tinselR – An RShiny Application for Annotating Outbreak Trees

Running Title:

Jennafer A. P. Hamlin[current,1,2,#] Teofil Nakov[3], and Amanda Williams Newkirk[2]

[1] Association of Public Health Laboratories Bioinformatics, Silver Springs, MD, USA

[2] Enteric Diseases Laboratory Branch, Centers for Disease Control and Prevention, Atlanta, GA, USA

[3] Department of Biological Sciences, University of Arkansas, Fayetteville, Arkansas, USA

[#] Address correspondance to Jennafer A. P. Hamlin ptx4@cdc.gov

[current] Respiratory Diseases Laboratory Branch, Centers for Disease Control and Prevention, Atlanta, GA, USA

ABSTRACT

Across the United States, public health laboratories perform whole-genome sequencing for many pathogens. This high-resolution data determines the relationships between isolates via phylogenetics. In combination with other epidemiological data, epidemiologists use this data to inform investigations to confirm an outbreak and identify potential transmission routes. Our goal was to develop an open-source user-friendly graphical user interface (GUI) for phylogenetic tree visualization and annotation. Here, we present tinselR for that purpose and which is availble at https://github.com/jennahamlin/tinselR.

ANNOUNCEMENT

Given that the R programming language contains some of the gold standard packages for phylogenetic analyses and visualization (e.g., ape (Paradis, Claude, and Strimmer 2004), and ggtree (Yu et al. 2017)), we

used the Rshiny framework (Chang et al. 2017) to develop **tinselR** (pronounced tinsel-er) to provide GUI access to the tools in ape, ggtree, and other vital packages. tinselR's minimum input requirement is a Newick formatted phylogenetic tree. Once loaded, user-selected inputs change the appearance of the displayed tree. For example, a user can quickly transform tip label formatting. By adding a genetic distance matrix or metadata file or both, the user can include annotations on the image, relabel tips, or add a heatmap to the phylogenetic tree. These modified tree images are downloadable in various formats (pdf, png, or tiff) for presentations, publications, or other communications with collaborators. Below we detail how to install the application and describe the example data pre-loaded so that new users can familiarize themselves with the application.

The genetic distance matrix file must contain a square matrix of single nucleotide polymorphism (SNP) differences between the tree tips. The metadata file is a table of additional information to be changed or displayed on the tree. The tip labels in the Newick tree, distance matrix, and metadata files must match before upload, or tinselR will report an error. The primary function of the metadata file is to relabel the tips on the tree image. The header of the first column must be Tip.labels, and it must contain the labels for all tree tips in the uploaded Newick file. The alternative identification labels can be provided in the metadata file using the column header Display.labels in column two. If desired, users may include additional columns in the metadata file, such as the collection site, and display the information in a heatmap next to the tree. Headers for these other columns in the metadata file are flexible because they are not automatically recognized and used by tinselR. Acceptable formats include CSV, TSV, and TXT for the genetic distance and metadata files. Users can set file types independently for each input.

INSTALLATION AND EXAMPLE DATA

To install tinselR from GitHub, users will need to install the R package devtools (Wickham and Chang 2016). The R packages ggtree (Yu et al. 2017) and treeio (Wang et al. 2020) is also required and can be installed from Bioconductor using BiocManager (Morgan 2019). With the installation of these dependencies, tinselR is installable via the install_github command from devtools. Explicit installation commands are below (Figure 1), and the final command (run_app()) will launch the application. Note that install_github will also install other missing R dependencies. tinselR will accept Newick tree files from any program, e.g., RAxML (Stamatakis 2014), as input. Although it is possible to host ShinyR applications on a server, to date tinselR has only been tested by single users running the application locally. We recommend testing to ensure tinselR performs as expected under multi-user conditions before providing access from a server for production purposes.

After launching tinselR, new users can explore the application using one of the pre-loaded datasets located in

the 'Example Data' tab. We provide three datasets (i.e., Newick formatted tree, genetic distance matrix, and metadata file). These data are either *Escherichia coli* (from NCBI Bioproject: PRJNA218110) or *Salmonella enterica* (from NCBI Bioproject: PRJNA230403) with the number of isolates ranging from 14 - 19. The genomic data used in the example data sets were generated and used under the CDC IRB protocol 7172. After clicking on the 'Example Data' tab, users can select one of the datasets (e.g., example data 1, 2, and 3) from the drop-down menu. We highlight the capabilities of tinselR (Figure 1) using example data 1 below. Run the below code in your R console -

1). Install devtools package

```
install.packages("devtools", dep=T)
```

2). Install ggtree and treeio

```
if (!requireNamespace("BiocManager", quietly = TRUE))
install.packages("BiocManager")
BiocManager::install("ggtree")
```

Note that installing ggtree will also install treeio

3). Install and launch the tinselR shiny application

```
devtools::install_github("jennahamlin/tinselR")
library(tinselR)
run_app()
```

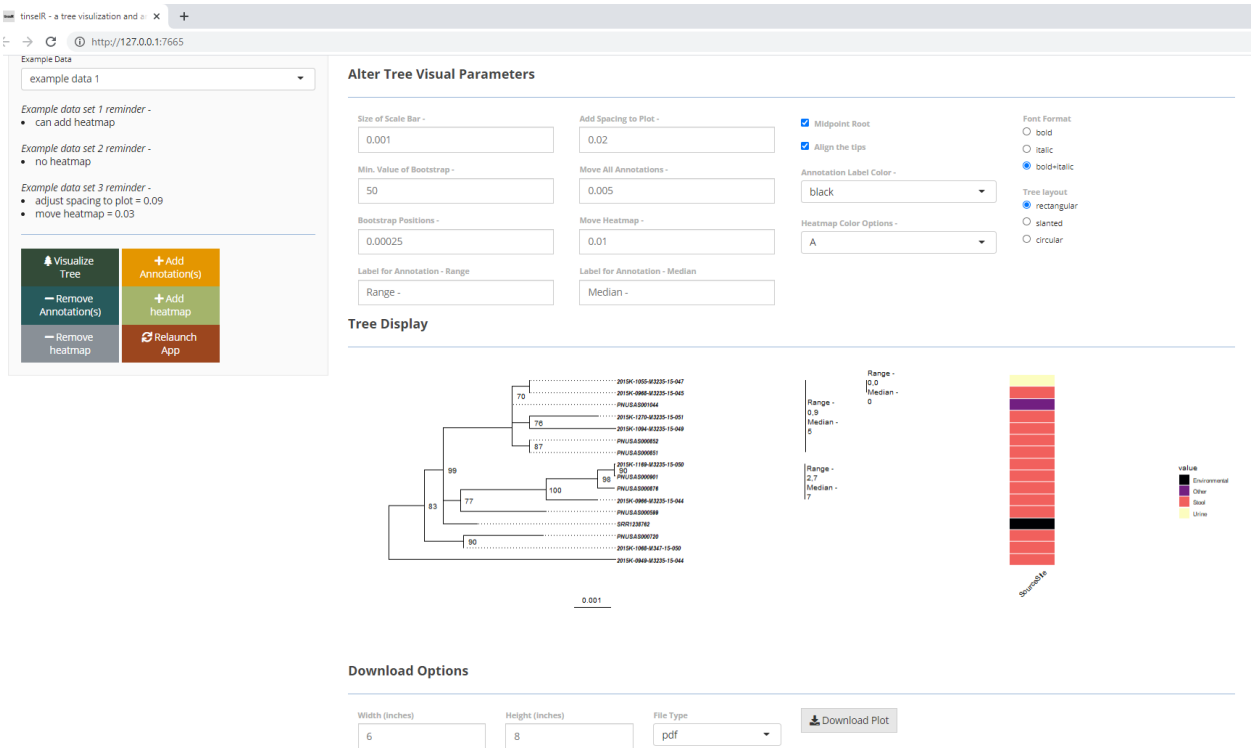

Figure 1: Example dataset 1 displayed with annotations and a heatmap indicating collection source.

Acknowledgements

We would like to thank those who participated in testing the application and providing valuable feedback during code review including the Biome team at CDC. We also thank J. Notoma for help getting tinselR up and running on the CDC internal server. This publication was supported by Cooperative Agreement Number 60OE000103, funded by Centers for Disease Control and Prevention through the Association of Public Health Laboratories. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of Centers for Disease Control and Prevention or the Association of Public Health Laboratories.

References

Chang, Winston, Joe Cheng, J Allaire, Yihui Xie, Jonathan McPherson, and others. 2017. “Shiny: Web Application Framework for R.” *R Package Version* 1 (5).

Morgan, M. 2019. “BiocManager: Access the Bioconductor Project Package Repository. R Package Version 1.30. 10.”

Paradis, Emmanuel, Julien Claude, and Korbinian Strimmer. 2004. “APE: Analyses of Phylogenetics and Evolution in R Language.” *Bioinformatics* 20 (2): 289–90.

Stamatakis, Alexandros. 2014. “RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies.” *Bioinformatics* 30 (9): 1312–3.

Wang, Li-Gen, Tommy Tsan-Yuk Lam, Shuangbin Xu, Zehan Dai, Lang Zhou, Tingze Feng, Pingfan Guo, et al. 2020. “Treeio: An R Package for Phylogenetic Tree Input and Output with Richly Annotated and Associated Data.” *Molecular Biology and Evolution* 37 (2): 599–603.

Wickham, Hadley, and Winston Chang. 2016. “Devtools: Tools to Make Developing R Packages Easier.” *R Package Version* 1 (0): 9000.

Yu, Guangchuang, David K Smith, Huachen Zhu, Yi Guan, and Tommy Tsan-Yuk Lam. 2017. “Ggtree: An R Package for Visualization and Annotation of Phylogenetic Trees with Their Covariates and Other Associated Data.” *Methods in Ecology and Evolution* 8 (1): 28–36.