

T-BAS: A Tool for Real-time Tracking of Biodiversity Across the Tree of Life

Ignazio Carbone

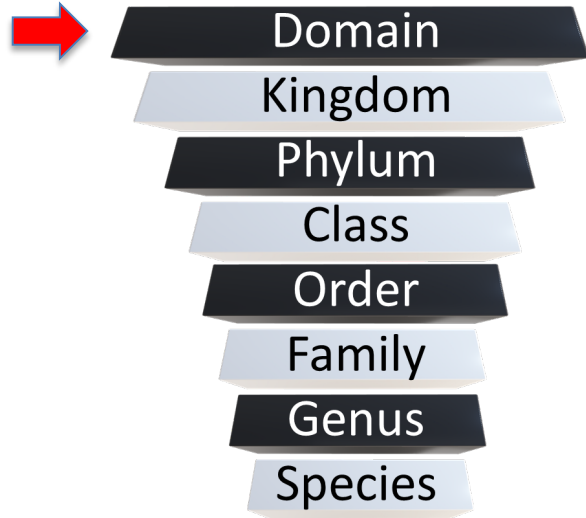
Department of Entomology and Plant Pathology

NC STATE Center for Integrated Fungal Research

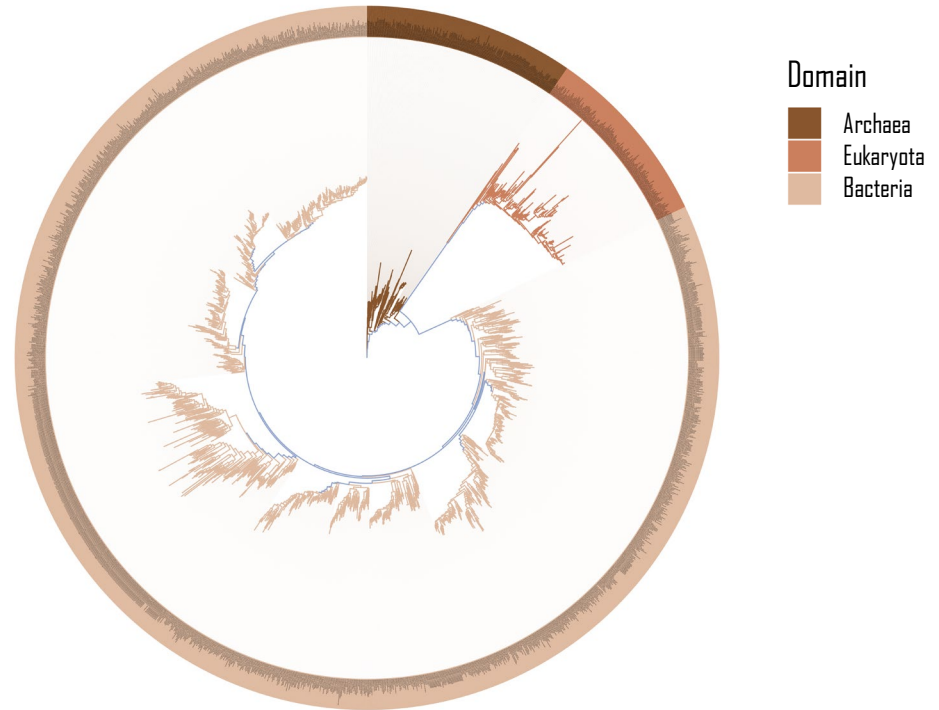
<https://tbas.cifr.ncsu.edu/>

> Taxonomy should reflect phylogeny

Biological Classification



Branches on the Tree of Life



➤ Current tree of life initiatives synthesize published phylogenetic trees along with taxonomic data

No sequence alignments

NCBI Taxonomy Browser

Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy BioCollections

Search for Xylariales as complete name lock Go

Display 0 levels using filter: none

Lineage (full): cellular organisms; Eukaryota; Opisthokonta; Fungi; Dikarya; Ascomycota; saccharomyceta; Pezizomycotina; leotiomyceta; sordariomyceta; Sordariomycetes; Xylariomycetidae

- Xylariales
 - Amphisphaeriaceae
 - Amphisphaeria
 - Amphisphaeria acericola
 - Amphisphaeria camelliae
 - Amphisphaeria curviconidia
 - Amphisphaeria flava
 - Amphisphaeria mangrovei
 - Amphisphaeria micheliae
 - Amphisphaeria parvispora

Open Tree of Life

Xylariales

Legend Zoom tree view

- Xylariaceae
 - Xylaria carpophila
 - Xylaria hypoxylon
 - Acrosphaeria collabens
 - Lichenagaricus crustaceus
 - Lichenagaricus cupressiformis
 - Lichenagaricus scutellus
 - Lichenagaricus scutellus var. albus
 - Lichenagaricus scutellus var. scutellus
 - Moelleroclivus penicilliopeis
 - Penzigia ectinomorpha
 - Penzigia amtzehii
 - Penzigia berteri
 - Penzigia cantarelensis
 - Penzigia carabayensis
 - Penzigia cretacea
 - Penzigia discolor
 - Penzigia enteroleuca
 - Penzigia eterio
 - Penzigia fuscoareolata
 - Penzigia handeli
 - Penzigia hawaiiensis
 - Penzigia indica
 - Penzigia leonensis
 - Penzigia lyogaloides
 - Penzigia mauritanica
 - Penzigia microspora
 - Penzigia orientalis
 - Penzigia olacenta

Royal Botanic Gardens Kew Tree of Life Explorer

9,833 specimens

View options Download tree

- Gnetales Gnetaceae *Gnetum montanum*
- Ephedrales Ephedraceae *Ephedra sinica*
- Ginkgoales Ginkgoaceae *Ginkgo biloba*
- Cycadales Cycadaceae *Cycas micholitzii*
- Pinales Araucariaceae *Araucaria rulei*
- Pinales Cupressaceae *Thuja plicata*
- Pinales Cupressaceae *Sequoia sempervirens*
- Pinales Pinaceae *Cedrus libani*
- Pinales Pinaceae *Picea engelmannii*
- Pinales Pinaceae *Pinus ponderosa*
- Amborellales Amborellaceae *Amborella trichopoda*
- Amborellales Amborellaceae *Amborella trichopoda*
- Amborellales Amborellaceae *Amborella trichopoda*
- Nymphaeales Nymphaeaceae *Nymphaea*

- ID Label
- Sclerotinia_sclerotiorum
 - Botryotinia_fuckeliana
 - Cercospora_sojina
 - Magnaporthe_riisea
 - Verticillium_alfalfae
 - Beauveria_bassiana
 - Neurospora_crassa
 - Myceliophthora_thermophila
 - Thielavia_terrestris
 - Chaetomium_globoseum
 - Cochliobolus_kusanoi
 - Pyrenophora_tritici
 - Neosartorya_fischeri
 - Penicillium_sp
 - Aspergillus_fumigatus
 - Arthroderma_otae
 - Arthroderma_gypseum
 - Tcinocarpus_reesii
 - Ajellomyces_capsulatus
 - Paracoccidioides_brasiliensis



➤ Current tree of life initiatives synthesize published phylogenetic trees along with taxonomic data

No specimen metadata

NCBI Taxonomy Browser

Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy BioCollections

Search for Xylariales as complete name lock Go

Clear

Display 0 levels using filter: none

Lineage (full): cellular organisms; Eukaryota; Opisthokonta; Fungi; Dikarya; Ascomycota; saccharomyceta; Pezizomycotina; leotiomyceta; sordariomyceta; Sordariomycetes; Xylariomycetidae

- Xylariales *Click on organism name to get more information.*
 - Amphisphaeriaceae

Open Tree of Life Add / browse trees Feedback About

Xylariales

Show comments Legend Zoom tree view

- Xylariales
 - Xylaria carpophila
 - Xylaria hypoxylon
 - Acrosphaeria collabens
 - Lichenagaricus crustaceus
 - Lichenagaricus cupressiformis
 - Lichenagaricus scutellus
 - Lichenagaricus scutellus var. albus
 - Lichenagaricus scutellus var. scutellus
 - Moelleroclavus penicilliopeis
 - Penzigia sedinomorpha
 - Penzigia amtzehii
 - Penzigia berfordi
 - Penzigia cantarelrensis
 - Penzigia carabayensis
 - Penzigia cretacea

Royal Botanic Gardens Kew Tree of Life Explorer

9,833 specimens View options Download tree

- Gnetales Gnetaceae *Gnetum montanum*
- Ephedrales Ephedraceae *Ephedra sinica*
- Ginkgoales Ginkgoaceae *Ginkgo biloba*
- Cycadales Cycadaceae *Cycas micholitzii*
- Pinales Araucariaceae *Araucaria rulei*
- Pinales Cupressaceae *Thuja plicata*

specimen_metadata_MixS_definitions.xlsx

strain_name	biotic_relationship	specific_host	geo_loc_name	project_name	source_mat_id	collection_date
Name assigned to genetic variant of a microorganism.	Is it free-living or in a host and if the latter what type of relationship is observed	If there is a host involved, please provide its taxid (or environmental if not actually isolated from the dead or alive host - i.e. pathogen could be isolated from a swipe of a bench etc) and report whether it is a laboratory or natural host). From this we can calculate any number of groupings of hosts (e.g. animal vs plant, all fish hosts, etc)	The geographical origin of the sample as defined by the country or sea name followed by specific region name. Country or sea names should be chosen from the INSDC country list (http://insdc.org/country.html), or the GAZ ontology (v 1.512) (http://purl.bioontology.org/ontology/GAZ)	Name of the project within which the sequencing was organized	A unique identifier assigned to a material sample (as defined by http://rs.tdwg.org/dwc/terms/materialSampleID), and as opposed to a particular digital record of a material sample) used for extracting nucleic acids, and subsequent sequencing. The identifier can refer either to the original material collected or to any derived subsamples. The INSDC qualifiers /specimen_voucher, /bio_material, or /culture_collection may or may not share the same value as the source_mat_id field. For instance, the /specimen_voucher qualifier and source_mat_id may both contain 'UAM:Herps:14', referring to both the specimen voucher and sampled tissue with the same identifier. However, the /culture_collection qualifier may refer to a value from an initial culture (e.g. ATCC13775) while source_mat_id would refer to an identifier from some derived culture from which the nucleic acids were extracted (e.g. xatc123 or ark:/2154/R2).	The time of sampling, either as an instance (single point in time) or interval. In case no exact time is available, the date/time can be right truncated i.e. all of these are valid times: 2008-01-23T19:23:10+00:00; 2008-01-23T19:23:10; 2008-01-23; 2008-01; 2008; Except: 2008-01; 2008 all are ISO8601 compliant

> Current tree of life initiatives synthesize published phylogenetic trees along with taxonomic data

No placement of unknown sequences

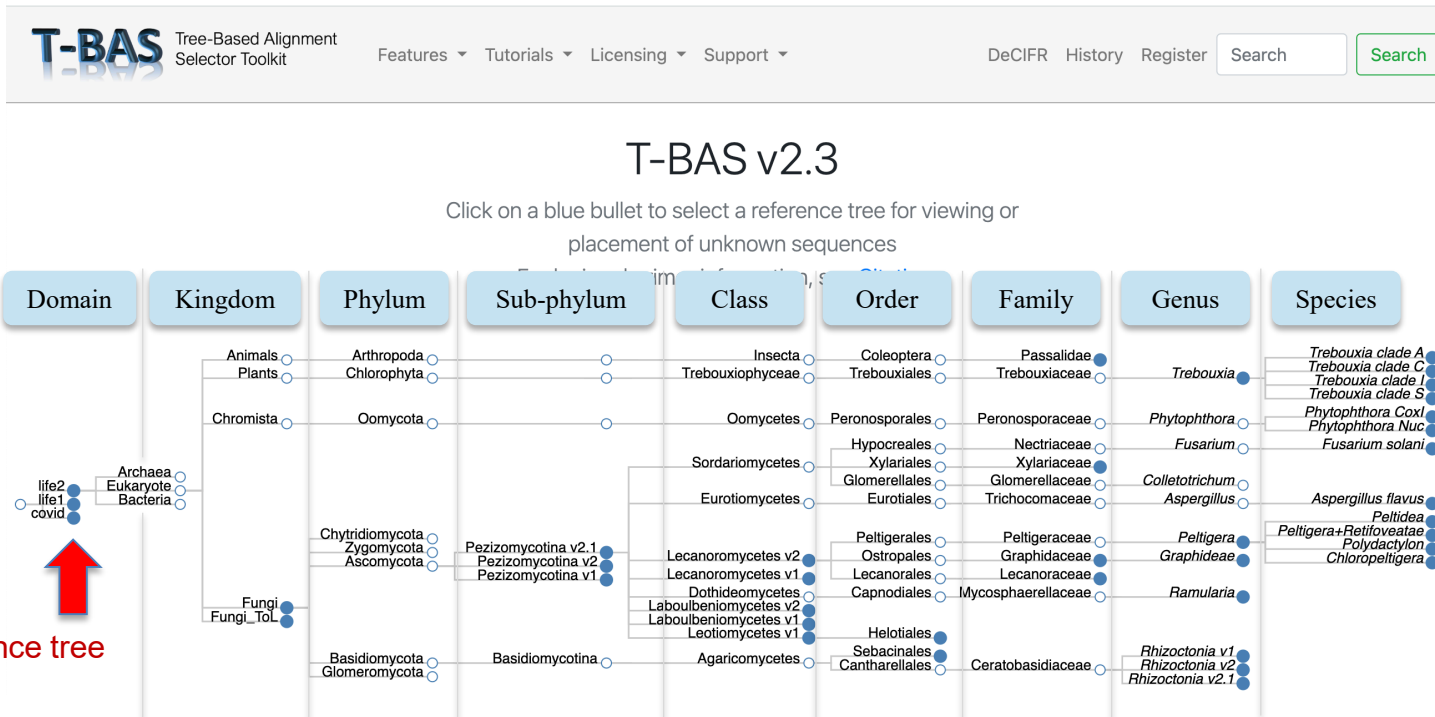
Lineage (full): cellular organisms; Eukaryota; Opisthokonta; Fungi; Dikarya; Ascomycota; saccharomyceta; Pezizomycotina; leotiomyceta; sordariomyceta; Sordariomycetes; Xylariomycetidae

o **Xylariales** *Click on organism name to get more information.*

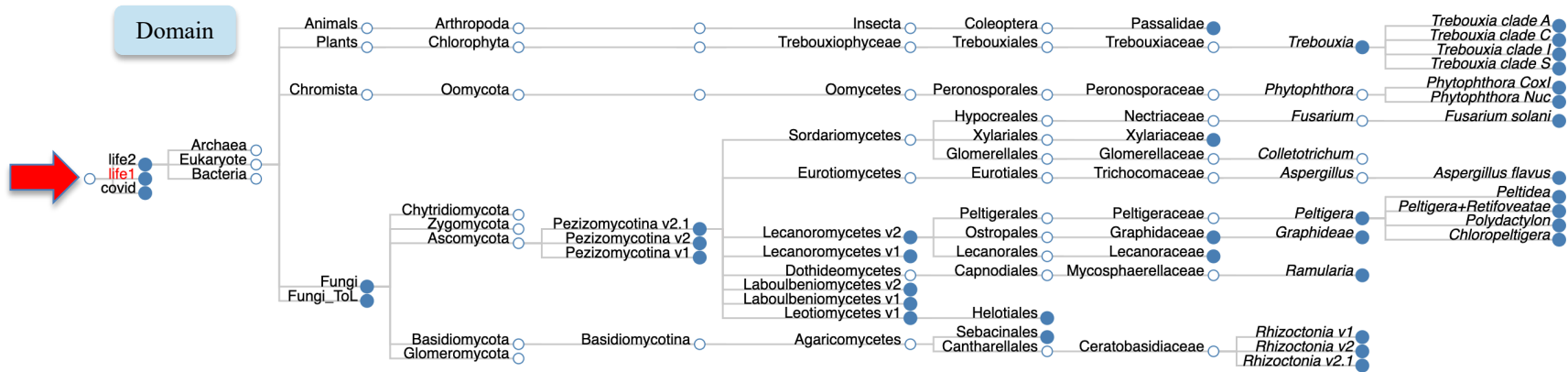
- o **Amphisphaeriaceae**
 - o **Amphisphaeria**
 - o **Amphisphaeria acericola**
 - o **Amphisphaeria camelliae**
 - o **Amphisphaeria curviconidida**
 - o **Amphisphaeria flava**
 - o **Amphisphaeria mangrovei**
 - o **Amphisphaeria micheliae**
 - o **Amphisphaeria parvispora**
 - o **Amphisphaeria sorbi**
 - o **Amphisphaeria thailandica**
 - o **Amphisphaeria umbrina**
 - o **Amphisphaeria unispinata**
 - o **Amphisphaeria yunnanensis**
 - o **unclassified Amphisphaeria**
- o **Arecophila**
 - o **Arecophila australis**
 - o **Arecophila bambusae**
 - o **Arecophila clypeata**
 - o **Arecophila miscanthei**
 - o **Arecophila muroiana**
 - o **unclassified Arecophila**
- o **Capsulospora**
 - o **unclassified Capsulospora**
- o **Ciferriascosea**
 - o **Ciferriascosea fluctamurum**
 - o **Ciferriascosea rectamurum**
- o **Discostroma**
 - o **Discostroma stoneae**
 - o **Discostroma tostum**
 - o **unclassified Discostroma**



➤ Phylogenetic integration of unknown sequences with reference trees, alignments and metadata from cultured specimens

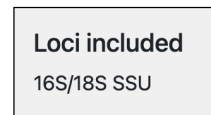
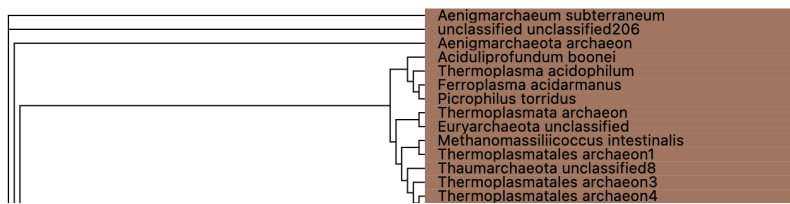
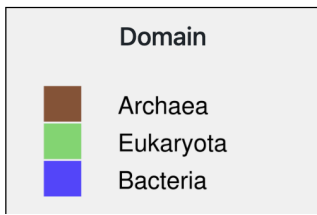


> Single-locus phylogeny-based placement of 16S/18S rRNA

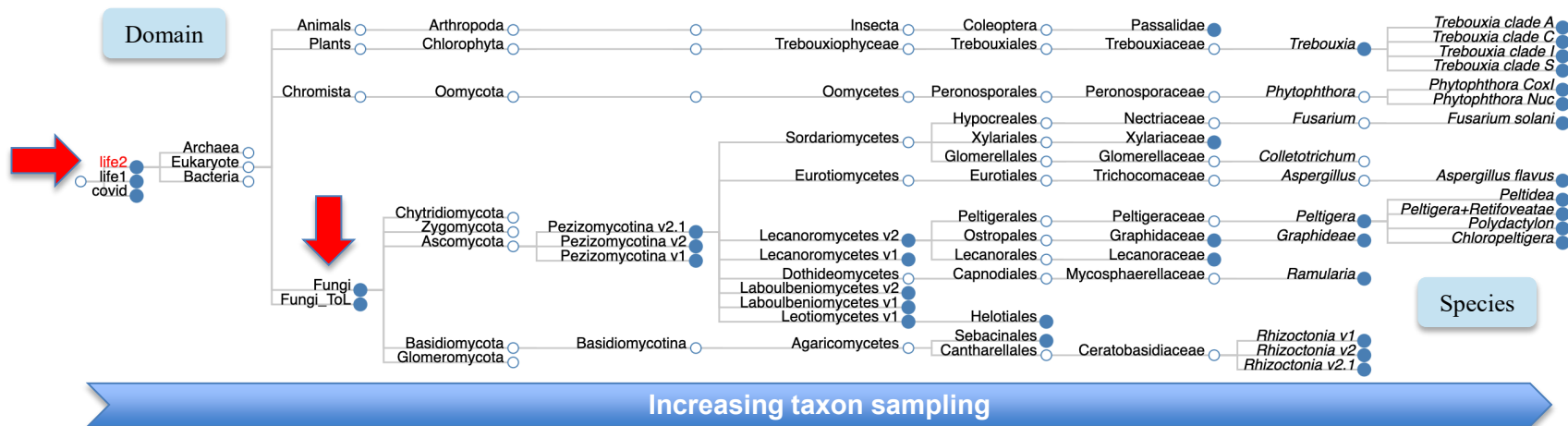


- Targeted Amplicon Sequencing
- Genomes
- Metagenomes

16S/18S SSU reference tree



> Multi-locus phylogeny-based placement of 16 ribosomal proteins



- Multi-locus phylogeny
- Genome-scale phylogeny

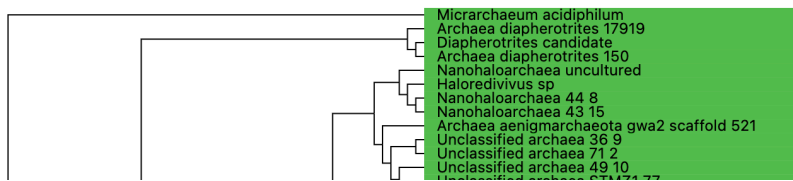
multilocus ribosomal protein reference tree

[View Tree Data](#)

[Place Unknowns](#)

Domain

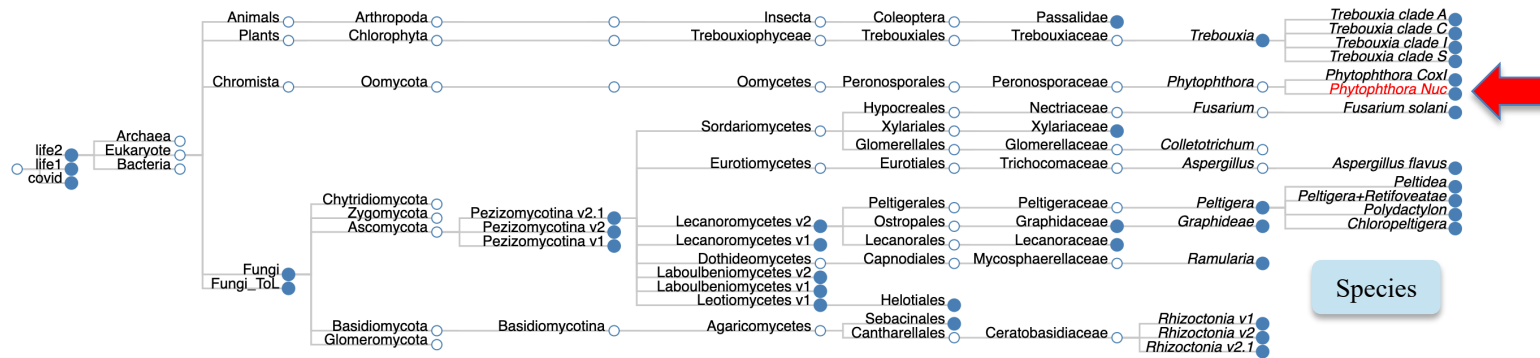
- Archaea
- Eukaryota
- Bacteria



Loci included

- L2p_L8e (16 total)
- L3p_L3e
- L4p_L1e
- L5p_L11e

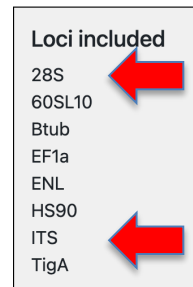
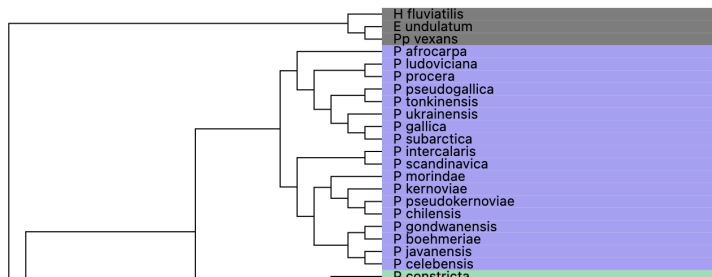
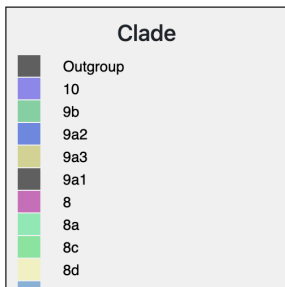
> Multi-locus phylogeny-based placement for species delimitation



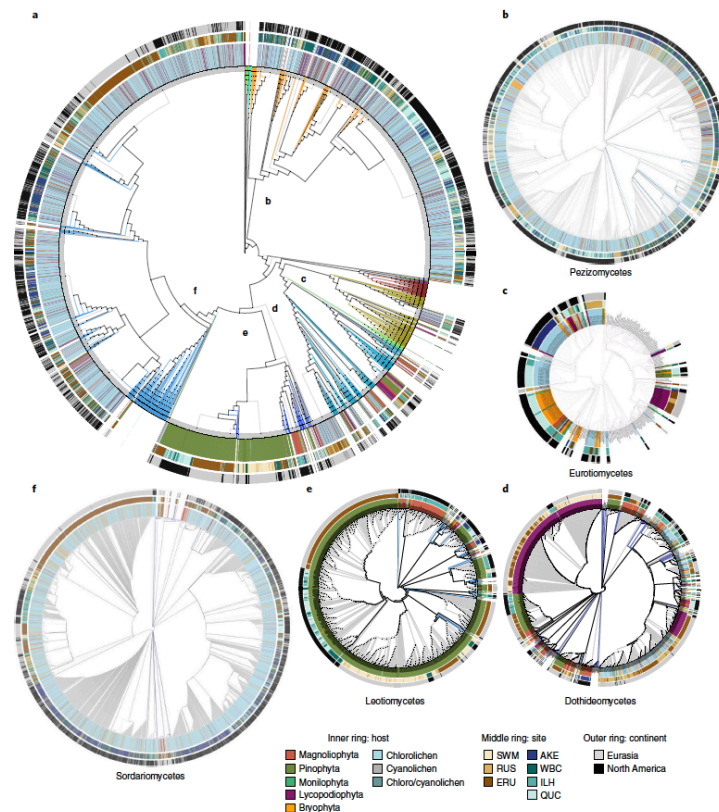
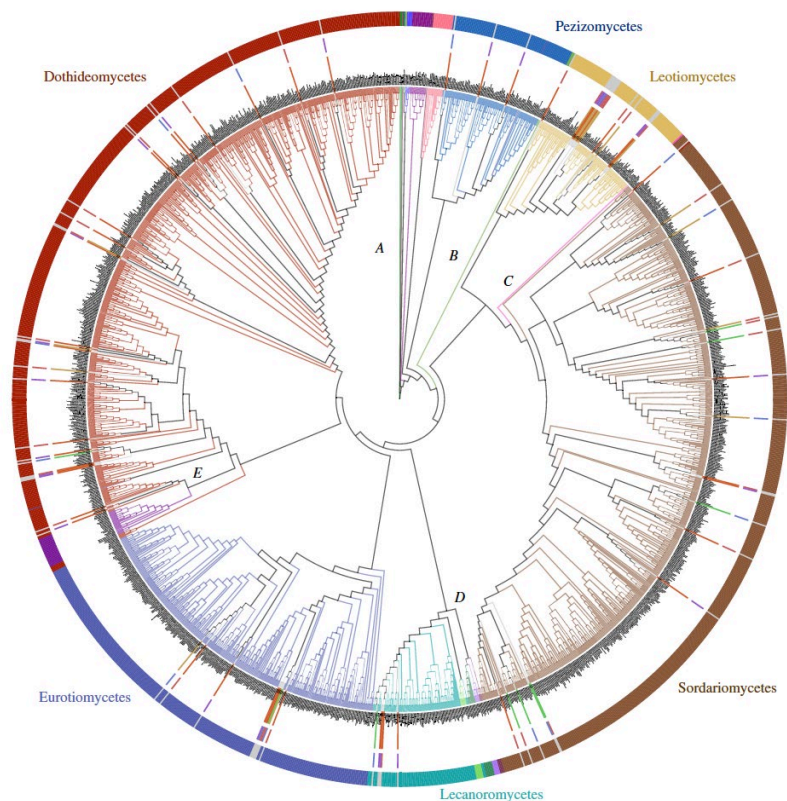
- Multi-level placement
- Population genetics/genomics

Phytophthora Nuclear

Coomber, A., Saville, A., Carbone, I. and J. B. Ristaino.
2023. An open-access T-BAS phylogeny for emerging
Phytophthora species. PLoS One In Press



> Integration of taxonomic information, alignments, and collection metadata



> T-BAS tracks metadata in phylogenies using MEP data standard

T-BAS Tree-Based Alignment Selector Toolkit

Features ▾ Tutorials ▾ Licensing ▾ Support ▾

DeCIFR History Register

	A	B	C
1	strain_name		
2	biotic_relationship		
3	specific_host		
4	geo_loc_name		
5	project_name		
6	source_mat_id		
7	collection_date		
8	env_material		
9	env_biome		
10			
11			

specimen_metadata_MixS_definiti +

Metadata Enhanced PhyloXML (MEP)

Authors

- Jim White
- Ignazio Carbone



Description

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. MEP is based on XML, 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 for validating MEP files. PhyloXML includes a phylogeny element that saves the tree information and associated alignments. MEP extends this by adding (1) a tag to each clade that is a leaf in the tree and saving the metadata for that leaf, (2) 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, and (3) an OTUs tag that saves the taxonomic assignments, associated query metadata and sequences for each OTU.

MEP uses two associated schema definitions:

<https://decifr.cifr.ncsu.edu/schema.php>

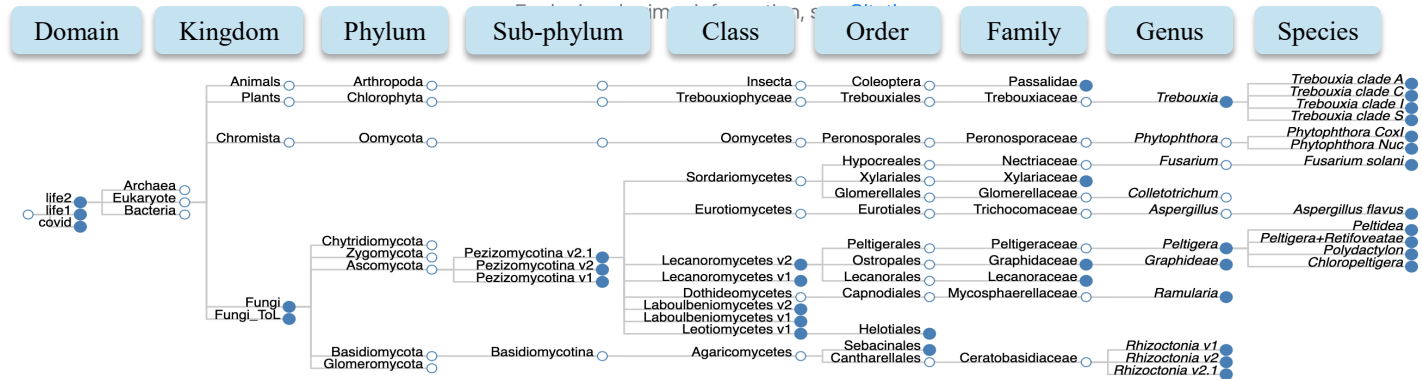
The screenshot shows the homepage of the Genomic Standards Consortium (GSC). The header includes the GSC logo and navigation links: Home, About, Board, Projects, Publications, Meetings, Jobs, News, Contact. The main content area features a welcome message: "The Genomic Standards Consortium (GSC) is an open-membership working body formed in September 2005. The aim of the GSC is making genomic data discoverable. The GSC enables genomic data integration, discovery and comparison through international community-driven standards." Below this, there are three columns of updates: "MixS" with links to learn about MixS standards and download them; "News" with several announcements including GSC22 (Thailand) postponement, Save The Date for GSC22, GSC21 Meeting, GSC20 Agenda & Logistics, and GSC20 Observations of Helicobacter; and "Twitter" with a tweet from @genstandards about the missing ORCID IDs.

> T-BAS is a dynamic and open-access tree of life

- > Features a phylogenomic database for taxonomy and metadata validation
- > Real-time tracking of biodiversity across taxonomic ranks
- > Species delimitation and population genetic analysis based on multi-locus genomic data
- > Can be used to track and catalog the uncultivated microbial diversity at all taxonomic ranks

T-BAS v2.3

Click on a blue bullet to select a reference tree for viewing or placement of unknown sequences



T-BAS version 2.3: phylogenetic-data repository and tool for phylogeny-based placement

Funding

- > NSF Dimensions of Biodiversity
- > NSF Genealogy of Life
- > NSF Ecology and Evolution of Infectious Diseases
- > Novo Nordisk Foundation Collaborative Crop Resilience Program
- > NSF Predictive Intelligence for Pandemic Prevention

Collaborators

- > Cyberinfrastructure for Phylogenetic Research (CIPRES)



novo nordisk fonden

 **ACCESS** | Advancing Innovation

