ggtreeExtra: Compact visualization of richly annotated phylogenetic data

Guangchuang Yu (gcyu1@smu.edu.cn)
Southern Medical University
https://orcid.org/0000-0002-6485-8781

shuangbin xu
Zehan Dai
Southern Medical University

Pingfan Guo
Southern Medical University

Xiaocong Fu
Southern Medical University

Shanshan Liu
Southern Medical University

Lang Zhou
Southern Medical University

Wenli Tang
Southern Medical University

Tingze Feng
Southern Medical University

Meijun Chen
Southern Medical University

Li Zhan
Southern Medical University

Tianzhi Wu
Southern Medical University

Erqiang Hu
Southern Medical University

Brief Communication

Keywords:

DOI: https://doi.org/10.21203/rs.3.rs-155672/v1
Abstract

We present the ggtreeExtra package for visualizing heterogeneous data with a phylogenetic tree (https://www.bioconductor.org/packages/ggtreeExtra). It supports more data types and visualization methods than other tools and has many features that are not available elsewhere. The ggtreeExtra package is a universal tool for tree data visualization. It extends the applications of phylogenetic tree in different disciplines by making more domain specific data to be available to visualize and interpret on the evolutionary context.

Introduction

Phylogenetic tree is widely used in a number of biological fields, including comparative genomics, epidemiology and microbiome. Integrating and visualizing phylogenetic tree with multi-dimensional associated data sets helps identifying patterns and generating new hypotheses. For example, a recent research constructed a phylogenetic tree of SARS-CoV-2 and integrated the state information of initial diagnosis of Australian SARS-CoV-2 genomes and country information of the origin of the GISAID genomes to investigate origins and transmission pathways of the COVID-19 strains in Australia\(^1\). The development of high-throughput experimental technologies has expanded the scales of phylogenetic trees and associated data sets. For instance, a microbiome study may collect hundreds of samples and reconstruct a phylogenetic tree that represents the evolutionary relationships of the microbial community composed of thousands of species. Associated data sets, such as the species abundance in each sample and the number or status of target genes of each species, can be incorporated and visualized on a phylogenetic tree to reveal new insights into factors that influence microbial community dynamics\(^2,3,4\).

However, integrating and visualizing multi-dimensional data with phylogenetic tree is still not an easy task. Over the past decade, several packages and web tools have been developed to integrate external data to phylogenetic tree, such as iTOL\(^5\), Evolview\(^6\), Microreact\(^7\), ETE3\(^8\) and GraPhlAn\(^9\). But these tools are mainly developed for certain fields and hard to be applied to other research domains. We previously proposed two general methods for mapping and visualizing associated data on phylogeny and the methods were implemented in ggtree\(^10\). The geom_facet function provided in ggtree\(^11\) employs modular design by separating tree visualization, data integration and graph alignment\(^10\). It allows visualizing multiple associated data sets in different panels and serves as a general tool as there is no prerequisite of the input data type\(^10,12\). As the types and scales of biological data keeps expanding, there is a new challenge to visualize richly layered phylogenetic data in circular layout, which allows more data to be displayed in a given space. However, the geom_facet function does not work with circular layout. To fully extend ggtree to support visualizing phylogenetic data from multiple sources in the big data era especially for circular layout, we developed the ggtreeExtra package. It follows the ideas and designs we proposed for phylogenetic tree data integration and visualization and inherits all the merits and benefits of ggtree\(^11\). The ggtreeExtra package allows progressively representing taxon-specific features on external panels of a phylogenetic tree and helps user to explore and compare different heterogeneous
data sets in evolutionary context. The \textit{ggtreeExtra} package has been released within the Bioconductor project\textsuperscript{13} and it is available at \url{https://bioconductor.org/packages/ggtreeExtra}.

\section*{Results}

The \textit{ggtreeExtra} package implemented a layer function, \texttt{geom\_fruit}, which is a universal function for aligning graphic layer to phylogenetic tree (Supplemental Fig. S1, S2 and Tab S1). It can internally reorder associated data based on the structure of a phylogenetic tree, visualize the data using specific geometric layer function with user provided aesthetic mapping and non-variable setting and the graphic layer will be displayed with the tree side by side (\textit{i.e.}, right hand side for rectangular layout or external ring for circular layout, Supplemental Fig. S3) with perfect alignment. Different data graph layers can be progressively added to a tree. For example, \texttt{geom\_fruit} is able to display a heatmap and a bar plot to the outer rings of an annotated phylogenetic tree to compare microbial abundance across different body sites of human (Supplemental Fig. S7). These two layers aligned automatically to the circular phylogenetic tree and were displayed on different external rings. The number of external rings is not strictly limited and user is free to visualize a number of associated data sets using different geometric layers on different external rings. Each data set is visualized on an independent ring layer and multiple ring layers are stacked on the circular phylogenetic tree. This makes the \textit{ggtreeExtra} package particularly useful for layering different data sets to create highly informative tree graphics. For example, multiple heatmap and bar chart layers were compactly displayed on the circular tree to represent the status of gene, metabolic capacity and genome size of 963 bacteria and archaea species (Supplemental Fig. S8).

Unlike other tools, \textit{ggtreeExtra} was developed based on grammar of graphic\textsuperscript{14} and allows user to map variables of associated data to visual attributes of the outer ring graphic layer with high level of abstraction (Supplemental Fig. S3, S4 and S7). The geometric layers defined in ggplot\textsuperscript{2} and its extensions can be used in the \texttt{geom\_fruit} function. For example, the \texttt{geom\_phylopic} implemented by the \textit{ggimage} package can be used to overlay silhouette images on the external layers to compare morphological characteristics with other attributes (\textit{e.g.} taxonomy order, dietary preferences and environmental variables) (Supplemental Fig. S5B). With this feature, more data types and visualization methods are supported in \textit{ggtreeExtra} compared to other tools, since there are a lot of geometric layers developed by the \textit{ggplot2} community (Supplemental Tabel S1, S2 and Fig. S2). For instance, taxon-specific infographics can be added as insets in \textit{ggtreeExtra} using the \texttt{geom\_plot} layer provided by the \textit{ggpmics} package (Supplemental Fig. S5A). The \textit{ggtreeExtra} package makes no assumption about user data. Given a suitable geometric layer, \textit{ggtreeExtra} can be used to incorporate and visualize any kind of information with a tree. This unique feature guarantees versatility of \textit{ggtreeExtra} and makes it easy to represent heterogeneous data from different disciplines.
One unique advantage of circular layout is to create chord diagram to reveal complex relationships. Couple with the inward circular tree layout supported by ggtree\textsuperscript{11}, ggtreeExtra allows displaying flows or connections between taxa, such as syntenic linkage among genes and genomes and reticulate evolutionary relationships including horizontal gene transfer, hybridization and interspecific recombination. This makes ggtreeExtra an ideal tool for exploring relationships or interactions between taxa in a compact way and it is extremely powerful and uniquely suitable for microbiome research to present microbial correlation or interaction network with phylogenetic tree and other associated data. To demonstrate this unique feature, we used ggtreeExtra and ggtree\textsuperscript{11} to integrate and visualize several data sets from Arabidopsis leaf microbiome\textsuperscript{16} on the phylogenetic tree, including directional interactions among different bacteria strains, number of target genes, strain abundance, taxonomy information and the biosynthetic potential of the isolates. The phylogenetic tree was visualized using inward circular layout and the interaction data was visualized as a chord diagram that connecting corresponding isolates of the tree leaves. Other information was displayed as stacked bar chart, heatmaps and symbolic points on the tree (Fig. 1). With ggtreeExtra incorporating all these information, some of the evolutionary patterns that are not straightforward becomes more obvious. We found that inhibitor interactions are more widely found at strains from Firmicutes and Grammproteobacteria, whereas, strains from Alphaproteobacteria and Betaproteobacteria prefer sensitivity interactions. These strains might have more BGCs (biosynthesis gene clusters) of ribosomal synthesized and post-translationally modified peptide. To our knowledge, there is no other software tools that can easily produce a figure like this and the visualization do help us to explore the data and generate new insights as our findings were not revealed in the original paper\textsuperscript{17}.

The ggtreeExtra is a sub-package of the ggtree package suite and takes all the advantages of other ggtree sub-packages. Phylogenetic data imported by the treeio\textsuperscript{18} package can be used in ggtreeExtra. This allows evolutionary inferences (\textit{e.g.} clade support, molecular dating and selection pressure) from commonly used software to be linked to other associated data (\textit{e.g.} observational and experimental data) for integrative and comparative study (Supplemental Fig. S6). Tree data can be processed using the tidytree package and the phylogenetic tree visualized by ggtree with fully annotation can be further annotated in ggtreeExtra with data layers especially in circular layout (Supplemental Fig. S5-S8 and Fig. 1). The ggtreeExtra package extends the capabilities of ggtree and fully supports grammar of graphics implemented in ggplot2\textsuperscript{15} (Supplemental Fig. S1). It supports aesthetic mapping (Supplemental Fig. S3-6) and a layered grammar of graphics (Supplemental Fig. S7-9). User can use scale functions to specify how the data was mapped to visual values and theme functions to adjust graphic appearance (Supplemental Fig. S3-9). Moreover, it takes all benefits of the ggplot2 community. Geometric layers defined in ggplot2 and other extension packages can be used in ggtreeExtra to visualize tree data (Supplemental Table S2 and Fig. S2-5). We proposed and implemented this framework design originally in ggtree\textsuperscript{10} and ggtreeExtra fully embrace the design concept. This is the beauty of the ggtree and ggtreeExtra and lays the foundation for displaying tree annotated data layers. It allows ggtreeExtra to
support more visualization methods and has no assumption of the input data types (Supplemental Table S1 and S2). As the ggplot2 community keeps expanding, there will be more methods implemented which can be employed to create tree data layer in ggtreeExtra. Furthermore, the combination of these methods allows ggtreeExtra to create more possibilities than other tools to integrate more diverse data sets for novel exploratory data analysis (Fig. 1 and Supplemental Fig. S2). Therefore, it has more potential to reveal systematic patterns and insights of our data than other tools. The versatility of this package ensures its applications in different research areas such as population genetics, molecular epidemiology and microbiome.

Methods

Code development. The ggtreeExtra package was designed with four parts: data importation, data scaling, external layer position adjustment and the API interface. External data integrated to the tree or ggtree objects can be directly used in ggtreeExtra. The geom_fruit function also allows directly importing data and the data will be reordered according to the tree structure. External data either integrated to the tree or imported by the geom_fruit function need to be rescaled, since the ranges of the data and the tree are not in the same order of magnitude. Before visualizing the data using specified geometric layer, the data scaling will be performed automatically to standardize the data range according to the magnitude of the tree scale. The data scaling process only changes the range of the data while keeping the distribution of the data unchanged. This ensures the control of the panel widths and avoids the tree and any external panel to be squeezed (Supplemental Fig. S1).

The position of the data graph (i.e. on right-hand side or external ring of the tree, see also Supplemental Fig. S3) is controlled by the position parameter, which accepts a position object. Follow the object-oriented framework of the ggplot2 package, ggtreeExtra implemented several position objects including position_identityx, position_stackx, position_dodgex and position_dodgex2 to shift the position of corresponding geometric layer horizontally. To increase user experience, the default value of the position parameter is set to “auto” and the geom_fruit function will guess and determine a suitable position object for the specified geometric layer. That means using position_stackx for geom_bar, position_dodgex for geom_boxplot and geom_violin, and geom_identityx for the others (e.g. geom_point, geom_tile, geom_star and geom_phylic etc.). A geometric layer defined in ggplot2, ggtree or other ggplot2 extensions that has a position parameter should be compatible with the geom_fruit function, since it allows using position object defined in the ggtreeExtra package to adjust the output layer position (Supplemental Table S2).

The interface of the geom_fruit is intuitive. It is easy to follow and working in the way users expect it to. External data can be imported via the data argument. The data will be visualized using a geometric layer, which can be specified via the geom argument. The mapping argument is designed to construct aesthetic mappings that describe how variable in the data are mapped to visual characteristics (e.g. color, shape and size etc.). The y aesthetic is required to be mapped to a variable that stored the taxa names in the external data set to allow geom_fruit to link through the data and the tree. The data will be reordered
based on the tree structure. Other aesthetic mappings and non-variable setting can be specified to control the appearance of the data graph. User can use the offset and pwidth arguments respectively to specify the distance between the external panels and the width of the data graph layer. Besides, the geom_fruit function allows setting axis and background grid lines for the data graph layer using the grid.params and axis.params arguments respectively. The data graph layer produced by geom_fruit will be displayed on external panel and perfectly aligned to the tree. Visualizing data in ggtreeExtra is as simple as visualizing the data in ggplot2. The critical elements are the same – a data frame and a geometric layer to visualize the data with specific aesthetic mapping. Users don’t need to care about how the tree was linked to the data, since the geom_fruit function will do all the data manipulation for us. The geom_fruit was developed using object oriented programming and the output data layer can be added to the tree graph using the plus sign that followed the layered grammar of graphics. Different data sets can be progressively added and displayed on different external panels.

**Code Availability.** The ggtreeExtra R package is open source and freely available on Bioconductor (https://bioconductor.org/packages/ggtreeExtra) and github (https://github.com/YuLab-SMU/ggtreeExtra). R markdown files and data sets that used to generate the supplemental file, are available on github (https://github.com/YuLab-SMU/plotting-tree-with-data-using-ggtreeExtra). We also provide an online book, https://yulab-smu.top/treedata-book/, that document all the details of the ggtree package suite including ggtreeExtra with many step-by-step examples.

**Declarations**

**Supplementary Material**

Supplementary data are available at the online.

**Acknowledgments**

This work was supported by the startup fund from Southern Medical University (G618289088).

**References**

1. Rockett, R.J., Arnott, A., Lam, C. et al. Revealing COVID-19 transmission in Australia by SARS-CoV-2 genome sequencing and agent-based modeling. Nat Med 26, 1398–1404 (2020).

2. D Ainsworth, T. et al. The coral core microbiome identifies rare bacterial taxa as ubiquitous endosymbionts. *The ISME Journal* 9, 2261–2274 (2015).

3. Morgan, X. C., Segata, N. & Huttenhower, C. Biodiversity and functional genomics in the human microbiome. *Trends in Genetics* 29, 51–58 (2013).
4. Segata, N., Börnigen, D., Morgan, X. C. & Huttenhower, C. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nature Communications* **4**, 2304 (2013).

5. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v4: Recent updates and new developments. *Nucleic acids research* **47**, W256–W259 (2019).

6. Subramanian, B., Gao, S., Lercher, M. J., Hu, S. & Chen, W.-H. Evolview v3: A webserver for visualization, annotation, and management of phylogenetic trees. *Nucleic acids research* **47**, W270–W275 (2019).

7. Argimón, S. *et al.* Microreact: Visualizing and sharing data for genomic epidemiology and phylogeography. *Microb Genom* **2**, e000093 (2016).

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

9. Asnicar, F., Weingart, G., Tickle, T. L., Huttenhower, C. & Segata, N. Compact graphical representation of phylogenetic data and metadata with graphlan. *PeerJ* **3**, e1029 (2015).

10. Yu, G., Lam, T. T.-Y., Zhu, H. & Guan, Y. Two methods for mapping and visualizing associated data on phylogeny using ggtree. *Molecular Biology and Evolution* **35**, 3041–3043 (2018).

11. Yu, G., Smith, D., Zhu, H., Guan, Y. & Lam, T. T.-Y. Ggtree: An r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution* **8**, 28–36 (2017).

12. Yu, G. Using ggtree to visualize data on tree-like structures. *Current Protocols in Bioinformatics* **69**, e96 (2020).

13. Gentleman, R. C. *et al.* Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology* **5**, R80 (2004).

14. Wilkinson, L. The grammar of graphics. in *Handbook of computational statistics: Concepts and methods* (eds. Gentle, J. E., Härdle, W. K. & Mori, Y.) 375–414 (Springer Berlin Heidelberg, 2012). doi:10.1007/978-3-642-21551-3_13.

15. Wickham, H. *Ggplot2: Elegant graphics for data analysis.* (Springer-Verlag New York, 2016).

16. Bai, Y., Müller, D., Srinivas, G. *et al.* Functional overlap of the Arabidopsis leaf and root microbiota. *Nature* 528, 364–369 (2015).

17. Helfrich, E. J. N. *et al.* Bipartite interactions, antibiotic production and biosynthetic potential of the arabidopsis leaf microbiome. *Nature Microbiology* 3, 909–919 (2018).

18. Wang, L.-G. *et al.* Treeio: An r package for phylogenetic tree input and output with richly annotated and associated data. *Molecular Biology and Evolution* 37, 599–603 (2020).

19. Yu, G. Ggtree: a serialized data object for visualization of phylogenetic tree and annotation data. bioRxiv,2020.