The Journal of Open Source Software

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

Fabrícia F. Nascimento [1], **Vinicius B. Franceschi** [1], **and Erik M. Volz** [1¶]

1 MRC Centre for Global Infectious Disease Analysis, Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, London, UK ¶ Corresponding author

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor:

Submitted: 03 October 2025
Published: unpublished

Summary

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 is not available or when the potential population structure is not easily observed. Methods developed to detect population structure have been developed, such as *CaveDive* (Helekal et al., 2022) and *fastbaps* (Tonkin-Hill et al., 2019), and applied to the detection of outbreaks and variant surveillance (Binney et al., 2025; Reimche et al., 2023).

Here we present `treestructure`, an R package that has been previously described (Volz et al., 2020) and used in a variety of studies, such as detection of lineages with different demographic histories in SARS-CoV-2 (Fountain-Jones et al., 2020); detection of fitness advantage on clades that showed similar demographic histories in *Neisseria gonorrhoea* (Joseph et al., 2022); and understanding a population of *Vibrio parahaemolyticus* in Latin America (Campbell et al., 2024).

The 'treestructure R package implements a statistical test based on the coalescent theory to detect unobserved population structure in a time-scaled phylogenetic tree. A time-scaled phylogenetic tree shows the evolutionary relationship between organisms in units of calendar time. We have now added new features to treestructure for the detection of population structure and for adding additional samples to a previous treestructure object.

Statement of need

`treestructure` is an R package developed to find clusters within a time-scaled phylogeny that are likely to show a distinct population structure, such as, demographic or epidemiological history (Volz et al., 2020). The `treestructure` R package also groups clusters showing similar population structure into partitions. Here, we describe new functionalities added to the package, enhancing its practical utility and statistical robustness: 1) Methods to automatically choose clustering hyperparameters; 2) use of branch support values (*e.g.* bootstrap and posterior clade credibility) to filter out clusters with low phylogenetic confidence, and 3) adding new tips to a previous `treestructure` object, allowing clusters to be updated in an online fashion as new data becomes available.

For details of the main algorithm used in `treestructure` see Volz et al. (2020).

For details on installation, documentation and tutorial using the new features see the package website.

Clustering significance level

Clustering methods require the specification of hyperparameters that specify how aggressively a method will partition data. The `treestructure` algorithm makes use of a *rank-sum significance level*, and clusters are defined when a coalescent-based statistical test detects a difference according to this 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 & 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. If the user decided to use the CH-index, the option `level` in the *trestruct* function should be set to NULL and a lower and upper bound for optimizing the significance level should be provided. A step-by-step tutorial on how to run such analysis can be found here.

Branch support

Whereas 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. We have also 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 value less than the threshold value will not be tested. This feature is very useful to filter out clusters that may not correspond to real phylogenetic splits.

Figure 1 shows an example on how the use of node support can filter out clusters with low phylogenetic confidence.

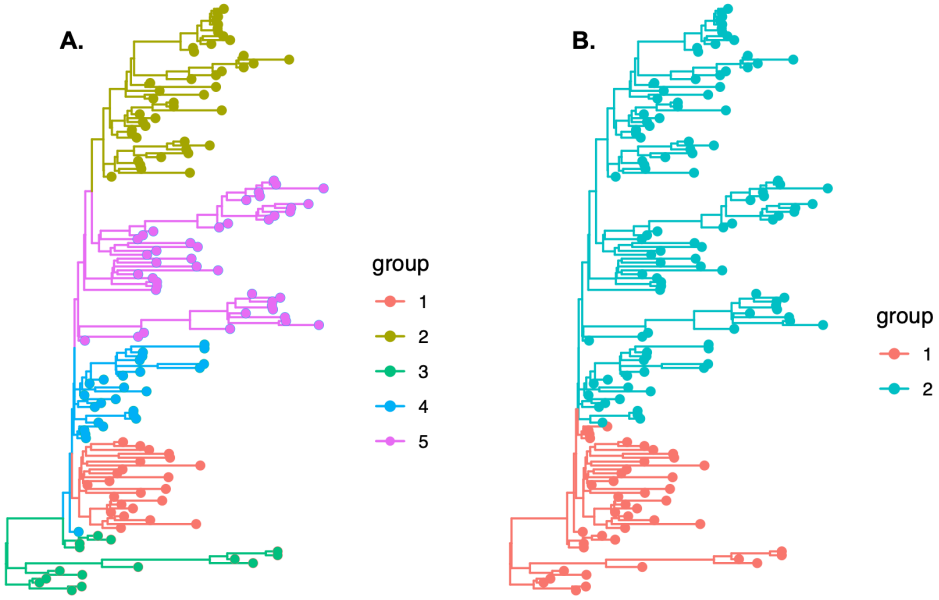

Figure 1: Down-sampled time-scaled phylogenetic tree for Ebola publicly available with 150 sequences. Clusters obtained by running `treestructure` **A.** without the use of node support and **B.** using node support threshold of 95. For both analyses we used a significance level of 0.01 and minimum clade size of 15 sequences. For an example analysing the complete Ebola dataset, a tutorial can be found here.

Online inference by adding new samples to previous a 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. 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 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 its most recent common ancestor in the new phylogeny. A step-by-step tutorial on how to use this feature can be found here.

Acknowledgements

We would like to acknowledge Oliver Stirrup for previous contribution to the package. EV acknowledges support from the UK Health Security Agency CARAA 104683ED "Development of phylogenetic analysis tools to track HIV transmissions for public health surveillance purposes" and CARAA 5126118 "Provision of expert advice and development support for emerging infections genomics and metagenomics analysis". We also thank support from the Wellcome Trust (Investigator Award 220885/Z/20/Z "Population genomic analysis of HIV transmission fitness" awarded to EV).

References

Binney, B. M., Gias, E., Foxwell, J., Little, A., Biggs, P. J., French, N., Lambert, C., Ha, H. J., Carter, G. P., Gyuranecz, M., Pardon, B., De Vliegher, S., Boyen, F., Bokma, J., Krömker, V., Wente, N., Mahony, T. J., Gibson, J. S., Barnes, T. S., … McCulley, M.

(2025). Genomic analysis of the 2017 Aotearoa New Zealand outbreak of *Mycoplasma bovis* and its position within the global population structure. *Frontiers in Microbiology*, *16*, 1600146. https://doi.org/10.3389/fmicb.2025.1600146

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, *3*, 1–27. https://doi.org/10.1080/03610927408827101

Campbell, A. M., Gavilan, R. G., Abanto Marin, M., Yang, C., Hauton, van A., C. and, & Martinez-Urtaza, J. (2024). Evolutionary dynamics of the successful expansion of pandemic *Vibrio parahaemolyticus* ST3 in Latin America. *Nature Communications*, *15*, 7828. https://doi.org/10.1038/s41467-024-52159-y

Fountain-Jones, N. M., Appaw, R. C., Carver, S., Didelot, X., Volz, E., & Charleston, M. (2020). Emerging phylogenetic structure of the SARS-CoV-2 pandemic. *Virus Evolution*, *6*, veaa082. https://doi.org/10.1093/ve/veaa082

Helekal, D., Ledda, A., Volz, E., Wyllie, D., & Didelot, X. (2022). Bayesian inference of clonal expansions in a dated phylogeny. *Systematic Biology*, *71*, 1073–1087. https://doi.org/10.1093/sysbio/syab095

Joseph, S. J., Thomas, J. C., Schmerer, M. W., Cartee, J. C., St Cyr, S., Schlanger, K., Kersh, E. N., Raphael, B. H., Gernert, K. M., & Antimicrobial Resistant Neisseria gonorrhoeae Working Group. (2022). Global emergence and dissemination of *Neisseria gonorrhoeae* ST-9363 isolates with reduced susceptibility to Azithromycin. *Genome Biology and Evolution*, *14*, evab287. https://doi.org/10.1093/gbe/evab287

Reimche, J. L., Clemons, A. A., Chivukula, V. L., Joseph, S. J., Schmerer, M. W., Pham, C. D., Schlanger, K., St Cyr, S. B., Kersh, E. N., Gernert, K. M., & Antimicrobial-Resistant Working Group. (2023). Genomic analysis of 1710 surveillance-based *Neisseria gonorrhoeae* isolates from the USA in 2019 identifies predominant strain types and chromosomal antimicrobial-resistance determinant. *Microbial Genomics*, *9*, mgen001006. https://doi.org/10.1099/mgen.0.001006

Tonkin-Hill, G., Lees, J. A., Bentley, S. D., Frost, S. D. W., & Corander, J. (2019). Fast hierarchical Bayesian analysis of population structure. *Nucleic Acids Research*, *47*, 5539–5549. https://doi.org/10.1093/nar/gkz361

Volz, E. M., Carsten, W., Grad, Y. H., M., D. A., & X., D. (2020). Identification of hidden population structure in time-scaled phylogenies. *Systematic Biology*, *69*, 884–896. https://doi.org/10.1093/sysbio/syaa009