Tree-Based Alignment Selector (T-BAS)

v. 2.1.1

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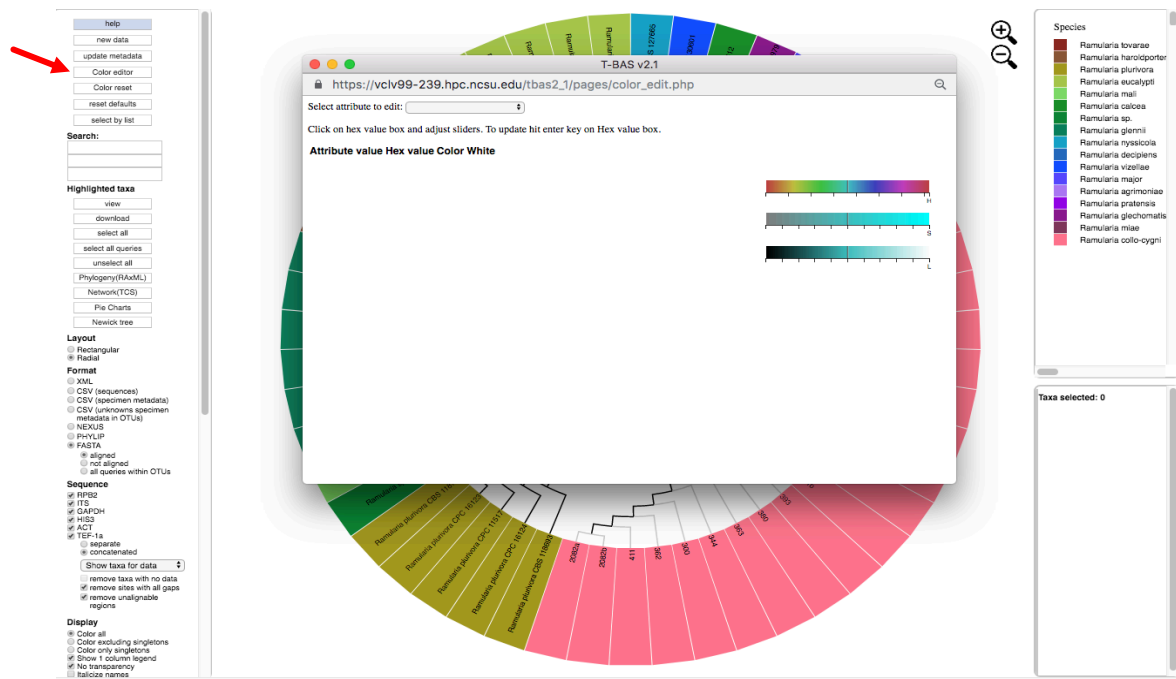
There are two ways to change the colors. Colors can be selected on the color bars or inputting a known hex color value.

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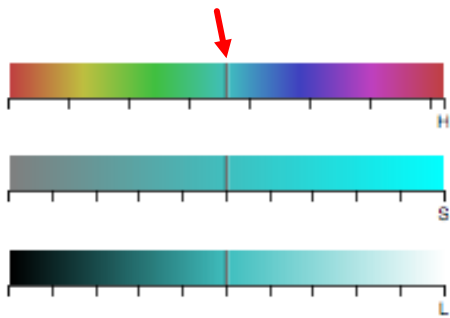
Selecting a specific attribute in the pull-down menu will display the current color arrangement on the tree. Here the hex values can be changed, if known. Hex values can be searched online or can be viewed here. Enter the value into the box and press Enter/Return.

| Attribute value	Hex value	Color	White
Ramularia_miae	8C2722		
Ramularia_collo-cygni	FE768C		
Ramularia_tovarae	8C2722		
Ramularia_haroldporteri	8A5738		
Ramularia_plurivora	A1971F		
Ramularia_eucalypti	A5C54A		
Ramularia_mali	7DDB65		
Ramularia_calcea	1C8F2B		
Ramularia_sp.	0C8535		
Ramularia_glennii	0E825D		
Ramularia_nyssicola	18A0C6		
Ramularia_decipiens	286ABC		
Ramularia_vizellae	164EFD		
Ramularia_major	5848FD		
Ramularia_agrimoniae	AC76F5		
Ramularia_pratensis	9204E4		
Ramularia_glechomatis	8A198D		

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.

Clicking the color reset button will undo all changes.
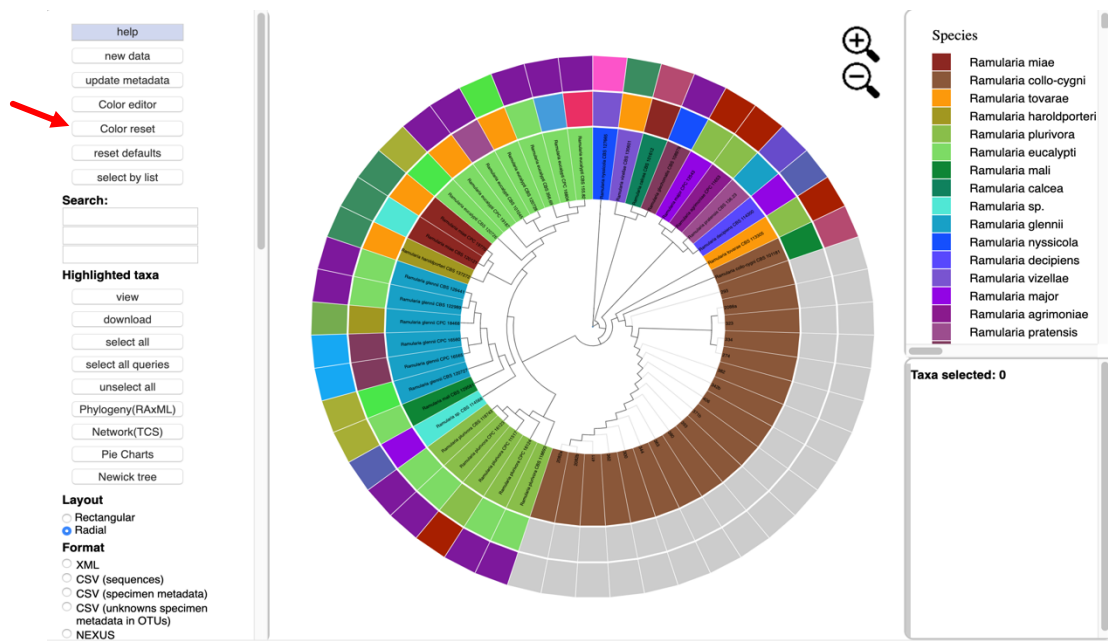
To copy a color scheme from one tree to another, copy hex values and then enter them manually in the color editor on the next tree.

DATA STANDARDIZATION

In T-BAS, DNA sequences and associated specimen metadata are phylogenetically placed on curated multilocus reference trees and the placement results are saved as Metadata Enhanced PhyloXML (MEP) files. The MEP format allows placements and associated specimen attributes (e.g. host, locality, environmental traits) to be readily viewed, archived and importantly analyzed within a phylogenetic context. MEP files are structured to adhere to the minimum information about any (x) sequence (MIxS) family of standards defined by the Genomic Standards Consortium. A template is provided for users to fill in and submit when performing a phylogeny-based placement in T-BAS. Additional categories of metadata information can be added. MIxS headers and metadata are saved in MEP files as defined in the XML schemas below. The use of MEP files ensures interoperability and retrieval of relevant sequences and metadata for downstream applications. This standardization provides a consistent handling of the data and is currently used by T-BAS and other tools in the DeCIFR toolkit. MEP is based on XML (Han & Zmasek 2009), a widely used markup language for representing and sharing information, and PhyloXML, an extension of XML with custom tags for describing evolutionary trees or networks.

The standard pre-defined XML schema for phyloXML is used as a starting point fo validating MEP files. PhyloXML includes a phylogeny element that saves the tree information and associated alignments. MEP extends this by adding (1) an OTUs tag that saves the taxonomic assignments, associated query metadata and sequences for each OTU, (2) a tag to each clade that is a leaf in the tree and saving the metadata for that leaf, and (3) a gene tag that saves the locus name, the number of sequence characters, and the positions of the excluded unaligned character set (i.e. exset) for each alignment.

MEP uses two associated schema definitions:
cifr_phyloxml.xsd to show how custom tags are added to PhyloXML.

cifr.xsd to define custom tags in the http://www.cifr.ncsu.edu namespace.

The MEP schema 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
<?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
<?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
<?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>
```

DᴇCIFR REST Sᴇʀᴠᴇʀ ᴀɴᴅ API

REST Server:

The Representational State Transfer Application Program Interface (REST) server can be used to retrieve tree information from Metadata Enhanced PhyloXML (MEP) files. This was released on github (https://github.com/ncsu-decifr/decifr-rest) using an open source BSD 3-Clause License. The DeCIFR REST server allows a user to share information about placements from T-BAS with other users via the web. The program is written to run in a Python 3 virtual environment and uses Flask (https://palletsprojects.com/p/flask/). Installation instructions are included.

REST API:

The DeCIFR REST API service allows access to results of a previous run of T-BAS v2.1.1 via HTTP with a browser request, programmatically, or using a command line tool such as CURL. More information is available at https://tools.decifr.hpc.ncsu.edu/rest .

For Docker Users:

Opening the URL to /list returns a list of run IDs of all the XMLs in the folder.

Clicking on a Run ID link (i.e.; 3F7THARX) will allow the leaves, queries, and OTUs to be viewed without opening the tree.
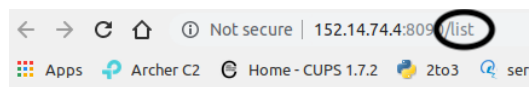
Click on 'leaves' to see the sample names that are present in a tree.

3F7THARX

leaves

queries

otus

Click on the metadata link for Ramulaira_calcea_CBS_101612 (#2).

3F7THARX

1. Ramularia_glechomatis_CBS_108979 - metadata
2. Ramularia_calcea_CBS_101612 - metadata
3. Ramularia_vizellae_CBS_130601 - metadata
4. Ramularia_nyssicola_CBS_127665 - metadata
5. Ramularia_major_CPC_12543 - metadata
6. Ramularia_agrimoniae_CPC_11653 - metadata
7. Ramularia_pratensis_CBS_136.23 - metadata
8. Ramularia_decipiens_CBS_114300 - metadata
9. Ramularia_tovarae_CBS_113305 - metadata
10. 411 - metadata
11. Ramularia_collo-cygni_CBS_101181 - metadata
12. Ramularia_plurivora_CBS_118693 - metadata
13. Ramularia_plurivora_CPC_16124 - metadata
14. Ramularia_plurivora_CPC_11517 - metadata
15. Ramularia_plurivora_CPC_16123 - metadata
16. Ramularia_plurivora_CBS_118743 - metadata
17. Ramularia_sp._CBS_114568 - metadata
18. Ramularia_mali_CBS_129581 - metadata
19. Ramularia_glennii_CBS_120727 - metadata
20. Ramularia_glennii_CPC_16565 - metadata
21. Ramularia_glennii_CPC_16560 - metadata
22. Ramularia_glennii_CPC_18468 - metadata
23. Ramularia_glennii_CBS_122989 - metadata
24. Ramularia_glennii_CBS_129441 - metadata
25. Ramularia_haroldporteri_CBS_137272 - metadata
26. Ramularia_miae_CBS_120121 - metadata
27. Ramularia_miae_CPC_19770 - metadata
28. Ramularia_eucalypti_CBS_120726 - metadata
29. Ramularia_eucalypti_CPC_19187 - metadata
30. Ramularia_eucalypti_CBS_101045 - metadata
31. Ramularia_eucalypti_CBS_120728 - metadata
32. Ramularia_eucalypti_CBS_356.69 - metadata
33. Ramularia_eucalypti_CPC_16804 - metadata
34. Ramularia_eucalypti_CBS_155.82 - metadata

The metadata for that sample will be shown.
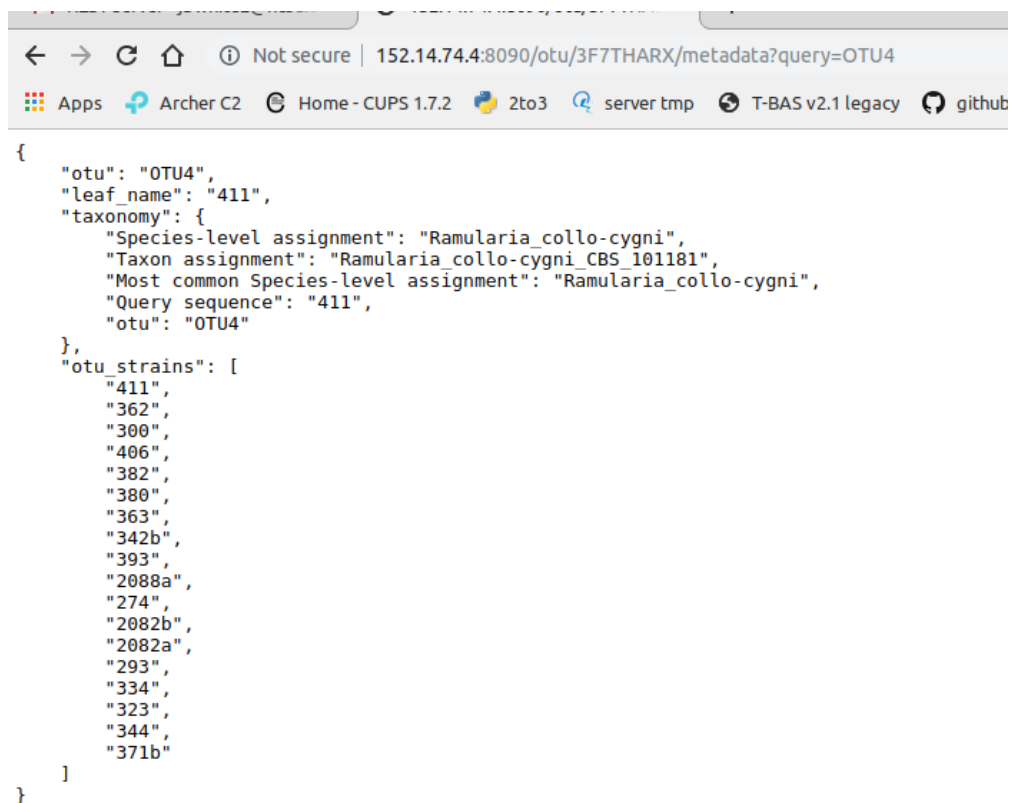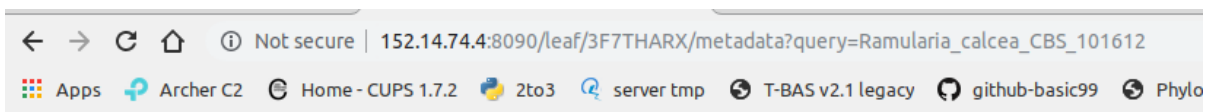
```
{
    "name": "Ramularia_calcea_CBS_101612",
    "GB_TEF1": "KJ504700",
    "Host_ID_3F7THARX": null,
    "Country": "Germany",
    "otu_size": "1",
    "Host_isolation_source": "Symphytum sp.",
    "otu": null,
    "GB_HIS3": "KJ504612",
    "GB_ITS": "KJ504785",
    "Date_collected_3F7THARX": null,
    "GB_ACT": "KJ504449",
    "GB_GAPDH": "KJ504568",
    "Collector": "G. Arnold",
    "Species": "Ramularia_calcea",
    "GB_RPB2": "KJ504656"
}
```

Clicking on "queries" will display the query and the tree placement information:
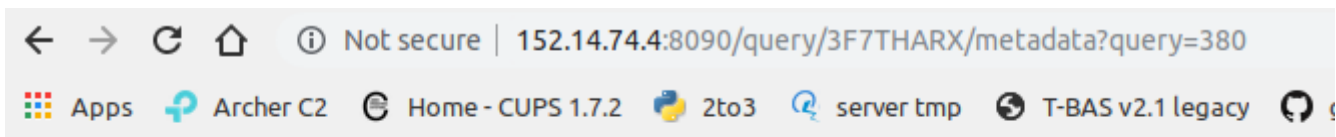
```
{
    "query": {
        "name": "380",
        "Strain": "380",
        "Date collected": "8-Apr-15",
        "Host ID": "SP "
    },
    "placement": {
        "Species-level assignment": "Ramularia_collo-cygni",
        "Taxon assignment": "Ramularia_collo-cygni_CBS_101181",
        "Most common Species-level assignment": "Ramularia_collo-cygni",
        "Query sequence": "411",
        "otu": "OTU4"
    }
}
```

Clicking on OTU will display the following information:

```
{
    "otu": "OTU4",
    "leaf_name": "411",
    "taxonomy": {
        "Species-level assignment": "Ramularia_collo-cygni",
        "Taxon assignment": "Ramularia_collo-cygni_CBS_101181",
        "Most common Species-level assignment": "Ramularia_collo-cygni",
        "Query sequence": "411",
        "otu": "OTU4"
    },
    "otu_strains": [
        "411",
        "362",
        "300",
        "406",
        "382",
        "380",
        "363",
        "342b",
        "393",
        "2088a",
        "274",
        "2082b",
        "2082a",
        "293",
        "334",
        "323",
        "344",
        "371b"
    ]
}
```

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).

RAxML options:

RAxML analysis:

○ EPA with likelihood weights

Fast, only with bifurcating reference tree.

○ Backbone constraint tree with bootstraps

Slow, bifurcating or multifurcating reference tree.

● De novo single or multi-locus phylogenetic analysis

Infer best tree for reference and unknown query sequences. more

Number of bootstrap replicates:

| 100 |

Rate heterogeneity model:

GTRCAT is recommended on large datasets with many taxa.

| GTRGAMMA ⬍ |

DNA substitution model:

Specifying a model here will apply to all DNA partitions and override other models.

| ⬍ |

☑ Use outgroup

Outgroup:

| Ramularia_nyssicola_CBS_127665 |

Ladderize tree:

Sort clades in-place according to the number of terminal nodes. Deepest clades are placed last by default. Use reverse=True to sort clades deepest-to-shallowest.

| ladderize (reverse=False) ⬍ |

Submit   Reset

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.

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
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
MEP	Metadata Enhanced PhyloXML format that is a valid phyloXML with added tags for use in T-BAS and DeCIFR

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