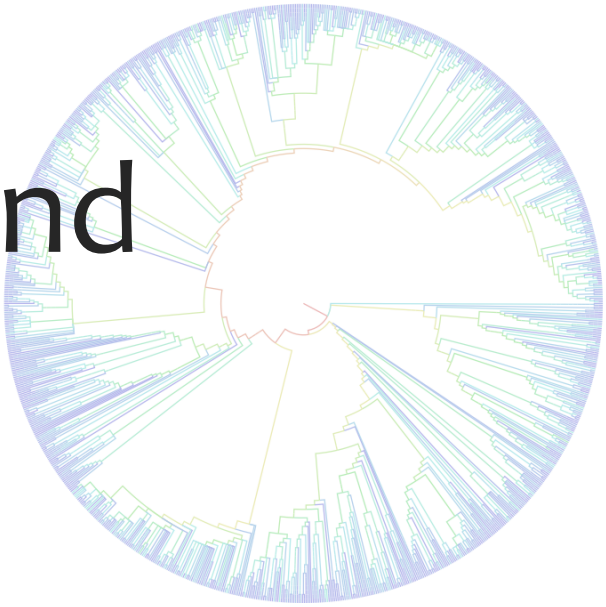
A **comprehensive** and **dated** seed plant phylogeny



STEPHEN A. SMITH

AND JOSEPH BROWN

UNIVERSITY OF MICHIGAN

I use phylogenies to address:

Empirical


Biodiversity: The origin and evolution of biodiversity

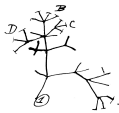

Biogeography: Why do species live where they live


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

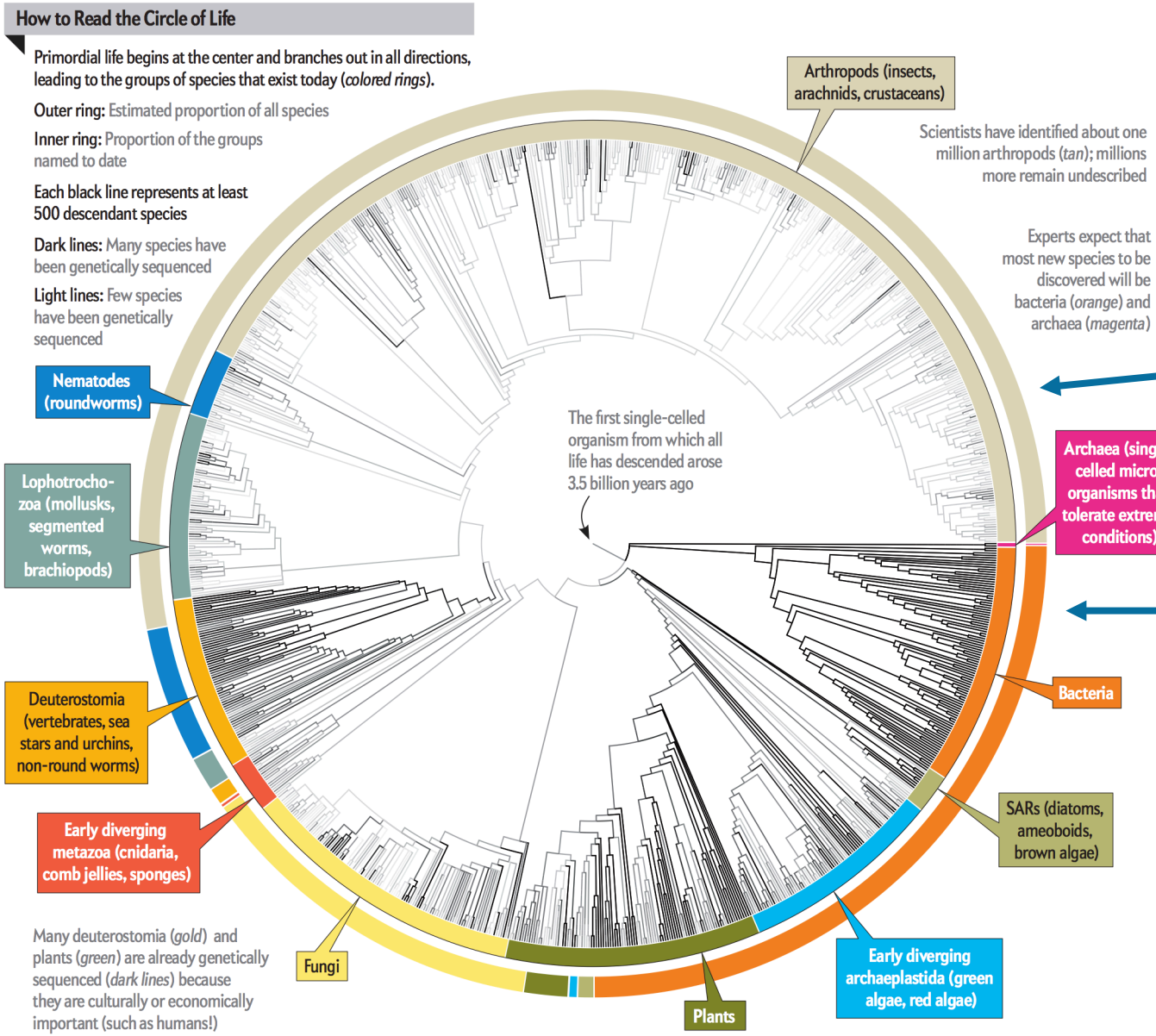
- Accurate and precise
- Updated
- Have branch lengths relative to time
- Comprehensive

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

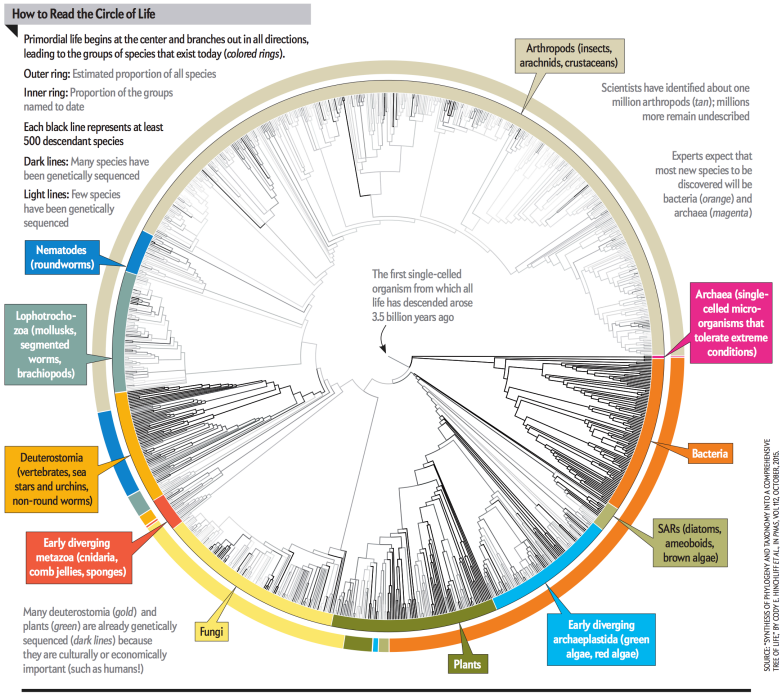


Informed by taxonomy

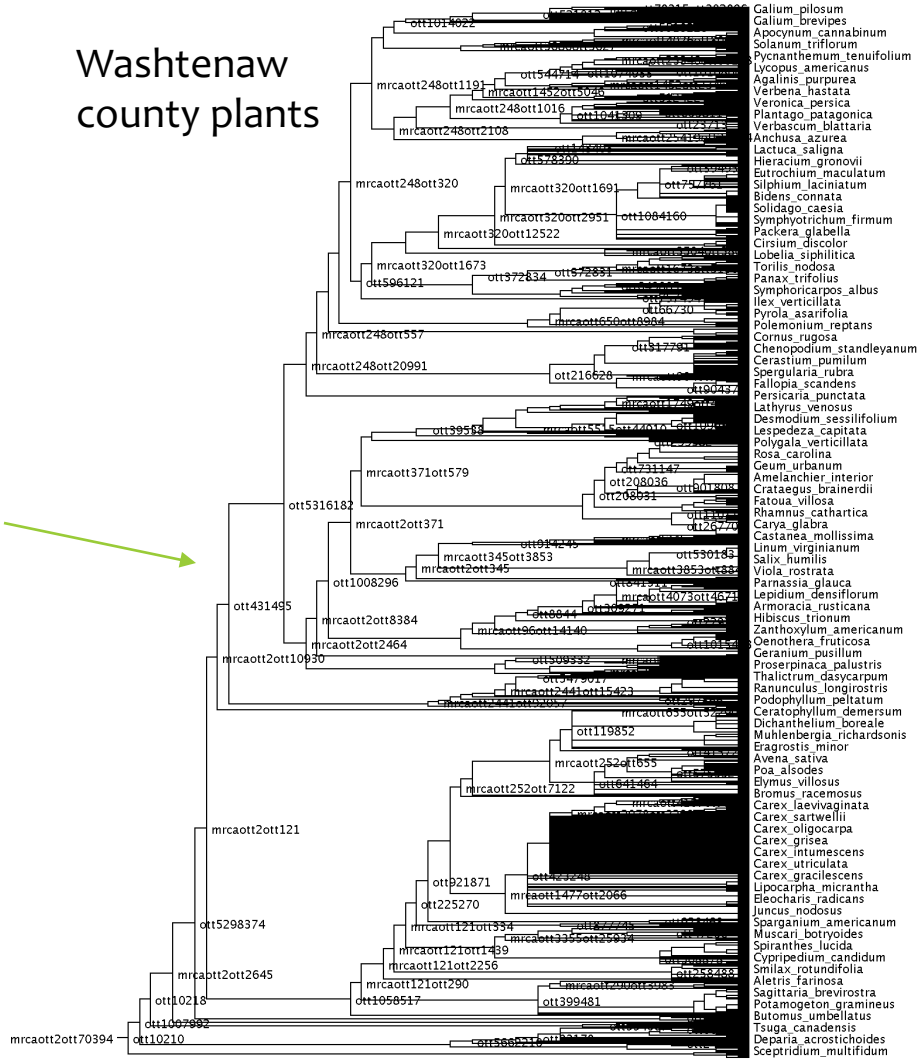
Informed by phylogenies

Hinchliff et al. 2015

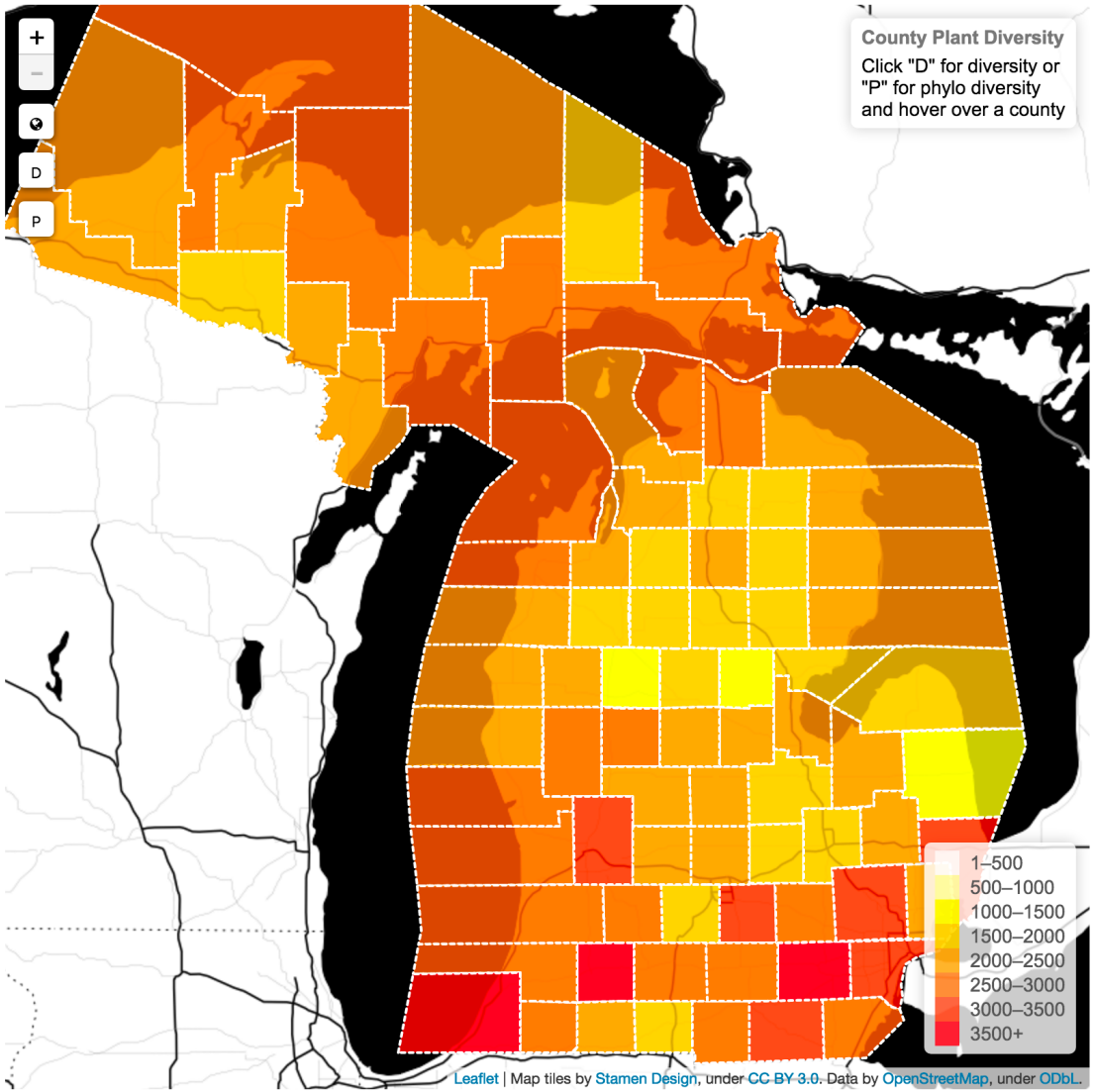
Phylogenies of the plants of Michigan



Washtenaw county plants



Michigan phylogenetic diversity



Comprehensive and dated tree for seed plants

LET'S SEE A BIG PLANT TREE

Stephen A. Smith

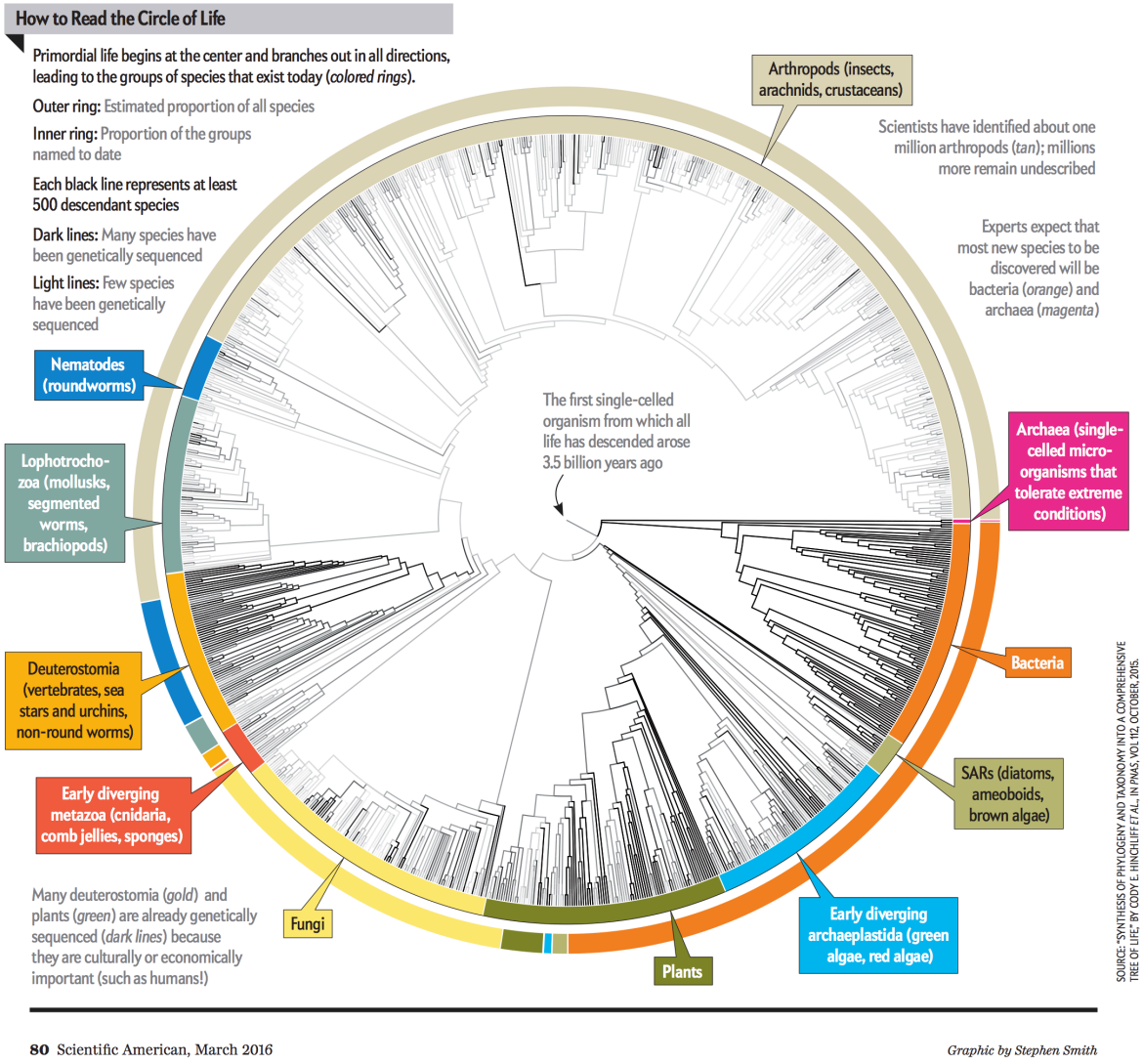
Tree from Open Tree of Life

Benefits

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

Limitations

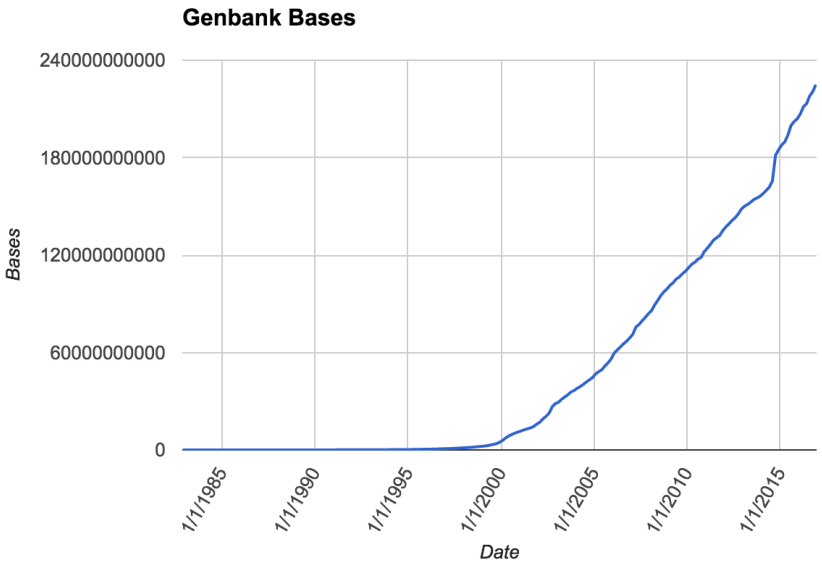
- Resolution may be poor
- No divergence times or branch lengths



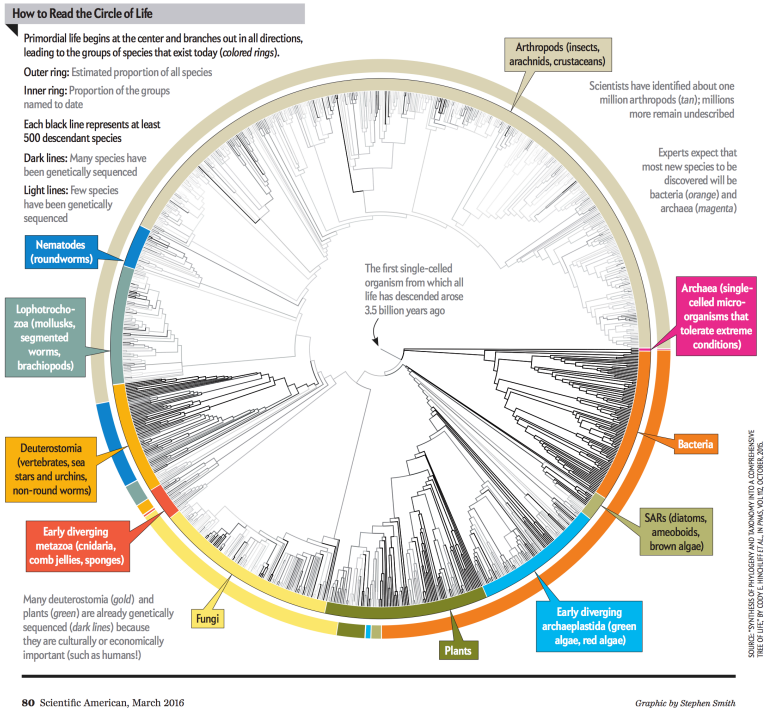
Combine GenBank and Open Tree

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)

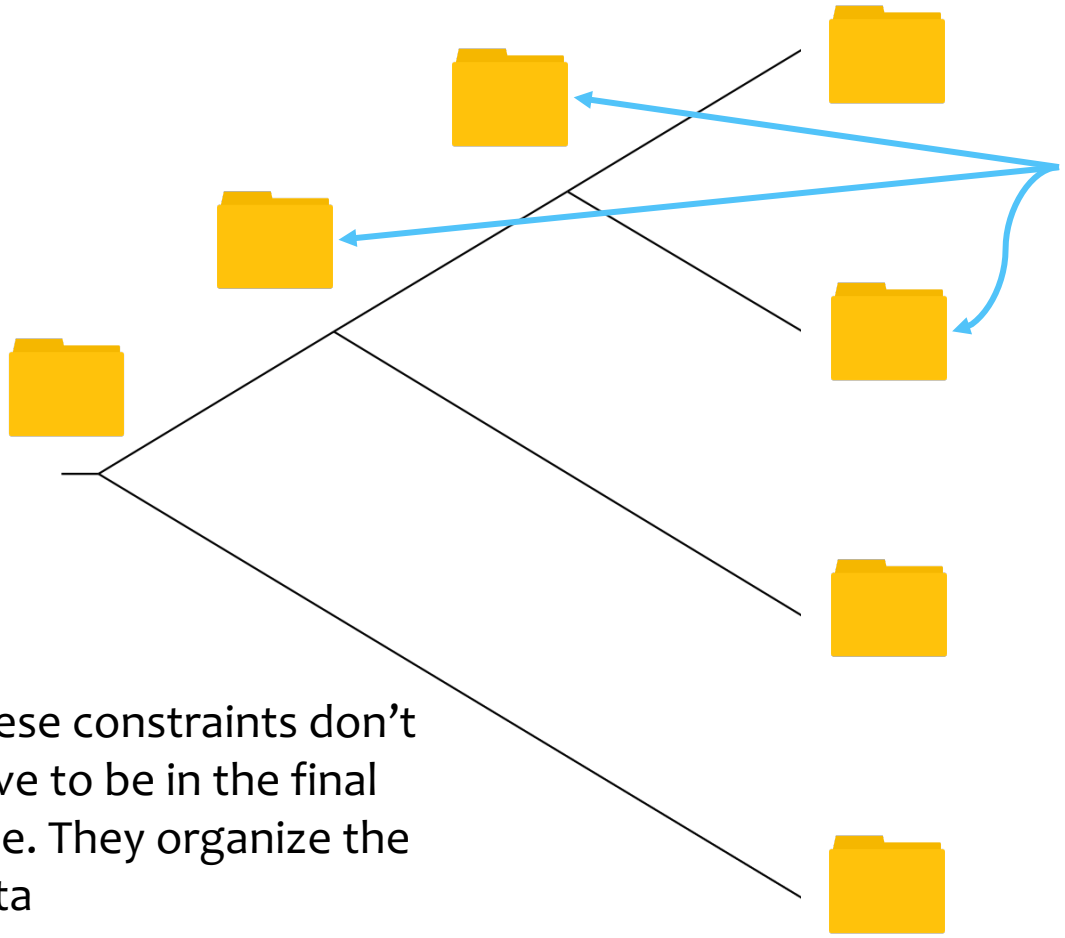
For each clade

- 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

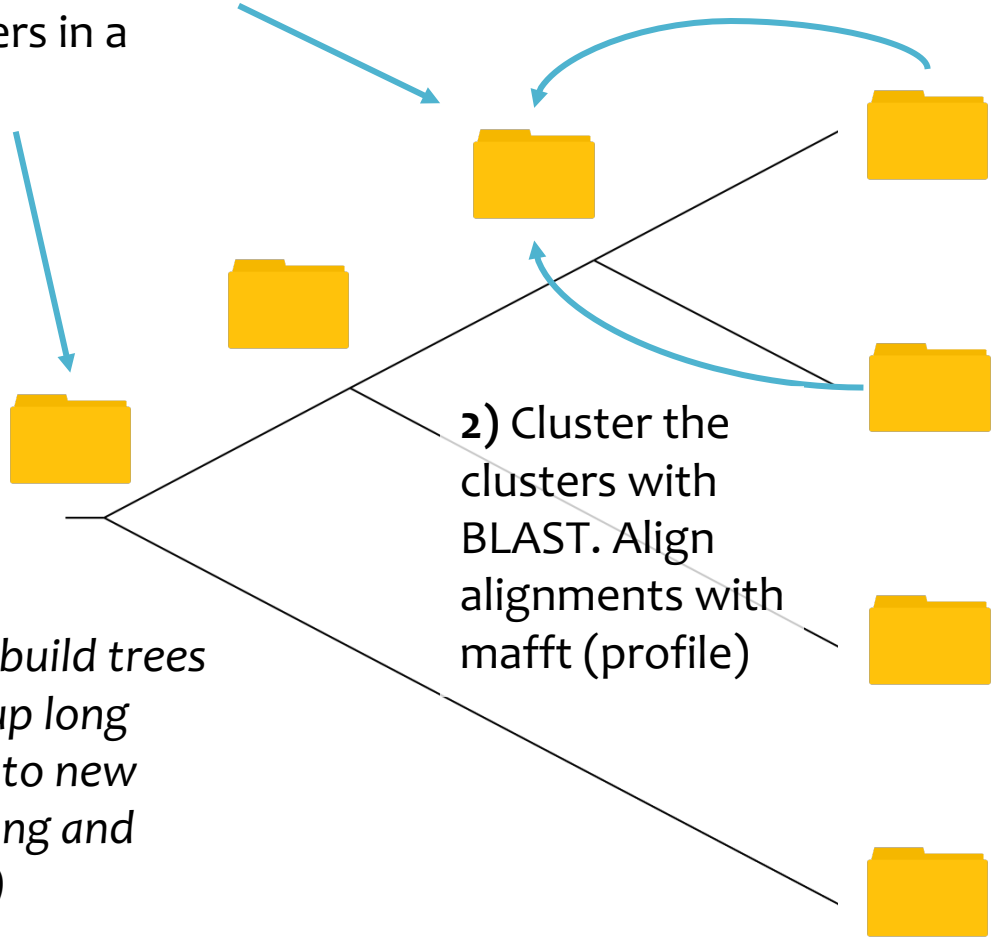


Each folder represents some putative monophyletic taxonomic group

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

Cluster and identify homologs

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



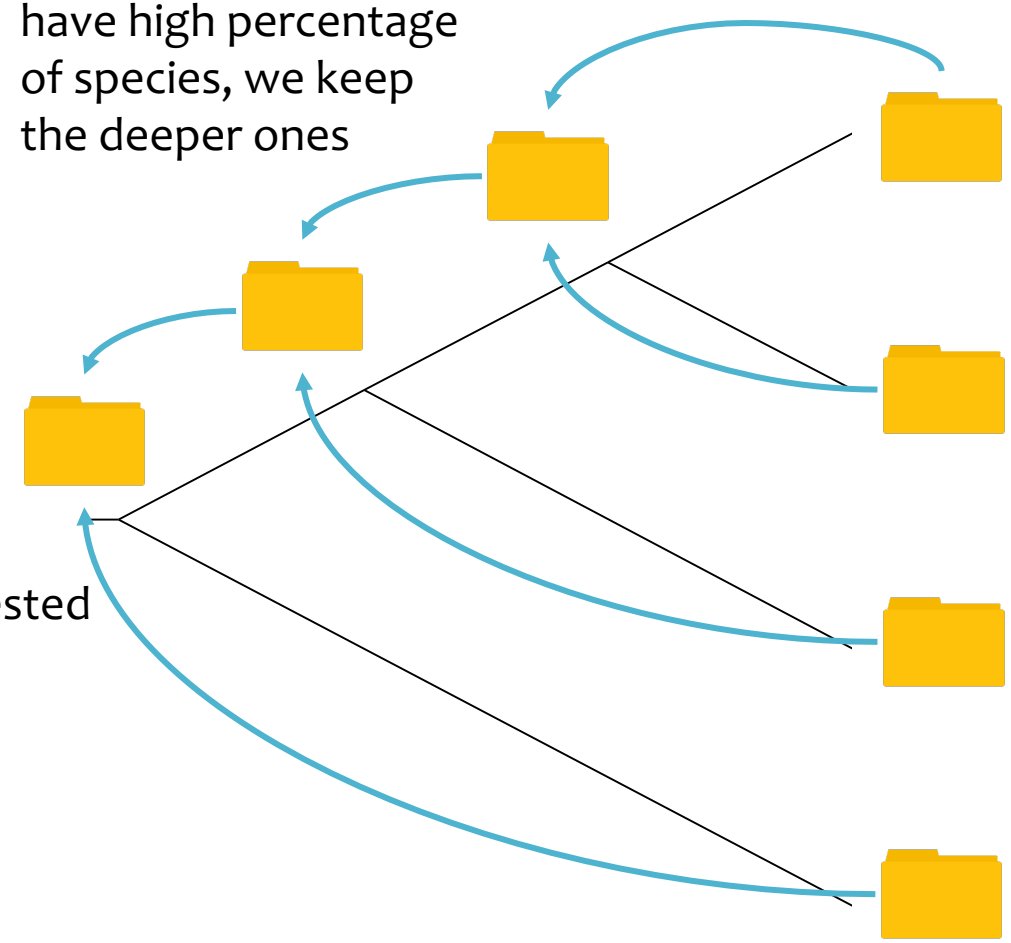
1) At the tips, we populate the folder with sequences and perform a clustering analysis (MCL)

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

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

Constructing datasets

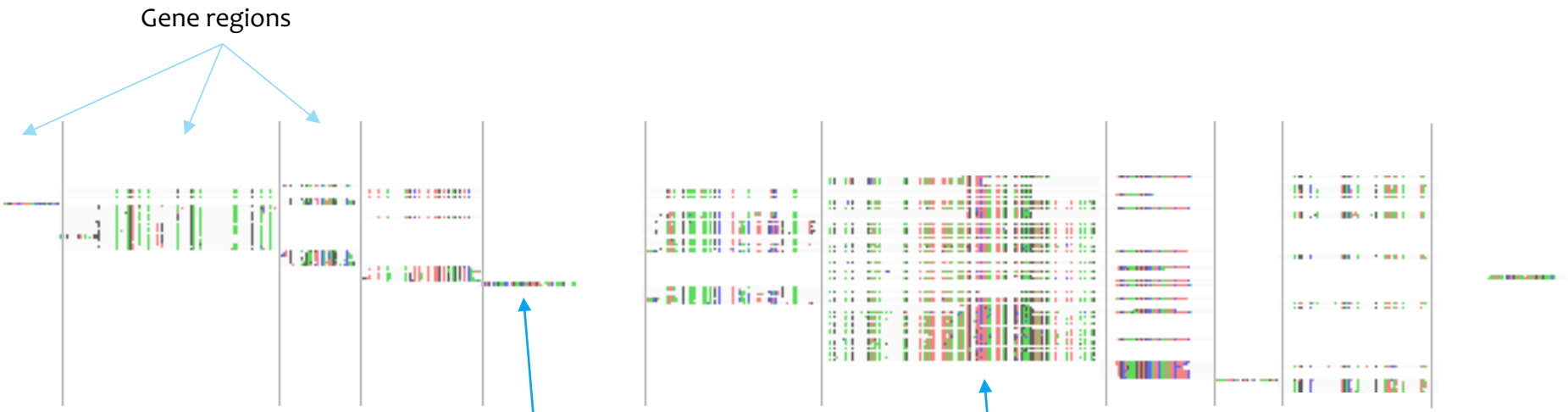
2) If clusters at internal nodes still have high percentage of species, we keep the deeper ones



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

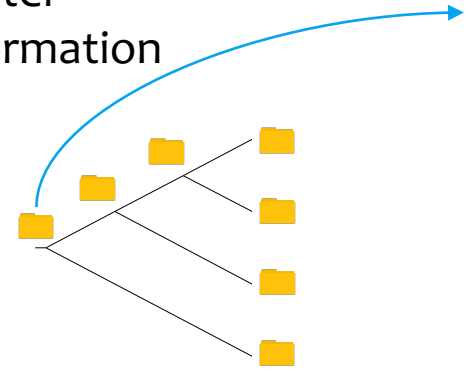
3) Keep good nested and good deep clusters

An alignment example



Website for each node/folder

Each clade has a website with cluster information



Ericales_41945

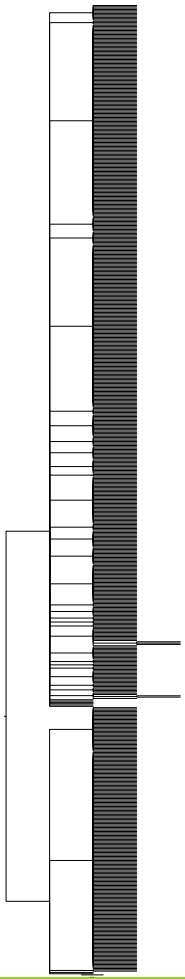
Actinidiaceae_3623	name	num_species	avg unaln len	defline
Balsaminaceae_25692	cluster1825.fa	2297	940.468437092	Eriastrum sparsiflorum voucher KANU:354712 tRNA-Leu (trnL) gene, partial sequence; trnL-trnF intergenic spacer, complete sequence; and tRNA-Phe (trnF) gene, partial sequence; chloroplast.
Clethraceae_16611				
Cyrillaceae_4339				
Diapensiaceae_16673				
Ebenaceae_19955	cluster2228.fa	2213	1417.07455942	Agapetes moorei chloroplast DNA, trnK intron including the matK gene, complete sequence, specimen_voucher: MBK:Kuroiwa et.al 030333.
Ericaceae_4345				
Fouquieriaceae_24902				
Lecythidaceae_3642	cluster2168.fa	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.
Marcgraviaceae_55360				
Mitrastemonaceae_91826				
Myrsinaceae_16614				
Pentaphylacaceae_125045	cluster1797.fa	940	1518.97340426	Schima superba NADH dehydrogenase (ndhF) gene, partial cds; chloroplast gene for chloroplast product.
Polemoniaceae_24584				
Primulaceae_4335	cluster1850.fa	938	1213.72601279	Micropholis obscura voucher P00610293 ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit (rbcl) gene, partial cds; plastid.
Roridulaceae_91900				
Sapotaceae_3737				
Sarraceniaceae_4353	cluster2360.fa	680	936.613235294	Lysimachia remyi subsp. remyi voucher Marr 424 (NY) ribosomal protein L16 (rpl16) gene, intron; chloroplast.
Sladeniaceae_235238				
Styracaceae_20008				
Symplocaceae_20019	cluster2211.fa	635	780.009448819	Aubregrinia taiensis atpB-rbcL intergenic spacer, partial sequence; chloroplast.
Ternstroemiaceae_91898				
Tetrameristaceae_91901	cluster900.fa	501	792.896207585	Erica alexandri subsp. alexandri isolate alexandri_EO12449 trnT-trnL intergenic spacer region, partial sequence; chloroplast.
Theaceae_27065				

Assume monophyly until proven otherwise

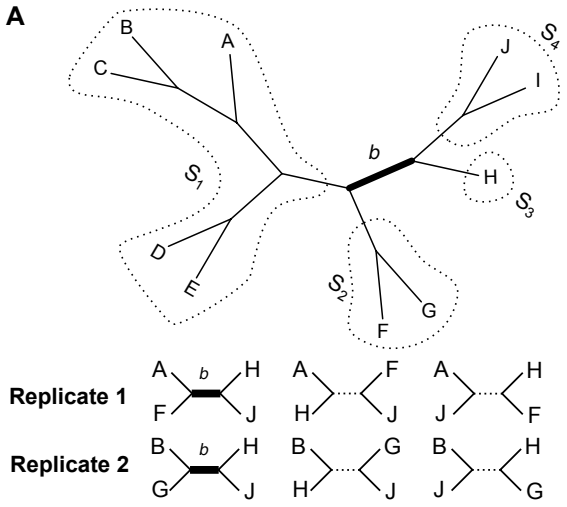
- Some edges are *impossible* to reconstruct without genomic data or otherwise difficult to assemble datasets
- Can we assume that things are monophyletic?
  - Certain clades?
  - Major Linnaean groups? Orders, Families, Genera?
- Do we have to confirm this every run?

Constraint procedure

1) Calculate phylogeny with a constraint tree based on taxonomy

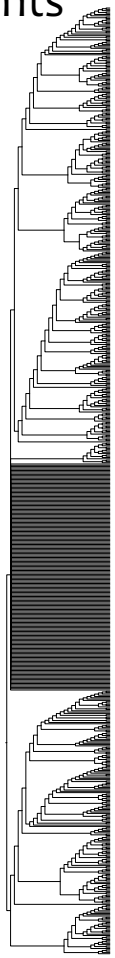


2) Test constraints using a quartet procedure (when there is strong conflict, constraints are removed)



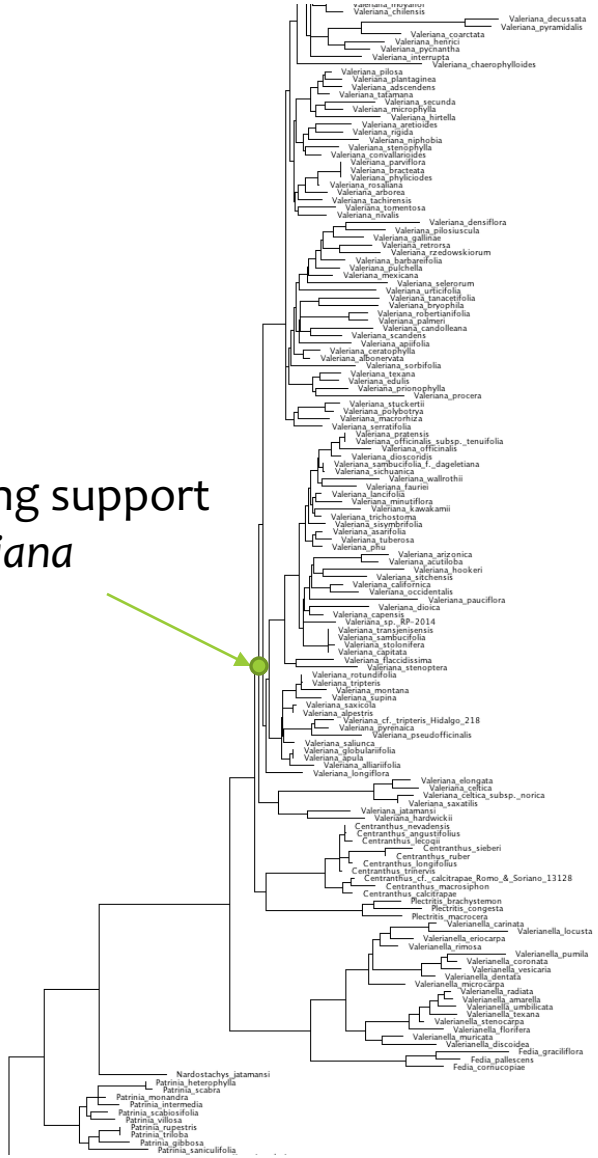
Pease et al. *in review*

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

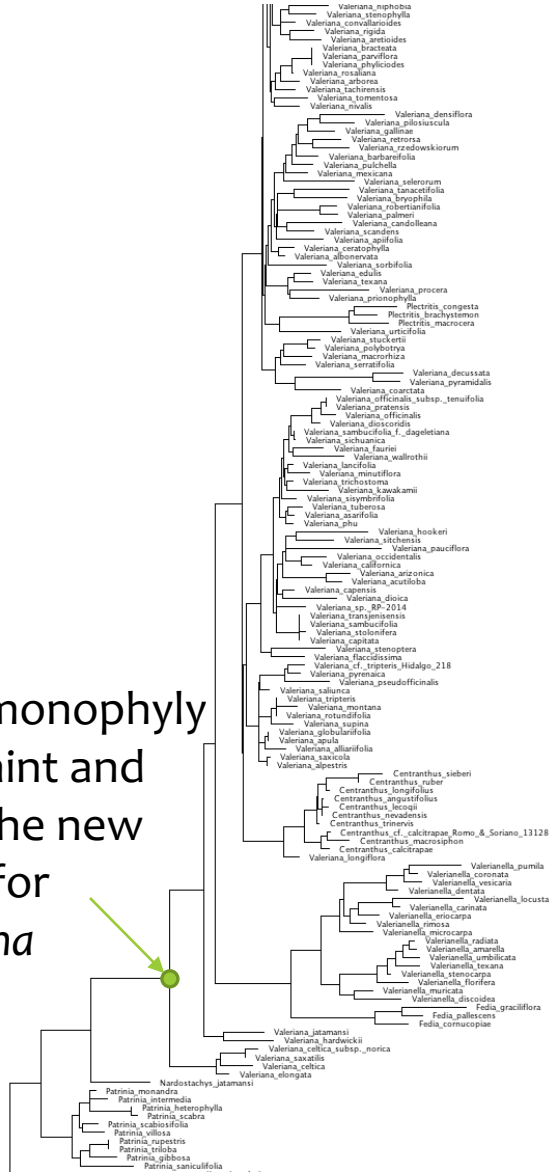


Example with Dipsacales

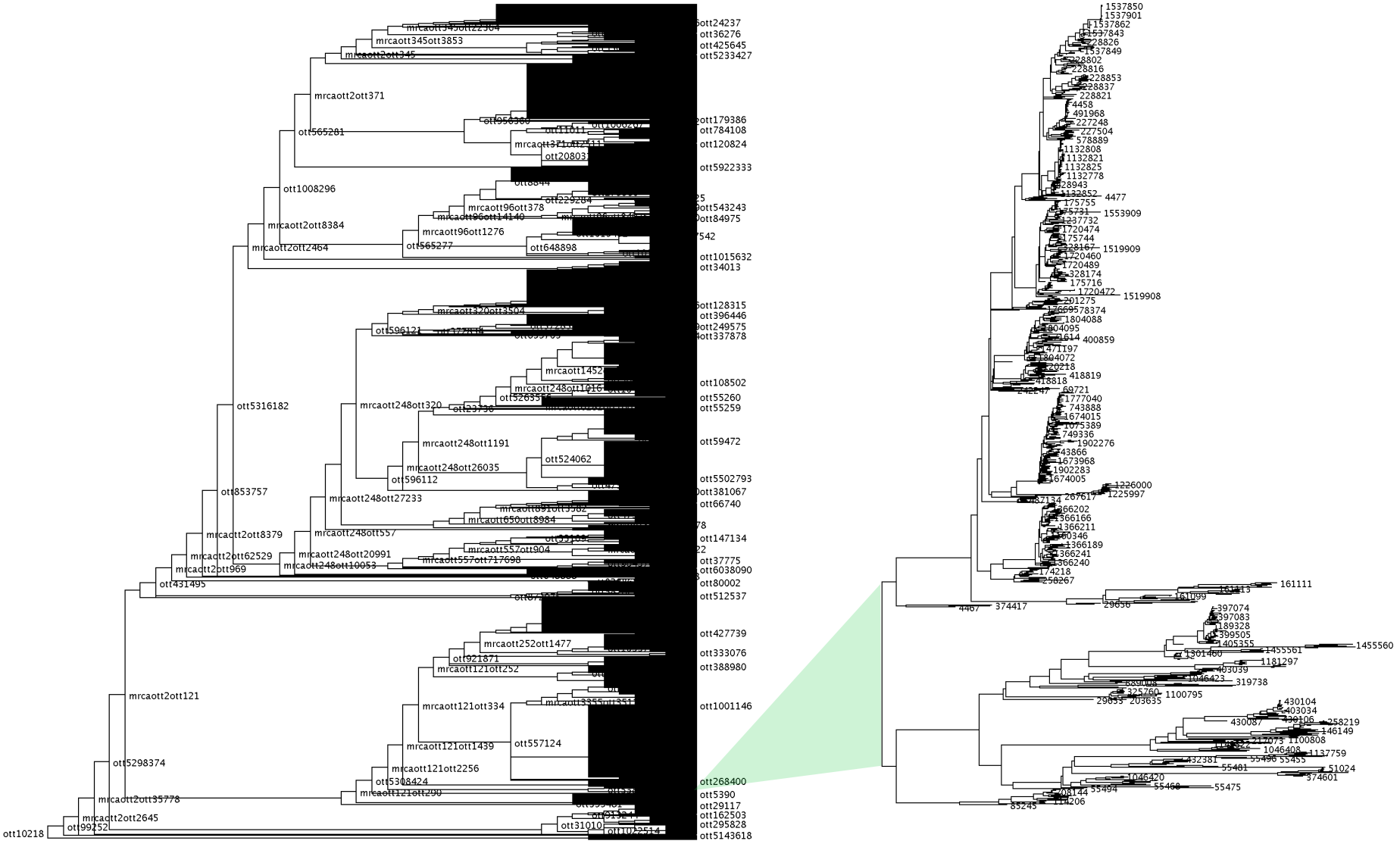
Conflicting support for *Valeriana*



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



Replace clades in the backbone/synth tree



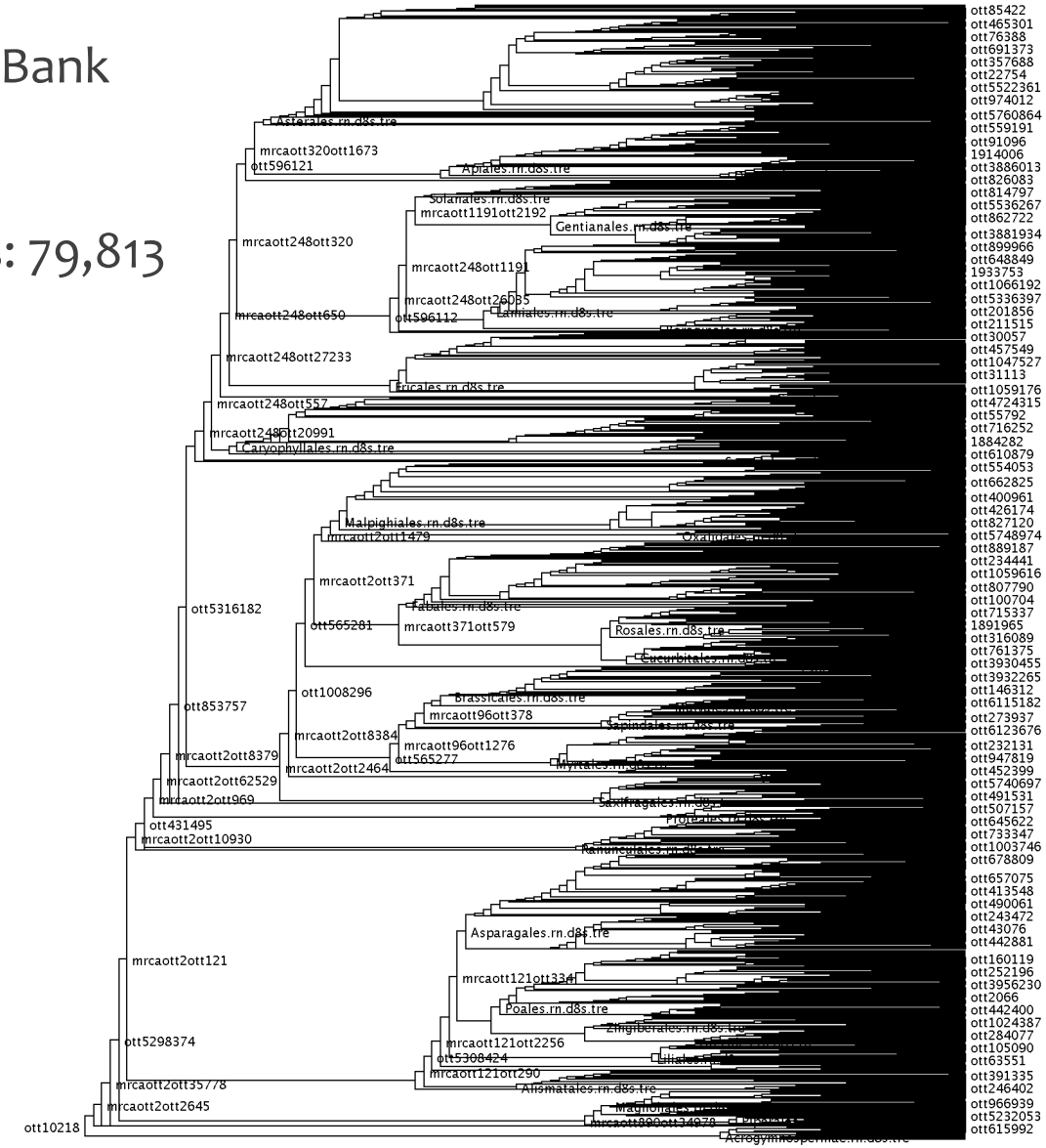
What does the data look like?

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

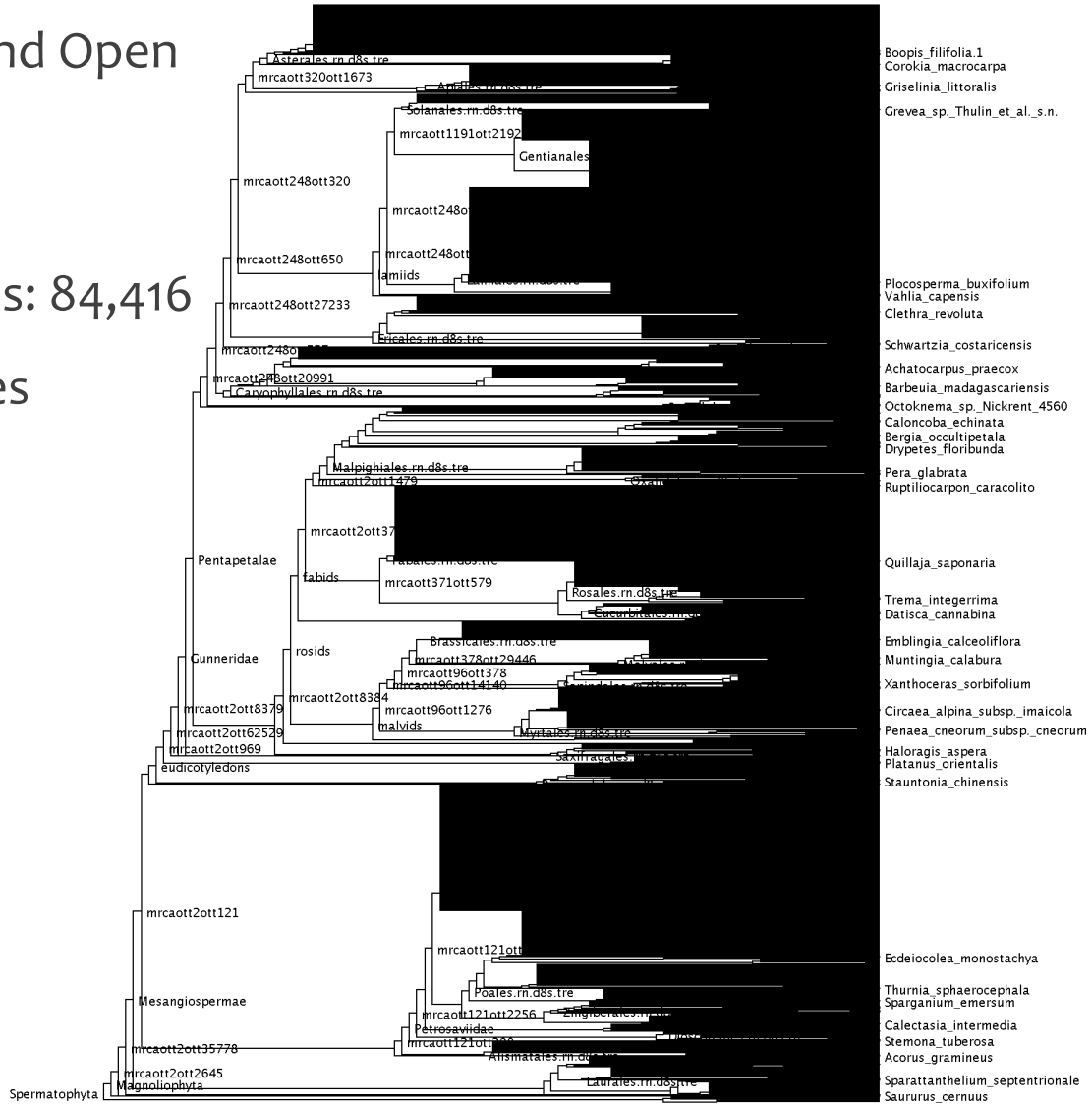
What does the data look like?

356,807 with GenBank and Open Tree

Number of tips: 356,807

Number of internal nodes: 84,416

272,390 unresolved nodes

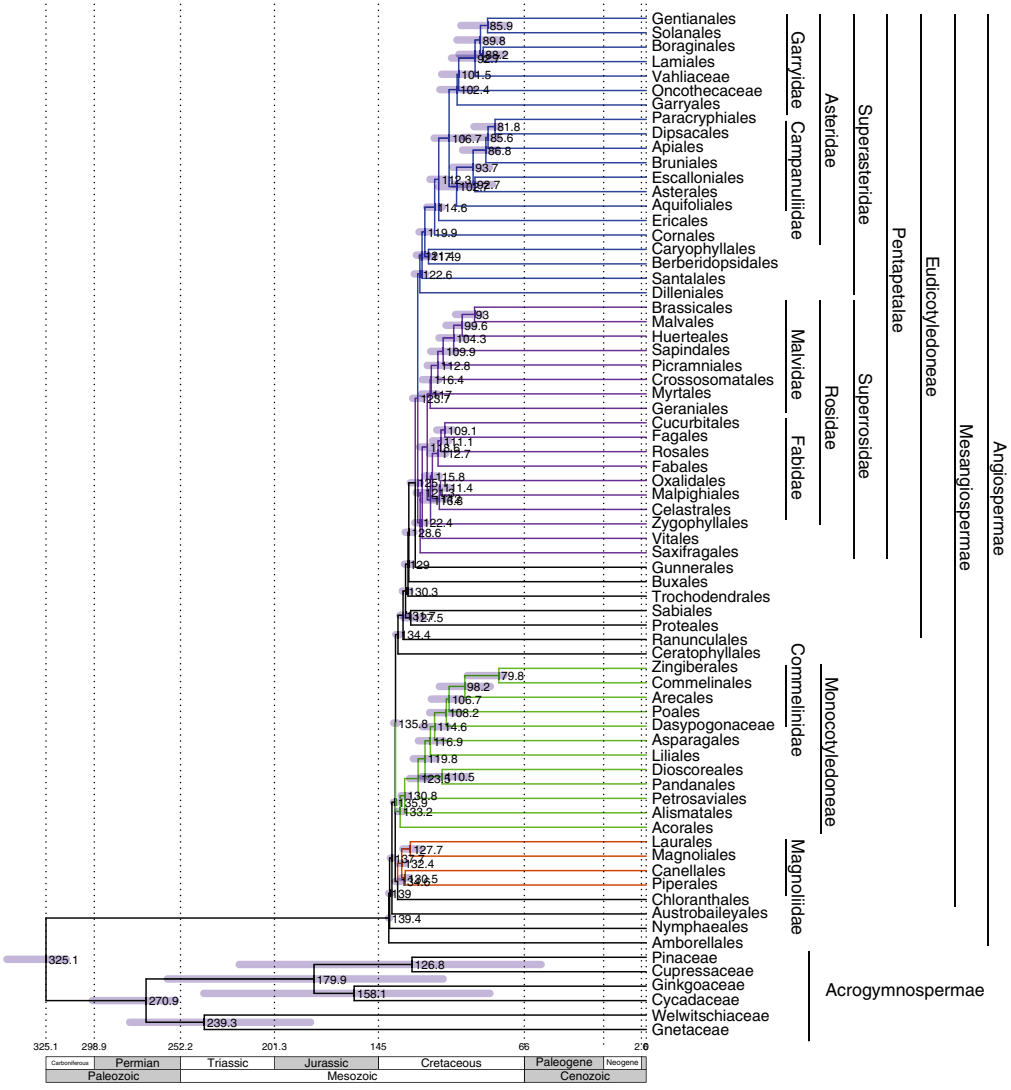


Divergence time estimation

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

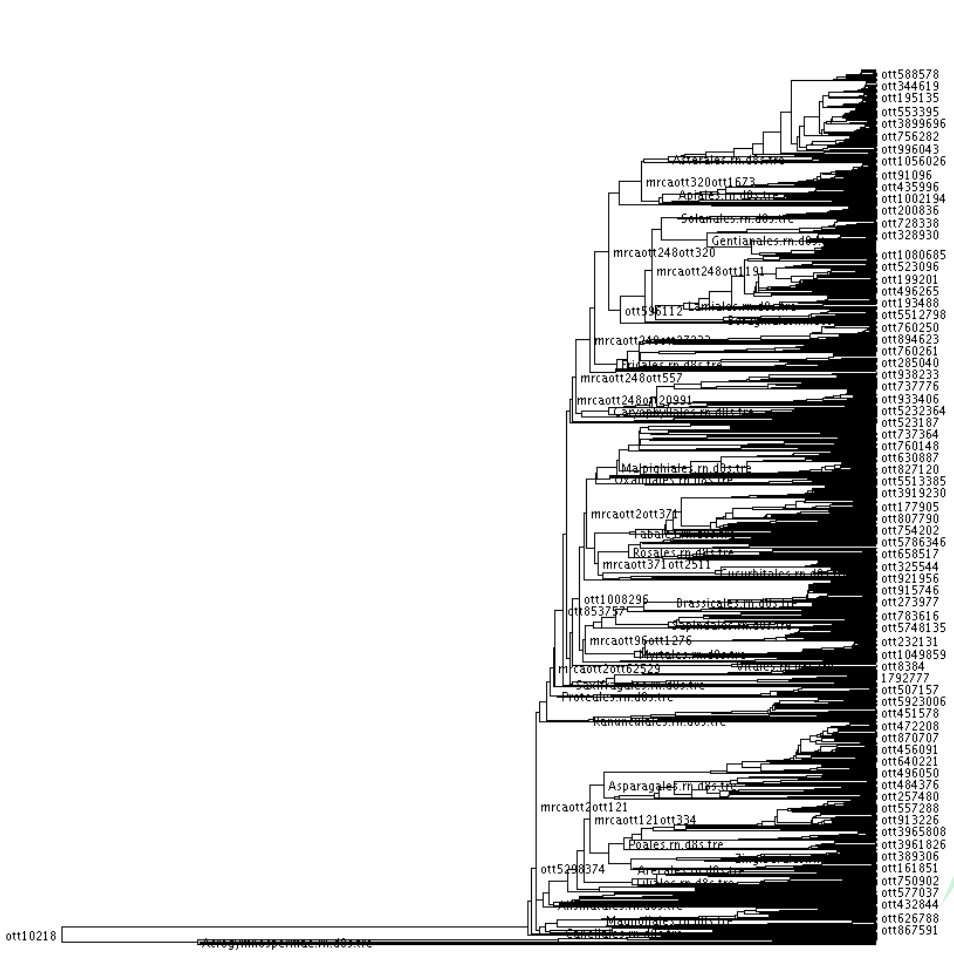
We run with treePL (Smith and O'meara 2012) on each clade

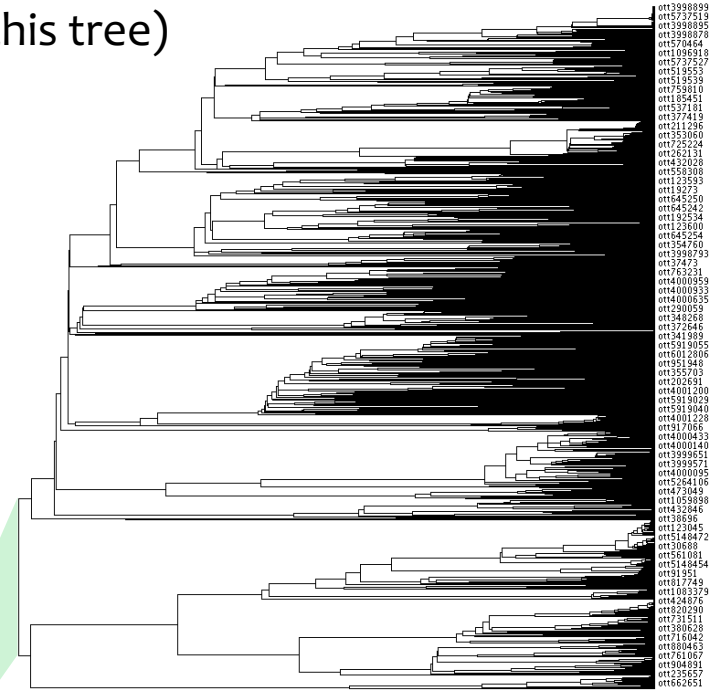


Magallon et al. 2015

Dated Alismatales

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

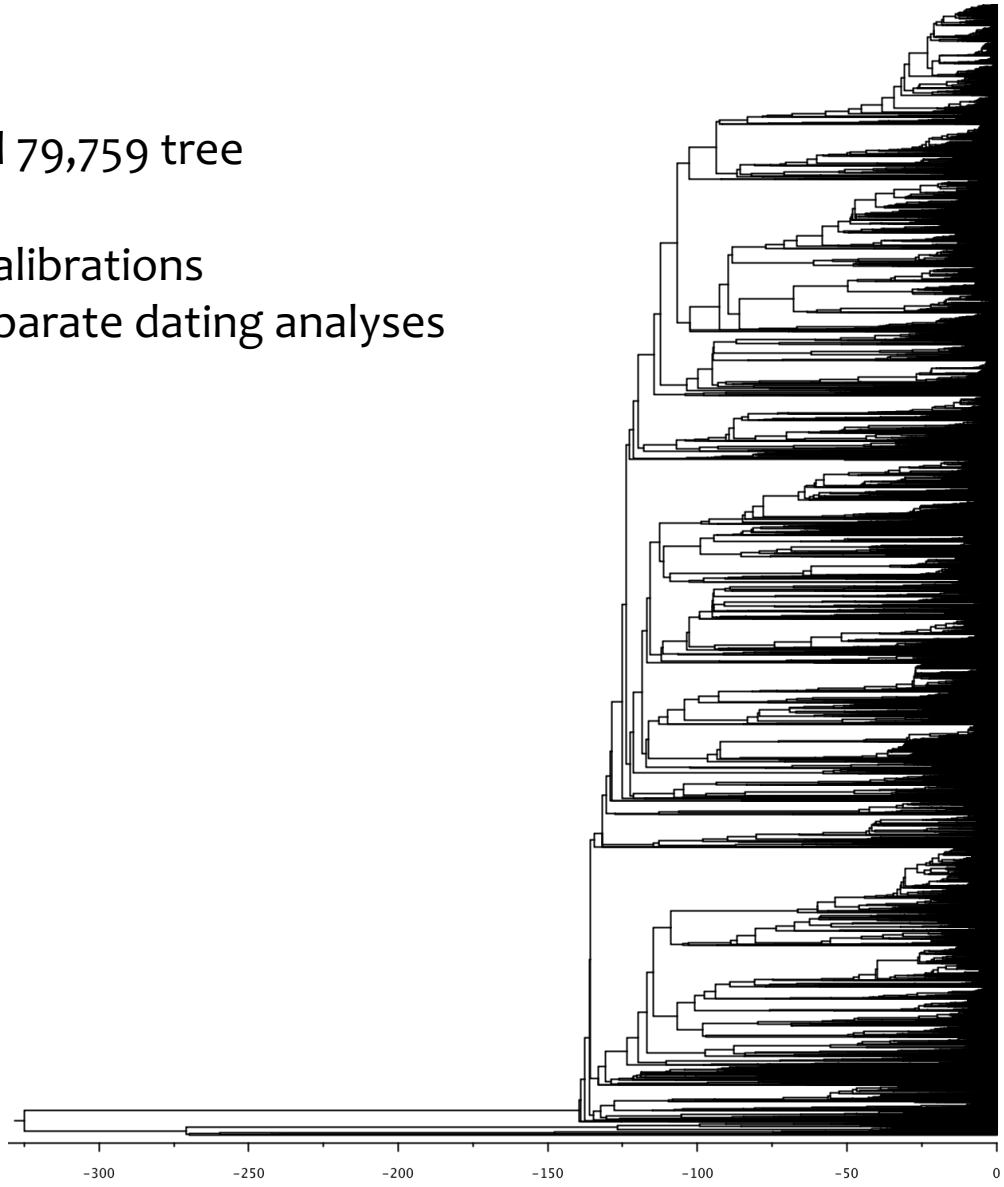




With divergence times

Dated 79,759 tree

590 calibrations
60 separate dating analyses

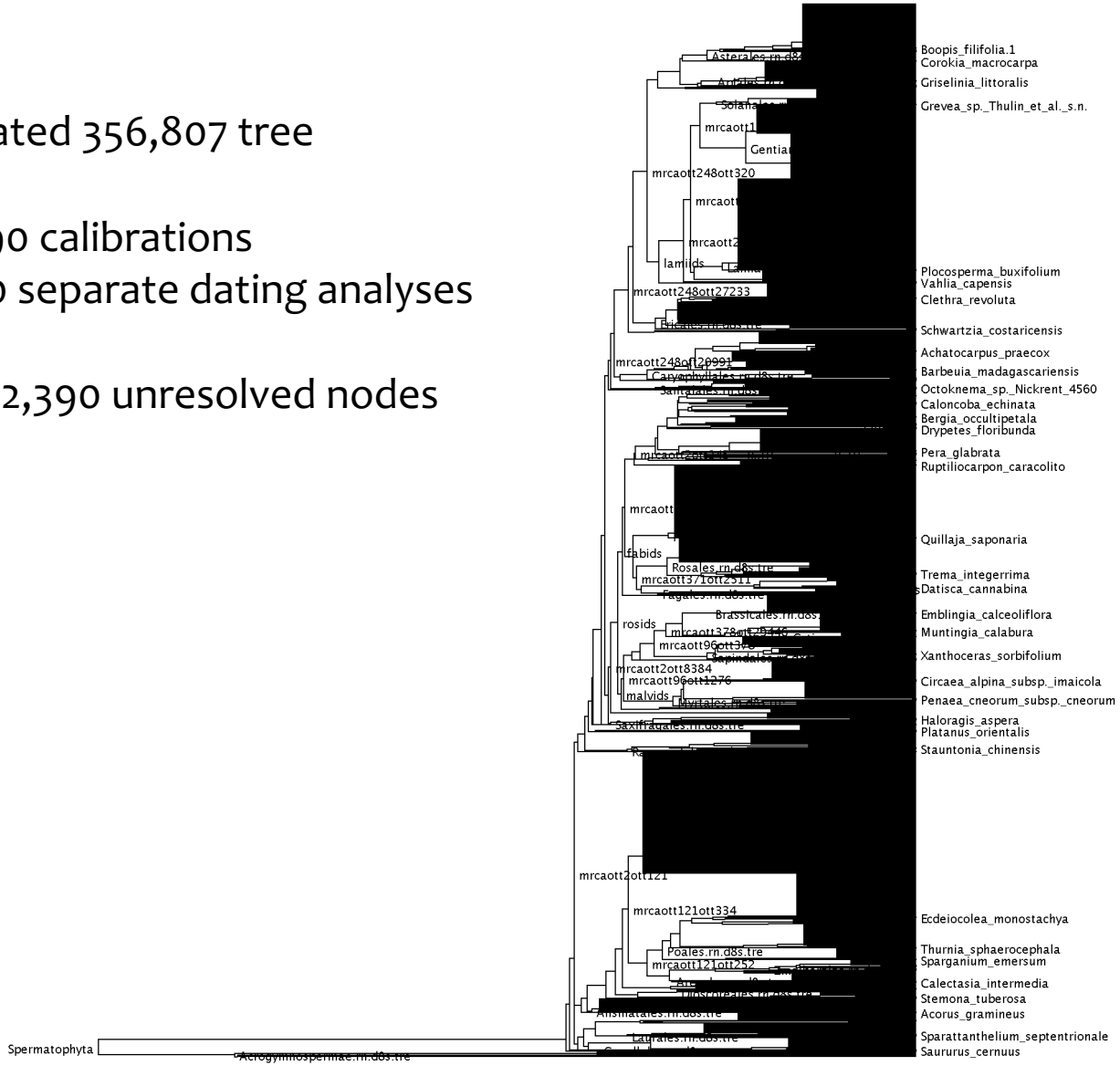


With divergence times and Open Tree

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

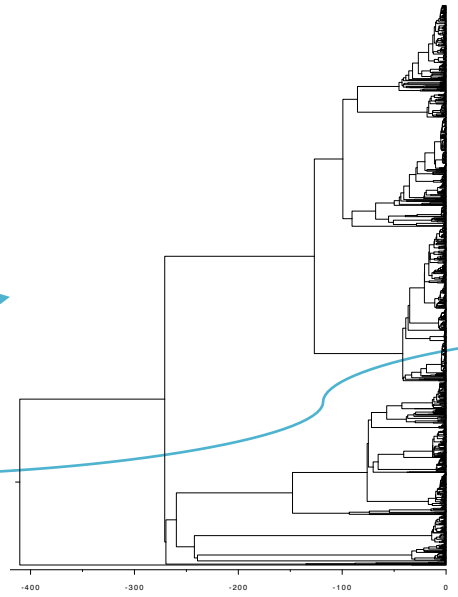
Pros

- Reasonable placement of taxa
- Reasonable divergence time estimates

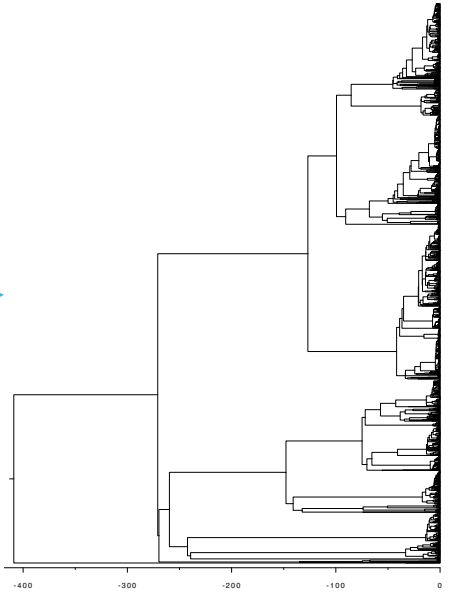
Cons

- 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

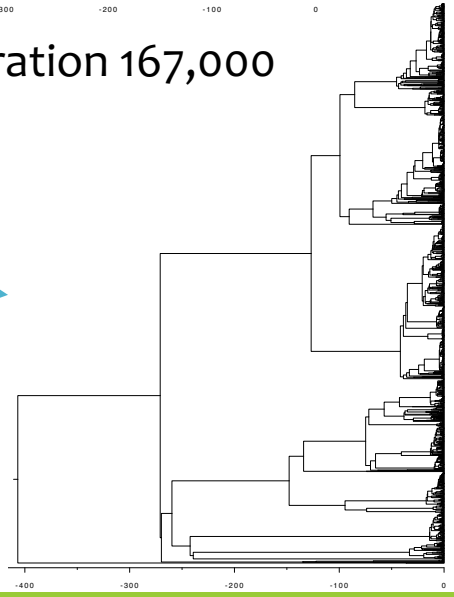
Example with Acrogymnospermae



Generation 167,000



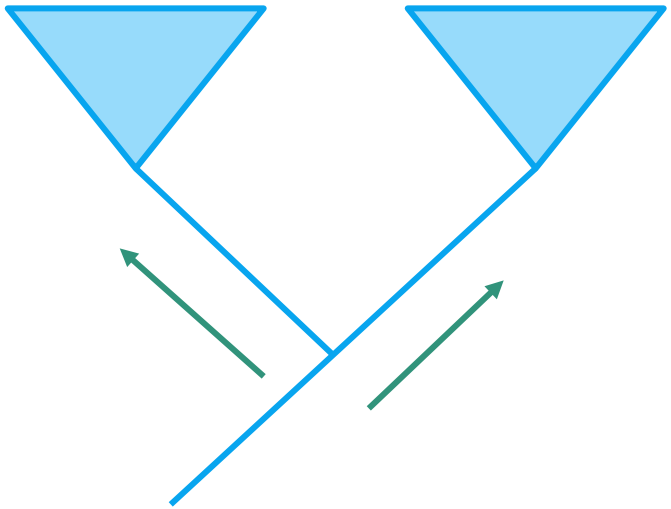
Generation 435,000



Generation 935,000

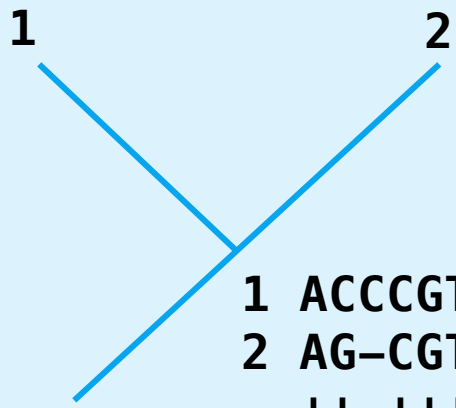
How about data overlap

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



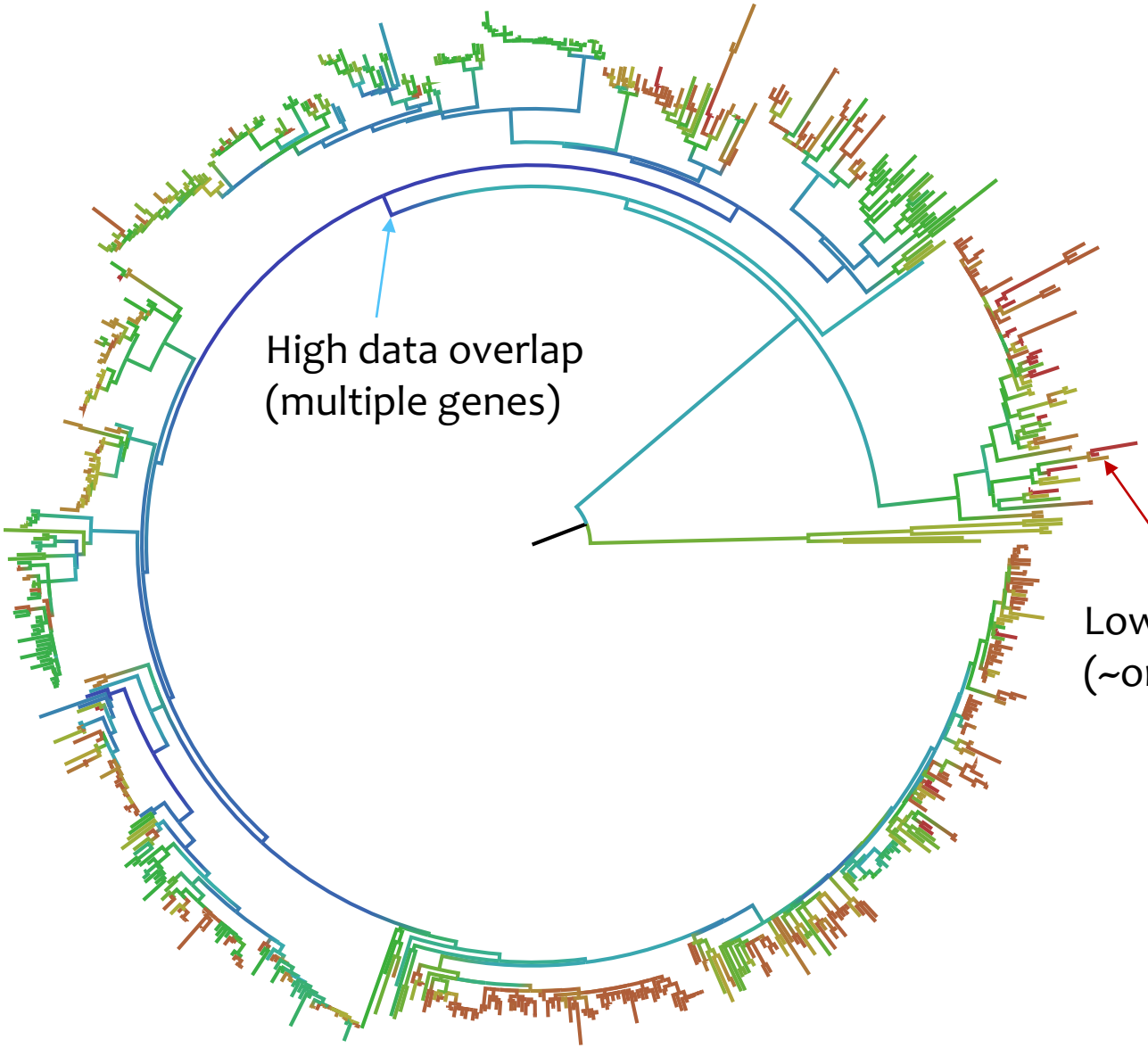
How many sites overlap between any taxa on the left and any taxa on the right?

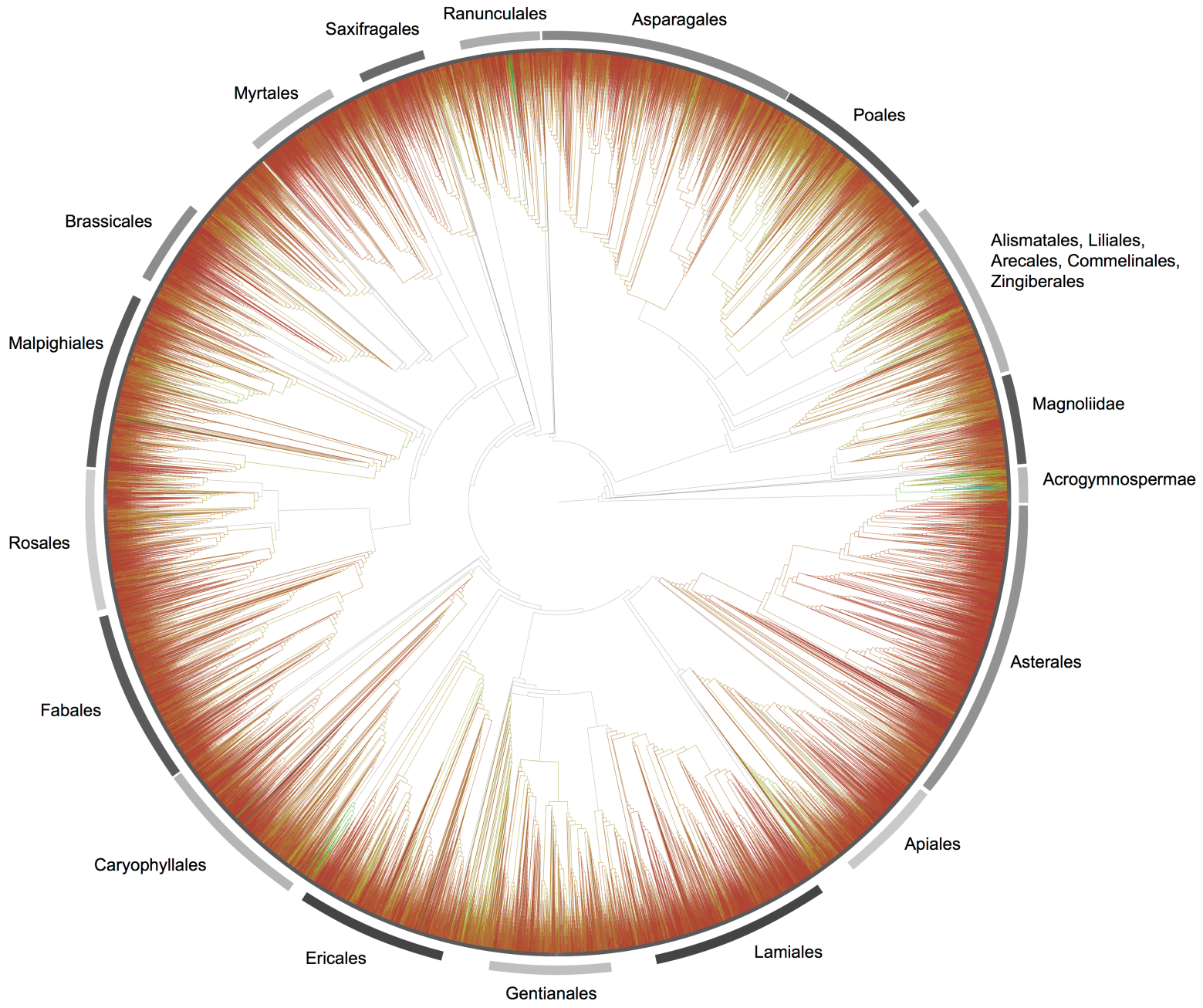
Example

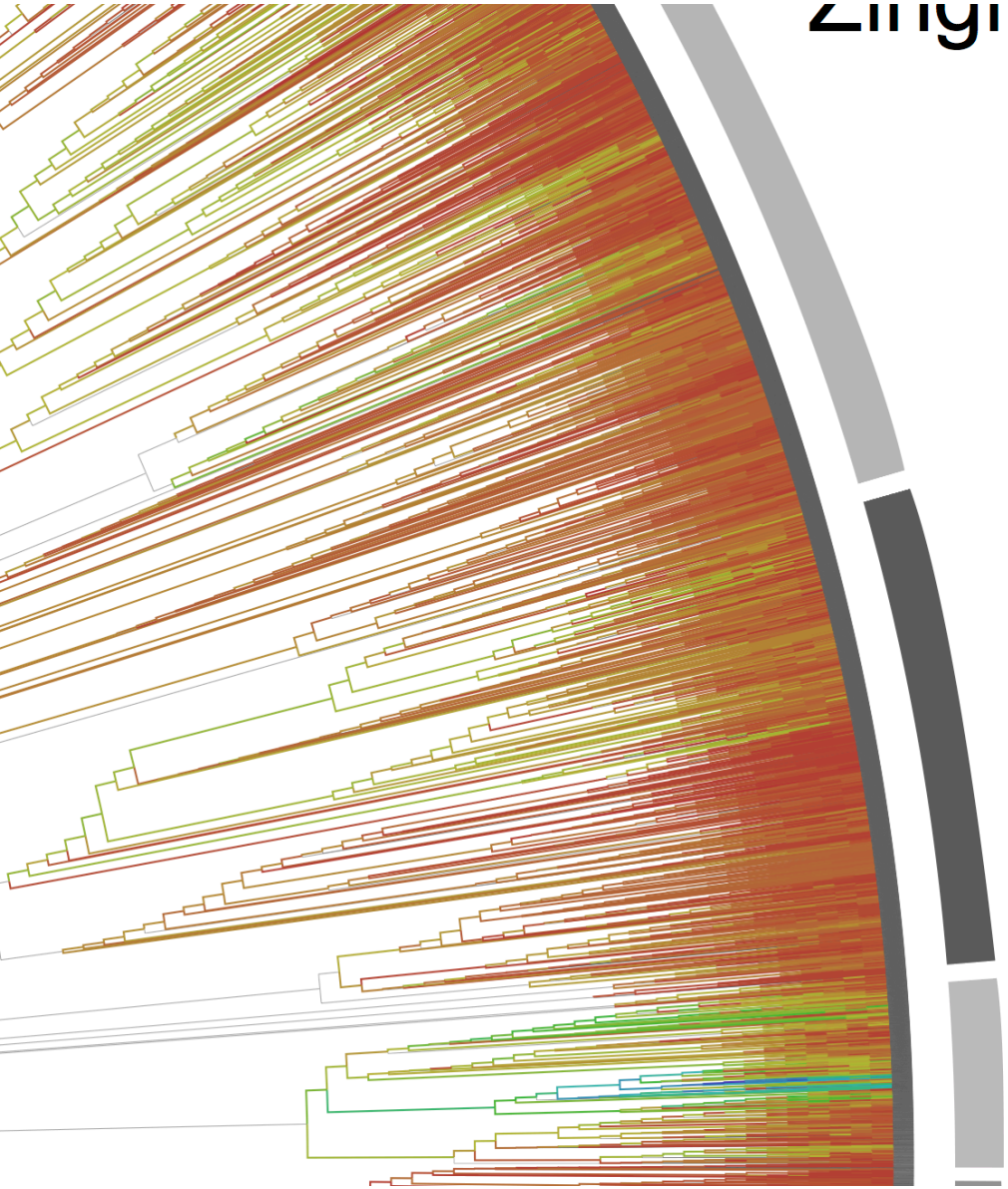



```
1 ACCCGTTT----GTGG
2 AG-CGTTT----G---
  ++ +++++    +   = 8
```

Arecales data overlap




Ranunculales
Asparagales
Saxifragales
Myrtales
Poales
Brassicales
Alismatales, Liliales,
Arecales, Commelinales,
Zingiberales
Malpighiales
Magnoliidae
Acrogymnospermae
Rosales
Asterales
Fabales
Apiales
Caryophyllales
Lamiales
Ericales
Gentianales



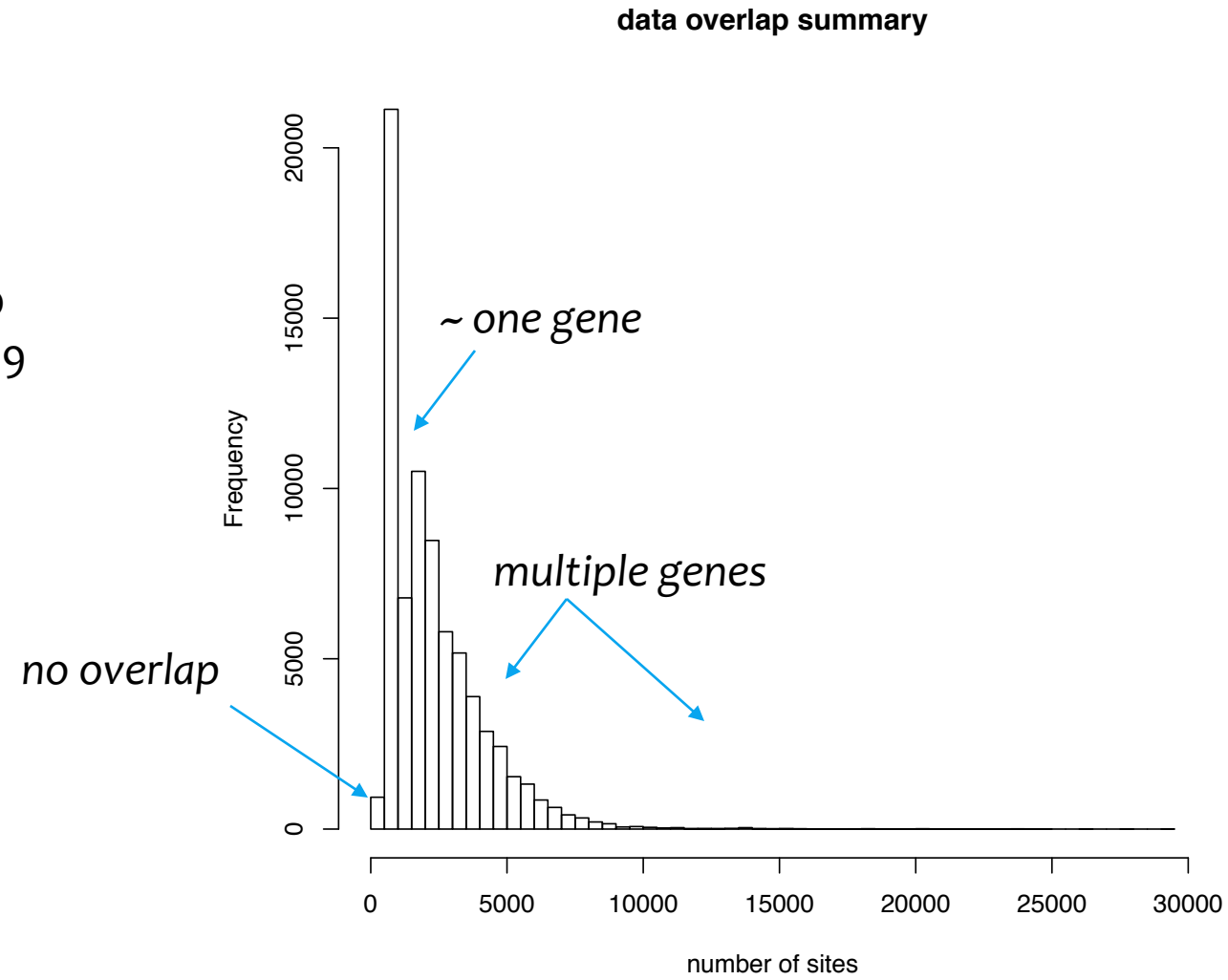
Zingiberales

Magnoliidae

Acrogymnospermae

Summary

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



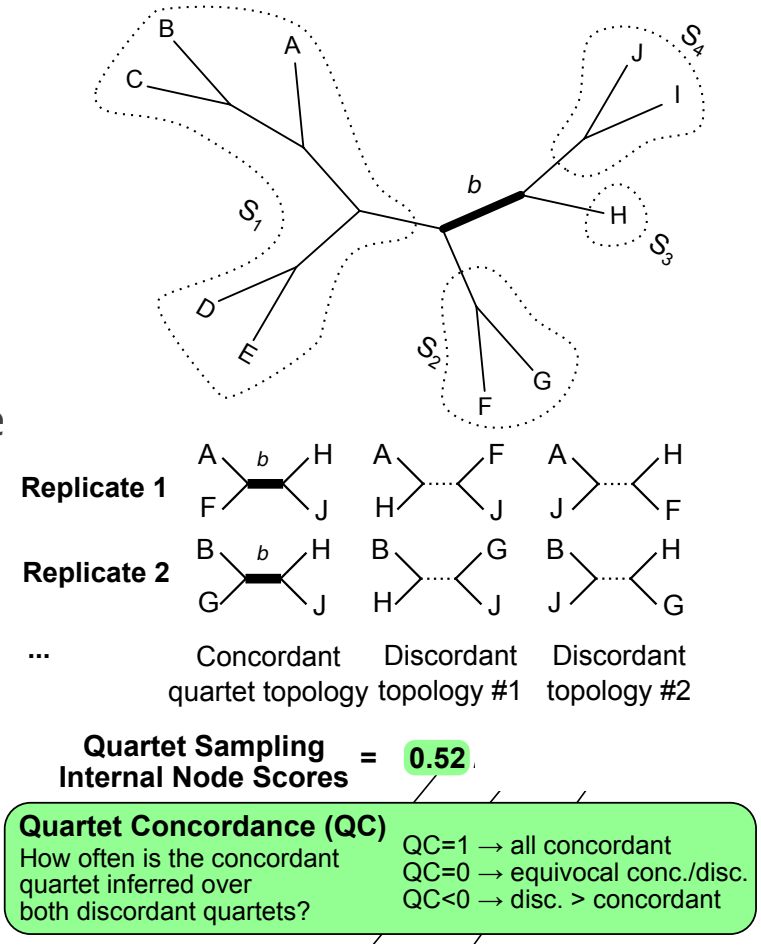
How about support?

We would like to ascertain support

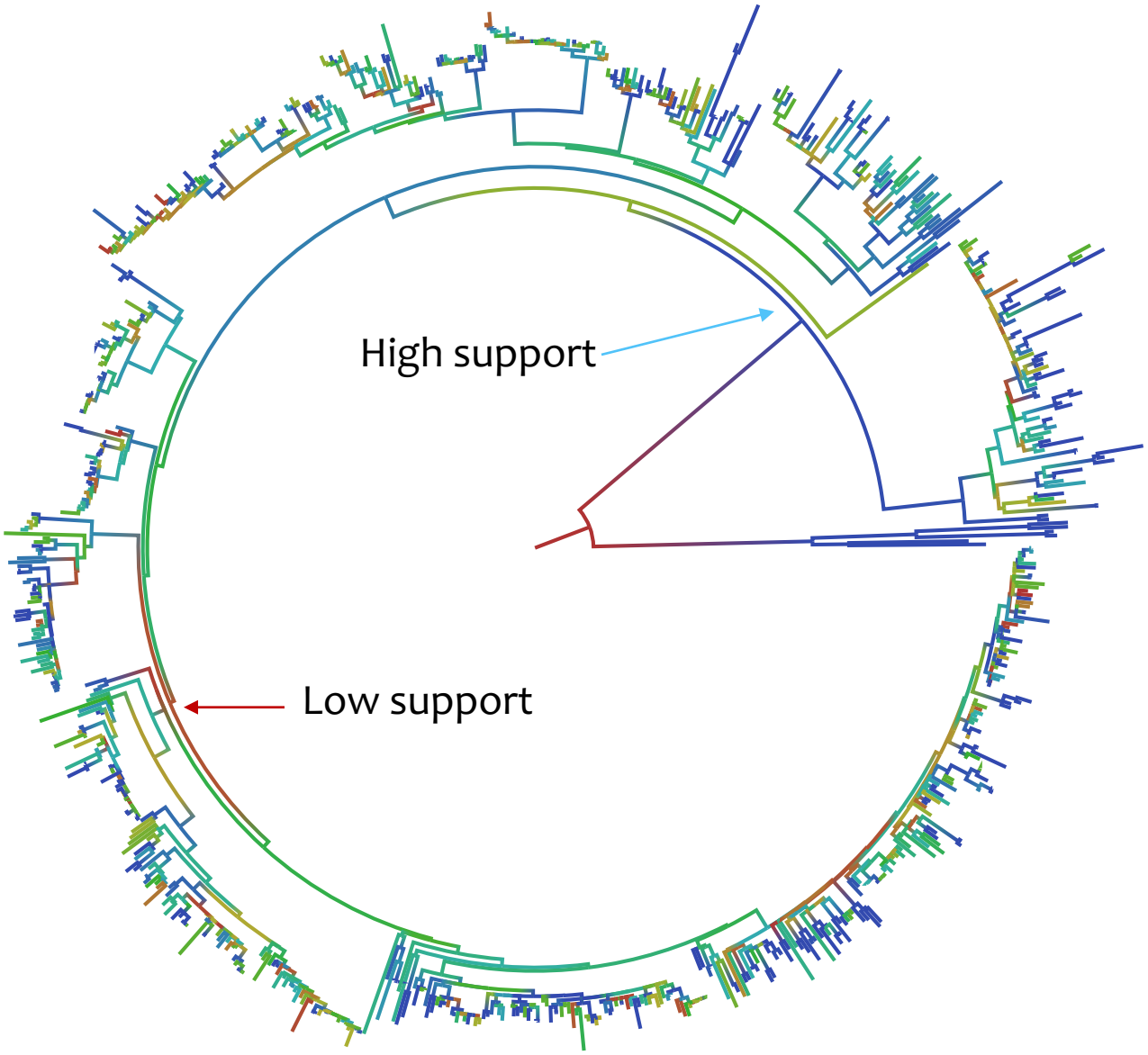
- Bootstrap is too slow and maybe not what we want
- Bayesian analysis is not going to happen
- aBayes and SH-Like are fine but aBayes is limited and SH-Like requires that you have the ML tree (we don't because of constraints)

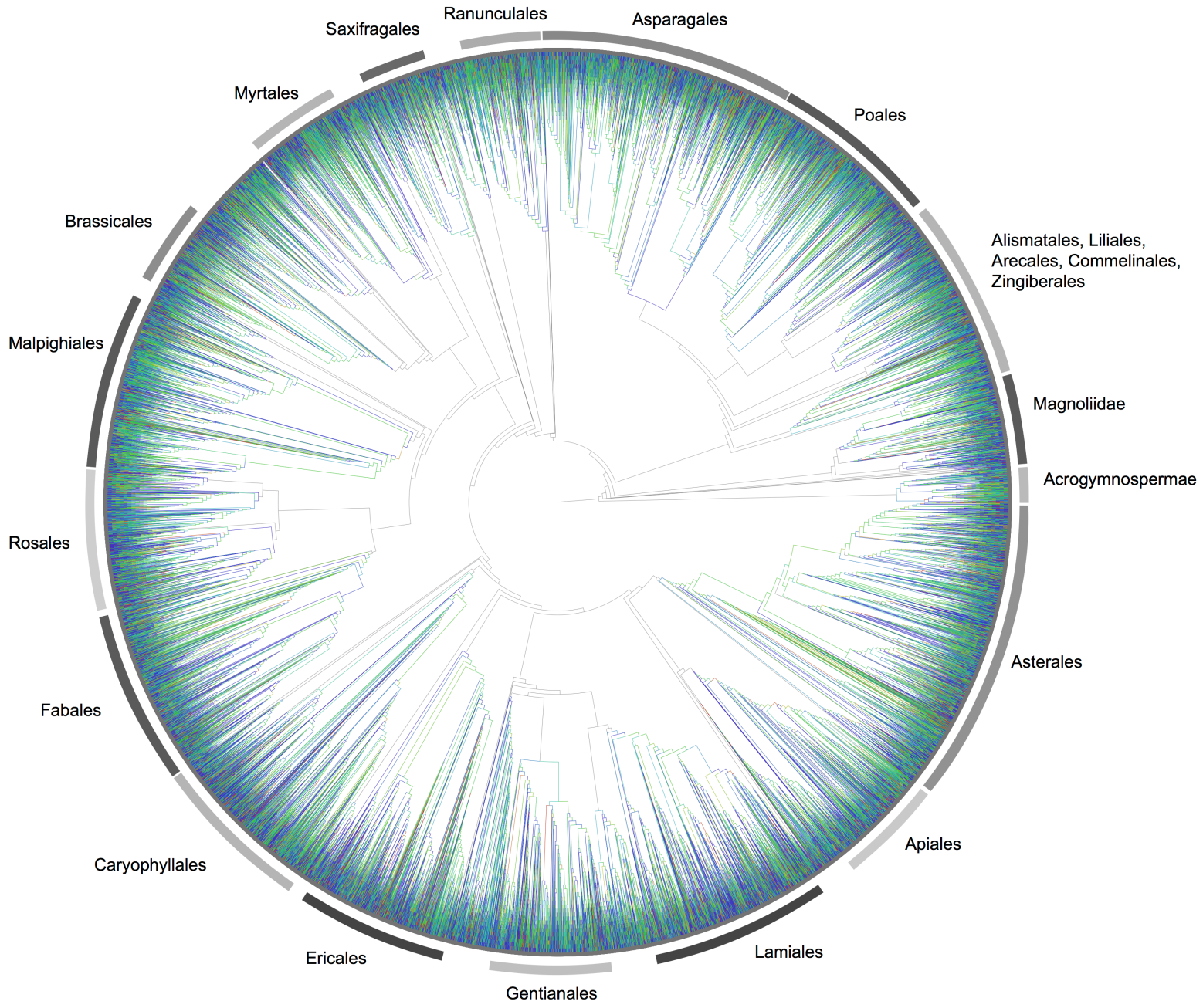
To do this, we are using Quartet Concordance (QC)

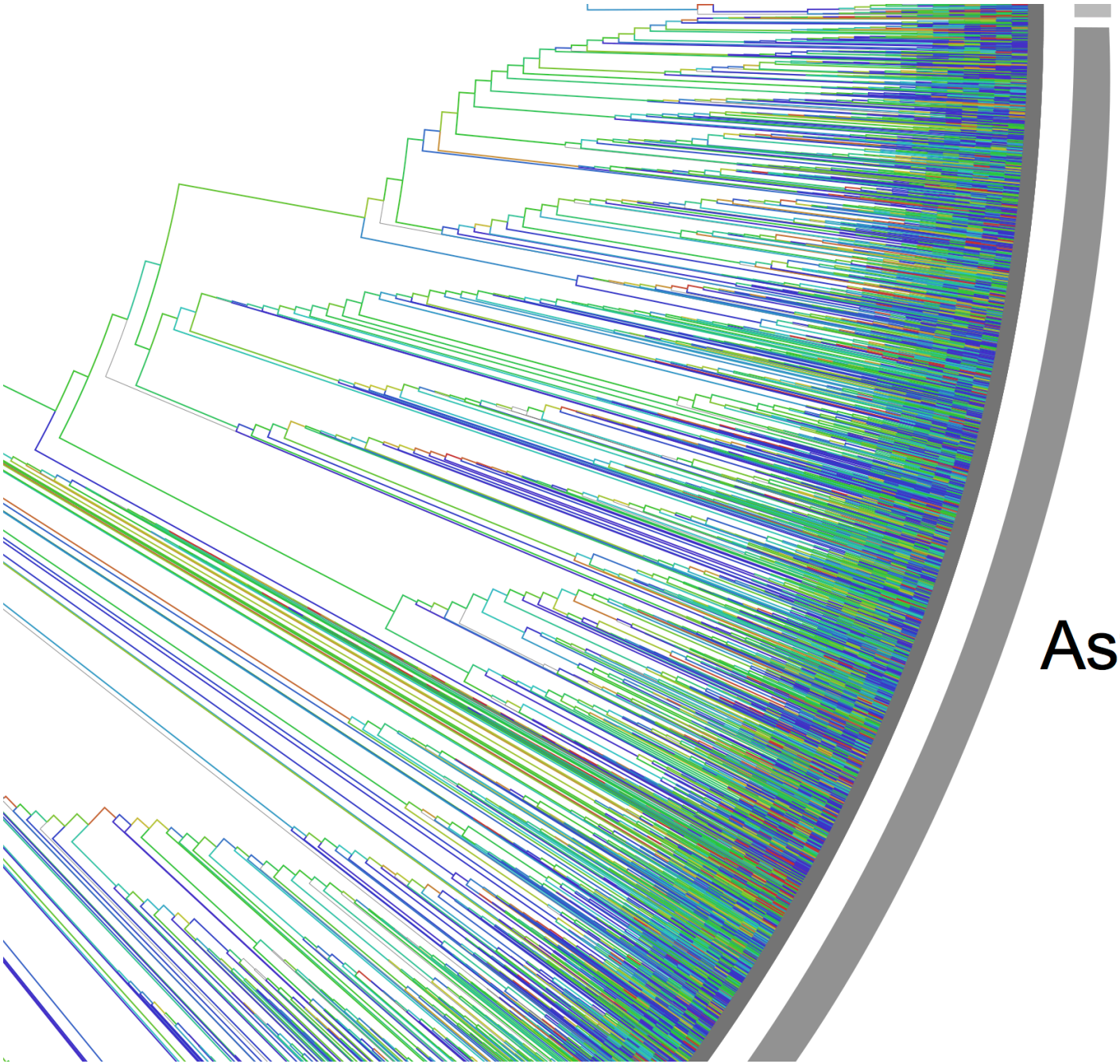
Pease et al. *in review*



Arecales support



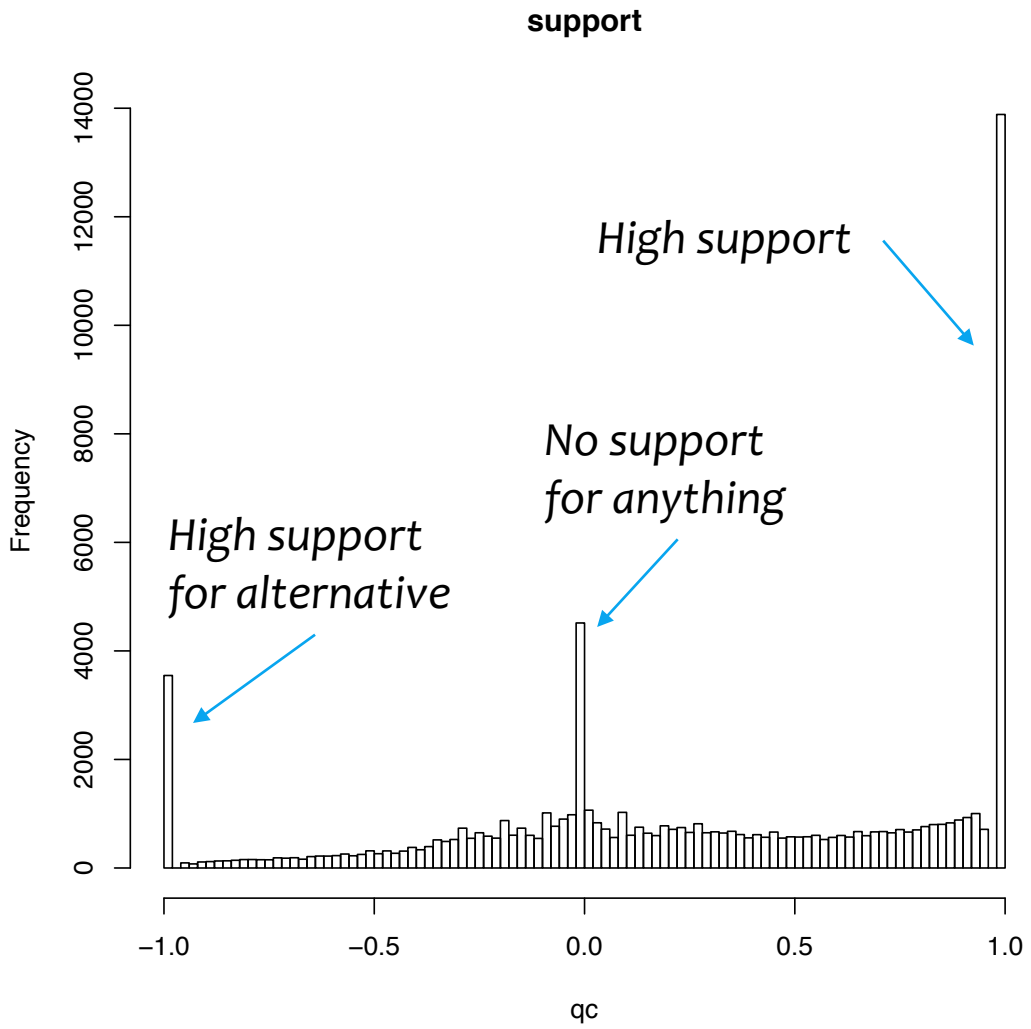

Ranunculales
Asparagales
Saxifragales
Myrtales
Poales
Brassicales
Alismatales, Liliales,
Arecales, Commelinales,
Zingiberales
Malpighiales
Magnoliidae
Acrogymnospermae
Rosales
Asterales
Fabales
Apiales
Caryophyllales
Lamiales
Ericales
Gentianales



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 misidentified
- It is difficult to get NCBI to correct these

No phylogenetic information

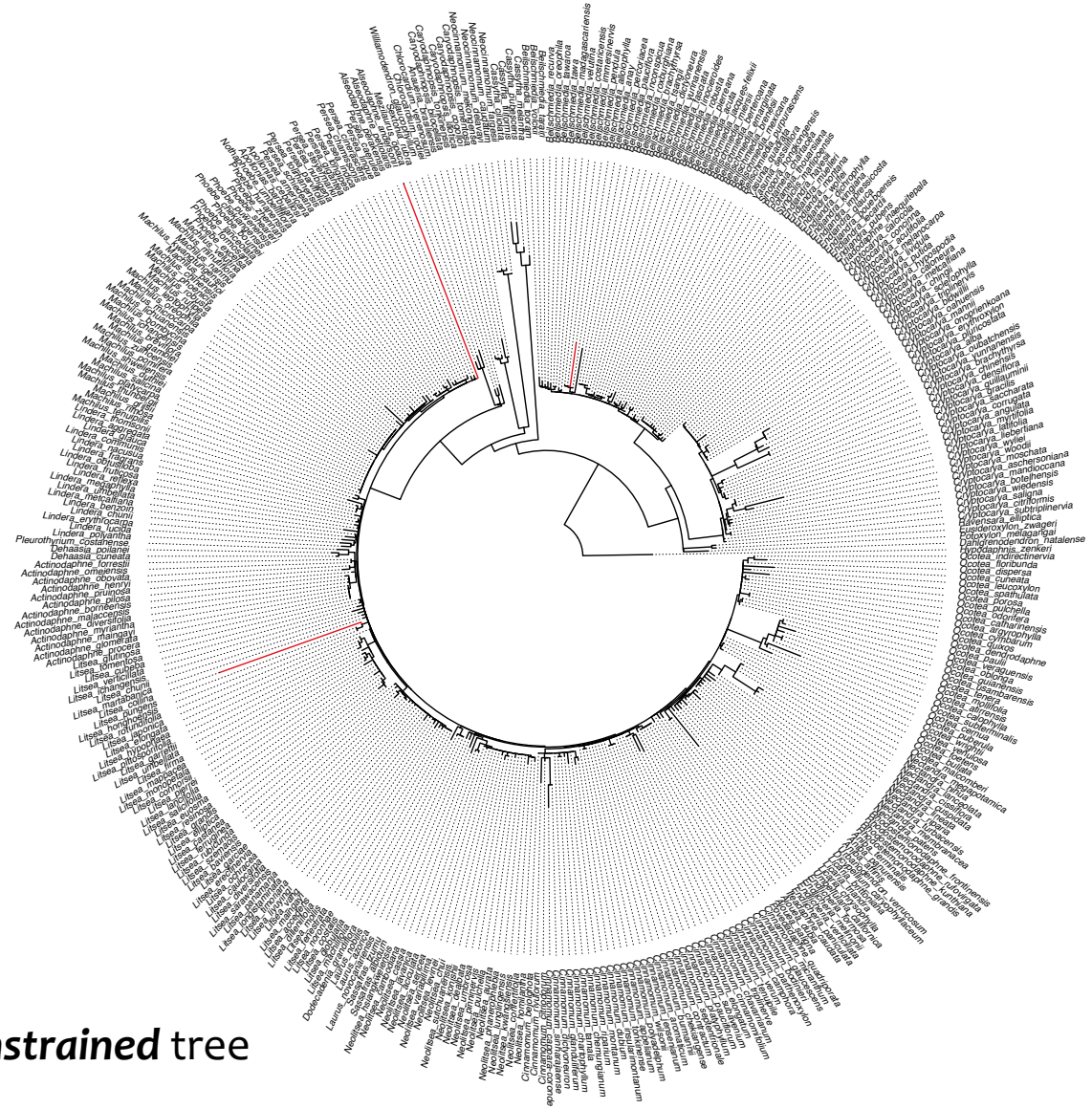
- Some of the sequences can be very short and not carry much if any phylogenetic information

The challenge of barcoding for phylogenetics

Outlying species

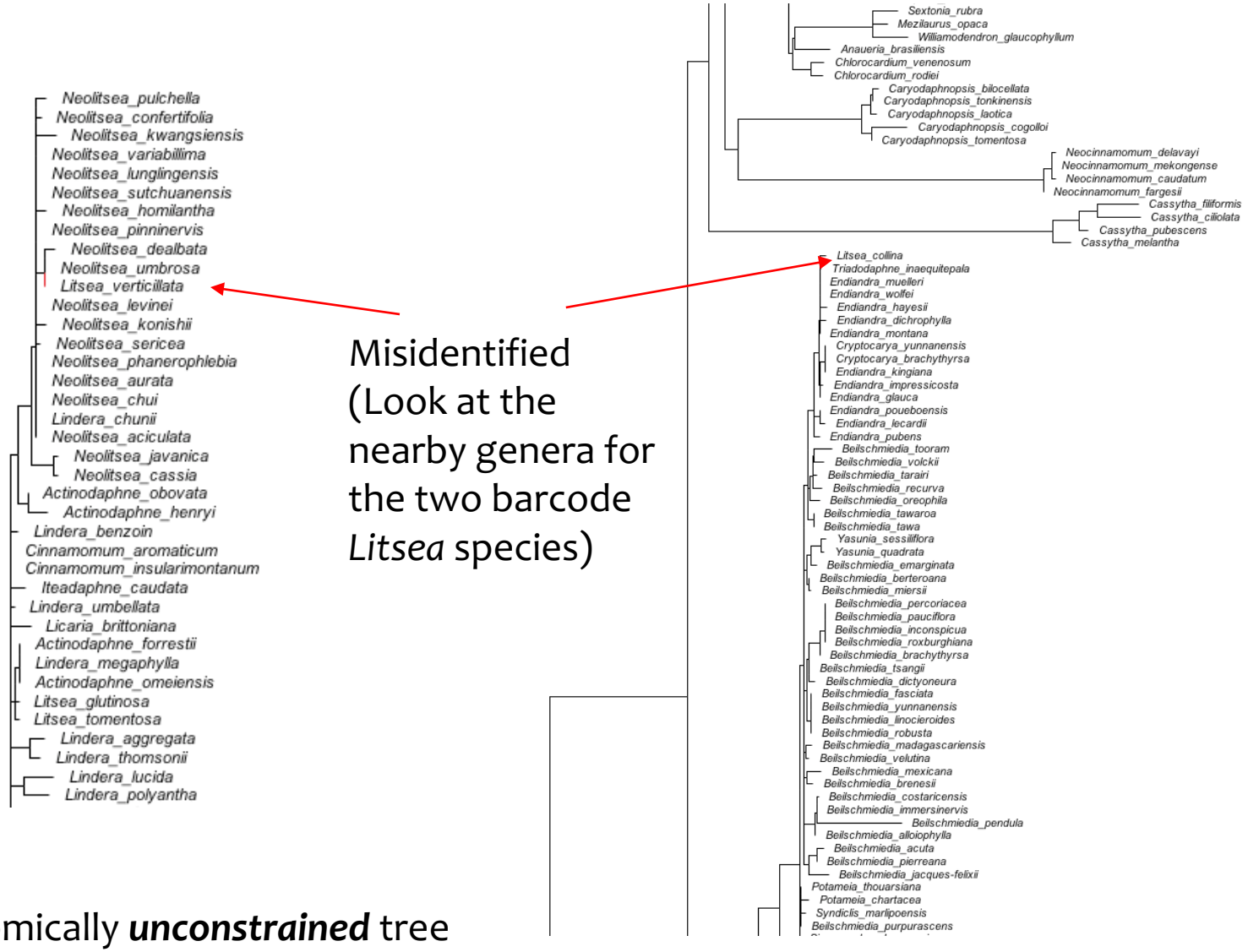
- *Litsea collina*
- *Litsea verticillata*
- *Alseodaphne andersonii*
- *Beilschmiedia pendula*

Clear pattern from constrained tree



Taxonomically ***constrained*** tree

The challenge of barcoding for phylogenetics

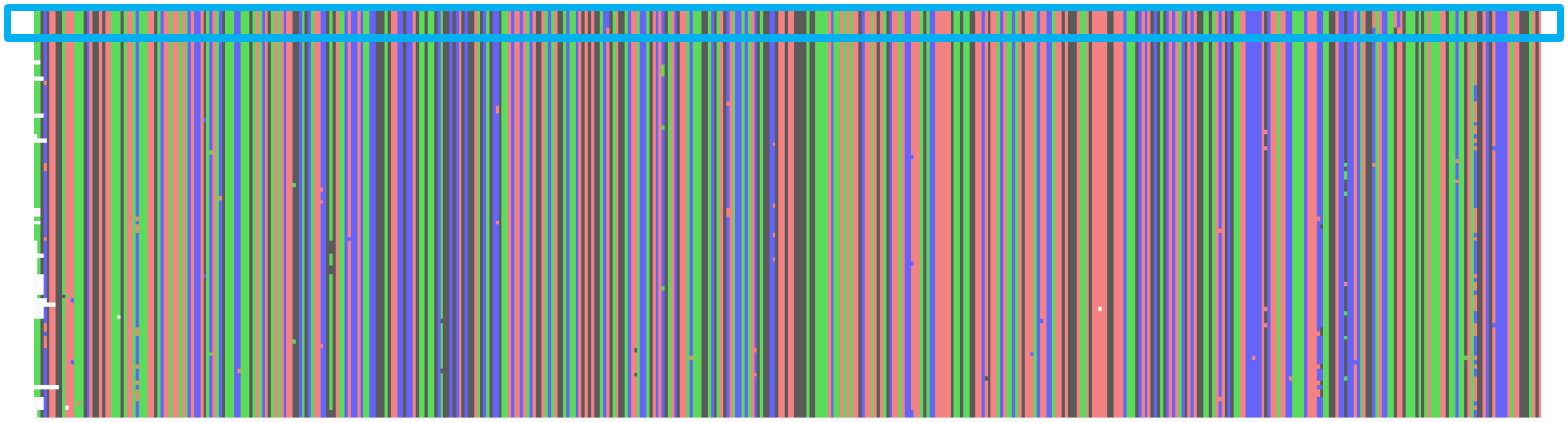


Misidentified (Look at the nearby genera for the two barcode *Litsea* species)

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

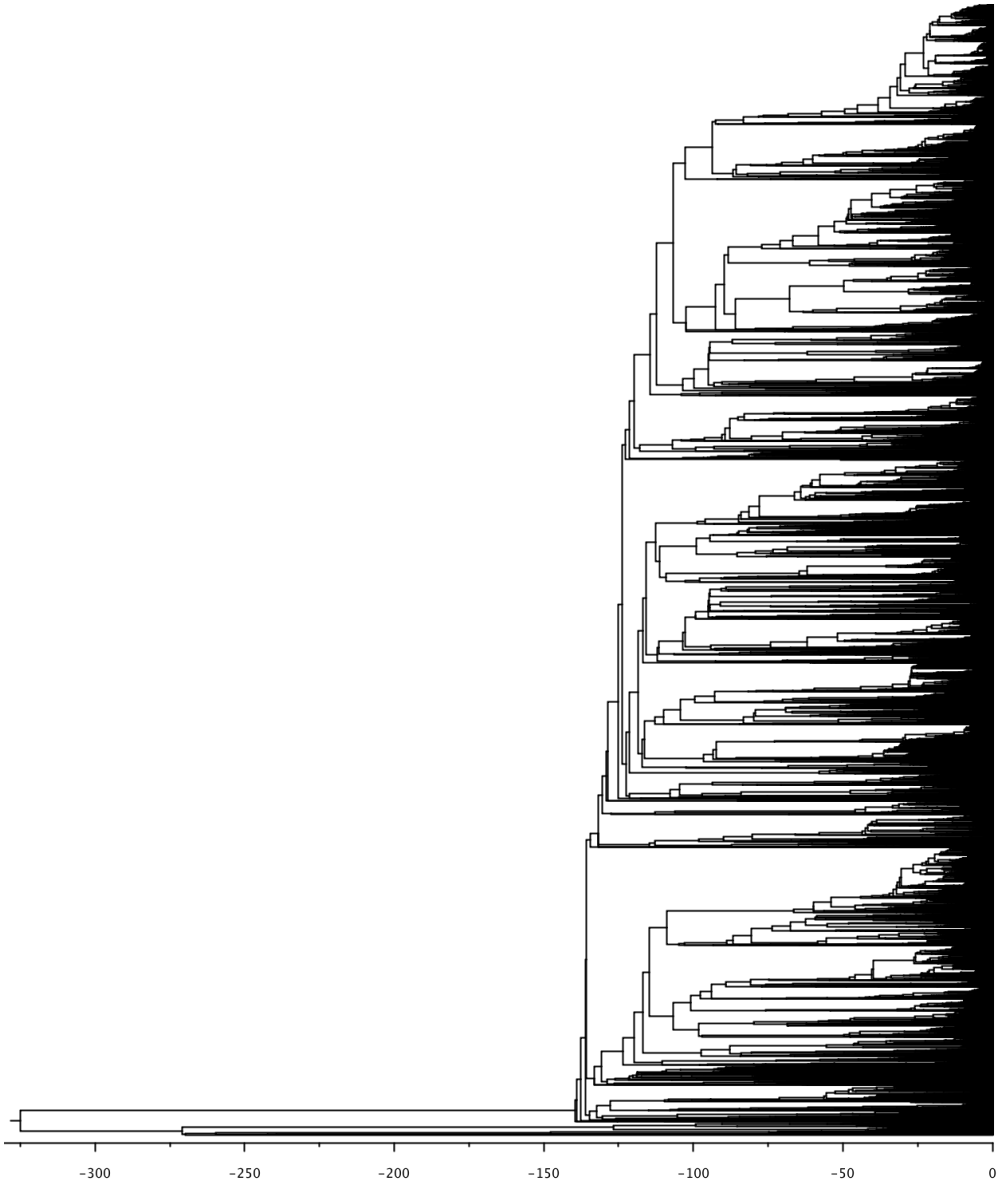
No phylogenetic information

- 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

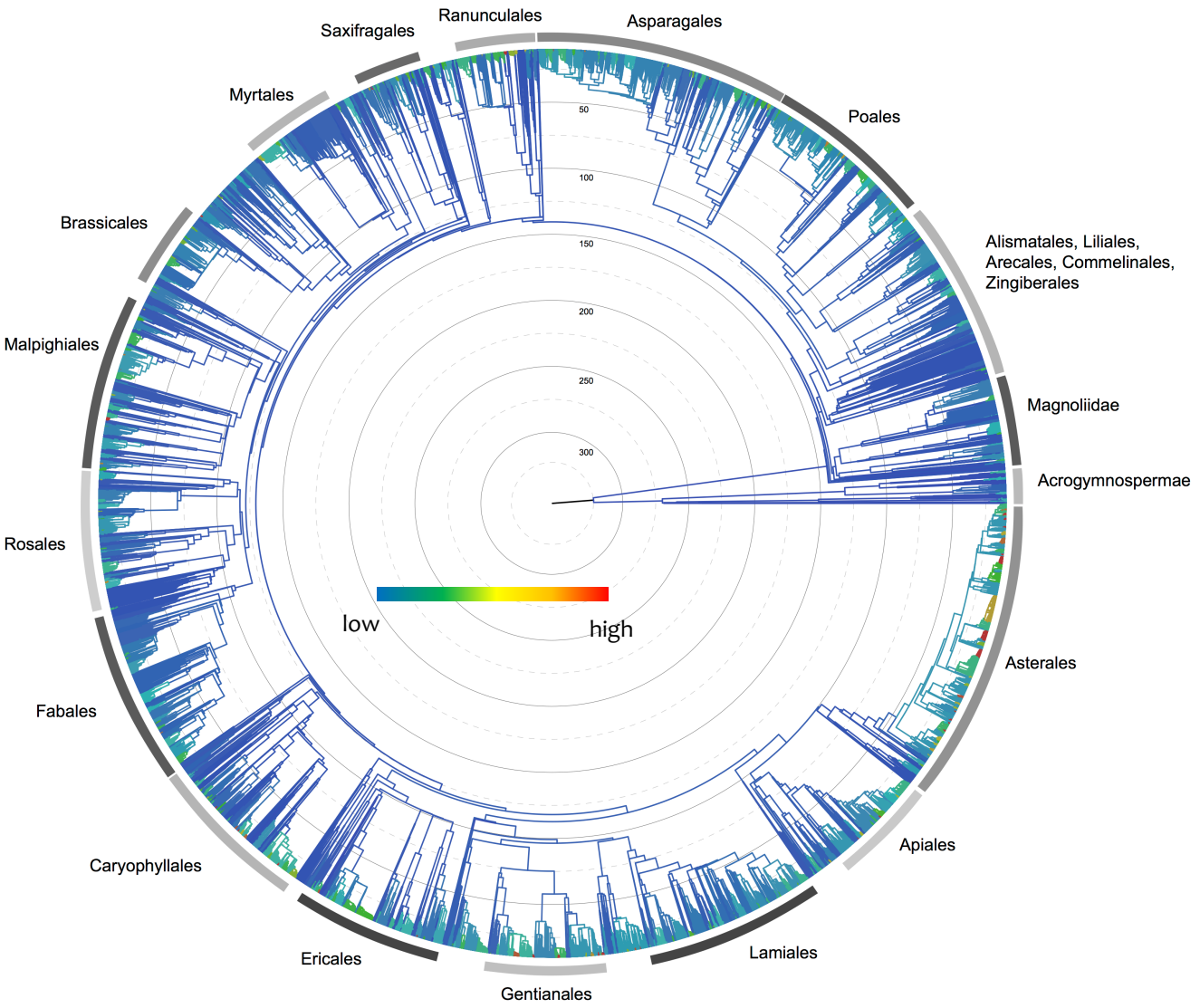


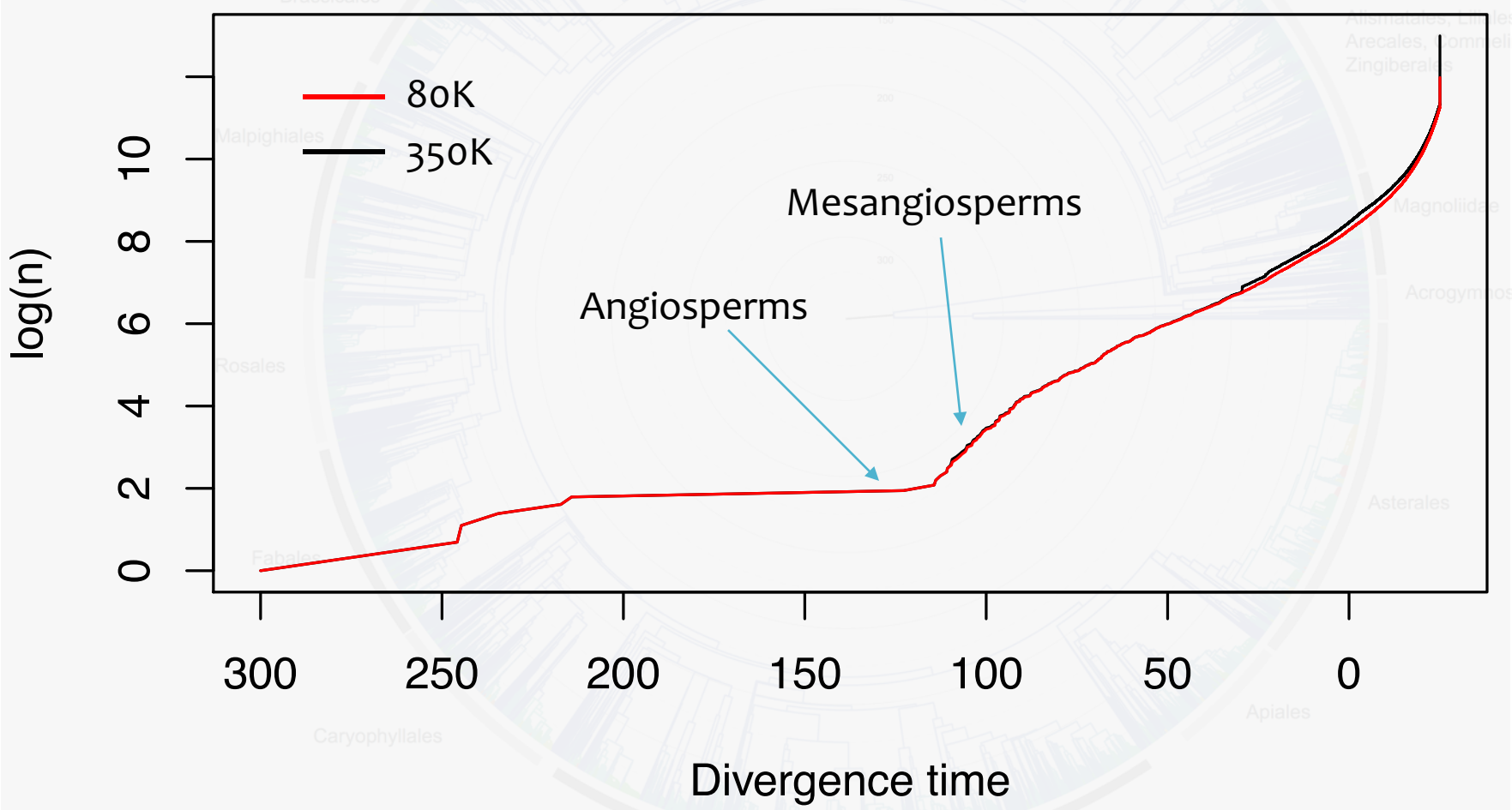
Rates of diversification

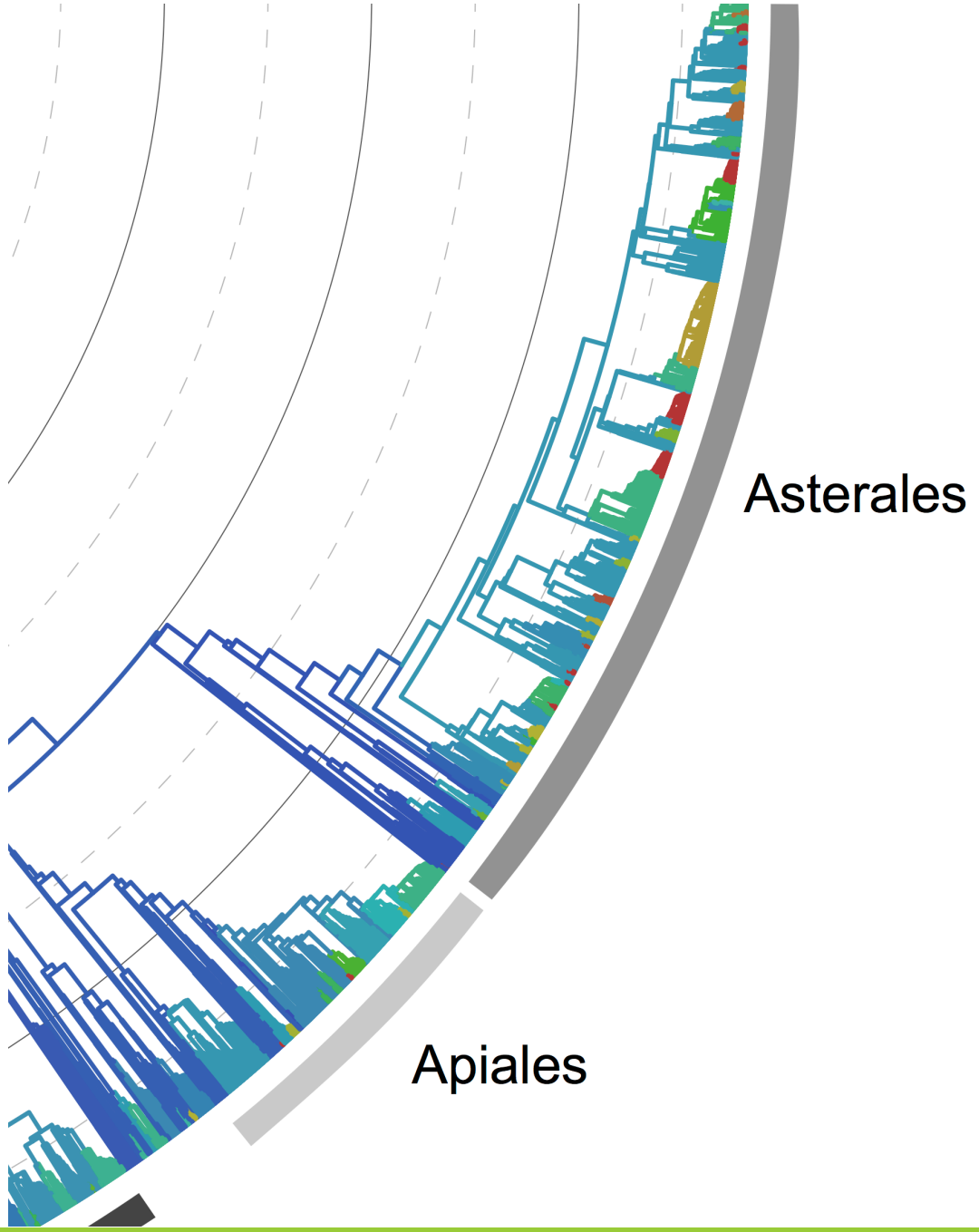
MEDUSA analyses of rates of evolution using birth-death models

Mean rate of evolution plotted

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

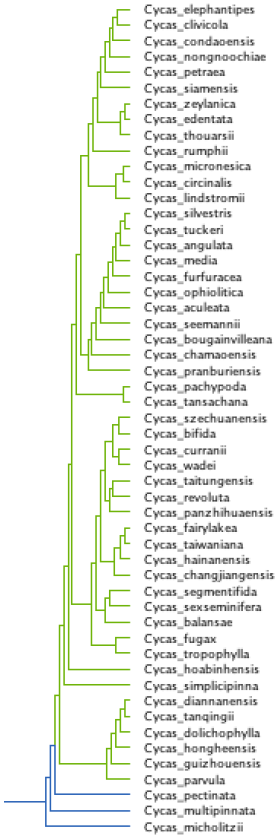


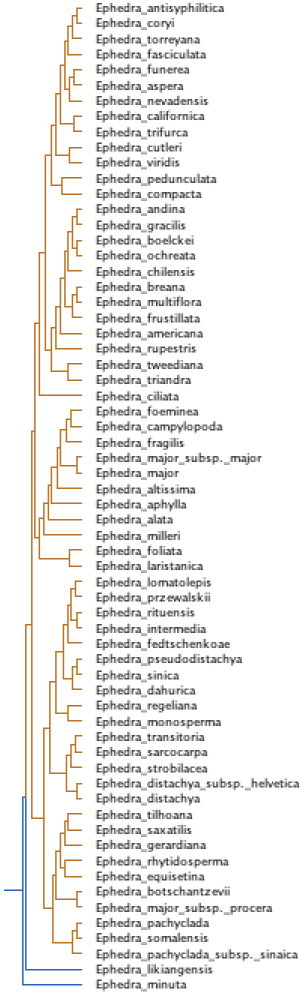

80K
350K
Mesangiosperms
Angiosperms
log(n)
0
2
4
6
8
10
300
250
200
150
100
50
0
Divergence time


Asterales
Apiales

Many repeated patterns

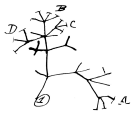
Many repeated patterns of depauperate lineages sister to very diverse lineages (Donoghue and Sanderson 2015)





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