TREE-BASED ALIGNMENT SELECTOR
(T-BAS)

v. 2.1

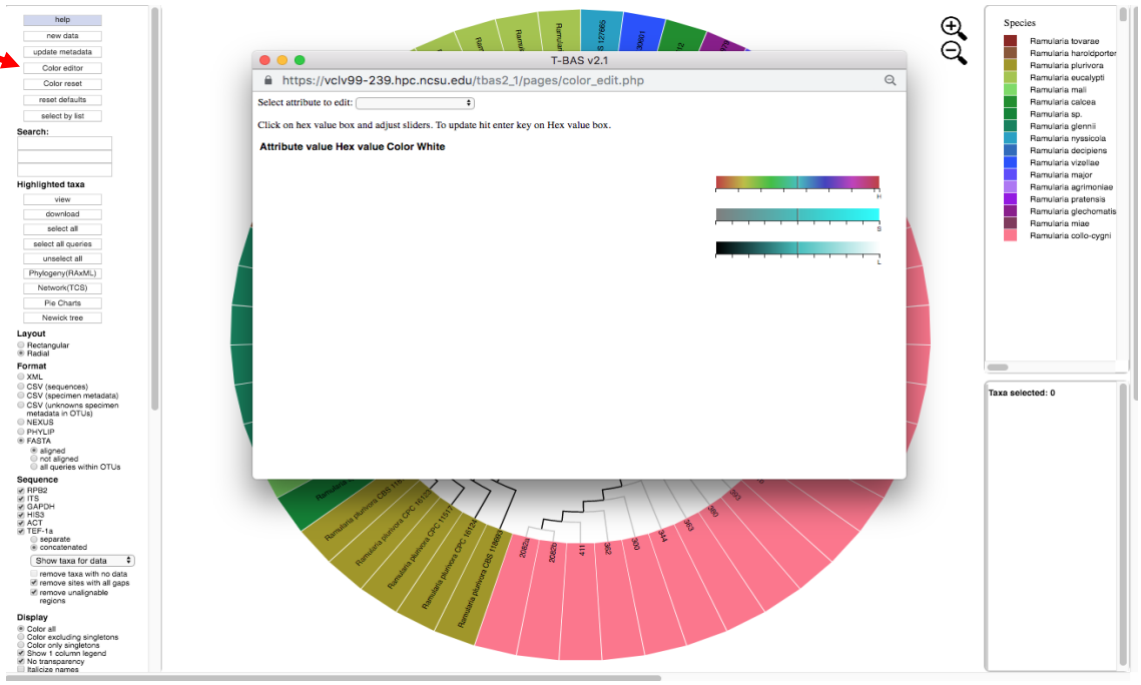
A TOOLKIT FOR EVOLUTIONARY PLACEMENT OF DNA SEQUENCES, VIEWING ALIGNMENTS AND SPECIMEN METADATA ON CURATED AND CUSTOM TREES

TABLE OF CONTENTS

COLOR EDITOR

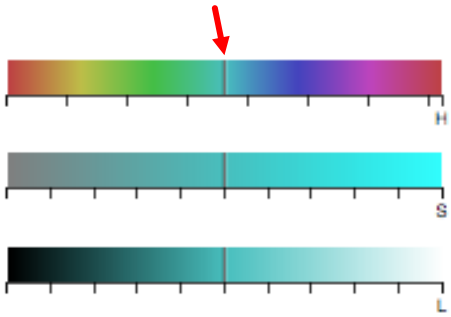
The purpose of the color editor is to allow the user to select preferred colors for the layout of the tree. When T-BAS creates a tree it randomly assigns colors to attributes from all colors in the spectrum. For each attribute, the rows in the legend are arranged by color so that the user can find the label of a color by looking in the legend. The colors can be changed in the color editor. However, the order of entries in the legend remains as for the original colors assigned. There is no limit to how many values or attributes can be edited.

To change the colors, click the color editor button and the color editor window will pop up.

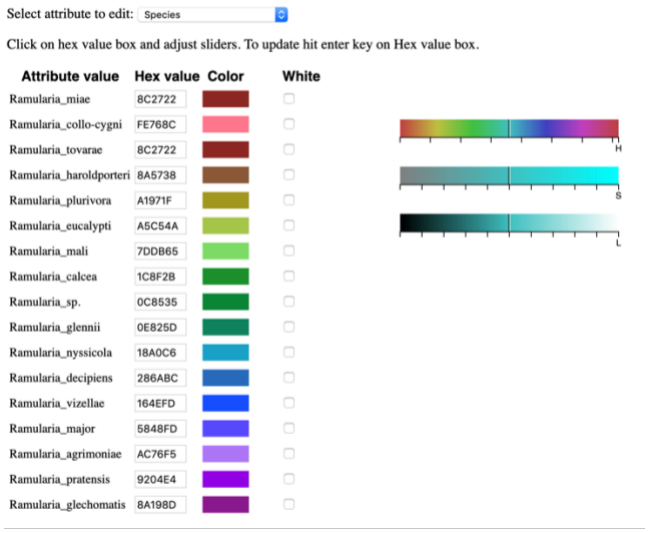


There are two ways to change the colors. Colors can be selected on the color bars or inputting a known hex color value.

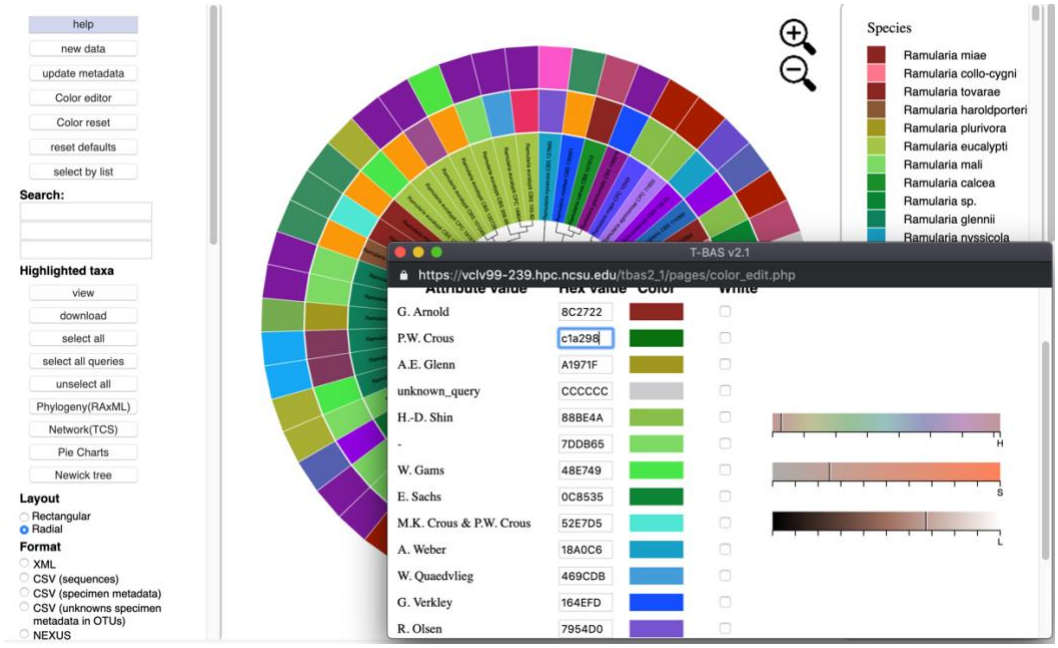
To change the color using the HSL (hue, saturation, lightness) color bars, slide the center vertical black line (while holding down the left mouse button) on one of the 3 bars to the left or right. One or all three bars can be modified in order to display the desired color. The letter under the corresponding bar indicate the following: H (hue), S (saturation), L (lightness).



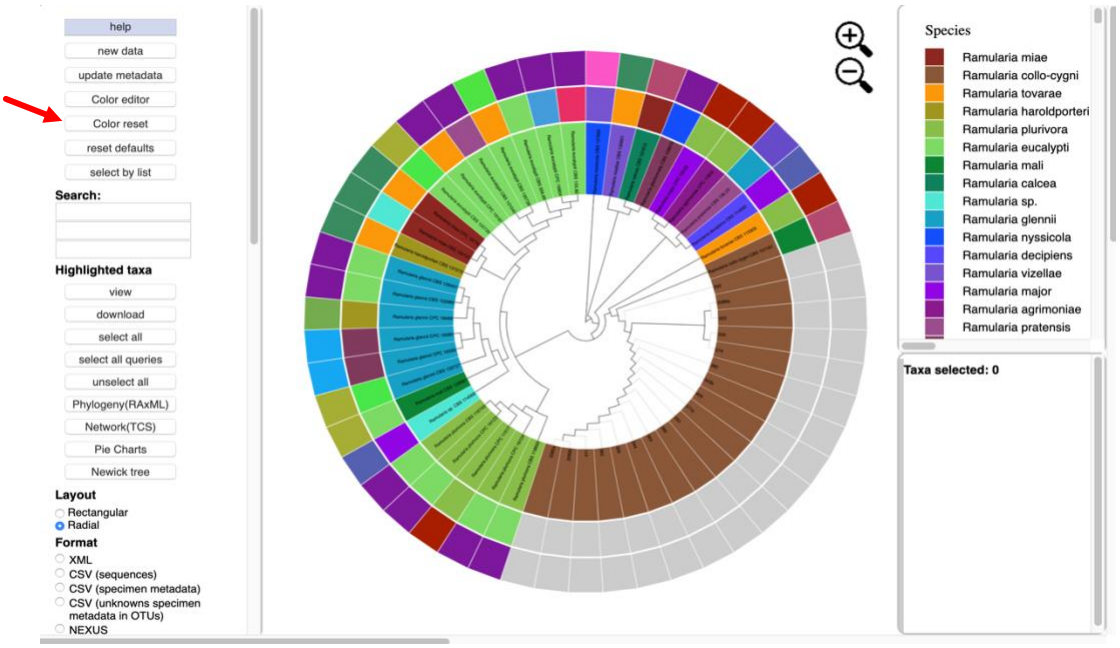
Selecting a specific attribute in the pull-down menu will display the current color arrangement on the tree. Here the hex values can be changed, if known. Hex values can be searched online or can be viewed here. Enter the value into the box and press Enter/Return.



If the value is not known, click inside a box of an attribute to be changed, and select a new color on the color bar or adjust the vertical black lines until a desired color appears. For the change to take effect, the cursor must be inside the box that has the edited color value and press Enter/Return. The colors will then be updated in the color editor, in the tree, and in the legend. To select the color white, click the box in the last column.


help
new data
update metadata
Color editor
Color reset
reset defaults
select by list
Search:
Highlighted taxa
view
download
select all
select all queries
unselect all
Phylogeny(RAxML)
Network(TCS)
Pie Charts
Newick tree
Layout
Rectangular
Radial
Format
XML
CSV (sequences)
CSV (specimen metadata)
CSV (unknowns specimen metadata in OTUs)
NEXUS
Species
Ramularia miae
Ramularia collo-cygni
Ramularia tovarae
Ramularia haroldporteri
Ramularia plurivora
Ramularia eucalypti
Ramularia mali
Ramularia calcea
Ramularia sp.
Ramularia glennii
Ramularia nyssicola
T-BAS v2.1
https://vclv99-239.hpc.ncsu.edu/tbas2_1/pages/color_edit.php
Attribute Value
Hex Value
Color
White
G. Arnold 8C2722
P.W. Crous c1a298
A.E. Glenn A1971F
unknown_query CCCCCC
H.-D. Shin 88BE4A
- 7DDB65
W. Gams 48E749
E. Sachs 0C8535
M.K. Crous & P.W. Crous 52E7D5
A. Weber 18A0C6
W. Quaedvlieg 469CDB
G. Verkley 164EFD
R. Olsen 7954D0
H
S
L

Clicking the color reset button will undo all changes.



To copy a color scheme from one tree to another, copy hex values and then enter them manually in the color editor on the next tree.

DATA STANDARDIZATION

PhyloXML is an XML language for describing evolutionary trees or networks and data associated with them (Han & Zmasek 2009). In T-BAS, DNA sequences and associated specimen metadata are phylogenetically placed on curated multi-locus reference trees and the placement results are standardized using an extended PhyloXML format. This standardization allows placements and associated specimen attributes (e.g. host, locality, environmental traits) to be readily viewed, archived and importantly analyzed within a phylogenetic context. Unlike other data standards used for phylogenetic trees, PhyloXML can be adapted and extended to integrate disparate forms of data. To this end we added additional PhyloXML elements to accommodate raw sequence data and alignments that are associated with each taxonomic scale in the tree – from phylum to individuals in populations. This standardization provides a consistent handling of the data and is currently used by T-BAS and other tools in the DeCIFR toolkit.

Our new PhyloXML schema named "cifr PhyloXML" includes new tags: cifr:otus, cifr:attributes, and cifr:genes.

cifr:otus

A cifr:otu tag saves all the information in the OTUs of the submitted samples.

A cifr:otu tag contains a cifr:name, cifr:leaf_name, and a cifr:taxon tag.

The cifr_taxon tag contains cifr:taxon_level and cifr:taxon_val tags with placement information for this OTU.

Also in the cifr:otu are cifr:placement tags with attributes and unaligned sequences for each sample in the OTU.

<?xml version="1.0" encoding="UTF-8"?>
<phyloxml xmlns:cifr="http://www.cifr.ncsu.edu" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns="http://www.phyloxml.org" x
<cifr:genes>
<cifr:otus similarity="1.0">
<cifr:otu>
<cifr:otu>
<cifr:otu>
<cifr:otu>
<cifr:otu>
<cifr:name>OTU4</cifr:name>
<cifr:leaf_name>411</cifr:leaf_name>
<cifr:taxon>
<cifr:taxon_level>Taxon assignment</cifr:taxon_level>
<cifr:taxon_val>Ramularia_collo-cygni_CBS_101181</cifr:taxon_val>
</cifr:taxon>
<cifr:taxon>
<cifr:taxon_level>Species-level assignment</cifr:taxon_level>
<cifr:taxon_val>Ramularia_collo-cygni</cifr:taxon_val>
</cifr:taxon>
<cifr:taxon>
<cifr:taxon_level>Likelihood weight</cifr:taxon_level>
<cifr:taxon_val>0.999991</cifr:taxon_val>
</cifr:taxon>
<cifr:taxon>
<cifr:taxon>
<cifr:taxon>
<cifr:placement>
<name>411</name>
<cifr:attributes/>
<sequence>
<gene_name>ITS</gene_name>
<mol_seq is_aligned="false">TTACTGAGTGAGGGAGCAATCCCGACCTCCAACCCTTTGTGAACGCATCACGTTGCTTCGGGGGCGACCCTGCC
</sequence>
</cifr:placement>
<cifr:placement>
<name>362</name>
<cifr:attributes/>
<sequence>
<gene_name>ITS</gene_name>
<mol_seq is_aligned="false">TTACTGAGTGAGGGAGCAATCCCGACCTCCAACCCTTTGTGAACGCATCACGTTGCTTCGGGGGCGACCCTGCC
</sequence>
</cifr:placement>
<cifr:placement>
<cifr:placement>
<cifr:placement>
<cifr:placement>
<cifr:placement>
<cifr:placement>

cifr:attributes

A cifr:attributes tag contains information for specimen metadata in the tree structure.

The cifr:attributes tag contains cifr:attribute, which contains cifr:name and cifr:value.

```
<?xml version="1.0" encoding="UTF-8"?>
<phyloxml xmlns:cifr="http://www.cifr.ncsu.edu" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns="http://www.phyloxml.org"
 <cifr:genes>
 <cifr:otus similarity="1.0">
 <phylogeny rooted="true">
  <clade id_source="tbas_id_0">
   <clade id_source="tbas_id_1">
    <name>Ramularia_nyssicola_CBS_127665</name>
    <branch_length>0.0</branch_length>
    <sequence>
    <sequence>
    <sequence>
    <sequence>
    <sequence>
    <sequence>
    <cifr:attributes>
     <cifr:attribute>
      <cifr:name>GB_TEF1</cifr:name>
      <cifr:value>KJ504680</cifr:value>
     </cifr:attribute>
     <cifr:attribute>
      <cifr:name>Country</cifr:name>
      <cifr:value>USA: Maryland</cifr:value>
     </cifr:attribute>
     <cifr:attribute>
      <cifr:name>otu_size</cifr:name>
      <cifr:value>1</cifr:value>
     </cifr:attribute>
     <cifr:attribute>
      <cifr:name>Host_isolation_source</cifr:name>
      <cifr:value>Nyssa ogeche _ sylvatica hybrid</cifr:value>
     </cifr:attribute>
     <cifr:attribute>
      <cifr:name>otu</cifr:name>
      <cifr:value/>
     </cifr:attribute>
     <cifr:attribute>
      <cifr:name>GB_HIS3</cifr:name>
      <cifr:value>KJ504592</cifr:value>
     </cifr:attribute>
     <cifr:attribute>
      <cifr:name>GB_ITS</cifr:name>
      <cifr:value>KJ504765</cifr:value>
     </cifr:attribute>
     <cifr:attribute>
      <cifr:name>GB_ACT</cifr:name>
      <cifr:value>KJ504429</cifr:value>
     </cifr:attribute>
```

cifr:genes

The cifr:gene tags saves metadata of the alignments.

The cifr:genes tag contains cifr:gene, which contains cifr:locus, cifr:nchar, and cifr:exset.

```
<?xml version="1.0" encoding="UTF-8"?>
<phyloxml xmlns:cifr="http://www.cifr.ncsu.edu" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns="http://www.phyloxml.org"
  xsi:schemaLocation="http://www.phyloxml.org phyloxml.xsd">
 <cifr:genes>
  <cifr:gene>
   <cifr:locus>rpb2</cifr:locus>
   <cifr:nchar>670</cifr:nchar>
   <cifr:exset>666-670</cifr:exset>
  </cifr:gene>
  <cifr:gene>
   <cifr:locus>his3</cifr:locus>
   <cifr:nchar>363</cifr:nchar>
   <cifr:exset>359-363</cifr:exset>
  </cifr:gene>
  <cifr:gene>
   <cifr:locus>gapdh</cifr:locus>
   <cifr:nchar>556</cifr:nchar>
   <cifr:exset>55-69 552-556</cifr:exset>
  </cifr:gene>
  <cifr:gene>
   <cifr:locus>tef</cifr:locus>
   <cifr:nchar>389</cifr:nchar>
   <cifr:exset/>
  </cifr:gene>
  <cifr:gene>
   <cifr:locus>act</cifr:locus>
   <cifr:nchar>182</cifr:nchar>
   <cifr:exset>178-182</cifr:exset>
  </cifr:gene>
  <cifr:gene>
   <cifr:locus>its</cifr:locus>
   <cifr:nchar>610</cifr:nchar>
   <cifr:exset>1-8 36-100 117 147-156 171-174 202-203 230-231 246-260 463-474 501-503 558 569-572 605-610</cifr:exset>
  </cifr:gene>
 </cifr:genes>
 <cifr:otus similarity="">
 <phylogeny rooted="false">
</phyloxml>
```

DECIFR REST SERVER

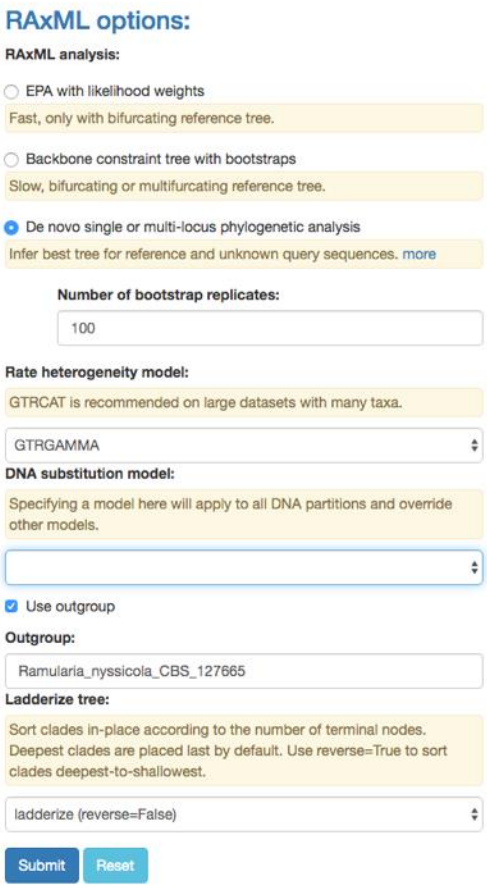
The code for a REST server that can be used to share results stored in a cifr PhyloXML over the internet is available at https://github.com/ncsu-decifr/decifr-rest.

The server uses the Python framework Flask – https://palletsprojects.com/p/flask/. Installation instructions are included.

To run the server, edit the location of the parameter TMP_FOLDER to the folder holding the cifr phyloXML files. Opening the URL to /list returns a list of run IDs of all the XMLs in the folder.

DE NOVO SINGLE OR MULTI-LOCUS PHYLOGENETIC ANALYSIS

This feature under the RAxML options can be used to Infer best tree for reference and unknown query sequences. Potential applications include: (1) inferring trees for species delimitation using the Genealogical Concordance Phylogenetic Species Recognition (GCPSR) concept (Taylor et al 2000), and (2) inferring an input tree for Poisson Tree Processes (PTP) model to delimit putative species (Zhang et al 2013).



References

Color Editor
https://www.compuhelpts.com/Color_Codes_1.pdf

Data Standardization
Han MV, Zmasek CM (2009) phyloXML: XML for evolutionary biology and comparative genomics. BMC bioinformatics 10, 356.

De novo single or multi-locus phylogenetic analysis
Taylor, J.W., D.J. Jacobson, S. Kroken, T. Kasuga, D.M. Geiser, D.S. Hibbett, et al. 2000. Phylogenetic species recognition and species concepts in fungi. Fungal Genet Biol 31: 21-32. doi:10.1006/fgbi.2000.1228.

Zhang, J., P. Kapli, P. Pavlidis and A. Stamatakis. 2013. A general species delimitation method with applications to phylogenetic placements. Bioinformatics 29: 2869-2876. doi:10.1093/bioinformatics/btt499.

APPENDIX

Description of Terms

Term	Description
Backbone constraint tree with bootstraps	RaxML method
Bifurcating tree	Tree where each node has 2 children
BLAST	Basic Local Alignment Search Tool, used to match unknown sequences to known sequences in database
Cifr phyloXML	Custom XML format that is valid phyloXML with added tags for use in T-BAS
De novo phylogenetic analysis	RaxML method
EPA with likelihood weights	RaxML method that places sequence on edges of existing tree
FASTA	A file sequence format for unaligned data
Genetic distance cutoff	Value used by custom algorithm to exclude divergent species from placement
GTRCAT (Rate heterogeneity model)	Faster model than GTRGAMMA that uses a different approximation to capture rate heterogeneity across sites
GTRGAMMA (Rate heterogeneity model)	General Time Reversible (GTR) model with Gamma distributed rates across sites
ITS	Internal transcribed spacer locus
Labels: Display Names	Node-click context menu, display leaf names in selected clade in large trees. Tree with greater than 2000 leaves do not display names for performance reasons.
Labels: Likelihood Weight	Node-click context menu, click on leaf of EPA placement will show all leaves attached to the edge that gives 95% cumulative weight."
Ladderize tree	Sort tree leaves from deepest to shallowest or reverse
Locus (Loci)	A location on a chromosome
LSU	Large subunit locus

Metadata: Download	Node-click context menu, download data of selected according to selections in format and sequence
Metadata: View	Node-click context menu, view data of selected according to selections in format and sequence in pop-up window
Multifurcating tree	Tree where each node can have multiple children
NEWICK	A standard for representing trees
NEXUS	A file format with multiple uses, can contain trees and alignments
OTUs	A grouping of sequences into percent similarity by the program QIIME
Outgroup	Leaves of a tree placed in a distinct clade, used to root tree
PHYLIP	A file format for aligned sequence data
PhyloXML	XML language designed to describe phylogenetic trees (or networks) and associated data
Query sequences	Unaligned unknown sequence data
Rate heterogeneity model	A phylogenetic model that accounts for evolutionary rate heterogeneity
RAxML	Software tool used to place alignment on a tree, plus some other utilities
Reference set	A set of tree, alignments, and metadata of known species at a specific taxonomic level used for placement
Taxa: Select All	Node-click context menu, select all leaves on tree
Taxa: Select(unselect)	Node-click context menu, select or unselect all leafs in clade
Taxa: Unselect All	Node-click context menu, unselect all leaves on tree
Tree: Collapse(expand)	Node-click context menu, collapse clade into a single node. Collapsed clade appears as a small circle. Click on this circle to restore clade.
Tree: Network (TCS)	Node-click context menu, create TCS network of all query strains in clade
Tree: Newick tree	Node-click context menu, download newick tree of selected clade in either phylip or NEXUS format

Tree: Phylogeny (RaxML)	Node-click context menu, create de novo tree of selected clade
Tree: Pie Charts	Node-click context menu, create pie charts to show relationships of selected attributes
Tree: Subtree (new window)	Node-click context menu, view subtree of selected clade in new window
Tree: Subtree(tree)	Node-click context menu, view subtree of selected clade
UNITE	Database of fungal ITS for BLAST