TaxonomR: Taxonomic Decision Support Tool

User Manual

1.   Introduction

TaxonomR is an interactive application designed to facilitate taxonomic ranking decisions and make them more transparent, explicit, and accessible to interpretation and comparison across alternative classifications. The program is developed using the Shiny package for R programming language and is available as an online application (https://oksanav.shinyapps.io/TaxonomR/). The source code and user instructions and manual are available through GitHub: https://github.com/OksanaVe/TaxonomR.

2.   Input files

Before using TaxonomR, you need to prepare two types of input files – (1) a tree file and (2) a data file. Examples of input data files are available at https://github.com/OksanaVe/TaxonomR.

a) *Tree file* – The purpose of a tree file is to provide information about evolutionary relationships within the group of interest as well as additional information related to the evolutionary history of the group such as divergence times of lineages and degree of their divergence (amount of genetic or morphological differences). Additionally, characteristics of phylogenetic trees themselves can be weighted into making taxonomic ranking decisions, such characteristics include robustness of a tree in terms of the statistical support of its nodes and degree of the tree resolution. TaxonomR supports nexus and newick tree formats and accepts output tree files produced by most popular phylogenetic software (e.g. MrBayes, BEAST, IQ-TREE, RAxML, PAUP, etc.). It is important to note that TaxonomR is not software for phylogenetic inference and does not reconstruct topologies from data files; instead, it requires an input tree file produced by an external program for phylogenetic inference. Each input tree file should contain a single tree topology with branch lengths and support values. Optional annotations such as node ages can also be included. If support values and node ages were estimated separately for a tree, a second tree file can be uploaded into the program; in this case, both files should include the same topology and the first uploaded tree file must contain node support values while the second tree file must contain node age information.

b) *Data file* – Information about names and taxonomic composition of suggested groupings of taxa (e.g., genera, families, orders) as well as additional information about those groups such as geographical range, usage of the taxonomic name, and information about ecological traits is provided via a CSV-formatted data file. The overall structure of the data file (names and order of the input data columns) must be kept the same as in the template data file available at https://github.com/OksanaVe/TaxonomR. We recommend that you download a template data file from the TaxonomR GitHub page and fill it in with your information to

ensure proper formatting of the data file (note that all entries in the data file are case-sensitive). Alternatively, if you decide to create a data file from the scratch, make sure the headings for the data columns are as shown in Figure 1.

	A	B	C	D	E	F	G	H	I
1	Group	Taxa	Usage	Morph	Ecology	Distribution	Age	Misc_1	Misc_2
2									
3									
4									
5									
6									
7									
8									
9									
10									
11									

Figure 1. Example of a proper heading format for an input data file.

Below is a brief description of the information that should be specified in each column of the input data file:
- *Group*: name of a taxonomic group (e.g., genus name);
- *Taxa*: list of terminal taxa included in a group (e.g., list of species included in the genus). Taxon names should be spelled exactly as they appear in the tree file;
- *Usage*: this field includes information on the current usage of the proposed taxonomic group – it is scored as "new" for a newly proposed taxonomic group name, "new_use" for a previously used taxonomic name but with a modified taxonomic composition, "minor_use" for an existing but rarely used taxonomic name, and "major_use" when using a previously established and commonly used taxonomic name;
- *Morph*: this field indicates presence or absence of distinct morphological traits characterizing a proposed group. For operational convenience, this field is scored on a discrete scale with possible states defined as either "none" for the absence of distinguishing phenotypic traits (e.g., cryptic taxa that are genetically differentiated but are not morphologically distinguishable), "plastic", if morphological features that characterize a taxon are plastic continuous traits such as colouration or highly plastic meristic and morphometric characters prone to homoplasy (e.g., number of gill rakers and body ratios in fish), "combo" for a unique combination of non-unique phenotypic traits (e.g., trout-perches of the genus *Percopsis* are characterized by the presence of an adipose fin, ctenoid scales, and fin spines; individually, these features are not unique to the group but together form a unique diagnostic combination), and "synap" for the presence of one or more unambiguous uniquely derived morphological traits that characterize a group.
- *Age*: this value specifies the divergence time for the proposed taxonomic group. A single value is entered in this column, which initiates an additional slider bar in the program interface where standard deviations can be indicated. The age value should be in the same units as the nodal age values in the tree file (e.g. million years);
- *Distribution*: geographical range of the group. Distribution can be specified as a list of biogeographical realms, countries, regions, or areas in which members of the group occur. Range units should not include extensively overlapping or mutually inclusive regions (e.g.,

New World and South America); they can include areas of any scale, from provincial regions to countries and entire continents. Because defining distribution units can be highly subjective and variable across studies (Vences et al., 2013), the distribution division scheme should be clearly identified in the study.

- *Ecology*: ecological trait indicative of the group's adaptive zone. It can include information on the diet, foraging habit, reproductive strategy, type of locomotion, nesting and migratory behaviour, host association or other traits that may be important in defining taxonomic groups. This criterion has to be scored as discrete traits and input into the information spreadsheet using short and descriptive keywords without any spaces, e.g. type of habitat can be scored as "woodlands", "tall_grasslands", "tropical_rainforest", "tundra", etc.
- *Miscellaneous custom traits:* current implementation of TaxonomR allows users to specify up to two additional traits in the "Misc_1" and "Misc_2" columns. While the choice of traits is determined entirely by the user, it should be clearly defined in the study. Scoring and formatting of these traits are similar to that of "Ecology".

Data fields for any column except Group and Taxa can be left blank if the information for a given criterion is not available or is not considered important.

3.   Using TaxonomR

TaxonomR has a very simple and accessible interface for users of all levels. The general workflow includes only two steps – uploading input files and specifying a prioritization scheme using interactive sliders to set preferred weights for different decision-making criteria. The second step is a dynamic procedure – TaxonomR reacts to the information provided by the user and outputs updated scores for each taxonomic group included in the analysis. The output is updated every time user changes any of the input weight values or data files.

       a.   Launch online application available at https://oksanav.shinyapps.io/TaxonomR/
   b.   Get familiar with the user interface (Fig. 2):
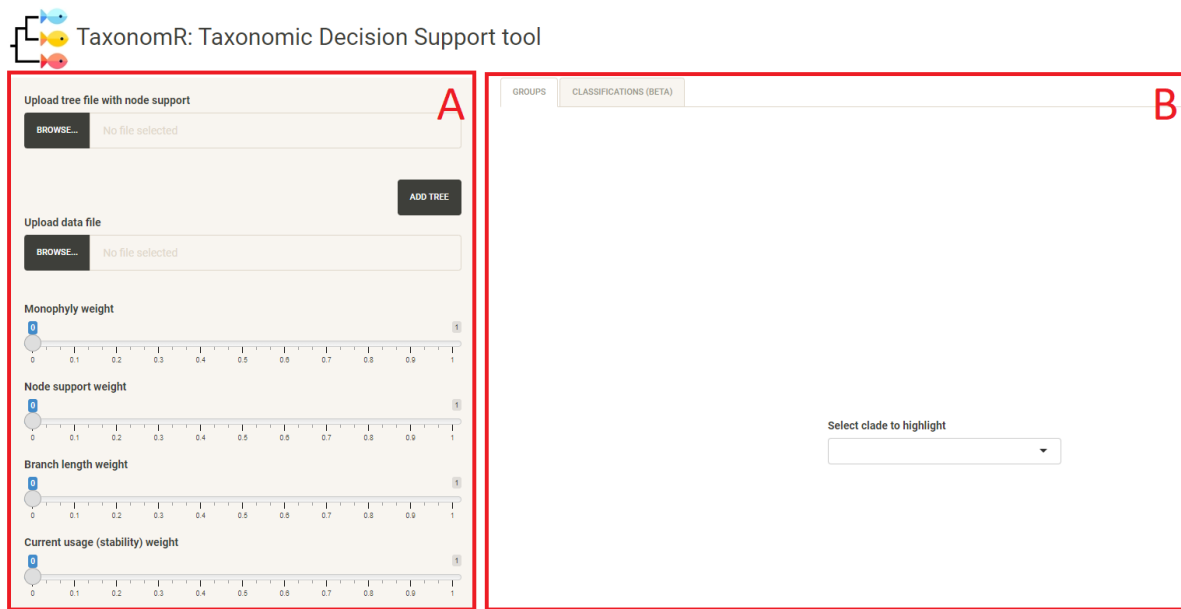
Figure 2. TaxonomR interface. A – user input panel; B – output field.

c. Upload tree file(s) by clicking the "Browse" button (Fig. 3A) at the top of the user input panel and navigating to the tree file saved on your computer. If you need to upload another tree file with additional information, click the "Add Tree" button and repeat previous steps to upload additional file (Fig. 3B). If you have a single tree file which contains all necessary information (branch lengths, support values, and divergence dates), you do not need to use the "Add Tree" option.
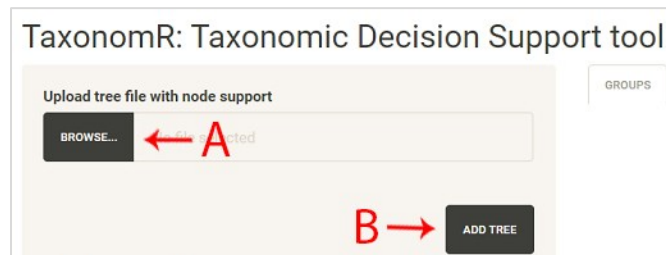
Figure 3. Upload a tree with branch length and node support values by clicking the "Browse" button (A). If a second tree file needs to be uploaded, click the "Add Tree" button (B) and upload another file in the new tree upload line.

d. Next, upload data file by clicking the "Browse" button in the data file upload field and navigating to your csv-formatted file. Once both tree and data files are uploaded, TaxonomR will update the output view field to show the tree and a list of taxonomic groups provided in the data file (Fig. 4). You can choose a taxonomic group to highlight and show as a zoom-in section by selecting that group from the "Select Group" drop-down menu. In the zoomed view, taxa that do not belong to the group are collapsed and shown as number of taxa in square brackets.
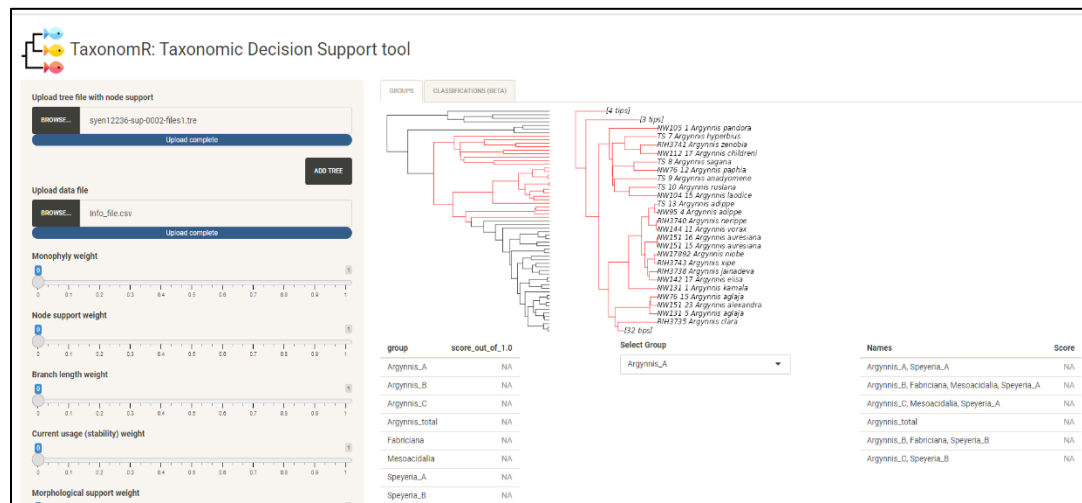
Figure 4. A screenshot of the TaxonomR interface with uploaded tree and data files. The output field shows the overall view of the supplied phylogenetic tree (tree view on the left) and a zoomed-in portion of the tree highlighting a group selected from the "Select Group" drop-down menu (clade view in the centre).

Below the tree view panel, you will see a list of all taxonomic groups specified in the data file. Initially, groups do not have score values; instead, each group has an 'NA' value which is updated to a calculated numeric value once you start specifying weights of the decision-making criteria in the user input panel on the left.

e. The last step is to start specifying weights of the decision-making criteria by adjusting sliders in the user input panel on the left. Each criterion is weighted on a scale from 0 to 1. Criteria that are considered to be of primary importance to the taxonomic ranking should be assigned higher weights than criteria of secondary importance. Figure 5 shows a simple example of weighting scheme that prioritizes monophyly as well as puts high weight on the timing of divergence of the groups. Under this scheme, the output suggested classification includes a single group (Argynnis_total) that originated within a narrow time band around 9 Mya (age range from 8.7 to 9.5 Mya).
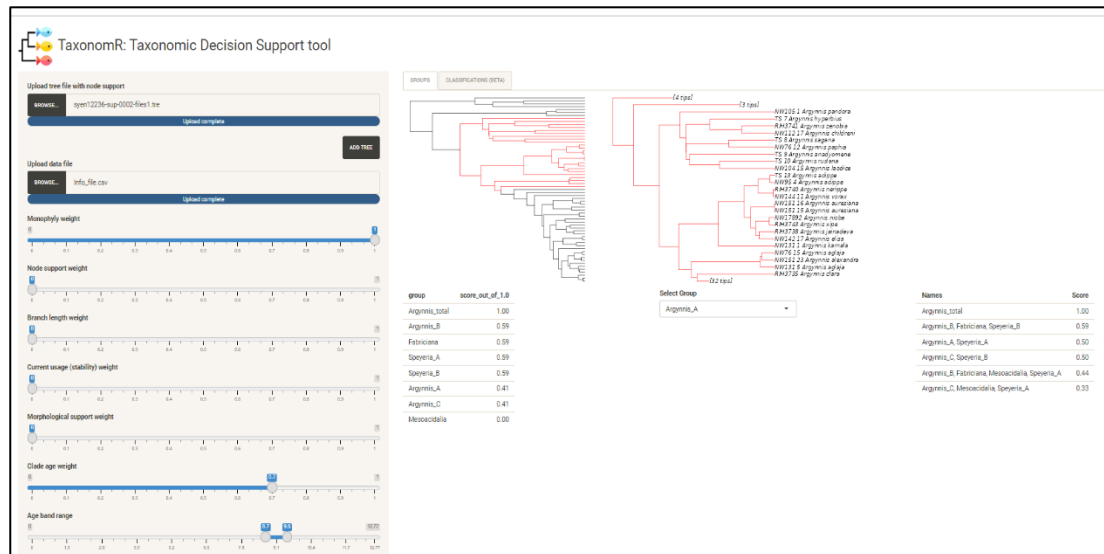
Figure 5. Screenshot of TaxonomR showing a simple example of a weighting scheme that prioritizes monophyly and divergence time. Suggested classification is shown on the right.

Finally, remember that taxonomic ranking decisions are inherently subjective, and it is important to make them transparent and explicit to reduce vagueness and make decisions accessible to a wide audience to engage interest of specialists and anyone interested in the topic.