

T-BAS reference tree datasets

Ignazio Carbone and James B. White, 2025

Center for Integrated Fungal Research, Department of Entomology and Plant Pathology,
North Carolina State University, Raleigh, NC 27695, USA

Corresponding author: Ignazio Carbone (icarbon@ncsu.edu)

DOI: <https://doi.org/10.52750/634503>

Platform URL: <https://tbas.cifr.ncsu.edu/>

Cite as: Carbone, I., & White, J.B. (2025). T-BAS reference tree datasets [Data set]. T-BAS.
<https://doi.org/10.52750/634503>

Abstract

T-BAS is a curated repository of reference tree datasets that supports phylogeny-based placement, taxonomic interpretation, and reproducible comparative analyses for DNA, RNA, and protein sequences. The resource centers on versioned reference trees, alignments, accessioned datasets, and linked specimen, strain, voucher, and study metadata that can be cited and reused across studies. In Version 2.4, T-BAS is positioned explicitly as a sustainable NC State-hosted data repository within the Center for Integrated Fungal Research (CIFR), providing stable, citable reference datasets for barcode, multilocus, and other phylogeny-based workflows while retaining integrated tools for placement, visualization, and downstream analysis.

Document type: Data repository record

This document describes the T-BAS reference tree repository and associated tools. The canonical dataset citation is: Carbone, I., & White, J.B. (2025). T-BAS reference tree datasets [Data set]. T-BAS. <https://doi.org/10.52750/634503>

Features and capabilities

T-BAS v2.4 organizes phylogenetically informed reference resources as accessioned, versioned datasets that can be searched, cited, and reused for taxon identification, sequence placement, and comparative analyses across a wide range of marker systems and taxonomic groups. Each dataset combines a curated reference tree with its associated alignment and linked biological metadata, allowing users to place single-locus or multilocus query sequences in the context of an explicitly defined and reusable reference framework. This repository-centered design supports transparent comparisons across studies because the underlying tree dataset, not just the software environment, is preserved as the primary research object.

In addition to serving as a data repository, T-BAS provides web-based tools for interactive use of these reference tree datasets. Users can visualize placements alongside existing diversity, inspect alignments, and launch downstream analyses such as phylogeny-based community summaries, UniFrac distance calculations (Lozupone and Knight 2005), network and clustering views based on gene genealogies and graph analyses (Clement et al. 2000; Templeton et al. 1992; Hagberg et al. 2008), ClaaTU lineage-level association analyses (Washburne et al. 2017), and PopPhy-CNN deep-learning analyses (Reiman et al. 2020). Trees, metadata tables, and geographic maps of

sampling localities can be exported in multiple formats for publication, archiving, or reuse in external software environments.

The repository maintains stable accessions and citations for curated reference tree datasets, enabling analyses to be reproduced and compared through time as the collection grows. T-BAS v2.4 builds on earlier T-BAS releases (Carbone et al. 2017; Carbone and White 2019) by emphasizing durable dataset stewardship, expanded reference coverage, and sustainable institutional hosting at NC State. All source code is maintained in a public GitHub repository, and Dockerized deployments can be mirrored internally or in external cyberinfrastructure such as the CyVerse Discovery Environment (Swetnam et al. 2024), strengthening resilience and long-term access to both the repository and its associated tools.

T-BAS v2.4 operates on dedicated compute and storage infrastructure at NC State that is maintained for teaching and research, providing long-term stability, regular backups, and high availability. Integration with the DeCIFR platform further supports movement between data-processing workflows and the T-BAS repository, allowing users to generate, place, visualize, and export results within a connected biodiversity informatics environment.

Availability

DOI: <https://doi.org/10.52750/634503>

Source code and documentation: <https://github.com/ncsu-decifr>

Repository: <https://tbas.cifr.ncsu.edu/>

References

- Carbone, I., White, J.B., Miadlikowska, J., Arnold, A.E., Miller, M.A., Kauff, F., et al. (2017). T-BAS: Tree-Based Alignment Selector toolkit for phylogenetic inference. *Bioinformatics* 33(8): 1160–1168.
- Carbone, I., and White, J.B. (2019). T-BAS Version 2.1: Tree-Based Alignment Selector toolkit for phylogenetic placement and visualization. *Microbiology Resource Announcements* 8(30): e00328-19.
- Lozupone, C., and Knight, R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology* 71(12): 8228–8235.
- Washburne, A.D., Silverman, J.D., Leff, J.W., Bennett, D.J., Darcy, J.L., Mukherjee, S., et al. (2017). Phylogenetic factorization of compositional data yields lineage-level associations in microbiome datasets. *mBio* 8(1): e01391-16. doi:10.1128/mBio.01391-16.
- Reiman, D., Metwally, A.A., Sun, J., and Dai, Y. (2020). PopPhy-CNN: A phylogenetic tree embedded architecture for convolutional neural networks to predict host phenotype from metagenomic data. *IEEE Journal of Biomedical and Health Informatics* 24(10): 2993–3001. doi:10.1109/JBHI.2020.2993761.
- Clement, M., Posada, D., and Crandall, K.A. (2000). TCS: a computer program to estimate gene genealogies. *Molecular Ecology* 9(10): 1657–1660.
- Templeton, A.R., Crandall, K.A., and Sing, C.F. (1992). A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. *Genetics* 132(2): 619–633.
- Hagberg, A.A., Schult, D.A., and Swart, P.J. (2008). Exploring network structure, dynamics, and function using NetworkX. In: Varoquaux, G., Vaught, T., and Millman, J. (eds), *Proceedings of the 7th Python in Science Conference (SciPy2008)*, Pasadena, CA, pp. 11–15.
- Swetnam, T.L., et al. (2024). CyVerse: Cyberinfrastructure for open science. *PLOS Computational Biology* 20(1): e1011270. doi:10.1371/journal.pcbi.1011270.