

Open Tree and phylogenetic updating

Emily Jane McTavish

University of California, Merced
ejmctavish@ucmerced.edu, Twitter: @snacktavish



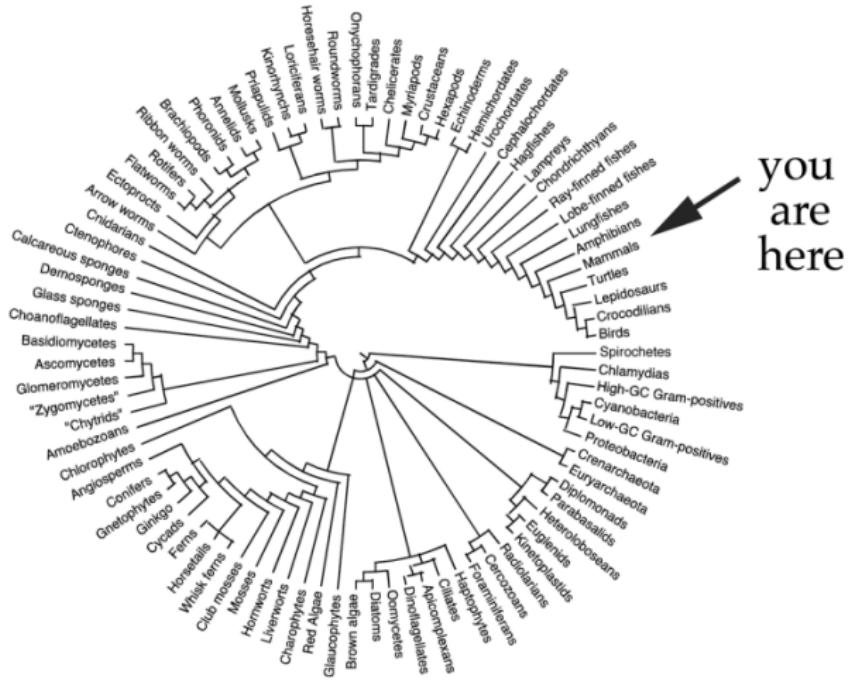


Image Ethan Hein

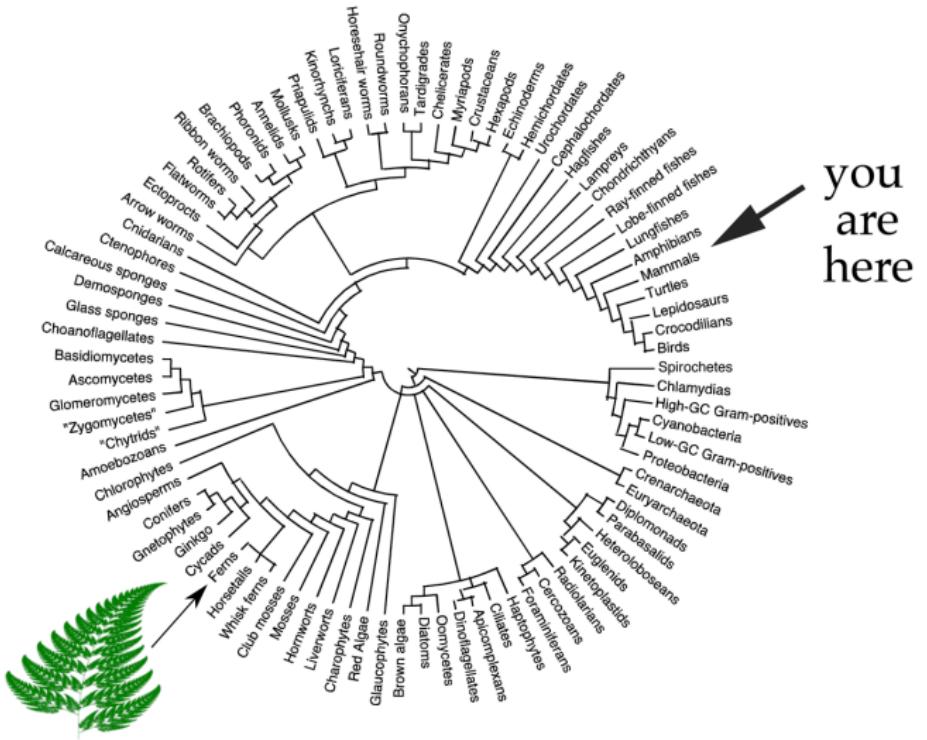
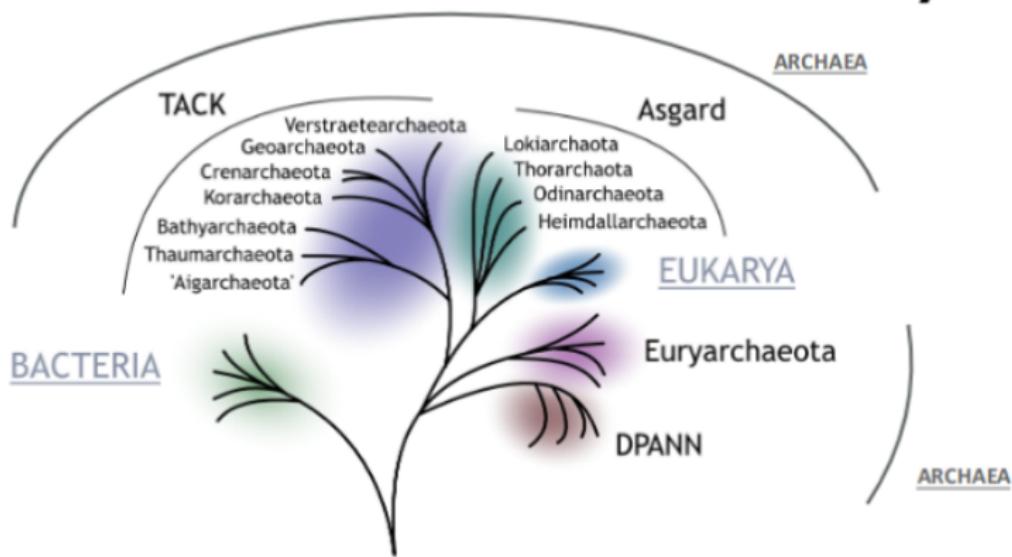


Image Ethan Hein

Why do we need phylogenies?

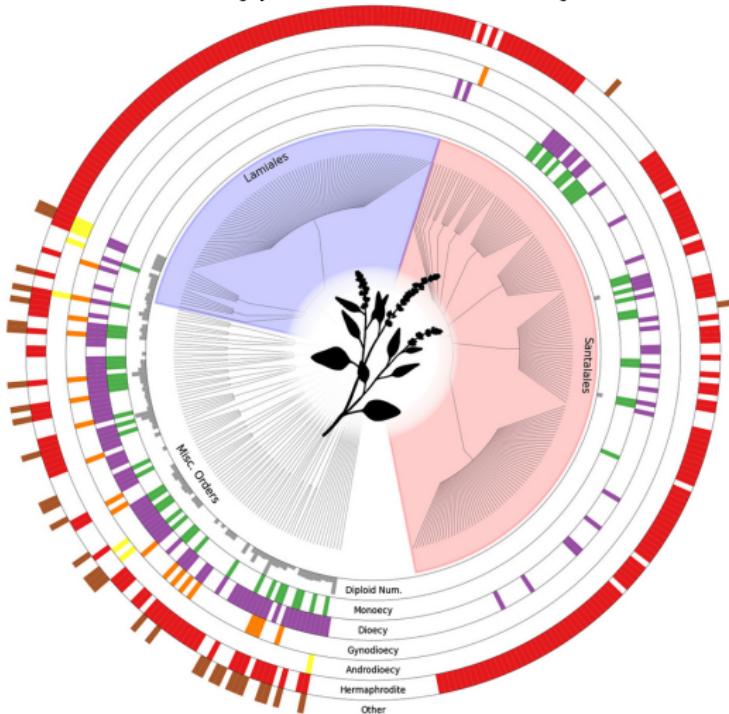
To understand the shared evolutionary history of life on earth



2017

Figure from Laura Eme

To understand rates and types of evolutionary transitions



The Tree of Sex Consortium, (2014) Scientific Data



Goal: Build a tree of all life.

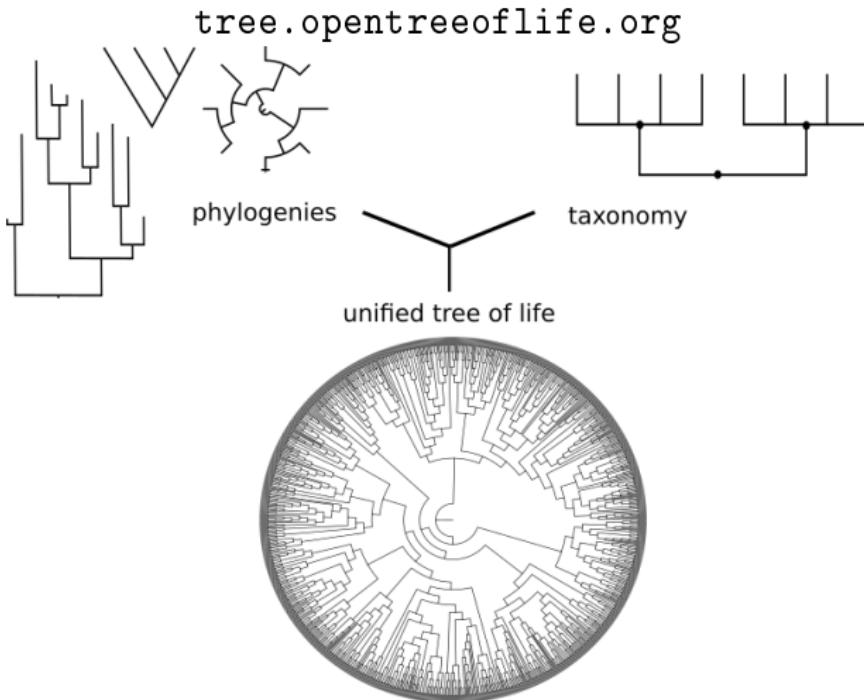


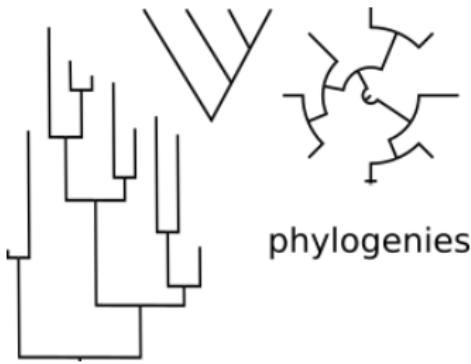
Goal: Build a tree of all life.

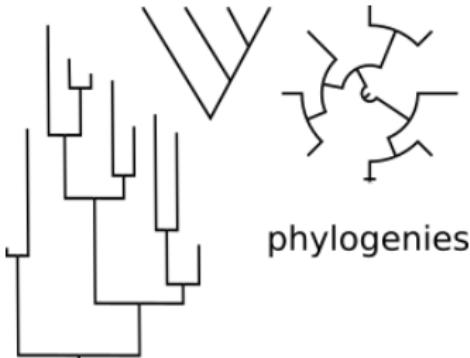
Every named species

Updated as new data becomes available

Freely and easily accessible







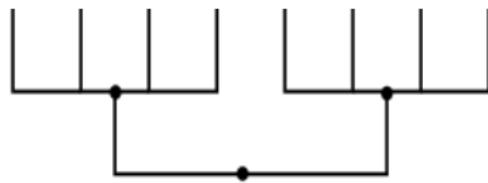
Current synthetic tree

987 representative phylogenies

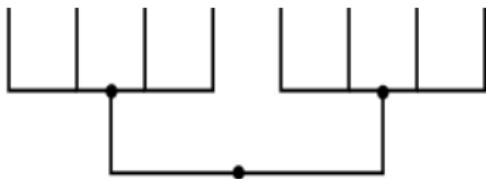
65,662 tips from phylogeny

New draft with more input trees every few months

Redelings and Holder, PeerJ 2017



taxonomy



taxonomy

2.7 million named taxa

Merges NCBI, Index fungorum, Silva, IRMNG, GBIF and other taxonomies

Scaffold for combining ranked phylogenetic estimates

New drafts released as inputs change

Rees and Cranston, Biodiversity Data Journal 2017



Open Tree taxonomy: **Metrosideros robusta**

The current taxonomy version is [ott3.0 \(click for more information\)](#). See the OTT wiki for [an explanation of the taxon flags used](#) below, e.g., `extinct`

Taxon details

species **Metrosideros robusta** [ncbi:101983](#) ([gbif:3185294](#)) (OTT id 284291)

[View this taxon in the current synthetic tree](#)

Synonym(s)

[Nania robusta](#), [Metrosideros florida](#)

Lineage

[life](#) > [cellular organisms](#) > [Eukaryota](#) > [Archaeplastida](#) > [Chloroplastida](#) > [Streptophyta](#) > [Embryophyta](#) > [Tracheophyta](#) > [Euphylllophyta](#) > [Spermatophyta](#) > [Magnolophyta](#) > [Mesangiospermae](#) > [eudicotyledons](#) > [Gunneridae](#) > [Pentapetalae](#) > [rosids](#) > [malvids](#) > [Myrales](#) > [Myrtaceae](#) > [Myrtoideae](#) > [Metrosidereae](#) > [Metrosideros](#)

Taxonomic amendments

New taxa can be added from uploaded trees, and will be included in future synthetic trees

Opportunity to feed-back to input taxonomic resources

Adding new taxa

Once added, these taxa will appear in the Open Tree Taxonomy, and possibly in the synthetic tree, with links to your curator profile, the current study, and any additional sources that you provide below. [Hide](#)

Selected label 1 of 1 [Previous label](#) [Next label](#) * required fields

Original label [Use as taxon name](#)

New taxon name * No duplicates found.

Taxonomic rank *

Parent taxon * [Zygodontomys — Open in OTT browser](#)
 in
 Use this parent taxon for all labels (un-check to edit)

Source(s) for this taxon *

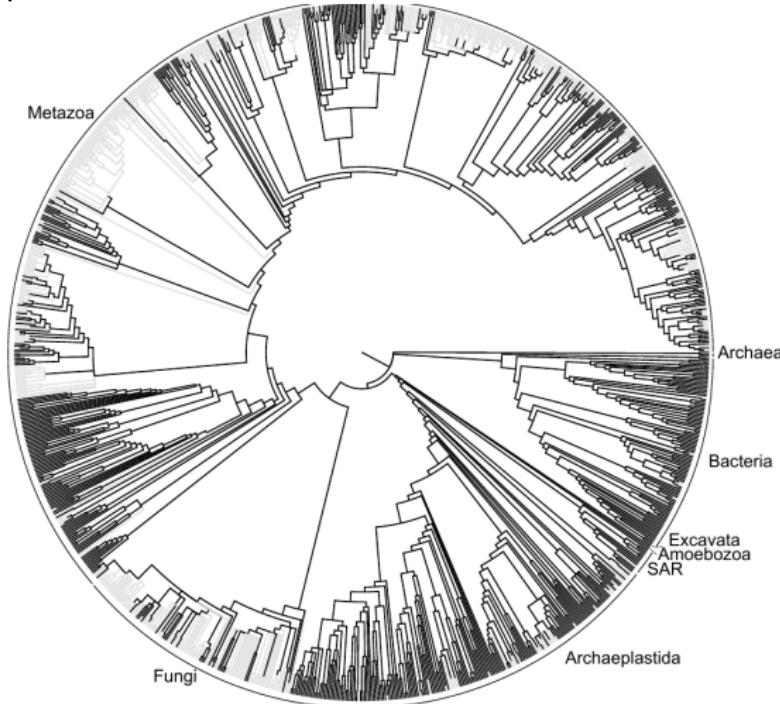
The taxon is described in this study

Use this source information for all labels (un-check to edit) [Add another source](#)

[Cancel](#) [Submit](#)

The synthetic tree (Hinchliff et al., PNAS 2015)

2.4 million species



Dark lineages have at least one representative in an input source tree

How will we fill in the gaps?

Need to build on existing phylogenetic information.

Need to build on existing phylogenetic information.

OPEN  ACCESS Freely available online



Perspective

Lost Branches on the Tree of Life

Bryan T. Drew^{1*}, Romina Gazis², Patricia Cabezas^{3,4}, Kristen S. Swithers⁵, Jiabin Deng¹, Roseana Rodriguez¹, Laura A. Katz⁵, Keith A. Crandall⁴, David S. Hibbett², Douglas E. Soltis^{1,6}

1 University of Florida, Gainesville, Florida, United States of America, **2** Clark University, Worcester, Massachusetts, United States of America, **3** Brigham Young University, Provo, Utah, United States of America, **4** George Washington University, Washington, DC, United States of America, **5** Smith College, Northampton, Massachusetts, United States of America, **6** Florida Museum of Natural History, Gainesville, Florida, United States of America

Drew et al. PLoS Biology 2013

only 16% of phylogenies published 2000-2013 are digitally available

Drew et al. PLoS Biology 2013

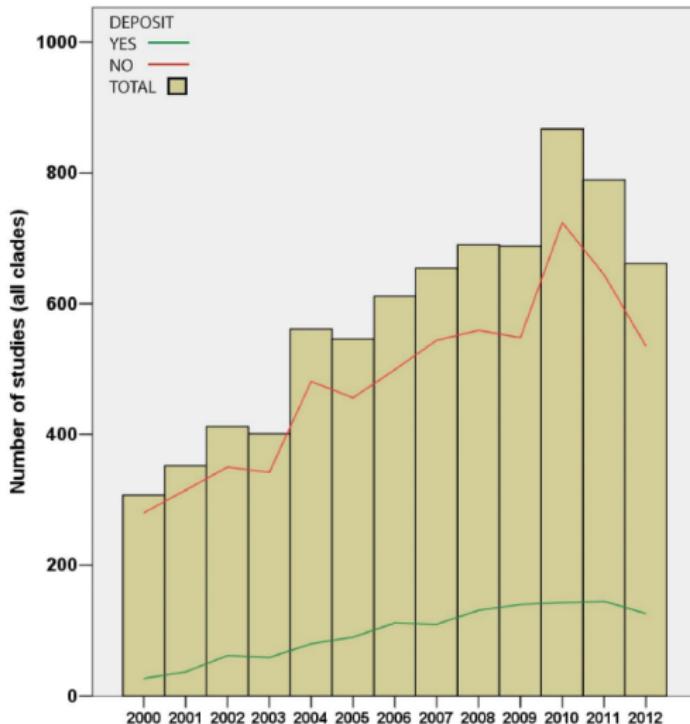


Figure 1. Overview of total number of publications surveyed from animal, fungus, seed plant, microbial eukaryote, archaea, and bacteria literature (indicated in red), and the number of those publications that archived their trees and alignments in either Dryad or TreeBASE (indicated in green).

doi:10.1371/journal.pbio.1001636.g001

only 16% of phylogenies published 2000-2013 are digitally available (Drew et al. PLoS Biology 2013)

20% of phylogenies published 2013-2018 (McTavish et al. BioEssays 2018)

Adding phylogenetic data

- Trees can be uploaded from any source, does not have to be own data.
- Easy to use browser based interface
- Track curation attribution by name or pseudonym
- Files are json representation of NeXML phylogenetic data format
- Data store is hosted publicly on GitHub

github.com/OpenTreeOfLife/phylesystem-1

McTavish et al. Bioinformatics 2015



Community Curation

253 individual curators of 4,431 uploaded studies

Community Curation

253 individual curators of 4,431 uploaded studies

Rapid curation progress at taxon focused in-person working groups, in collaboration with FuturePhy

Community Curation

253 individual curators of 4,431 uploaded studies

Rapid curation progress at taxon focused in-person working groups, in collaboration with FuturePhy

Currently a several month lag for incorporation into synthetic tree, will begin daily builds in the next year

Automated tree updating

Rapid accumulation of sequence data

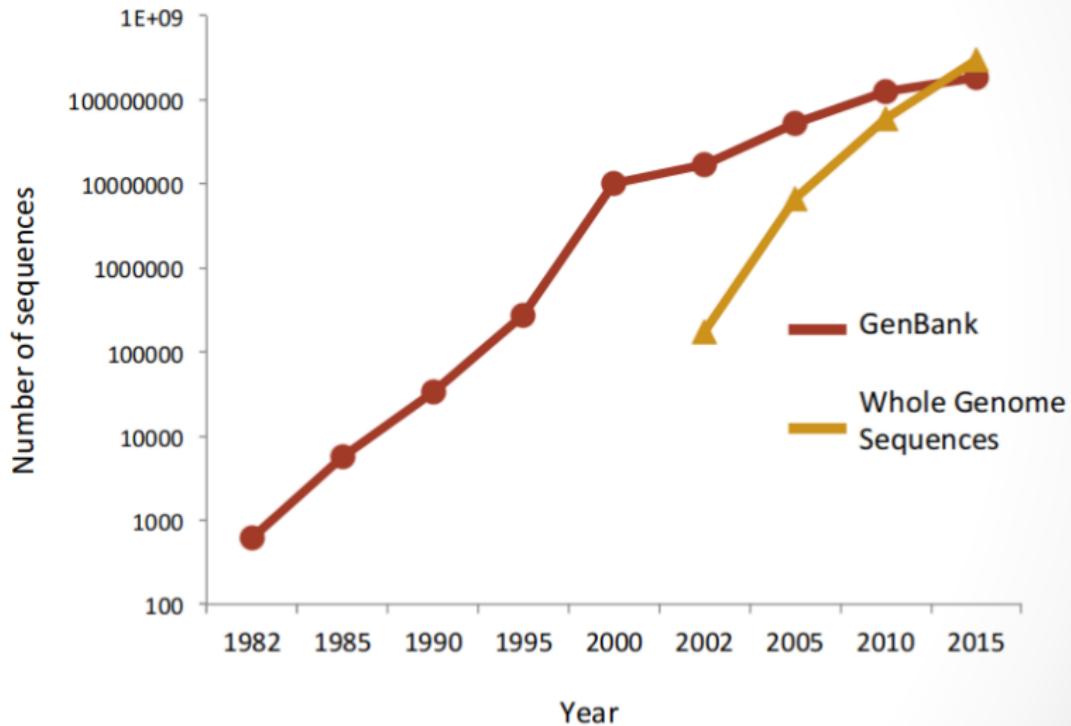


Figure from Belinda Chang

Often time lag of months to years between sequencing and inclusion in phylogenetic estimates!

Streamlining the process

Inputs: An existing alignment and phylogeny, and a database of sequences or reads

Output: A maximum likelihood phylogenetic estimate including new taxa

Problems with automated phylogenetics

Homology/paralogy

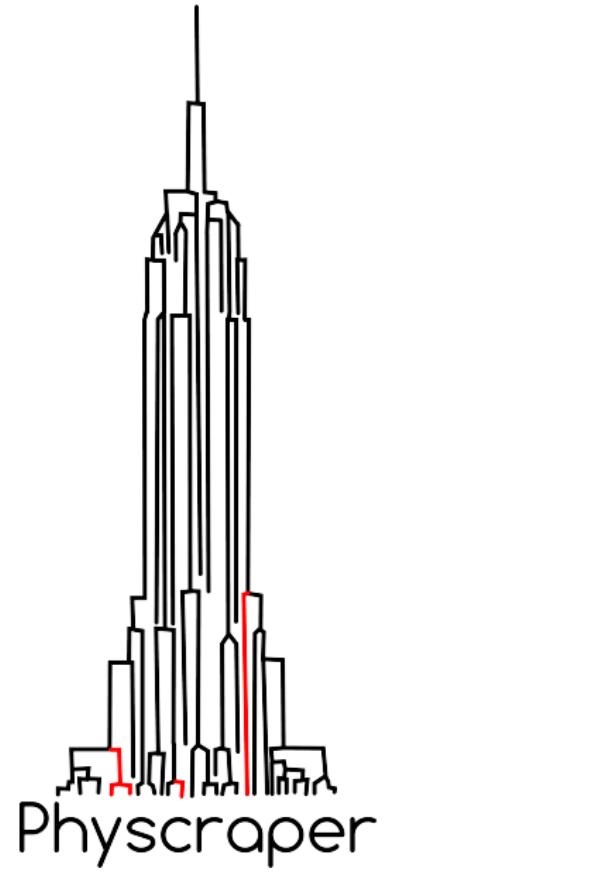
Data collection and assembly

Alignment

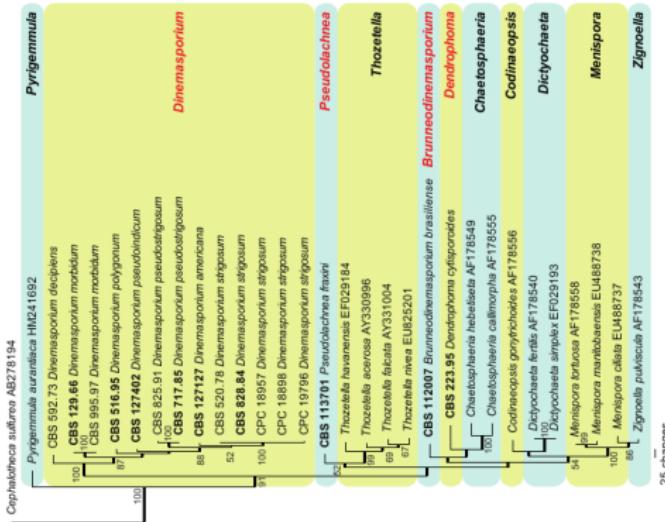
Physcraper

Automated updating of
existing phylogenies

[https://github.com/
McTavishLab/
physcraper](https://github.com/McTavishLab/physcraper)



Restrict taxon addition to in-group
Uses loci that have been developed by taxon experts



e.g. using ITS to understand relationships within ascomycota
Crous et al. 2012 Persoonia

Data collection and assembly

Blast existing loci for assembled sequences and use their NCBI taxon id as label

The screenshot shows the NCBI BLAST search interface. At the top, there's a navigation bar with links for Home, Recent Results, Saved Strategies, and Help. Below the navigation bar, a banner says "BLAST finds regions of similarity between biological sequences." A search input field is present with a placeholder "Enter organism name or id—completions will be suggested". To the right of the input field is a "GO" button. On the left, under "BLAST Assembled Genomes", there's a list of organisms with checkboxes: Human, Mouse, Rat, Cow, Pig, Dog, Rabbit, Chimp, Guinea pig, Fruit fly, Honey bee, Zebrafish, Clawed frog, Arabidopsis, Rice, Yeast, and Microbes. On the right side of the interface, there's a sidebar titled "Your Recent Results" with a link to "All Recent results...". Below that is a "News" section featuring "SmartBLAST" and a "Tip of the Day" section.

Basic BLAST

Choose a BLAST program to run.

- | | |
|----------------------------------|---|
| nucleotide blast | Search a nucleotide database using a nucleotide query
Algorithms: blastn, megablast, discontiguous megablast |
| protein blast | Search protein database using a protein query
Algorithms: blastp, psi-blast, phi-blast, delta-blast |
| tblastx | Search protein database using a translated nucleotide query |
| tblastn | Search translated nucleotide database using a protein query |
| tblastz | Search translated nucleotide database using a translated nucleotide query |

Specialized BLAST

Choose a type of specialized search (or database name in parentheses.)

- Get faster protein results with a graphical view using [SmartBLAST](#)
- Make specific primers with [Primer-BLAST](#)
- Cluster multiple sequences together with their database neighbors using [MOLE-BLAST](#)
- Find [conserved domains](#) in your sequence (cds)
- Find sequences with similar [conserved domain architecture](#) (cdat)
- Search sequences that have [gene expression profiles](#) (GEO)
- Search [immunoglobulins](#) and [T cell receptor sequences](#) (igBLAST)
- Screen sequence for [vector contamination](#) (vecscren)

Data collection and assembly

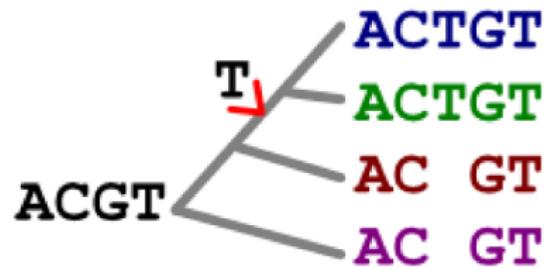
Alternately, use aligned sequences to directly assemble loci of interest from whole genome sequencing reads.

- Decreases bias due to selection of reference

- Whole genome assembly is computationally expensive, and results in data being discarded

- Captures read quality and polymorphism

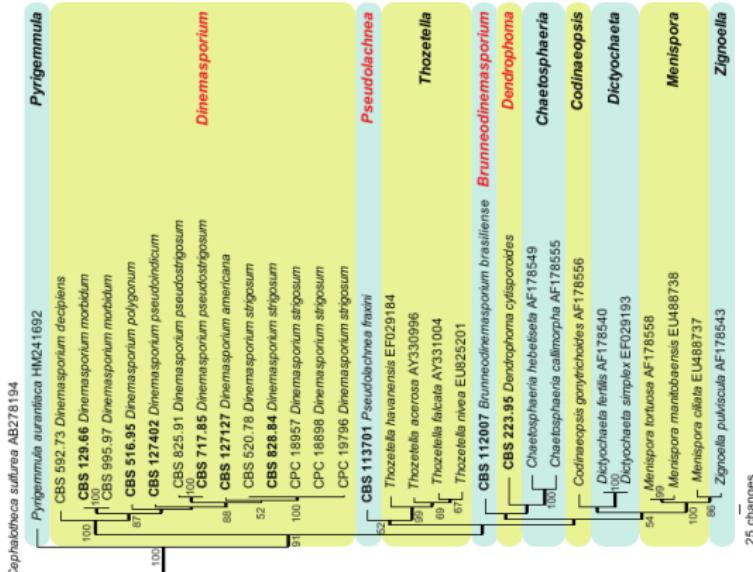
Using existing phylogeny as a guide tree can inform multiple sequence alignment



<http://www.ebi.ac.uk/goldman-srv/prank/differences/>

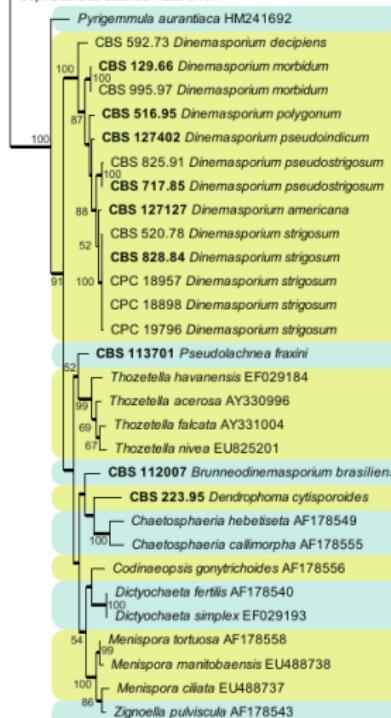
Perform a full maximum likelihood tree search

Improve convergence using previous estimate as starting tree



Original tree (36 taxa)

Cephalotheca sulfurea AB278194



25 changes

Pyrigemmula

Dinemasporium

Pseudolachnea

Thozetella

Chaetosphaeria

Codinaeopsis

Dictyochaeta

Menispora

四

ZigBee

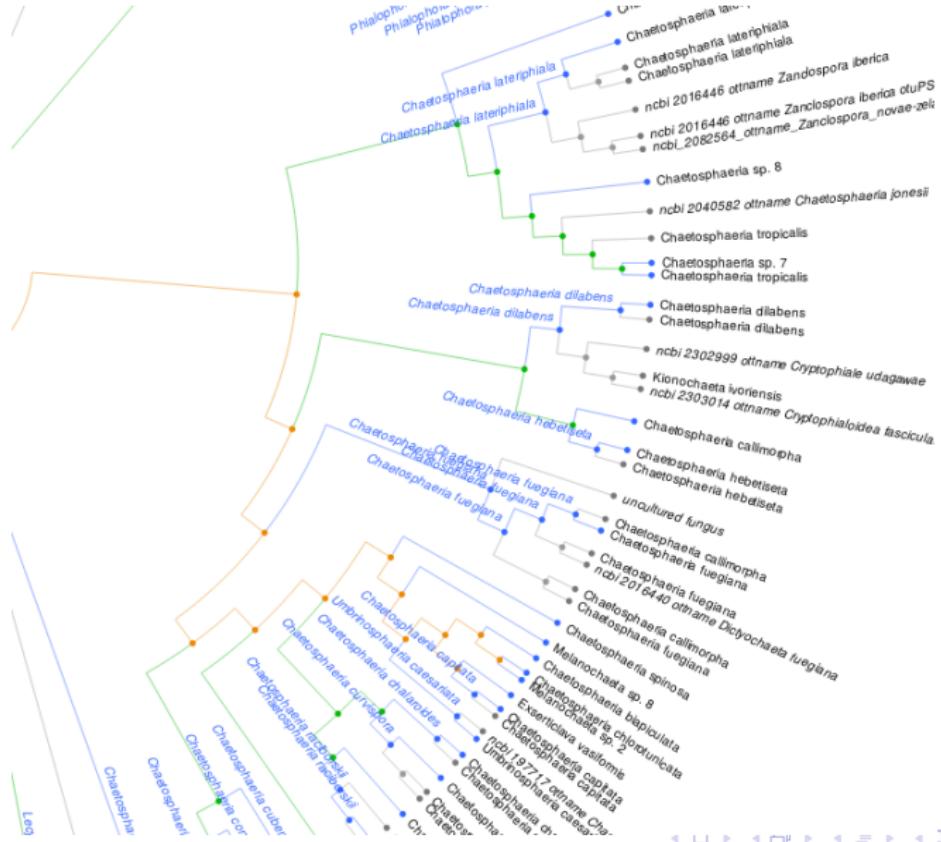
Updated tree (100 taxa)



This is exploratory!

but can flag potential areas of interest.

e.g.



Advantages

- Rapid data-to-phylogeny loop
- Prioritize further data collection
- Apply data collected for other projects
- Minimize researcher time input
- Stream taxa into draft of synthetic tree

Future directions

- Test effect of new data on alignment
- Develop new tree search algorithms that leverage previous search results
- Use placements to inform a divide-and-conquer approach

Phylogenetic updating code

available at:

<https://github.com/>

McTavishLab/physcraper

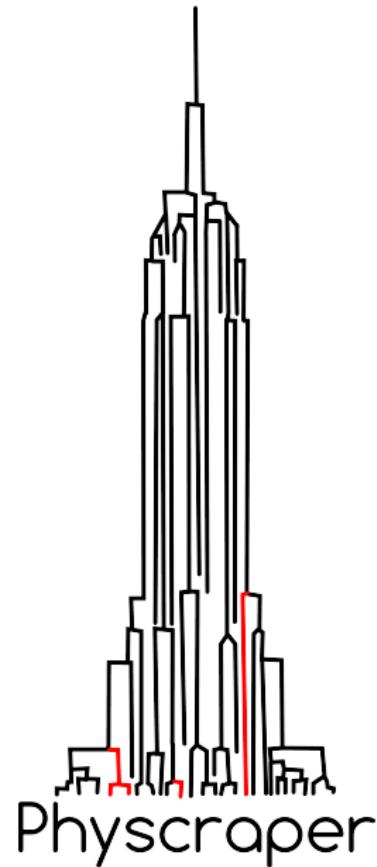
(searches GenBank for
homologous loci)

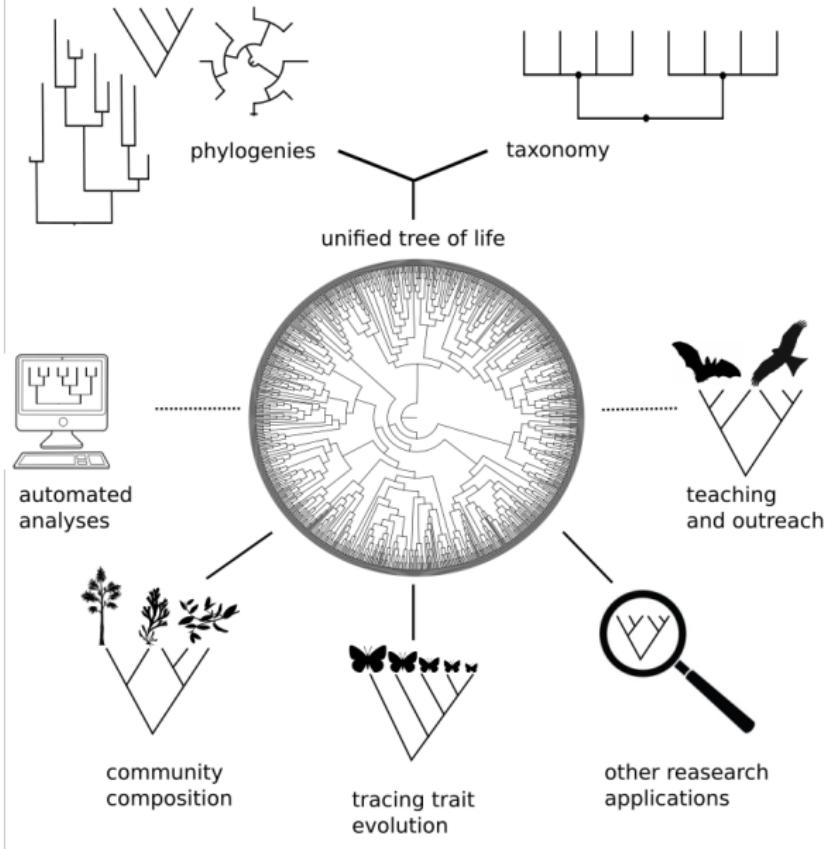
<https://github.com/>

McTavishLab/phycorder

(assembles homologous loci from
short read data)

github
SOCIAL CODING





(McTavish et al. Bioessays 2017)

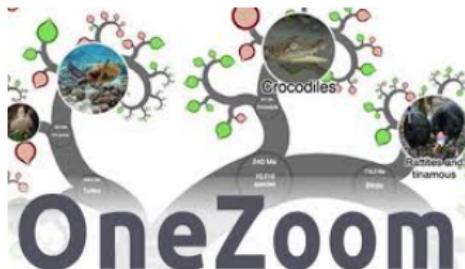
Open Tree resources are available via a range of implementations

- Browser interface, `tree.opentreeoflife.org`
- Open Tree of Life API
-  python wrapper
- R Open Tree of Life (rotl) 

CC0 license provides fully open access for downstream re-usability

CC0 license provides fully open access for downstream re-usability

Open Tree provides the tree backend for:



New this year! <https://glouwa.github.io/>

The phylogenetic tree illustrates the evolutionary relationships within the Felidae family and its sister groups. The tree is rooted at the bottom and branches upwards, with the Felidae family being the most recent common ancestor of all the genera shown. The tree is color-coded by family: Mustelidae (green), Canidae (orange), Viverridae (blue), and Felidae (grey). Small icons of animal illustrations are placed near their respective branches.

Panthera

Panthera is a genus within the Felidae family that was named and first described by the German naturalist Oken in 1816.^[2] The British taxonomist Pocock revised the classification of this genus in 1916 as comprising the species lion, tiger, jaguar, and leopard on the basis of cranial features.^[3] Results of genetic analysis indicate that the snow leopard also belongs to the Panthera, a classification that was accepted by IUCN assessors in 2008.^{[4][5]}

Panthera^[1]
Temporal range: Late Miocene – present, 5.95–0 Ma

Tiger (*Panthera tigris*), the largest species of the genus *Panthera*

Radial bone of *Panthera* fossil

Scientific classification

Kingdom: Animalia

Coming soon!

Automated updating with new genetic data

Branch lengths / Node ages

Private data stores

Custom synthesis

Infrastructure improvements

Holder, McTavish  ABI

Cranston 

Conclusions

Phylogenetic estimates should be freely accessible and reusable
Open Tree cross-links phylogenetic and taxonomic information
A variety of tools and approaches provides wide access to
Open Tree resources

Lab today:

- Browser interface, tree.opentreeoflife.org
- Standardizing taxon names
- Getting existing trees for arbitrary sets of taxa
- Visualizing conflict between estimates, by uploading to OpenTree
- Updating an existing phylogeny with new data from GenBank

Contribute your knowledge!
tree.opentreeoflife.org/curator



Thank You



NSF ABI 1759846

Mark Holder

Karen Cranston



NSF AVATOL 1208809

AVATOL PI'S: Burleigh,
Crandall, Cranston, Gude,
Hibbett, Holder, Katz, Ree,
Smith, Soltis, Williams

Dendropy Jeet Sukumaran

Lab group:

Martha Kandziora

Luna Luisa Sanches Reyes

Lesly Lopez Fang

Jasper Toscani-Field



