

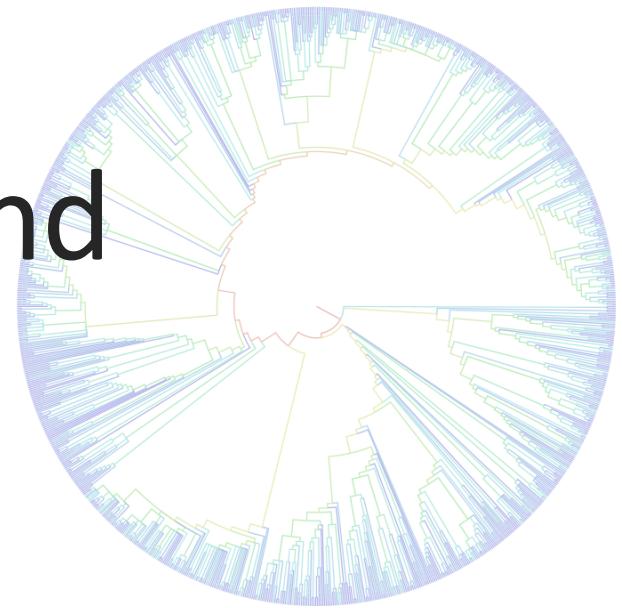
# A comprehensive and dated seed plant phylogeny

STEPHEN A. SMITH



AND JOSEPH BROWN

UNIVERSITY OF MICHIGAN

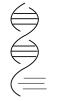
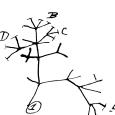


# I use phylogenies to address:

## Empirical

-  **Biodiversity:** The origin and evolution of biodiversity
-  **Biogeography:** Why do species live where they live
-  **Tempo and mode:** What are the rates of evolution
-  **Phylogenomics and evo/eco:** Molecular evolution and ecological adaptations

## Methodological

-  Genomes and transcriptomes for evolutionary biology and ecology
-  **Phylogenetic methods**
  - How do we know what we know
  - Divergence time estimation
  - Biogeography, niches, and phylogenies

# What would we like to do?

We would like phylogenies that are

- Accurate and precise
- Updated
- Have branch lengths relative to time
- Comprehensive

# Open Tree of Life

- Open Tree of Life (NSF AVATOL)
  - 11 PIs working collaboratively
- > 2.3 million species, extinct and extant
- Synthesize data from existing sources (trees and datasets)
- Allow for annotations, comments, and comparison



# Open Tree of Life synthetic tree

- constructed taxonomy (OTT) with only evolutionary lineages
- combined 484 studies (from our database of over 7000)
- 2,339,460 tips
- 37,525 tips are informed by phylogeny (many inform deeper edges)
- opentreeoflife.org for adding trees or browsing synthesis
  - Continues to be updated and improved

## How to Read the Circle of Life

Primordial life begins at the center and branches out in all directions, leading to the groups of species that exist today (colored rings).

Outer ring: Estimated proportion of all species

Inner ring: Proportion of the groups named to date

Each black line represents at least 500 descendant species

Dark lines: Many species have been genetically sequenced

Light lines: Few species have been genetically sequenced

Nematodes (roundworms)

Lophotrochozoa (mollusks, segmented worms, brachiopods)

Deuterostomia (vertebrates, sea stars and urchins, non-round worms)

Early diverging metazoa (cnidaria, comb jellies, sponges)

Many deuterostomia (gold) and plants (green) are already genetically sequenced (dark lines) because they are culturally or economically important (such as humans!)

Fungi

Plants

Arthropods (insects, arachnids, crustaceans)

Scientists have identified about one million arthropods (tan); millions more remain undescribed

Experts expect that most new species to be discovered will be bacteria (orange) and archaea (magenta)

Archaea (single-celled micro-organisms that tolerate extreme conditions)

Bacteria

SARs (diatoms, amoeboids, brown algae)

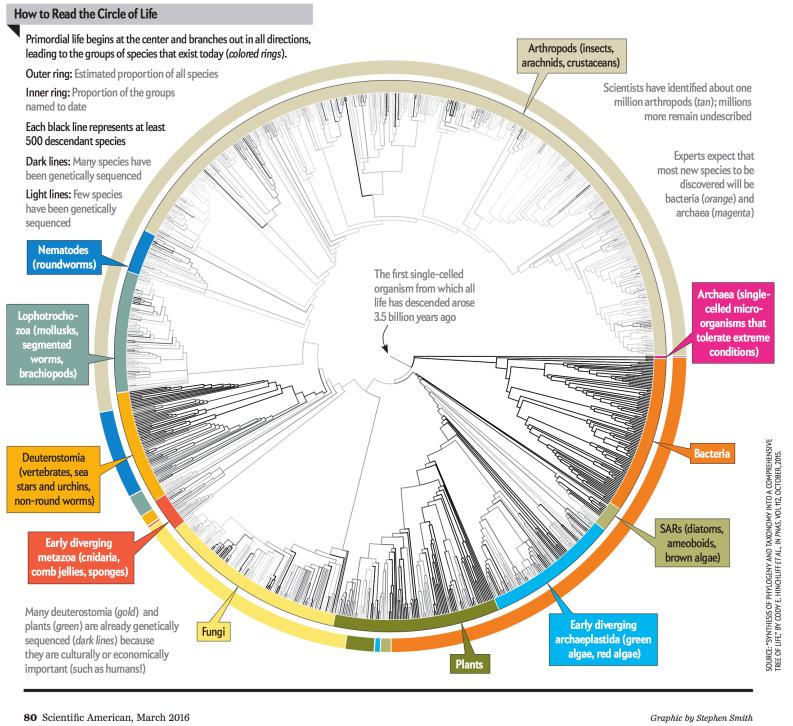
Early diverging archaeplastida (green algae, red algae)

Informed by taxonomy

Informed by phylogenies

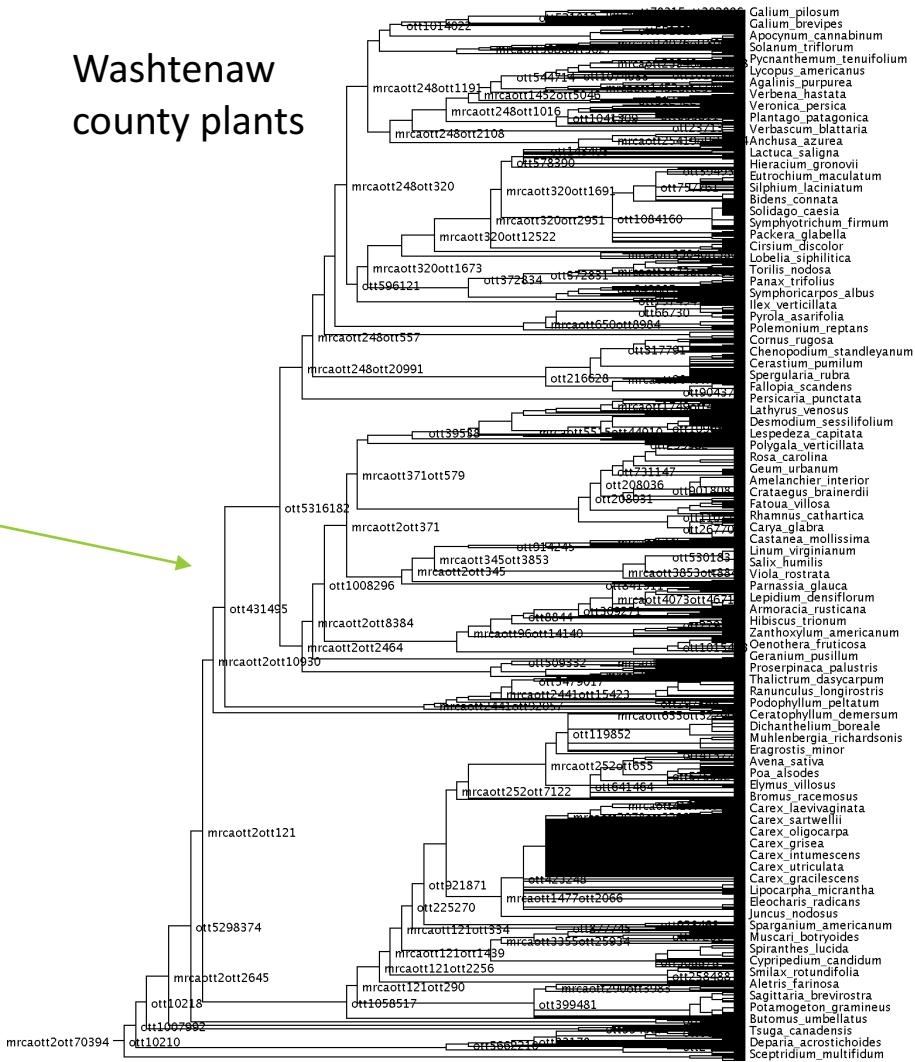
SOURCE: "SYNTHESIS OF PHYLOGENY AND TAXONOMY INTO A COMPREHENSIVE TREE OF LIFE," BY COOKE HINCHLIFF ET AL., IN *PNAS*, VOL 112, OCTOBER 2015.

# Phylogenies of the plants of Michigan

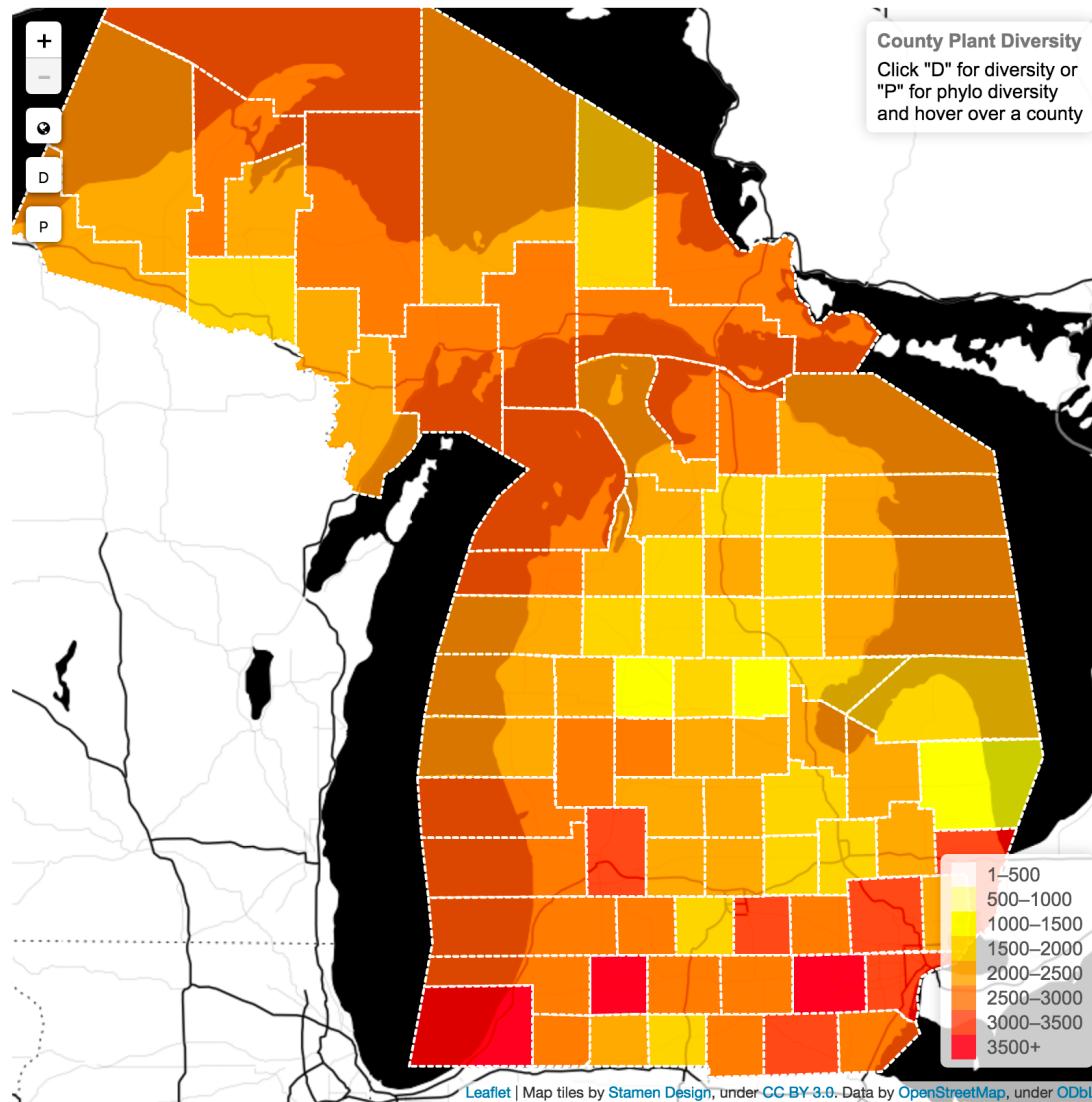


80 Scientific American, March 2016

## Washtenaw county plants



# Michigan phylogenetic diversity



# Comprehensive and dated tree for seed plants

---



LET'S SEE A BIG PLANT TREE

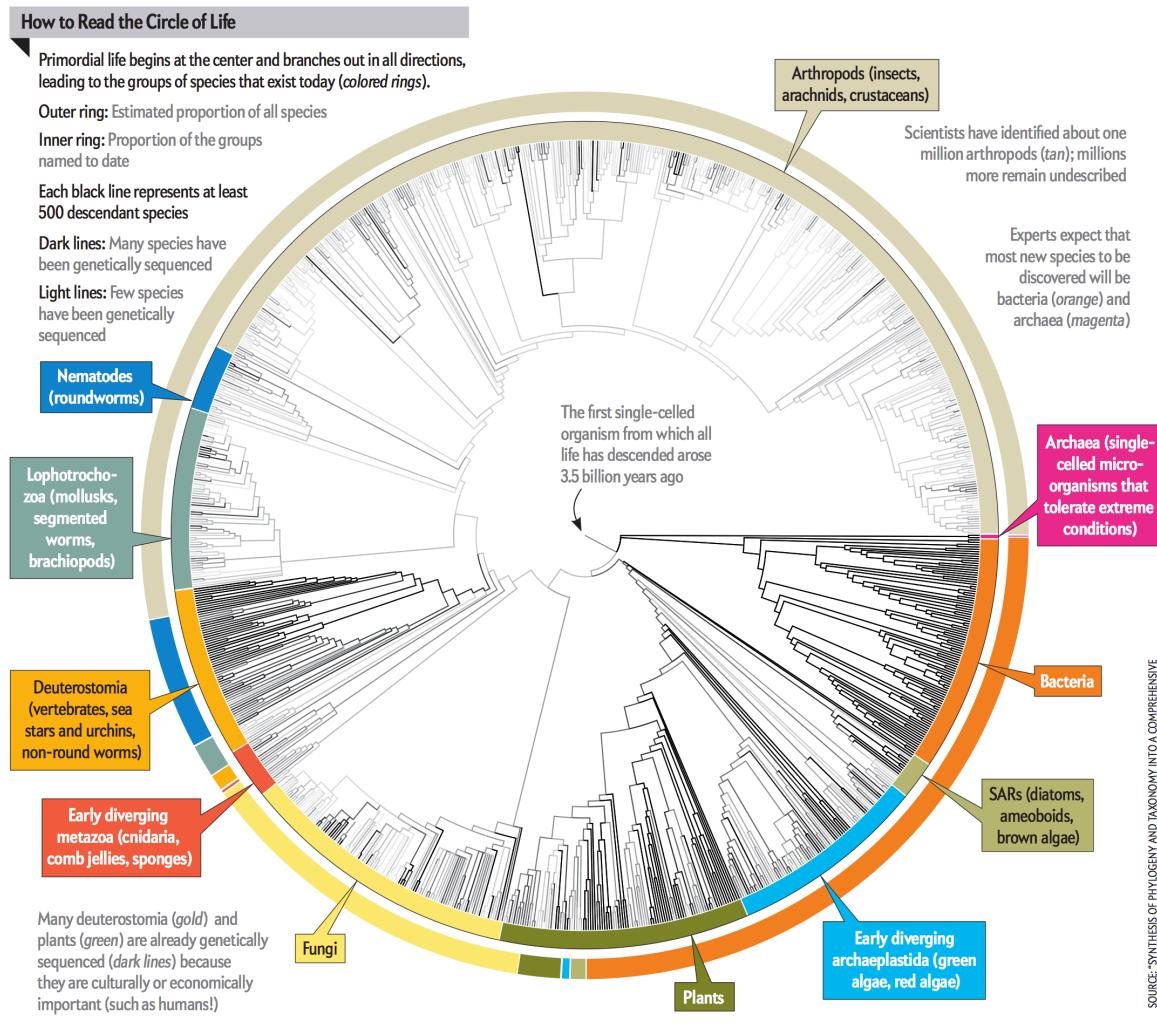
# Tree from Open Tree of Life

## Benefits

- Comprehensive (or nearly so)
- Resolution in many well sampled areas
- Can be updated

## Limitations

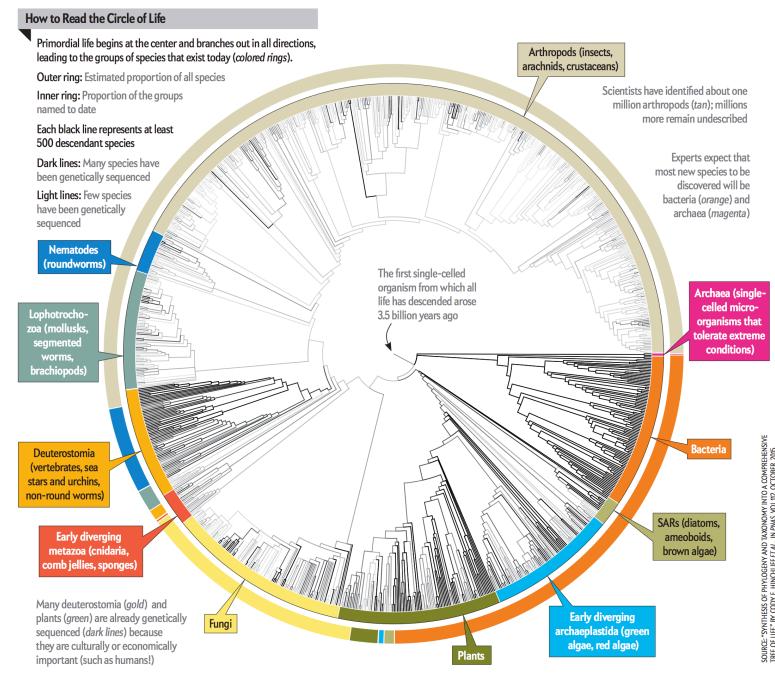
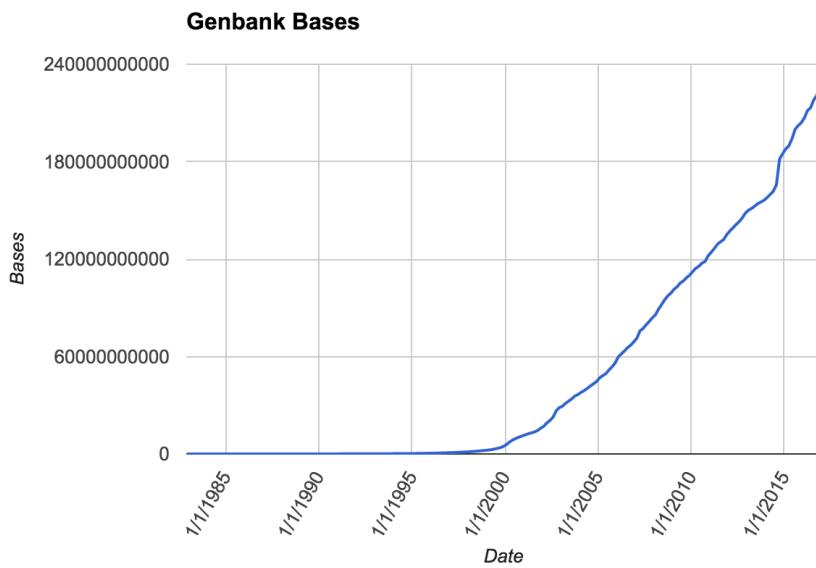
- Resolution may be poor
- No divergence times or branch lengths



# Combine GenBank and Open Tree

We developed a technique that will allow us to combine the efforts of the Open Tree of Life and the public data on GenBank

This is implemented in PyPHLAWD (Smith et al. in prep; Smith and Brown submitted)



# Procedure

## Get a list of clades

- We construct trees for monophyletic orders (roughly)

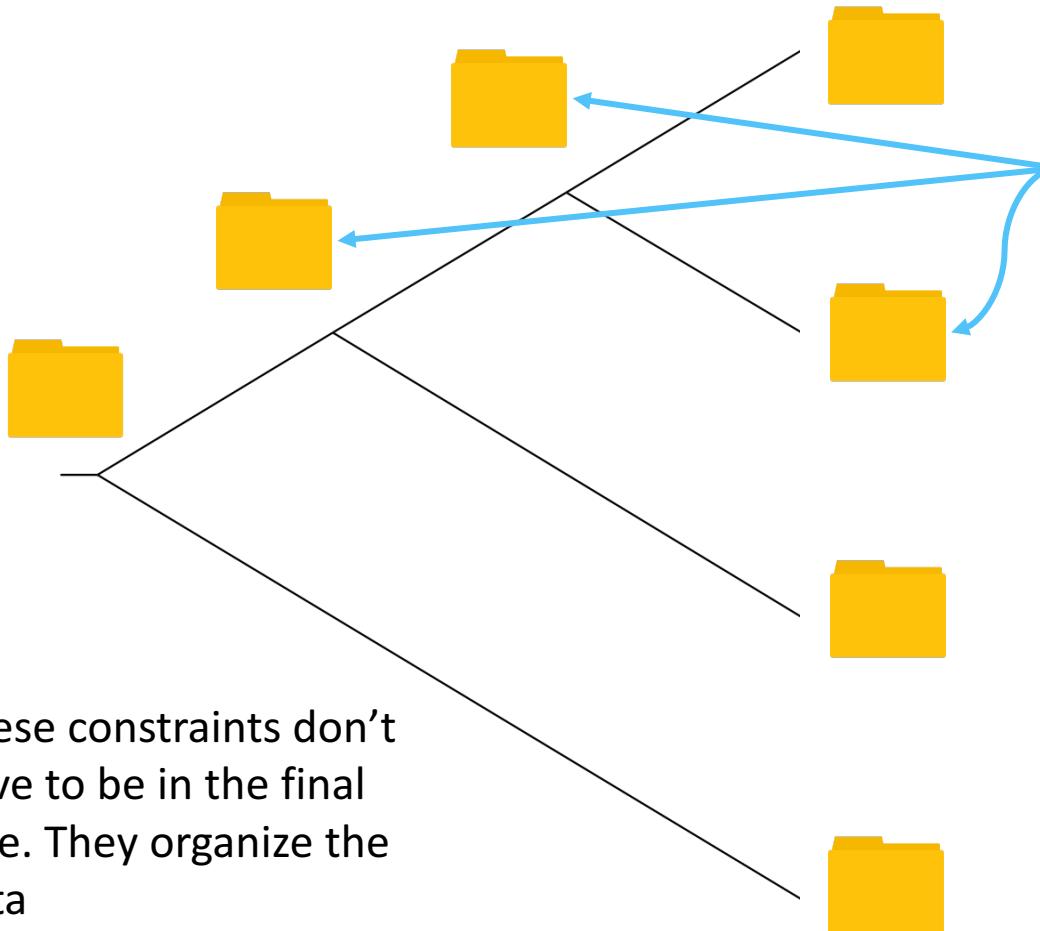
## For each clade

- Cluster and identify homologs
- Construct phylogenies (assuming monophyly)
  - Test monophyly constraints and remove if necessary
  - Re-estimate phylogeny with constraints removed
- Estimate support
- Date

## Add these new clades to a backbone

- OpenTree synth tree for comprehensive sampling
- Magallon dated tree
- Construct our own (work with Joe Walker – speaking on Monday)

# Create folders that replicate the constraint tree



These constraints don't have to be in the final tree. They organize the data

Each folder represents some putative monophyletic taxonomic group

# Cluster and identify homologs

**3)** At each node in the tree, we record the clusters in a folder

**1)** At the tips, we populate the folder with sequences and perform a clustering analysis (MCL)

*Optionally, build trees and break up long branches into new clusters (Yang and Smith 2014)*

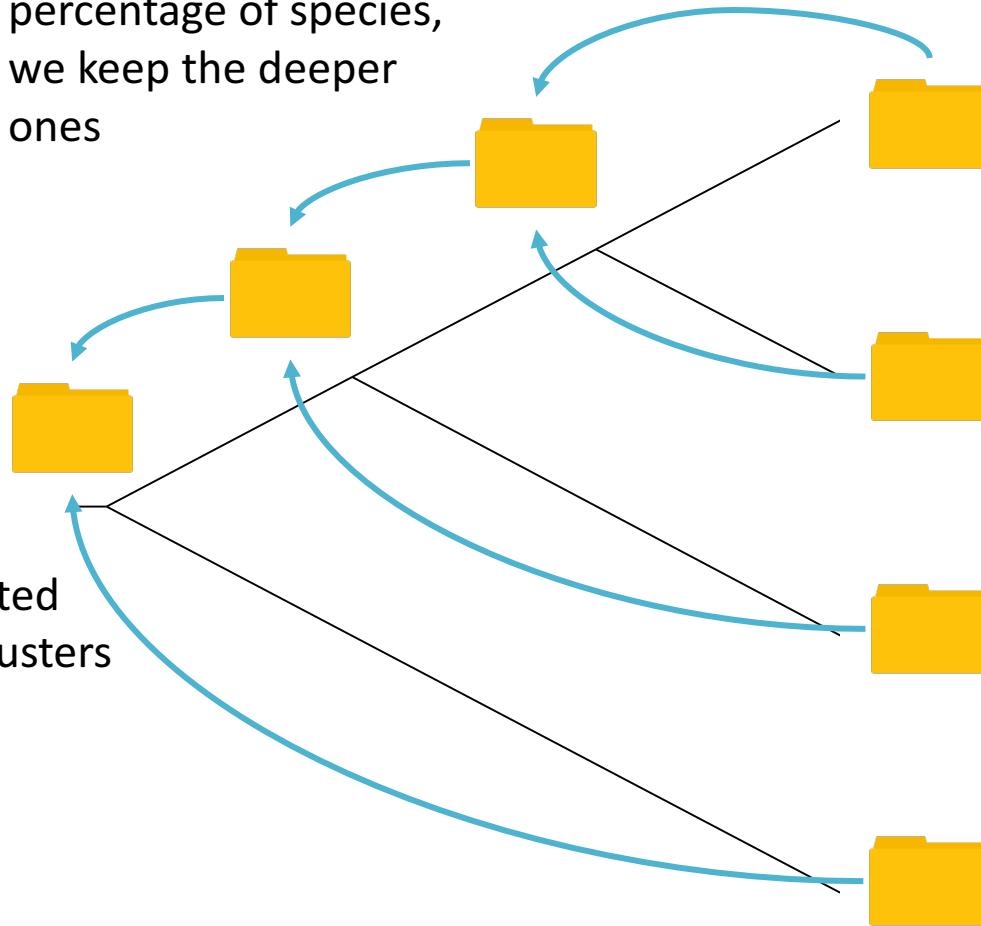
**2)** Cluster the clusters with BLAST. Align alignments with mafft (profile)

# Constructing datasets

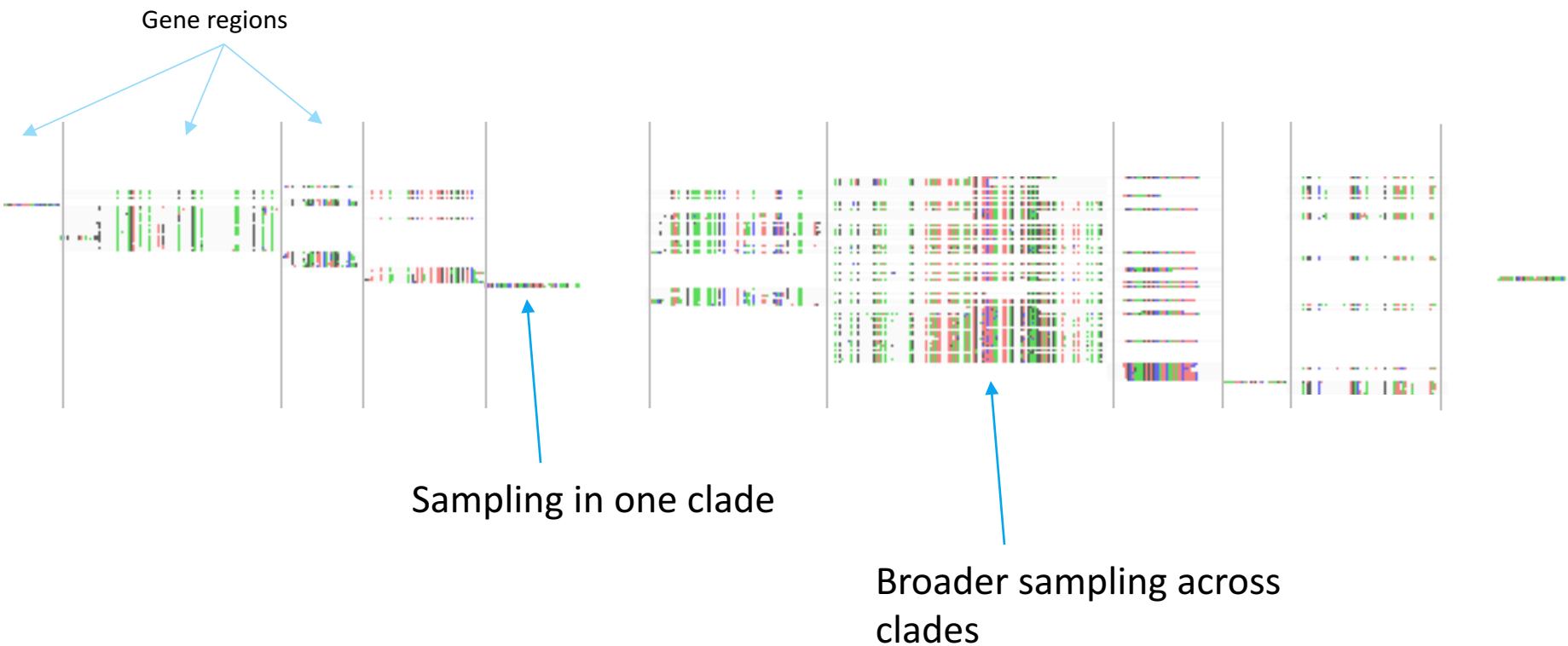
**2)** If clusters at internal nodes still have high percentage of species, we keep the deeper ones

**1)** Start at the tip folders and get the number of species and those clusters with high percentage of species are kept

**3)** Keep good nested and good deep clusters



# An alignment example

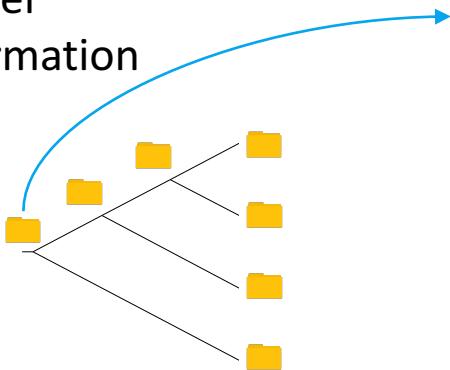


# Website for each node/folder

## Ericales\_41945

	<a href="#">name</a>	<a href="#">num_species</a>	<a href="#">avg unaln len</a>	<a href="#">defline</a>
<a href="#">Actinidiaceae_3623</a>				
<a href="#">Balsaminaceae_25692</a>				Eriastrum sparsiflorum voucher
<a href="#">Clethraceae_16611</a>	<a href="#">cluster1825.fa</a>	2297	940.468437092	KANU:354712 tRNA-Leu (trnL) gene, partial sequence; trnL-trnF intergenic spacer, complete sequence; and tRNA-Phe (trnF) gene, partial sequence; chloroplast.
<a href="#">Cyrillaceae_4339</a>				
<a href="#">Diapensiaceae_16673</a>				
<a href="#">Ebenaceae_19955</a>	<a href="#">cluster2228.fa</a>	2213	1417.07455942	Agapetes moorei chloroplast DNA, trnK intron including the matK gene, complete sequence, specimen_voucher: MBK:Kuroiwa et.al 030333.
<a href="#">Ericaceae_4345</a>				
<a href="#">Fouquieriaceae_24902</a>				
<a href="#">Lecythidaceae_3642</a>	<a href="#">cluster2168.fa</a>	1397	802.29706514	Aubregrinia taiensis 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 26S ribosomal RNA gene, partial sequence.
<a href="#">Marcgraviaceae_55360</a>				
<a href="#">Mitragastropodiaceae_91826</a>				
<a href="#">Myrsinaceae_16614</a>				
<a href="#">Pentaphylacaceae_125045</a>	<a href="#">cluster1797.fa</a>	940	1518.97340426	Schima superba NADH dehydrogenase (ndhF) gene, partial cds; chloroplast gene for chloroplast product.
<a href="#">Polemoniaceae_24584</a>				
<a href="#">Primulaceae_4335</a>				
<a href="#">Roridulaceae_91900</a>	<a href="#">cluster1850.fa</a>	938	1213.72601279	Micropholis obscura voucher P00610293 ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit (rbcl) gene, partial cds; plastid.
<a href="#">Sapotaceae_3737</a>				
<a href="#">Sarraceniaceae_4353</a>				
<a href="#">Sladeniaceae_235238</a>	<a href="#">cluster2360.fa</a>	680	936.613235294	Lysimachia remyi subsp. remyi voucher Marr 424 (NY) ribosomal protein L16 (rpl16) gene, intron; chloroplast.
<a href="#">Styracaceae_20008</a>				
<a href="#">Symplocaceae_20019</a>	<a href="#">cluster2211.fa</a>	635	780.009448819	Aubregrinia taiensis atpB-rbcL intergenic spacer, partial sequence; chloroplast.
<a href="#">Ternstroemiaceae_91898</a>				
<a href="#">Tetrameristaceae_91901</a>	<a href="#">cluster900.fa</a>	501	792.896207585	Erica alexandri subsp. alexandri isolate alexandri_EO12449 trnT-trnL intergenic spacer region, partial sequence; chloroplast.
<a href="#">Theaceae_27065</a>				

Each clade has a website with cluster information

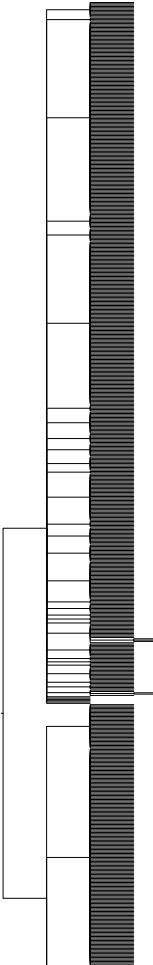


## Assume monophyly until proven otherwise

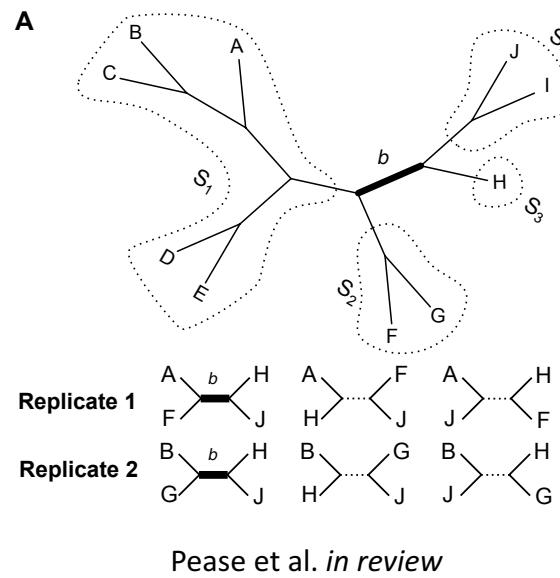
- Some edges are *impossible* to reconstruct without genomic data or otherwise difficult to assemble datasets
- Can we assume that things are monophyletic?
  - Certain clades?
  - Major Linnaean groups? Orders, Families, Genera?
- Do we have to confirm this every run?

# Constraint procedure

1) Calculate phylogeny with a constraint tree based on taxonomy



2) Test constraints using a quartet procedure (when there is strong conflict, constraints are removed)

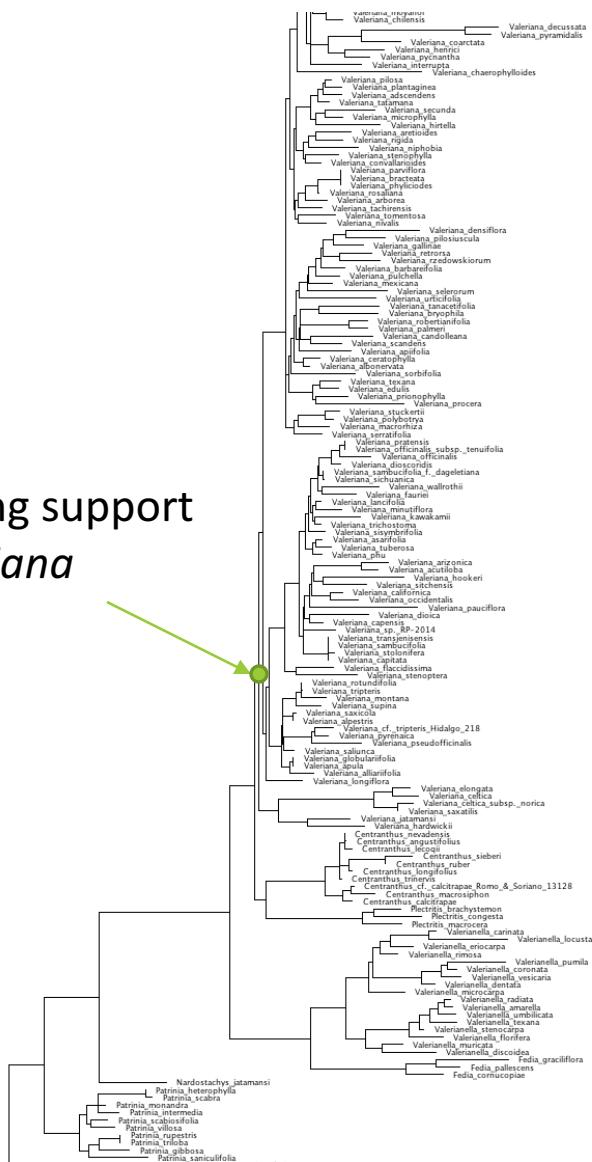


2) Recalculate the tree with the constraint tree as the ML tree with constraints removed



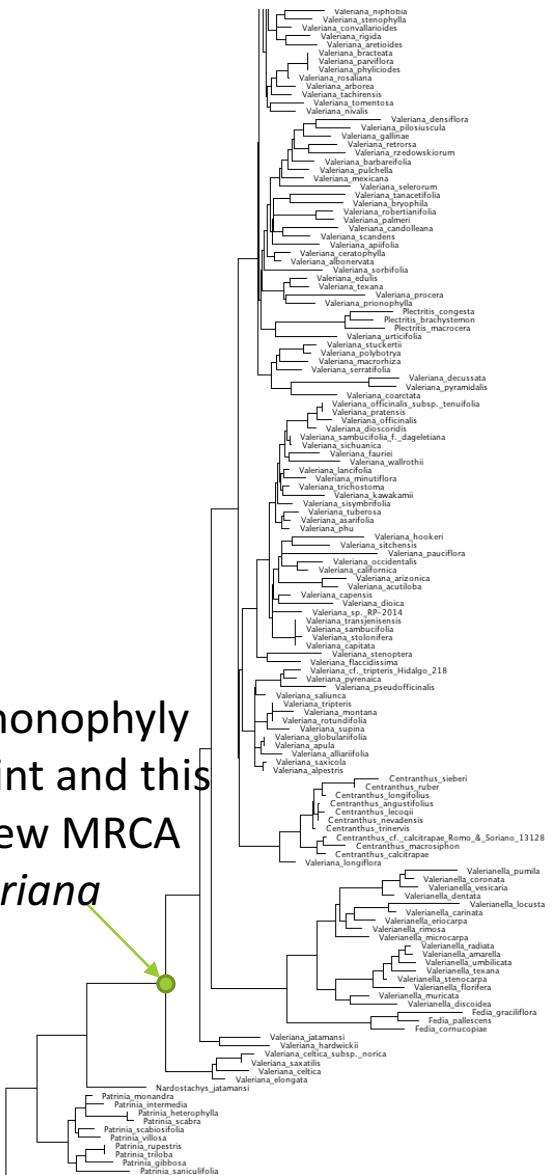
# Example with Dipsacales

Conflicting support  
for *Valeriana*

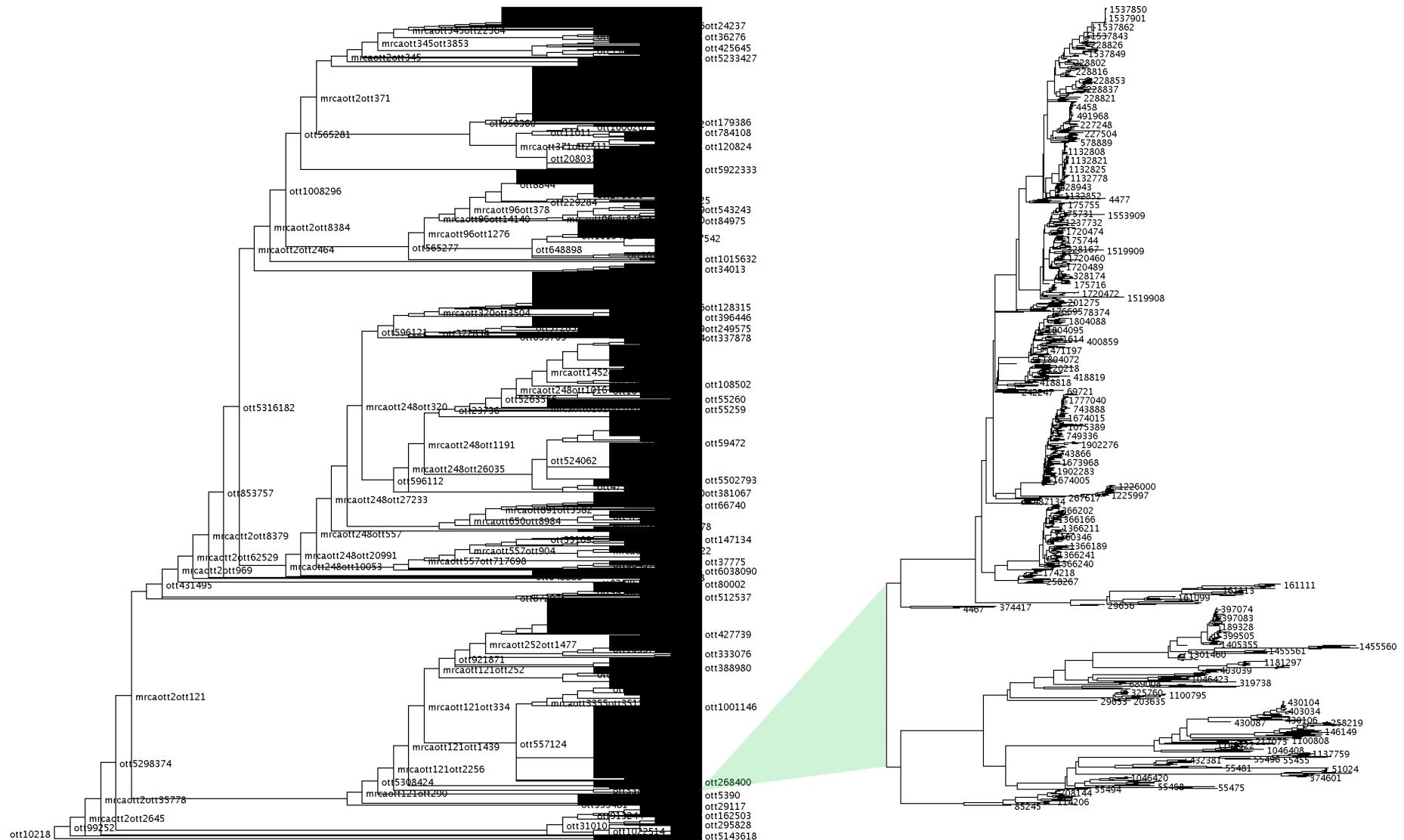


Stephen A. Smith

Break monophyly  
constraint and this  
is the new MRCA  
for *Valeriana*



# Replace clades in the backbone/synth tree



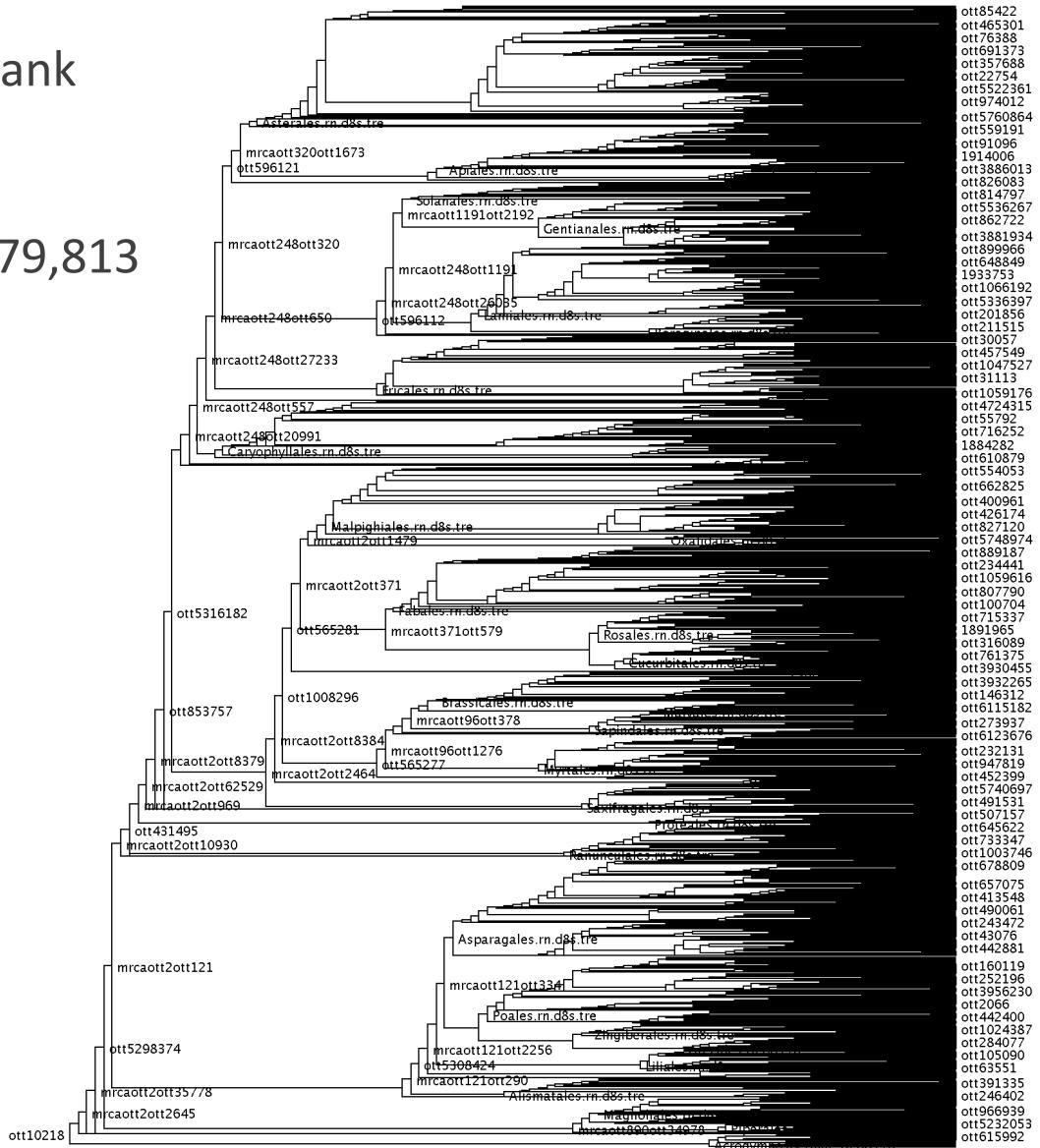
# What does the data look like?

79,759 sampled from GenBank

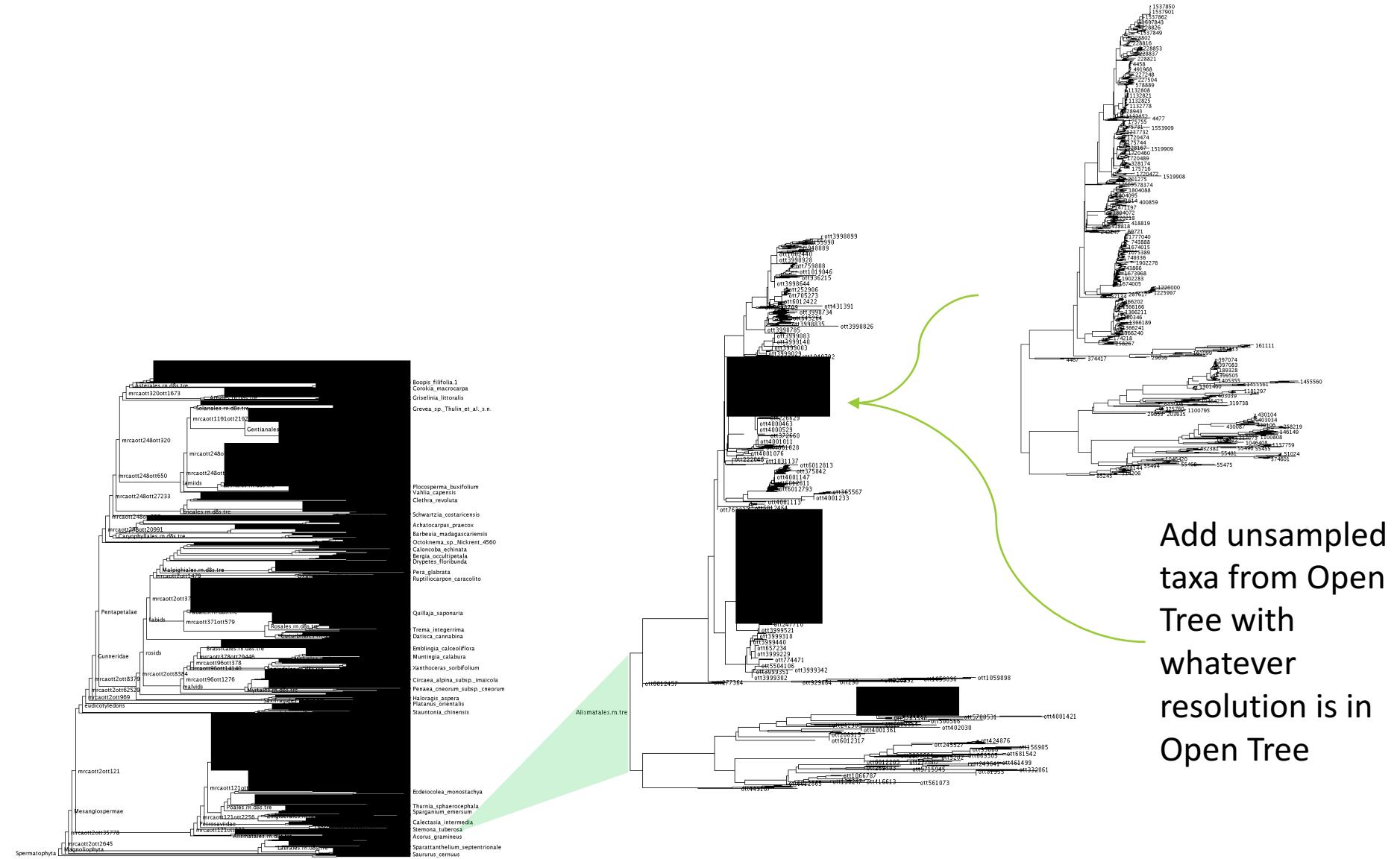
Number of tips: 79,875

Number of internal nodes: 79,813

61 unresolved nodes



# Place clades in the backbone/synth tree



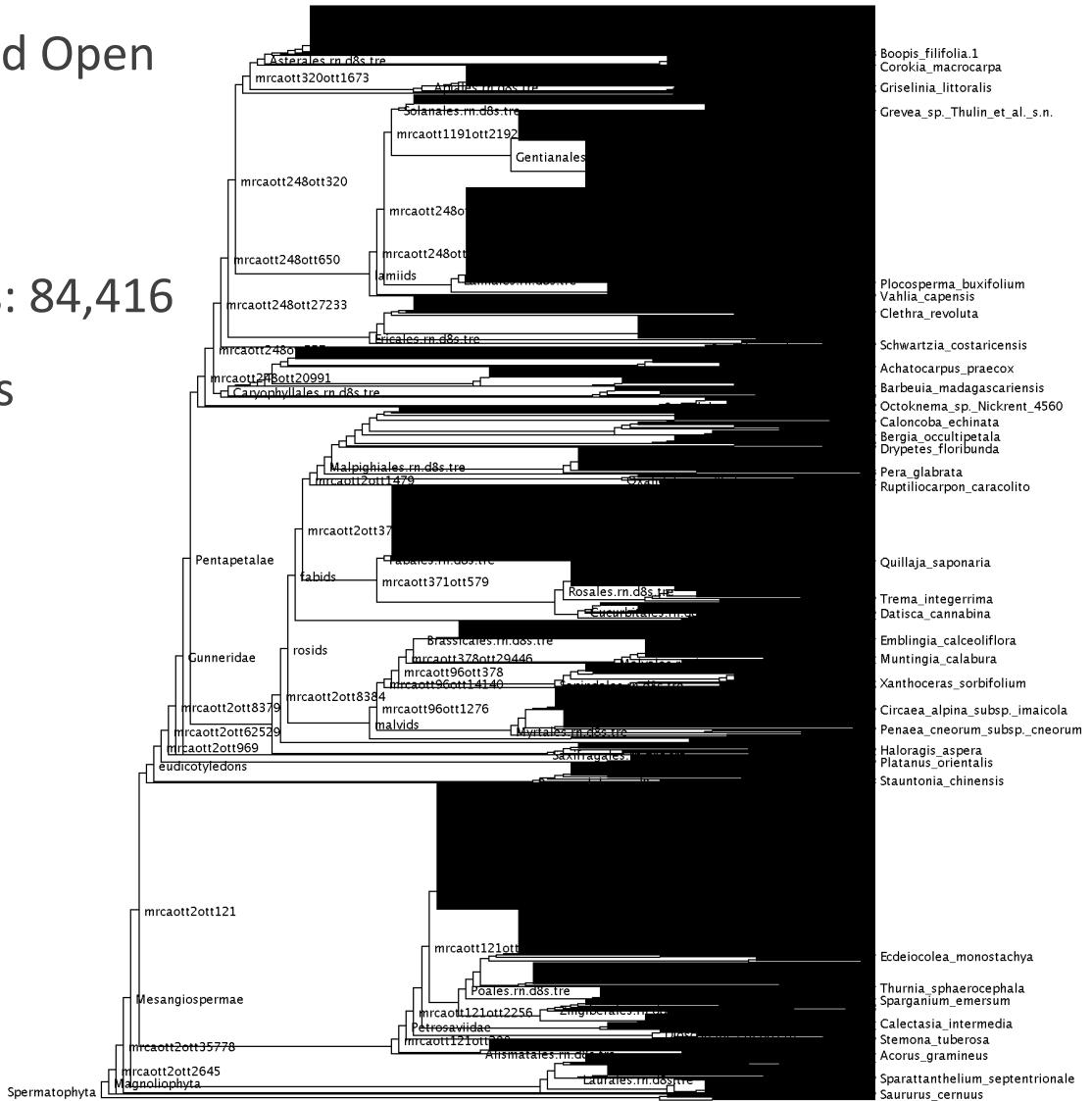
# What does the data look like?

356,807 with GenBank and Open Tree

Number of tips: 356,807

Number of internal nodes: 84,416

272,390 unresolved nodes



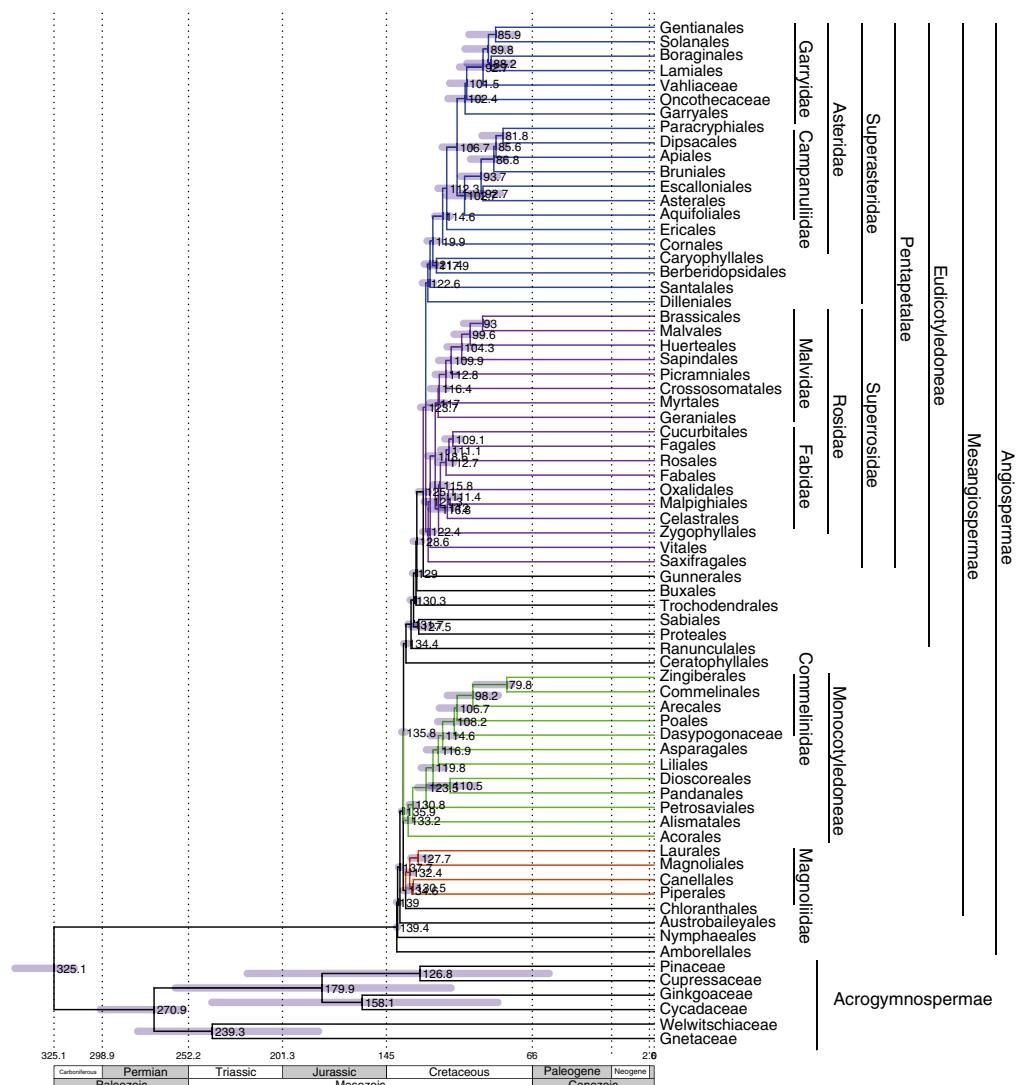
# Divergence time estimation

We use the Magallon tree as a backbone for dates and resolution or just dates

## Details:

We use hard constraints for clades that overlap with Magallon

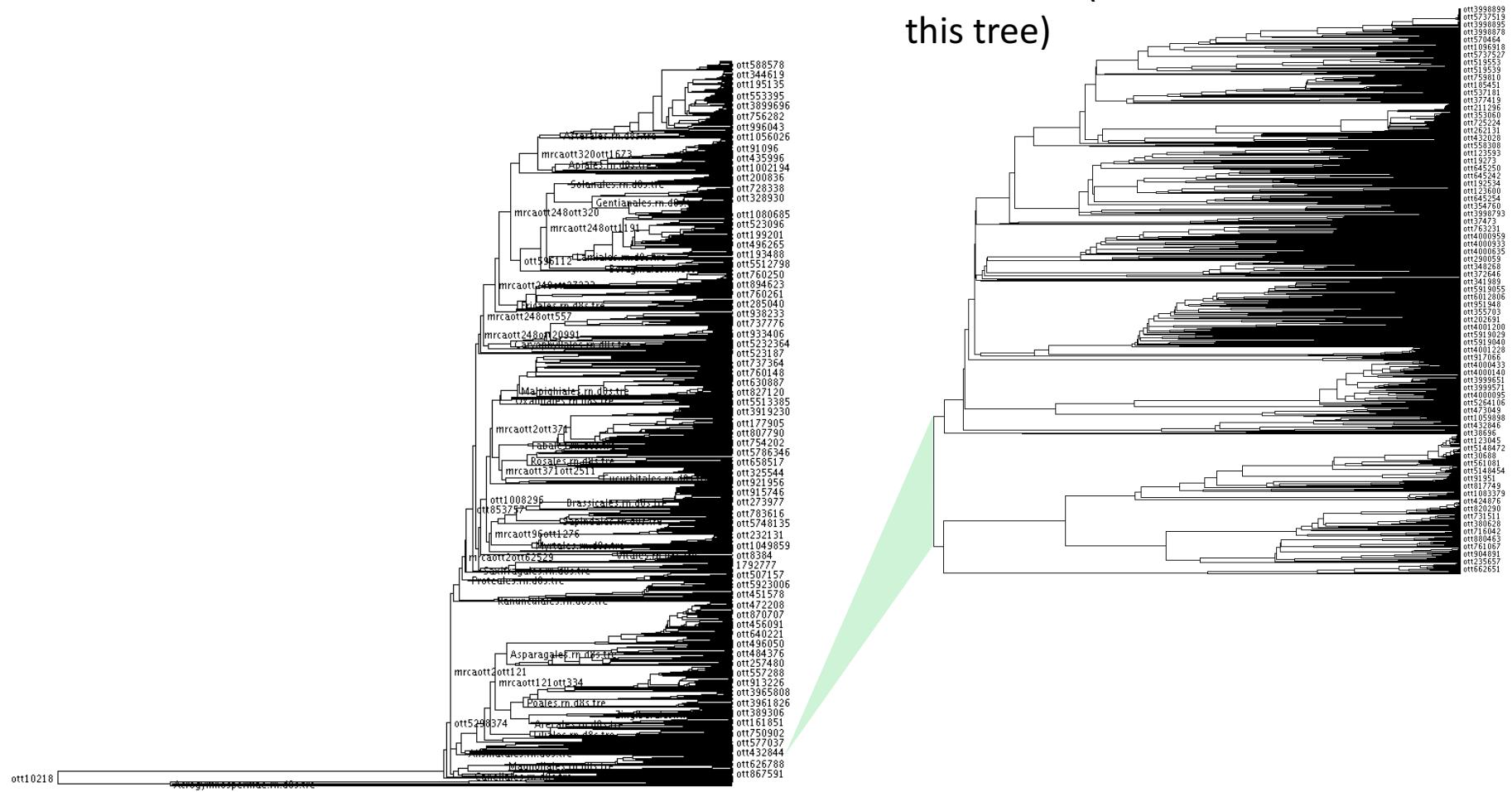
We run with treePL (Smith and O'meara 2012) on each clade



Magallon et al. 2015

# Dated Alismatales

Each clade that is found here and in Magallon et al. gets a calibration (there are 10 for this tree)

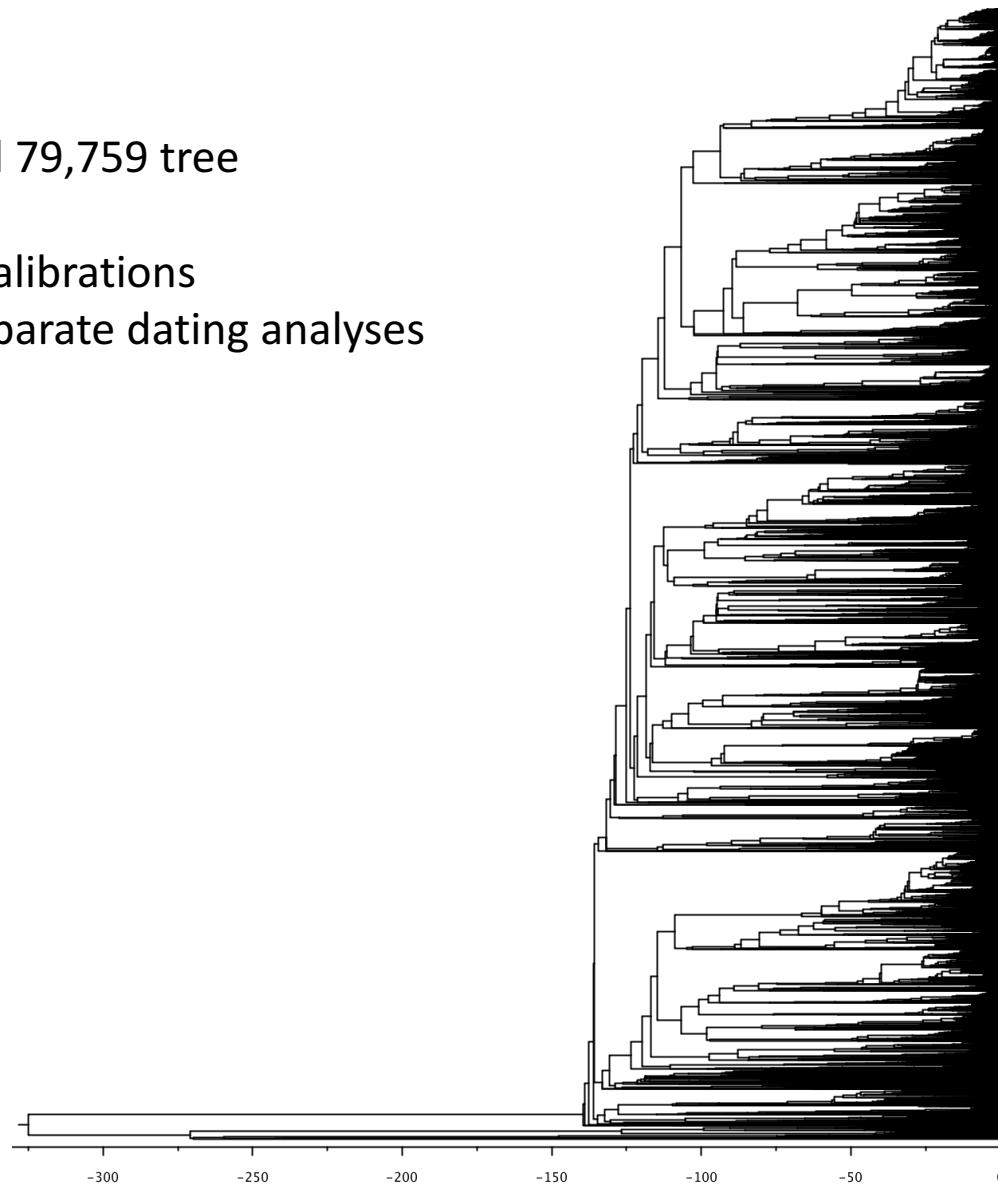


# With divergence times

Dated 79,759 tree

590 calibrations

60 separate dating analyses



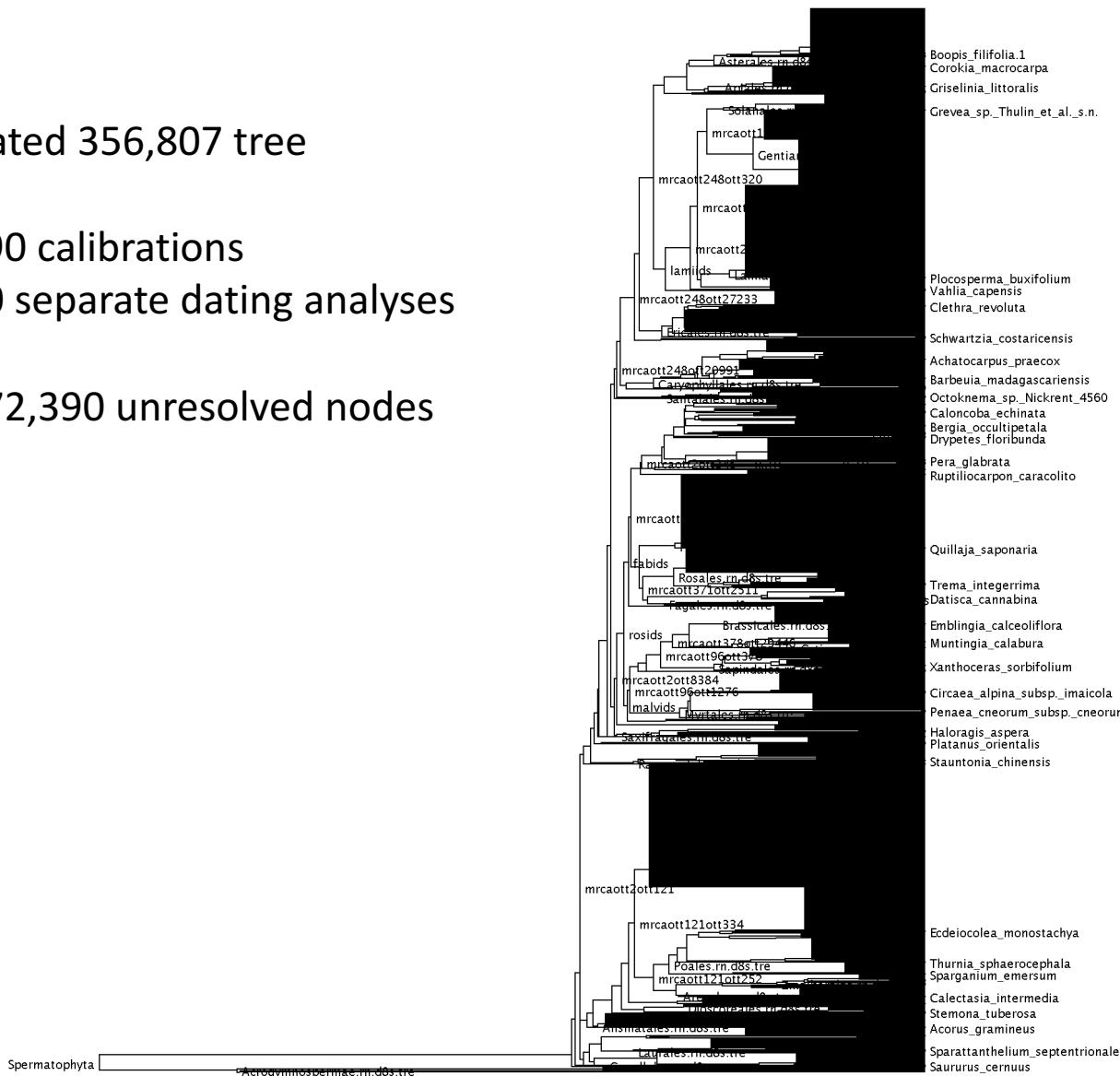
# With divergence times and Open Tree

Dated 356,807 tree

590 calibrations

60 separate dating analyses

272,390 unresolved nodes



# Resolving the polytomies

## Birth-Death placement of unresolved taxa from Open Tree

- We use MrBayes and our own scripts to construct the input file with constraints (topo and dates)
- Described in Jetz et al. ; Kuhn et al. 2011

### Pros

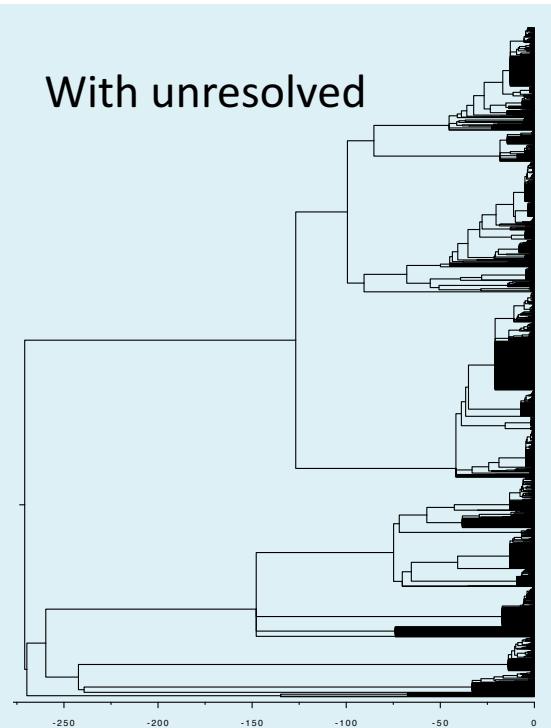
- Reasonable placement of taxa
- Reasonable divergence time estimates

### Cons

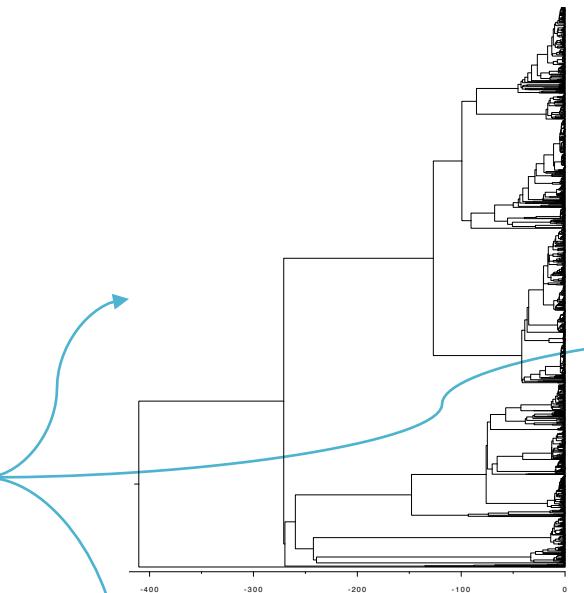
- Single rate estimates for each tree of birth / death parameters
  - Although we get an estimate for each large clade (so actually >60 different rates)
- Convergence can take some time
- Probably shouldn't be doing a lot of diversification analyses on these

# Example with *Acrogymnospermae*

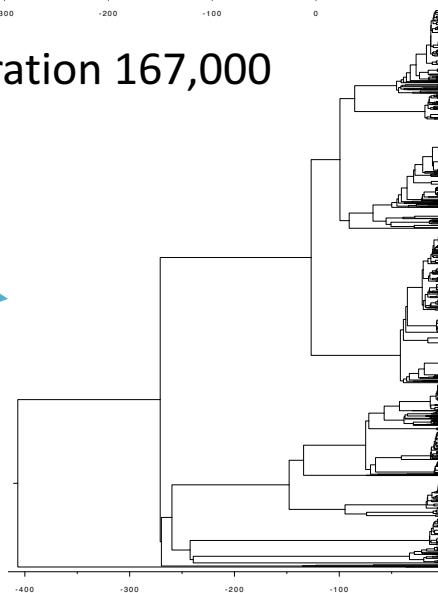
With unresolved



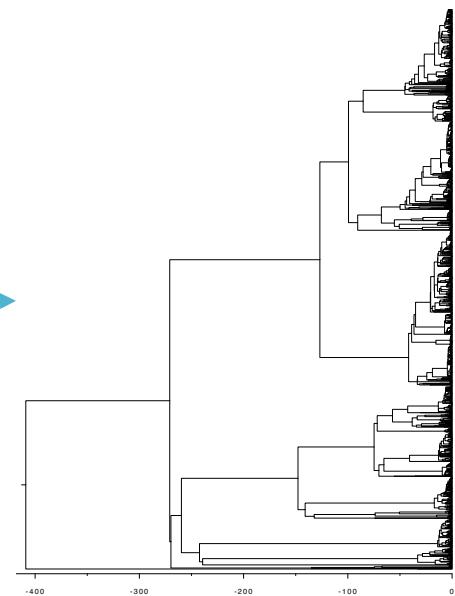
Generation 167,000



Generation 935,000

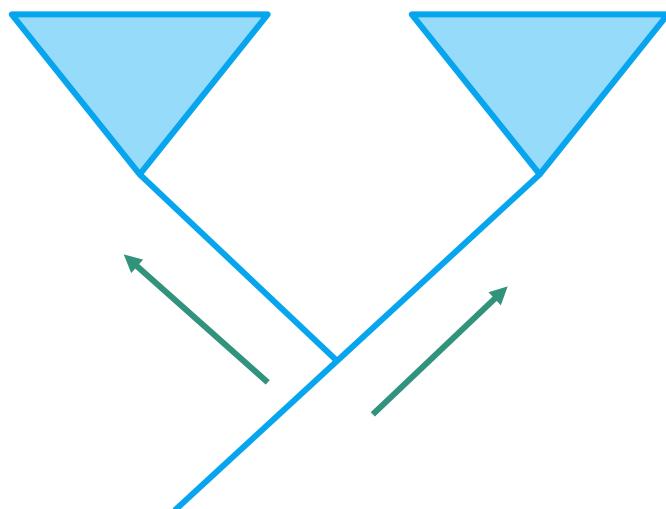


Generation 435,000



# How about data overlap

We measured each edge for data overlap between subtending left and right edges



How many sites overlap between any taxa on the left and any taxa on the right?

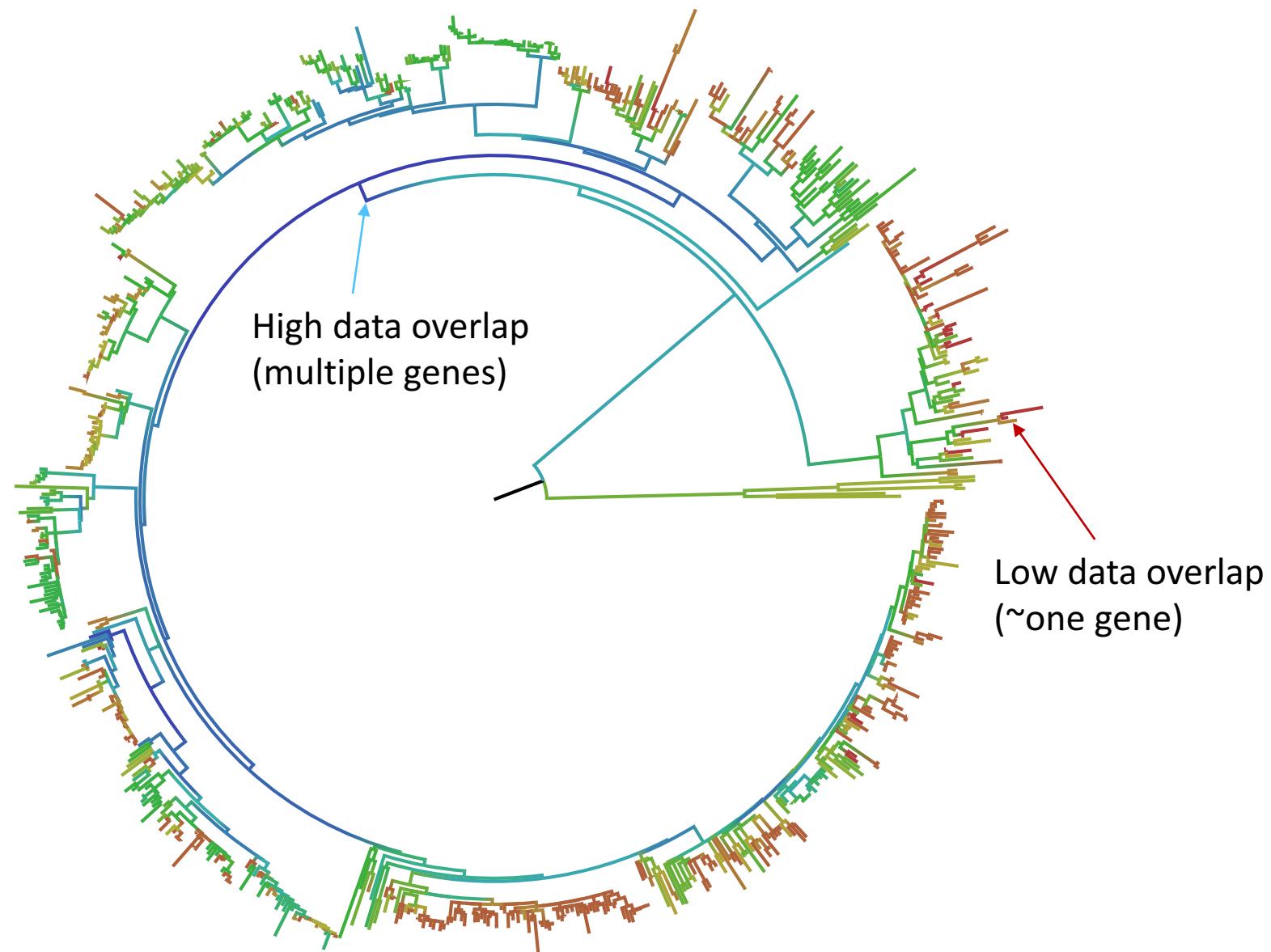
## Example

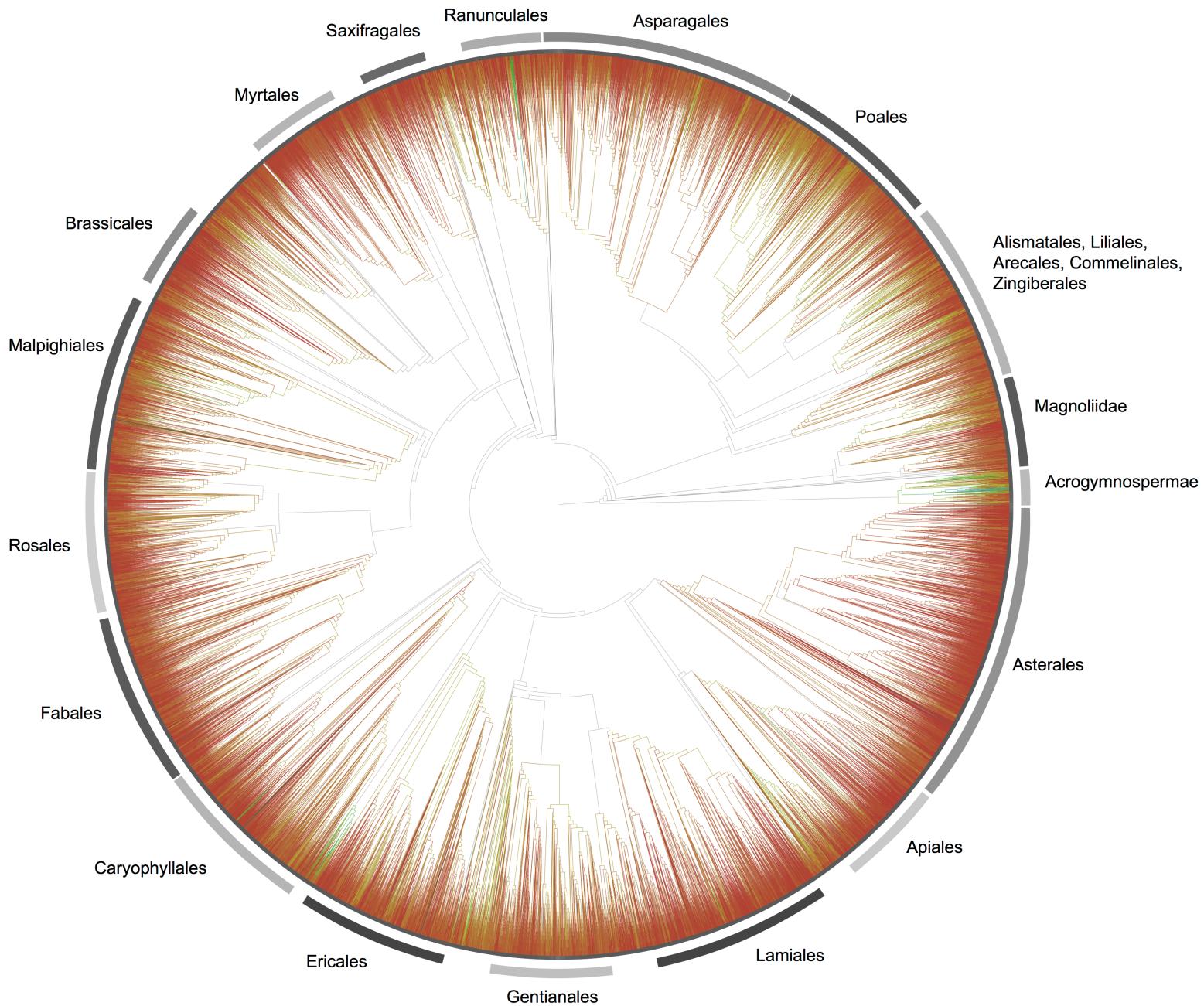
1

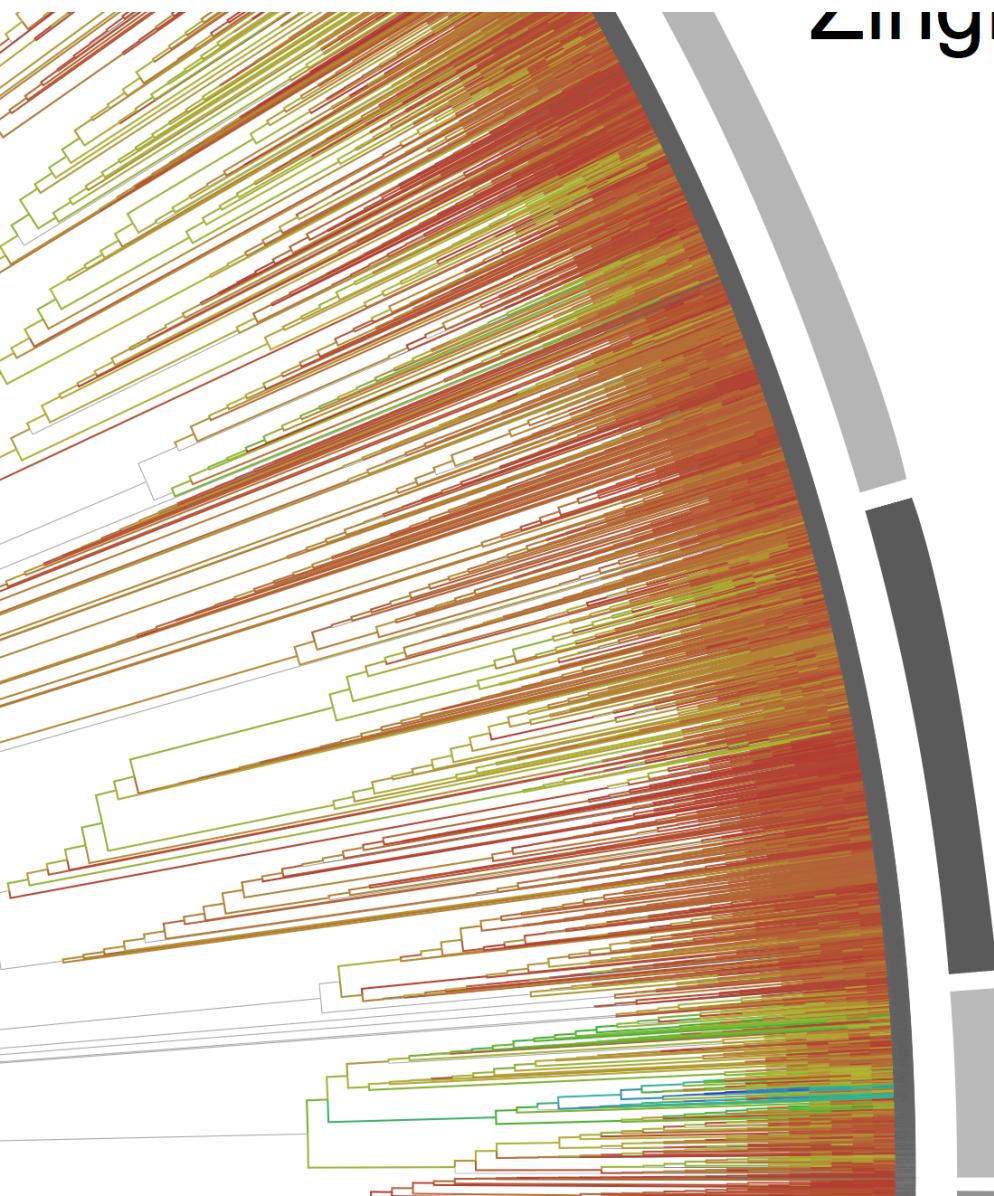
2

1 ACCCGTTT----GTGG  
2 AG-CGTTT----G---  
++ +++++ + = 8

# Arecales data overlap







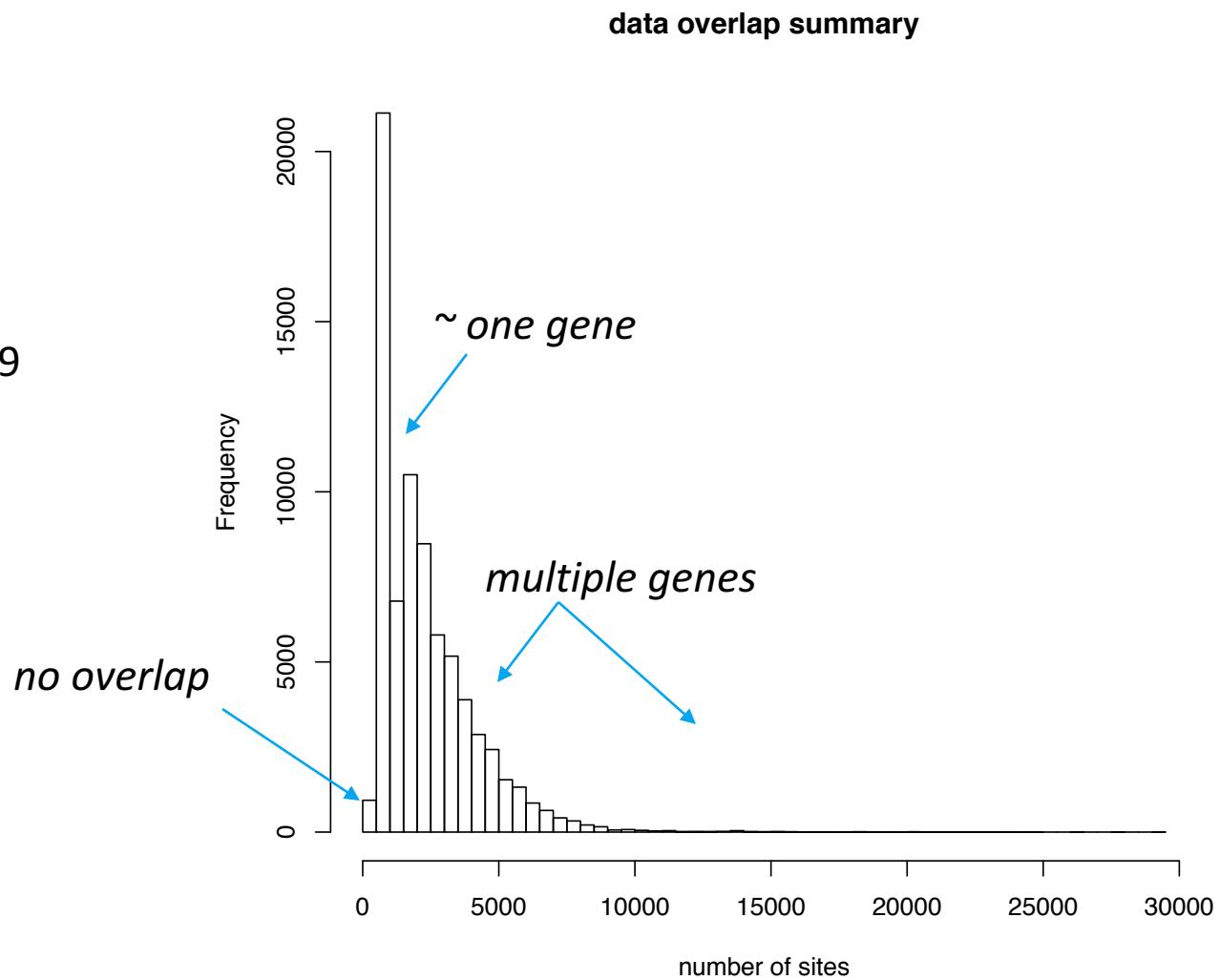
Liliidae

Magnoliidae

Acrogymnospermae

# Summary

*Minimum:* 0  
*Median:* 1792  
*Mean:* 2340  
*Maximum:* 29229



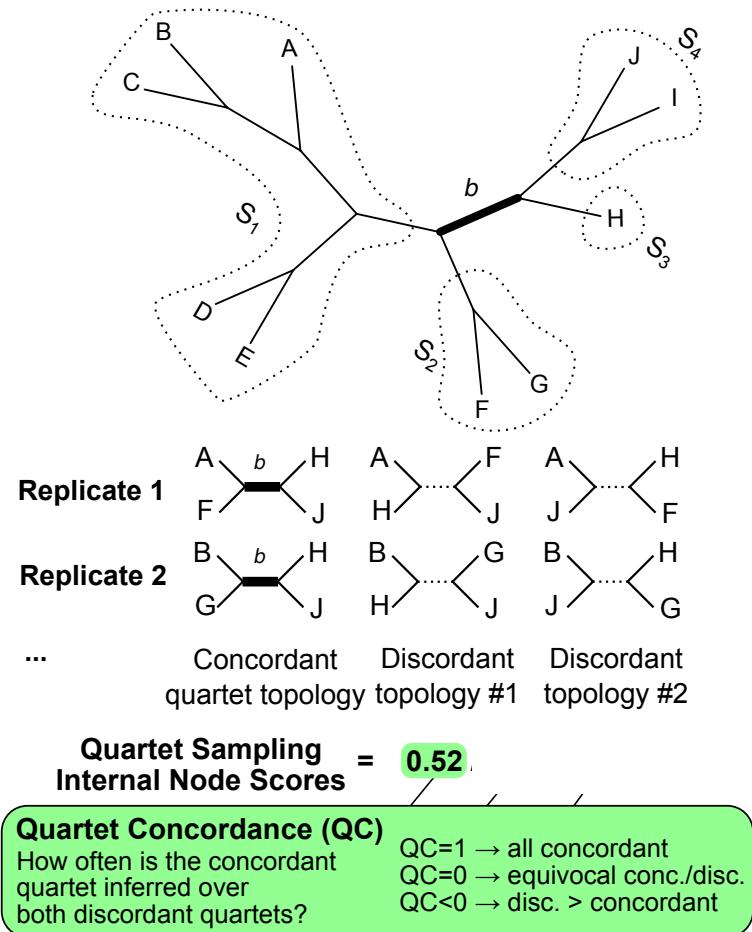
# How about support?

We would like to ascertain support

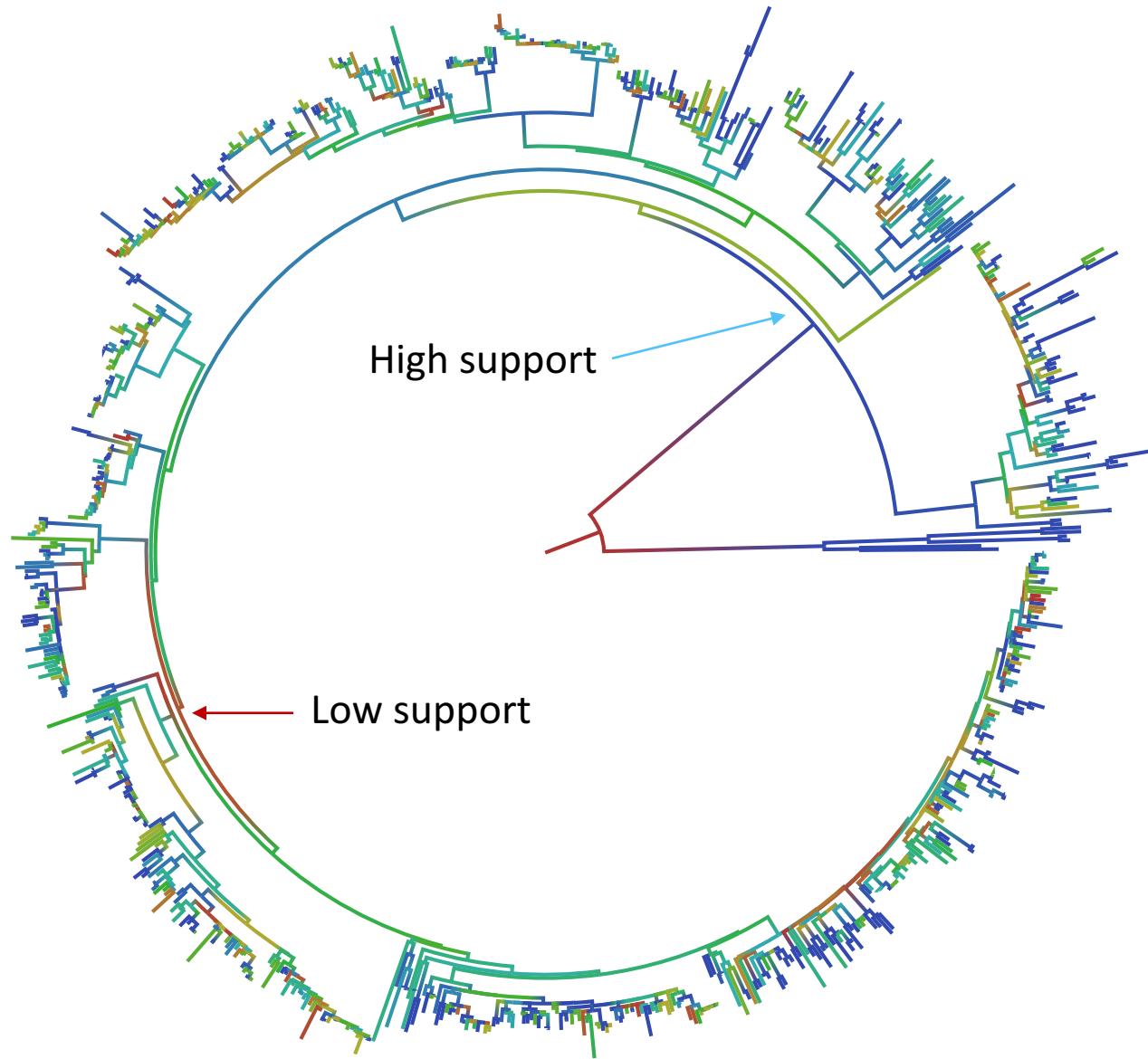
- Bootstrap is too slow and maybe not what we want
- Bayesian analysis is not going to happen
- aBayes and SH-Like are fine but aBayes is limited and SH-Like requires that you have the ML tree (we don't because of constraints)

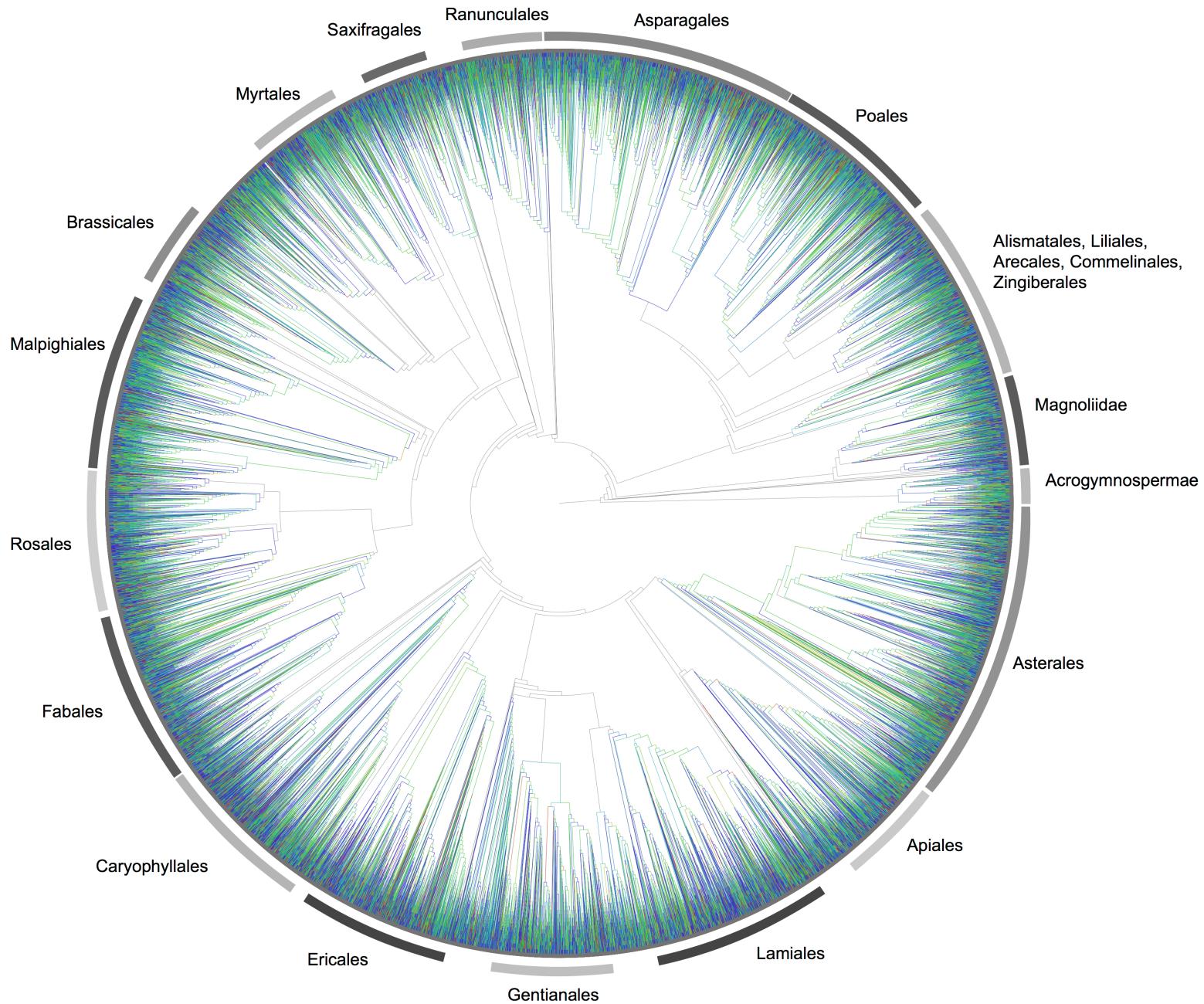
To do this, we are using Quartet Concordance (QC)

Pease et al. *in review*



# Arecales support

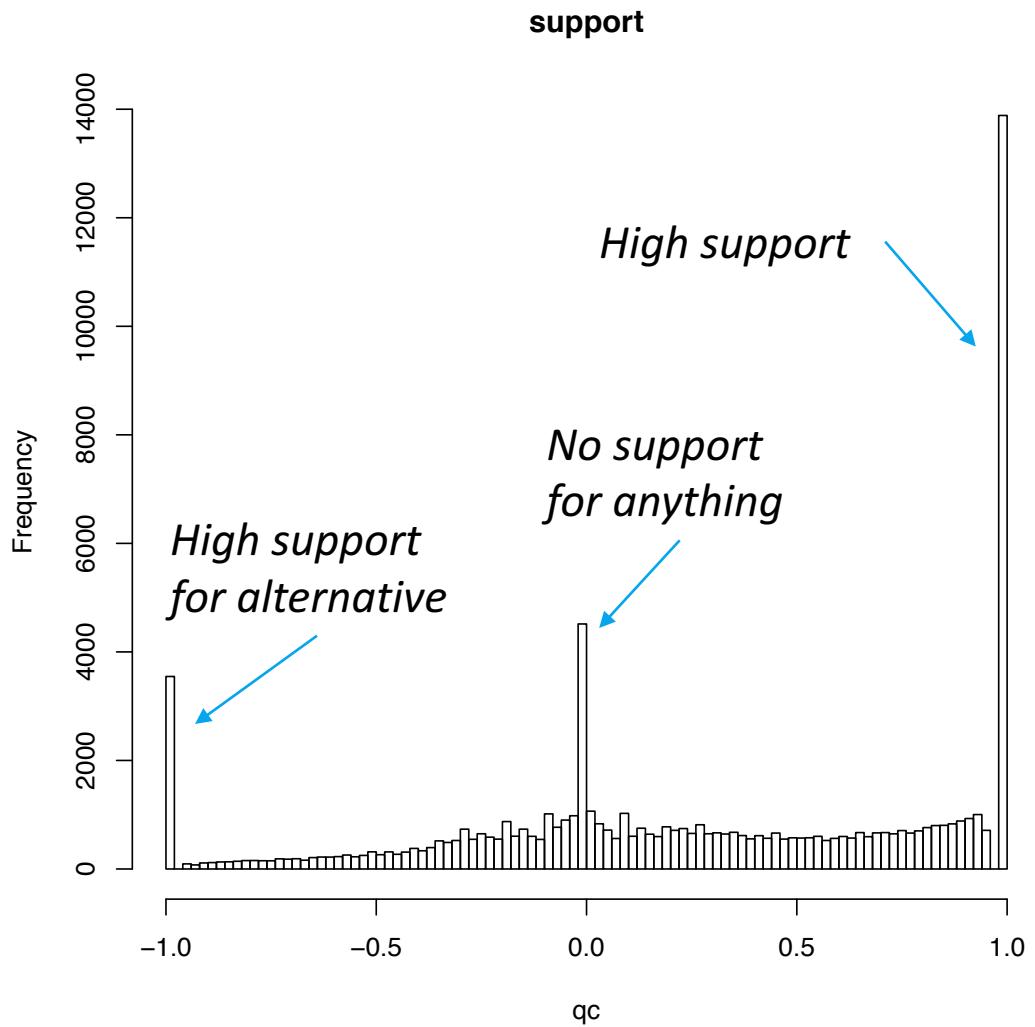




# Asterales

# QC support distribution

*Minimum:* -1  
*Median:* 0.29  
*Mean:* 0.285  
*Maximum:* 1



---

# Why not more taxa?

# The challenge of barcoding for phylogenetics

## Misidentification

- Submitted sequences may be misidentified
- It is difficult to get NCBI to correct these

## No phylogenetic information

- Some of the sequences can be very short and not carry much if any phylogenetic information

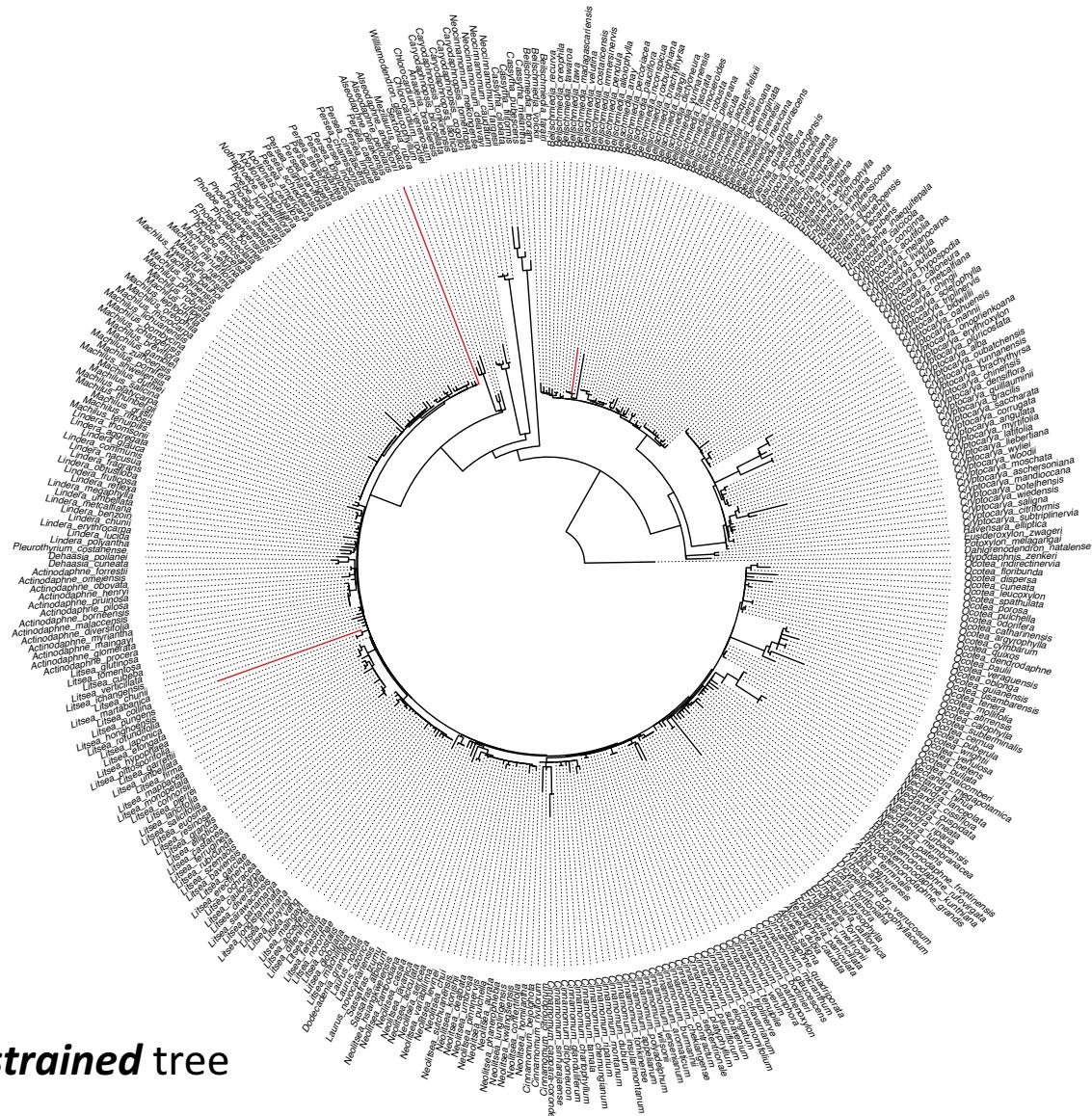
# The challenge of barcoding for phylogenetics

## Outlying species

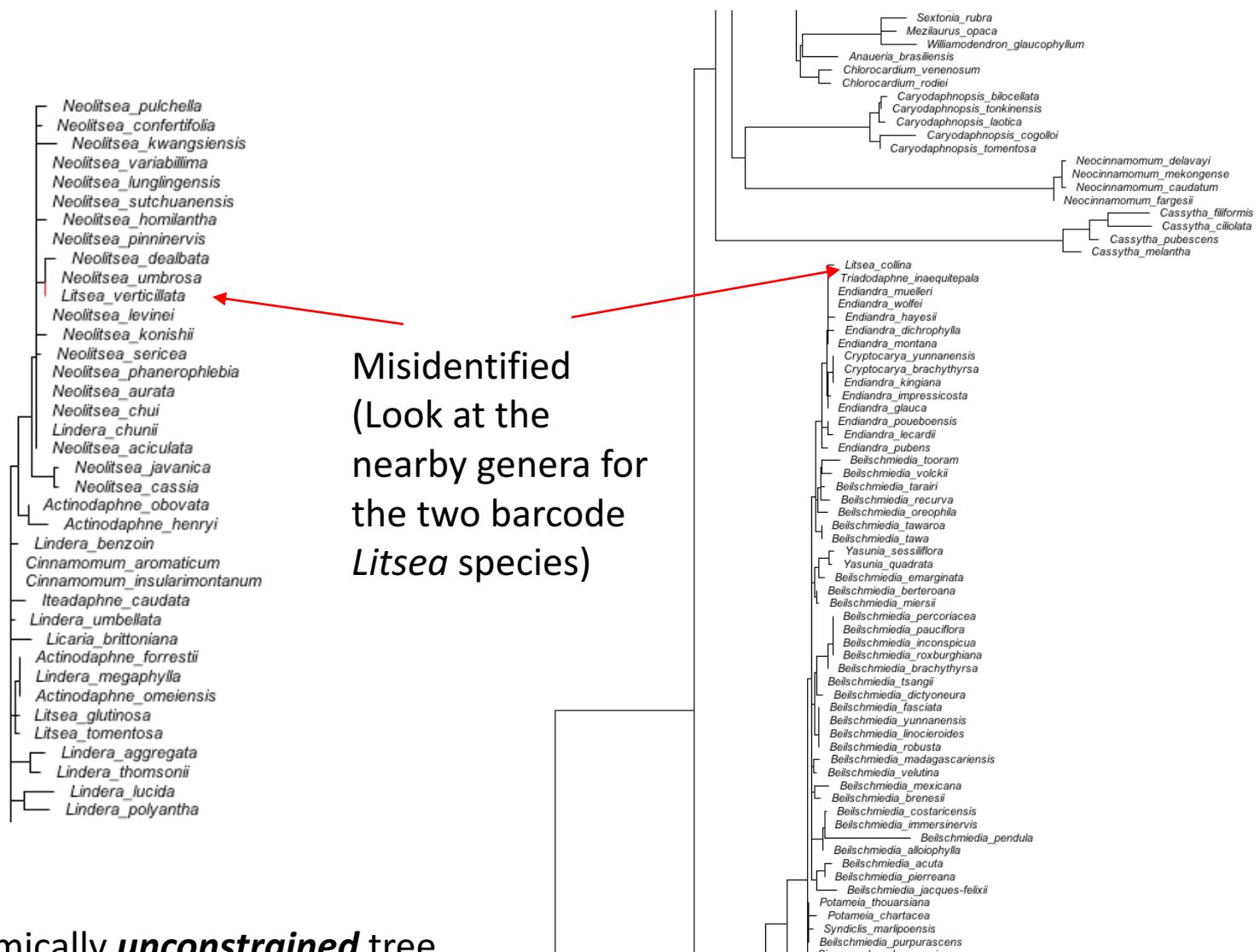
- *Litsea collina*
- *Litsea verticillata*
- *Alsoedaphne andersonii*
- *Beilschmiedia pendula*

Clear pattern from constrained tree

Taxonomically **constrained** tree



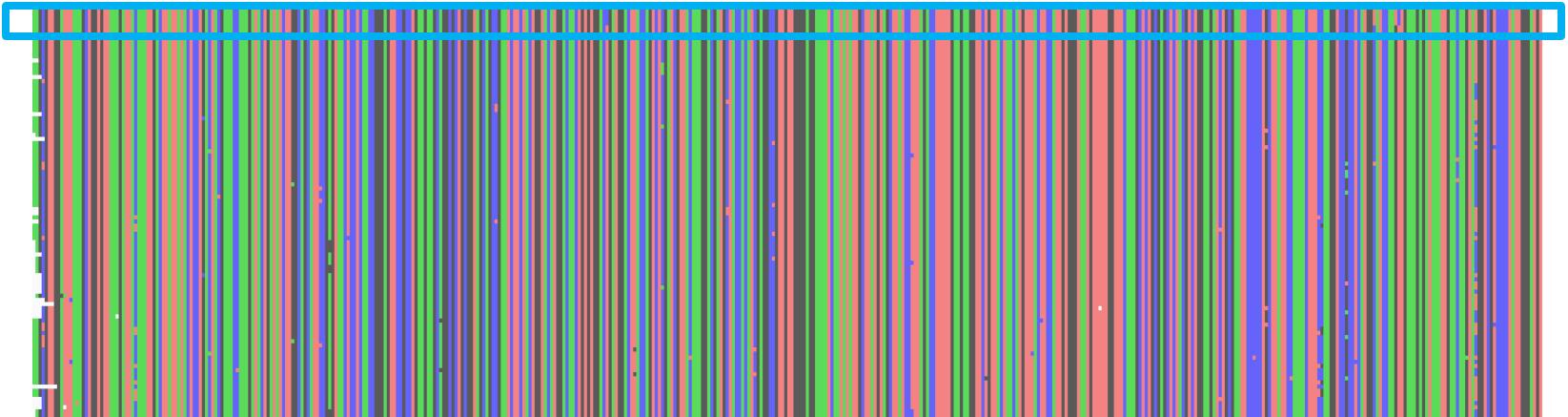
# The challenge of barcoding for phylogenetics



Taxonomically *unconstrained* tree

# Little to no phylogenetic information

*Two identical barcodes*



Identical sequences in maximum likelihood have distance (branch length) = 0

We have to either

- integrate over the uncertainty
- add these back later
- constrain

# The challenge of barcoding for phylogenetics

## Misidentification

- There is little that can be done for this on NCBI
  - NCBI requires the original submitter correct the taxonomy

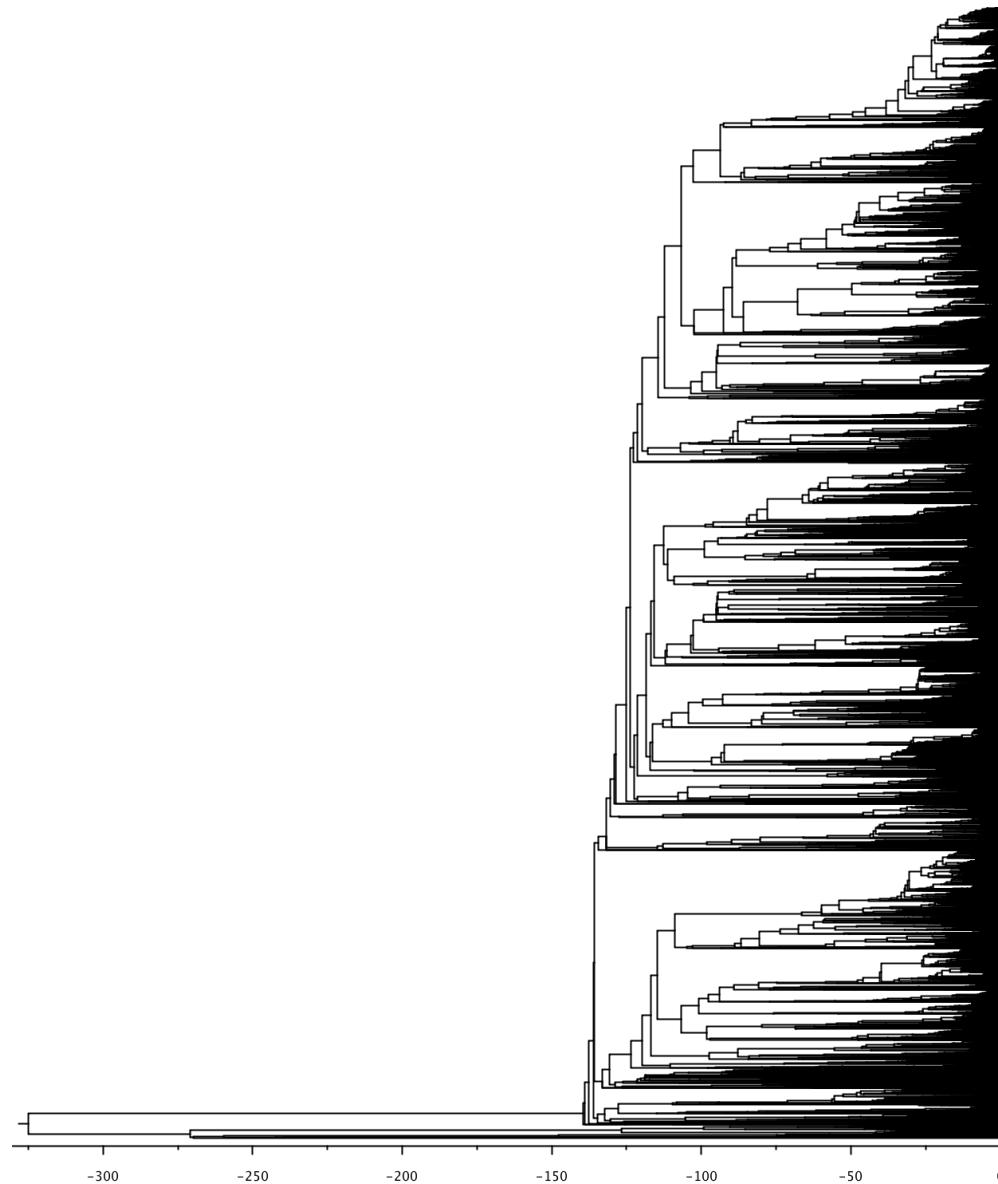
## No phylogenetic information

- Maximum likelihood may be positively misleading if uncertainty is not incorporated
- Perhaps this data is not going to be very useful for phylogenetics

## Ways forward

- Automate the discovery of these and filter them (with PyPHLAWD but communicated on GitHub)
- Can we collect this data so that it is more useful?

# Don't despair

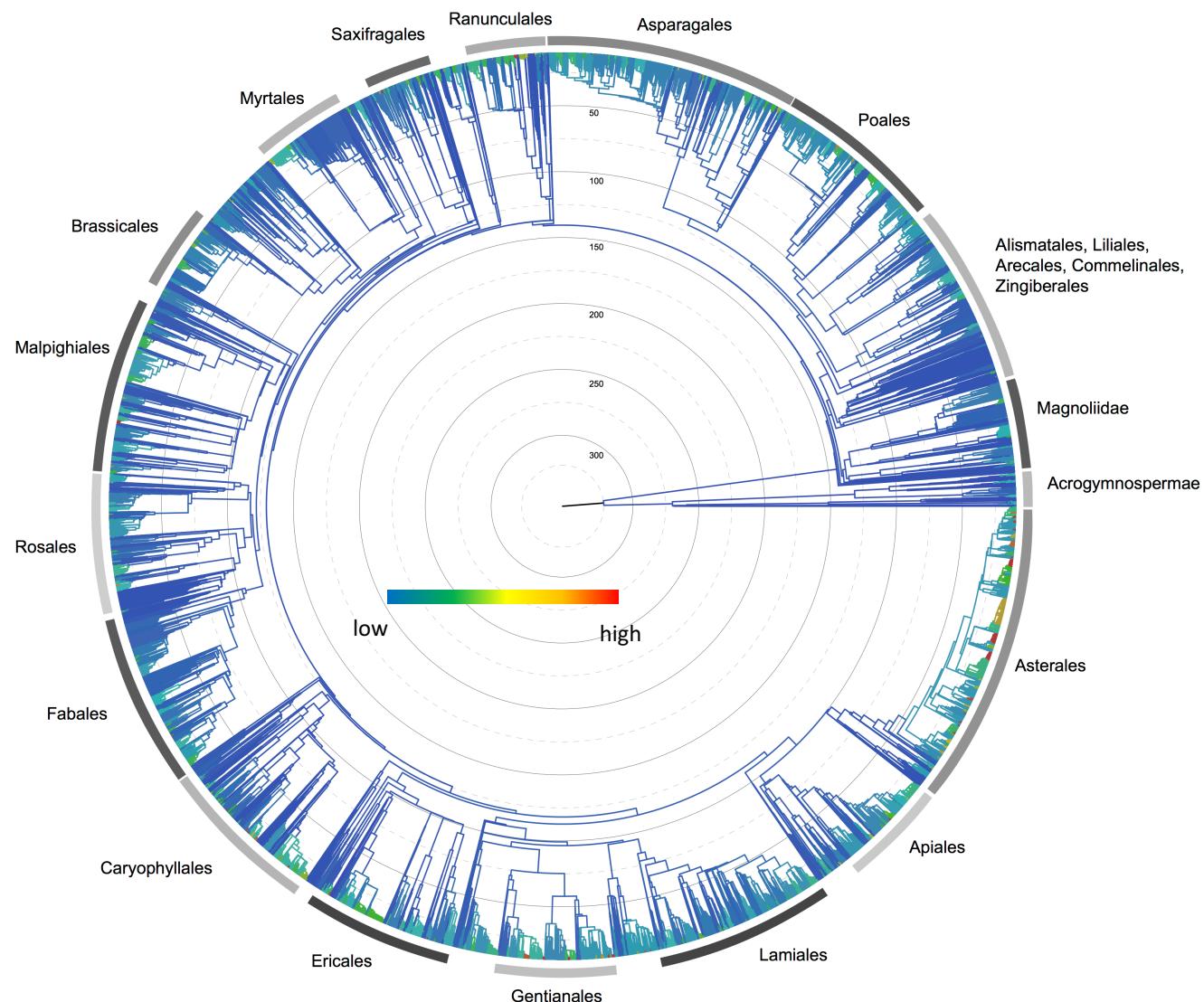


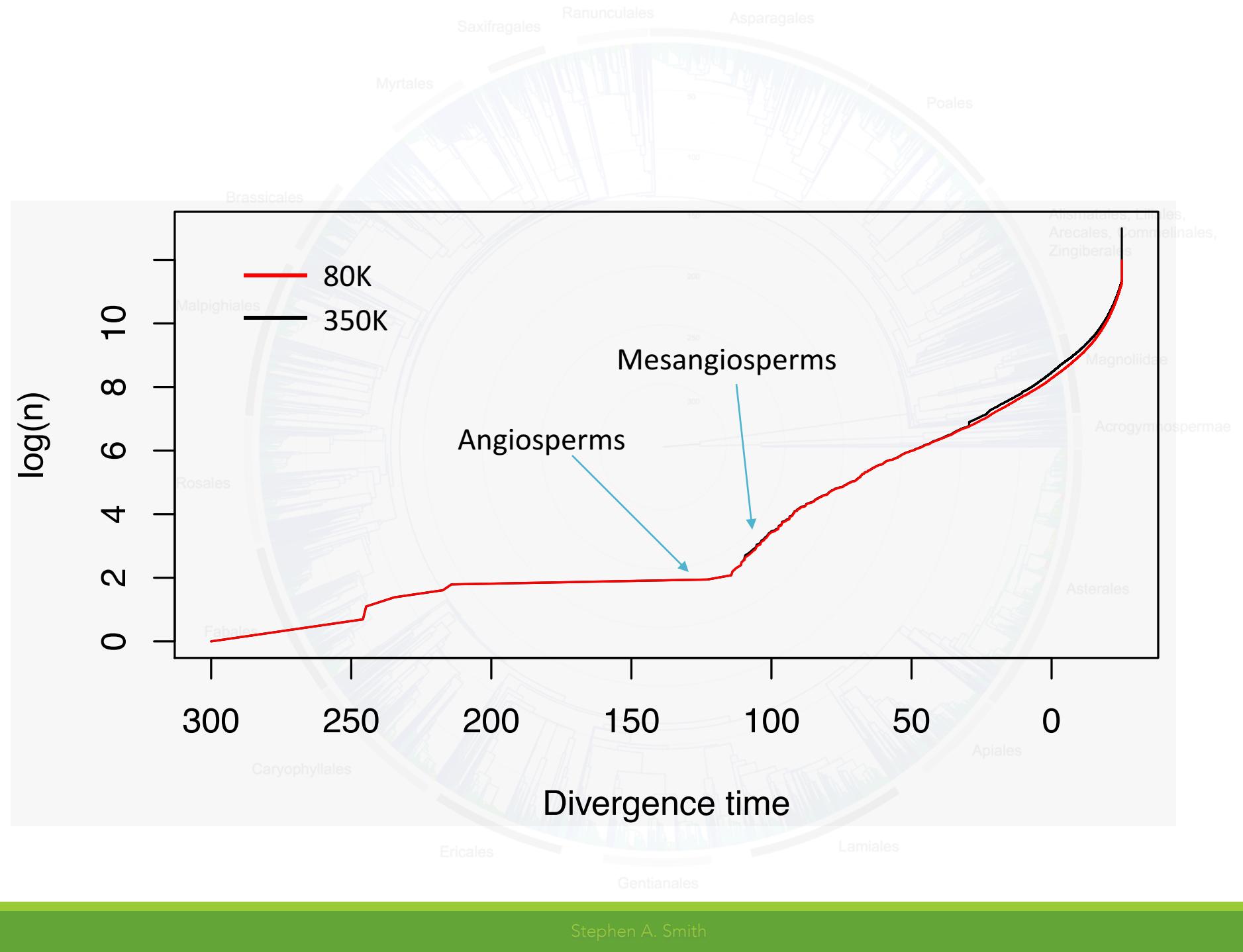
# Rates of diversification

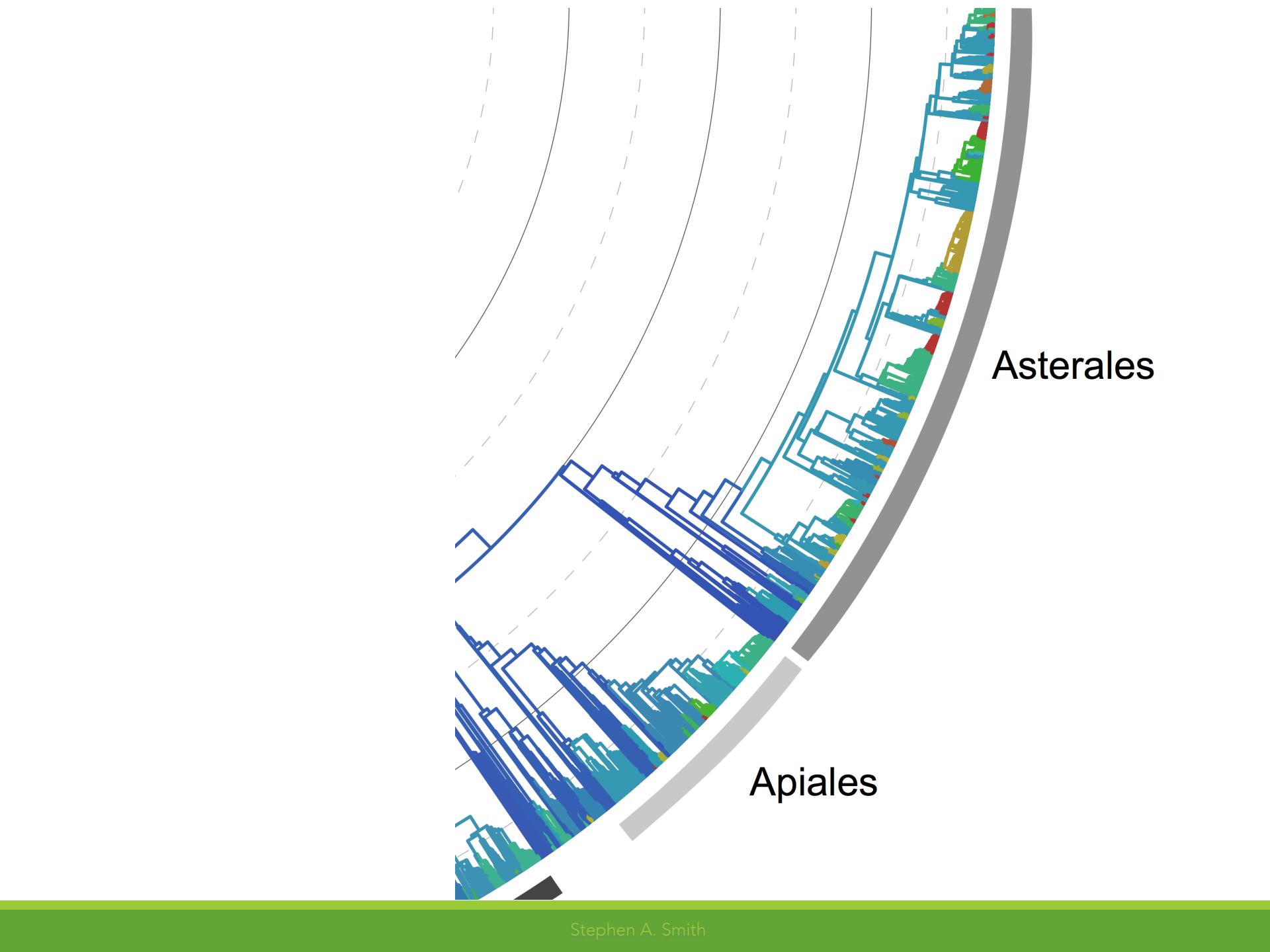
MEDUSA analyses of rate of evolution using birth-death models

Mean rate of evolution plotted

(highest rate [2.9] capped [1.5] for easier viewing)





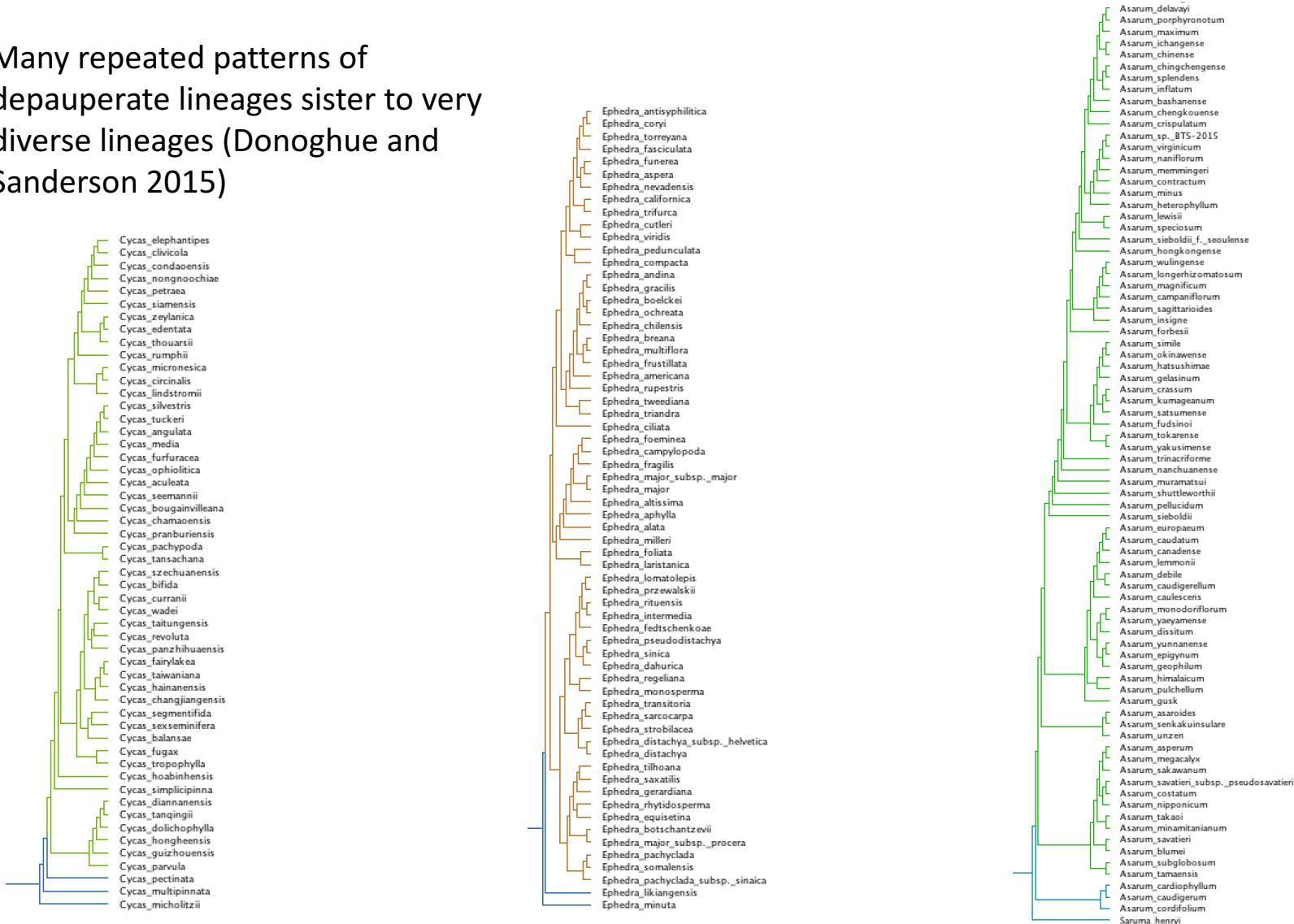


Asterales

Apiales

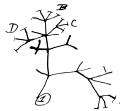
# Many repeated patterns

Many repeated patterns of depauperate lineages sister to very diverse lineages (Donoghue and Sanderson 2015)

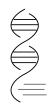


# General conclusions

## Methodological



With some methodological advances, ***we can construct comprehensive trees*** with branch lengths



We may need to combine large tree, single genes, and, genomic studies to do simultaneous analyses

## Empirical



We can construct ***a reasonable tree for seed plants*** that can be used for many analyses. We can also add back taxa not sampled in GenBank for comprehensive analyses.



Diversification analyses yield results suggesting ***many nested shifts throughout the seed plant tree***

# Acknowledgements

- Funding sources

- National Science Foundation
- University of Michigan

- Collaborators

- Smith lab

- Grad students
    - Joseph Walker
    - Drew Larson
    - Lijun Zhao
  - Postdocs
    - Greg Stull
    - Joseph Brown
    - Ning Wang
    - Oscar Vargas
  - Undergrads
    - Sonia Ahluwalia
    - Jordan Shore



- Former postdocs

- Ya Yang
  - James Pease
  - Cody Hinchliff

- Michael Moore

- Sam Brockington

- Douglas Soltis

- Pam Soltis

- All the Open Tree of Life folks

