Lemoine, F. (2018). Renewing Felstein's phylogenetic Bootstrap in the era of big data.
Nature, 556: 452 – 456.
http://dx.doi.org/10.1038/s41586-018-0043-0

Renewing Felsenstein's phylogenetic Bootstrap in the era of big data

F. Lemoine, J.-B. Domelevo Entfellner, E. Wilkinson, D. Correia, M. Dávila Felipe, T. De Oliveira and O. Gascuel

Abstract
Felsenstein's application of the bootstrap method to evolutionary trees is one of the most cited scientific papers of all time. The bootstrap method, which is based on resampling and replications, is used extensively to assess the robustness of phylogenetic inferences. However, increasing numbers of sequences are now available for a wide variety of species, and phylogenies based on hundreds or thousands of taxa are becoming routine. With phylogenies of this size Felsenstein's bootstrap tends to yield very low supports, especially on deep branches. Here we propose a new version of the phylogenetic bootstrap in which the presence of inferred branches in replications is measured using a gradual 'transfer' distance rather than the binary presence or absence index used in Felsenstein's original version. The resulting supports are higher and do not induce falsely supported branches. The application of our method to large mammal, HIV and simulated datasets reveals their phylogenetic signals, whereas Felsenstein's bootstrap fails to do so.

The bootstrap method is a widely used statistical approach to study the robustness, bias and variability of numerical estimates[1,2]. It involves resampling with replacement from the original dataset to obtain replications of the original estimate and then, typically, computing the variance and distribution of this estimate. In 1985, Joseph Felsenstein proposed the use of the bootstrap to assess the robustness, or repeatability, of phylogenetic trees[3]. Given a sequence alignment and a reference tree inferred from it, the procedure is: (i) resample, with replacement, the sites of the alignment to obtain pseudo-alignments of the same length, (ii) infer pseudo-trees using the same inference method and (iii) measure the support of every branch in the reference tree as the proportion of pseudo-trees containing that branch. The usefulness, simplicity and interpretability of this method has made it extremely popular in evolutionary studies, to the point that it is generally required for publication of tree estimates in a wide variety of domains (molecular biology, genomics, systematics, ecology, epidemiology and so on). As a result, Felsenstein's article has been cited more than 35,000 times and is ranked in the top 100 of the most cited scientific papers of all time[4]. However, the use of Felsenstein's bootstrap has been questioned on biological grounds, notably regarding assumptions of site independence and homogeneity[5]. Furthermore, the statistical meaning of Felsenstein's bootstrap proportions

(FBPs) has been the subject of intense debate[6], the main questions being whether FBPs can be seen as the confidence levels of some test and whether or not they are biased[7–10]. Several methods[9,11,12] have been proposed to correct FBP to better agree with standard ideas of confidence levels and hypothesis testing. These works have greatly contributed to the understanding of what Felsenstein's bootstrap is and what it is not. However, FBP correction methods are limited to relatively small datasets for mathematical and computational reasons (for example, double bootstrapping), and the original method is still often used; a Google Scholar search reveals about 2,000 citations of Felsenstein's paper in 2017. As has previously been stated[13], "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". The alternatives to FBPs are the Bayesian posterior probabilities of the tree branches[14]—which are difficult to obtain with large datasets for computational reasons—and the approximate branch supports[15,16], which are computed quickly but provide only a local view. The bootstrap is also computationally heavy, but is easily parallelized and fast algorithms have been designed[17,18].

It is commonly acknowledged[13] that Felsenstein's bootstrap is not appropriate for large datasets that contain hundreds or thousands of taxonomic units (taxa), which are now common as a result of high-throughput sequencing technologies. Though such datasets generally contain a lot of phylogenetic information, the bootstrap proportions tend to be low, especially when the tree is inferred from a single gene, or only a few genes, as illustrated in Fig. 1a with a dataset of approximately 9,000 HIV-1 group M (HIV-1M) *DNA polymerase* (*pol*) sequences. The strongest signal in such a phylogeny generally corresponds to the deep branching of the subtypes. This signal is immediately visible here and is consistent with the common belief regarding subtype branching[19] but some of the HIV-1M subtypes (A, B, D and G) are not supported, and neither is their branching (for example, the grouping of C and H). When using a medium-sized dataset of about 550 randomly selected sequences the FBPs are higher, with most sub-types supported at 70% or more. However, their deep branching is still unresolved (Extended Data Fig. 5).

The reason for such degradation is explained by the core methodology of Felsenstein's bootstrap. A replicated branch must match a reference branch exactly to be accounted for in the FBP value. A difference of just one taxon—which is highly likely to be the case in large datasets— is sufficient for the replicated branch to be counted as absent, even though it is nearly identical to the reference branch[13,20]. There are many biological and computational reasons for the existence of 'rogue' taxa with unstable phylogenetic positions: convergence, recombination, sequence and tree errors, and so on. The standard approach[20–24] is to remove these taxa and relaunch the analysis, but this is statistically questionable and computationally expensive. Furthermore, with a large number of taxa and a low number of sites the phylogenetic signal is weak. The inferred branches are then likely to have errors

and a large fraction of taxa may be unstable, even in the absence of model misspecification of any sort and without long branches.

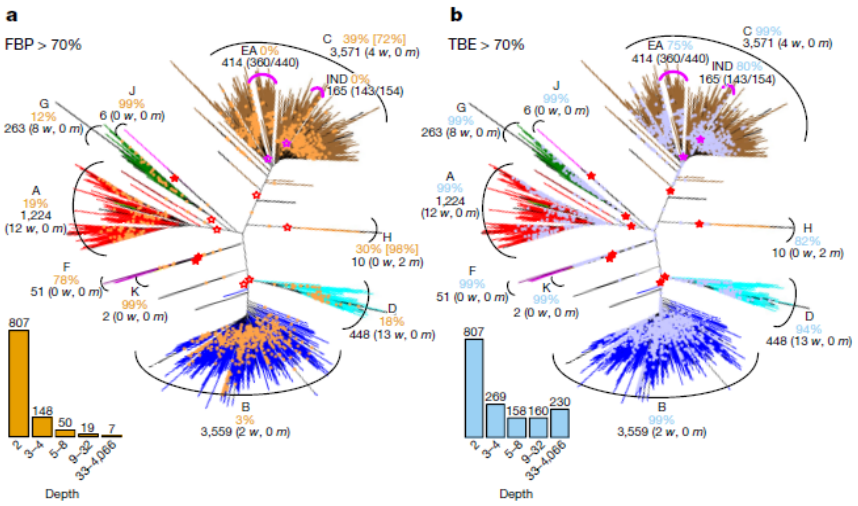


Fig. 1 | FBP and TBE bootstrap supports on the same phylogeny with 9,147 HIV-1M *pol* sequences. a, FBP. **b**, TBE. Subtypes are colourized; recombinant sequences are black; dots correspond to branches with support > 70%[7]. Supports (orange and blue percentages in **a** and **b**, respectively) are given for the tree clades that are closer to the subtypes (red stars, filled when support > 70%); for each of these clades we use jpHMM predictions to provide the number of wrong taxa (*w*) that do not belong to the corresponding subtype, and the number of missing taxa (*m*) that belong to the subtype but not to the clade. The C and the H clades are not supported by FBP but there exist neighbouring clades with FBP supports larger than 70%, which are shown in square brackets. The same approach is applied to the C sub-epidemics in India (IND) and East Africa (EA). The ratio provides the coverage of the clade: the number of studied taxa (for example, from India) in the clade versus the total number of these taxa in the dataset. The South American clade (SA, included in EA, not shown) is supported by TBE but not by FBP (73% versus 14%, respectively, for 15 taxa in total, with above-defined ratio of 14/14). The histograms provide the number of branches with > 70% support depending on branch depth, which is measured by the number of taxa in the smaller of the two clades defined by the given branch.

A statistical approach

Our approach has a simple but sound statistical basis that is partly inspired by Sanderson's monophyly index[25] and partly by our work on gene clusters obtained from expression data[26], both of which are tailored for rooted trees. We replace the branch presence proportion— that is, the expectation of a {0,1} indicator function—of Felsenstein's bootstrap, by the expectation of a refined, gradual function in the [0,1] range, quantifying the branch presence in the bootstrap trees. In doing so, we admit that the inferred branch is not simply correct or incorrect (as with FBP), but that it may contain some errors. Our ultimate aim is to quantify these errors and the presence of the inferred branch in the true tree, using the plug-in principle (see below). We use the transfer distance[27–29], in which the distance $\delta(b,b^*)$ between a branch b of the reference tree T and a branch b^* of a bootstrap tree T^* is equal to the number of taxa that must be transferred (or removed) to make both branches identical (that is, both branches split the set of taxa identically). To measure the presence of b in T^*, we search the branch in T^* that is closest to b and use the 'transfer index', $\varphi(b,T^*)=\mathrm{Min}_{b^*\in T^*}\{\delta(b,b^*)\}$.

This index has several important and useful properties. Any branch b splits the taxa into two subsets. If l is the number of taxa and p is the size of the smaller subset induced by b, we have the following properties (see Methods): $\varphi(b,T^*) = 0$ if and only if b belongs to T^*;

$\varphi(b,T^*) \leq p-1$; $\varphi(b,T^*)/(p-1)$ is very close to 1 when T^* is random and l is large (>100); and $\varphi(b,T^*)$ is computed recursively in time proportional to l, just as is FBP. On the basis of these properties, we define the transfer just as is FBP. On the basis of the bootstrap expectation (TBE) as:

$$\mathrm{TBE}(b) = 1 - \frac{\overline{\phi(b,T^*)}}{p-1}$$

in which the numerator is the average transfer index among all bootstrap trees. It can easily be seen that TBE ranges from 0 to 1, in which 0 means that the bootstrap trees are random regarding b and 1 means that b appears in all bootstrap trees. Considering the same set of bootstrap trees, TBE(b) is necessarily larger than FBP(b) and the difference is substantial for deep branches, whereas TBE(b) = FBP(b) when b defines a (shallow) 'cherry', which is a clade comprising only two taxa (that is, $p=2$). Importantly, we shall see that TBE supports very few branches showing substantial contradictions with the true tree when used with common thresholds (typically 70%[7] or higher; Fig. 2c, d and Extended Data Figs. 2, 3, 7, 8).

These properties—easy computation, higher supports than FBP and a low number of falsely supported branches—are all highly desirable. Furthermore, TBE has a simple and natural interpretation; for instance, with l = 1,000 and p = 200, TBE(b) = 95% means that, on average, (200 − 1) × 0.05 ≈ 10 taxa have to be transferred to recover b in bootstrap replicate trees. This interpretation is radically different from that of FBP, in which b is assessed globally as correct or erroneous. With TBE, branches that are nearly correct are also likely to be supported. Moreover, we can define an instability score for each taxon based on the number of times it is transferred in TBE computations.

TBE uses the same procedure of resampling with replacement as does FBP and thus inherits some of the statistical properties of FBP[3,6,9], as well as the usual properties of the bootstrap method[1,2]. Notably, TBE relies on the same assumptions as FBP regarding site independence and homogeneity, but these assumptions can be relaxed[6], for instance, by using block bootstrapping[30]. Just as with FBP[3], TBE(b) cannot be interpreted as the probability for the branch b to belong to the true phylogeny. Although deep mathematical approaches[6,9–12,31] have previously been proposed to connect FBP to hypothesis-testing theory, TBE should not be interpreted as the confidence level of some statistical test (with null and alternative hypotheses, distribution of test statistics under the null and so on). TBE is better and more simply interpreted in terms of repeatability: TBE(b) estimates the extent to which branches identical or similar to b would be recovered when applying the same tree inference method to a new sample of the same size drawn from the same site distribution as the original sample. With large samples, the empirical distribution obtained from observed data comes close to the unknown underlying distribution of this data, and sampling with replacement in the empirical distribution is asymptotically equivalent to drawing samples from the underlying distribution[1,2].

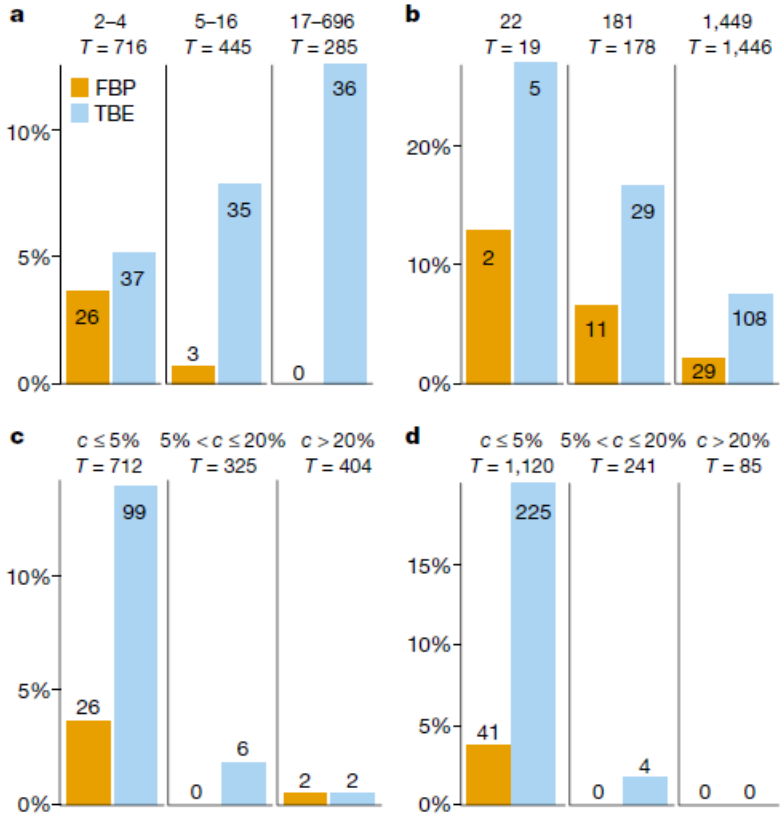


Fig. 2 | FBP and TBE bootstrap supports on the same phylogeny with 1,449 COI-5P mammal sequences using FastTree. The graphs in **a–d** refer to branches with supports > 70%[7], with the vertical axes denoting the percentage of these branches in a given condition (in **b** for example, 22-taxon trees contain 19 internal branches (*T*), and 2/19 ≈ 10% of branches have FBP > 70%). **a**, Supports regarding branch depth (see definition in Fig. 1). **b**, Supports regarding tree size (that is, number of taxa). **c**, Supports regarding percentage of quartet conflicts (*c*) with NCBI taxonomy ($c \leq 5\%$, low level of conflict; $5\% < c \leq 20\%$, moderate level of conflict; $c > 20\%$, high level of conflict). **d**, As in **c** but regarding the true tree used for simulations (noisy condition). *T*, number of internal branches.

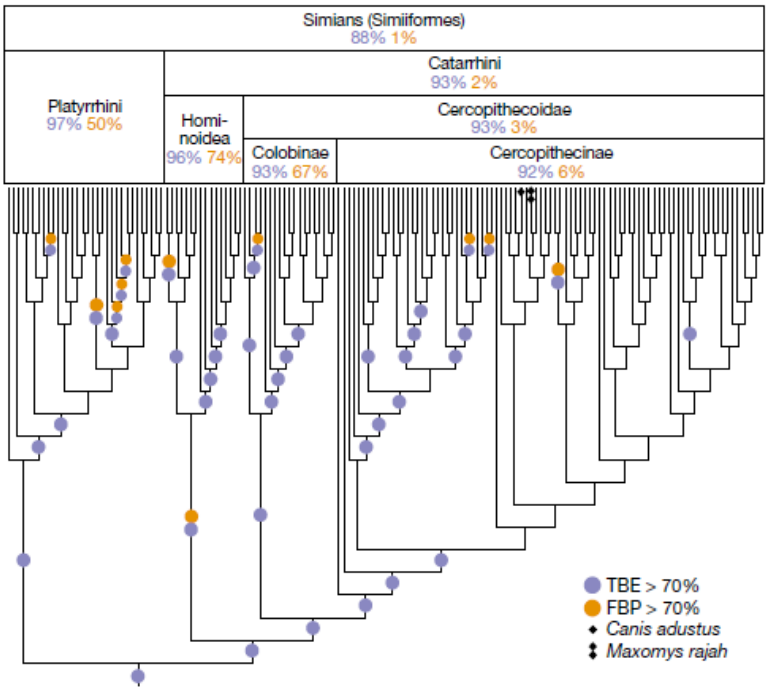


Fig. 3 | FBP and TBE bootstrap supports on the simian clade. The complete tree inferred by FastTree using 1,449 COI-5P mammal sequences is the same as in Fig. 2, but we focus on the simian clade. All simian sequences are included in this clade and two additional non-simian sequences are added, one rogue taxon (*Maxomys rajah*, detected by TBE) and one stable but erroneous taxon with partial sequence (*Canis adustus*); this simian tree is very close to the NCBI taxonomy (< 2.5% of contradicted quartets, when both erroneous taxa are pruned). Platyrrhini, New World monkeys; Cercopithecoidae, Old World monkeys.

The convergence rate is unknown with models as complex as the ones used in phylogenetics, but our simulation results show that moderate sample sizes suffice to obtain good approximations (Extended Data Fig. 10). When the sample size is extremely large, as in phylogenomic studies using genome-scale sequence alignments[32], both FBP and TBE are expected to be nearly equal to 1 for all branches. Again, this should not be interpreted in terms of absolute truth regarding the phylogenetic inferences, but it simply reflects the closeness of the empirical and underlying distributions and the very small variability of tree estimates. In fact, a high level of repeatability is necessary to trust phylogenetic inferences, but it may be not sufficient. Felsenstein[3] states that the bootstrap "may be misleading if the method used to infer phylogenies is inconsistent". This applies both to FBP and TBE, and is typical for inference methods subject to long-branch attraction. With a consistent, unbiased inference method, we expect the plug-in principle[1,2,6,9] to apply; this principle states that the distribution of the distance between the true tree and the inferred tree can be well-approximated by the distribution of the distance between the inferred and bootstrap trees. Using both real and simulated data, here we show that this principle does apply with maximum-likelihood estimation, a phylogenetic inference method that is typically consistent[33]. In this setting, TBE provides information on the (transfer, quartet-based) distance between the inferred branch and the true tree, and rarely supports poor

branches. Moreover, the ability of TBE to identify rogue taxa makes it possible to study them further, to understand why they are phylogenetically unstable and to revise the branch supports.

Analysis of mammal data

We first studied the advantage of using TBE on a large phylogeny of 1,449 mammals, obtained from a usual barcoding marker (COI-5P). The reference and bootstrap trees were inferred by maximum likelihood from the protein alignment (527 sites) using both FastTree[34] and RAxML with rapid bootstrap[17] to check that similar conclusions were drawn with different inference methods. To study the effect of the number of taxa, we randomly selected small- (22 taxa) and medium-sized (181 taxa) datasets and performed the same analyses. The results were compared to the NCBI taxonomy (https://www.ncbi.nlm.nih.gov/ taxonomy, accessed April 2016), which represents current thinking about the evolutionary history of mammals. To cope with the low resolution of the NCBI taxonomy, we used a quartet-based topological distance rather than the transfer distance. For all inferred branches, we measured the quartet-based percentage of conflicts with the NCBI taxonomy, and the same approach was used to assess the topological accuracy of FastTree and RAxML phylogenies. As expected in this type of study based on a unique marker, the inferred topologies were relatively poor, and thus challenging for branch support methods. However, RAxML was more accurate than FastTree and had higher branch supports, as is generally observed with rapid bootstrap[16] (Extended Data Figs. 2, 3).

Our results (Fig. 2a–c and Extended Data Figs. 2, 3) indicate clearly that TBE provides some support for deep branches, whereas FBP does not. As expected, the supports for shallow branches are similar between the two methods, and the advantage of TBE is more pronounced with a large number of taxa but still of interest with medium-sized datasets. Comparisons with the NCBI taxonomy show that TBE supports a larger number of weakly contradicted branches than FBP—which fulfils one of the objectives of TBE (nearly correct branches must be supported)— and the number of supported branches with moderate-to-high quartet conflicts remains very low. These results are confirmed by simulations (Fig. 2d and Extended Data Figs. 7, 8). The advantage of TBE appears clearly when inspecting the tree clades. For example (Fig. 3), the simian clade inferred by FastTree has a strong support with TBE; by contrast, when using FBP, support for this clade is nearly null owing to a large number of rogue taxa in the bootstrap trees, and the same holds true for several sub-clades. The simian clade includes all 152 simian sequences plus two non-simian taxa (*Maxomys rajah* and *Canis adustus*).

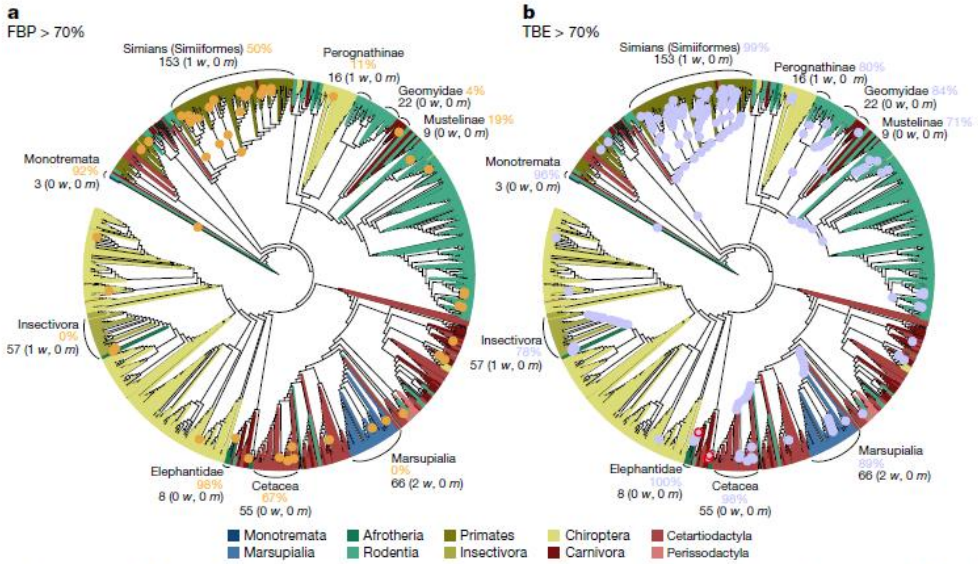


Fig. 4 | FBP and TBE supports on the same phylogeny with 1,449 COI-5P mammal sequences using RAxML with rapid bootstrap. **a**, FBP. **b**, TBE. This phylogeny is more accurate than the one inferred using FastTree (27% versus 38% of contradicted quartets, respectively) but is still relatively poor, especially regarding deep nodes and larger groups. For example, rodents and chiropterans are not monophyletic and are distributed in several subtrees. However, some parts of the tree are more accurate. A few clades are highlighted, corresponding almost exactly to the NCBI taxonomy. For example, all Elephantidae taxa are recovered by RAxML in a single clade that contains only Elephantidae, and Insectivora taxa are included in a clade containing one extra taxon. To select these clades, we minimized the transfer distance with the NCBI taxonomy, in case of ambiguity. See Fig. 1 and Methods for details.

The latter is not a rogue taxon: its sequence is incomplete and very close to the simian sequences for the part that is available, and its position is very stable in the bootstrap trees. By contrast, *Maxomys rajah* is a rogue taxon and is detected as such by TBE (transferred in 659 out of 1,000 bootstrap trees when computing the support of the simian clade). Similar results were found with other well-established clades when using RAxML (Fig. 4). Both FBP and TBE support some small clades, namely the Monotremata and Elephantidae. However, FBP does not support any deep branches, except for the Cetacea (67%) and, to some extent, the simians (50%). TBE provides strong supports for these two groups, but also for five other groups, including the Marsupiala and Insectivora. The latter clade (FBP: 0%, TBE: 78%) contains all Insectivora of the NCBI taxonomy, plus one extra taxon (*Plecotus strelkovi*), which again is detected by TBE as a rogue taxon (transferred in 965 out of 1,000 bootstrap trees). By comparison, the removal of rogue taxa[24] does not substantially improve FBP: eight and three taxa are removed with FastTree and RAxML, respectively, but the number of branches with FBP > 70% remains the same. This is explained by hundreds of taxa, which are relatively unstable but not removed.

Analysis of HIV data

We applied our method to a large dataset of 9,147 HIV-1M *pol* sequences. Datasets of this size are increasingly common in molecular epidemiology and phylodynamics[35]. We retained only sequences that were annotated as non-recombinant by the Los Alamos HIV-1 database using a fast-filtering approach. Among these sequences, 48 recombinant sequences were detected by jpHMM[36]. These 48 sequences were kept in the analyses to study the effect of recombinant sequences, as their presence is inevitable in any HIV dataset. In contrast to that of mammals, the tree topology of HIV-1M strains is essentially

unknown. Moreover, it is intrinsically unstable because reconstructing a tree with so many relatively short and possibly recombinant sequences is challenging. Thus, the main expectation is to observe a clear separation between the subtypes. We built the reference and bootstrap trees using FastTree on the DNA sequence alignment (1,043 sites), and performed the same analyses using smaller subsets of 35 and 571 sequences. Although the deep branching of the subtypes[19] is poorly supported by FBP (Fig. 1a), it becomes apparent with TBE, as when using this approach all subtypes have a support larger than 80% and close to 100% in most cases (Fig. 1b and Extended Data Fig. 5). For example, the subtype B clade (3,559 taxa) has a support of only 3% using FBP, but a support of 99% using TBE. This clade contains all subtype B sequences, plus two taxa detected as recombinant by jpHMM; this means that both supports are likely to be correct insofar as they state that this clade is incorrect (FBP) or nearly correct (TBE). However, FBP fails to detect any phylogenetic signal, whereas TBE reveals that this signal is very strong. The same holds true with other well-described clades. For example, TBE supports the identification of regional variants of HIV-1 subtypes that are of epidemiological importance (such as the East African, Indian and South American subtype C variants), which FBP fails to support. TBE provides a substantial support to a much larger number of deep branches. Again, the advantage of using TBE is higher with large datasets (Extended Data Fig. 4), but is still apparent when using 571-taxon datasets, for which the deep subtype branching and C sub-epidemics are supported by TBE but not FBP (Extended Data Fig. 5). An important feature of TBE is that the supports may be non-local, but attached to 'caterpillar-like' paths, in which the main phylogenetic backbone is connected to a few isolated taxa (Fig. 1b; for example, subtype C). With HIV-1M data, this corresponds to the fact that the subtype roots are usually not well defined owing to recombinant and ancient sequences, which tend to be isolated in basal position. Moreover, the instability score among recombinant sequences is clearly higher than in the sequences that were not detected as recombinant (Extended Data Fig. 6), which supports the biological soundness of the approach and its power to detect recombinant and rogue taxa.

Analysis of simulated data

To check that TBE does not support erroneous branches, we performed extensive computer simulations with various tree sizes and phylogenetic signal levels. We also added unstable taxa that had a weaker phylogenetic signal than the others. The results are highly similar to those obtained using real data, regarding the support of deep branches and the tree size (Extended Data Figs. 7, 8). In all the conditions we examined, TBE supported very few branches that showed substantial contradictions with the true tree, and the rogue taxa exhibited lower stability (Extended Data Fig. 9). In the absence of rogue taxa (Extended Data Fig. 7), the gain of TBE was still substantial compared to FBP, with almost twice as many branches with support > 70%, thus demonstrating the importance of accounting for the global instability of the inferred tree. Furthermore, we checked the interpretation of TBE as a measure of repeatability (Extended Data Fig. 10) by comparing TBE to its counterpart computed from simulated alignments, rather than bootstrap pseudo-alignments; both simulation- and bootstrap-based supports are highly correlated (Pearson's ρ = 0.85) with alignments of moderate length (about 500) and have

analogous performance in detecting rogue taxa. Lastly, we checked the validity of the plug-in principle by comparing TBE to the similarity—measured using the normalized transfer index—between the inferred branch and the true tree (Extended Data Fig. 10). Again a high correlation (Pearson's ρ = 0.74) was found. When performing the same experiments with FBP similar or slightly lower correlations were observed, probably owing to the discontinuous nature of FBP.

Discussion

The transfer bootstrap thus provides a measure of branch repeatability, or robustness. Our results clearly demonstrate its usefulness, especially with deep branches and large datasets, for which branches known to be essentially correct are supported by TBE but not by FBP. Furthermore, when combined with consistent maximum-likelihood tree estimation, TBE rarely supports poor branches. Importantly, TBE supports are easily interpreted as fractions of unstable taxa. Although our results suggest that 70% is a reasonable threshold from which to start (Extended Data Figs. 2, 3, 4, 8), we suggest that it is better to interpret TBE values depending on the data and the phylogenetic question being addressed; for example, using a lower TBE support threshold with HIV and possibly recombinant sequences, than with mammals. Moreover, our experiments demonstrate the ability of the transfer index to detect unstable taxa responsible for low supports. Lastly, the approach is applicable to rapid bootstrap[17,18] (Fig. 4 and Extended Data Fig. 3) and could be extended to parametric bootstrap[2] and Bayesian branch supports[14].

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586- 018-0043-0.

1. Efron, B. Bootstrap methods: another look at the jackknife. *Ann. Stat.* **7**, 1–26 (1979).
2. Efron, B. & Tibshirani, R. J. An Introduction to the Bootstrap (Chapman & Hall, New York, 1993).
3. Felsenstein, J. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**, 783–791 (1985).
4. Van Noorden, R., Maher, B. & Nuzzo, R. The top 100 papers. *Nature* **514**, 550–553 (2014).
5. Sanderson, M. J. Objections to bootstrapping phylogenies: a critique. *Syst. Biol.* **44**, 299–320 (1995).
6. Holmes, S. Bootstrapping phylogenetic trees: theory and methods. *Stat. Sci.* **18**, 241–255 (2003).
7. Hillis, D. M. & Bull, J. J. An empirical test of bootstrapping as a method for assessing confidence in phyogenetic analysis. *Syst. Biol.* **42**, 182–192 (1993).
8. Felsenstein, J. & Kishino, H. Is there something wrong with the bootstrap on phylogenies? A reply to Hillis and Bull. *Syst. Biol.* **42**, 193–200 (1993).
9. Efron, B., Halloran, E. & Holmes, S. Bootstrap confidence levels for phylogenetic trees. *Proc. Natl Acad. Sci. USA* **93**, 7085–7090 (1996).
10. Susko, E. Bootstrap support is not first-order correct. *Syst. Biol.* **58**, 211–223 (2009).
11. Zharkikh, A. & Li, W.-H. Estimation of confidence in phylogeny: the complete- and-partial bootstrap technique. *Mol. Phylogenet. Evol.* **4**, 44–63 (1995).
12. Susko, E. First-order correct bootstrap support adjustments for splits that allow hypothesis testing when using maximum likelihood estimation. *Mol. Biol. Evol.* **27**, 1621–1629 (2010).
13. Soltis, D. E. & Soltis, P. S. Applying the bootstrap in phylogeny reconstruction. *Stat. Sci.* **18**, 256–267 (2003).
14. Huelsenbeck, J. & Rannala, B. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. *Syst. Biol.* **53**, 904–913 (2004).
15. Anisimova, M. & Gascuel, O. Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. *Syst. Biol.* **55**, 539–552 (2006).
16. Anisimova, M., Gil, M., Dufayard, J. F., Dessimoz, C. & Gascuel, O. Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. *Syst. Biol.* **60**, 685–699 (2011).
17. Stamatakis, A., Hoover, P. & Rougemont, J. A rapid bootstrap algorithm for the RAxML Web servers. *Syst. Biol.* **57**, 758–771 (2008).
18. Minh, B. Q., Nguyen, M. A. T. & von Haeseler, A. Ultrafast approximation for phylogenetic bootstrap. *Mol. Biol. Evol.* **30**, 1188–1195 (2013).
19. Hemelaar, J. The origin and diversity of the HIV-1 pandemic. *Trends Mol. Med.* **18**, 182–192 (2012).
20. Sanderson, M. J. & Shaffer, H. B. Troubleshooting molecular phylogenetic analyses. *Annu. Rev. Ecol. Syst.* **33**, 49–72 (2002).
21. Wilkinson, M. Majority-rule reduced consensus trees and their use in bootstrapping. *Mol. Biol. Evol.* **13**, 437–444 (1996).
22. Thorley, J. L. & Wilkinson, M. Testing the phylogenetic stability of early tetrapods. *J. Theor. Biol.* **200**, 343–344 (1999).

23. Thomson, R. C. & Shaffer, H. B. Sparse supermatrices for phylogenetic inference: taxonomy, alignment, rogue taxa, and the phylogeny of living turtles. *Syst. Biol.* **59**, 42–58 (2010).
24. Aberer, A. J., Krompass, D. & Stamatakis, A. Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and webservice. *Syst. Biol.* **62**, 162–166 (2013).
25. Sanderson, M. J. Confidence limits on phylogenies: the bootstrap revisited. *Cladistics* **5**, 113–129 (1989).
26. Bréhélin, L., Gascuel, O. & Martin, O. Using repeated measurements to validate hierarchical gene clusters. *Bioinformatics* **24**, 682–688 (2008).
27. Charon, I., Denoeud, L., Guénoche, A. & Hudry, O. Maximum transfer distance between partitions. *J. Classif.* **23**, 103–121 (2006).
28. Day, W. H. E. The complexity of computing metric distances between partitions. *Math. Soc. Sci.* **1**, 269–287 (1981).
29. Lin, Y., Rajan, V. & Moret, B. M. E. A metric for phylogenetic trees based on matching. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **9**, 1014–1022 (2012).
30. Künsch, H. R. The jackknife and the bootstrap for general stationary observations. *Ann. Stat.* **17**, 1217–1241 (1989).
31. Billera, L. J., Holmes, S. P. & Vogtmann, K. Geometry of the space of phylogenetic trees. *Adv. Appl. Math.* **27**, 733–767 (2001).
32. Kumar, S., Filipski, A. J., Battistuzzi, F. U., Kosakovsky Pond, S. L. & Tamura, K. Statistics and truth in phylogenomics. *Mol. Biol. Evol.* **29**, 457–472 (2012).
33. Truszkowski, J. & Goldman, N. Maximum likelihood phylogenetic inference is consistent on multiple sequence alignments, with or without gaps. *Syst. Biol.* **65**, 328–333 (2016).
34. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
35. Grenfell, B. T. et al. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* **303**, 327–332 (2004).
36. Schultz, A.-K. et al. jpHMM: improving the reliability of recombination prediction in HIV-1. *Nucleic Acids Res.* **37**, W647–W651 (2009).

Acknowledgements We thank F. Delsuc, S. Holmes, L. Chindelevitch and E. Susko for help and suggestions. This work was supported by the EU-H2020 Virogenesis project (grant number 634650, to E.W., T.D.O. and O.G.), by the INCEPTION project (PIA/ANR-16-CONV-0005, to F.L., D.C., M.D.F. and O.G.), by the Institut Français de Bioinformatique (IFB - ANR-11-INBS-0013, to D.C.), by the Flagship grant from the South African Medical Research Council (MRC-RFA-UFSP-01-2013/UKZN HIVEPI to E.W., T.D.O. and J.-B.D.E.) and by the H3ABioNet project (NIH grant number U41HG006941 to J.-B.D.E. and U24HG006941 to E.W. and T.D.O.).

Reviewer information *Nature* thanks E. Susko and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions O.G. designed the research; F.L., J.-B.D.E., M.D.F. and O.G. performed the research; F.L. and J.-B.D.E. implemented the algorithms; F.L. and

D.C. realized the website and GitHub repositories; F.L. performed the analyses and graphics, with the help of E.W. and T.D.O. for HIV; O.G. wrote the paper with the help of all co-authors.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0043-0.

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0043-0.

Reprints and permissions information is available at http://www.nature.com/reprints.

Correspondence and requests for materials should be addressed to O.G.

Methods

Definitions and properties of transfer distance and index.

The transfer distance[27], also called R-distance[28], was introduced to compare partitions in cluster analysis. In this context, the transfer distance is equal to the minimum number of elements to be transferred (or removed) to transform one partition into the other. Tree branches are commonly seen as bipartitions or splits, as a branch divides the taxa into two subsets situated on its two sides. The most used topological distance between two trees is the Robinson–Foulds distance[37], which is equal to the number of bipartitions that belong to one tree but not the other. The bipartition distance is overly sensitive to some small tree changes, possibly involving a unique taxon[29]. Previous authors[29] have proposed using the transfer distance and designed algorithms to compute a more robust 'matching' distance between trees; although a different task, this is related to the aim of this article. In the following, we first provide basic definitions—following the standard text book for phylogenetic trees[38]—and then demonstrate the properties of the transfer distance in a bootstrap context.

Let X be a fixed set of l taxa. An X-tree is a phylogenetic tree with l leaves labelled by the taxa of X. All reference and bootstrap trees discussed here are X-trees, meaning that they are labelled by the same set of l taxa. Any branch of an X-tree defines a bipartition of X, and the topology of an X-tree can be recovered from its bipartition set. Thus, we will use the terms branch and bipartition to mean the same thing in different contexts. Any bipartition, b, of X can be encoded by a {0,1} vector $\boldsymbol{v}(b)$ of length l, in which the taxa on the same side of the bipartition are encoded by the same value. Note that b is also encoded by $\bar{\boldsymbol{v}}(b)$, the negation of $\boldsymbol{v}(b)$ (that is, the zeros are turned into ones, and vice versa). Moreover, the smaller of the two subsets induced by a bipartition b will be called here the 'light side' of b, and p will denote the size of the light side of b ($p \leq l - p$). A bipartition is 'trivial' when it has a unique taxon in its light side ($p = 1$). An X-tree defines l trivial bipartitions corresponding to each of the taxa. These

trivial bipartitions are contained in every X-tree, while the other non-trivial bipartitions define the core of the tree topology and are the central subject of phylogenetic studies.
The transfer distance $\delta(b,b^*)$ between a bipartition b of the reference tree T and a bipartition b^* of a bootstrap tree T^* is equal to the number of taxa that must be transferred (or removed) to make both bipartitions identical. The transfer distance is easily defined and computed using the Hamming distance, H, between $\boldsymbol{v}(b)$ and $\boldsymbol{v}(b^*)$:

$$\delta(b,b^*) = \mathrm{Min}\{H(v(b),v(b^*)),H(\bar{v}(b),v(b^*))\}$$

To measure the presence of b in T^*, we search the bipartition in T^* that is closest to b and use the transfer index $\varphi(b,T^*) = \mathrm{Min}_{b^* \in T^*}\{\delta(b,b^*)\}$. Based on above definitions, $\delta(b,b^*) = 0$ if and only if $\boldsymbol{v}(b)$ and $\boldsymbol{v}(b^*)$ define the same bipartition of X. Thus, the transfer index satisfies $\varphi(b,T^*) = 0$ if and only if $b \in T^*$. Moreover, let b be any given bipartition of T and t be a taxon on the light side of b. The trivial bipartition $b^* = \{t\}|X - \{t\}$ is found in any bootstrap tree T^* and $\delta(b,b^*) = p - 1$. There may well be another bipartition closer to b in T^*, but at least this ensures that $\varphi(b,T^*) \leq p - 1$, and thus the transfer support, TS, satisfies:

$$\mathrm{TS}(b,T^*) = 1 - \frac{\phi(b,T^*)}{p-1} \in [0,1]$$

and $\mathrm{TS}(b,T^*) = 1$ if and only if $b \in T^*$. Let $1_b(T^*)$ be the indicator function equal to 1 when $b \in T^*$ and 0 otherwise. For any bipartition b and tree T^*, we have $1_b(T^*) \leq \mathrm{TS}(b,T^*)$. The FBP is equal to the average of $1_b(T^*)$ over the set of bootstrap trees, while the TBE is equal to the average of $\mathrm{TS}(b,T^*)$. Thus, when using the same set of bootstrap trees, we necessarily have $\mathrm{FBP}(b) \leq \mathrm{TBE}(b)$. When b is a cherry ($p = 2$), we have $1_b(T^*) = \mathrm{TS}(b,T^*)$ and thus $\mathrm{FBP}(b) = \mathrm{TBE}(b)$. With deeper bipartitions, we generally observe that in the presence of a clear phylogenetic signal, only a small number of taxa need to be transferred to make b identical to a bipartition in T^*, while the strict presence of b in T^* can be relatively rare; the difference between $\mathrm{FBP}(b)$ and $\mathrm{TBE}(b)$ can then be substantial. The transfer distance and index are related to parsimony. The branch b is equivalent to a binary {0,1} character; assuming that the tips of T^* are labelled accordingly, we can define $\mathrm{PA}(b,T^*)$, which is the minimum number of changes along T^* branches required to explain the labels of the tips. When b belongs to T^*, we have $\mathrm{PA}(b,T^*) = 1$, and the more shuffled the zeros and ones among the tips of T^*, the higher is $\mathrm{PA}(b,T^*)$. It is easy to see that $\mathrm{PA}(b,T^*) \leq \varphi(b,T^*) + 1$. Indeed, let b^* be a branch in T^* such that $\varphi(b,T^*) = \delta(b,b^*)$ and assume, without loss of generality, that $\delta(b,b^*)$ is equal to the number of tips labelled 1 in the light side of b^* plus the number of tips labelled 0 in the heavy side of b^* (in other words, the light side of b^* is mostly 0 and the heavy side is mostly 1). Now consider that all internal nodes in the light side are 0 and all internal nodes in the heavy side are 1; this implies a number of changes equal to $\varphi(b,T^*) + 1$, which by the definition of parsimony is larger than or equal to $\mathrm{PA}(b,T^*)$. Parsimony is thus another option to measure branch presence, but it is inappropriate in our context. For example, consider a reference branch $b = AB|CD$, in which A, B, C and D are four large 'corner' subtrees, and a tree T^* with an internal branch b^* grouping the corner subtrees the other way around (for example, $b^* = AC|BD$, meaning that A and C sit on one side of b^*,

and B and D on the other side). Then, PA(b,T^*) is equal to 2, a very low value, whereas T^* is phylogenetically very different from b because both clades defined by b are mixed. In this case, the transfer index between b and T^* is much larger and equal to the minimum size of A, B, C and D.

Recursive computation of the transfer index. A recursive algorithm to compute all transfer distances between any given bipartition b of T and all bipartitions of another tree T' has previously been described[26,29]. This algorithm is easily transformed to compute the transfer index. The principle is as follows:

1. Map all the leaves of the light side of b to 0, the others to 1 and apply the same mapping to the leaves of T^*. Furthermore, root T^* at any internal node.
2. With a single post-order tree traversal, one can compute the number of leaves labelled 0 and the number of leaves labelled 1 for every subtree in T^*.
3. Let l_0 be the number of leaves labelled 0 and l_1 be the number of leaves labelled 1 in the subtree attached below a given bipartition b^*. The transfer distance between b and b^* is given by $\delta(b,b^*) = \mathrm{Min}\{p - l_0 + l_1, l - p - l_1 + l_0\}$ (think to the missing zeros and the ones to be removed in b^* below subtree, and vice versa). This distance can be computed during the post-order traversal as well as the transfer index $\varphi(b,T^*)$, which is the minimum of $\delta(b,b^*)$ for all bipartitions of T^*. This algorithm has linear time complexity, and thus computing TBE for all bipartitions in T with r bootstrap replicates has a time complexity in $O(rl^2)$. FBP has the same time complexity, but very efficient implementations have been developed (for example, using bit vectors to encode bipartitions). In practice, computing all TBE supports with 4,000 taxa and 1,000 replicates requires less than one hour (5 core Intel Xeon 3.5 GHz), which is negligible compared to the time required to infer the reference and bootstrap trees.
4.

Expected transfer index with random trees and TBE distribution. We have seen that the transfer index satisfies $\varphi(b,T^*) \leq p - 1$. We show here that the expected transfer index is very close to this upper bound with random 'bootstrap' trees when the number of taxa is large enough. Consequently, the transfer bootstrap expectation of any branch b ($\mathrm{TBE}(b) = 1 - \varphi(b, T^*)/(p-1)$) is close to 0 when the bootstrap trees seem to be random and do not contain any signal regarding b. This property explains why moderate supports—for example, 70% as used throughout this paper—are sufficient to reject poor branches, as a branch support of 70% cannot be observed by chance. We first provide a simple argument to explain this result, based on the expected transfer distance between a fixed bipartition b and a random bipartition b^* with fixed light-side size p^*. Let $x = p/l$ denote the proportion of taxa in the light side of b ($x \leq 1 - x$ because $p \leq l - p$). Both bipartitions b (fixed) and b^* (random) define four taxon subsets, the sizes of which follow hypergeometric distributions with expectations: E(light side of $b \cap$ light side of b^*) $= xp^*$; E(light side of $b \cap$ heavy side of b^*) $= x(l - p^*)$; E(heavy side of $b \cap$ light side of b^*) $= (1 - x)p^*$; and E(heavy side of $b \cap$ heavy side of b^*) $= (1 - x)(l - p^*)$. It is easily seen that under these assumptions, the expected transfer distance between b and b^* is equal to the sum of the second and third (anti-diagonal) terms: that is, $E[\delta(b,b^*)] = (1 - 2x)p^* + p$. As $p^* > 0$ and $(1 - 2x) \geq 0$, we have: $E[\delta(b,b^*)] \geq p$. This result shows that the expected transfer distance between b and b^* is larger than or equal to p, for any value of p and p^*. Moreover, with

a lower p^*, the expected transfer distance is closer to p. As a first approximation, we thus see that the transfer index should be close to its upper-bound $p - 1$, because it is equal to the minimum of distances which taken separately are all expected to be larger than p. However, these distances fluctuate around their expected values, and their minimum may be lower than the minimum of their individual expectations, especially when using small samples (that is, low number of taxa). We performed computer simulations to measure the extent of this phenomenon and the validity of the $E[\varphi(b,T^*)] \approx p - 1$ approximation. We used four tree sizes: l = 16, 128, 1,024 and 8,192 taxa, and four models of random phylogenetic trees: caterpillars (fully imbalanced), PDA, Yule–Harding and perfectly balanced[38]. For the bipartition b, all possible integer values of p in the $[2, l/2]$ range were used. The number of random bootstrap trees was equal to 1,000, and we performed 100 runs per tree size. Results are displayed in Extended Data Fig. 1. With $l \geq 1{,}024$, the average transfer index with random trees is very close in relative value to the upper bound $p - 1$, and the approximation is already satisfying with l = 128. Moreover, the results are nearly the same for the four random tree models, suggesting that the property holds in a number of settings. As expected, the approximation is better with small p. Indeed, note that the upper bound $p - 1$ is obtained with a trivial bipartition b^* made of a unique taxon belonging to the light side of b. When a cherry in T^* contains two taxa from the light side of b, then $\varphi(b,T^*) \leq p - 2$. Similar deviations are observed with subtrees in T^* containing a large fraction of taxa belonging to the light side of b. With a larger p, there is a higher probability for such an event to occur. Note, however, that large values of p (that is, $p \approx 2$) are relatively rare for most tree models (for example, Yule–Harding). Looking at the distribution of TBE, we see that having TBE larger than a moderate threshold (such as 50%) is very unlikely, even with 16 taxa, thus explaining that TBE rarely supports poor branches with real and simulated data (Fig. 2c, d and Extended Data Figs. 2, 3, 7, 8).

Software programs and web server. We developed several tools to compute the transfer bootstrap. We first implemented a command line tool in C, 'Booster' (open source, available at https://github.com/evolbioinfo/booster). This tool computes TBE as well as FBP supports, and the stability scores of the taxa (globally or per branch). It takes two files as input: (1) a reference tree file in Newick format and (2) a bootstrap tree file in Newick format, containing all bootstrap trees. A number of software programs can be used to infer trees from multiple sequence alignments (MSAs) and produce these reference and bootstrap files in the desired format; these include RAxML, FastTree and PhyML—used in this article—as well as many others (see examples in Booster GitHub repository). We also developed 'BoosterWeb' (http://booster.c3bi.pasteur.fr), a freely available web interface that enables users to compute bootstrap supports (TBE and FBP) easily without installing any tool on their own computer. Computations are launched on the Institut Pasteur cluster throughout a Galaxy instance. As with the command line tool, this includes the option to input reference and bootstrap trees inferred using any phylogenetic program. Another option is to upload an MSA and then run PhyML-SMS[39] (for medium-size datasets) or FastTree (for large datasets) to infer the trees. We propose a basic visualization of the resulting tree highlighting highly supported branches at a given threshold. The resulting tree can be uploaded in one-click on iTOL[40] for further

manipulation. Moreover, BoosterWeb is self-contained and can be easily installed on any desktop computer (Windows, MacOS and Linux) by downloading the BoosterWeb executable.

For the sake of reproducibility, all analyses described in this article were implemented in the NextFlow workflow manager[41], and are accessible along with all our data at https://github.com/evolbioinfo/booster-workflows. The software programs that we developed to manipulate data are available for download at http://github. com/fredericlemoine/goalign and http://github.com/fredericlemoine/gotree, for manipulating alignments and trees, respectively.

Mammal dataset and analyses. We downloaded all aligned mammals COI-5P amino acid sequences from the Barcode of Life Data System (http://www. barcodinglife.org, accessed September 2015). We removed all sequences shown to be identical among several species, kept one sequence per species (several gene versions are available for some species, but no paralogues), and converted the resulting multiple alignment (1,449 sequences, 527 sites) into FASTA format. This alignment was subsampled to study the effect of tree size. We randomly drew 8 samples with 1/8th of the sequences (that is, 181) and 64 samples with 1/64th of the sequences (that is, 22). We then generated 1,000 bootstrap alignments for the full alignment and each of the 72 subsampled alignments by drawing sites with replacement. We used FastTree[34] (options: -nopr -nosupport -wag -gamma) to infer trees from each of these reference ($1 + 8 + 64 = 73$) and bootstrap (73,000) alignments. To ensure that the results and conclusions were independent of the tree inference method, we also performed the same analyses using RAxML with rapid boot- strap[17] (options: -f a -m PROTGAMMA -c 6 -T 10 -p $RANDOM -x $RANDOM -#1000). The FBP and TBE supports for the (73×2) reference trees were computed using Booster (command-line version written in C). All trees were drawn using iTOL and are available in the Booster GitHub repository, along with the sequence alignments. To assess whether rogue taxa removal improves FBP supports, we ran RAxML rogue-detection tool[24] (options: -J MR_DROP -z bootstrap_trees -m PROTGAMMAWAG -c 6 -T 4) and recomputed FBP supports without the detected taxa.

The FastTree and RAxML complete tree topologies were compared to the NCBI taxonomy (https://www.ncbi.nlm.nih.gov/taxonomy), which was converted to Newick format and reduced to the 1,444 taxa common to both our alignment and the NCBI taxonomy. This NCBI tree is not fully resolved and summarizes common belief about the evolutionary history of mammals, resulting from a number of phylogenetic studies based on numerous markers. The unresolved part of the NCBI tree (~4.35 descendants per node on average, instead of 2 for a fully resolved tree) corresponds to the unknown or uncertain part of that history. To cope with uncertainty, we used quartets to compare the (fully resolved) inferred trees to the NCBI tree. A quartet is a tree topology with four taxa; $AB|CD$ is the standard notation for quartets, indicating that taxa A and B form a cherry separated by an internal branch from the cherry formed by C and D; a quartet is unresolved when the four taxa are connected to a single central node. A bipartition b induces a quartet $AB|CD$ when A and B belong to the same side of b, and C and D to the other side. We used tqDist[42] to count the number of quartets induced by

the reference branches, which appeared to contradict the quartets induced by the NCBI tree and its bipartitions; for example, *AB|CD* was found in the studied branch, whereas *AC|BD* was found in the NCBI tree. Unresolved quartets of the NCBI tree were not counted as contradictory, as they represent an unknown evolutionary truth and the inferred resolution could be correct. Such an approach would be difficult to implement with the transfer distance. The number of contradicted quartets was divided by the total number of quartets induced by the studied branch, to obtain a normalized measurement in the [0,1] range (0: no contradiction; 1: all induced quartets are contradicted). We used the same approach to check the accuracy of the FastTree and RAxML tree topologies, comparing the whole set of quartets induced by the inferred tree to those induced by the NCBI tree.

HIV dataset and analyses. From the HIV database (https://www.hiv.lanl.gov/content/index) we retrieved *pol* sequences of the nine 'pure' subtypes of HIV-1 group M, corresponding to positions 2258–3300 relative to the HXB2 reference strain (accessed September 2014). The 'one sequence per patient' option was used and we randomly selected samples of the over-sampled subtypes (A1, B, C, D and G), resulting in a final dataset of 9,147 sequences. These sequences are annotated as 'pure' (that is, non-recombinant) in the database, using a fast filtering approach. However, 48 recombinant sequences were still detected using the standalone version of jpHMM[36] (version March 2015; options: -v HIV, with default input and priors). These 48 sequences were kept in the analyses to study the effect of recombinant sequences, as their presence is inevitable in any HIV dataset. jpHMM was also used to annotate the whole set of sequences depending on their subtype or recombinant status.

Sequences were aligned using MAFFT[43] (version 7.0; default parameters) along with the HXB2 reference strain. Codon positions associated with major drug resistance mutations were removed before tree inference, resulting in an alignment of 1,043 DNA sites (R source code available at https://github.com/olli0601/big. phylo). This alignment was subsampled to study the effect of tree size. We randomly drew 16 samples with 1/16th of the sequences (that is, 571), and 256 samples with 1/256th of the sequences (that is, 35). Then, we generated 1,000 bootstrap alignments for the full alignment and each of the 272 subsampled alignments, by drawing sites with replacement. We used FastTree[34] (options: -nopr -nosupport -gtr -nt -gamma) to infer trees from each of these reference (1 + 16 + 256 = 273) and bootstrap (273,000) alignments. The FBP and TBE supports for the 273 reference trees were computed using Booster (command-line version written in C). All trees were drawn using iTOL (http://itol.embl.de/) and are available on the Booster-workflows GitHub repository, along with the sequence alignments. The instability score was computed considering the reference branches with TBE > 70% (the signal becomes noisy when incorporating branches with lower supports in the calculation, as these branches may be erroneous and thus non-informative about taxon stability). For every taxon, the instability score is equal to the average number of times it has to be transferred to recover these branches from the bootstrap trees, divided by the number of these branches.

The most representative clades for each of the subtypes in the reference trees (Fig. 1 and Extended Data Fig. 5) were obtained by minimizing the transfer distance. For example, in Fig. 1 with the full dataset, we obtained a clade very close to subtype B, with 3,559 taxa, 2 wrong taxa (that is, non-B), and all (3,557 taxa) B taxa included, resulting in the values 3,559, 2 wrong (w) and 0 missing (m) shown in this figure.

A similar approach was used for the regional variants of subtype C, which is responsible for approximately 50% of the HIV-1 infections in the world. Three monophyletic variants of subtype C have been identified by phylogenetic analysis in East Africa[44], South America[45] and India[46]. Furthermore, the South American epidemic was shown to originate in the East African cluster[44]. To identify these variants in the inferred trees (Fig. 1 and Extended Data Fig. 5) we again used the transfer distance. Following previous publications[44–46], we extracted three groups of C sequences from the whole dataset, based on their geographic origins: East Africa (EA: 440 sequences, originating from Burundi (288), Djibouti (1), Ethiopia (9), Kenya (41), Somalia (1), Sudan (11), Tanzania (78) and Uganda (11)), India (IND: 154 sequences, originating from India (133), Nepal (13) and Myanmar (8)) and South America (SA: 14 sequences, originating from Brazil (12), Uruguay (1) and Argentina (1)). We then searched for the tree clades that were closer to these three sets of sequences. The South American sequences were not accounted for in transfer distance computations when searching for the East African clade, as they originate from East Africa. Moreover, we checked that no neighbouring, nearly optimal clade was supported by FBP. In all three cases, we found clades closely related to the sequence sets. As expected, the South American clade was included in the East African clade. The features of these clades are displayed in Fig. 1 and Extended Data Fig. 5. The fractions correspond to the number of studied sequences included in these clades, versus the total number of such sequences in the whole dataset (for example, 360 East African sequences in the East African clade in Fig. 1, among 440 in the whole tree). The 'wrong' sequences were expected in most cases. For example, the Indian clade (167 sequences, 143 from IND among 154 in the whole tree) contains 19 sequences from China corresponding to the spread of the virus in Asia via heroin trafficking routes[46].

Simulated data and analyses. The aim of our simulation experiments was to check that the results observed with the mammal and HIV-1 datasets are reproducible and quantifiable when the simulation conditions and correct tree are known, notably regarding the support of poor branches and the ability to detect rogue taxa. Simulated data mimicked the mammal dataset. We used the tree inferred by PhyML[47] (options: -b 0 -m WAG -a e -t e -o tlr -d aa) from the full COI-5P protein alignment with 1,449 taxa. Protein sequences were evolved along this tree using INDELible[48], which was launched with options and parameter values derived from the PhyML analysis, and similar to previously conducted experiments[24] to assess the accuracy of rogue-taxon detection. The length of the root sequence was 250 AAs; the substitution model was WAG; amino acid frequencies were estimated from the COI-5P alignment; the rates across-sites model used 4 gamma categories with 'alpha' = 0.441

and no invariant sites; and the indel model used 'power law', 'parameter' = 1.5, 'indel max size' = 5, and 'indel rate' = 0.02.

In this manner, we obtained a first 'non-noisy' MSA of length ~500 with ~50% gaps. Noise was added to this MSA to mimic rogue taxa and homoplasy. We shuffled the amino acids vertically for 50% of the sites (MSA columns), thus making these sites homoplasic. For 5% of the sequences (MSA rows), 25% additional sites were shuffled vertically, thus making these sequences unstable and 'rogue', as they contained half of the phylogenetic signal compared to the other (95%) sequences. Both noisy and non-noisy MSAs were used to compare FBP and TBE. To measure the effect of tree size, both MSAs (comprising 1,449 sequences) were sampled to obtain 8 MSAs with 181 sequences (~1/8 of the full sequence set) and 64 MSAs with 22 sequences (~1/64 of the full sequence set). For each of these reference MSAs we sampled with replacement 1,000 pseudo-alignments to compare the two bootstrap methods. All trees were inferred using FastTree (options: -nopr –nosupport -wag -gamma). Just as with the mammal dataset, for each of the branches in the reference trees we computed the percentage of quartet-based conflicts with the correct (PhyML) tree used to generate the data. We also computed the instability score of all taxa in the complete noisy MSA, using only the branches with TBE > 70%.

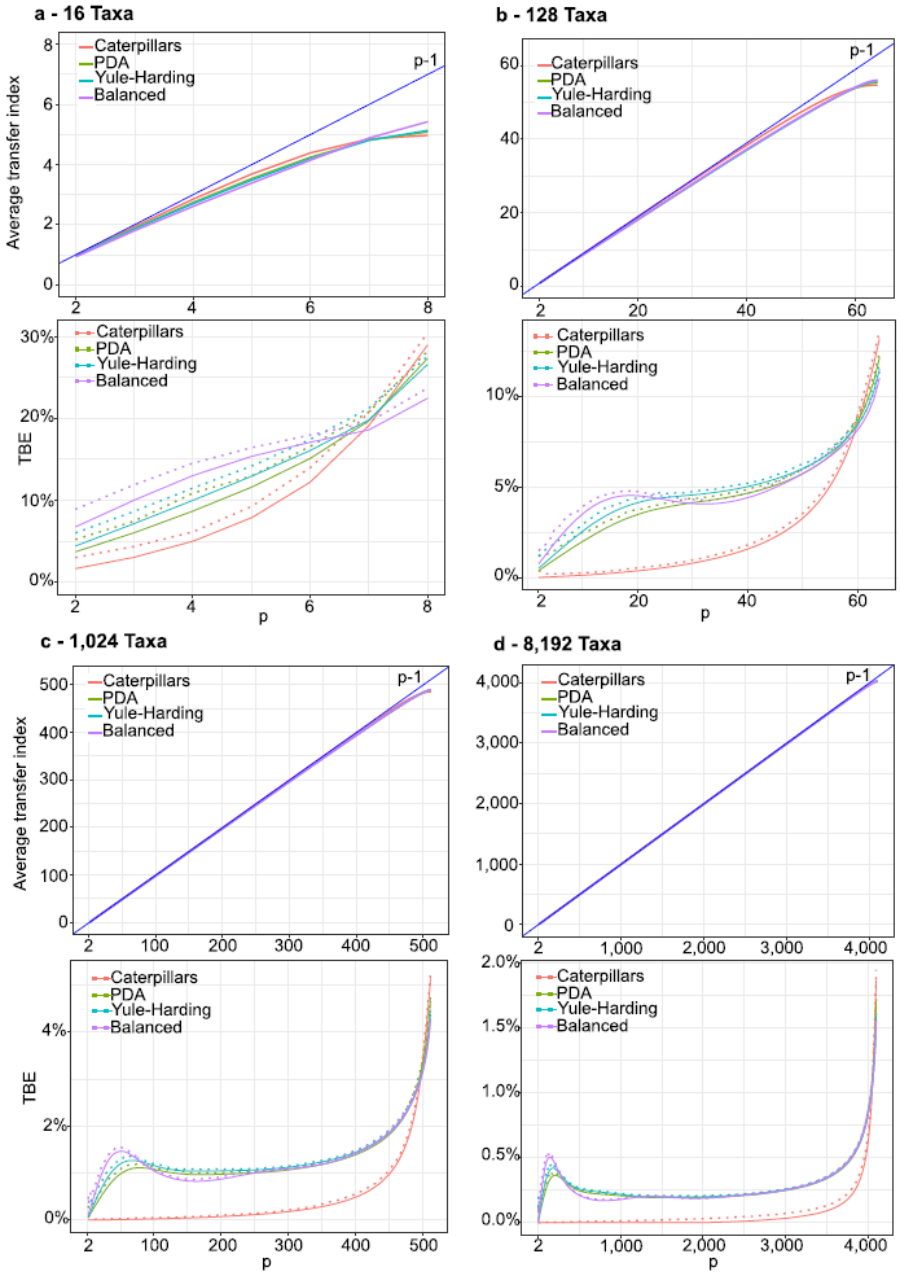
To check the repeatability of FBP and TBE, we generated 1,000 noisy MSAs using the same phylogenetic tree, simulation procedure and set of rogue taxa as the reference noisy MSA (1,449 sequences, ~500 sites and ~50% gaps). We then compared the branch supports of the inferred branches computed using the pseudo-alignments to those obtained using the simulated MSAs. The bootstrap theory[2] indicates that both types of supports are close when the sample size is large enough. The goal was thus to check that 500 sites are enough to obtain a good approximation, and that the bootstrap-based and simulation-based supports are clearly correlated (Pearson's and Spearman's coefficients), as well as the instability score (again computed using branches with TBE > 70%). This experiment was performed with FBP and TBE, with both FastTree (options: -nopr –nosupport -wag -gamma) and RAxML (options: -f d -m PROTGAMMAWAG -c 6). Lastly, the same experiment was used to check the validity of the plug-in principle: we compared the FBP and TBE supports of every inferred branch (both FastTree and RAxML) to the presence or absence (1/0) of that branch in the true tree (FBP), and the normalized transfer distance between that branch and the true tree (TBE).

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

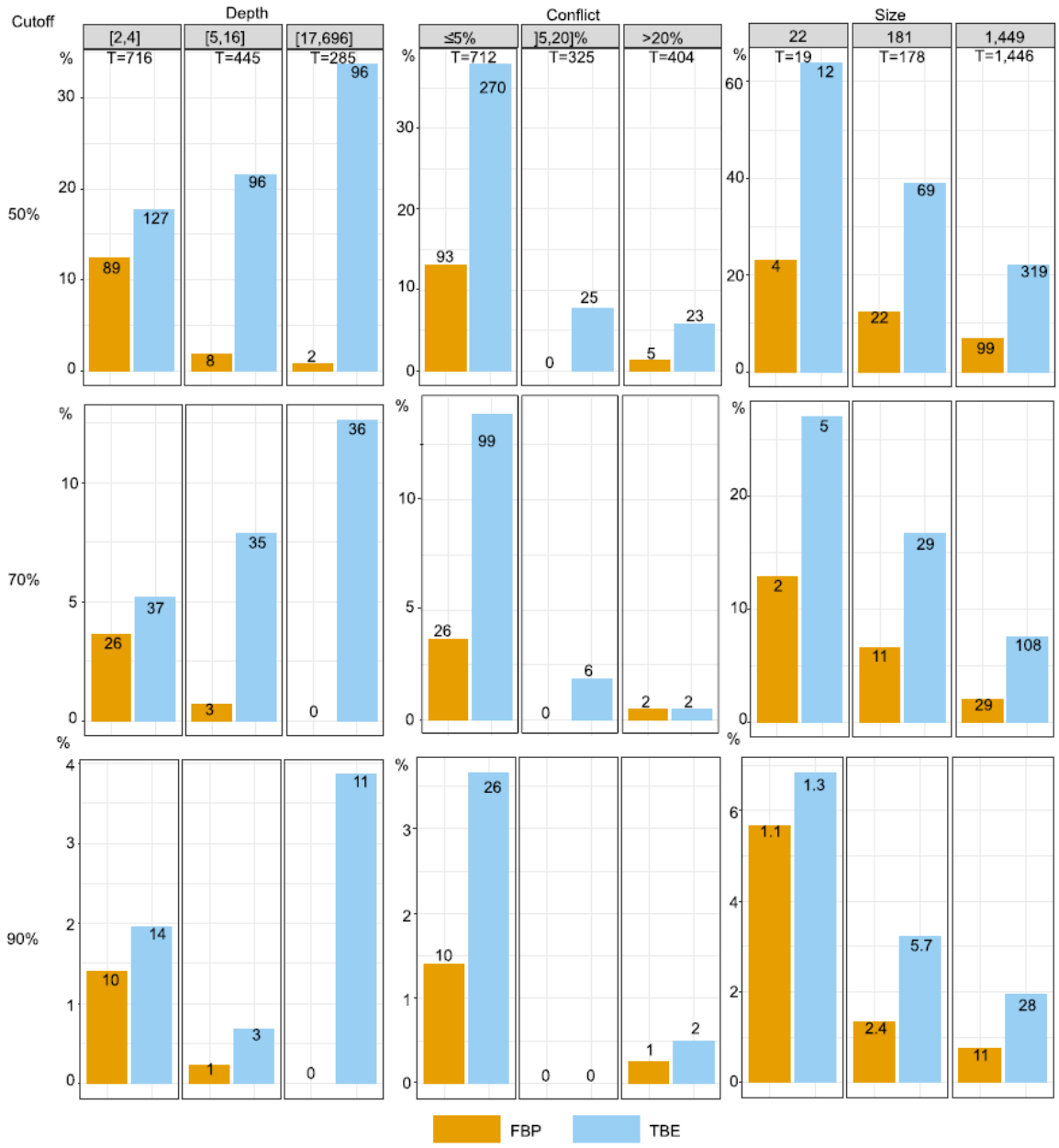
Code availability. Our web interface and software programs are available from Booster website (http://booster.c3bi.pasteur.fr) and GitHub (https://github.com/evolbioinfo/booster). The transfer bootstrap is available in several phylogenetic programs, including PhyML, SeaView, RAxML-NG and others (see http://booster. c3bi.pasteur.fr).

Data availability. All our multiple alignments, phylogenetic trees and workflows are available as Source Data. This material is also available from Booster website (http://booster.c3bi.pasteur.fr). All other data are available from the corresponding author upon reasonable request.

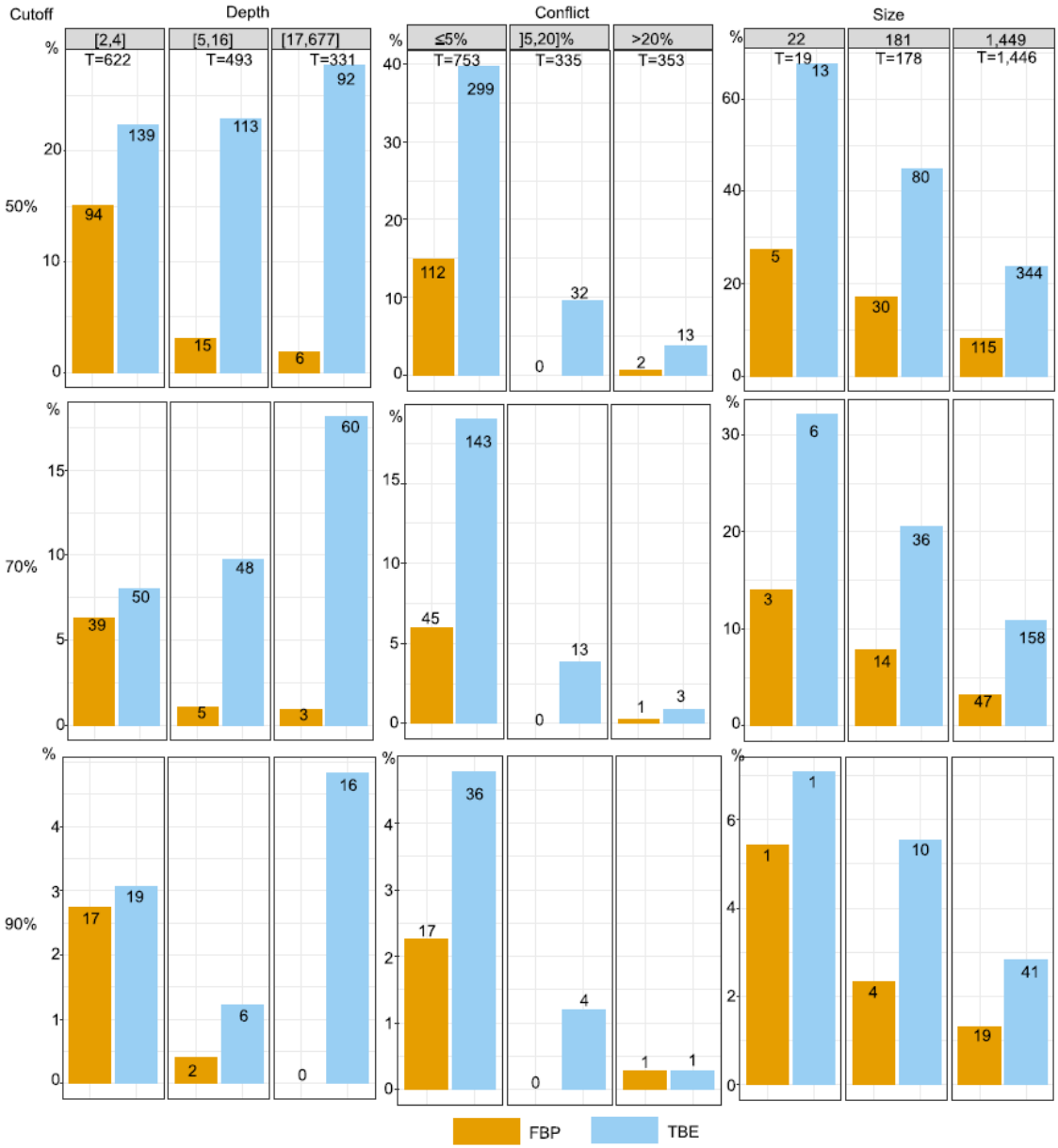
1. Robinson, D. F. & Foulds, L. R. Comparison of phylogenetic trees. *Math. Biosci.* **53**, 131–147 (1981).
2. Semple, C. & Steel, M. A. *Phylogenetics* (Oxford Univ. Press, Oxford, 2003).
3. Lefort, V., Longueville, J. E. & Gascuel, O. SMS: smart model selection in PhyML. *Mol. Biol. Evol.* **34**, 2422–2424 (2017).
4. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44**, W242–W245 (2016).
5. Di Tommaso, P. et al. Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319 (2017).
6. Sand, A. et al. tqDist: a library for computing the quartet and triplet distances between binary or general trees. *Bioinformatics* **30**, 2079–2080 (2014).
7. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
8. Delatorre, E. O. & Bello, G. Phylodynamics of HIV-1 subtype C epidemic in east Africa. *PLoS ONE* **7**, e41904 (2012).
9. Soares, M. A. et al. A specific subtype C of human immunodeficiency virus type 1 circulates in Brazil. *AIDS* **17**, 11–21 (2003).
10. Siddappa, N. B. et al. Identification of subtype C human immunodeficiency virus type 1 by subtype-specific PCR and its use in the characterization of viruses circulating in the southern parts of India. *J. Clin. Microbiol.* **42**, 2742–2751 (2004).
11. Guindon, S. et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
12. Fletcher, W. & Yang, Z. INDELible: a flexible simulator of biological sequence evolution. *Mol. Biol. Evol.* **26**, 1879–1888 (2009)



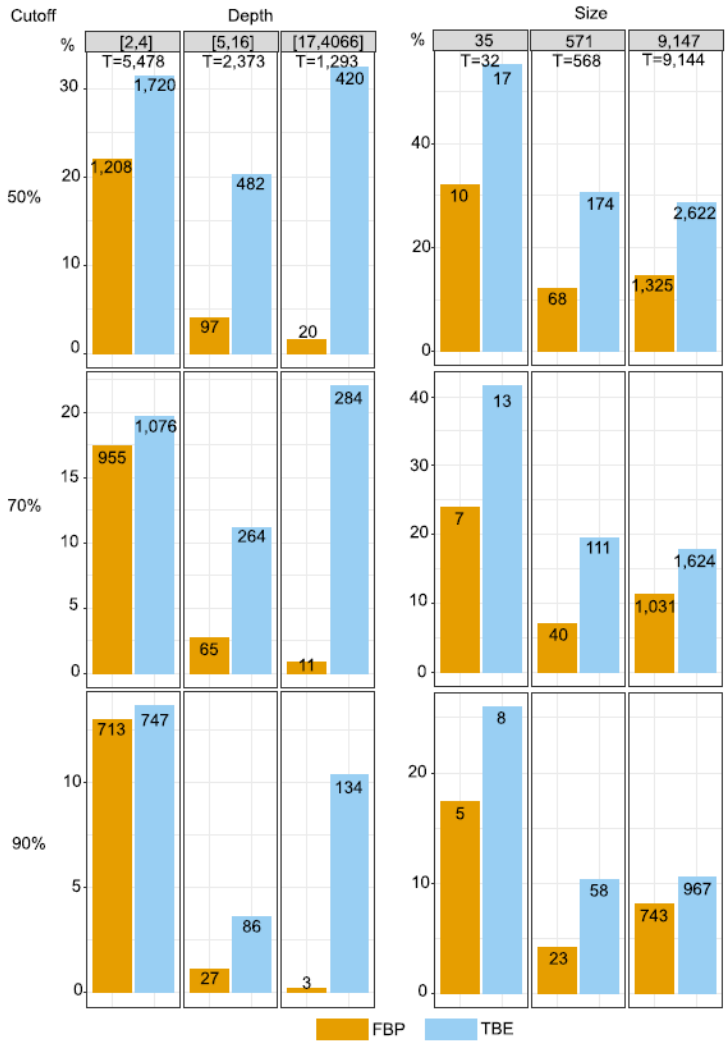
Extended Data Fig. 1 | Transfer index expectation and TBE support with random trees. a–d, For each number of taxa (16, 128, 1024 and 8,191 in **a**, **b**, **c** and **d**, respectively) and random tree model, we compare the transfer index average over 100 runs with the upper-bound $p - 1$ (top graphs in each panel). We also compare the average transfer bootstrap support (TBE) to 0, and provide the maximum value observed among 100 runs (dashed lines), thus approximating the 1% quantile of the distribution (bottom graphs). In these experiments, the number of random 'bootstrap' trees is equal to 1,000. With $l \geq 1{,}024$ (**c**), the average transfer index with random trees is very close in relative value to the upper-bound $p - 1$ and the approximation is already satisfying with $l = 128$ (**b**). Furthermore, the results are nearly the same for the four random tree models, suggesting that the asymptotic behaviour holds in a number of settings. As expected, the approximation of the transfer index over random bootstrap trees by $p - 1$ is better with small values of p. These results explain why moderate TBE supports—for example, 70% as used in this article—are sufficient to reject poor branches, as a TBE branch support of 70% cannot be observed by chance, even with a small number of taxa (for example, 16, as in **a**).



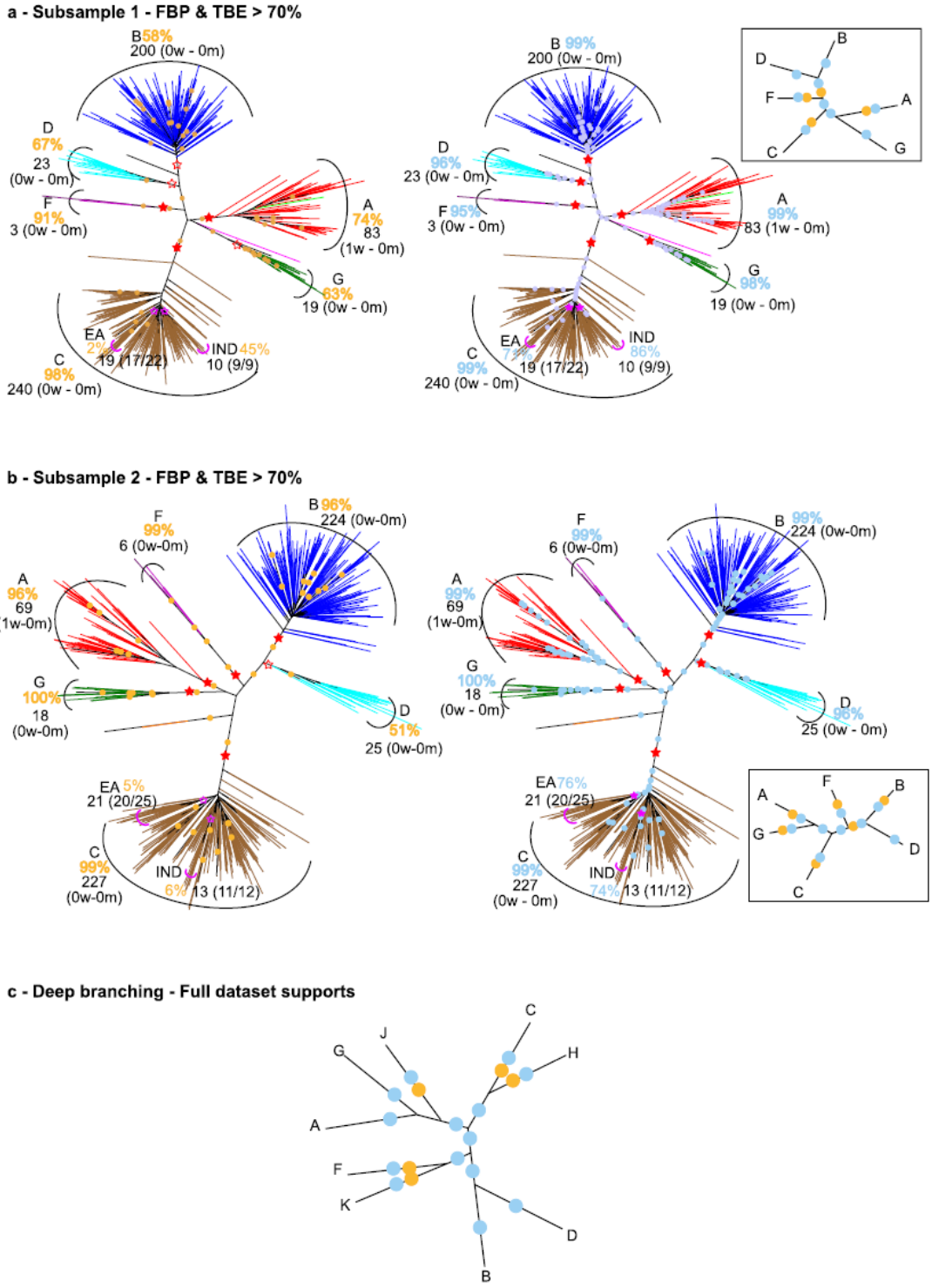
Extended Data Fig. 2 | Comparison of FBP and TBE using the mammal dataset and FastTree phylogeny. FBP and TBE supports are compared with respect to branch depth, quartet conflicts with the NCBI taxonomy and tree size (see main text and legends of Figs. 1, 2 for explanations). Three support cut-offs are used to select the branches: 50%, 70% and 90% (for example, 28 branches among the 1,446 in total have TBE ≥ 90% and 11 have FBP ≥ 90%). The FastTree topology is poor, with 38% of quartets contradicted by the NCBI taxonomy, and 404 of the 1,441 branches with contradictions above 20%. Despite this difficulty, FBP and TBE perform well: they give supports larger than 70% to a very low number of moderately ((5,20]%) and highly (> 20%) conflictual branches. FBP supports very few deep branches, whereas TBE supports a larger number of branches and is especially useful with large trees. Comparing the three cut-offs, we see that with a 50% cut-off the selected branches are still weakly contradicted, especially with FBP; as expected, with TBE the fraction of contradicted branches (> 5%) is a bit higher but still low (~7%). With a cut-off of 90% very few branches are selected (~2% with TBE), thus justifying the use of the 70% threshold for TBE—as is standard with FBP.



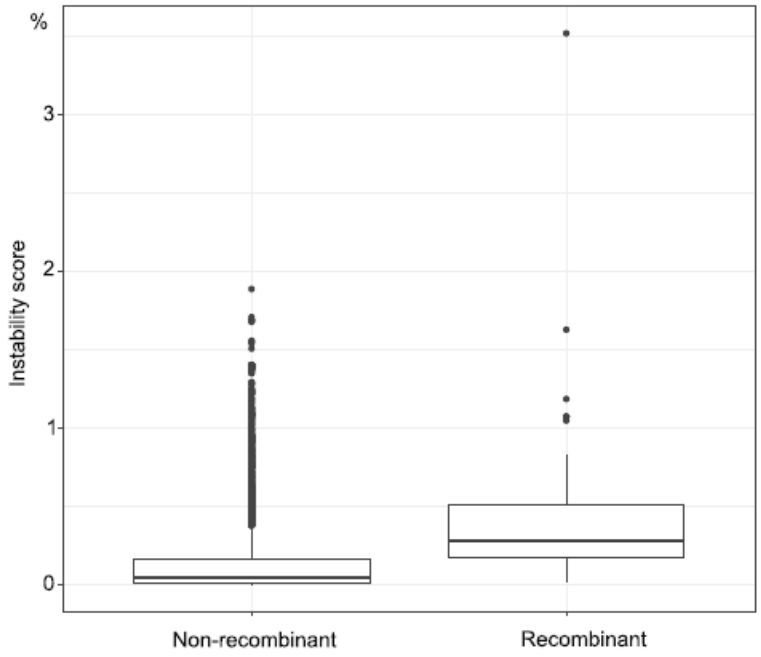
Extended Data Fig. 3 | Comparison of FBP and TBE using the mammal dataset and the phylogeny inferred by RAxML with rapid bootstrap. FBP and TBE supports are compared with respect to branch depth, quartet conflicts with the NCBI taxonomy and tree size (see main text and legends of Figs. 1, 2 for explanations). Three support cut-offs are used to select the branches: 50%, 70% and 90% (for example, 41 branches among the 1,446 in total have TBE $\geq$ 90% and 19 have FBP $\geq$ 90%). The RAxML topology is closer to the NCBI taxonomy than is the FastTree topology (27% versus 38% of contradicted quartets, and 353 versus 404 branches with contradiction > 20%, respectively). However, the RAxML topology is still relatively poor, as expected in this type of phylogenetic study based on a unique marker (Fig. 4 and main text). Despite this difficulty, FBP and TBE perform well as they give supports larger than 70% to a very low number of moderately ((5,20]%) and highly (> 20%) conflictual branches. The supports obtained with RAxML are higher than those obtained with FastTree (47 versus 29 branches with FBP > 70% for RAxML and FastTree, respectively; 158 versus 108 branches with TBE > 70% for RAxML and FastTree, respectively). Part of the explanation could be that the RAxML tree is more accurate than that of FastTree, and is thus better supported. Another factor is that the rapid bootstrap tends to be more supportive than the standard procedure, as shown in previous publications[16]. Indeed, the rapid bootstrap uses already inferred trees to initiate tree searching, and therefore tends to produce less diverse bootstrap trees than the standard, slower procedure, which restarts tree searching from the very beginning for each replicate. Despite these differences between FastTree and RAxML with rapid bootstrap, similar conclusions are drawn when comparing FBP and TBE: FBP supports very few deep branches, whereas TBE supports a larger number of them; TBE is especially useful with large trees; and both methods support a very low number of contradicted branches. Comparing the support cut-offs, 70% again appears as a good compromise for both FBP and TBE.



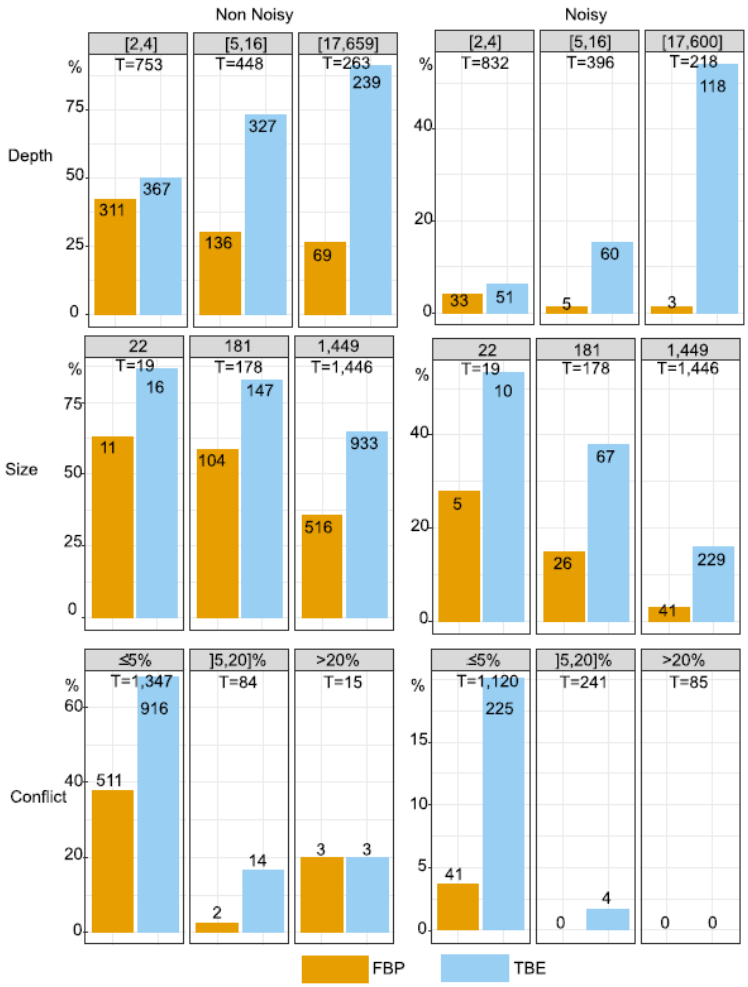
Extended Data Fig. 4 | Comparison of FBP and TBE using the HIV dataset and FastTree phylogeny. FBP and TBE supports are compared with respect to branch depth, and tree size (see main text and legends of Figs. 1, 2 for explanations). Three support cut-offs are used to select the branches: 50%, 70% and 90% (for example, 1,624 branches among the 9,144 in total have TBE > 70% and 1,031 have FBP > 70%). Results are for the most part similar to those observed with the mammal dataset. We see a major effect of depth on FBP supports: with the full dataset, less than 1% of the deep ($p > 16$) branches have FBP support larger than 70%, whereas this percentage is higher than 20% with TBE. The effect of tree size is less pronounced. The fraction of supported branches decreases when the tree size increases from 35 to 571 taxa, but is analogous between 571 and 9,147 taxa. Furthermore, the gap between FBP and TBE remains similar, probably owing to the very large number of cherries and small clades, for which TBE and FBP are nearly equivalent. Regarding the support cut-off, 70% again appears as a good compromise for TBE, though there is no way to evaluate the fraction of supported branches that is actually erroneous. The interpretability of TBE will be a major asset for choosing the support level depending on the phylogenetic question being addressed. Here, as recombinant sequences are inevitable, lower supports than with mammals are likely to be acceptable.



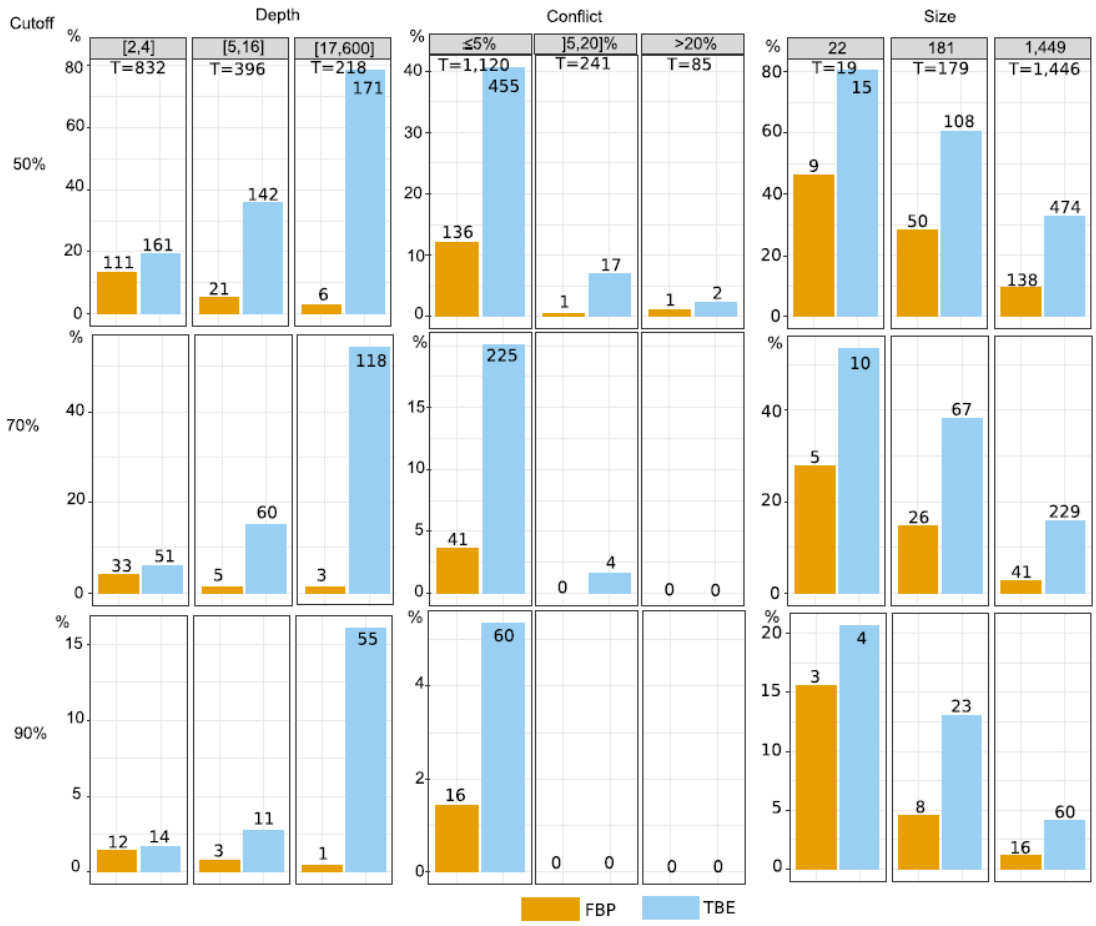
Extended Data Fig. 5 | Subtype deep branching and comparison of FBP and TBE using medium-sized HIV datasets. As the taxa were randomly drawn from the full dataset, the supports and findings show some fluctuations. **a**, **b**, Trees obtained with two of the medium-sized datasets; branches with FBP > 70%: yellow dots; branches with TBE > 70%: blue dots; subtype clades: red stars, filled if support > 70% (see Methods and Fig. 1 legend for further details). **c**, Deep branching of the subtypes[19] and supports obtained on the full dataset (see also Fig. 1). Rare subtypes (H, J and K) are absent in the medium-sized datasets, and the subtype clades are almost perfectly recovered (only one incorrect taxon in A clade for both trees). FBP supports are higher when using medium-sized datasets than when using the full dataset (for example, 58% and 99% for subtype B, versus 3% in Fig. 1). However, some subtype clades (for example, D) have moderate FBP support, though the clade matches the subtype perfectly. When using TBE, all subtype supports are higher than 95%. The deep branching is the same for all full and medium-sized datasets, and is identical to that found in a previous study[19], but is not supported by FBP, whereas TBE is larger than 70% for every branch (or path in Fig. 1). Again, the Indian and East African sub-epidemics of subtype C are supported by TBE, but not by FBP.



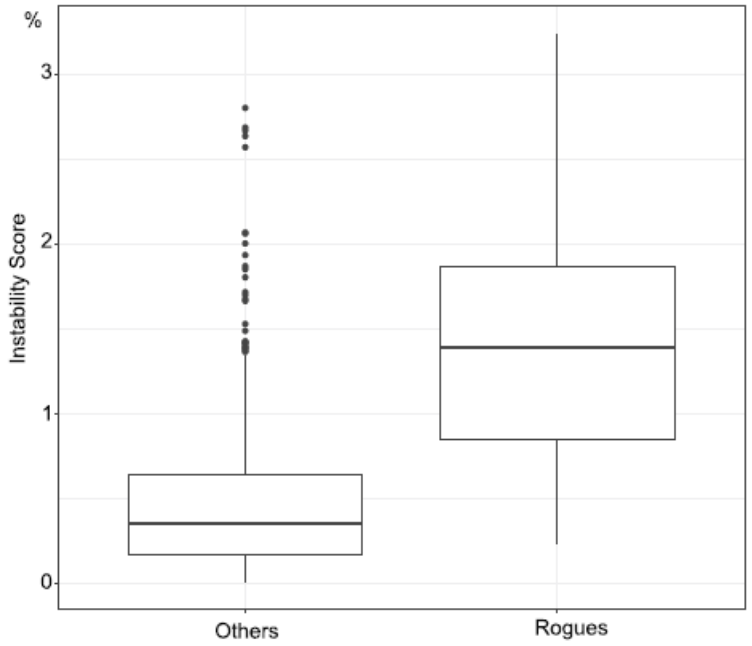
Extended Data Fig. 6 | Distribution of the instability score in HIV recombinants. We see a clear difference between the distributions of the instability score for the recombinant and non-recombinant sequences, which means that this score can be used to detect or confirm the recombinant status of sequences (box quantiles: 25%, 50% and 75%). See main text for details.



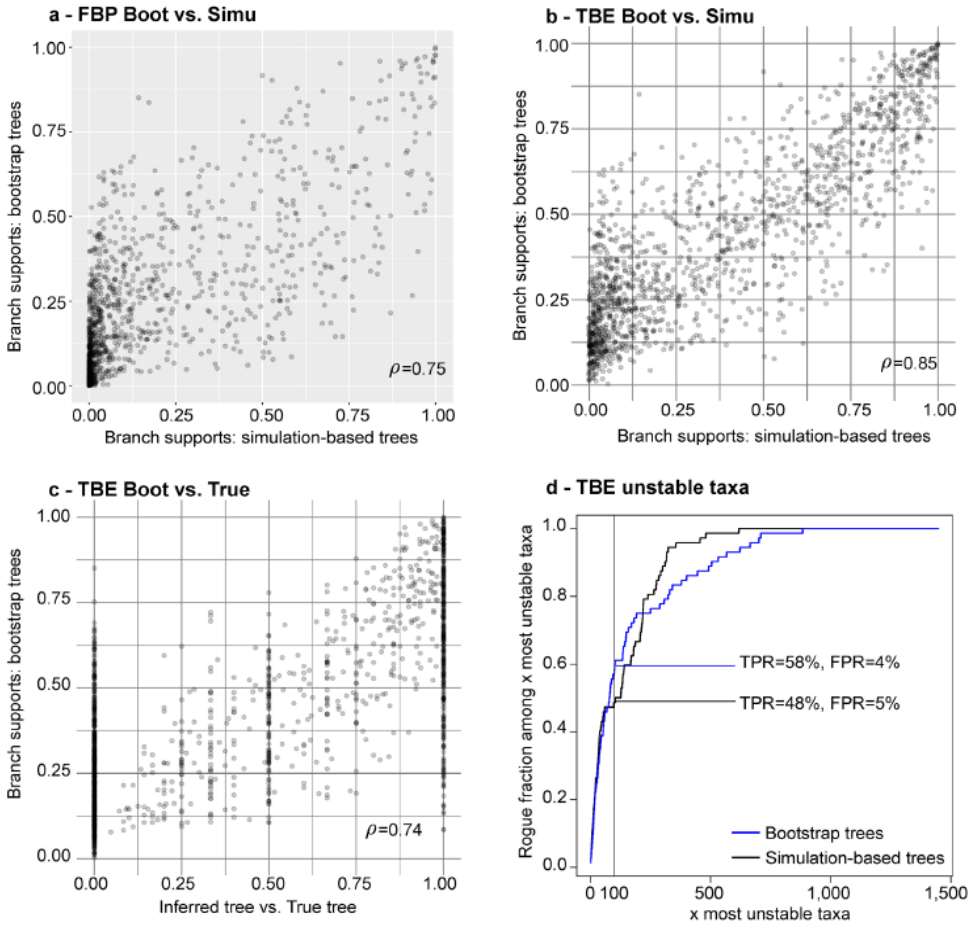
Extended Data Fig. 7 | Comparison of FBP and TBE using non-noisy and noisy simulated data. Noisy data include rogue taxa and homoplasy and non-noisy data do not (see Methods for details). The graphs display the distribution of branches with FBP or TBE support > 70%. Supports are compared regarding branch depth, tree size and quartet conflicts with the model tree used for simulations (see main text and legends of Figs. 1, 2 for explanations). Results are fully congruent with those obtained with real datasets. TBE supports more deep branches than FBP, especially with noisy data. The effect of tree size is also more visible with noisy MSA, and the number of supported branches with moderate ((5,20]%) and high (> 20%) conflict levels is very low, for both FBP and TBE.



Extended Data Fig. 8 | Comparison of FBP and TBE at different support cut-offs using simulated, noisy data. Comparison of FBP and TBE with respect to branch depth, quartet conflicts and tree size, at different support cut-offs (see main text and legends of Figs. 1, 2 for explanations). A cut-off of 50% seems to be acceptable, as neither FBP nor TBE support highly contradicted branches. However, this could be due to the low level of contradiction compared to real datasets (85 branches with contradiction > 20%, versus about 400 in the mammal dataset in Extended Data Figs. 2, 3).



Extended Data Fig. 9 | Distribution of the instability score in rogue taxa using simulated, noisy data. TBE again appears to be useful for detecting and confirming rogue taxa (box quantiles: 25%, 50% and 75%). See main text for details.



e - Correlations: Pearson(Spearman)

		FBP		TBE		
	%Correct	Simu/Boot	True/Boot	Simu/Boot	True/Boot	%Rogues
RAxML	322/1,446	0.75(0.83)	0.59(0.54)	0.85(0.83)	0.74(0.70)	58
FastTree	261/1,446	0.72(0.82)	0.54(0.48)	0.85(0.82)	0.75(0.72)	60

Extended Data Fig. 10 | Repeatability and accuracy of FBP and TBE using simulated data. The bootstrap theory[1,2] indicates that with large samples the supports estimated using bootstrap replicates should be close to supports obtained with datasets of the same size drawn from the same distribution as the original sample. We used simulated data to check that this property holds with protein MSAs of 1,449 taxa and about 500 sites (see main text for details). **a**, **b**, Comparison of the two supports (**a**, FBP; **b**, TBE) for all branches in the tree inferred by RAxML from the original MSA. We observe a clear correlation, which is higher for TBE ($\rho = 0.85$) than for FBP ($\rho = 0.75$) using Pearson's linear correlation coefficient, but identical (0.83) using Spearman's rank coefficient, which is better suited to the discontinuous nature of FBP. These results appear to contradict previous conclusions[7] that the bootstrap is a highly imprecise measure of repeatability. However, this previous work measured the probability of inferring the correct tree (not the supports of inferred branches, as consistent in the bootstrap context) and its main result was based on 50 sites, which is probably too low for the bootstrap theory to apply. The bootstrap also relies on the plug-in principle[2,3,6,9], which states that the distribution of the distance between the true tree and the inferred tree can be well-approximated by the distribution of the distance between the inferred and bootstrap trees. **c**, The accuracy of TBE in predicting the topological distance between b and the true tree as measured using the normalized transfer index, for every branch b inferred by RAxML from the original MSA. Again, we observe a clear correlation ($\rho = 0.74$, Spearman's rank coefficient = 0.70). We performed the same experiment with FBP, seeking to predict the presence or absence (1/0) of the inferred branch in the tree true; a lower but significant correlation was found ($\rho = 0.59$, Spearman's rank coefficient = 0.54). **d**, Comparison using RAxML of the performance of simulation-based and bootstrap-based instability scores in detecting rogue taxa; both are nearly identical. TPR, true positive rate; FPR, false positive rate. **e**, Table summarizing the results described above and those of FastTree, which are nearly identical to those of RAxML, except regarding topological accuracy (%Correct, fraction of correct branches) for which RAxML is again more accurate than FastTree.