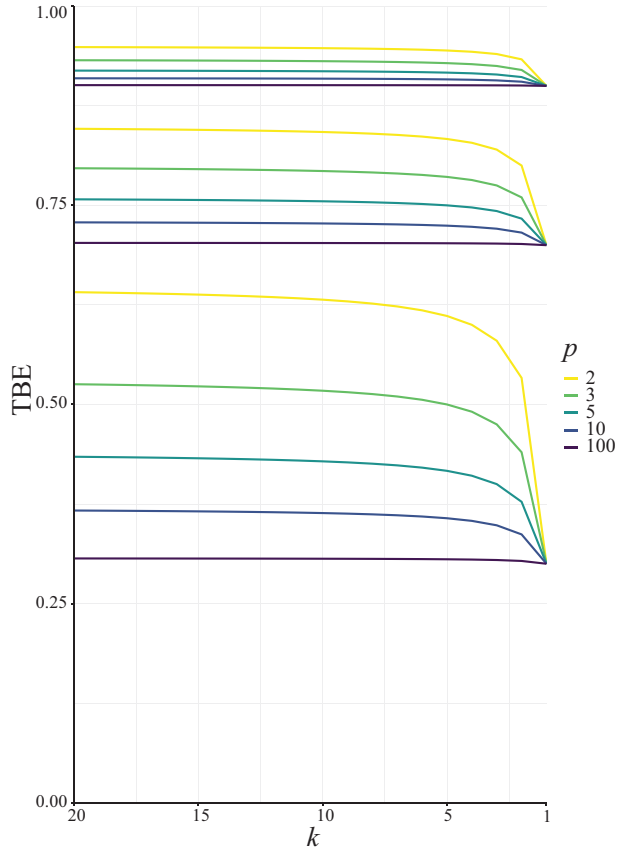

FIGURE 1. Variation of TBE supports in the presence of duplicates. Horizontal axis: $k = 1$ corresponds to the original tree with no duplicates, whereas with $k = 20$ each taxon is duplicated 20 times; the largest value of k is on the left to facilitate understanding of results with real data, where we progressively remove taxa (see below). Vertical axis: TBE support. Colorings: p corresponds to the depth of the branch in the original tree (e.g., $p = 2$ is a cherry, in yellow). The 3 sets of curves correspond to different values of the original TBE support without duplicates: $\sigma(1) = 0.9$ on top, 0.7 in the middle, and 0.3 on bottom.

not in our next example, which has no duplicates and a strong signal, but where we introduce a few rogue taxa.

Presence of Rogue Taxa When the Tree is Globally Supported

As in the previous example, let us consider again a bipartition $P|Q$, where P and Q have sampling size p and q, respectively, $p \leq q$ and $p+q = n$. We assume now that the signal is globally strong but affected by the presence of a few rogue taxa. Hence, the bipartition $P|Q$ is "almost" found in bootstrap trees, but every taxon has a small probability π (in the order of $1/n$) of being rogue and randomly placed in P or Q in the bootstrap trees, following a uniform model with a probability of being assigned to $P = p/n$ and probability of being assigned to $Q = q/n$. In this model, we can easily compute FBP and TBE supports as a function of π, p, and n.

If a taxon of P in the reference tree is rogue, its probability of being wrongly placed in a bootstrap tree is $q/n = (n-p)/n$. Conversely, the probability for a rogue taxon of Q to be wrongly placed is p/n. Thus, the total probability for a rogue to be wrongly placed is equal to $2p(n-p)/n^2$ and the probability for a taxon to be rogue and wrongly placed is equal to $2\pi p(n-p)/n^2$. Therefore, the expected FBP support is equal to the probability that no taxon is rogue and wrongly placed: $\mathrm{FBP}(\pi, p, n) = (1 - 2\pi p(n-p)/n^2)^n$. As for TBE, the expected number of rogues is equal to πn, and so the TBE support is approximated by $\mathrm{TBE}(\pi, p, n) \approx 1 - 2\pi p(n-p)/n(p-1)$ (we assume here that π represents a small fraction of the taxa, making the perturbed-by-rogues version of $P|Q$ the closest bipartition to $P|Q$ in bootstrap trees).

Figure 2 shows the evolution of supports for FBP and TBE when varying π {0.001, 0.005, 0.01, 0.05} and p {2, 3 ... 500} in a 1000-taxon tree ($n = 1000$). With cherries ($p = 2$) both supports are the same, as explained before. For FBP, the support depends on the number of rogues (π), but also strongly on the sampling size p of the clade P. Figure 2 confirms that FBP drops extremely fast when the number of rogues is high (i.e., ~50 rogues with $\pi = 0.05$). Furthermore, FBP also drops for large clades even with a very low number of rogues (e.g., ~1 rogue with $\pi = 0.001$; then FBP $< 70\%$ when $p > 232$). The TBE support, on the other hand, behaves in a diametrically opposed way to the FBP support, but with much less variation; TBE rapidly increases when p is very low (between 2 and 5), and then slowly increases as p keeps increasing. Figure 2 highlights the very different nature of FBP and TBE supports in the presence of randomly distributed rogue taxa in large trees. TBE remains mostly stable no matter the depth of the reference branch (bipartition) when the number of rogue taxa is reasonable (i.e., 1–10 among 1000 taxa). On the contrary, the FBP support will always be considered low if there is even a single rogue when the topological depth of the reference branch is high (i.e., a large clade). These behaviors are clearly visible in our analyses of empirical datasets.

In light of the two models presented above, there is reason to be concerned for both branch supports: for FBP, with deep branches and large clades in case of rogue taxa; for TBE, with shallow branches and small clades in case of sampling issues. However, real data are much more complex than these two theoretical examples, and thus we will further explore the properties of the two branch supports on a series of large biological datasets that supposedly have more or less sampling disequilibrium and rogue taxa.

RESULTS ON EMPIRICAL DATASETS

Overview

Our dataset-picking strategy consists in taking large empirical datasets from the literature, for which we expect some degree of heterogeneity in phylogenetic signal and sampling. This would imply that the genetic

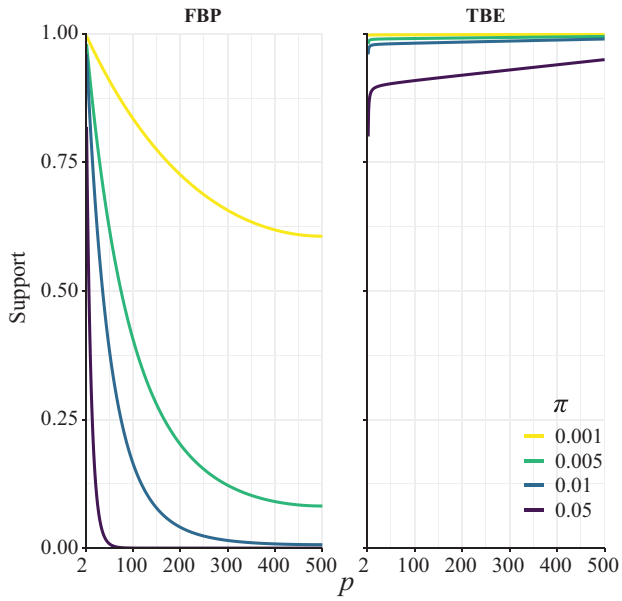

FIGURE 2. Theoretical results with strong signal but a few rogues. Evolution of support values (FBP on the left, TBE on the right) depending on sampling p of clade P, in a 1000-taxon tree. The color gradient represents the probability (π) of being rogue as the color darkens (e.g., $\pi = 0.005$ in green corresponds to ~5 rogues among 1000 taxa).

locus is expected to return good phylogenetic signal for some clades, but will fail to fully resolve a large portion of the phylogeny due to a lack of (informative) sites. For example, barcode markers like the ~660 bp segment of cytochrome *c* oxidase I (COI) gene in most metazoans, or 16S ribosomal ribonucleic acid (16S rRNA; ~1500 bp) in prokaryotes, are typically adequate phylogenetic markers to get a rough idea of the relationships between taxa, but are not sufficient to build reliable phylogenies, especially with thousands of taxa as we are considering here. Another striking example of a weak phylogenetic signal is that of SARS-CoV-2, for which we have complete genome sequences (~29,500 bp), but very few sites are informative, which prevents any complete resolution of the tree (Morel et al. 2021). Moreover, unbalanced sampling is expected, with some "charismatic" clades being over-represented (e.g., Primates and Cetaceans in Mammals), with many more taxa and sequences available in the databases and published phylogenies, and some others under-represented (e.g., Rodents; see our numbers below for Mammals).

We selected the two empirical datasets of the original TBE study (Lemoine et al. 2018) consisting of a COI Mammals dataset (1449 sequences; ~266 aa) and an HIV dataset (9147 *pol* sequences; ~1036 bp). Additionally, we used the full SARS-CoV-2 dataset from Zhukova et al. (2021) comprising 11,316 genomes (~29,500 bp). Finally, we selected 10 aligned nucleotide barcode datasets from Delsuc and Ranwez (2020) comprising between 1000 and 2000 mitochondrial COI gene sequences (~660 bp). All of these published datasets were already aligned and had in common to be large (>1000 sequences), with heterogeneous phylogenetic signal and unbalanced taxonomic sampling.

Our methodology consisted in studying the evolution of FBP and TBE supports, starting from a reference tree with N taxa and unbalanced sampling, to reach a reduced and re-balanced target tree with n ($<< N$) taxa. In addition, along this path, we studied progressively more and more balanced intermediate trees. The sampling of these intermediate trees was defined by: $n_f = N - f * (N - n)$, with $0 \leq f \leq 1$. For example, to get the midpoint between N and n, we choose $f = 0.5$, while $f = 0$ and $f = 1$ correspond to the starting and target trees, respectively. When we wanted to target the sampling of specific clades (e.g., HIV subtypes), then the representativeness of each clade was multiplied by n to obtain the required number n_X of sequences to be retained for clade X. Let N_X be the initial size of clade X, the intermediate clade samplings were then defined by $n_{X_f} = N_X - f * (N_X - n_X)$. The target samplings ($n_X$) in the reduced tree were defined by different criteria based on the specifics of the dataset: representativeness for Mammals, worldwide prevalence for HIV subtypes, sampling density per country for SARS-CoV-2, and uniform sampling per delimited species for the barcode datasets.

The objective was to assess the stability and robustness of each branch-support measure throughout the subsampling. We selected a set of relevant clades for which one expects a strong phylogenetic signal. For instance, long-branch clades like Cetacea or Marsupialia in the Mammals dataset are expected to be recovered even with a simple COI marker. The same applies to

HIV subtypes and the *pol* gene. For such clades, with strong phylogenetic signal, a stable and robust branch support should give a high score, with more or less the same value, whatever the sampling size and balance.

To get supports for the target and intermediate trees, the ideal approach would be re-estimating the reference and bootstrap trees at each sampling point. However, this procedure can be quite computationally intensive, especially for very large datasets with multiple replicates. Another much faster approach is to prune the leaves of the starting reference and bootstrap trees, and calculate FBP/TBE supports on the reduced trees. The "pruned" approach is the one we favored to reduce computational costs, but in parallel, we also used the "re-estimated" approach on the target sampling to check if the results matched those obtained with the "pruned" approach. In the two approaches, all branches with length smaller than the expectation of having 0.5 mutations in total among all sites were collapsed in both the reference and bootstrap trees to prevent adding noise in the support values (Guindon et al. 2010). The minimum branch length was defined by $\ell = 0.5/L$, where L was the number of sites in the alignment. As such, ℓ is a simple form of confidence interval: branches with length less than ℓ are estimated to carry 0 mutations (i.e., no phylogenetic signal) and are collapsed, while branches with length greater than ℓ carry at least 1 mutation and are retained.

We looked for clades of interest in the reference trees (T) using well-established classifications. For Mammals, we chose the NCBI taxonomy, with a focus on Primates, Cetaceans, Marsupials, and so on. For HIV, we used subtype annotation and Nextstrain clades for SARS-CoV-2. These clades may (or not) be fully recovered in the reference tree, which is inferred from the data. To account for this, we traversed all branches in the reference tree and looked for the branch that minimized the transfer distance with the clade of interest (e.g., Primates with Mammals, or subtype B with HIV). We distinguished wrong taxa (i.e., taxa that do not belong to the clade of interest) from missing taxa (i.e., taxa that belong to the clade of interest, but are not found in the closest clade of T). The numbers of wrong and missing taxa are denoted w and m, respectively, and $w+m$ is the transfer distance between the clade of interest and the closest clade in T that is selected. Hence, a clade in the reference tree that had no wrong and no missing taxa ($w = m = 0$) was perfectly recovered, with respect to the NCBI taxonomy, HIV subtypes, Nexstrain clades, and so on. Some other clades were not perfectly recovered in our reference trees, with a few wrong and missing taxa. In both cases, we measured the FBP and TBE supports of these reference clades, having in mind that a clade that is (almost) perfectly recovered should be highly supported. The same approach was used in Lemoine et al. (2018), for example with HIV, where several subtypes were not fully (but almost perfectly) recovered in the large tree estimated from the 9147 available *pol* sequences (e.g., $w = 2$ and $m = 0$ with subtype B that comprises >3500 sequences).

To summarize our overall strategy (Fig. 3), we begin with a large reference tree of sampling N (along with 1000 bootstrap trees), and define an overall target sampling n. The reference and bootstrap trees of sampling N are then progressively pruned until achieving target sampling n, with multiple replicates to account for random effects. At each intermediate point, FBP/TBE values are computed from the pruned reference and bootstrap trees. Once target sampling is achieved, we re-estimate reference and bootstrap trees from sampling n and compute again FBP/TBE values for comparison with the "pruned" approach. To implement this workflow, we extensively used Gotree/Goalign (Lemoine and Gascuel 2021), a toolkit that implements more than 120 user-friendly commands dedicated to multiple sequence alignment and phylogenetic tree manipulations (e.g., tree pruning, collapsing branches, MSA de-duplicating). The list of Gotree/Goalign commands used for FBP/TBE support exploration is provided in the Supplementary Appendix.

Prior to focusing on the selected clades, we briefly analyzed the overall average support in the Mammals, HIV, and SARS-CoV-2 datasets, for different samplings and depths (Supplementary Fig. SF1). Our results mainly show that FBP and TBE average supports remain stable (with a few exceptions) throughout the successive subsamplings ("pruned" experiments), even when the number of taxa is drastically reduced. Furthermore, we observe that the difference in average scores between the "pruned" and "re-estimated" experiments is low, thus validating our overall strategy (Fig. 3). Average TBE supports are on average higher than those of FBP are, especially for deep branches (as expected, see above). Nevertheless, average FBP and TBE supports are low, even for shallow branches (<0.6 for FBP and <0.8 for TBE), as expected with these datasets. However, all these are average supports, and the results for the biologically important clades tell a very different story, as shown in the following.

Mammals Dataset

The Mammals dataset consists of 1449 aligned COI protein sequences used in Lemoine et al. (2018). A few clades, almost without contradiction with the NCBI taxonomy ($w \approx m \approx 0$; for details, see Lemoine et al. 2018), had been highlighted to illustrate the differences between the FBP and TBE supports (Cetacea, Elephantidae, Geomyidae, "Insectivora," Marsupialia, Monotremata, Mustelinae, Perognathinae, and Simians). We selected these clades but discarded those where the number of available species was insufficient for our experiments (i.e., Elephantidae and Monotremata). As expected, these clades have been unevenly sampled across the mammalian diversity, either over-sampled or under-sampled. For example, the Cetacea sampling

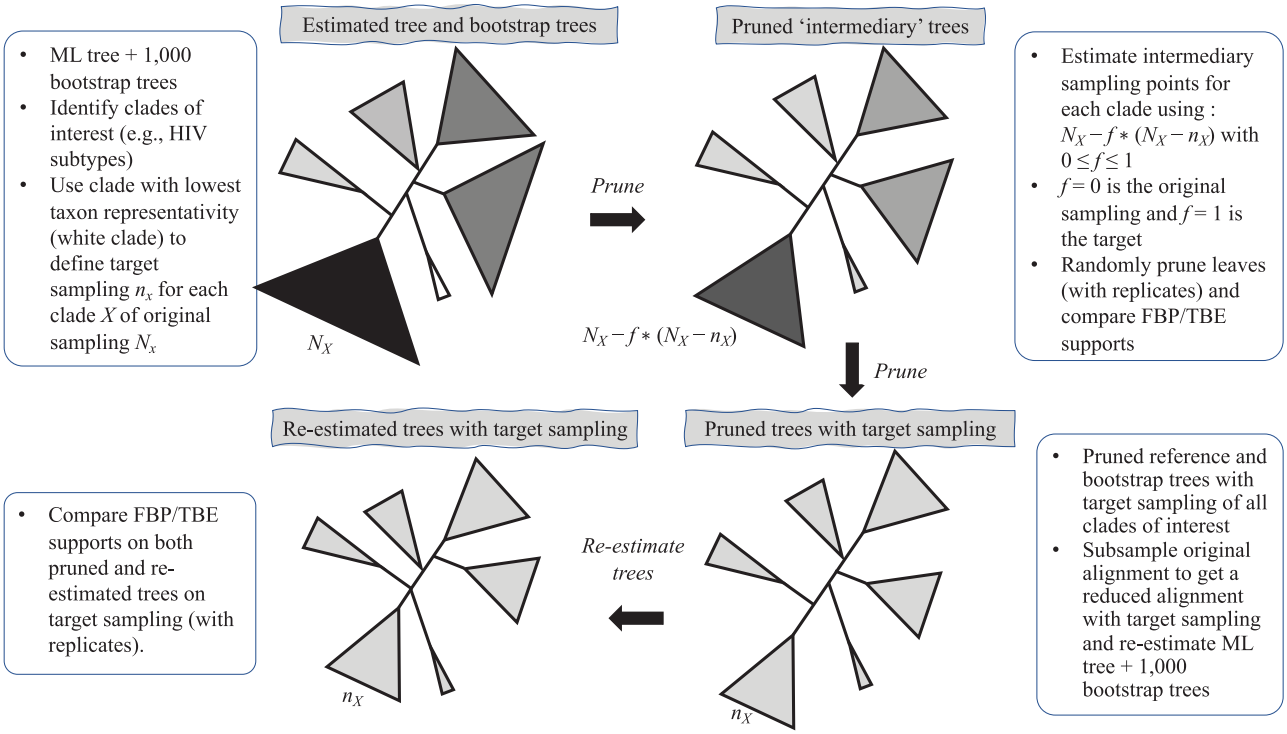

FIGURE 3. Summary of the overall sampling strategy. We start from a large reference tree of unbalanced sampling (N, top-left) and we define a target sampling (n). We then define intermediate sampling points (top-right) and prune the reference and bootstrap trees accordingly, across multiple replicates. When target sampling is achieved (bottom-right), we then subsample the original alignment accordingly and re-estimate a reference tree along with 1000 bootstrap trees.

represents 61% of all cetacean species, while the Soricomorpha ("Insectivora" in Lemoine et al. 2018) sampling represents only 13% of the Soricomorpha species.

We used the Integrated Taxonomic Information System to obtain the taxon representativeness R_X of each selected clade X (Supplementary Table ST1). The target sampling n for the reduced tree from the initial sampling $N(= 1449)$ was estimated with the following steps. First, we selected the clade with the lowest sampling (i.e., Mustelinae, with $N_X = 9$). To obtain the smallest possible target sampling n while keeping all the selected clades, the target sampling n_X of Mustelinae was set to 2 (a cherry, the smallest possible clade). Then, we divided 2 by the Mustelinae taxon representativeness $R_X = 0.0003$ (i.e., 0.03% of Mammals diversity), to obtain the final target sampling of the reduced tree, $n = 669$. We estimated the target sampling of each clade X using $n_X = n \times R_X$. Finally, we estimated 3 intermediate samplings points with $f = 0.25$, 0.5, and 0.75 (while $f = 0$ and $f = 1$ correspond to the starting and target trees, respectively). The species retained in the target sampling as well as in the intermediate points were randomly drawn from the initial species set, and this subsampling was iterated to obtain 30 replicates. This procedure allowed us to follow the trend of FBP and TBE scores while subsampling from initial unequal clade sampling (N_X) to more accurate taxon representativeness (n_X).

In Lemoine et al. (2018), bootstrap trees were initially estimated with RAxML version 8 (Stamatakis 2014), but using rapid bootstrapping (Stamatakis et al. 2008), and thus bootstrap trees was lacking branch lengths. We kept the initial reference tree from Lemoine et al. (2018) but re-estimated 1000 "traditional" bootstrap trees (see the Supplementary Appendix for command-line options).

Reference and bootstrap trees were also re-estimated on alignments with target sampling using the same RAxML command. Prior to each FBP and TBE support computation, branches shorter than 0.000949 (= 0.5/527) were collapsed (resulting in ~53% of internal branches collapsed). This value corresponds to the expectation of having less than 0.5 mutations on the branch, given that the alignment has 527 sites. FBP and TBE support values were computed at each intermediary point with the pruned approach (Fig. 3). We retrieved the selected clades in the re-estimated reference trees as in Lemoine et al. (2018, see above and the Supplementary Appendix for command-lines options).

The initial results from Lemoine et al. (2018) showed that TBE significantly supports the selected clades while FBP does not, even when these clades were perfectly recovered in the reference tree (e.g., Cetacea, with $w = m = 0$). Our experiment confirms and complements these results by showing similar trends,

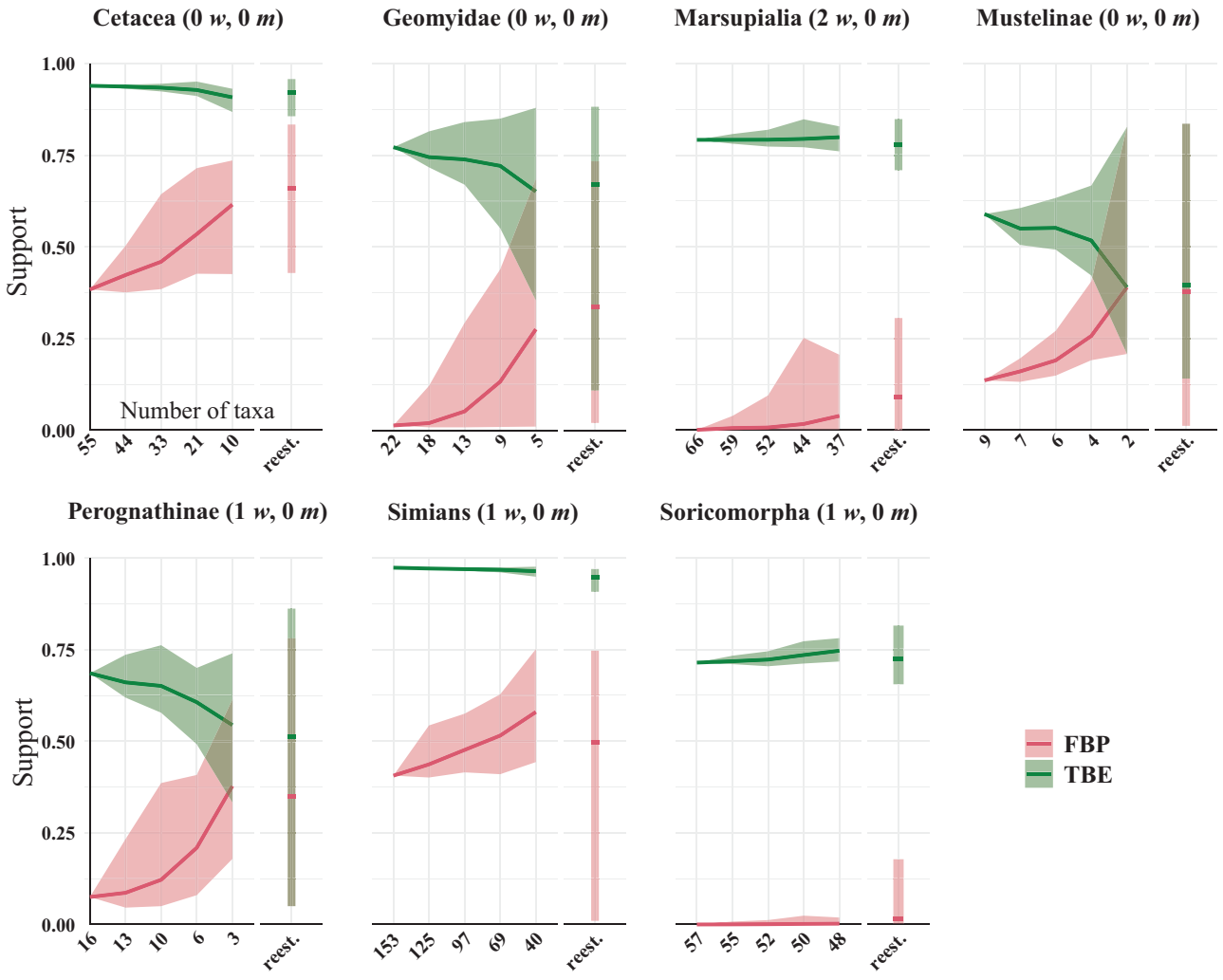

FIGURE 4. Results on the Mammals dataset. For each clade, the first value on the x-axis corresponds to the initial number of taxa and the last value to the number of taxa in the target sampling. We also provide the intermediate values, with proportions $f = 0.25$, 0.50, and 0.75. Trees were re-estimated on the target sampling (reest.). The number of wrong (w) and missing (m) taxa is indicated next to each clade name. To obtain the subsamples, taxa were randomly removed. The ribbon shows the maximum and minimum values over 30 replicates, the thick line is the mean value.

when subsampling the initial species to achieve a (balanced) target sampling of the Mammals tree (Fig. 4). FBP tends to increase with reduced sampling, but in most cases, it fails to reach the conventional threshold of 0.7 (e.g., the best score is Cetacea with original support ≈40% and average support with reduced sampling ≈60%).

Two types of behavior can be observed for TBE, as predicted by our theoretical analyses (see above). When the number of taxa is sufficiently high (i.e., more than 10 in the reduced target sampling), TBE scores are high (>70%) and stable across all sampling points, as observed for the clades Cetacea, Marsupialia, Simians, and Soricomorpha. However, when the number of taxa in the target sampling is low (<10), then TBE tends to decrease slightly and behave more like FBP as the number of taxa reduces (i.e., Geomyidae, Mustelinae, Perognathinae). When there are only 2 taxa left in the clade (i.e., Mustelinae), then FBP and TBE scores are equal, as expected and explained in the "Introduction" section. Noteworthy, FBP shows more variability in maximum and minimum values (red ribbons) than TBE (green ribbons), even when there is no signal for the presence of rogues ($w = m = 0$, e.g., Cetacea and Geomyidae, Fig. 4). These results suggest a low robustness of FBP in the face of a heterogeneous phylogenetic signal, even when the clade is perfectly recovered in the reference tree. Finally, the trees re-estimated at the target sampling (n) confirm our results obtained with the "pruned" approach but tend to show greater variability in FBP scores across replicates, indicating that the robustness of FBP is even weaker in a realistic (full computation) setting with these mammalian data.

The HIV Dataset

Like the Mammals dataset, the HIV dataset was used in Lemoine et al. (2018) to highlight the differences between FBP and TBE. It consists of 9147 HIV-1 group M *pol* sequences representing the 9 subtypes, and includes 50 recombinants detected using jpHMM (Schultz et al. 2009), that is, sequences that contain DNA from at least two different subtypes in the *pol* region. In this experiment, we achieved a representative sampling based on the current prevalence of each subtype in the worldwide population. The prevalence values of each subtype were obtained from Cassan et al. (2016, Fig. 3 and Supplementary Table ST2). Then, we applied the same approach as used with the Mammals dataset. We divided 2 by the prevalence of the most under-represented subtype (i.e., subtype K, with $R_X = 0.1261\%$) to calculate n (= 1599). We estimated 4 intermediate sampling points with $f = 0.2$, 0.4, 0.6, and 0.8, and achieved 30 replicates by randomly subsampling. With the target sampling, we re-estimated the trees and retrieved the selected clades as Lemoine et al. (2018, see above, the Supplementary Appendix and Supplementary Table ST2 for details). All trees (reference, bootstrap, subsampled, and so on) were estimated using FastTree version 2.1.11 Double precision (Price et al. 2010). The precision of branch length estimation was improved using RAxML-NG v. 1.1.0 (Kozlov et al. 2019). Then, branches shorter than 0.000479 (= 0.5/1043, where 1043 is the number of alignment sites) was collapsed (resulting in ~12% of internal branches collapsed). In addition, we computed the average FBP/TBE support of all branches for different depths. This additional experiment allowed us to follow the evolution of the average FBP/TBE supports at different depths while reducing the sampling, instead of focusing only on HIV-1 group M subtypes that are known to be well supported in most datasets.

With the initial sampling, Lemoine et al. (2018) showed that all 9 subtypes were found in the reference tree and highly supported by TBE (although the average TBE support across the tree is relatively low; Supplementary Fig. SF1). On the opposite, only the sparsely represented subtypes (i.e., F, H, J, and K) were supported by FBP. A prominent example is the FBP/TBE supports of the B subtype, a clade comprising 3559 sequences and which is almost perfect in the reference tree since it contains all B sequences plus 2 B-recombinants (as detected by jpHMM). The FBP score of this clade is 0.03, indicating almost no signal, while the TBE score is 0.99, indicating a strong signal.

In the re-equilibrated trees using prevalence information, we see that the overall tendency observed in Lemoine et al. (2018) remains true, even when the sampling of some subtypes is considerably reduced (e.g., subtype B, from 3559 to 226 sequences) and most recombinants are removed by random subsampling (Fig. 5; Supplementary Table ST2). In all cases, average TBE values are high (>90%) and with very little variation across replicates (green ribbons). In contrast, evolution of FBP supports across pruned trees does not follow the same pattern for all subtypes. In subtypes A, C, D, and G, the FBP scores increase as sampling is reduced, but we observe a high variability (red ribbons) across replicates. Interestingly, those are the four subtypes (A, C, D, and G) with a high number of recombinants (as detected with jpHMM; see Supplementary Table ST2), with 4–12 wrong/missing taxa each (w and m; Fig. 5). Extreme variability across replicates is likely explained by the presence/absence of recombinants that are randomly subsampled. Quite differently, subtypes B and F (2 wrong ones in B corresponding to B-recombinants, none in F, and no missings in both F and B) show little evolution of FBP support, and little variability across replicates, at least for the "pruned" strategy. This suggests that subtypes B and F could be affected by taxa other than recombinants (or undetected recombinants), and that these taxa are relatively numerous and unstable, so that replicate scores are not affected by the randomness of subsampling. One could thus distinguish two forms of taxon instability (with the whole spectrum of intermediate cases): single rogue terminal taxa prone to instability (e.g., recombinants) and taxa belonging to globally unstable clades.

Our results again illustrate the low robustness of FBP to rogue taxa and show how robust TBE is, whether or not some rogues are present (as predicted by our theoretical model with rogues, see above). One could argue that TBE might be over-supporting those clades, but our experiment on the overall support values across the tree shows another story (Supplementary Fig. SF1). The overall FBP/TBE support in the phylogeny is low for all ranges of branch depth (by the standards of each metric, that is, less than 0.4 for FBP and less than 0.6 for TBE) and shows little change as trees are increasingly subsampled, except for cherries and deep branches where a trend emerges with this HIV dataset. Indeed, average FBP/TBE support for cherries ($p = 2$) slowly decreases as we remove more and more sequences. Conversely, average TBE supports for deep branches ($p > 9$) tend to increase. These overall trends contradict the results with selected clades/subtypes (Fig. 5), but are slight (less than 10% point difference between initial and target samplings) and are not found in the analysis of the SARS-CoV-2 dataset that follows.

The SARS-Cov-2 Dataset

We retrieved the SARS-CoV-2 dataset from Zhukova et al. (2021) and processed the sequences through the Nextclade Web 2.3.0 interface (Aksamentov et al. 2021); https://clades.nextstrain.org) to assign each sequence to a Nextstrain "clade." Some Nextstrain "clades" are paraphyletic, as they do not include the sequences of new child clades. When a "clade" was paraphyletic (e.g., clade 20B), we added all child clades in the sampling to make it monophyletic (i.e., 20D in the case of 20B). The result is a reference classification composed of 5 clades: 19B, 20A (= 20A + 20B + 20C + 20D Nextstrain "clades"),

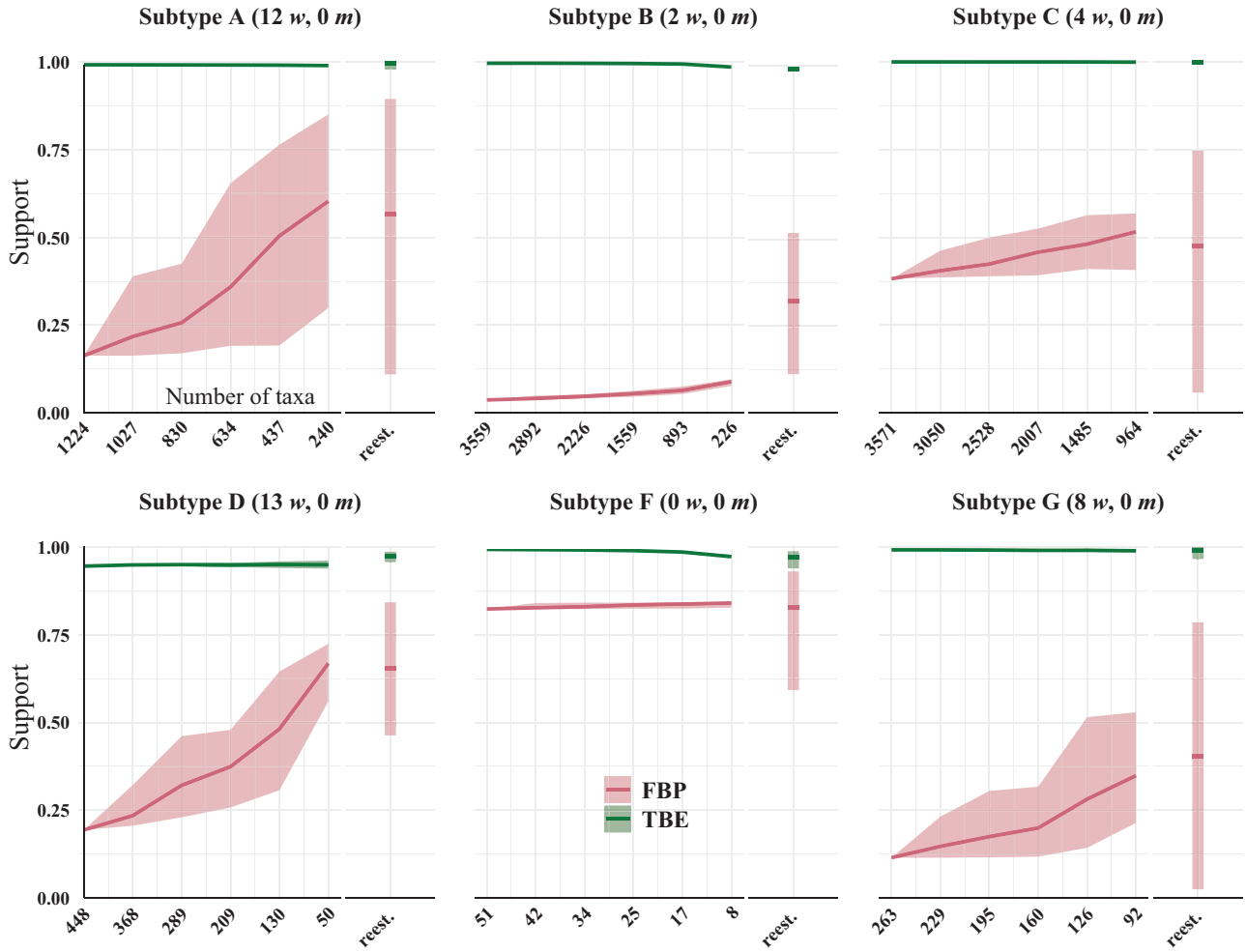

FIGURE 5. Results on the HIV dataset. We removed taxa within each HIV subtype based on prevalence. For each subtype (clade), the first value on the x-axis corresponds to the initial number of taxa, and the last value to the number of taxa in the target sampling. We also provide intermediate values, corresponding to proportions $f = 0.2$, 0.4, 0.6, and 0.8 (see text for details). Trees were re-estimated on the target sampling (reest.). The number of wrong (w) and missing (m) taxa is indicated next to each subtype identifier. To obtain the subsamples, taxa were randomly removed. The ribbon shows the maximum and minimum values over 30 replicates, the thick line is the mean value.

20C, 20B (= 20B + 20D Nextstrain "clades"), and 20D. Sampling re-equilibrium had already been achieved by Zhukova et al. (2021), starting from the complete 11,316-sequence tree to 5 subsampled and re-balanced trees of size ~2000. The target sampling for these 5 datasets was defined to reflect the number of cases reported by country over time during the early months of the pandemic.

For the SARS-CoV-2 dataset, we ran multiple tests to determine what would be the best approach to infer the tree on the 8541 genomes—after removing duplicate genomes from the initial 11,316 genomes dataset of Zhukova et al. (2021). We looked at the best trade-off for running time and tree accuracy between FastTree, IQ-TREE 2 (Minh et al. 2020), IQ-TREE 2 in fast mode, and RAxML-NG (results not shown). Tree accuracy was evaluated by counting the number of taxa to transfer from the estimated tree to recover the Nextstrain clades from our reference taxonomy. The fastest approaches were by far FastTree and IQ-TREE 2-fast, but the total number of taxa to transfer for FastTree (309) was much higher than for IQ-TREE (82). The total number of taxa to transfer was not much different from other, more time consuming, methods; hence, IQ-TREE 2-fast was selected as the best approach for the SARS-CoV-2 dataset. Branches shorter than 0.000017 (= 0.5/29,726, where 29,726 is the number of alignment sites) were collapsed (resulting in ~84% of internal branches collapsed, as expected due the very low number of mutations in this dataset).

We successively randomly pruned taxa by batches of ~1500 in the starting tree to retrieve the sampling of the 5 trees of size ~2000 from Zhukova et al. (2021), using 20 replicates each time, for a total of $20 * 5 = 100$ replicates. We extracted the FBP/TBE supports for the 5 selected clades (Fig. 6) and the average of these

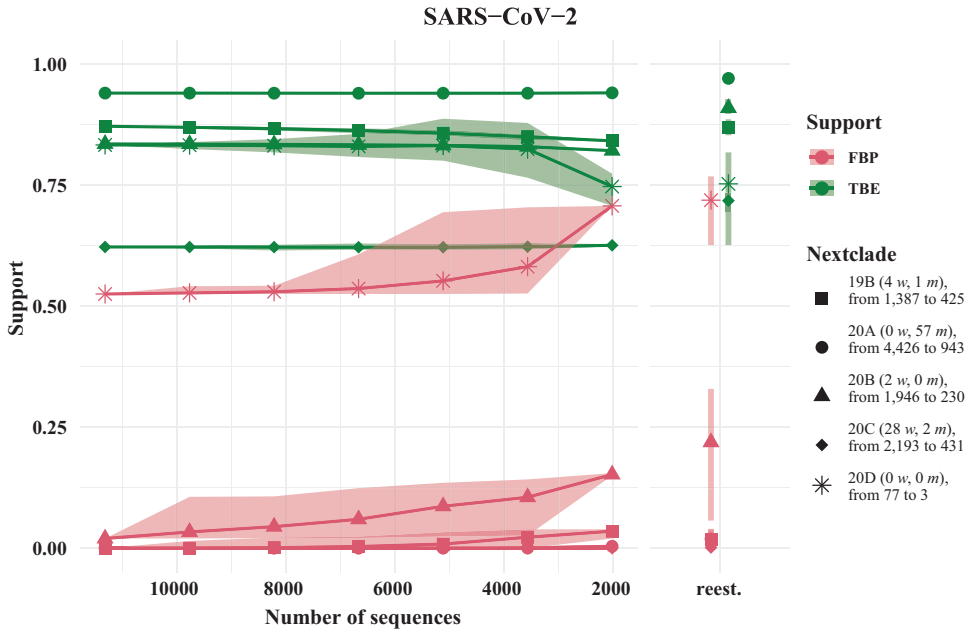

FIGURE 6. Results on the SARS-CoV-2 dataset. We randomly removed taxa from the original SARS-CoV-2 tree (11,316 tips) in order to achieve a better sampling equilibrium on trees with about 2000 taxa. Trees were re-estimated on target sampling (reest.). The number of wrongly (w) and missing (m) taxa is given next to each clade identifier, and we provide the original and target samplings for each clade (e.g., "77 to 3" with 20D). The ribbon indicates the maximum and minimum values, and the thick line corresponds to the mean value across all replicates.

supports for different branch depth intervals (Supplementary Fig. SF1), from the pruned and estimated trees. As with the other datasets, we retrieved the selected clades in the estimated trees by minimizing the number of wrong (w) and missing (m) taxa (see above and the Supplementary Appendix for details).

The difference with the previously analyzed dataset (HIV) is that the overall phylogenetic signal is scarce, with clades supported by only one or two mutations (https://nextstrain.org; Supplementary Table ST3). As for the previous datasets, most Nextstrain clades are well supported by TBE and not by FBP, even when the clade is well retrieved in the reference tree (e.g., 20D). Interestingly, support values are little affected by the number of wrongly placed and missing sequences ($w + m$), and the FBP/TBE scores remain stable across most clades upon subsampling (Fig. 6). For example, clade 20C is not well supported by TBE (~0.62). Indeed, with 2193 sequences in this clade, a score of 0.62 means that on average $(2193 - 1) \times 0.38 = 833$, sequences need to be transferred to recover that clade in the bootstrap trees; this number of 833 is much greater than the number of wrong/missing sequences in the reference clade (=30, Fig. 6). As opposed to HIV, results with SARS-CoV-2 seem to indicate a global instability of a large number of taxa. This is likely due to the scarcity of the signal and the presence of "rogue clades," which would be responsible for the transfer of hundreds of taxa in the bootstrap trees (thus greatly affecting TBE as well as FBP), rather than a few rogue terminal taxa as with HIV (e.g., subtype B, Fig. 5).

The evolution of average FBP/TBE support at different ranges of branch depth (Supplementary Fig. SF1) indicates again that TBE is higher than FBP on average. FBP returns (again) heterogeneous average support values depending on the depth range (low FBP support for shallow branches, very low FBP support for deep branches), while TBE average supports are more homogeneous. However, these average results (Supplementary Fig. SF1) confirm the finding in clade-specific Figure 6 that FBP and TBE supports are little affected by sampling biases. Moreover, both average supports are higher compared to HIV (Supplementary Fig. SF1). In fact, the signal is scarce that makes it impossible to infer a fully resolved tree (Morel et al. 2021), hence the high number of collapsed branches (~84%) that are not supported by any mutation, but the remaining branches are easy to infer, even with parsimony (Kramer et al. 2023), and are relatively well supported on average. In HIV trees, only ~12% of branches are collapsed and the reconstruction is based on a single marker, which is more prone to saturation and other phylogenetic biases.

The Barcode Datasets

We selected 10 aligned nucleotide barcode datasets from Delsuc and Ranwez (2020) comprising between 1000 and 2000 mitochondrial COI sequences (Acanthocephala, Archaeognatha, Bryozoa, Dermaptera, Megaloptera, Onychophora, Siphonaptera, Tardigrada, Testudines, Uraniidae). For all these barcode datasets, we first removed duplicated (i.e., strictly identical)

sequences from the initial alignment. We estimated reference and 1000 bootstrap trees on each dataset using RAxML-NG. Trees were then "repopulated" using Gotree (see the Supplementary Appendix for command-line options) to add back the duplicated sequences. We ran the ML heuristic of the multi-rate Poisson Tree Processes (mPTP) method (Kapli et al. 2017) in default mode to delimit putative species in each dataset. The output of mPTP was used to annotate species branches in each reference tree. Some of so delimited species have many sampled sequences (~8% of retained species have >30 sequences), while others are poorly sampled (~22% of retained species have 2 sequences; see Supplementary Fig. SF2). We computed FBP/TBE supports for each dataset on the "complete" trees with all sequences, and on the "deduplicated" trees where all strictly identical sequences from the reference and bootstrap trees were removed. Our target sampling was defined by keeping a maximum of 5 different sequences ("max5") for each species so delimited. The reference and bootstrap trees were fully re-estimated using RAxML-NG from this reduced and rebalanced set of sequences (the pruning approach was not used here). Sequences were not removed randomly, but using the greedy algorithm of (Pardi and Goldman 2005; Steel 2005) implemented in the Phylogenetic Diversity Analyzer (PDA; Chernomor et al. 2015). This algorithm consists of iteratively deleting the taxon associated with the shortest branch; in doing so, we maximize the phylogenetic diversity of the tree while reducing the sampling. Using RAxML-NG, we re-estimated trees where a maximum of 5 samples per species (selected using PDA) were kept. Finally, we compared species edges between 3 alternative samplings ("full," "dedup" and "max5") and considered only the edges that were found in all 3 samplings. In total, our 10 barcode datasets add up to 15,111 COI gene sequences, of which 5700 (~38%) are duplicates and a total of 1390 putative species have been delimited. On average, ~59% of internal branches was collapsed in the "full" trees. Some species clades were not retained in the analyses, because they were singletons in the complete or deduplicated trees, or because they were not found in the "max5" reference trees due to phylogeny re-estimation. This filtering downsized to 938 the number of putative species used in this experiment. We distinguished several categories of branches, based on the initial number of sequences in the delimited species, as well as supra-specific branches (i.e., above the species level).

On average, the FBP and TBE support values for the putative species branches are high, that is, above 0.7. This result was expected and is explained by our use of mPTP. This method counts the number of substitutions per branch and estimates the rates of branching events to detect which parts of the tree follow as speciation model (interspecific) and that follow a coalescent model (intraspecific). Thus, the delimited species branches usually correspond to multiple substitutions and carry a strong signal.

As expected, FBP scores do not change when removing duplicate sequences since complete and deduplicated tree topologies are identical (Fig. 7). We also notice that TBE scores only slightly decrease when removing duplicates (except for cherries). Thus, despite removing nearly 38% of the sequences, TBE shows support stability. The same stability in TBE scores is observed in the target sampling ("max5") dataset when reducing all species' sampling to a maximum of 5 sequences. In contrast, FBP is much more affected than TBE by subsampling, especially for species that initially have many sequences (i.e., more than 30 samples). To confirm this observation, we calculated the mean absolute difference Δ between the initial and target ("max5") samplings for all comparable branches in the 10 datasets for each depth category. Results indicate a high Δ in FBP scores between the initial and target samplings, particularly for species that are highly represented in the initial sampling (e.g., $\Delta = 0.19$ with >30 samples, versus 0.07 for TBE; Fig. 7).

This last experiment differs not only from the previous three in its subsampling strategy but also in terms of the quantity of data and results. Instead of highlighting a few clades, we assume that most delimited species that are retrieved in all three modes of sampling should have a reasonably strong phylogenetic signal, be well supported (on average), should be little affected by the subsampling procedure. While most species are well supported by both metrics, our results indicate a weak robustness of FBP to subsampling, particularly for species branches that initially contain many samples, while there is no reason to believe that these species, in particular, should be less supported than the others. This observation also holds for the supra-specific branches (Fig. 7), where TBE scores are higher and show a better robustness ($\Delta = 0.05$) than FBP scores ($\Delta = 0.07$), even if the global tendency is consistent with the theoretical analyses, with TBE supports slightly decreasing with re-equilibrated ("max5") sampling, while FBP increases by a rather larger margin in this condition.

Discussion

Similar to what was shown in Lemoine et al. (2018), our theoretical and empirical results demonstrate the usefulness of TBE, especially for large trees with heterogeneous phylogenetic signal. Deep branches that are known to be essentially correct are generally supported by TBE, but FBP supports will generally be low, whether a phylogenetic signal is expected or not. In this study, we explored the impact of sampling biases on FBP/TBE support values. Through numerous datasets (1 Mammals, 1 HIV, 1 SARS-CoV-2, and 10 barcode), we found no evidence that TBE falsely supports poor branches due to oversampling. In fact, most branches in these large datasets with heterogeneous and relatively low phylogenetic signals are on average poorly supported

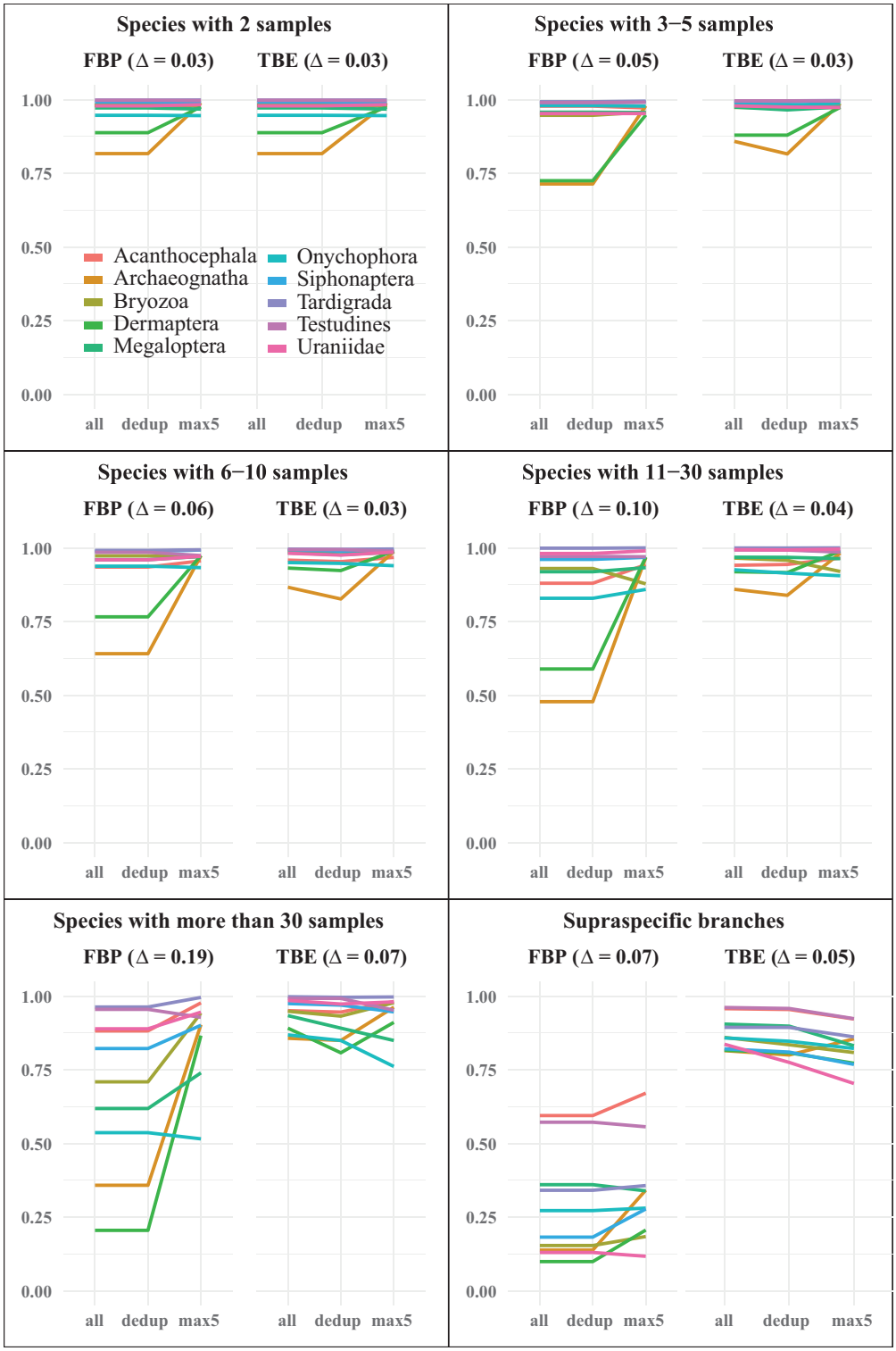

FIGURE 7. Results on the barcode datasets. Boxes correspond to the initial number of samples in the delimited species in the tree with all available sequences. For example, if we consider the bottom-left panel "Species with more than 30 samples," it means that for each tree on the full set of sequences ("all"), we retain all delimited species with more than 30 samples and report the mean support of the species branches on the first x-axis value. Second x-axis value takes the exact same set of species, but this time, we compute supports using the trees with identical sequences removed ("dedup"). Finally, third x-axis value takes again the same set of species, but for re-estimated trees for which a maximum of 5 samples ("max5") is kept by delimited species. "Supraspecific branches" are all branches above the species level, that is, we exclude species and infra-specific branches. Delta values correspond to the average absolute difference of supports between the initial ("all") and target ("max5") conditions, that is, a high delta value indicates a large difference in FBP/TBE supports from a highly heterogeneously sampled dataset to a more homogeneous dataset.

by TBE. Furthermore, our results indicate that TBE scores of the clades with significant signal are little affected by sampling biases, unless those clades are very small. Basically, in our experiments, the TBE support remained unchanged when over-sampled clades were downsampled to achieve balanced sampling, except for small clades whose TBE support tended to decrease. Indeed, for small clades the FBP and TBE supports are similar (they are identical for cherries). With medium-to-large clades, FBP supports were found to be not only lower but also much less robust than TBE supports, with an overall tendency to decrease with increasing sampling. Furthermore, depending on the (rogue) taxa sampled, the same clade with the same sample size can have low or high FBP support, revealing great variability in FBP support.

Based on these experiments, we suggest a few guidelines on how to conduct a routine bootstrap analysis to assess branch support in large trees:

- First, it should be remembered that this study and TBE, in general, are mainly aimed at large trees with relatively low phylogenetic signal (typically gene or virus trees with very many taxa of the order of a thousand or more). In the opposite configuration, with a smaller number of taxa but many informative sites (typically with concatenations of multiple gene alignments), phylogenetic signal increases and we are faced with the opposite problem: FBP supports are generally very (sometimes too) high, and TBE supports are usually of little interest. In fact, when FBP is close to 100%, TBE is also close to 100% as TBE is always higher than or equal to FBP.
- Remove duplicates. Duplicated (i.e., strictly identical sequences) are quite common in large datasets and can have an impact on TBE scores. Most ML software (e.g., IQ-TREE, RAxML-NG) already remove duplicate sequences prior to ML inference and then add back those sequences by creating near-zero length branches. We suggest that duplicates should also be removed from reference and bootstrap trees prior to computing TBE scores (and then added back).
- Collapse insignificant branches. In the obtained reference and bootstrap trees, many branches can be short and have no real biological meaning because they essentially correspond to no substitution event among all sites in the MSA.
- Systematically calculate both FBP and TBE. Bootstrap trees take a long time to compute while calculating FBP/TBE is much faster in comparison, meaning that once bootstrap trees have been calculated, they should be used for both FBP and TBE. If the tree is highly supported by FBP (i.e., almost 100% for all branches), then TBE will not teach anything new. However, FBP should be systematically compared to TBE even with branches that appear to be "well" supported (e.g., 70%), because the interpretation of these two supports is very different. In this case, when TBE is close to 100%, this likely means that a few rogues are perturbing FBP analyses (as with HIV), whereas when TBE is also low (as with SARS-CoV-2), it is likely that many taxa are unstable and the overall phylogenetic signal is weak.
- Do not stop at the 70% threshold! For many phylogeneticists, 70% (or 80% for some) of FBP support has become the rule of thumb for assessing well-supported branches. As stated in Soltis and Soltis (2003) and already quoted "consensus has been reached among practitioners, if not among statisticians and theoreticians" and "many systematists have adopted Hillis and Bull's '70%' value as an indication of support." In fact, the "70%" rule has no statistical basis and is likely inappropriate for many situations. For example, finding a deep branch with FBP = 70% in a large tree with thousands of tips generally means that this branch has strong support and is likely correct (unless model miss-specification, long-branch attraction, or any other reconstruction bias), whereas having FBP = 70% for a cherry in a small tree will generally not be considered very strong. For more information on the long debate about the meaning of FBP and the best selection threshold, see Sanderson (1995); Berry and Gascuel (1996); Soltis and Soltis (2003); Lemoine et al. (2018) and Simon (2022).
- With TBE, the support has a very different meaning. With FBP, we interpret the support roughly as the probability that the inferred branch is entirely correct (assuming there is no reconstruction bias, etc.), whereas with TBE, the support is a measure of the correctness of the inferred branch, accepting that, in this inferred branch, some taxa are misplaced and must be transferred to recover the genuine branch (see Extended Data Fig. 10 in Lemoine et al. (2018) for further results along these lines). With such an interpretation of support, the question of the support threshold becomes very different and must be put into perspective with the sampling of the clade studied. With a clade of size p and of support s, one can easily calculate the average number of taxa that need to be transferred in the bootstrap trees to recover the reference clade, using the formula $(p-1)*(1-s)$. Thus, a score of 80% or even 95% should not necessarily be considered as high support for TBE if the clade is large, as this means that many taxa must be transferred in the bootstrap trees to recover the inferred clade. For example, with $p = 1000$ and $s = 95\%$, 50 taxa need to be transferred on average, which may be considered acceptable with virus data due to the inherent weak signal, recombinants, and so on, but becomes considerable if we think, for example, to mammals or birds where the ultimate aim is to decipher the true evolutionary origin of all the species studied. These calculations on the number of taxa to be transferred are available in certain

software packages and websites (e.g., BOOSTER, https://booster.pasteur.fr/) and enable a better understanding of TBE supports, which must be interpreted by users depending on the data they are analyzing and their expectations regarding this data.

With now a better understanding of FBP and TBE behaviors under various sampling conditions, one of the major challenges in phylogenetics is to better interpret and use these branch supports. In the era of large-scale datasets, understanding the causes of a branch support to be low or high, through a better characterization of rogue taxa, would allow phylogeneticists to better comprehend their data and their flaws.

Acknowledgments

Computations have been performed on the Plateforme de Calcul Intensif & Algorithmique (PCIA, UAR 2700 2AD—Muséum National d'Histoire Naturelle, Centre National de la Recherche Scientifique, Paris, FRANCE) and on the cluster of the Institut Pasteur (Paris, France).

Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.ncjsxkt05.

Funding

PZ and OG are supported by the Paris Artificial Intelligence Research Institute (PRAIRIE, ANR-19-P3IA-001)

Data Availability

Supplementary Information (Supplementary Fig. SF1 and SF2 and Supplementary Tables ST1–ST3) and the Online Appendix as well as all our data and scripts are available from https//doi.org/10.5061/dryad.ncjsxkt05.

References

Aksamentov I., Roemer C., Hodcroft E.B., Neher R.A. 2021. Nextclade: clade assignment, mutation calling and quality control for viral genomes. J. Open Source Softw. 6(67):3773.

Anisimova M., Gascuel O. 2006. Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. Syst. Biol. 55:539–552.

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. Syst. Biol. 60:685–699.

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.

Cassan E., Arigon-Chifolleau A.-M., Mesnard J.-M., Gross A., Gascuel O. 2016. Concomitant emergence of the antisense protein gene of HIV-1 and of the pandemic. Proc. Natl. Acad. Sci. U.S.A. 113:11537–11542.

Chernomor O., Minh B.Q., Forest F., Klaere S., Ingram T., Henzinger M., von Haeseler A. 2015. Split diversity in constrained conservation prioritization using integer linear programming. Methods Ecol. Evol. 6:83–91.

Dávila Felipe M., Domelevo Entfellner J.-B., Lemoine F., Truszkowski J., Gascuel O. 2019. Distribution and asymptotic behavior of the phylogenetic transfer distance. J. Math. Biol. 79:485–508.

Delsuc F., Ranwez V. 2020. Accurate alignment of (meta) barcoding data sets using MACSE. In: Delsuc, F., Scornavacca, C., Galtier, N., editors, Phylogenetics in the genomic era, Chapter 2.3, pp. 2.3:1–2.3:31 (hal-02541199).

Douady C.J., Delsuc F., Boucher Y., Doolittle W.F., Douzery E.J.P. 2003. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. Mol. Biol. Evol. 20:248–254.

Efron B., Halloran E., Holmes S. 1996. Bootstrap confidence levels for phylogenetic trees. Proc. Natl. Acad. Sci. U.S.A. 93: 13429–13434.

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

Felsenstein J., Kishino H. 1993. Is there something wrong with the bootstrap on phylogenies? A reply to Hillis and Bull. Syst. Biol. 42:193–192.

Gascuel O., Lemoine F. 2022. Chapter 9 phylogénétique: quelles mesures de support pour les branches d'un arbre? In: Didier, G., Guindon, S., editors. Modèles et méthodes pour l'évolution biologique. ISTE Editions Ltd. pp. 223–246.

Gouy M., Tannier E., Comte N., Parsons D.P. 2021. Seaview version 5: a multiplatform software for multiple sequence alignment, molecular phylogenetic analyses, and tree reconciliation. In: Katoh K., editor. Multiple sequence alignment: methods and protocols. New York (NY): Springer US. pp. 241–260.

Guindon S., Dufayard J.-F., Lefort V., Anisimova M., Hordijk W., Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59:307–321.

Hillis D.M., Bull J.J. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst. Biol. 42:182–192.

Hoang D.T., Chernomor O., von Haeseler A., Minh B.Q., Vinh L.S. 2018. UFBoot2: improving the ultrafast bootstrap approximation. Mol. Biol. Evol. 35:518–522.

Janssens S., Couvreur T.L.P., Mertens A., Dauby G., Dagallier L.-P., Abeele S.V., Vandelook F., Mascarello M., Beeckman H., Sosef M., Droissart V., Bank M. van der, Maurin O., Hawthorne W., Marshall C., Réjou-Méchain M., Beina D., Baya F., Merckx V., Verstraete B., Hardy O. 2020. A large-scale species level dated angiosperm phylogeny for evolutionary and ecological analyses. Biodivers. Data J. 8:e39677.

Kapli P., Lutteropp S., Zhang J., Kobert K., Pavlidis P., Stamatakis A., Flouri T. 2017. Multi-rate Poisson tree processes for single-locus species delimitation under maximum likelihood and Markov chain Monte Carlo. Bioinformatics 33:1630–1638.

Kozlov A.M., Darriba D., Flouri T., Morel B., Stamatakis A. 2019. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics 35:4453–4455.

Kramer A., Thornlow B., Ye C., De Maio N., McBroome J., Hinrichs A.S., Lanfear R., Turakhia Y., Corbett-Detig R. 2023. Online phylogenetics with matoptimize produces equivalent trees and is dramatically more efficient for large SARS-CoV-2 phylogenies than de novo and maximum-likelihood implementations. Syst. Biol, syad031, https://doi.org/10.1093/sysbio/syad031.

Lemoine F., Domelevo Entfellner J.-B., Wilkinson E., Correia D., Dávila Felipe M., De Oliveira T., Gascuel O. 2018. Renewing Felsenstein's phylogenetic bootstrap in the era of big data. Nature 556:452–456.

Lemoine F., Gascuel O. 2021. Gotree/Goalign: toolkit and Go API to facilitate the development of phylogenetic workflows. NAR Genom. Bioinform. 3(3):lqab075.

Lewitus E., Bittner L., Malviya S., Bowler C., Morlon H. 2018. Clade-specific diversification dynamics of marine diatoms since the Jurassic. Nat. Ecol. Evol. 2:1715–1723.

Lutteropp S., Kozlov A.M., Stamatakis A. 2020. A fast and memory-efficient implementation of the transfer bootstrap. Bioinformatics 36:2280–2281.

Minh B.Q., Nguyen M.A.T., von Haeseler A. 2013. Ultrafast approximation for phylogenetic bootstrap. Mol. Biol. Evol. 30:1188–1195.

Minh B.Q., Schmidt H.A., Chernomor O., Schrempf D., Woodhams M.D., von Haeseler A., Lanfear R. 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. 37:1530–1534.

Morel B., Barbera P., Czech L., Bettisworth B., Hübner L., Lutteropp S., Serdari D., Kostaki E.-G., Mamais I., Kozlov A.M., Pavlidis P., Paraskevis D., Stamatakis A. 2021. Phylogenetic analysis of SARS-CoV-2 data is difficult. Mol. Biol. Evol. 38:1777–1791.

Pardi F., Goldman N. 2005. Species choice for comparative genomics: being greedy works. PLoS Genet. 1(6):e71.

Price M.N., Dehal P.S., Arkin A.P. 2010. FastTree 2—approximately maximum-likelihood trees for large alignments. PLoS One 5:e9490.

Rabosky D.L., Chang J., Title P.O., Cowman P.F., Sallan L., Friedman M., Kaschner K., Garilao C., Near T.J., Coll M., Alfaro M.E. 2018. An inverse latitudinal gradient in speciation rate for marine fishes. Nature 559:392–395.

Rannala B., Yang Z. 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. J. Mol. Evol. 43:304–311.

Sanderson M.J. 1995. Objections to bootstrapping phylogenies: a critique. Syst. Biol. 44:299–320.

Schultz A.-K., Zhang M., Bulla I., Leitner T., Korber B., Morgenstern B., Stanke M. 2009. jpHMM: improving the reliability of recombination prediction in HIV-1. Nucleic Acids Res. 37:W647–W651.

Sharma S., Kumar S. 2021. Fast and accurate bootstrap confidence limits on genome-scale phylogenies using little bootstraps. Nat. Comput. Sci. 1:573–577.

Simon C. 2022. An evolving view of phylogenetic support. Syst. Biol. 71:921–928.

Soltis P.S., Soltis D.E. 2003. Applying the bootstrap in phylogeny reconstruction. Stat. Sci. 18:256–267.

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

Stamatakis A., Hoover P., Rougemont J. 2008. A rapid bootstrap algorithm for the RAxML web servers. Syst. Biol. 57:758–771.

Steel M. 2005. Phylogenetic diversity and the greedy algorithm. Syst. Biol. 54:527–529.

Susko E. 2009. Bootstrap support is not first-order correct. Syst. Biol. 58:211–223.

Turakhia Y., Thornlow B., Hinrichs A., McBroome J., Ayala N., Ye C., Smith K., De Maio N., Haussler D., Lanfear R., Corbett-Detig R. 2022. Pandemic-scale phylogenomics reveals the SARS-CoV-2 recombination landscape. Nature 609(7929): 994–997.

Van Noorden R., Maher B., Nuzzo R. 2014. The top 100 papers. Nat. News 514:550–553.

Varga T., Krizsán K., Földi C., Dima B., Sánchez-García M., Sánchez-Ramírez S., Szöllősi G.J., Szarkándi J.G., Papp V., Albert L., Andreopoulos W., Angelini C., Antonín V., Barry K.W., Bougher N.L., Buchanan P., Buyck B., Bense V., Catcheside P., Chovatia M., Cooper J., Dämon W., Desjardin D., Finy P., Geml J., Haridas S., Hughes K., Justo A., Karasiński D., Kautmanova I., Kiss B., Kocsubé S., Kotiranta H., LaButti K.M., Lechner B.E., Liimatainen K., Lipzen A., Lukács Z., Mihaltcheva S., Morgado L.N., Niskanen T., Noordeloos M.E., Ohm R.A., Ortiz-Santana B., Ovrebo C., Rácz N., Riley R., Savchenko A., Shiryaev A., Soop K., Spirin V., Szebenyi C., Tomšovský M., Tulloss R.E., Uehling J., Grigoriev I.V., Vágvölgyi C., Papp T., Martin F.M., Miettinen O., Hibbett D.S., Nagy L.G. 2019. Megaphylogeny resolves global patterns of mushroom evolution. Nat. Ecol. Evol. 3:668–678.

Wilkinson M. 1996. Majority-rule reduced consensus trees and their use in bootstrapping. Mol. Biol. Evol. 13:437–444.

Zhu Q., Mai U., Pfeiffer W., Janssen S., Asnicar F., Sanders J.G., Belda-Ferre P., Al-Ghalith G.A., Kopylova E., McDonald D., Kosciolek T., Yin J.B., Huang S., Salam N., Jiao J.-Y., Wu Z., Xu Z.Z., Cantrell K., Yang Y., Sayyari E., Rabiee M., Morton J.T., Podell S., Knights D., Li W.-J., Huttenhower C., Segata N., Smarr L., Mirarab S., Knight R. 2019. Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. Nat. Commun. 10:5477.

Zhukova A., Blassel L., Lemoine F., Morel M., Voznica J., Gascuel O. 2021. Origin, evolution and global spread of SARS-CoV-2. C.R. Biol. 344:57–75.