Bioinformatics, 2026, **42**(5), btag195
https://doi.org/10.1093/bioinformatics/btag195
Published: 22 April 2026
Applications Note | Phylogenetics

OXFORD

treestructure: an R package to detect population structure in phylogenetic trees

Fabrícia F. Nascimento, Vinicius B. Franceschi, Erik M. Volz*

MRC Centre for Global Infectious Disease Analysis, Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, London, W12 0BZ, United Kingdom
*Corresponding author. MRC Centre for Global Infectious Disease Analysis, Department of Infectious Disease Epidemiology, School of Public Health, White City Campus, 90 Wood Lane, Imperial College London, London W12 0BZ, United Kingdom. E-mail: e.volz@imperial.ac.uk.
Associate Editor: Peter Robinson

Abstract

Motivation: **How population structure can shape genetic diversity is a longstanding problem in population genetics. While the use of geographic locations, when available, can help answer some of these questions, it is still difficult to determine population structure when such metadata are not available or when the potential population structure is not easily observed. Here, we present an updated version of *treestructure*, an R package that implements a statistical test based on coalescent theory to detect unobserved population structure in a time-scaled phylogenetic tree.**

Availability: **treestructure is available at CRAN at https://cloud.r-project.org/web/packages/treestructure/ and at https://emvolz-phylodynamics.github.io/treestructure/.**

1 Introduction

The *treestructure* R package implements statistical tests based on coalescent theory to detect unobserved population structure in time-scaled phylogenetic trees (Volz et al. 2020). A time-scaled phylogenetic tree shows the evolutionary relationships between organisms in units of calendar time, and *treestructure* will detect clusters that are likely to be the result of distinct population structure with heterogeneous demographic or epidemiological history (Volz et al. 2020). The algorithm implemented in *treestructure* also groups clusters showing similar population structure into partitions. The initial version of the package has been used in a variety of studies, such as the detection of lineages with different demographic histories in SARS-CoV-2 (Fountain-Jones et al. 2020); the detection of a fitness advantage in clades that showed similar demographic histories in *Neisseria gonorrhoea* (Joseph et al. 2022); and the understanding of population of *Vibrio parahaemolyticus* in Latin America (Campbell et al. 2024).

In the new version of *treestructure*, we have added new functionalities that have not been previously described, enhancing its practical utility and statistical robustness. These new functionalities include (i) methods to automatically choose clustering hyperparameters, (ii) use of branch support values (e.g. bootstrap and posterior clade credibility) to filter out clusters with low phylogenetic confidence, and (iii) the addition of new tips to a previous *treestructure* object, allowing clusters to be updated in an online fashion as new data become available.

Other methods have been developed to detect population structure in phylogenetic trees, such as *CaveDive* (Helekal et al. 2022) and *fastbaps* (Tonkin-Hill et al. 2019), and have been applied to the detection of outbreaks and variant surveillance (Reimche et al. 2023, Binney et al. 2025). However, none of these methods can take into consideration the use of branch support to avoid detecting clusters on unsupported clades.

2 Implementation

treestructure is implemented as an R (R Core Team 2025) package and is available on CRAN (https://cloud.r-project.org/web/packages/treestructure/) and on GitHub (https://github.com/emvolz-phylodynamics/treestructure). For complete details on installation, documentation and tutorials using the new features, see the package website (https://emvolz-phylodynamics.github.io/treestructure).

Below, we give details on the new features and how these enhance the practical utility and statistical robustness of *treestructure* by using a publicly available time-scaled phylogeny based on Ebola virus whole genomes (https://github.com/ebov/space-time/blob/master/Data/Makona_1610_cds_ig.GLM.MCC.tree) collected during the 2014 epidemic in West Africa (Dudas et al. 2017).

2.1 Clustering significance level

Clustering methods, such as the one used in *treestructure*, require the specification of hyperparameters that specify how aggressively a method will partition data. The *treestructure* algorithm repeatedly tests subclades for deviations in their time of the most recent common ancestor (TMRCA) distributions

Received: 22 December 2025. **Revised:** 20 March 2026. **Accepted:** 10 April 2026

from expectations under a neutral coalescent model and divides the tree where such deviations occur. To perform this test, it uses a *rank-sum* statistic, and clusters are defined when this coalescent-based statistical test detects a difference according to a desired significance level. Decreasing the significance level in the *treestructure* algorithm will increase the number of clusters detected. However, detecting more clusters will also increase the number of false-positive detections.

To determine the significance level, users can use additional metadata associated with each sample and then select the significance level which gives a set of clusters that explains the most variance in the data of interest (e.g. use the cluster as a factor in an ANOVA).

If metadata information is not available, users can use the new feature that implements the Caliński–Harabasz index, or CH-index (Caliński and Harabasz 1974) which is a metric based on within- and between-cluster variance in a given statistic to select a quasi-optimal significance level. Within *treestructure* we use the node heights of the phylogeny itself, observed within each cluster, as the statistic that is used when computing the CH-index. Thus, clusters are selected such that there is high between-cluster variance in phylogenetic node heights. A step-by-step tutorial on how to run such an analysis can be found on the *treestructure* website (https://emvolz-phylodynamics.github.io/treestructure/articles/supportValues.html).

2.2 Branch support values

As there is often a great deal of uncertainty about individual phylogenetic splits, a user may not want to cluster their data along branches which are poorly supported. Because of that, we have implemented the use of branch support (e.g. bootstrap and posterior probability) to refine clusters in *treestructure*. To use this functionality, the time-scaled tree should be annotated with node support values, and the user will need to define a node support threshold value between 0 and 100. Nodes with support values less than the threshold value will not be tested. This feature is very useful for filtering out clusters that may not correspond to real phylogenetic splits, which was the case when analysing HIV-1 sequences published by our group (Franceschi et al. 2025).

Figure 1 shows an example of how the use of node support can filter out clusters with low phylogenetic confidence. Note that in the previous description of *treestructure*, the researcher would observe a higher number of clusters, many of which are in poorly supported clades as observed in this downsampled Ebola data (Fig. 1A).

2.3 Online inference by adding new samples to a previous *treestructure* object

Without the need to run multiple *treestructure* analyses, users can now update a *treestructure* object with new samples observed in a phylogenetic tree. This is particularly useful for classifying the cluster type of new samples without re-estimating and reassigning cluster membership. The updated tree does not need to be time-scaled or binary, reducing the need for expensive computation.

This new feature is implemented in the *addtips* function in the *treestructure* R package. The function *addtips* will compare

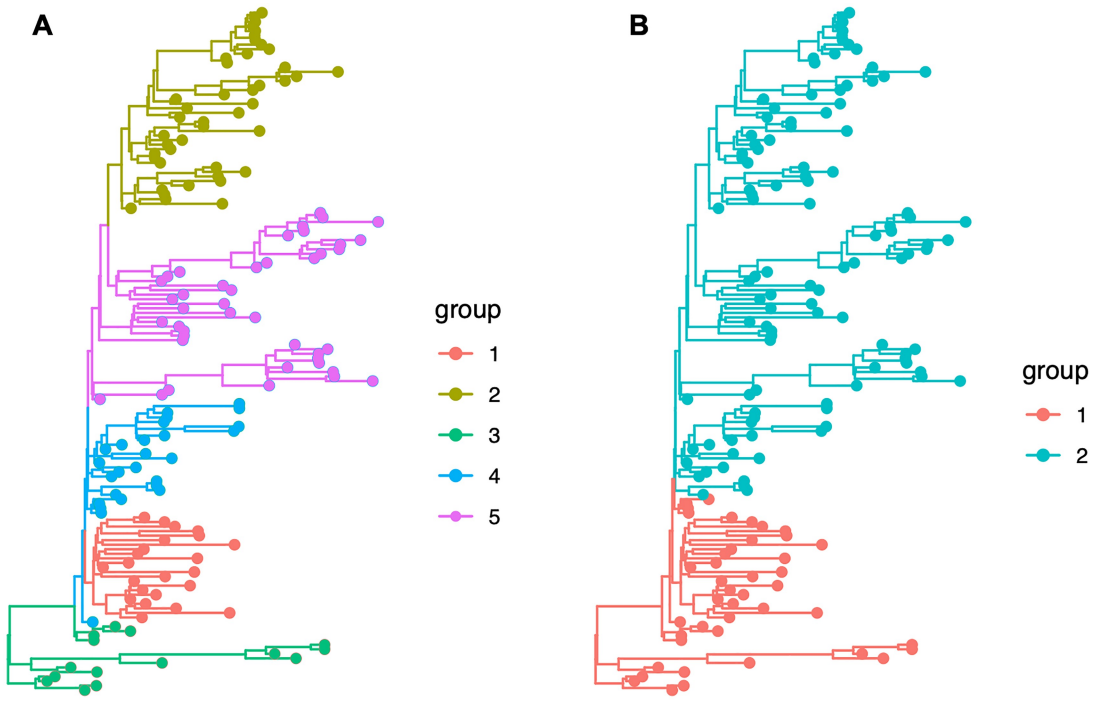

Figure 1 Clusters obtained by running *treestructure*, (A) without the use of branch support and (B) using a branch support threshold of 95. For both analyses, we used a significance level of 0.01 and a minimum clade size of 15 sequences. This figure was created using a publicly available Ebola time-scaled phylogenetic tree (https://github.com/ebov/space-time/blob/master/Data/Makona_1610_cds_ig.GLM.MCC.tree) downsampled to 150 sequences. An example analysing the complete Ebola dataset with 1610 sequences can be found on the *treestructure* website (https://emvolz-phylodynamics.github.io/treestructure/articles/supportValues.html).

the new phylogenetic tree to the old *treestructure* object, and it will merge the tips of the new tree into the *treestructure* object. Merging is carried out based on a phylogenetic criterion: New tips are added to the cluster which includes their most recent common ancestor in the new phylogeny. A step-by-step tutorial on how to use this feature can be found at https://emvolz-phylodynamics.github.io/treestructure/articles/updating_treestructure.html.

Acknowledgements

We would like to acknowledge Oliver Stirrup for previous contribution to the package. VBF acknowledges funding from the Imperial College London President's PhD Scholarship.

Author contributions

Fabrícia F. Nascimento (Data curation [equal], Formal analysis [equal], Software [equal], Writing—original draft [lead], Writing—review & editing [equal]), Vinicius B. Franceschi (Data curation [equal], Formal analysis [equal], Software [equal], Writing—review & editing [equal]), and Erik M. Volz (Conceptualization [lead], Funding acquisition [lead], Methodology [lead], Software [lead], Supervision [lead], Writing—review & editing [equal])

Conflicts of interest

None declared.

Funding

This work has been supported by the UK Health Security Agency CARAA 104683ED "Development of phylogenetic analysis tools to track HIV transmissions for public health surveillance purposes"; CARAA 5126118 "Provision of expert advice and development support for emerging infections genomics and metagenomics analysis"; and Wellcome Trust Investigator Award 220885/Z/20/Z "Population genomic analysis of HIV transmission fitness" awarded to E.M.V.

Data availability

Data and scripts used in this manuscript are available at https://github.com/emvolz-phylodynamics/treestructure/

Ebola phylogenetic tree used to generate Fig. 1 is available at https://raw.githubusercontent.com/ebov/space-time/master/Data/Makona_1610_cds_ig.GLM.MCC.tree

References

Binney BM, Gias E, Foxwell J *et al.* Genomic analysis of the 2017 Aotearoa New Zealand outbreak of *Mycoplasma bovis* and its position within the global population structure. *Front Microbiol* 2025;**16**:1600146.

Caliński T, Harabasz J. A dendrite method for cluster analysis. *Commun Stat* 1974;**3**:1–27.

Campbell AM, Gavilan RG, Abanto Marin M *et al.* Evolutionary dynamics of the successful expansion of pandemic *Vibrio parahaemolyticus* ST3 in Latin America. *Nat Commun* 2024;**15**:7828.

Dudas G, Carvalho LM, Bedford T *et al.* Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature* 2017;**544**:309–15.

Fountain-Jones NM, Appaw RC, Carver S *et al.* Emerging phylogenetic structure of the SARS-CoV-2 pandemic. *Virus Evol* 2020;**6**:veaa082.

Franceschi VB, Drake KO, Bibby DF *et al.* Evidence for circulation of high-virulence HIV-1 subtype B variants in the United Kingdom. *Virus Evol* 2025;**11**:veaf048.

Helekal D, Ledda A, Volz E *et al.* Bayesian inference of clonal expansions in a dated phylogeny. *Syst Biol* 2022;**71**:1073–87.

Joseph SJ, Thomas JC, IV, Schmerer MW *et al.*; Antimicrobial Resistant Neisseria gonorrhoeae Working Group. Global emergence and dissemination of *Neisseria gonorrhoeae* ST-9363 isolates with reduced susceptibility to azithromycin. *Genome Biol. Evol* 2022;**14**:evab287.

R Core Team 2025. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/ (27 April 2026, date last accessed).

Reimche JL, Clemons AA, Chivukula VL *et al.* Genomic analysis of 1710 surveillance-based *Neisseria gonorrhoeae* isolates from the USA in 2019 identifies predominant strain types and chromosomal antimicrobial-resistance determinants. *Microb Genom* 2023;**9**:mgen001006.

Tonkin-Hill G, Lees JA, Bentley SD *et al.* Fast hierarchical Bayesian analysis of population structure. *Nucleic Acids Res* 2019;**47**:5539–49.

Volz EM, Carsten W, Grad YH *et al.* Identification of hidden population structure in time-scaled phylogenies. *Syst Biol* 2020;**69**:884–96.

Bioinformatics, 2026, 42, 1–3
https://doi.org/10.1093/bioinformatics/btag195
Applications Note