TreeProfiler: A command-line tool for computing and visualizing phylogenetic profiles against large trees

Ziqi Deng[1], Ana Hernández-Plaza[1], and Jaime Huerta-Cepas[1,*]

[1]Centro de Biotecnología y Genómica de Plantas, Universidad Politécnica de Madrid (UPM) and Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA-CSIC), 28223 Madrid, Spain.

* To whom correspondence should be addressed. Tel: +34 910679202; Email: j.huerta@csic.es

ABSTRACT

Phylogenetic profiling is a common bioinformatics approach used to study the distribution of traits across species or, more broadly, along the topology of a phylogenetic tree. Originally proposed as a technique for establishing functional relationships between various genes through their correlated presence/absence patterns, phylogenetic profiling is currently considered a powerful exploration tool for investigating the evolutionary distribution of nearly any trait across both gene or species phylogenies. However, computing and visualizing phylogenetic profiles based on large trees is a laborious task. First, traits need to be mapped against tree nodes, which usually requires the implementation of *ad hoc* scripts or substantial manual work. Second, tree mapping is often limited to leaf nodes, neglecting the usefulness of profiling traits against internal tree nodes. Lastly, graphical exploration of large phylogenetic tree-based profiles is challenging, as plotting usually requires complex manual adjustments and most visualization tools cannot scale to the size and complexity of current datasets. To address these limitations, we have developed TreeProfiler, a command line tool designed to facilitate both the automated annotation of large trees using a wide range of data sources, and the visual exploration of these annotations as phylogenetic profiles. TreeProfiler enables seamless annotation of massive trees and data tables, including built-in support for functional data, taxonomic information, multiple sequence alignments and protein domain architectures. Importantly, TreeProfiler expands its annotation capabilities beyond tree leaves, enabling the annotation of internal nodes with summarized or consensus information under each branch. Additionally, TreeProfiler offers a predefined set of graphical layouts that are based on the ETE Toolkit v4.0 library. TreeProfiler offers unique capabilities that extend and facilitate the use of phylogenetic profiling techniques for a broad landscape of scenarios and for a wide community of researchers interested in evolutionary analysis. To demonstrate its value, we showcase TreeProfiler by automatically profiling multiple data sources across a large bacterial phylogeny including 45,555 species, and by exploring functional and taxonomic annotations over a large gene family tree. TreeProfiler is available at https://github.com/compgenomicslab/TreeProfiler.

KEYWORDS

phylogenetic profiling; phylogenetics; tree annotation; tree visualization;

INTRODUCTION

Phylogenetic profiling is a computational approach by which the presence/absence pattern of a given trait is calculated across various species. The technique was originally proposed by Pellegrini et al (Pellegrini et al., 1999) as a method to infer functional associations between two proteins (whose presence/absence calls were considered as traits), as having correlated phylogenetic profiles across different species is indicative of co-evolution and therefore potential functional interactions.

Since its original formulation, many implementations of the phylogenetic profiling method have been developed (Pellegrini, 2012), including more advanced approaches that incorporate trees to the analysis of profiles (Barker & Pagel, 2005). Tree-based profiling allowed researchers to extend the applications of the methodology, identify specific patterns of gene gain and loss, and calculate the likelihood of coevolution more accurately (Barker et al., 2007).

Today, phylogenetic profiling is not only used to correlate the presence/absence distribution of genes across species to infer functional associations (Dembech et al., 2023; Dey et al., 2015; Szklarczyk et al., 2023), but also to study the evolutionary patterns of nearly any trait that can be associated with phylogenetic tree nodes. Thus, phylogenetic profiles are used to identify synapomorphic traits (Peng et al., 2023), horizontal transfer events (Ravenhall et al., 2015), convergent evolution (Rey et al., 2018) and more. In microbial ecology, phylogenetic profiling is also commonly used to compare the relative abundances of different organisms and gene functions in a metagenome among different samples (Asnicar et al., 2015). In addition, phylogenetic profiles can be computed based on gene phylogenies rather than species trees, allowing to investigate many aspects of gene family evolution, such the emergence of gene function and the identification of neo-/sub-functionalization events (Dey et al., 2015).

Overall, phylogenetic profiling and all its variations have become a very common approach in comparative genomics. In fact, tree-based phylogenetic profiling is considered a powerful exploration tool for multiple research areas (Moi & Dessimoz, 2023). This has led to the development of numerous bioinformatic tools that assist in both mapping traits against tree nodes (also known as tree-annotation), and the graphical exploration of these features over the tree topology. However, given the increased volume and size of phylogenetic trees produced by current genomic and metagenomic projects, tree-based phylogenetic profiling is a laborious and relatively impractical task.

First, annotating large trees with different data sources often requires the development of custom bioinformatic pipelines. There are available bioinformatic tools for this purpose (Huerta-Cepas et al., 2016; Revell, 2012; Stamatakis, 2014; Suchard et al., 2018; Sukumaran & Holder, 2010), but they all demand programming skills and *ad hoc* implementations. Other tools have been developed that provide more straightforward and user-friendly interfaces, but they are usually restricted to specific scenarios such as profiling orthologous groups (e.g.,

eggNOG (Hernández-Plaza et al., 2022), Matreex (Rossier et al., 2023), PhyloProfile (Tran et al., 2018)), domain architectures (e.g., PhyloPro2.0 (Cromar et al., 2016)) or metagenomics profiling (e.g., Graphlan (Asnicar et al., 2015), TAMPA (Sarwal et al., 2023)).

Furthermore, a common limitation of current tools is that they predominantly focus on mapping traits against leaf nodes in the tree; despite the fact that annotating internal nodes would be equally useful for exploration purposes. For instance, while presence/absence matrices are typically linked to the terminal nodes in a tree, it would make most sense to automatically annotate internal branches with the summary or consensus version of the annotations grouped underneath. This idea could actually be applied to more complex annotations and scenarios than just presence/absence calls. Profiling sequence variations (e.g. point mutations), protein domain architectures or gene copy number per species over a gene-tree phylogeny, is also a very useful approach used to investigate the evolution of gene function (Nevers et al., 2017; van Dam et al., 2013). Hence, although these traits are only associated with leaf nodes in a tree (i.e. genes or proteins), they could also be used to infer consensus sequences and domain architectures for each internal branch in the tree, or calculating the average number of genes per species at each clade. In fact, the same concept could be exported to any feature for which a summarized or consensus version can be calculated, like functional terms associated with proteins, gene expression values or ecological aspects of the species represented in each tree.

Second, visualizing phylogenetic profiles over large trees is currently challenging. Different traits might be better represented using different graphical elements, such as text labels, color gradients, bar plots, heatmaps, presence/absence matrices, etc. A configuration that needs to be manually adjusted and adapted on a per-case basis. Bioinformatic software exists that allows for the visualization of annotated trees using custom graphical decorations, both using programmatic methods (Huerta-Cepas et al., 2016; Yu et al., 2017), standalone tools (Cantrell et al., 2021) or online resources (Letunic & Bork, 2021). However, creating custom tree images and visualizing them interactively is not a trivial process and it is often limited by the size of the underlying tree and the nature of annotations. Often, large annotated trees, containing tens of thousands nodes, fail at rendering or cannot be interactively explored. In addition, complex features such as protein domain architectures or sequence alignments require creating *ad hoc* graphical representations which are not available in most tools.

We also argue that the annotation of internal tree nodes, as explained above, is particularly critical when visualizing phylogenetic profiles based on large phylogenies, as leaf nodes are often not visible in the tree images produced. As a result, leaf node annotations (the actual phylogenetic profiles) are either hidden or shown as an superimposed graphical representation of multiple elements in the tree image, which does not account for an accurate consensus of the features found under each collapsed node.

Here, we addressed all these limitations by implementing TreeProfiler, a general purpose command line tool designed to streamline the annotation of large trees using a variety of source data and to facilitate their interactive visualization. TreeProfiler is available at

https://github.com/compgenomicslab/TreeProfiler.

IMPLEMENTATION AND MAIN FEATURES

TreeProfiler is implemented as a command-line tool that allows users to map any kind of data onto a phylogenetic tree, as well as interactively visualizing the resulting annotated trees. These two tasks are performed by means of two sub-commands: *treeprofiler-annotate* and *treeprofiler-plot*, respectively.

Computing phylogenetic profiles

The command *treeprofiler-annotate* is in charge of reading raw data tables and annotations and mapping them onto a tree structure, producing a tree-based phylogenetic profile for each annotation type or data column. *Treeprofiles-annotate* requires an input tree in Newick format and one or more data tables in tab-separated values format (TSV). TreeProfiler analyzes all columns in the data tables provided, automatically inferring their type and binding their values to the corresponding nodes in the tree(s). Each column in the data tables is considered an independent trait, and their values will be profiled against the different tree nodes. The only requirement is that the first column in each data table corresponds to a node name in the target tree.

Although the data type of each column in the annotation tables is automatically inferred, users have the flexibility to control how different columns in the input data tables should be interpreted. As for the current version, TreeProfiler can handle data columns containing numeric values, text labels, boolean values or even categorical multi-value items. For convenience, text-based columns can be automatically converted into presence-absence matrices, where each categorical value will become a boolean column in the annotation matrix. Importantly, TreeProfiler will automatically annotate internal tree nodes by summarizing the values under each branch taking into account the data type of each annotation. For instance numerical values will be summarized using descriptive statistics, while categorical values will be summarized using relative frequencies and term counters.

Given their overall value for comparative genomic analyses, various special data types are also built-in supported in the annotating process, namely: taxonomic annotations, functional terms, sequence alignments and domain architectures. For instance, TreeProfiler can automatically interpret taxonomic annotations based on NCBI taxonomy IDs (Sayers et al., 2018) and GTDB (Parks et al., 2021), which will be then used to infer full lineage tracks, species names and taxonomic ranks. Moreover, taxonomic annotations will be used to deduce the last common ancestor of every branch in the tree and automatically annotate internal nodes accordingly.

Similarly**,** *treeprofiler-annotate* can seamlessly integrate with the functional annotations and protein domain coordinates generated by eggNOG-mapper (Cantalapiedra et al., 2021), a widely used tool for functional annotation of genomes and metagenomes. This allows users to quickly explore gene phylogenies together with the distribution of, for instance, Gene

Ontology terms, enzymatic annotations, orthology assignments or Pfam (Mistry et al., 2020) and SMART (Letunic et al., 2020) protein domain architectures. Additionally, multiple sequence alignments can be seamlessly associated with their respective phylogenies, which will trigger the automatic annotation of consensus sequences for every internal node in the tree.

As a result, *treeprofiler-annotate* will produce a new annotated tree file where all annotations are already processed and encoded, facilitating the posterior analysis and visualization of each data source as a tree-based phylogenetic profile. The resulting annotated tree file is, by default, exported using the ETE Toolkit (Huerta-Cepas et al., 2016) v4.0 format, which allows users loading and operating with the data using Python scripts. However, annotated trees can also be exported as Extended Newick Format, so they can be processed by other software.

Visualizing phylogenetic profiles

The *treeprofiler-plot* command can be used to visually explore any annotated tree produced by the *treeprofiler-annotate* program, automatically adjusting the most suitable graphical representation for each data source. Alternatively, users can easily adjust which annotations should be shown and choose among different predefined layouts.

The *treeprofiler-plot* command starts an interactive interface to visualize the target tree with annotations. It provides a control panel (Figure 1A) for customizing the visualization layout and related features, where users can browse the topology of the tree (Figure 1B) and perform actions on specific nodes, including collapsing, pruning, rooting and other per-node options (Figure 1C). With TreeProfiler, tree annotations based on numerical data can be easily converted into text strings, heatmaps, barplots, color gradients, or symbols. Categorical values can be represented as branches or node labels with autocoloring scales, presence-absence matrices (converting each value into a column in the matrix), or colored rectangles (Figure 1D). Importantly, the annotations of internal nodes, visible only when collapsing branches, will be also shown whenever leaf nodes are not visible, providing graphical representation of the collapsed nodes. This will translate into showing the average, maximum, or minimum values of numeric data types; or pie chart distributions, horizontal stacked bars and color gradients for categorical data types (Figure 1E). Moreover, special annotation types have their own dedicated layouts and graphical style configurations. Thus, taxonomic classifications, including common ancestry, can be represented as vertical color bands aligned with their corresponding clades, providing a clear visual depiction of taxa distribution over the topology of the trees. Similarly, multiple sequence alignments and protein domain architectures will be automatically plotted using *ad hoc* graphical representations (Figure 1F).

Although node collapsing is automatically handled by the tree viewer in large trees, TreeProfiler allows users to have finer controls on how to collapse, prune or highlight specific branches on the tree, based on custom and more advanced criteria. Interestingly, such

criteria can involve the annotations present in the tree. For instance, users can force collapsing a large annotated tree at an specific taxonomic rank (e.g., phylum), which will be automatically inferred out of taxonomic annotations of leaf nodes. Similarly, users can highlight or prune specific parts of the tree based on custom annotation filters (examples are provided in the online documentation of TreeProfiler).

Finally, we consider it important to note that all the visualization options provided by TreeProfiler are implemented based on the latest version of the ETE Toolkit (v4), which provides support for efficient and interactive visualization of very large trees (e.g., >10,000 tips).

Figure 1

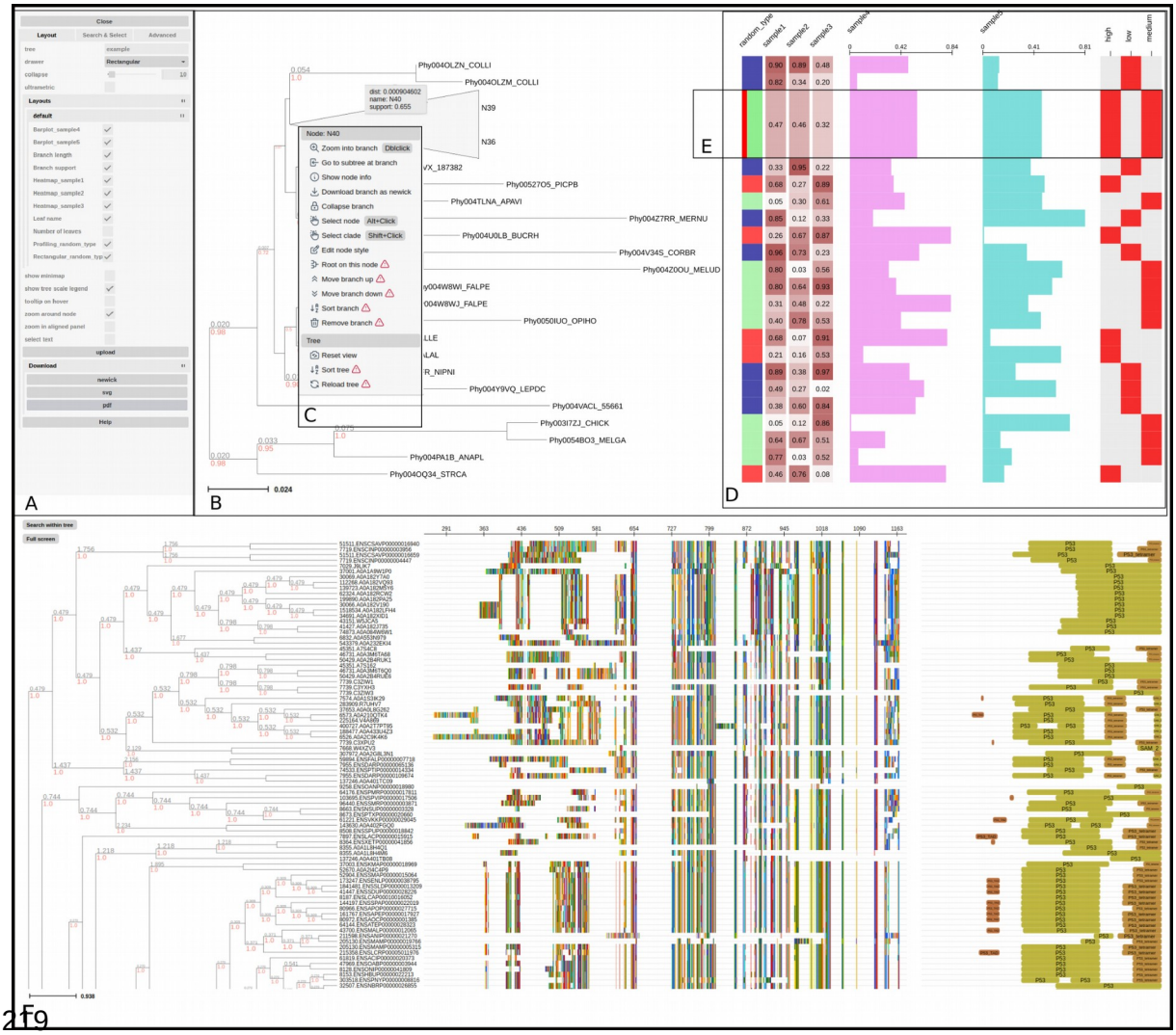


Figure 1. Overview of the treeprofiler-plot visualization interface A) The control panel for customizing visualization layout and features, providing also search options. **B)** Phylogenetic tree representation with support values (red) and branch distance (gray) displayed on top of branches. **C)** Right click on a tree node will show a context menu showing the node option panel, which offers collapsing, pruning, rooting and other per-node options. **D**) Various tree node annotations shown as

phylogenetic profiles. Each column represents a different data source (i.e. trait) displayed using different layouts, namely: rectangle (first column), binary (second column), heatmap (third to fifth columns), barplot (sixth and seventh column) and presence-absence matrices (last three columns). The name of each data source is shown as a column header. **E)** Summarized annotations of a collapsed branch. Note that horizontal stacked bars (column 1) represent the frequency of each categorical value found in the descendant nodes. Similarly, the heatmap columns represent average values from descendant nodes, and the presence-absence matrix shows the consensus value of descendant nodes. **F)** Another phylogenetic tree annotated with a multiple sequence alignment and domain architecture data, which are visualized using built-in TreeProfiler layouts. Data used to generate this figure and its panels are available as supplementary data.

USE CASE EXAMPLES

Automated annotation of GTDB species phylogeny

To illustrate the easiness and flexibility of TreeProfiler, we used it to annotate and visualize the release 202 of the GTDB bacterial phylogeny (Parks et al., 2021), which includes 45,555 representative bacterial species. The GTDB database provides the phylogenetic tree in plain Newick format, together with a massive data table including hundreds of columns informing different aspects of the genomes used to build the tree. Moreover, the ProGenomes database (Fullam et al., 2023; Parks et al., 2021) provides habitat information for many of the species included in the GTDB tree.

Hence, we used TreeProfiler to annotate the GTDB tree with both GTDB metadata and proGenomes data, which was performed with a single command line:

```
treeprofiler annotate \
      --tree bac120_r202.nw \
      --input-type newick \
      --metadata bac120_metadata_r202.tar.gz progenome3.tar.gz \
      --taxonomic-profile \
      --taxadb GTDB \
      --outdir ./
```

This command generated a new annotated tree (*bact1202_r202_annotated.ete*) where all data sources included in the metadata files are now profiled along the tree nodes, including summarized and consensus annotations of internal nodes.

Next, to visually explore such profiles, we used the following command:

```
treeprofiler plot \
      --tree bac120_r202_annotated.ete \
      --input-type ete \
      --barplot-layout genome_size,protein_count \
      --heatmap-layout gc_percentage \
      --binary-layout aquatic_habitat host_associated soil_habitat \
      --rectangle-layout ncbi_assembly_level ncbi_genome_category \
      --taxonclade-layout
```

Note that we specifically selected which data sources (profiles) should be visualized and under which graphical layouts. The resulting tree image (Figure 2), shows the annotations across all visible tree nodes, allowing to interactively browse, zoom, search and collapse branches along this large tree topology.

Figure 2

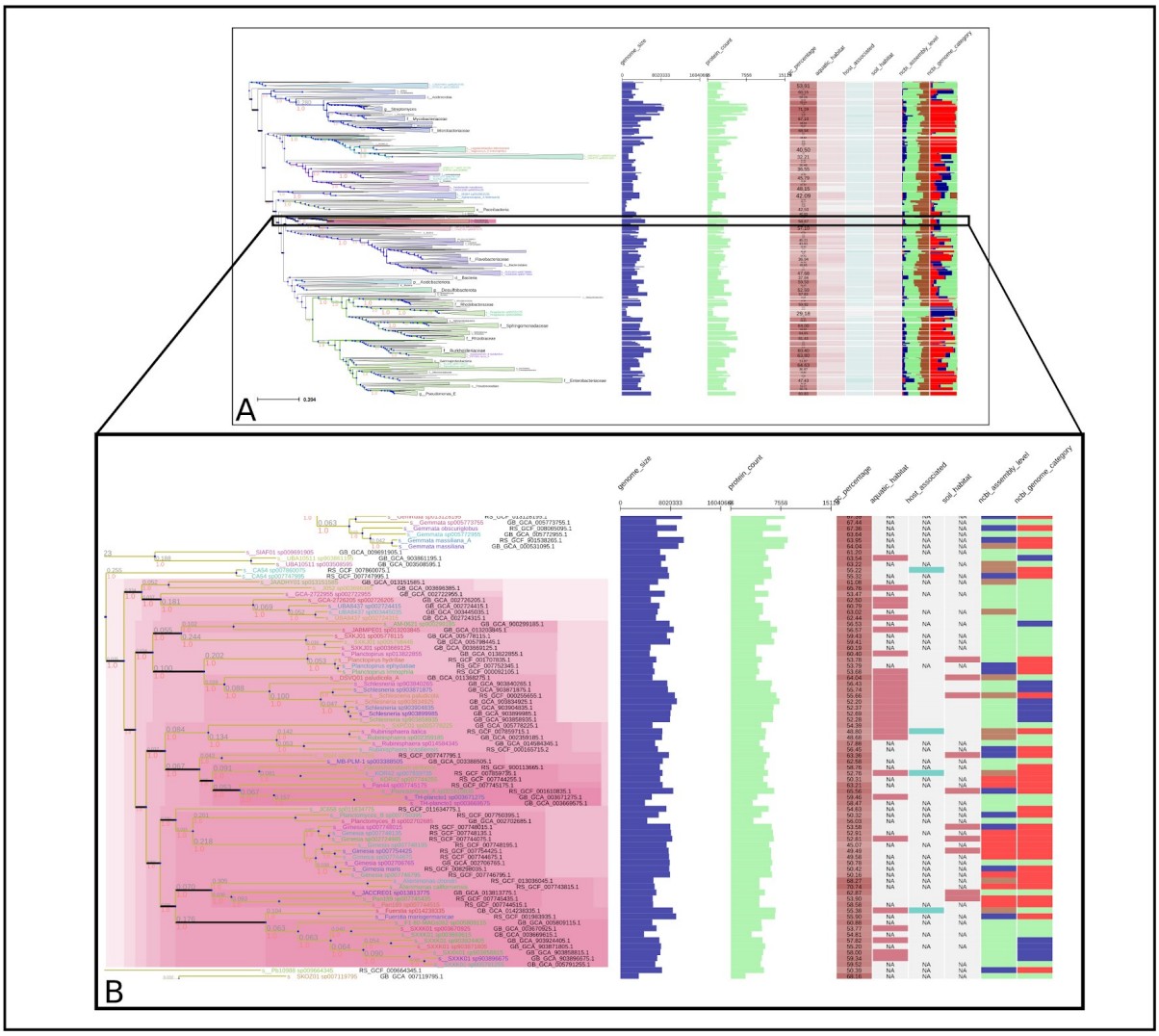


Figure 2. GTDB r202 taxonomy tree (bacteria, 45,555 leaves) represented in a rectangular tree layout and annotated with various data sources. In the example shown, *genome_size* and *protein_count* are visualized as barplots (columns 1 and 2), *gc_percentage* is shown as a heatmap in column 3, habitat information is shown in columns 4,5 and 6, and *ncbi_assembly_level* and *ncbi_genome_category* are visualized as rectangle layouts in columns 7 and 8. Tree branches are automatically colored based on the inferred taxonomic rank. **Panel A** shows a view of the entire GTDB tree with all the layouts activated, where only internal tree branches are visible due to the tree size and current zoom level. All phylogenetic profiles shown are therefore consensus or summary representations of the annotations of nodes underneath. **Panel B** shows an expanded version of the profiles, which are dynamically adjusted to the zoomed region.

Functional and taxonomic profiling of the NifH gene family

To demonstrate the capabilities of TreeProfiler in annotating large gene phylogenies, we analyzed the nitrogenase iron protein NifH gene family by using its corresponding phylogenetic in the eggNOG6 database (Hernández-Plaza et al., 2022), which contains 2,152 protein members. To perform the analysis, we annotated all protein sequences included in the nifH tree using the online version of eggNOG-mapper, a tool for functional annotation based on precomputed orthology assignments. Next we used *treeprofiler-annotate* to bind the functional and domain annotations obtained from eggNOG-mapper with the NifH phylogeny using a single command line:

```
treeprofiler annotate \
      --tree nifH.nw \
      --input-type newick \
      --emapper-annotation nifH.out.emapper.annotations \
      --emapper-pfam nifH.out.emapper.pfam \
      --alignment nifH.faa.aln \
      --taxonomic-profile \
      --taxadb NCBI \
      --taxon-delimiter . \
      --taxa-field 0 \
      --outdir ./
```

Note that besides providing the NifH tree as input, together with the eggNOG-mapper output files, we also provided the multiple sequence alignment used to generate the tree. This allows TreeProfiler to automatically compute domain positions in the sequence alignment and compute sequence consensus. As in previous examples, the functional annotations mapped to leaf nodes were used to automatically annotate internal branches of the tree, where the prevalence of each term was calculated.

Finally, to visually explore all functional annotations as tree profiles, we used the following *treeprofiler-plot* command:

```
treeprofiler plot \
      --tree nifH_annotated.ete \
      --input-type ete \
      --emapper-layout \
      --domain-layout \
      --taxonclade-layout \
```

Here, we just provided the annotated tree as input and activated the special *emapper*, *domain* and *taxonclade* layouts, which produced the tree images shown in Figure 3.

Figure 3

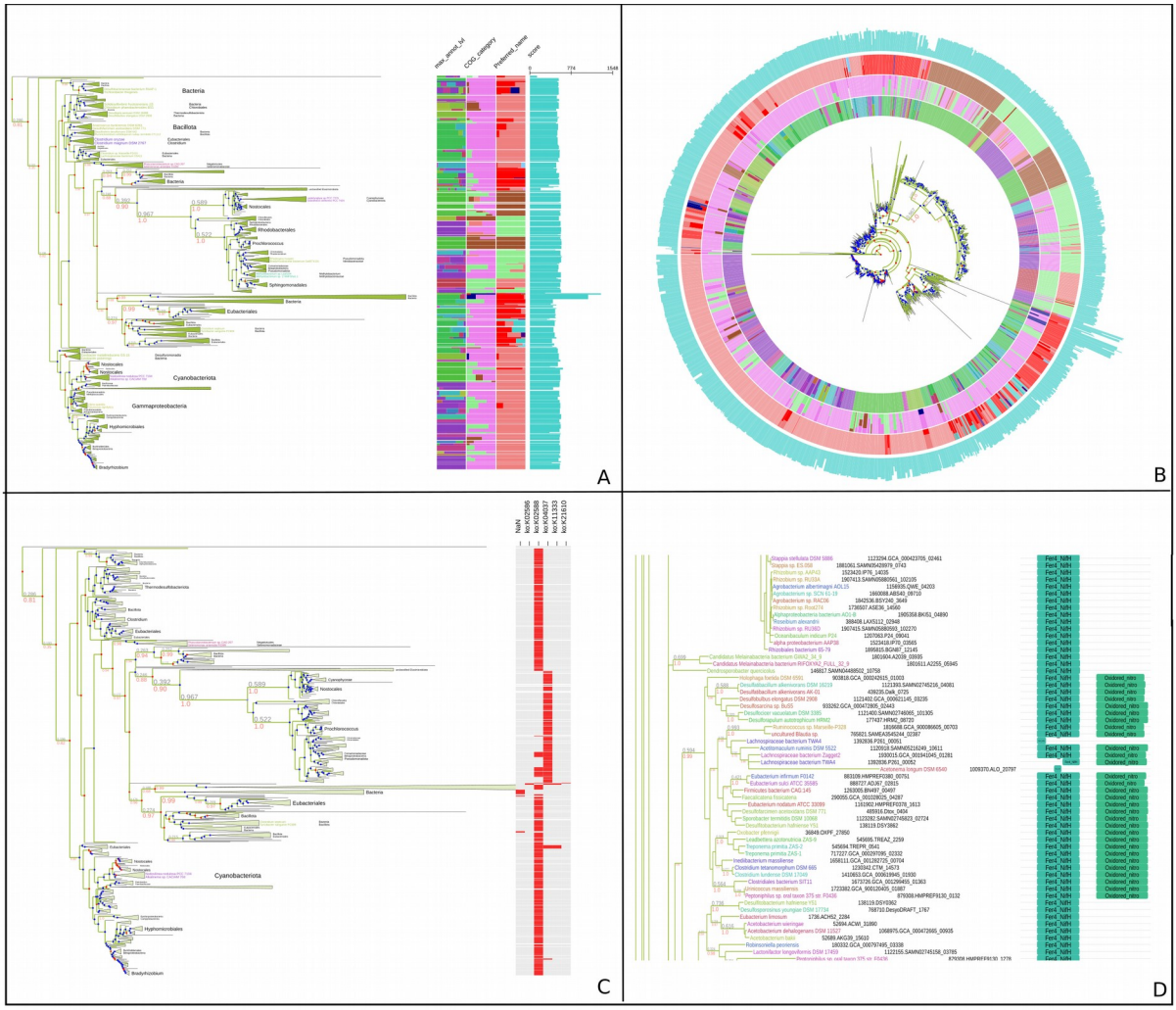


Figure 3. Functional and taxonomic gene annotations profiled against the NifH phylogeny. All panels show the NifH gene family phylogeny. **Panel A and B**, show the profiles of functional annotations, including COG categories and inferred gene names, in rectangular and circular mode, respectively. **Panel C** shows the phylogenetic profile of KEGG KO terms across different parts of the tree. Note that KEGG KO annotations from eggNOG-mapper results are automatically converted into a presence-absence matrix, which facilitates the visualization of each KO distribution. **Panel D** demonstrates the visualization of protein domain architecture along the tree, which was automatically inferred from eggNOG-mapper output files.

CONCLUSIONS

Phylogenetic profiling is a common and useful technique in comparative genomics. However, its application on large trees is not straightforward, especially when multiple traits are to be profiled and visualized. Our TreeProfiler tool provides a general and simple way of profiling nearly any kind of trait against large target phylogenies. It includes novel features like the automated annotation of internal tree nodes and interactive visualization of results. The whole process of computing and visualizing custom phylogenetic profiles can be performed with two simple command lines, avoiding the implementation of *ad hoc* pipelines and enabling fast and efficient exploration of large phylogenetic datasets. Additional examples and detailed

documentation can be found at https://github.com/compgenomicslab/TreeProfiler.

AUTHOR CONTRIBUTIONS

Z.D and J.HC conceived the project, designed the methodology and wrote the manuscript. Z.D wrote the programming code, testing and software documentation. A.HP contributed with the domain annotation algorithm. All authors read and approved the final manuscript.

ACKNOWLEDGEMENTS

The authors want to thank Jorge Botas and Jordi Burguet-Castell for their invaluable help developing and debugging many features of the ETE Toolkit software that are used by TreeProfiler. This study was supported by FPI-Severo Ochoa predoctoral fellowship [SEV-2016-0672-18-2:PRE2018-084075] to Z.D; Research Technical Support Staff Aid [PTA2019-017593-I/AEI/10.13039/501100011033] to A.HP; CZI grant [DAF2020-218584] DOI https://doi.org/10.37921/207480vhbyqe from the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation (funder DOI 10.13039/100014989); and the grant PID2021-127210NB-I00 MCIU/AEI/FEDER, UE, National Programme for Fostering Excellence in Scientific and Technical Research, to J.HC;

Conflict of interest statement

None declared.

REFERENCES

Asnicar, F., Weingart, G., Tickle, T. L., Huttenhower, C., & Segata, N. (2015). Compact graphical representation of phylogenetic data and metadata with GraPhlAn. *PeerJ*, *3*, e1029. https://doi.org/10.7717/peerj.1029

Barker, D., Meade, A., & Pagel, M. (2007). Constrained models of evolution lead to improved prediction of functional linkage from correlated gain and loss of genes. *Bioinformatics* , *23*(1), 14–20. https://doi.org/10.1093/bioinformatics/btl558

Barker, D., & Pagel, M. (2005). Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. *PLoS Computational Biology*, *1*(1), e3. https://doi.org/10.1371/journal.pcbi.0010003

Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., & Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution*, *38*(12), 5825–5829. https://doi.org/10.1093/molbev/msab293

Cantrell, K., Fedarko, M. W., Rahman, G., McDonald, D., Yang, Y., Zaw, T., Gonzalez, A., Janssen, S., Estaki, M., Haiminen, N., Beck, K. L., Zhu, Q., Sayyari, E., Morton, J. T., Armstrong, G., Tripathi, A., Gauglitz, J. M., Marotz, C., Matteson, N. L., … Knight, R. (2021). EMPress Enables Tree-Guided, Interactive, and Exploratory Analyses of Multi-omic Data Sets. *mSystems*, *6*(2). https://doi.org/10.1128/mSystems.01216-20

Cromar, G. L., Zhao, A., Xiong, X., Swapna, L. S., Loughran, N., Song, H., & Parkinson, J. (2016). PhyloPro2.0: a database for the dynamic exploration of phylogenetically conserved proteins and their domain architectures across the Eukarya. *Database: The Journal of Biological Databases and Curation*, *2016*, baw013. https://doi.org/10.1093/database/baw013

Dembech, E., Malatesta, M., De Rito, C., Mori, G., Cavazzini, D., Secchi, A., Morandin, F., & Percudani, R. (2023). Identification of hidden associations among eukaryotic genes through statistical analysis of coevolutionary transitions. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(16), e2218329120. https://doi.org/10.1073/pnas.2218329120

Dey, G., Jaimovich, A., Collins, S. R., Seki, A., & Meyer, T. (2015). Systematic Discovery of Human Gene Function and Principles of Modular Organization through Phylogenetic Profiling. *Cell Reports*, *10*(6), 993–1006. https://doi.org/10.1016/j.celrep.2015.01.025

Fullam, A., Letunic, I., Schmidt, T. S. B., Ducarmon, Q. R., Karcher, N., Khedkar, S., Kuhn, M., Larralde, M., Maistrenko, O. M., Malfertheiner, L., Milanese, A., Rodrigues, J. F. M., Sanchis-López, C., Schudoma, C., Szklarczyk, D., Sunagawa, S., Zeller, G., Huerta-Cepas, J., von Mering, C., … Mende, D. R. (2023). proGenomes3: approaching one million accurately and consistently annotated high-quality prokaryotic genomes. *Nucleic Acids Research*, *51*(D1), D760–D766. https://doi.org/10.1093/nar/gkac1078

Hernández-Plaza, A., Szklarczyk, D., Botas, J., Cantalapiedra, C. P., Giner-Lamia, J., Mende, D. R., Kirsch, R., Rattei, T., Letunic, I., Jensen, L. J., Bork, P., von Mering, C., & Huerta-Cepas, J. (2022). eggNOG 6.0: enabling comparative genomics across 12 535 organisms. *Nucleic Acids Research*, *51*(D1), D389–D394. https://doi.org/10.1093/nar/gkac1022

Huerta-Cepas, J., Serra, F., & Bork, P. (2016). ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Molecular Biology and Evolution*, *33*(6), 1635–1638.

https://doi.org/10.1093/molbev/msw046

Letunic, I., & Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Research*, *49*(W1), W293–W296. https://doi.org/10.1093/nar/gkab301

Letunic, I., Khedkar, S., & Bork, P. (2020). SMART: recent updates, new developments and status in 2020. *Nucleic Acids Research*, *49*(D1), D458–D460. https://doi.org/10.1093/nar/gkaa937

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., & Bateman, A. (2020). Pfam: The protein families database in 2021. *Nucleic Acids Research*, *49*(D1), D412–D419. https://doi.org/10.1093/nar/gkaa913

Moi, D., & Dessimoz, C. (2023). Phylogenetic profiling in eukaryotes comes of age [Review of *Phylogenetic profiling in eukaryotes comes of age*]. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(19), e2305013120. https://doi.org/10.1073/pnas.2305013120

Nevers, Y., Prasad, M. K., Poidevin, L., Chennen, K., Allot, A., Kress, A., Ripp, R., Thompson, J. D., Dollfus, H., Poch, O., & Lecompte, O. (2017). Insights into Ciliary Genes and Evolution from Multi-Level Phylogenetic Profiling. *Molecular Biology and Evolution*, *34*(8), 2016–2034. https://doi.org/10.1093/molbev/msx146

Parks, D. H., Chuvochina, M., Rinke, C., Mussig, A. J., Chaumeil, P.-A., & Hugenholtz, P. (2021). GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Research*, *50*(D1), D785–D794. https://doi.org/10.1093/nar/gkab776

Pellegrini, M. (2012). Using Phylogenetic Profiles to Predict Functional Relationships. In J. van Helden, A. Toussaint, & D. Thieffry (Eds.), *Bacterial Molecular Networks: Methods and Protocols* (pp. 167–177). Springer New York. https://doi.org/10.1007/978-1-61779-361-5_9

Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D., & Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(8), 4285–4288. https://doi.org/10.1073/pnas.96.8.4285

Peng, C., Guo, X.-L., Zhou, S.-D., & He, X.-J. (2023). Backbone phylogeny and adaptive evolution of Pleurospermum s. l.: New insights from phylogenomic analyses of complete plastome data. *Frontiers in Plant Science*, *14*, 1148303. https://doi.org/10.3389/fpls.2023.1148303

Ravenhall, M., Škunca, N., Lassalle, F., & Dessimoz, C. (2015). Inferring horizontal gene transfer. *PLoS Computational Biology*, *11*(5), e1004095. https://doi.org/10.1371/journal.pcbi.1004095

Revell, L. J. (2012). phytools: an R package for phylogenetic comparative biology (and other things). In *Methods in Ecology and Evolution* (Vol. 3, Issue 2, pp. 217–223). https://doi.org/10.1111/j.2041-210x.2011.00169.x

Rey, C., Guéguen, L., Sémon, M., & Boussau, B. (2018). Accurate Detection of Convergent Amino-Acid Evolution with PCOC. *Molecular Biology and Evolution*, *35*(9), 2296–2306. https://doi.org/10.1093/molbev/msy114

Rossier, V., Train, C., Nevers, Y., Robinson-Rechavi, M., & Dessimoz, C. (2023). Matreex: compact and interactive visualisation of large gene families provides evidence for loss of intraflagellar transport in a myxozoan. In *bioRxiv* (p. 2023.02.18.529053). https://doi.org/10.1101/2023.02.18.529053

Sarwal, V., Brito, J., Mangul, S., & Koslicki, D. (2023). TAMPA: interpretable analysis and visualization of metagenomics-based taxon abundance profiles. *GigaScience*, *12*, giad008. https://doi.org/10.1093/gigascience/giad008

Sayers, E. W., Cavanaugh, M., Clark, K., Ostell, J., Pruitt, K. D., & Karsch-Mizrachi, I. (2018). GenBank. *Nucleic Acids Research*, *47*(D1), D94–D99. https://doi.org/10.1093/nar/gky989

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* , *30*(9), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., & Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution*, *4*(1), vey016. https://doi.org/10.1093/ve/vey016

Sukumaran, J., & Holder, M. T. (2010). DendroPy: a Python library for phylogenetic computing. *Bioinformatics* , *26*(12), 1569–1571. https://doi.org/10.1093/bioinformatics/btq228

Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., Gable, A. L., Fang,

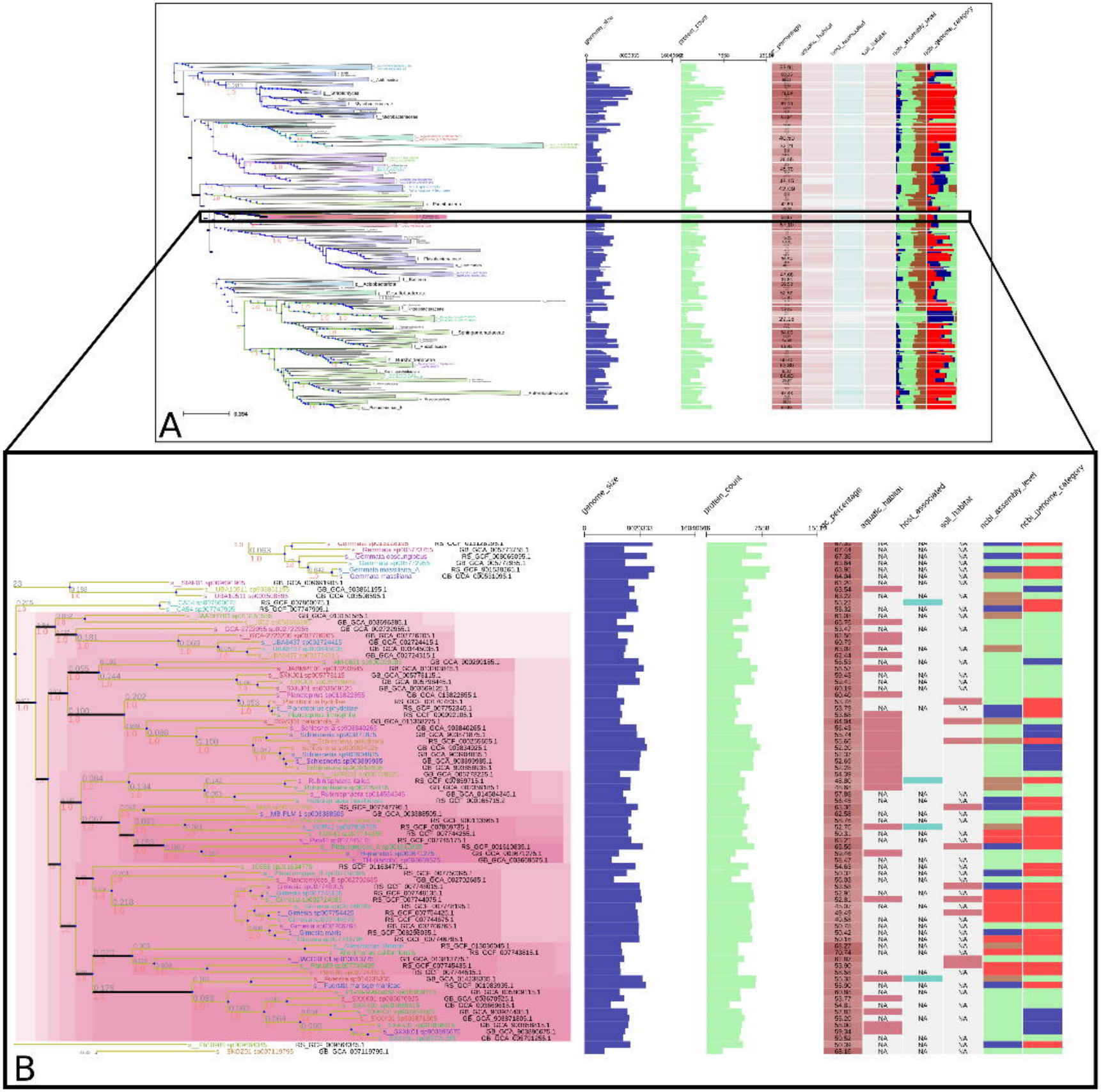
T., Doncheva, N. T., Pyysalo, S., Bork, P., Jensen, L. J., & von Mering, C. (2023). The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, *51*(D1), D638–D646. https://doi.org/10.1093/nar/gkac1000

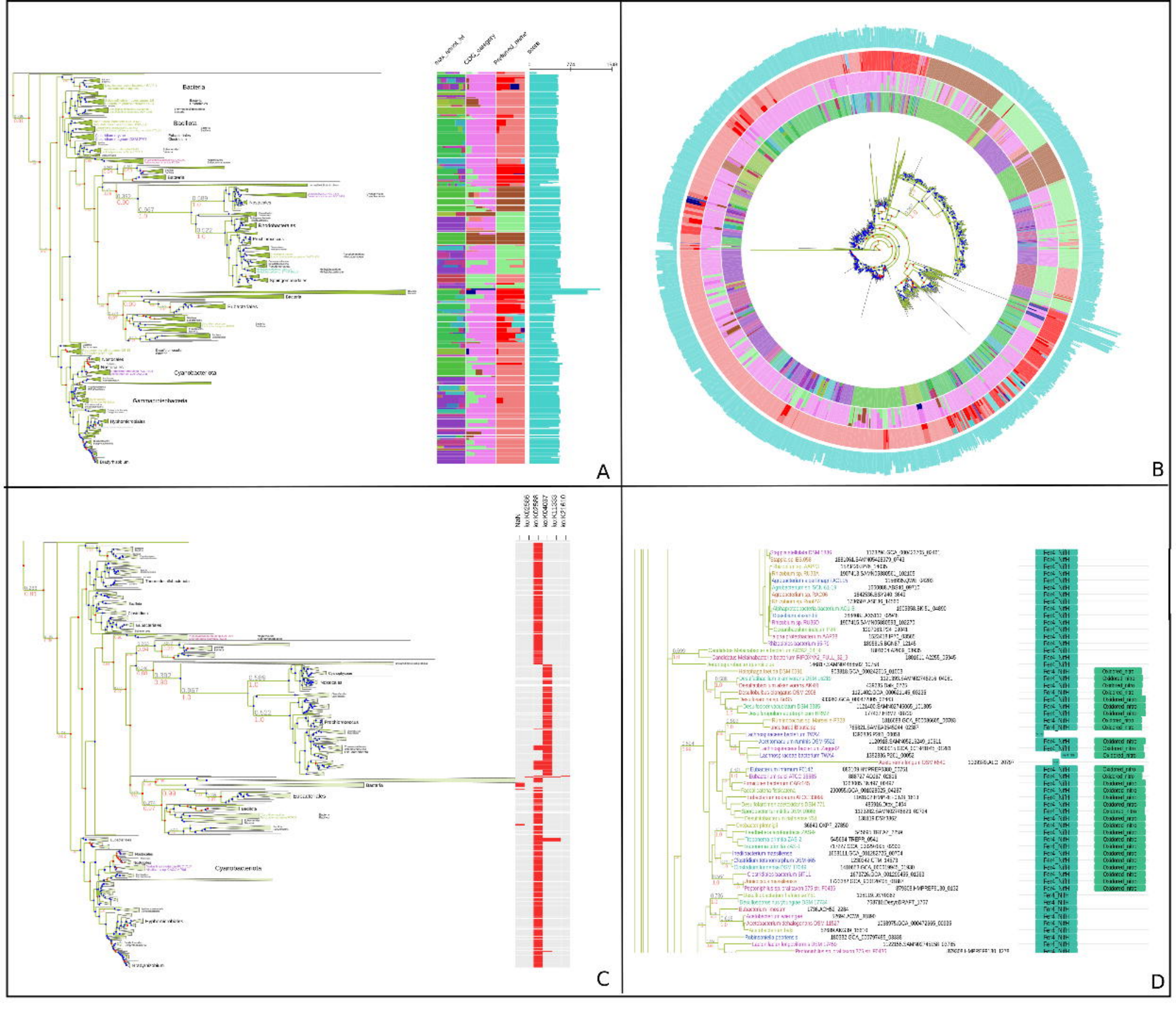
Tran, N.-V., Greshake Tzovaras, B., & Ebersberger, I. (2018). PhyloProfile: dynamic visualization and exploration of multi-layered phylogenetic profiles. *Bioinformatics* , *34*(17), 3041–3043. https://doi.org/10.1093/bioinformatics/bty225

van Dam, T. J. P., Townsend, M. J., Turk, M., Schlessinger, A., Sali, A., Field, M. C., & Huynen, M. A. (2013). Evolution of modular intraflagellar transport from a coatomer-like progenitor. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(17), 6943–6948. https://doi.org/10.1073/pnas.1221011110

Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T.-Y. (2017). Ggtree : An r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution / British Ecological Society*, *8*(1), 28–36. https://doi.org/10.1111/2041-210x.12628

A
B
C
D
E
F
Phy004OLZN_COLLI
Phy004OLZM_COLLI
Phy0052705_PICPB
Phy004TLNA_APAVI
Phy004Z7RR_MERNU
Phy004U0LB_BUCRH
Phy004V34S_CORBR
Phy004Z0OU_MELUD
Phy0050IUD_OPIHO
Phy004Y8VQ_LEPDC
Phy004VACL_55661
Phy003I7ZJ_CHICK
Phy0054BD3_MELGA
Phy004PA1B_ANAPL
Phy004DQ34_STRCA
Node: N40
Zoom into branch
Go to subtree at branch
Show node info
Download branch as newick
Collapse branch
Select node
Select clade
Edit node style
Root on this node
Move branch up
Move branch down
Sort branch
Remove branch
Tree
Reset view
Sort tree
Reload tree
random_type
sample1
sample2
sample3
sample4
sample5
high
low
medium


A
B
genome_size
protein_count
gc_percentage
aquatic_habitat
host_associated
soil_habitat
ncbi_assembly_level
ncbi_genome_category


Bacteria
Cyanobacteria
A
B
C
D