Virus Evolution, 2024, **10(1)**, veae081

DOI: https://doi.org/10.1093/ve/veae081
Advance access publication date: 15 October 2024
Reflections

On the importance of assessing topological convergence in Bayesian phylogenetic inference

Marius Brusselmans[1,*], Luiz Max Carvalho[2], Samuel L. Hong[1], Jiansi Gao[3], Frederick A. Matsen IV[4,5,6], Andrew Rambaut[7], Philippe Lemey[1], Marc A. Suchard[8], Gytis Dudas[9], Guy Baele[1]

[1]Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven, Leuven, Belgium
[2]School of Applied Mathematics, Getulio Vargas Foundation, Praia de Botafogo, 190, 22250-900 Rio de Janeiro, Brazil
[3]Computational Biology Program, Fred Hutchinson Cancer Center, Seattle, WA 98109, United States
[4]Howard Hughes Medical Institute, Computational Biology Program, Fred Hutchinson Cancer Research Center,Seattle, Washington, United States
[5]Department of Genome Sciences, University of Washington, Seattle, Washington, United States
[6]Department of Statistics, University of Washington, Seattle, Washington, United States
[7]Institute of Ecology and Evolution, University of Edinburgh, Edinburgh EH9 3FL, United Kingdom
[8]Department of Biostatistics, Fielding School of Public Health, University of California, Los Angeles, CA 90095, USA
[9]Institute of Biotechnology, Life Sciences Center, Vilnius University, Vilnius, Lithuania

*Corresponding author. Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven, Leuven, Belgium.
E-mail: marius.brusselmans@kuleuven.be

Abstract

Modern phylogenetics research is often performed within a Bayesian framework, using sampling algorithms such as Markov chain Monte Carlo (MCMC) to approximate the posterior distribution. These algorithms require careful evaluation of the quality of the generated samples. Within the field of phylogenetics, one frequently adopted diagnostic approach is to evaluate the *effective sample size* and to investigate trace graphs of the sampled parameters. A major limitation of these approaches is that they are developed for continuous parameters and therefore incompatible with a crucial parameter in these inferences: the *tree topology*. Several recent advancements have aimed at extending these diagnostics to topological space. In this reflection paper, we present two case studies—one on Ebola virus and one on HIV—illustrating how these topological diagnostics can contain information not found in standard diagnostics, and how decisions regarding which of these diagnostics to compute can impact inferences regarding MCMC convergence and mixing. Our results show the importance of running multiple replicate analyses and of carefully assessing topological convergence using the output of these replicate analyses. To this end, we illustrate different ways of assessing and visualizing the topological convergence of these replicates. Given the major importance of detecting convergence and mixing issues in Bayesian phylogenetic analyses, the lack of a unified approach to this problem warrants further action, especially now that additional tools are becoming available to researchers.

Keywords: effective sample size, topologies, Bayesian inference, phylogenetics, phylodynamics, convergence, mixing, EBOV, HIV

Introduction

Background

When performing Bayesian phylogenetic inference using Markov chain Monte Carlo (MCMC) algorithms, the standard practice to evaluate the quality of the generated samples is to visually inspect trace plots and to compute the corresponding effective sample size (ESS) of the sampled parameters. This can be done using a variety of software packages such as *Tracer* (Rambaut et al., 2018), *Beastiary* (Wirth and Duchene, 2022) or *CODA* (Plummer et al., 2020), for example. However, such software packages only produce diagnostics for simple, often univariate and continuous, model parameters, which the topology of the phylogenetic tree is not. If these diagnostics suggest the MCMC convergence and mixing of the simple model parameters are satisfactory, it is usually assumed that this will be the case for the topology as well. This is potentially problematic, as the tree topology is often of key interest in phylogenetic and phylodynamic studies, and obtaining a correct (consensus) phylogeny is essential when performing outbreak investigation and monitoring ongoing epidemics [see e.g. Attwood et al. (2022)].

Recent research has focused on convergence and mixing diagnostics applicable to the tree topology, including studies by Lanfear et al. (2016), Magee et al. (2023), and Guimarães Fabreti and Höhna (2021). The former two studies focus on finding ways to apply the principles of trace graphs and effective sample sizes to the topology as a whole, while the latter study considers the presence of each split in the tree as an individual parameter to be evaluated using classical diagnostics. We here explore the main topological methods introduced by Lanfear et al. (2016) and Guimarães Fabreti and Höhna (2021), as well as those from Magee et al. (2023) on the data from an Ebola virus (EBOV) study by Dudas et al. (2017) and an HIV study by Hong et al. (2020). Of note, we also include the multidimensional scaling (MDS) ESS metric—a more conservative tree ESS measure (Magee et al., 2023)—that also enables us to project the high-dimensional phylogenies onto a small number of dimensions suitable for visualization (Kruskal, 1964a; Kruskal, 1964b).

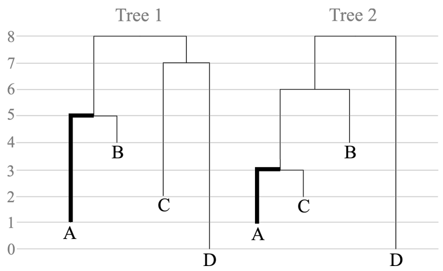
We find that the evaluation of convergence and mixing of samples in topological space can reveal issues not typically detected by standard diagnostics for (convergence and mixing of) continuous parameters. Furthermore, we find that decisions regarding the computation of these diagnostics can impact the conclusions regarding MCMC convergence and mixing. We selected the EBOV study because of its large size and the rich complexity of the models that were applied, making it a prime case study of a challenging phylogenetic analysis that could be susceptible to hidden convergence and mixing issues. Further, HIV phylogenies are known to be star-like (i.e. have short internal but long external branches), which could lead to a different set of issues from a topological perspective.

Phylogenetic distance metrics

A central concept in our exploration is that of phylogenetic distance, a quantitative measure of similarity between two phylogenetic trees. Such a distance can be computed in several ways, using what we will refer to as phylogenetic distance metrics. These distances are zero for two identical trees and are expected to increase as trees grow more dissimilar. The metrics considered in this paper are: the **Robinson–Foulds** distance (Robinson and Foulds, 1981) and its weighted counterpart (Robinson and Foulds, 1979), the **path difference** (Steel and Penny, 1993), the **branch score** (Kuhner and Felsenstein, 1994), the **Kendall–Colijn** distance (Kendall and Colijn, 2016) with $\lambda = 0$ (i.e. disregarding the branch lengths), and the rooted **subtree-prune-regraft (SPR) distance** (Whidden et al., 2013).

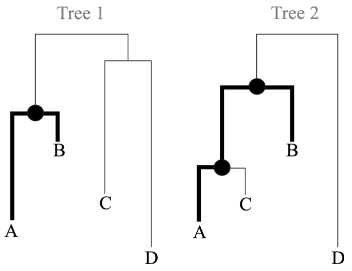
In order to provide the reader with some intuition of what aspects of topological differences these metrics convey, we use a toy example of two phylogenetic trees with the same set of time-calibrated taxa on which an example calculation of each of these metrics is performed.

Fig. 1 shows, for each branch in the tree, the partition defined by that branch. Each of these branches also has an associated branch length in both trees, which is reported as well. The Robinson–Foulds distance is simply the number of partitions present in one tree but not the other, which in this case is 2, since only {AB|CD} and {AC|BD} are such partitions. The weighted Robinson–Foulds distance would be the sum of the absolute differences in the branch lengths of branches defining corresponding partitions, which are shown in the |Δ| column. This distance is thus 17. Closely related is the branch score, which takes the square root of the sum of squares of |Δ|, which in this case equals 7.42.



Partition	Associated branch lengths		
	tree 1	tree 2	|Δ|
{A|BCD}	**4**	**2**	**2**
{B|ACD}	1	2	1
{C|ABD}	5	1	4
{D|ABC}	7	$8 + 2$	3
{AB|CD}	$3 + 1$	*none*	4
{AC|BD}	*none*	3	3

Figure 1. Partitions defined by each branch in the two trees and the associated branch lengths. The branches defining partition {A|BCD} and their associated branch lengths are **bolded** in both trees and the table. |Δ| is the absolute difference of two corresponding branch lengths. Certain partitions do not exist in one of either trees, in which case the branch length is shown as *none* and treated as 0 for the computation of |Δ|. The **(weighted) Robinson–Foulds distance** and the **branch score** can be computed from the information contained in the provided table.



Pair of tips	d_P			$d_{MRCA-root}$		
	tree 1	tree 2	|Δ|	tree 1	tree 2	|Δ|
{A-B}	**1**	**2**	1	1	1	0
{A-C}	2	1	1	0	2	2
{A-D}	2	2	0	0	0	0
{B-C}	2	2	0	0	1	1
{B-D}	2	1	1	0	0	0
{C-D}	1	2	1	1	0	1

Figure 2. Pairwise tip-to-tip path lengths d_P, defined as the number of internal nodes that must be crossed (not counting the root node) to get from one tip to another, as well as absolute difference between these. The path to go from tip *A* to B is **bolded** in both trees, as are the crossed internal nodes. Also shown is the closely related MRCA-to-root path length $d_{MRCA-root}$ for each pair of tips, defined as the number of internal nodes that must be crossed to go from the MRCA of the tips to the root node (MRCA node included). The **path difference** and **Kendall–Colijn** distance can be computed from the information contained in the provided table.

Fig. 2 shows—for both trees—all pairwise tip-to-tip path lengths, which is the number of internal nodes that must be crossed to go from one tip to the other. The path difference is the square root of the sum of squares of the differences in path lengths d_P between the two trees, which in this case equals 2. Fig. 2 also shows the path lengths from the most recent common ancestor (MRCA) of each pair of tips to the root node |Δ|. The Kendall–Colijn distance (with $\lambda = 0$, as considered in this manuscript) is the square root of the sum of squares of these |Δ| values, which in this case equals 2.45.

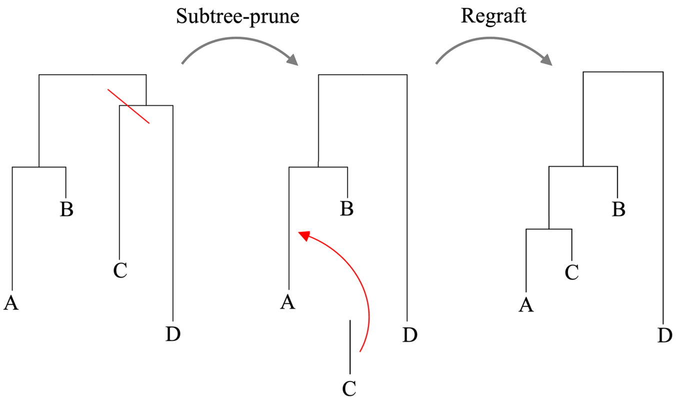


Figure 3. The subtree-prune-regraft (SPR) move required to transform tree 1 into tree 2, disregarding branch lengths. The move involves pruning the branch defining the {C|ABD} partition and regrafting it onto the branch defining the {A|BD} partition. The SPR distance is the number of SPR moves required to transform tree 1 into tree 2, in this case 1.

Thus, these distance metrics can be divided into two general categories. First are metrics defined by partitions and branch lengths (Robinson–Foulds, weighted Robinson–Foulds, branch score) as shown in Fig. 1. Second are metrics defined by path lengths between tips and/or nodes (path difference, Kendall–Colijn).

A last type of distance metric we consider is the SPR distance. SPR refers to a type of operation that can be performed on a tree to change its topology. This is closely related to the nature of the trees considered in this manuscript, since most Bayesian phylogenetic software packages make use of SPR-like moves to explore tree space. Fig. 3 shows the SPR move that would be required to transform tree 1 into tree 2. Since only one such move is needed, the SPR distance between the two trees shown equals 1.

Topological convergence diagnostics

We here briefly discuss the topological convergence diagnostics considered in this manuscript.

The **topology trace plot**, as defined by Lanfear et al. (2016), works very similarly to standard trace plots used for continuous parameters. The value of the trace being graphed is the phylogenetic distance from each sampled tree to a chosen reference tree. As a reference tree, one can choose between, for example, one of the posterior trees or a consensus tree. However, choosing a tree that is part of the chain will cause a 'slump' in the trace plot as the distance from this tree to itself is inevitably equal to zero. It is good practice to try several reference trees and compare the resulting trace plots. In this manuscript, we use the first tree as a reference and exclude it from the graphs to avoid scaling issues.

The **pseudo-ESS** (Lanfear et al., 2016) is the ESS of the vector of phylogenetic distances from an arbitrarily chosen focal tree to all other trees in the sample (note the analogy with the topology trace plot). Because there is randomness involved in the choice of the focal tree, the computation is repeated using each tree in the sample as a focal tree. The lowest value and median value are then reported.

The **approximate ESS** (Lanfear et al., 2016) derives an approximation of the common ESS calculation of dividing the actual sample size by the autocorrelation time, but uses a topological version of autocorrelation time estimated by determining the thinning interval at which the average phylogenetic distance between subsequently sampled trees ceases to increase as the thinning interval increases.

The **Fréchet correlation ESS** (Magee et al., 2023) is defined analogously to the standard one-dimensional continuous ESS, but substitutes Pearson autocorrelation for an alternative definition of autocorrelation between trees using Fréchet (co)variances, which make use of the relationship between covariance and Euclidian distance—here substituted with whatever phylogenetic distance is being used.

The **split frequency ESS** (Magee et al., 2023) is computed by treating each tree as a vector of binary split indicators (split is present/absent). Fréchet (co)variances can then be computed using the Euclidean norm (as the trees are now reduced to vectors of binary indicators), which enable computation of an ESS. Note that this is the only method that does not explicitly use any kind of phylogenetic distance metric.

The **multidimensional scaling ESS** (Magee et al., 2023) (MDS ESS) is computed by performing multidimensional scaling of the matrix of squared pairwise distances between trees. The first dimension of the resulting multidimensional scaling matrix is then used to compute an ESS on.

Materials and methods

Software

We computed the topological ESS estimators described in the previous section using the treess package version 1.0.1 (Magee et al., 2023) with R v4.3.0 (R-team, 2022), the phylogenetic distances required for these ESS estimators using the R packages phangorn v2.11.1 (Schliep et al., 2022) and TreeDist v2.6.1 (Martin et al., 2023), the per-split ESS values with the R package convenience (Guimarães Fabreti and Höhna, 2021), and the approximate subtree-prune-regraft (aSPR) distances using RSPR version 1.3.1 (Whidden et al., 2013). We constructed the tanglegrams using Baltic version 0.2.2, available at https://github.com/evogytis/baltic. Finally, we also made use of *Tracer* v1.7.2 (Rambaut et al., 2018) and the *TreeAnnotator* v1.10.4 tool associated with the BEAST 1.10.4 software package for summarizing maximum clade credibility (MCC) trees (Suchard et al., 2018), as well as the R package ggtree v3.8.2 (Xu et al., 2022) for visualization of these trees.

Data

The first dataset considered in this study is from a genomic epidemiology study of the largest Ebolavirus (EBOV) outbreak to date, which investigated the impact of several potential predictors, such as climate and demographic information, on EBOV spread in West Africa from 2014 to 2015. We refer the interested reader to the original publication for further details (Dudas et al., 2017). In summary, Dudas et al. (2017) performed Bayesian phylogenetic inference using MCMC on a total of 1610 EBOV genome sequences sampled between March 2014 and October 2015 using an HKY + Γ_4 nucleotide substitution model, an uncorrelated relaxed molecular clock model with an underlying lognormal distribution and a flexible non-parametric coalescent model (Gill et al., 2013) as the tree prior. We downloaded the 1000 posterior sample trees, which were sampled every 10 000 iterations, and log-files from the original publication from https://github.com/ebov/space-time/tree/master/Analyses/Phylogenetic. Of note, the burn-in was already discarded from the posterior sample trees file shared by Dudas et al. (2017).

We obtained the second dataset considered from a 2020 phylogeographic study of the spread of HIV-1 subtype B in the USA. As in the EBOV study, the HIV study also aimed to identify relevant covariates for the spread of the virus, but also studied the impact of different subsampling schemes on these inferences. We refer interested readers to the original publication for further details (Hong et al., 2020), from which we selected a data set consisting of 500 sequences that was constructed with the aim of maximizing phylogenetic diversity. The authors used a GTR + Γ_4 nucleotide substitution model, a strict molecular clock model, and a logistic population growth model acting as a tree prior. The output of the original analysis also consists of 1000 posterior sample trees. The burn-in was already discarded from the posterior sample trees file, which is available on https://github.com/hongsamL/HIV_trees.

Assessing topological convergence

The EBOV data

Fig. 4 shows the topology trace plots for the EBOV analysis using the six different topological distance metrics, as well as the corresponding topological ESS estimates. The results in this figure can be grouped in two sets based on the distance metric used: the (weighted) Robinson–Foulds distance, the branch score and the aSPR distance on the one hand, and the path difference and Kendall–Colijn distance on the other. Note that this grouping of metrics is closely related to the conceptual differences between them as shown in Figs 1 and 2.

In the former group of metrics, the combined traces can be clearly divided into three distinct parts at iterations 333 and 666. This suggests that three independent replicate analyses were combined to obtain the posterior sample of trees in the study of Dudas et al. (2017), which has been confirmed by the authors. Bayesian phylogenetic inference on large data sets indeed commonly employs the practice of concatenating the samples of several independent chains to both reduce computation time (by increasing the ESS values of continuous parameters) and assess convergence towards the same posterior. Four of the topology trace plots in Fig. 4 therefore suggest a discrepancy between the posterior space explored by the three chains. Whether this indicates failure to converge to the same posterior, a case of extremely slow/poor mixing, or something else entirely is not clear. The ESS estimates in this group tend to be substantially lower when computed for the entire sample than when computed for the three individual samples, which can be expected if the subsamples explore different parts of tree space. This is not entirely consistent though, as the approximate ESS only shows this behaviour when considering the aSPR distance, and the split-frequency ESS does not show this behaviour at all. It should be remarked that the split frequency ESS is invariant to the choice of phylogenetic distance metric, as it is computed on the vector of splits directly.

In the latter group of metrics—considering the path difference and the Kendall–Colijn distance—the three individual samples are indistinguishable in the topology trace plots. The ESS values are also substantially better across the board and do not decrease when considering the whole concatenated sample as opposed to the individual samples, but instead are higher than the ESS of each individual sample.

The HIV data

Fig. 5 shows the topology trace plots for the HIV analysis using the six different topological distance metrics, as well as the corresponding topological ESS estimators. Similar to the EBOV analysis, the full sample is a concatenated set of two samples [as confirmed by the authors of Hong et al. (2020)], but with each sample containing a different number of trees. However, unlike with the EBOV data, this is not immediately apparent from any of the topology trace plots (of the combined sample). All distance metrics show a clear slump at the beginning, which could be indicative of the discarded burn-in not having been set sufficiently high from a topological perspective (but indeed set sufficiently high from the current standard practice of only assessing the traces of continuous parameters), which suggests part of the sample is from a non-converged part of the analysis. Topological ESS values tend to be substantially lower across the board for this dataset, most often not even reaching the often-used ESS cut-off of 200 for continuous parameters. Further, the difference in behaviour between the different distance metrics as seen in the EBOV data is not apparent here.

Alternative visualizations

The topological trace graphs presented in the previous section require relatively few computational resources to generate, but only provide partial information as they only consider distances from a single reference tree. We present two alternative kinds of visualizations—generated from pairwise distance calculations (meaning all distances between all possible pairs of the 1000 trees in the samples: producing 449 500 distances) to produce a more comprehensive visual mapping of topological distances between sampled trees.

Pairwise heatmaps

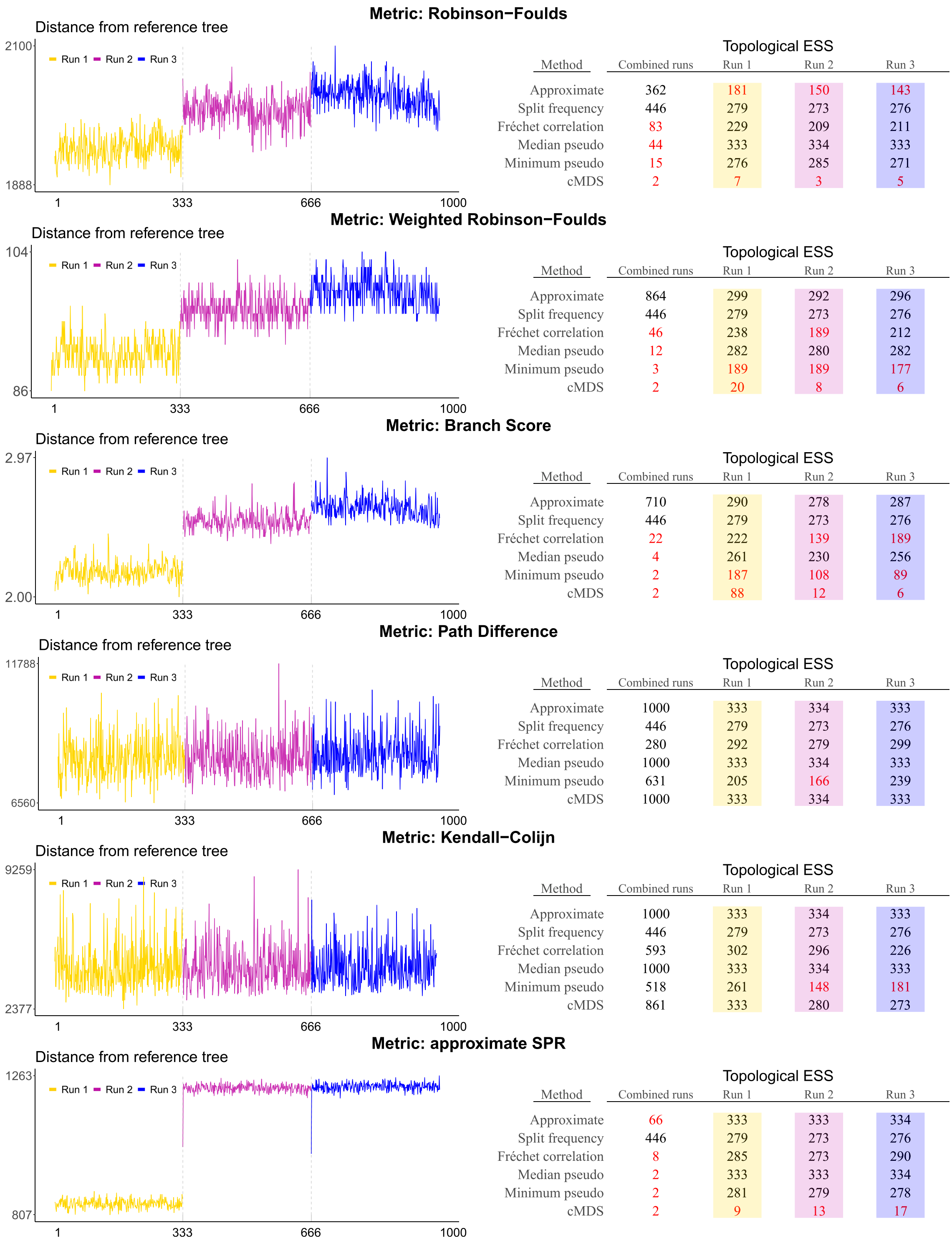
A first alternative visualization comes in the form of heatmaps of the pairwise distances. Fig. 6 shows heatmaps of the Robinson–Foulds distances for the EBOV trees and HIV trees, and of the path distances for the EBOV trees.

For the EBOV posterior tree samples, the patterns seen in the trace graphs of Fig. 4 are also seen in Fig. 6, i.e. distinguishable samples when using the Robinson–Foulds distance, but not when using the path difference. We refer to Supplementary Figures S2 and S4 for pairwise heatmaps using the other distance metrics, which also recreate the previously observed patterns.

For the HIV posterior tree samples, all heatmaps actually show a clear distinction between the two samples. This discrepancy—which was not visible in the trace graphs of Fig. 5—is thus only produced when considering all pairwise distances as opposed to only distances from the first tree sample.

Network graphs

The 1000 × 1000 matrix of pairwise distances between trees can be converted into a similarity matrix, done here by normalizing the distances to a range of 0 to 1 and subtracting them from 1. Using a force-directed algorithm (Fruchterman and Reingold, 1991), one can create a two-dimensional graph where each tree is represented by a node and relative distances between nodes reflect the pairwise distances between trees. Fig. 7 shows the network graphs of the Robinson–Foulds distances for the EBOV and HIV trees, and of the branch score for the HIV trees. All other network



Metric: Robinson−Foulds

		Topological ESS		
Method	Combined runs	Run 1	Run 2	Run 3
Approximate	362	181	150	143
Split frequency	446	279	273	276
Fréchet correlation	83	229	209	211
Median pseudo	44	333	334	333
Minimum pseudo	15	276	285	271
cMDS	2	7	3	5

Metric: Weighted Robinson−Foulds

		Topological ESS		
Method	Combined runs	Run 1	Run 2	Run 3
Approximate	864	299	292	296
Split frequency	446	279	273	276
Fréchet correlation	46	238	189	212
Median pseudo	12	282	280	282
Minimum pseudo	3	189	189	177
cMDS	2	20	8	6

Metric: Branch Score

		Topological ESS		
Method	Combined runs	Run 1	Run 2	Run 3
Approximate	710	290	278	287
Split frequency	446	279	273	276
Fréchet correlation	22	222	139	189
Median pseudo	4	261	230	256
Minimum pseudo	2	187	108	89
cMDS	2	88	12	6

Metric: Path Difference

		Topological ESS		
Method	Combined runs	Run 1	Run 2	Run 3
Approximate	1000	333	334	333
Split frequency	446	279	273	276
Fréchet correlation	280	292	279	299
Median pseudo	1000	333	334	333
Minimum pseudo	631	205	166	239
cMDS	1000	333	334	333

Metric: Kendall−Colijn

		Topological ESS		
Method	Combined runs	Run 1	Run 2	Run 3
Approximate	1000	333	334	333
Split frequency	446	279	273	276
Fréchet correlation	593	302	296	226
Median pseudo	1000	333	334	333
Minimum pseudo	518	261	148	181
cMDS	861	333	280	273

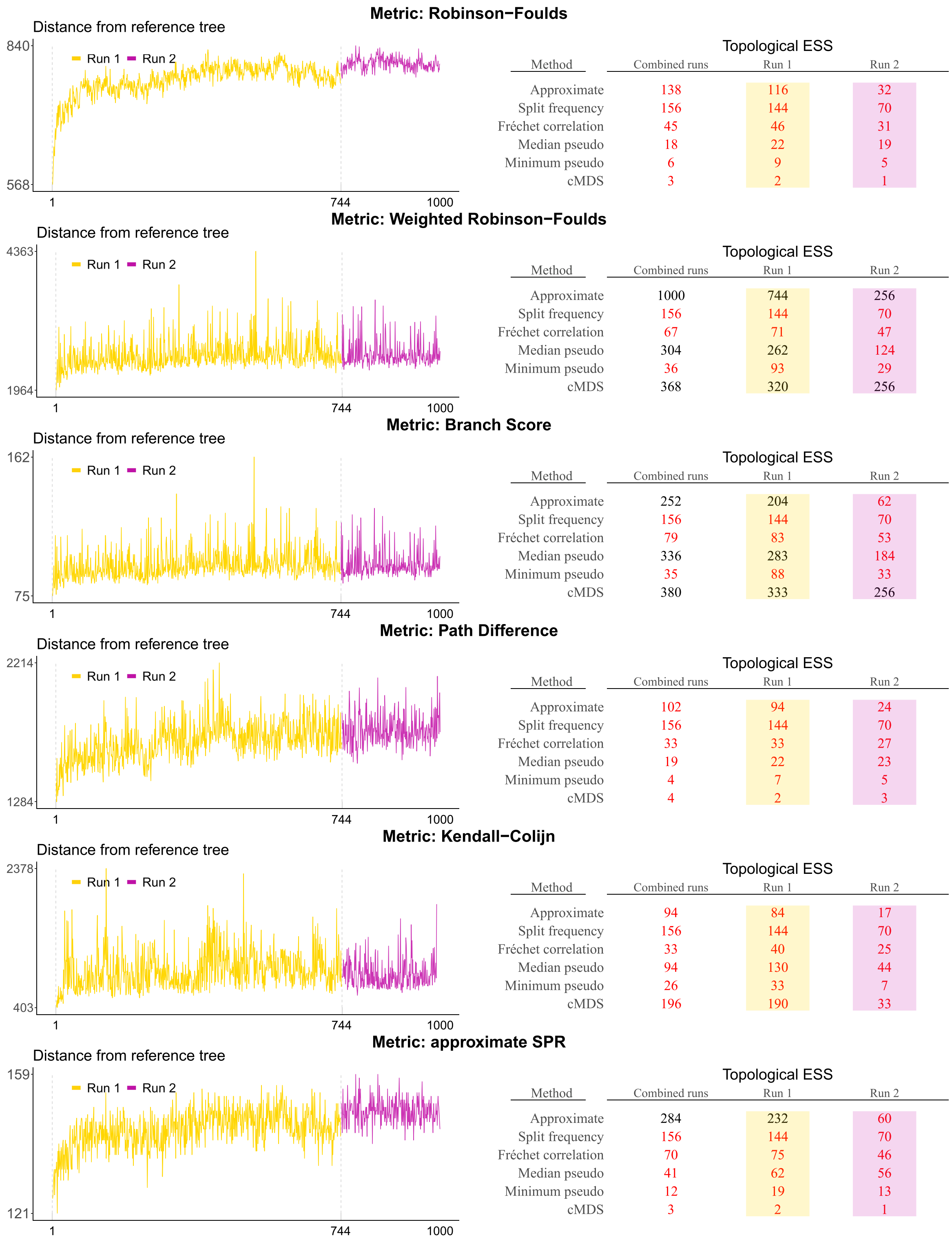
Metric: approximate SPR

		Topological ESS		
Method	Combined runs	Run 1	Run 2	Run 3
Approximate	66	333	333	334
Split frequency	446	279	273	276
Fréchet correlation	8	285	273	290
Median pseudo	2	333	333	334
Minimum pseudo	2	281	279	278
cMDS	2	9	13	17

Figure 4. Topology trace plots for the EBOV analysis using six different phylogenetic distance metrics, with associated topological ESS estimates. Color-coded trace plots correspond to color-coded columns in the topological estimates, with all of them generated from samples from the same three independent BEAST runs. The X-axis shows the index of the sampled tree (thinned sample), while the Y-axis shows the phylogenetic distance from the first tree. The distance from the first tree to itself (which is inevitably 0) is excluded from the graph to avoid scaling issues. ESS values below 200—an often-used cutoff in practice for terminating running analyses in Bayesian phylogenetics—are highlighted in red. This sample of 1000 trees is in fact a concatenated set of three samples from independent BEAST runs that yielded 333, 333, and 334 trees, respectively. Notice how certain topological distance metrics show clear jumps between the separate runs, while for others the traces seem entirely homogeneous.

graphs (for both datasets) can be found in Supplementary Figures S1 and S3.

For the EBOV trees, the patterns we came to expect from the trace graphs and heatmaps are reproduced. However, for the HIV trees, the two sets of samples can be clearly distinguished using the Robinson–Foulds distance, but not using the branch score. This is surprising, given how these two metrics are closely related (see Fig. 1). Furthermore, in the first subsample of the HIV trees (trees



Metric: Robinson−Foulds

Topological ESS

Method	Combined runs	Run 1	Run 2
Approximate	138	116	32
Split frequency	156	144	70
Fréchet correlation	45	46	31
Median pseudo	18	22	19
Minimum pseudo	6	9	5
cMDS	3	2	1

Metric: Weighted Robinson−Foulds

Topological ESS

Method	Combined runs	Run 1	Run 2
Approximate	1000	744	256
Split frequency	156	144	70
Fréchet correlation	67	71	47
Median pseudo	304	262	124
Minimum pseudo	36	93	29
cMDS	368	320	256

Metric: Branch Score

Topological ESS

Method	Combined runs	Run 1	Run 2
Approximate	252	204	62
Split frequency	156	144	70
Fréchet correlation	79	83	53
Median pseudo	336	283	184
Minimum pseudo	35	88	33
cMDS	380	333	256

Metric: Path Difference

Topological ESS

Method	Combined runs	Run 1	Run 2
Approximate	102	94	24
Split frequency	156	144	70
Fréchet correlation	33	33	27
Median pseudo	19	22	23
Minimum pseudo	4	7	5
cMDS	4	2	3

Metric: Kendall−Colijn

Topological ESS

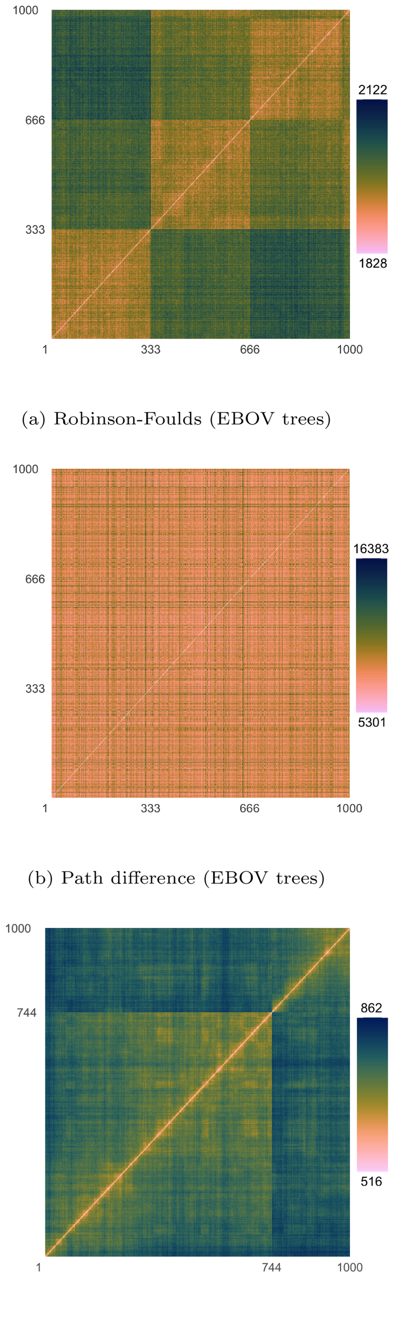
Method	Combined runs	Run 1	Run 2
Approximate	94	84	17
Split frequency	156	144	70
Fréchet correlation	33	40	25
Median pseudo	94	130	44
Minimum pseudo	26	33	7
cMDS	196	190	33

Metric: approximate SPR

Topological ESS

Method	Combined runs	Run 1	Run 2
Approximate	284	232	60
Split frequency	156	144	70
Fréchet correlation	70	75	46
Median pseudo	41	62	56
Minimum pseudo	12	19	13
cMDS	3	2	1

Figure 5. Topology trace plots for the HIV analysis using six different phylogenetic distance metrics, with associated topological ESS estimates. Color-coded trace plots correspond to color-coded columns in the topological estimates, with all of them generated from samples from the same two independent BEAST runs. The X-axis shows the index of the sampled tree (thinned sample), while the Y-axis shows the phylogenetic distance from the first tree. The distance from the first tree to itself (which is inevitably 0) is excluded from the graph to avoid scaling issues. ESS values below 200—an often used cutoff in practice for terminating running analyses in Bayesian phylogenetics—are coloured in red. This sample of 1000 trees is in fact a concatenated set of two samples from distinct *BEAST* runs that yielded 744 and 256 trees, respectively.



(a) Robinson-Foulds (EBOV trees)

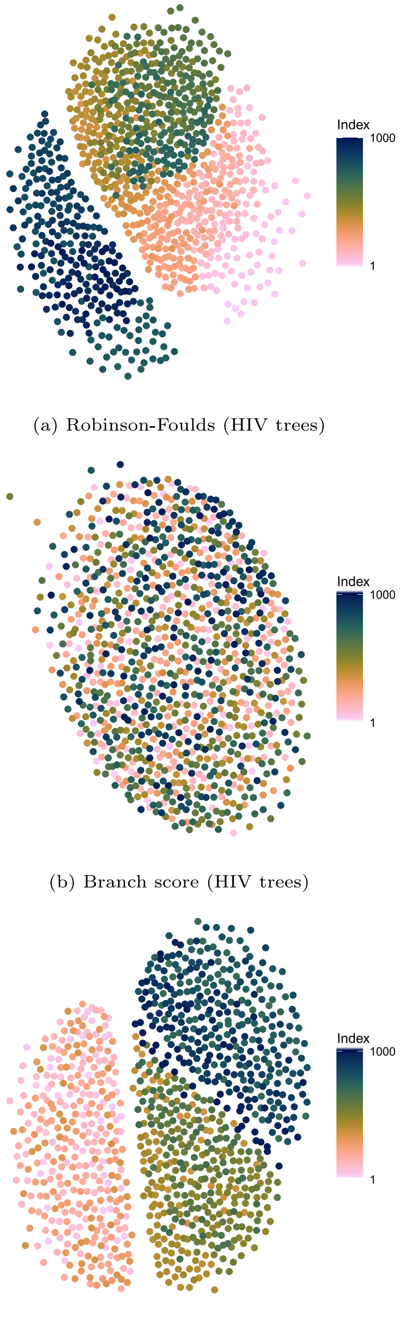
(b) Path difference (EBOV trees)

(c) Robinson-Foulds (HIV trees)

Figure 6. Heatmaps of pairwise distances between trees in the EBOV and HIV combined sets of samples. For the EBOV trees, the three sets of samples can be clearly distinguished using certain distance metrics (shown here: Robinson–Foulds) but not with others (shown here: path difference). For the HIV trees, the two sets of samples can be distinguished using the Robinson–Foulds distance, which was not the case in Fig. 5.

1–744), the colouring of the nodes show a clear gradient—in line with the shape of the trace graphs in Fig 5.

These visualizations may currently be hard to implement in practice, as the increase in computation time from a simple trace graph (which requires 999 distances in our 1000-tree samples) to a set of pairwise distances (499 500 distances) is substantial—going up to several days on a standard computer.



(a) Robinson-Foulds (HIV trees)

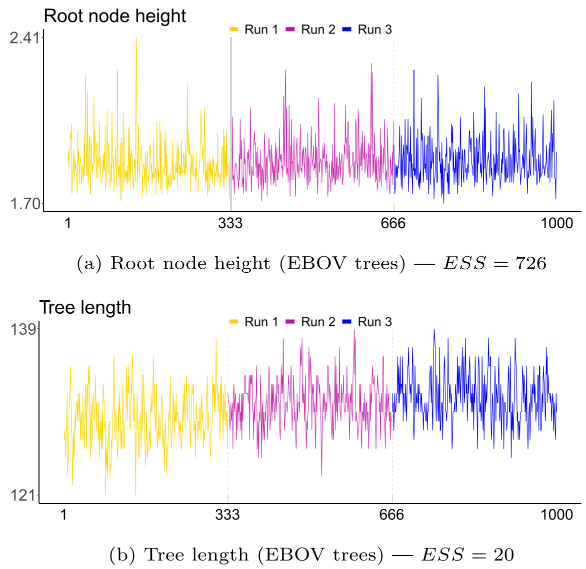
(b) Branch score (HIV trees)

(c) Robinson-Foulds (EBOV trees)

Figure 7. Network graphs of pairwise distances between trees in the EBOV and HIV combined sets of samples. Each node reflects a tree, and the relative distances between nodes reflect the pairwise distances between trees. Nodes are coloured by position in the sample, going from light to dark. For the HIV trees, the two sets of samples can be clearly distinguished using the Robinson–Foulds distance, but not using the branch score. For the EBOV trees, the network graph made with the Robinson–Foulds distance shows the three distinct runs clearly.

Non-distance metric based approaches

All approaches considered in the previous two sections—with the exception of the split frequency ESS—explicitly rely on computing phylogenetic distances between trees in a sample. We here discuss a few alternative explorations and diagnostics for topological convergence.



(a) Root node height (EBOV trees) — $ESS = 726$

(b) Tree length (EBOV trees) — $ESS = 20$

Figure 8. Trace plots for the root node height and tree length for the EBOV samples. Tree length refers to the sum of all tree branches, and is thus a statistic closely related to the topology of the tree. The root height, a commonly evaluated parameter, shows satisfactory mixing, while the tree length, which is less often included in convergence/mixing assessment, has an ESS of 20 which indicates problematic mixing. The ESS of individual replicates can be found in tables S1 and S2 for the EBOV and HIV samples, respectively.

Continuous parameters

An easy first step would be to consider whether the continuous parameters of the model show any signs of convergence or mixing issues. While this is standard practice in Bayesian phylogenetic inference, we here pay specific attention to those parameters that directly translate aspects of a phylogenetic tree into continuous values. Fig. 8 shows trace graphs of the height of the root node and the total tree length in the EBOV sample. The height of the root—a continuous parameter often evaluated using standard methodology, as it reflects the temporal distance between the youngest tip and the most recent common ancestor of all sequences—shows a well-mixed sample with a satisfactory ESS. The trace of the tree length, which is not often readily assessed despite it being readily implemented in the *BEAST* software package, shows three distinguishable subsamples—although the difference is not marked—as well as a very poor ESS of only 20.

Supplementary Figure S5 shows the same plots for the HIV sample. In this case, neither of the two statistics show a discrepancy between the two runs and the ESS values are very high.

Maximum clade credibility trees

The observed discrepancies between individual runs in the explored posterior topological space for both the EBOV and HIV samples raise the question of what these differences actually reflect in terms of phylogenetic inference. An often employed summary tree for such samples is the maximum clade credibility (MCC) tree. We can compute this tree for the total EBOV and HIV samples, as well as the individual subsamples, and compare them.

Fig. 9 shows a tanglegram comparing the MCC trees of the EBOV samples by linking the corresponding taxa to each other. Taxa are coloured by country of origin. The trees all have the same overall shape, but several clades end up in different locations depending on the individual run. Further, Supplementary Figure S6 shows a tanglegram of the MCC trees of the first and second half of the first EBOV run (so trees numbered 1–166 and trees numbered 167–333); these two trees thus reflect within-run variability. Although these trees have differences as well—as expected from the inherent randomness of the MCMC algorithm—they are substantially more similar than the MCC trees of different replicate analyses, as shown in Fig. 9, thus coinciding with our previous findings that suggested higher between-run variability than within-run variability.

Similarly, Fig. 10 shows a tanglegram comparing the MCC trees of the HIV samples by linking the corresponding taxa to each other. Taxa are not coloured by geographic origin but simply by location in the sample—as they were in the network graphs of Fig. 7. Again, there are substantial differences between the first and second run. Supplementary Figure S7 shows a tanglegram of the MCC trees of the first and second half of the first HIV run (so trees numbered 1–372 and trees numbered 373–744). Unlike for the EBOV trees, it is not apparent from the MCC trees that there would be more between-run than within-run variability in the sampled topologies. This is in line with the low ESS values for the HIV samples—as shown in Fig. 5—and can potentially be attributed to the typical star-like shape of HIV phylogenies.

ESS of individual tree splits

The approach by Guimarães Fabreti and Höhna (2021) does not aim to compute an ESS for the entire topology. Instead, each split in the tree is considered individually as a binary parameter (1 = present, 0 = absent) for which an ESS is computed using standard methodology.

Fig. 11 shows the cumulative density of ESS values of the 6 859 splits observed in the EBOV trees and the 2371 splits observed in the HIV trees. For the EBOV trees, the vast majority of these (94%) have an ESS above the often-used cutoff value of 200. Guimarães Fabreti and Höhna (2021) suggest a more stringent cutoff of 625, which is met by 90% of the splits. Although not shown here, the three independent samples that make up the full EBOV sample showed a nearly identically shaped distribution of individual split ESS values. Thus, the vast majority of splits show a satisfactory ESS both by according to the commonly used cutoff and more stringent cutoff suggested by Guimarães Fabreti and Höhna (2021). For the HIV trees, the ESS values are substantially lower, with only 51% of the splits having an ESS above 200 and 22% above 625. By both criteria, these ESS values suggest poor mixing of the HIV sample. Thus, when considering the approach towards topological convergence assessment suggested by Guimarães Fabreti and Höhna (2021), the EBOV trees suggest satisfactory results, while the HIV trees do not.

Discussion

In this case study, we applied an ensemble of diagnostics in topological space on a set of large trees, obtained through Bayesian phylogenetic inference under a set of complex models and through combining the output of at least two independent replicate analyses. Importantly, we find that the different approaches to evaluating topological convergence can lead to drastically different conclusions, a finding that to the best of our understanding has not been observed to this extent before. It is likely that a combination of the complexity of the models, the size of the datasets, and the fact that we looked at samples that result from two or three different/independent analysis replicates—as

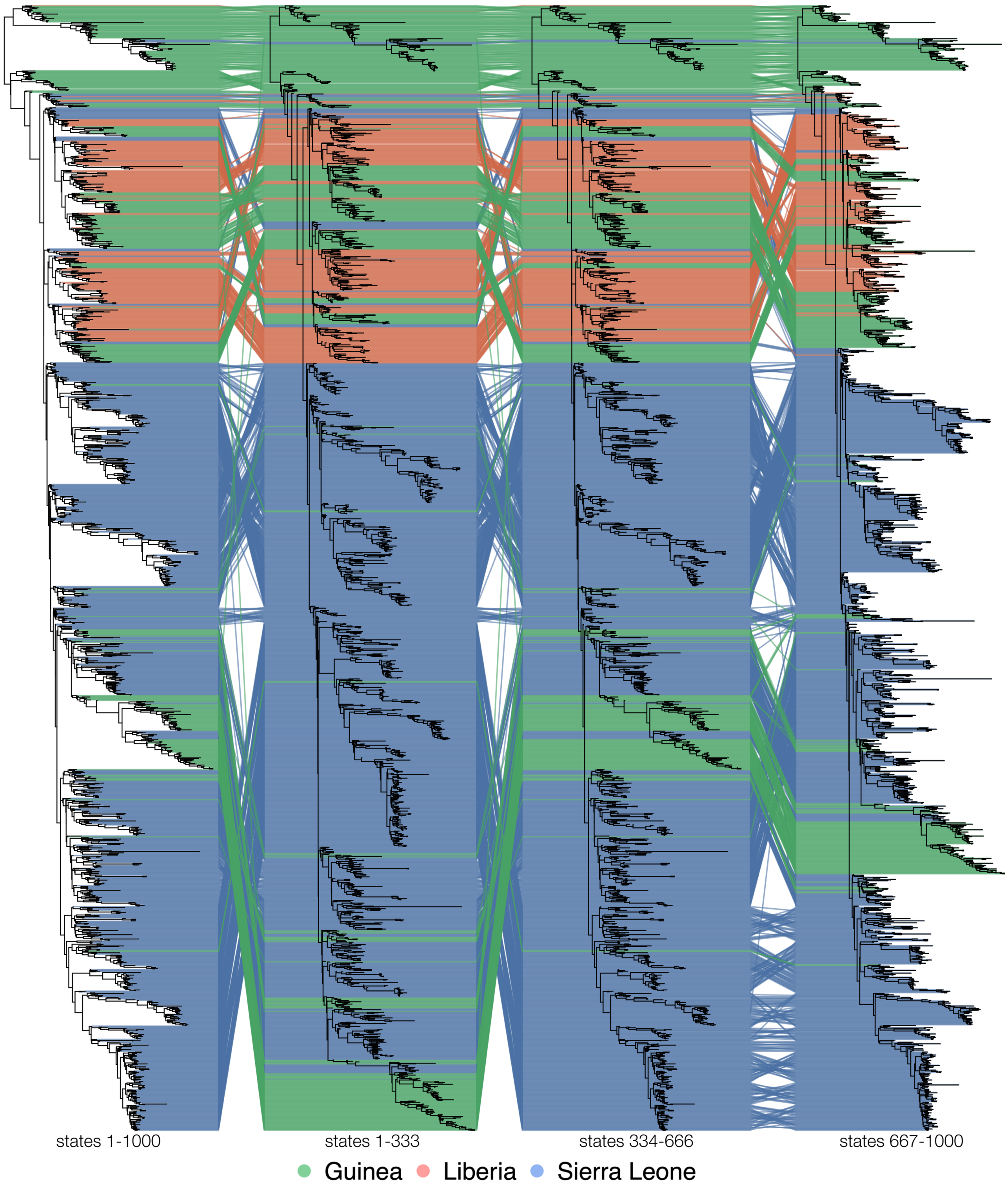


Figure 9. Tanglegram of the MCC trees of the EBOV sample. The tips of the trees are connected to each other by lines, coloured by the country of origin. Disagreement between the subsamples regarding the location of several clades is apparent by the fact that the lines connecting the tips of these clades are not parallel. The MCC tree of the total sample is identical to the MCC tree of the second subsample.

opposed to only just one—allowed us to observe these phenomena that went unnoticed before. These findings stress the importance of assessing topological convergence to the posterior and not merely continuous parameter and (joint) density convergence, which is the current approach in nearly all Bayesian phylogenetic and phylodynamic studies. Even when additional continuous parameters—such as the root height and the tree length—are to be logged for assessing their ESS, these still offer no guarantee at avoiding topological convergence difficulties. Whether the discrepancies we found affect inferences on estimates of parameters of interest downstream in the analysis is not yet clear, and warrants further research.

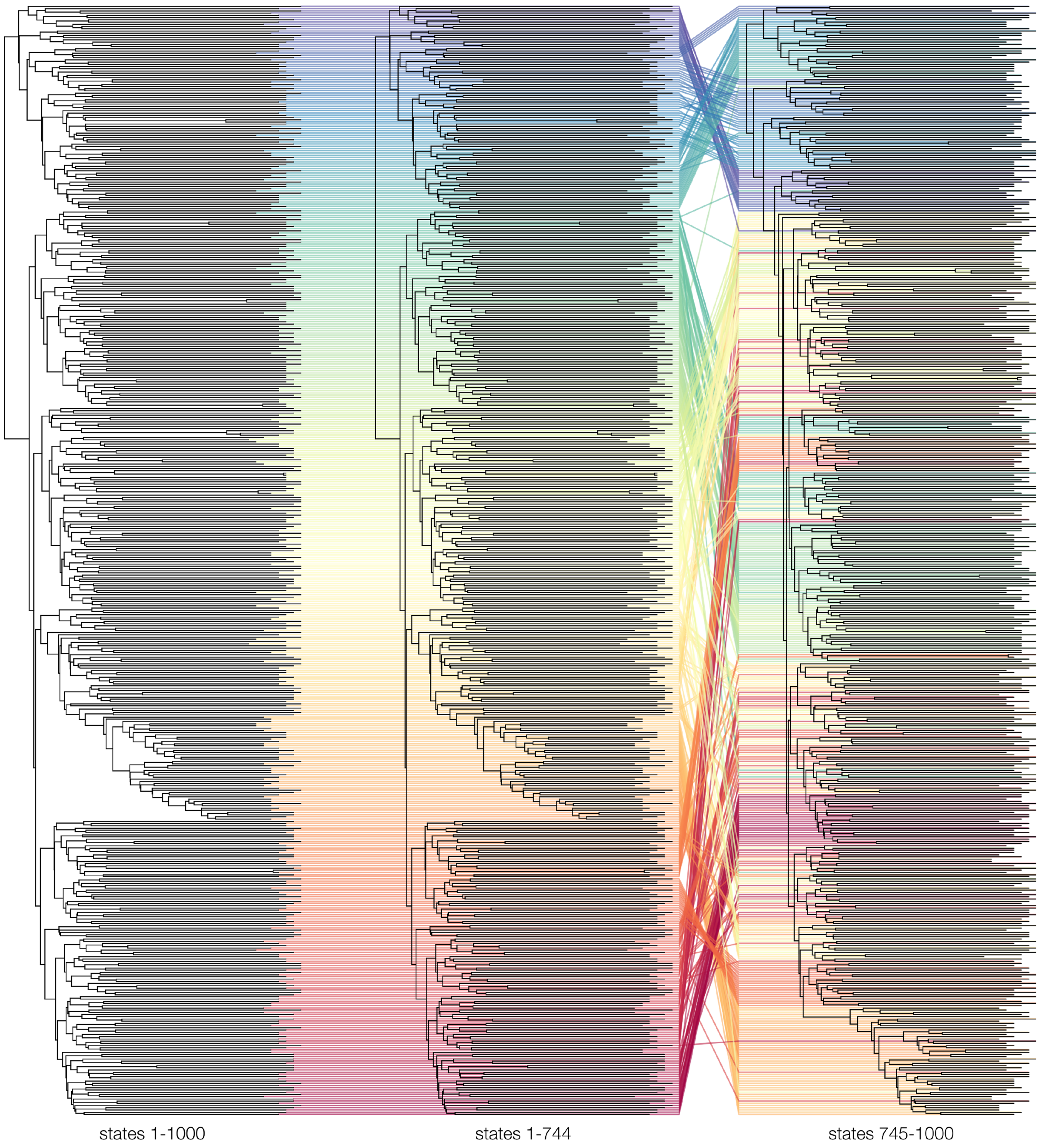


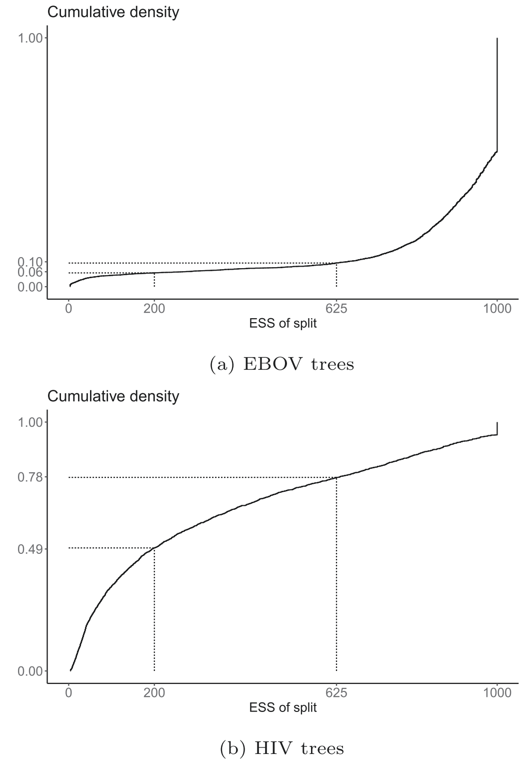
Figure 10. Tanglegram of the MCC trees of the HIV sample. The tips of the trees are connected to each other by lines, coloured by position in the first MCC tree. Disagreement between the subsamples regarding the location of several clades is apparent by the fact that the lines connecting the tips of these clades are not parallel. The MCC tree of the total sample is identical to the MCC tree of the first subsample.

The importance of performing replicate analyses

We emphasize the importance of running more than one replicate analysis—using different starting points—when performing Bayesian phylodynamic inference, as it is clear from our results that even a well-behaved sample from a single replicate may not be representative of the posterior topological space. When performing multiple replicate/independent analyses, it is important to favour a few long runs over many short runs (see section 1.11.3 in Brooks et al. (2011)), as many short runs can keep one from running the analysis long enough to detect pseudo-convergence (i.e. when the Markov chain appears to have converged but not to the true posterior distribution, possibly due to parts of the state space being poorly connected by the Markov chain dynamics which means that it takes many iterations to get from one part to another) or other problems.

Visualizing topological convergence

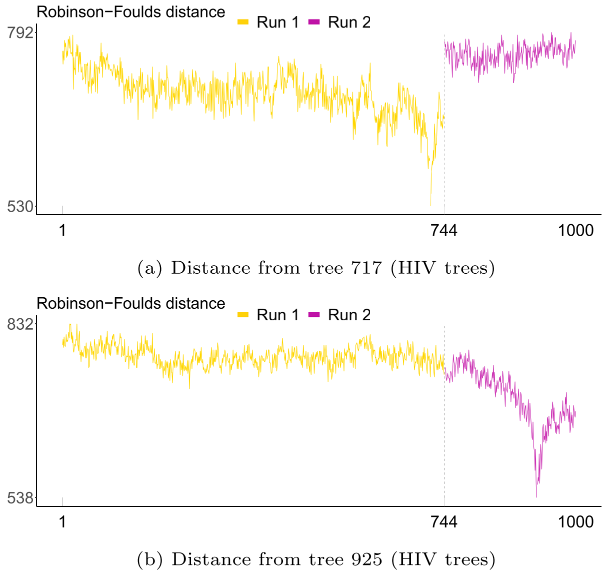
The most straightforward approach to visualizing topological convergence is the topological trace plot. Making such a graph requires the choice of a reference tree, which can have a substantial impact on the ability of the trace to discriminate between runs.



(a) EBOV trees

(b) HIV trees

Figure 11. Cumulative densities of the ESS values of individual splits in the sampled EBOV and HIV trees. ESS values of 200 and 625 are indicated and connected to their corresponding cumulative density values with dotted lines. These ESS values were computed using the `convenience` R package.



(a) Distance from tree 717 (HIV trees)

(b) Distance from tree 925 (HIV trees)

Figure 12. Trace plots of the Robinson–Foulds distance to each tree using the MCC trees of the first and second subsample (trees 717 and 925, respectively) as the reference tree for the HIV analysis.

For example, the two HIV runs could not be distinguished in any of the topology trace graphs in Fig. 5, which used the first tree in the sample as a reference tree from which to compute distances—but were clearly distinct when considering all pairwise distances in Figs 6 and 7. Fig. 12 shows the trace plots of the Robinson–Foulds distance to each tree using the MCC trees of the first and second subsamples as the reference trees, instead of simply the first. The discrepancy between individual runs is much more apparent here than in the trace plots of Fig. 5. This suggests that the choice of reference tree can have a substantial impact on the ability of topological traces to discriminate between runs, and is an important consideration when interpreting these diagnostics. Furthermore, the trace graphs of Fig. 5 had a trajectory reminiscent of undiscarded burn-in. However, the trace graphs of Fig. 12 show a pronounced slump around the reference (MCC) trees, which cannot be explained by burn-in. A possible explanation would be that—despite thinning of the chain and there being no sign of any problems with the transition kernels—the chain could exhibit very strong autocorrelation leading to the trees closer to the reference tree simply being more similar to it. Given that the reference tree in Fig. 5 was the first tree, it can be difficult to tell whether the slump is due to autocorrelation or burn-in.

The issue of selecting a reference tree is resolved by considering pairwise distance based visualizations such as heat maps and network graphs instead of a topological trace plot, making such visualizations preferable. However, the computational cost of calculating $n \times (n-1)/2$ distances for a sample of n trees quickly ramps up as n increases. A possible solution would be to reduce the sample size for the purposes of convergence assessment by only using a subset of equally spaced trees—such as was done in Supplementary Figures S1(e) and S3(e). This can substantially reduce the computational requirements, at the cost of not considering every tree in the sample.

The choice of distance metric

The impact of the choice of topological distance metric on the ability to detect discrepancies in the combined chains cannot be systematically determined in a case-study format such as this. Comparing trees using SPR distances—which are closely related to how the studied trees were generated since most implemented MCMC transition kernels use SPR-like moves for exploring tree-space—did suggest that the trees explored by the replicate analyses differ in a topologically substantial manner. In the EBOV data, the metrics which detected differences between replicates were those based on splits in the trees, while the metrics that could not were those based on tip-to-tip distance. In the HIV data however, the difference in performance between these two categories was much less clear, with metrics from both groups being both capable and incapable of detecting differences between replicates. We speculate that tree shape—with the HIV data in this analysis being characteristically 'star-like' with longer branches towards the tips, and the EBOV data being more 'ladder-like' (Colijn and Plazzotta, 2018)—may have an impact on how sensitive different topological distance metrics are to changes in topology.

While we would recommend using a variety of distance metrics when performing topological convergence assessment, the unweighted Robinson–Foulds distance and the SPR distance are good starting points given that they were the only ones capable of capturing the discrepancies between the independent replicates for both the EBOV and HIV data. As for which ESS estimator to use, it is difficult to make strong recommendations based on our current results, given that the behaviour of the different estimators depended both on the dataset and the distance metric used. We would restate the conclusions made by Magee et al. (2023) that the pseudo-ESS and the Fréchet correlation ESS are the optimal choice according to their experiments, although we recommend looking at the entire range of pseudo-ESS values (not just the median and minimum), since the choice of reference tree can have a large impact on its value.

On combining the output of independent replicates

Finally, our findings raise important questions as to how the output of replicate Bayesian phylogenetic and phylodynamic analyses should be combined when discrepancies in topological space are detected. The EBOV and HIV replicates produced samples from different regions of the posterior distribution, and the sampled trees from these replicates were systematically more different than trees from the same replicate, which suggests that these replicates were stuck in local modes of the topological space. A key open question stemming from our work is how to combine these sampled trees in such a way that the resulting summary tree accurately reflects the region of highest posterior topological probability as well as the uncertainty surrounding it.

Simply averaging the samples with weights proportional to their size—which is currently standard practice—might not produce estimates that properly reflect the multimodal posterior. As with the HIV example, where the two subsamples were of unequal size, the region explored by the first subsample represents almost three fourths of the total sample, a degree of representation that is unlikely proportional to the posterior mass of this region compared to the second subsample. From Fig. 10, the MCC tree of the total HIV sample is the MCC tree of the first subsample. However, when we artificially lengthen the second subsample by duplicating it two additional times, such that the first and second subsample are roughly of equal length, we find that the MCC tree of the total sample is no longer the initial estimate. Thus, if concatenated samples are exploring different regions of topological space, their relative weights in the total sample have a meaningful impact on downstream inferences. Techniques such as importance sampling could be employed to weigh samples proportionally to their posterior density (Yao et al., 2022), such that relative representation of different regions of the posterior are preserved, but we considered these to deserve additional attention and out of scope of the current manuscript.

Acknowledgements

The authors thank Barney Isaksen Potter for their help developing quality figures and Andrew Magee for their valuable technical support.

Author contributions

M.B. and G.B. initialised the study. M.B. performed the analyses and wrote the manuscript. G.D. created the tanglegrams and computed the aSPR distances. L.M.C., J.G., F.A.M. A.R., M.A.S., G.D., S.L.H., P.L., and G.B. provided valuable guidance and feedback. All authors reviewed the manuscript.

Supplementary data

Supplementary data are available at *Virus Evolution* online.

Conflict of interest: None declared.

Funding

S.L.H. and G.B. acknowledge support from the Research Foundation—Flanders ('Fonds voor Wetenschappelijk Onderzoek—Vlaanderen,' G0E1420N). G.B. acknowledges support from the Internal Funds KU Leuven (Grant No. C14/18/094), from the Research Foundation—Flanders ('Fonds voor Wetenschappelijk Onderzoek—Vlaanderen,' G098321N) and from the European Union Horizon 2023 RIA project LEAPS (Grant Agreement No. 101094685). P.L., M.A.S., and A.R. acknowledge support from the Wellcome Trust (Collaborators Award 206298/Z/17/Z, ARTIC network), the European Research Council (Grant Agreement No. 725422—ReservoirDOCS) and the National Institutes of Health (NIH) (R01 AI153044). F.A.M. and M.A.S. acknowledge further support from the NIH through R01 AI162611. P.L. acknowledges support from the Research Foundation, Flanders ('Fonds voor Wetenschappelijk Onderzoek—Vlaanderen,' G066215N, G0D5117N, and G0B9317N) and from the European Union Horizon 2020 project MOOD (Grant Agreement No. 874850). M.B. and G.B. acknowledge support from the DURABLE EU4 Health Project 02/2023-01/2027, which is co-funded by the European Union (call EU4H-2021-PJ4) under Grant Agreement No. 101102733. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Data availability

The data underlying this article are freely available through the respective GitHub repositories linked in the *Materials and Methods* section.

References

Attwood S, Hill S, Aanensen D *et al.* Phylogenetic and phylodynamic approaches to understanding and combating the early SARS-Cov-2 pandemic. *Nat Rev Genet* 2022;**23**:1–16.

Brooks S., Gelman A., Jones G. *et al. Handbook of Markov Chain Monte Carlo*. CRC Press, 2011.

Colijn C, Plazzotta G. A metric on phylogenetic tree shapes, *Syst Biol* 2018;**67**:113–126.

Dudas G, Carvalho LM, Bedford T *et al.* Virus genomes reveal factors that spread and sustained the Ebola epidemic, *Nature*, 2017;**544**:309–315.

Fruchterman TMJ, Reingold EM. Graph drawing by force-directed placement, *J Softw Pract Exp* 1129–1164, 1129–1164;**21**:1129–1164.

Gill M, Lemey P, Faria N *et al.* Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci, *Mol Biol Evol* 2013;**30**:713–724.

Guimarães Fabreti L, Höhna S. Convergence assessment for Bayesian phylogenetic analysis using MCMC simulation, *Methods Ecol Evol* 2021;**13**:77–90.

Hong SL, Dellicour S, Vrancken B *et al.* In search of covariates of HIV-1 subtype B spread in the United States—a cautionary tale of large-scale Bayesian phylogeography, *Viruses*, 2020;**12**:182.

Kendall M, Colijn C. Mapping phylogenetic trees to reveal distinct patterns of evolution, *Mol Biol Evol* 2016;**33**:2735–2743.

Kruskal JB. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, *Psychometrika*, 1964a;**29**:1–27.

Kruskal JB. Nonmetric multidimensional scaling: a numerical method, *Psychometrika*, 1964b;**29**:115–129.

Kuhner M, Felsenstein J. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates, *Mol Biol Evol* 1994;**11**:459–468.

Lanfear R, Hua X, Warren DL. Estimating the effective sample size of tree topologies from Bayesian phylogenetic analyses, *Genome Biol Evol* 2016;**8**:2319–2332.

Magee A, Karcher M, Matsen F *et al.* How trustworthy is your tree? Bayesian phylogenetic effective sample size through the lens of Monte Carlo error, *Bayesian Anal* 2023;**1**: 1–29.

Plummer M, Best N, Cowles K *et al.* Package 'coda', *CRAN*, 2020.

Rambaut A, Drummond A, Xie D *et al.* Posterior summarisation in Bayesian phylogenetics using Tracer 1.7 (available at http://beast.community/tracer), *Syst Biol* 2018;**67**:901–904.

D. Robinson and L. Foulds. Comparison of weighted labelled trees. In: Horadam AF, Wallis, WD, (eds) *Combinatorial Mathematics VI*, Vol. 748, 119–126, 1979.

Robinson D, Foulds L. Comparison of phylogenetic trees, *Math Biosci* 1981;**53**:131–147.

R-Team. R: A language and environment for statistical computing, 2022.

Schliep K, Paradis E, de Oliveira Martins L *et al.* Package 'phangorn', *CRAN*, 2022.

Smith M, Jonker R, Yang Y *et al.* Treedist: calculate and map distances between phylogenetic trees. *R package* 2023. https://github.com/ms609/TreeDist

Steel MA, Penny D. Distributions of tree comparison metrics - some new results, *Syst Biol* 1993;**42**:126–141.

Suchard MA, Lemey P, Baele G *et al.* Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10, *Virus Evol* 2018;**4**:vey016.

Whidden C, Beiko R, Zeh N Fixed-parameter algorithms for maximum agreement forests, *SIAM J Comput* 2013;**42**:1431–1466.

Wirth W, Duchene S Real-time and remote MCMC trace inspection with beastiary, *Mol Biol Evol* 2022;**39**:msac095.

Xu S, Li L, Luo X *et al.* Ggtree: a serialized data object for visualization of a phylogenetic tree and annotation data, *iMeta*, 2022;**1**:e56.

Yao Y, Vehtari A, Gelman A Stacking for non-mixing Bayesian computations: the curse and blessing of multimodal posteriors, *J Mach Learn Res*, 2022;**23**:1–45.

Virus Evolution, 2024, 10(1), veae081 , DOI: https://doi.org/10.1093/ve/veae081, Advance access publication date: 15 October 2024, Reflections