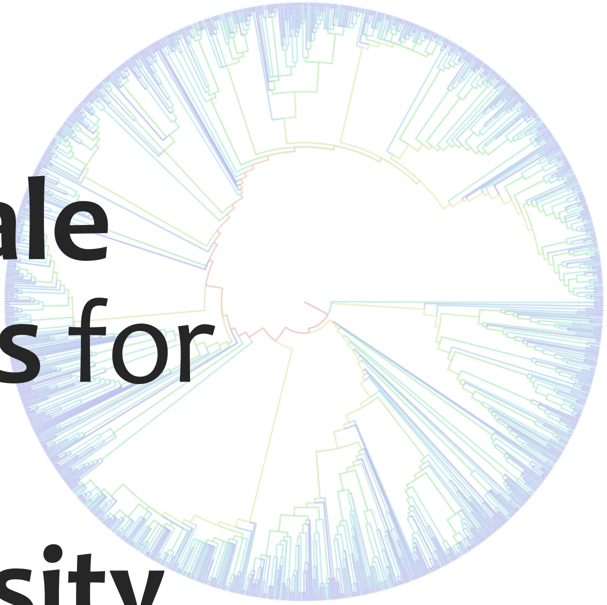
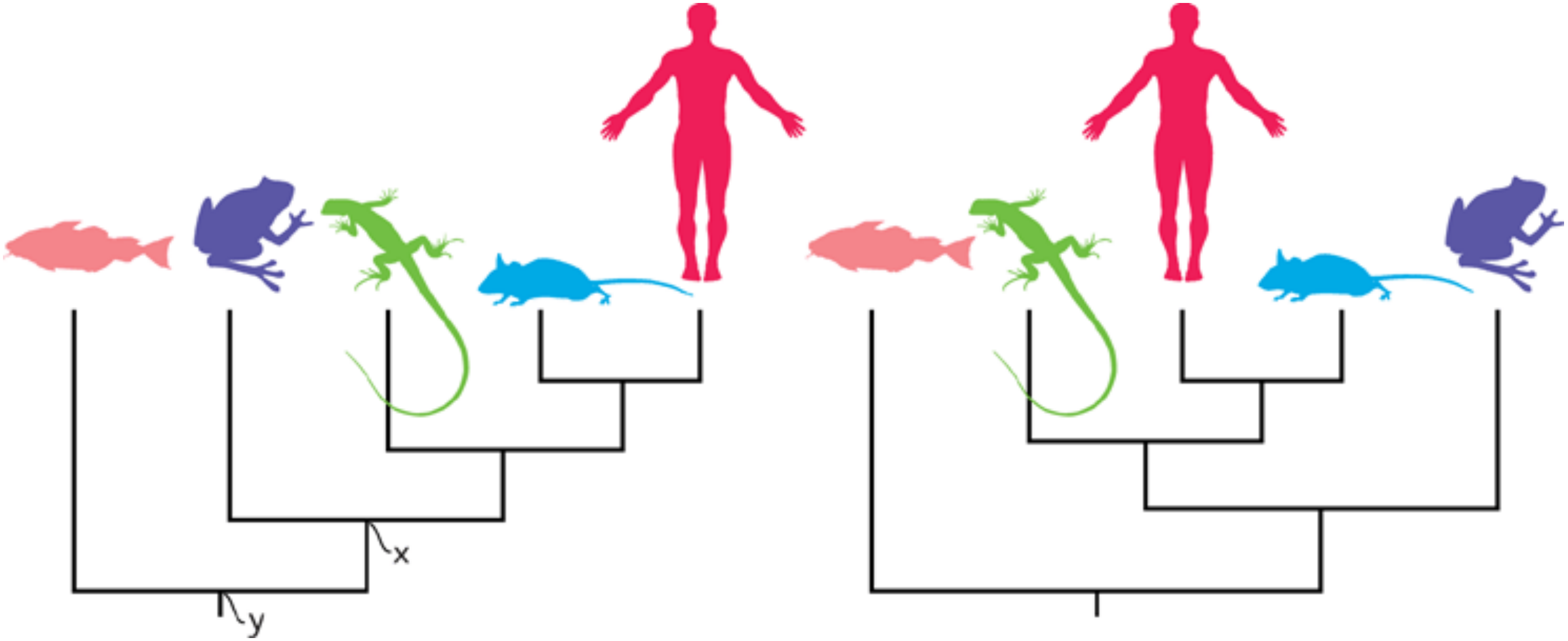


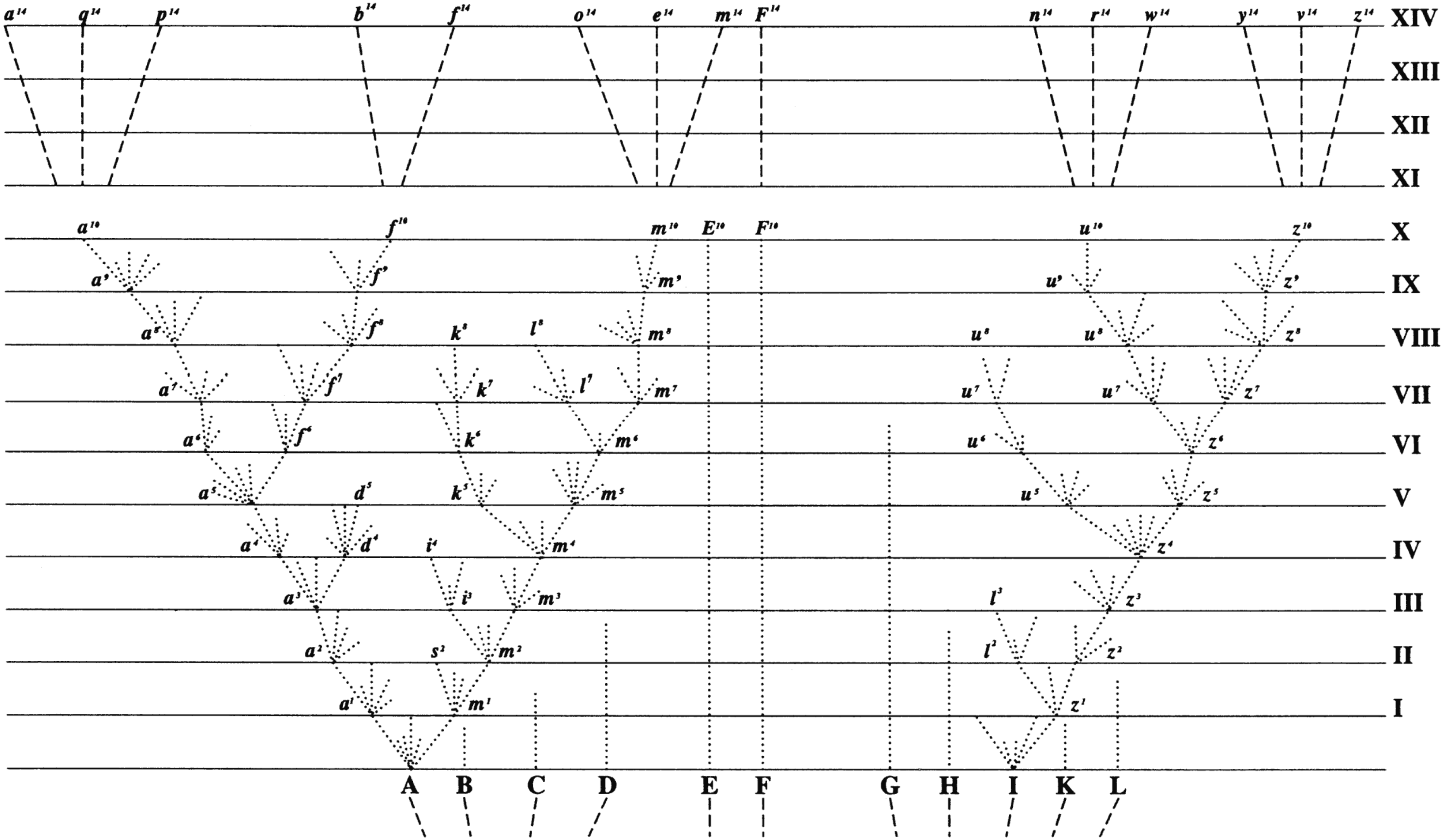
The utility of **large-scale phylogenetic analyses** for understanding the **evolution of biodiversity**

DR. STEPHEN A. SMITH

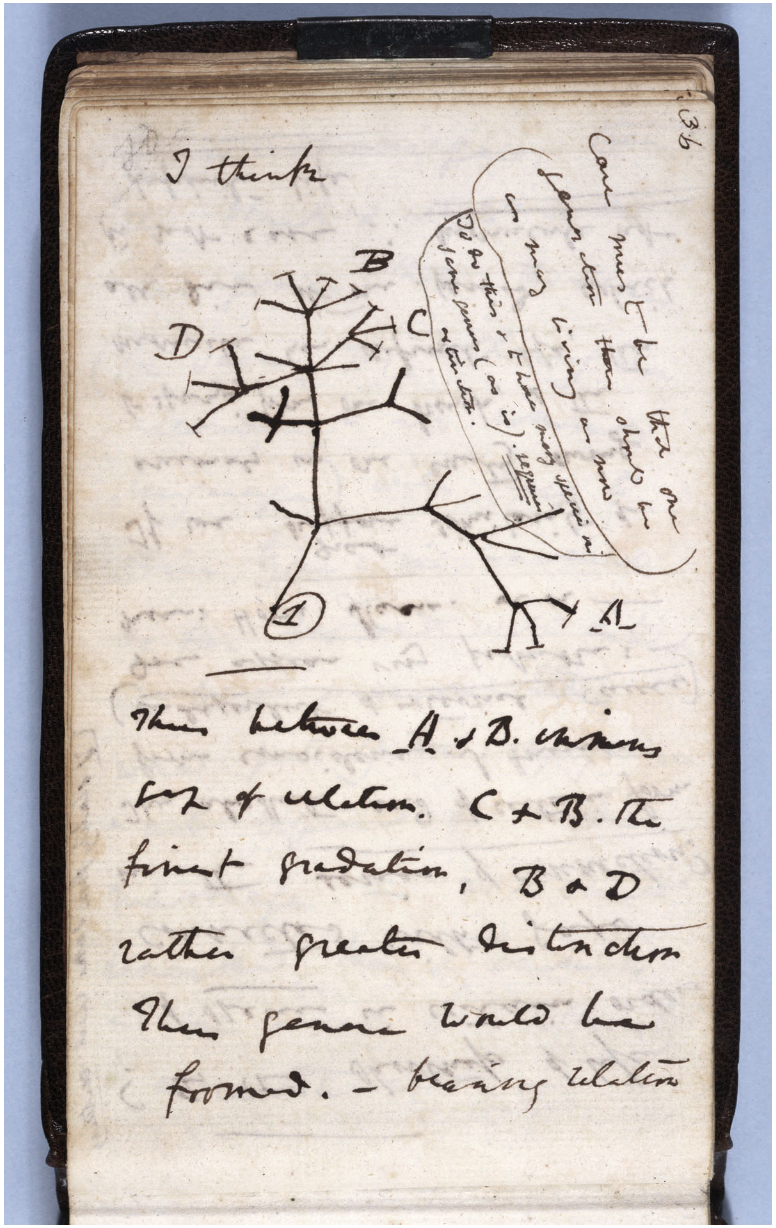
UNIVERSITY OF MICHIGAN

Phylogenetic tree: a visual representation of the relationship between species, taxa, individuals, or populations.





On the origin of species



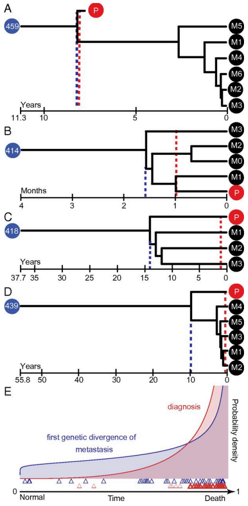
36

I think

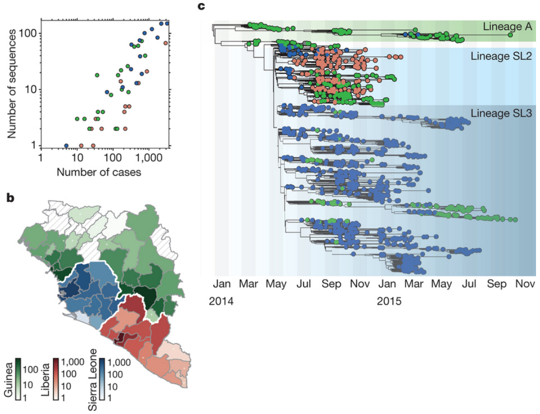
Case must be that one generation then should be as many living as now. To do this & to have many species in same genus (as is). requires extinction.

Thus between A & B. immens gap of relation. C & B. the finest gradation, B & D rather greater distinction Thus genera would be formed. — bearing relation

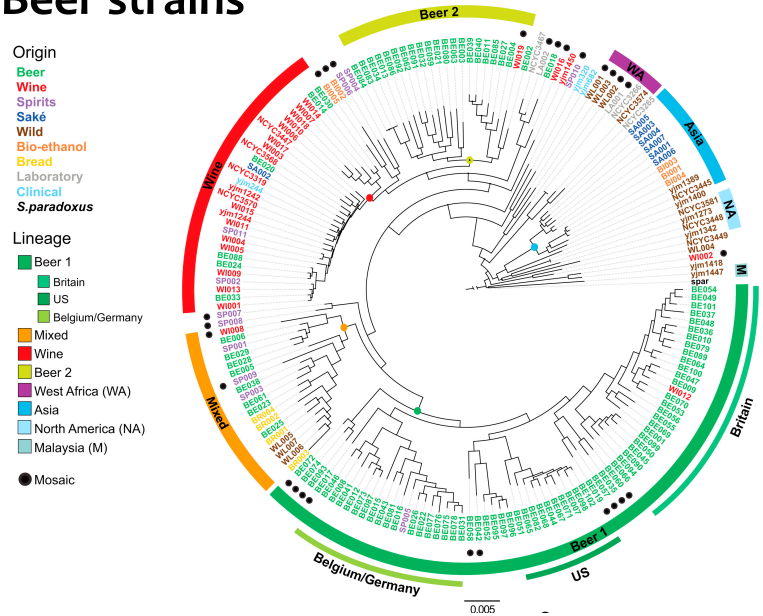
Cancer



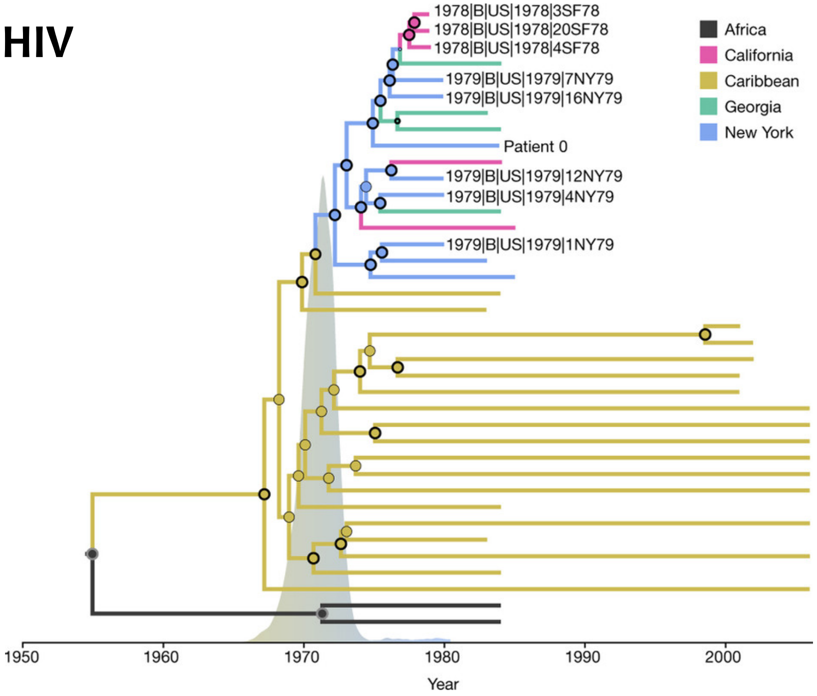
Ebola



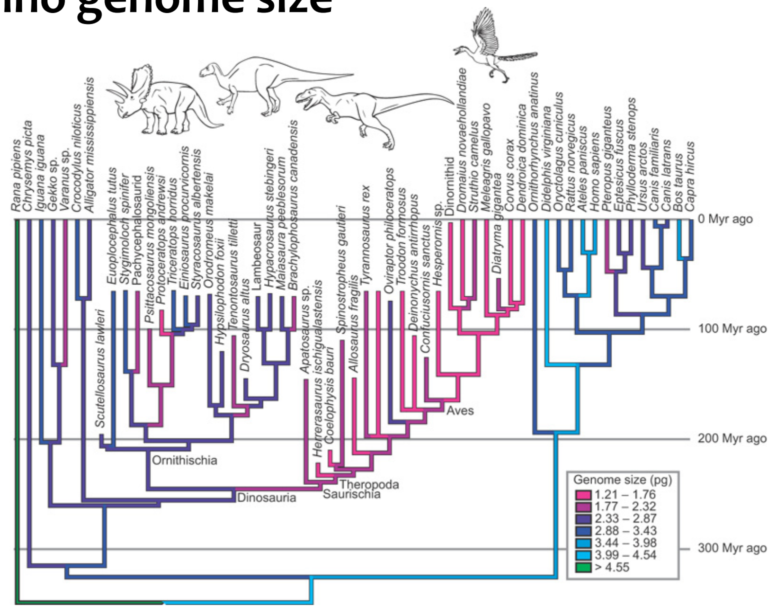
Beer strains



HIV



Dino genome size




WE ARE HERE

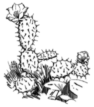
I use phylogenies to address:

Empirical

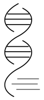

Biodiversity: The origin and evolution of biodiversity

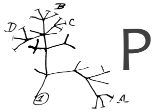

Biogeography: Why do species live where they live


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

“I would rate this as excellent 5 star service. I'll definitely be ordering phylogenies from you in the future :). Plus you offer up the biggest phylogenies with more dated nodes than anyone in the business!!! You need a Yelp page!”

Amy Zanne

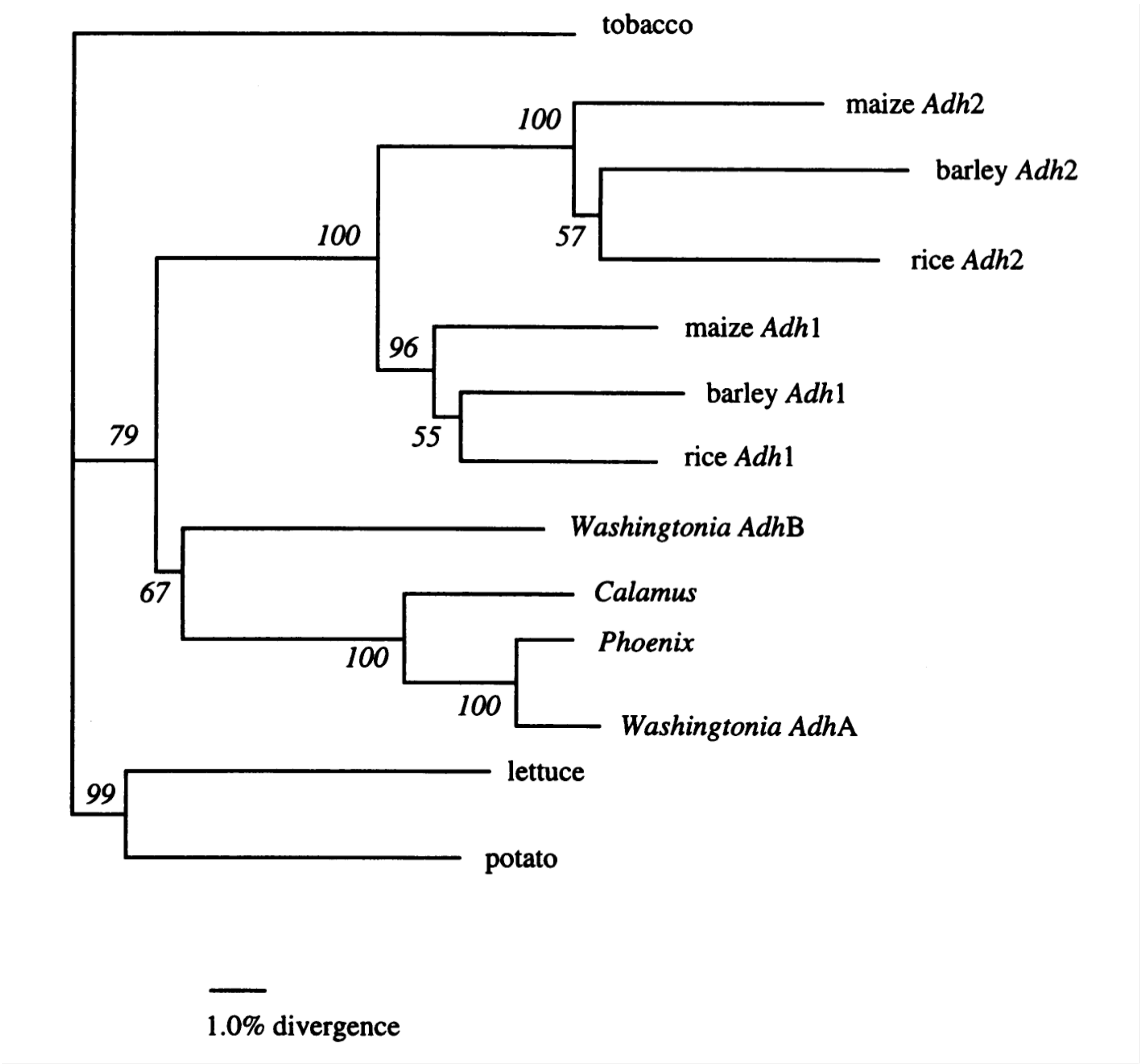
After returning a draft tree of plants in a record 5 minutes

Today

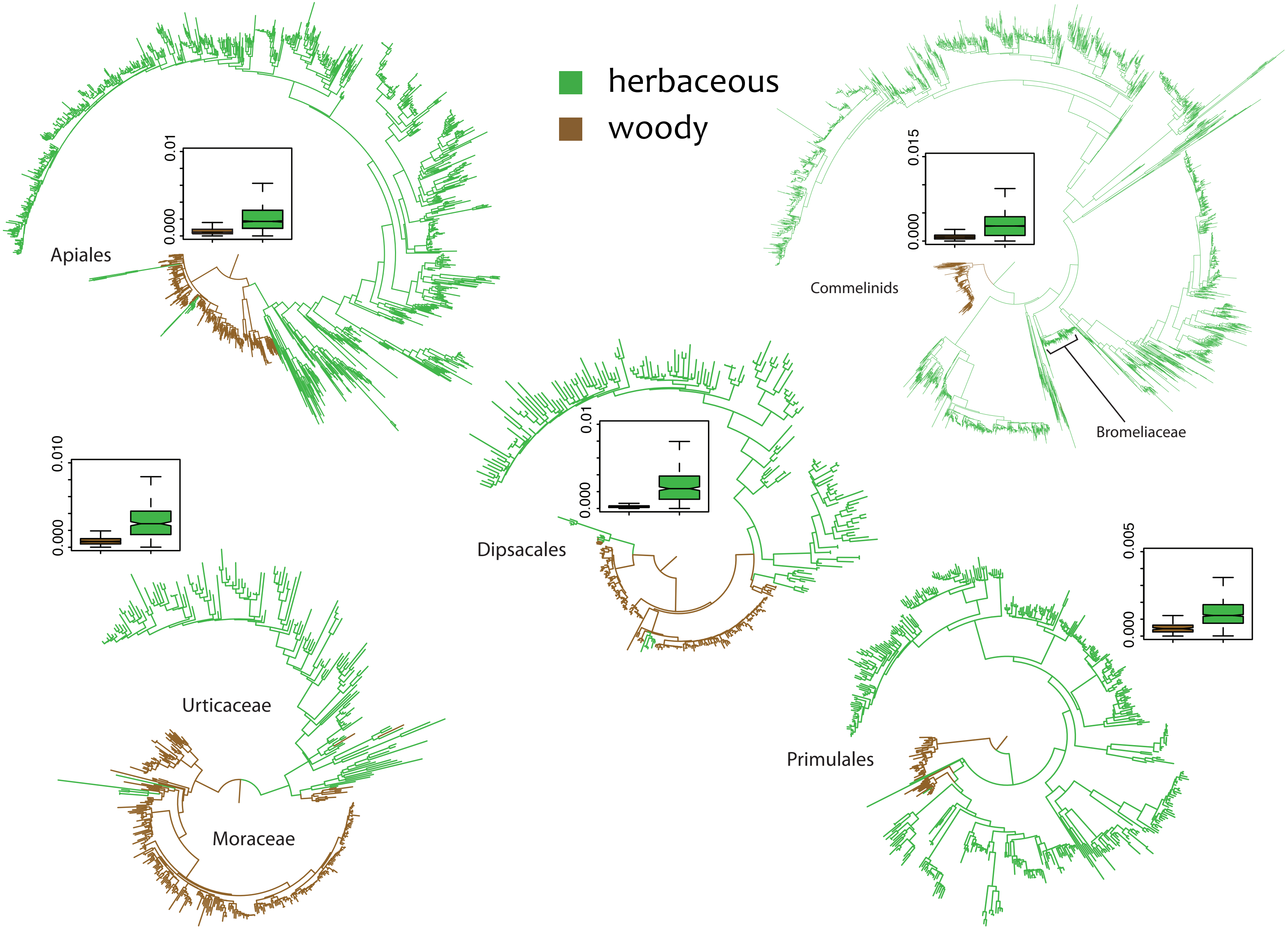
- Why build bigger trees?
- Building a better tree of life to address large scale questions about biodiversity
- A comprehensive dated seed plant phylogeny

If we build a *better* **tree of life**, can we learn more about **biodiversity**?

CAN WE CONSTRUCT A COMPLETE TREE OF LIFE?



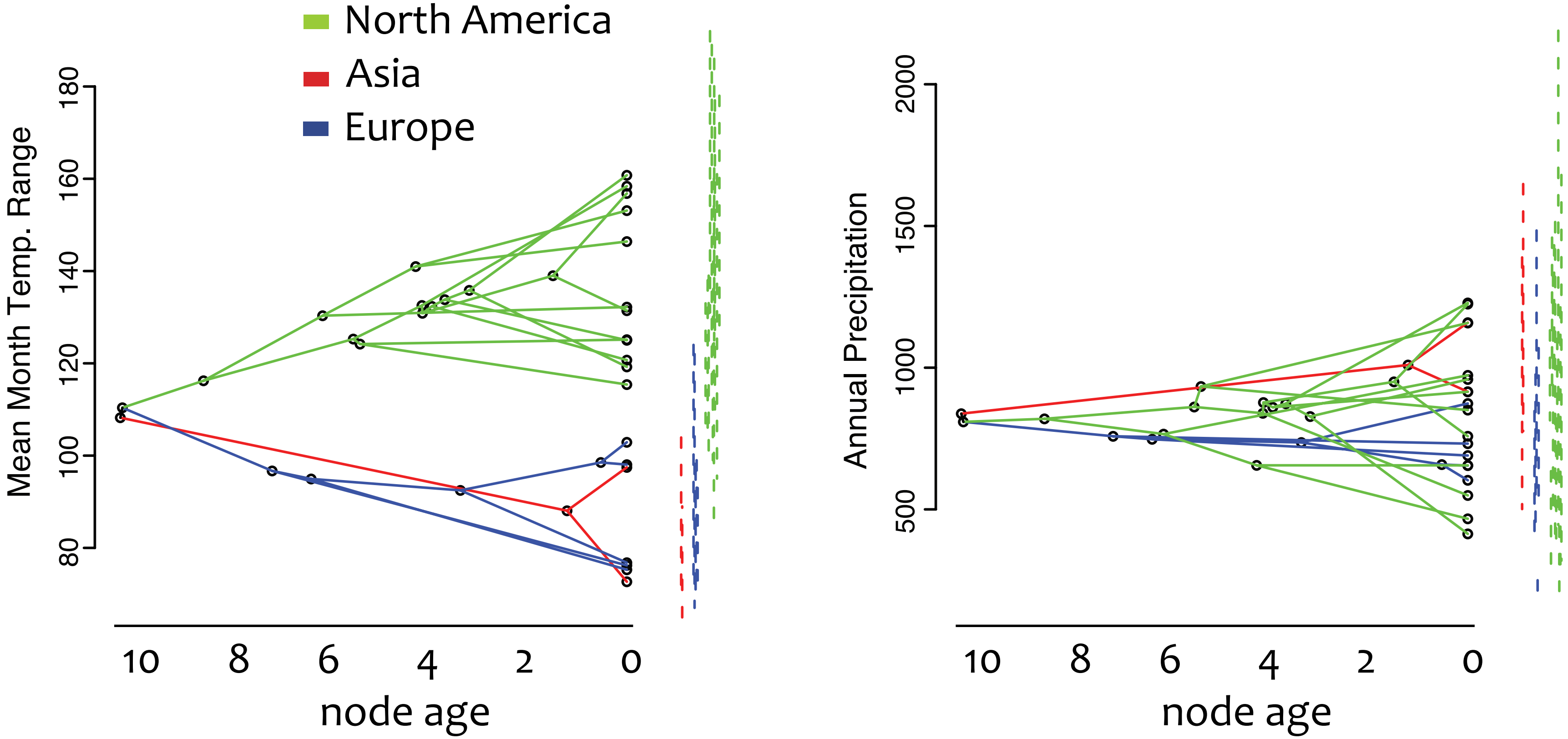
Gaut et al., 1996



Smith and Donoghue 2008

Lonicera (Honeysuckles)

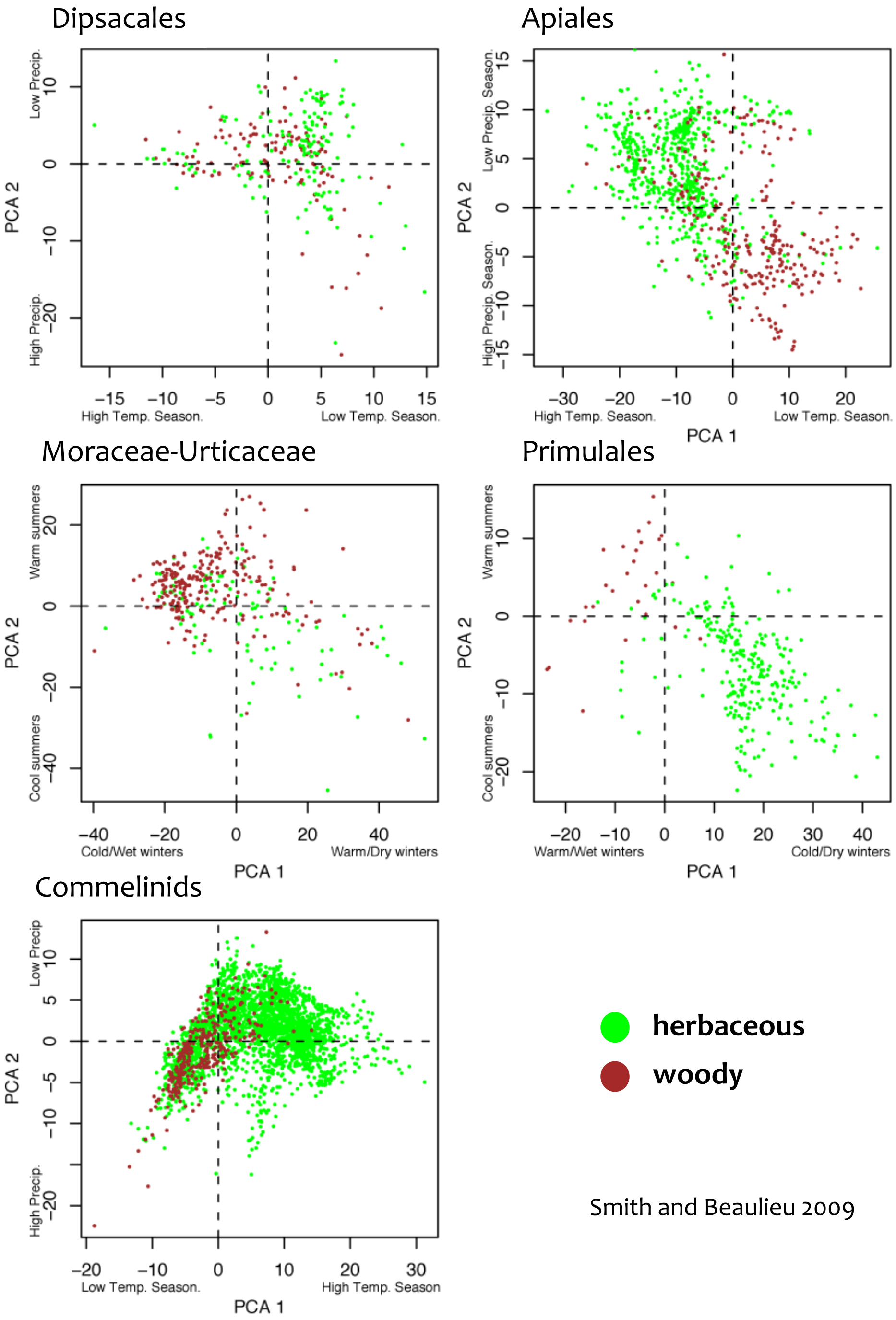
Lonicera climatic niche reconstruction



Smith and Donoghue 2010

Rates of niche evolution

- Phylogenetically corrected PCA
- Herbaceous plants occupy ***larger*** climatic niche space
- Woody plants occupy ***smaller*** climatic niche space
- Often occupy different spaces

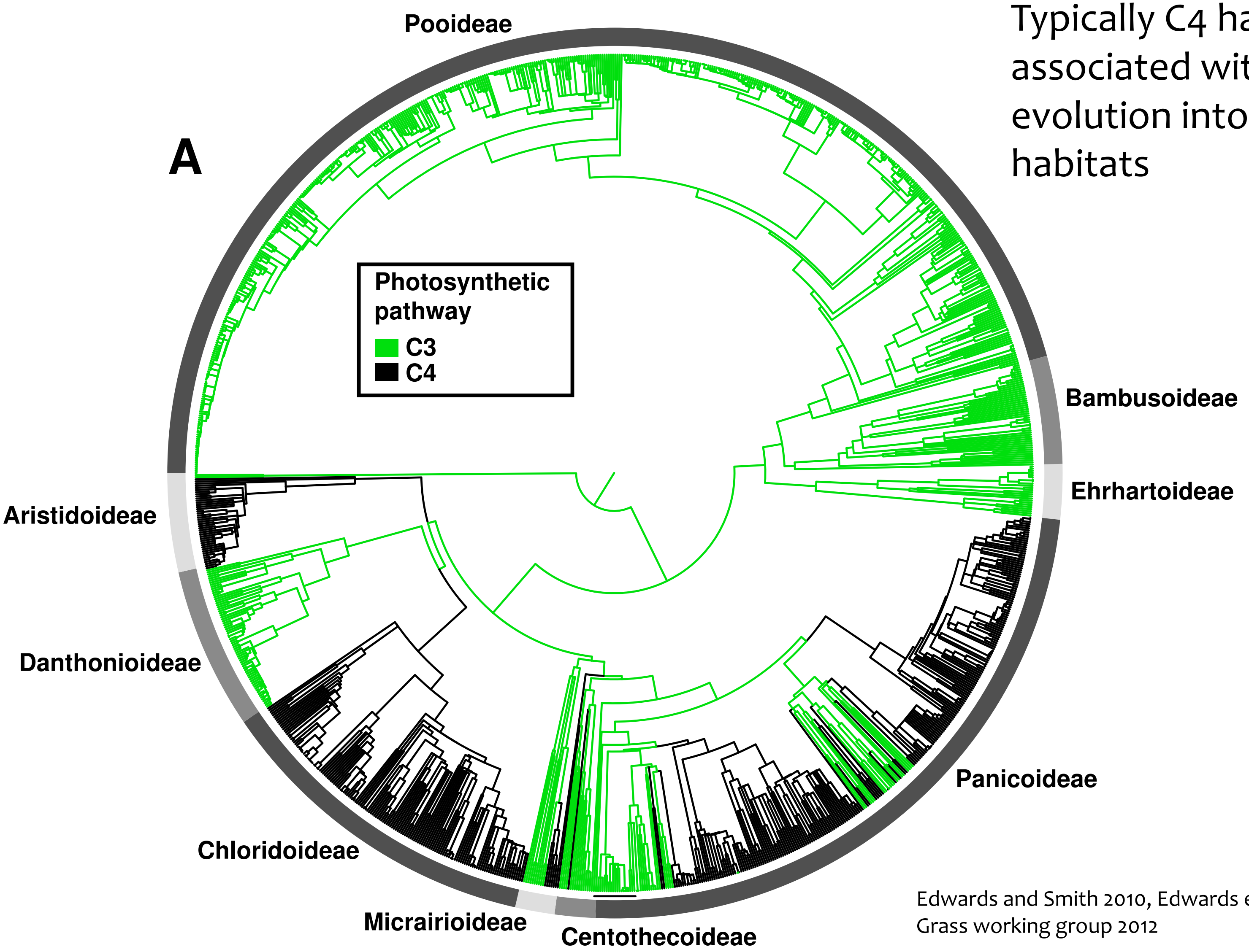


Smith and Beaulieu 2009

Grasses

- 5th most diverse angiosperm “family”
- over 800 genera and more than 10,000 species
- C3 and C4 photosynthetic pathways

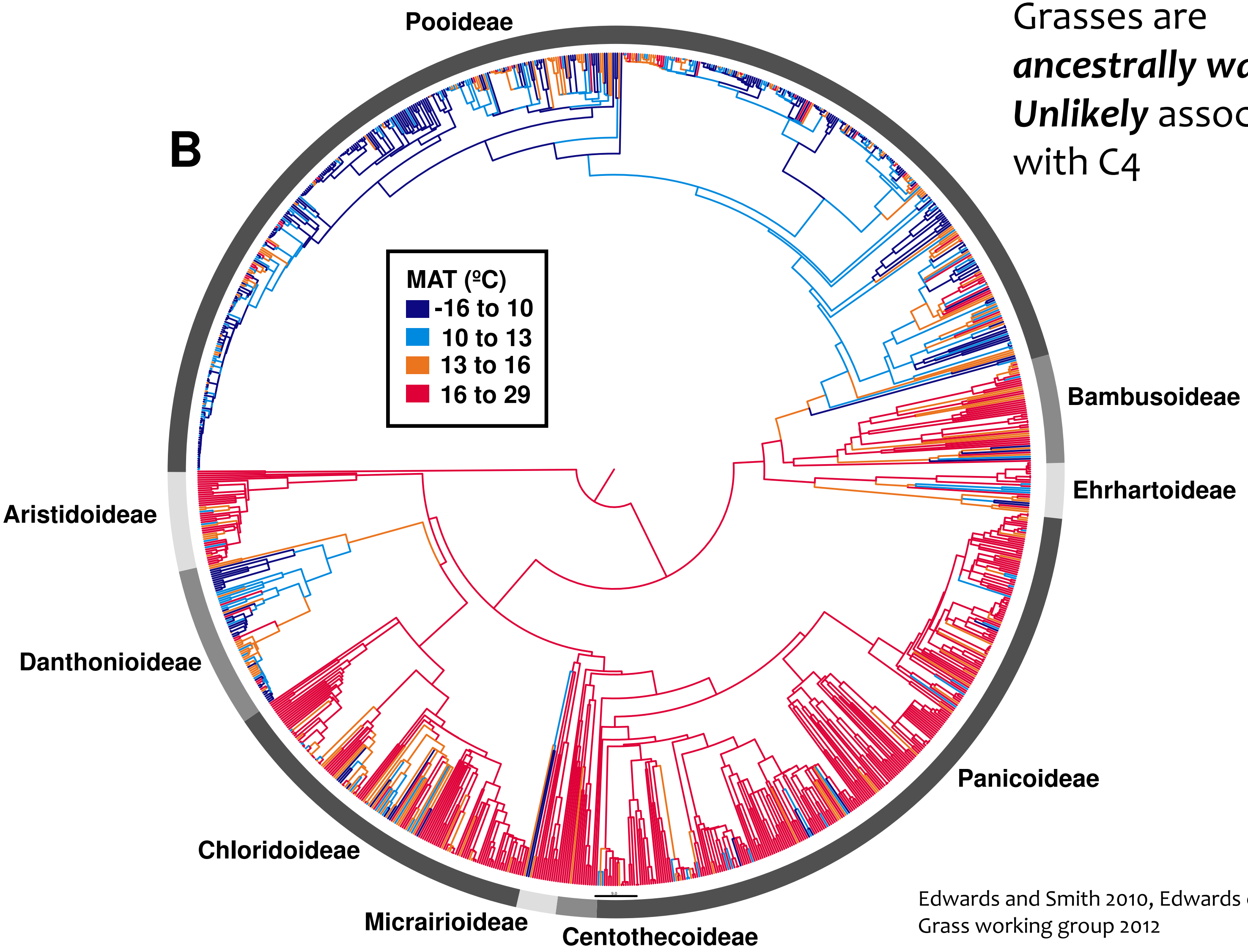
What about C4?



Typically C4 had been associated with evolution into ***warm*** habitats

Edwards and Smith 2010, Edwards et al. 2010, Grass working group 2012

What about climate (mean annual temperature)?



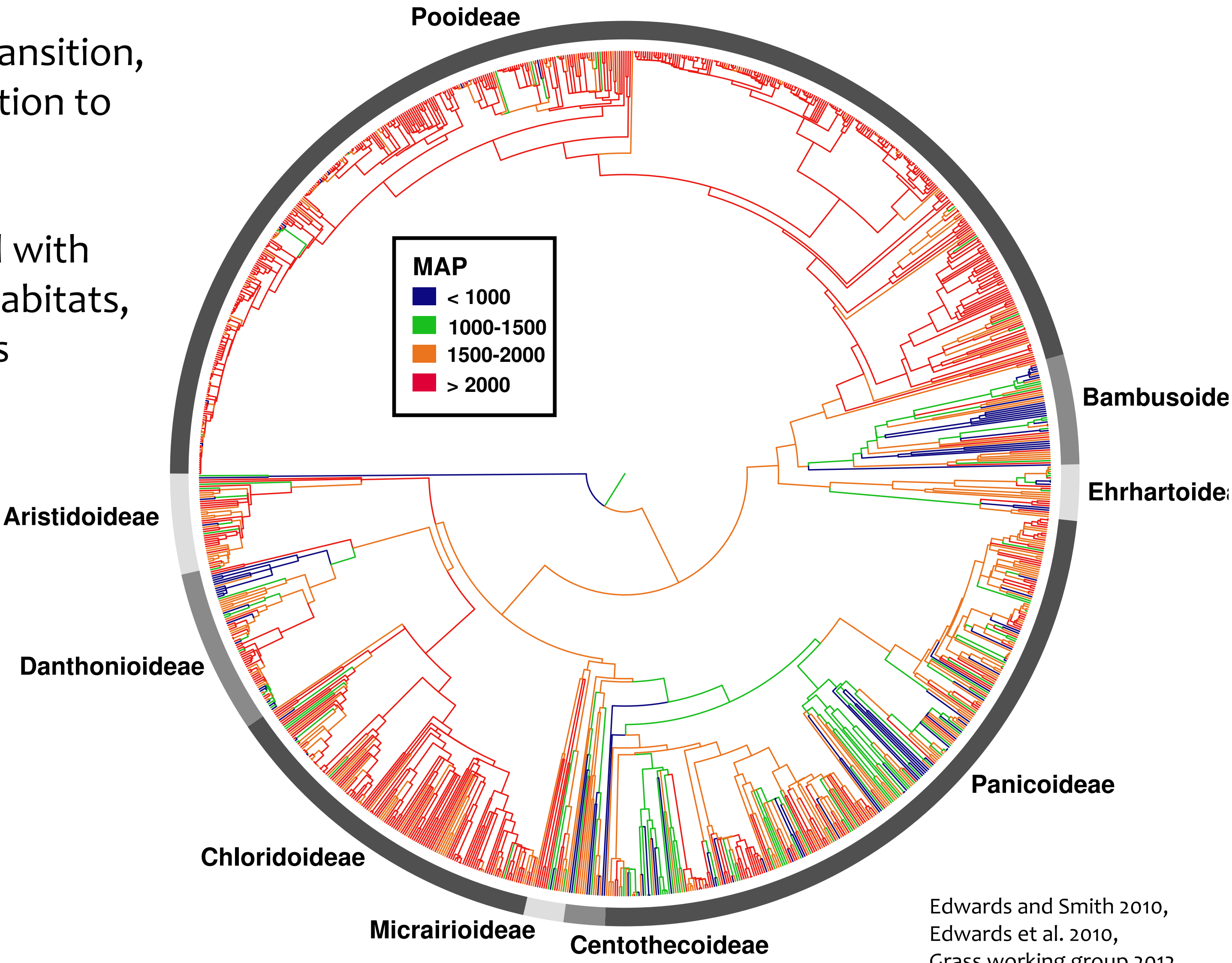
Grasses are ***ancestrally warm.*** ***Unlikely*** associated with C4

Edwards and Smith 2010, Edwards et al. 2010, Grass working group 2012

What about climate (mean annual precipitation)?

With each C4 transition, there is a transition to ***dry***

C4 is associated with shifts into ***dry*** habitats, **not *hot*** habitats



Edwards and Smith 2010,
Edwards et al. 2010,
Grass working group 2012

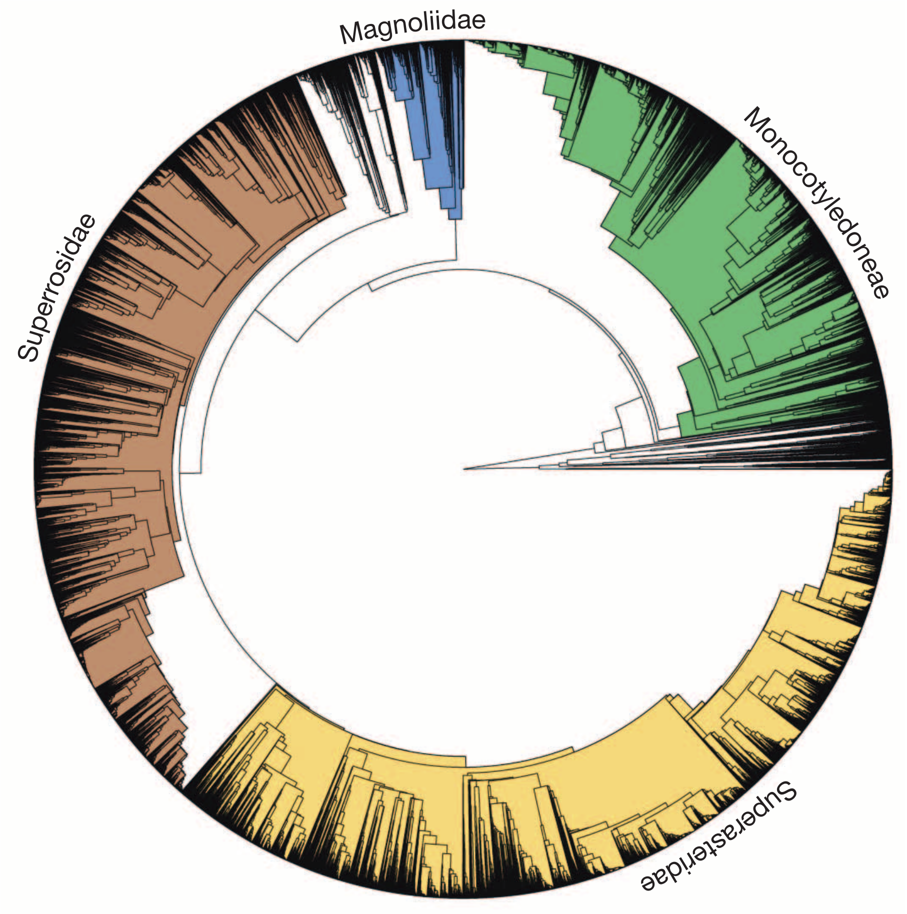
Did temperate plants adapt before or after they moved?

Tropical is likely ancestral for flowering plants

Temperate plants have many adaptations to colder and seasonal climates

- Deciduous
- Herbaceous
- Vessel size

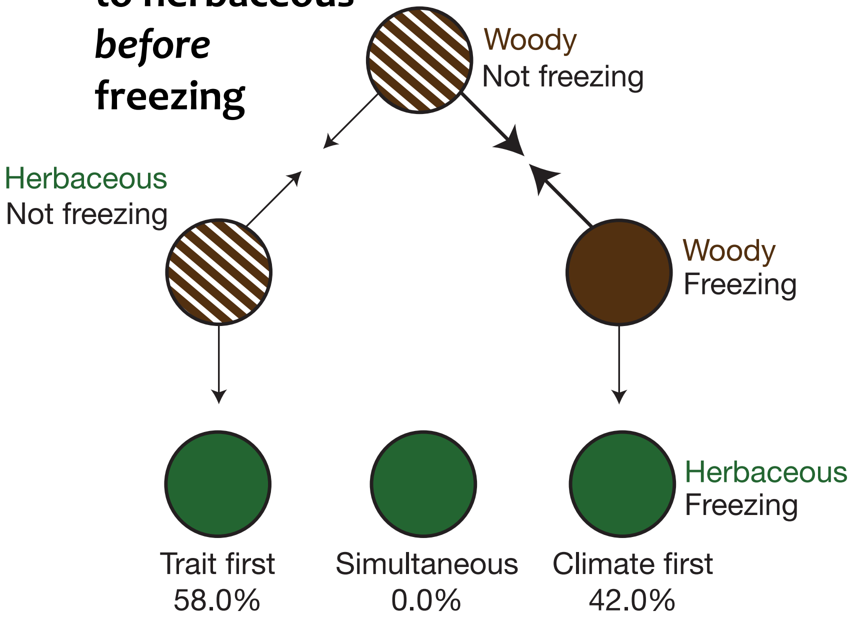
Did these adaptations occur *before* or *after* movement into temperate climates



Zanne et al. 2013, Zanne et al. 2014

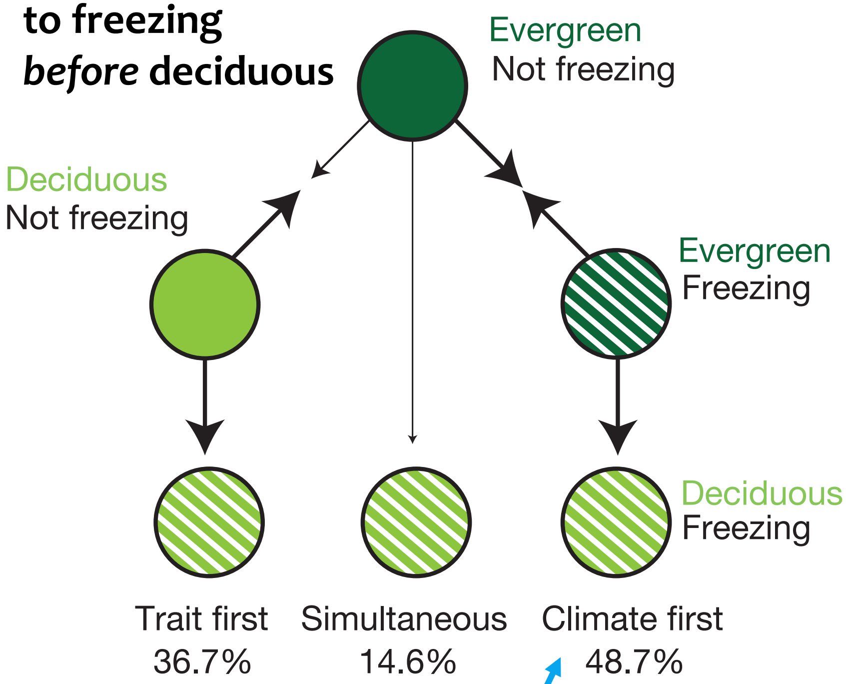
Did temperate plants adapt before or after they moved?

From woody to herbaceous *before* freezing



Herbaceous *before* moving

From evergreen to freezing *before* deciduous



Deciduous *after* moving

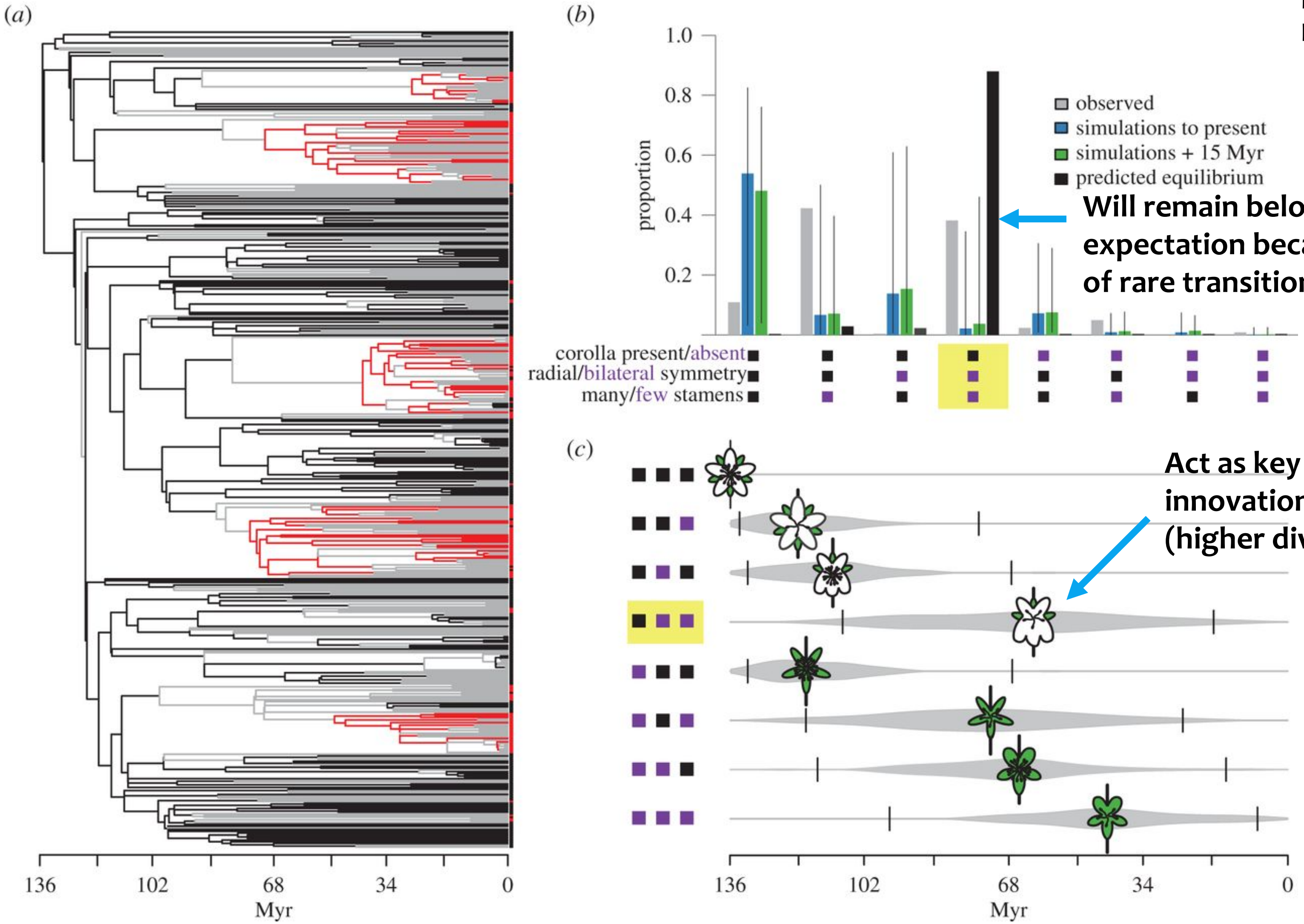
Zanne et al. 2013, Zanne et al. 2014

What flower types increase diversification?

Are there particular flower combinations that contribute to higher speciation?

- Corolla present vs corolla absent?
- Bilateral symmetry vs radial symmetry
- Stamens many vs stamens few

Evolution of biodiversity and flowers

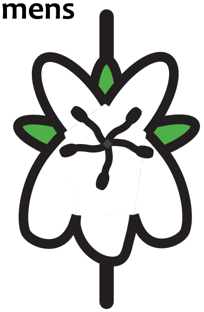


Will remain below expectation because of rare transitions

Act as key innovations (higher diversity)

O’meara et al. 2016

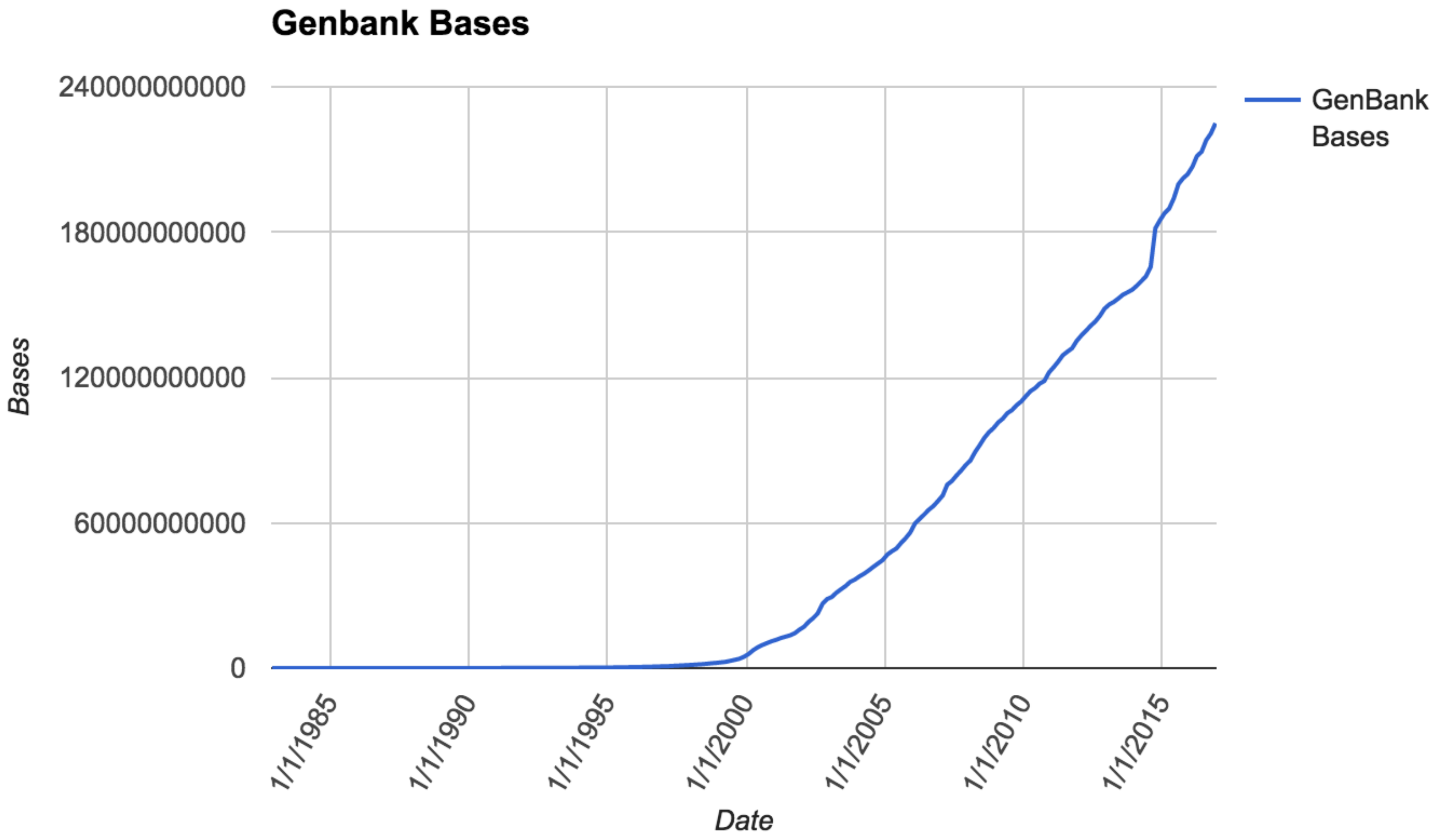
Corolla present
Bilateral symmetry
Few stamens



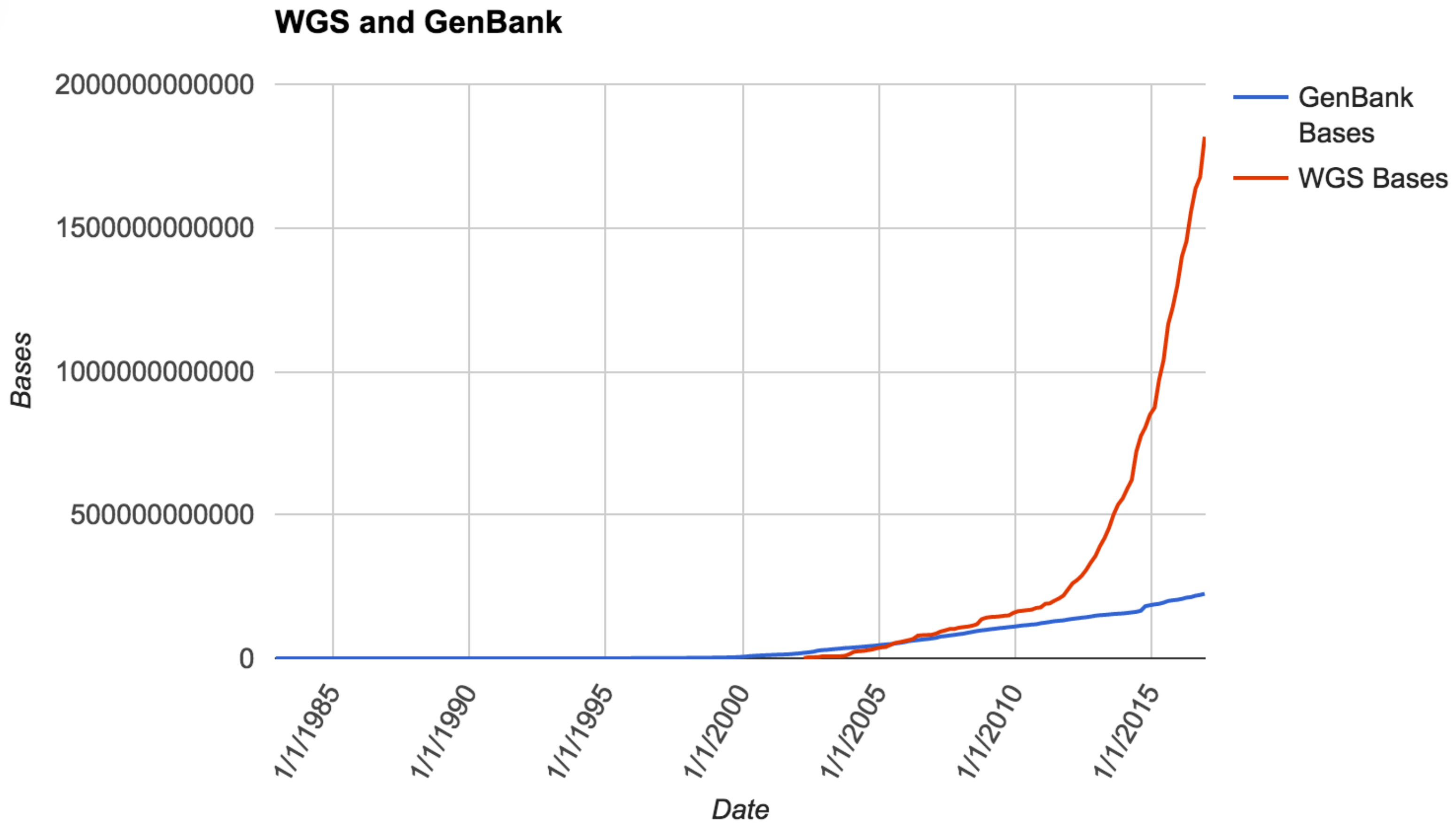
Transcriptomes for evolution and ecology

WHAT ARE THE PROCESSES THAT UNDERLIE WHAT WE SEE IN MOLECULAR DATA?

Molecular data availability



Molecular data availability (with whole genomes)



New data

Typical phylogenetic analyses

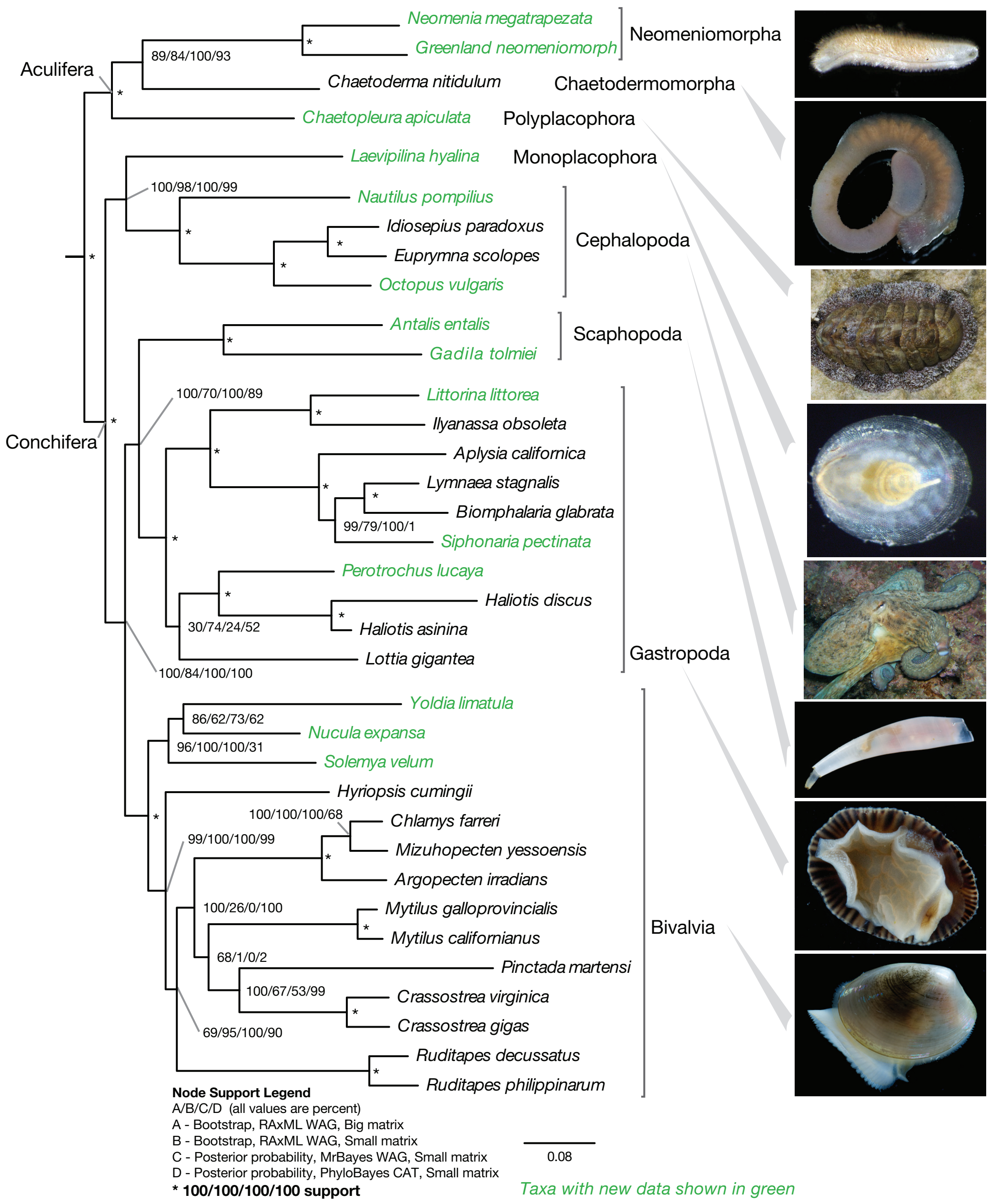
- 1-10 genes
- 17 genes. Plants (Soltis et al. 2011)
- 19 genes. Birds (Hackett et al. 2007)

Transcriptomic and genomic phylogenetic analyses

- 140 genes. Metazoa (Dunn et al. 2008)
- 242 genes. Metazoa (Ryan et al. 2013)
- 248 genes. Turtles (Chiari et al., 2012)
- 1185 genes. Molluscs (Smith et al. 2011)
- 1720 genes. Rice (Cranston et al. 2007)
- 2970 genes. Seed plants (Lee et al. 2011)
- >8000 genes. Birds (Jarvis et al. 2014)
- 259 genes. Birds (Prum et al. 2015)
- 859 genes. Seed plants (Wickett et al. 2014)

Mollusc phylogeny

- 14 transcriptomes added to existing genomes and EST datasets
- RAxML partitioned analyses
- PhyloBayes CAT model
- MrBayes unpartitioned
- general agreement between 301 and 1185 gene datasets



Node Support Legend
A/B/C/D (all values are percent)
A - Bootstrap, RAxML WAG, Big matrix
B - Bootstrap, RAxML WAG, Small matrix
C - Posterior probability, MrBayes WAG, Small matrix
D - Posterior probability, PhyloBayes CAT, Small matrix
*** 100/100/100/100 support**

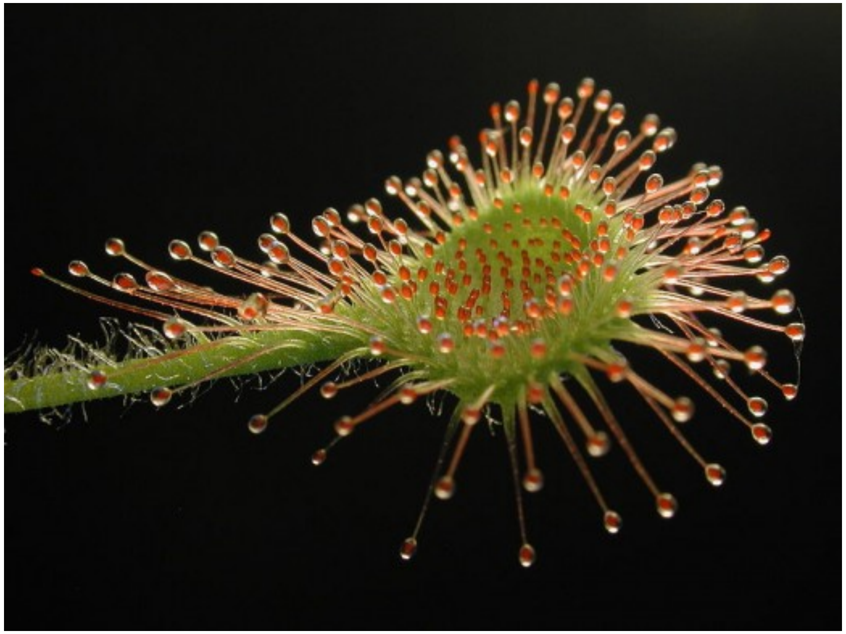
Taxa with new data shown in green

Smith et al., 2011

Caryophyllales

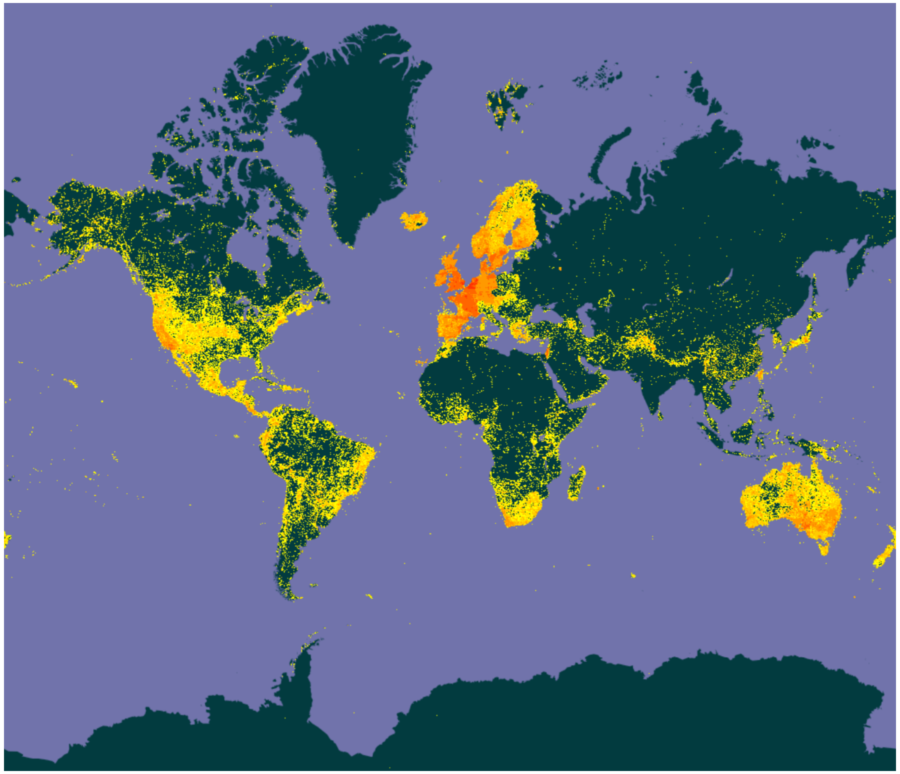
Caryophyllales

- >12,500 species in 39 families
- extreme disparity in life history and ecology



Carnivory

Cold environments

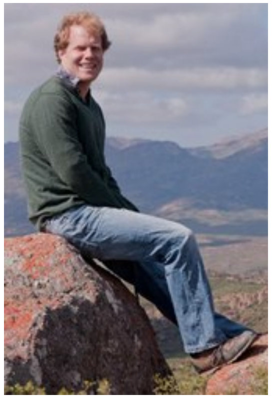


Photosynthetic modifications

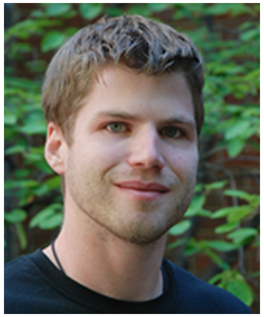
Morphological modifications

Caryophyllales

- NSF DEB funded
- 300 transcripomes broadly sampled
- investigate phylogeny and climate
- determine utility of transcriptomes for addressing broad evolutionary questions



Sam Brockington (postdoc – now curator at Ucambridge)



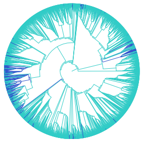
Joe Walker (grad student at UM)

Michael Moore (co-PI)

Ya Yang (postdoc – now employed at Uminn)

Driving biological questions

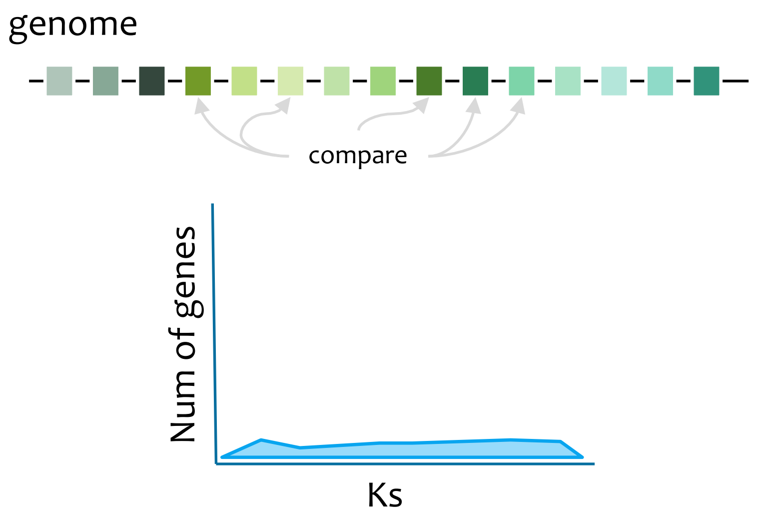
Is there a molecular signature of life history across the genome?

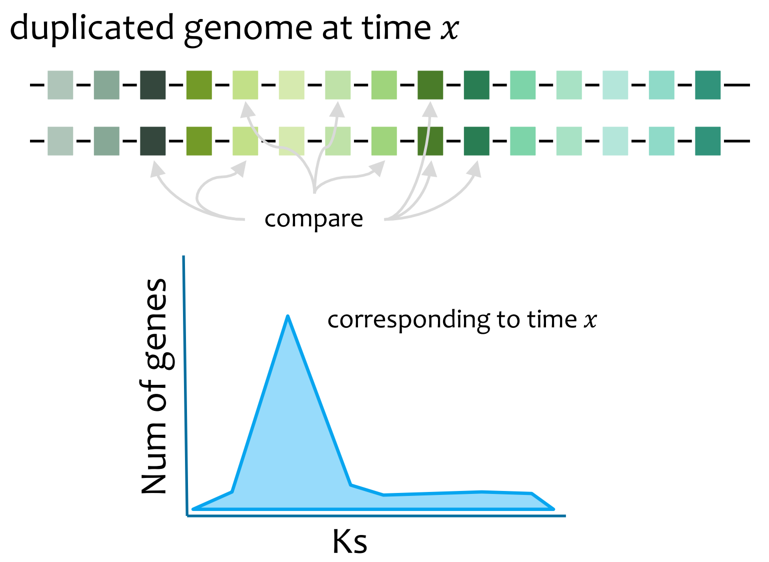


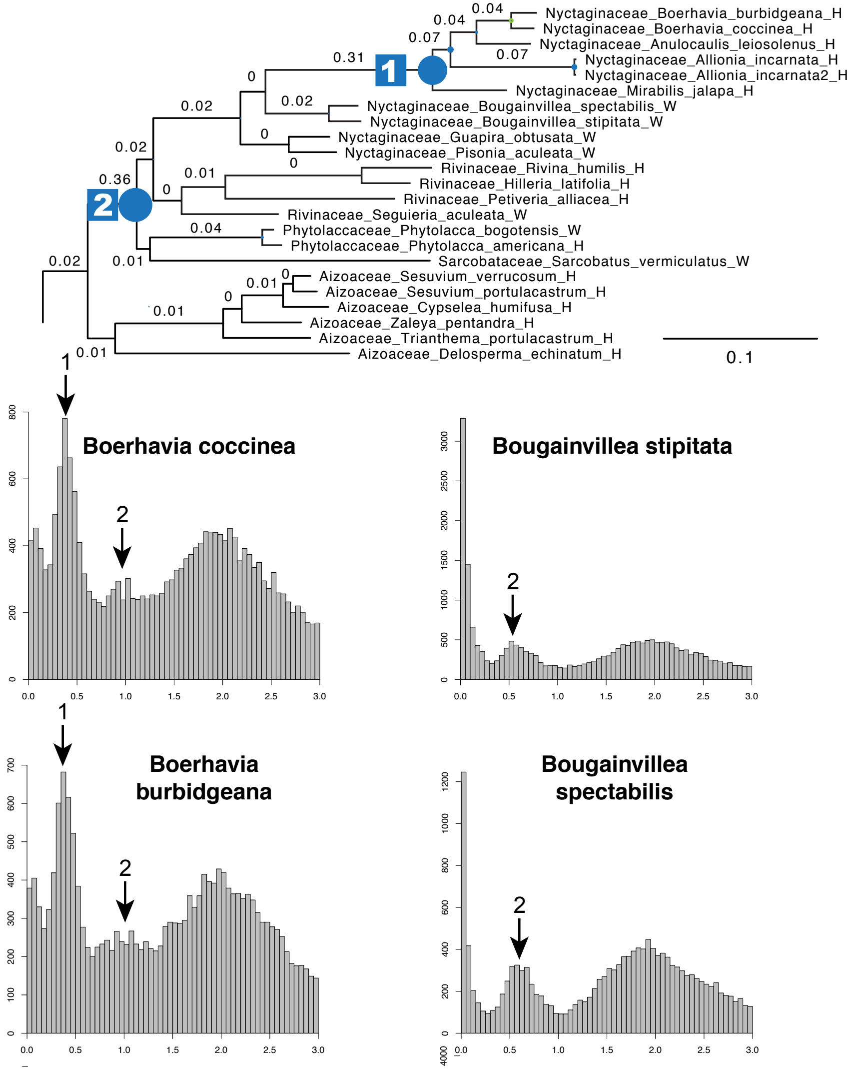
Are gene/genome duplications associated with diversification?

Can we determine gene families that have been important for new traits or ecological adaptation?

Genome duplications

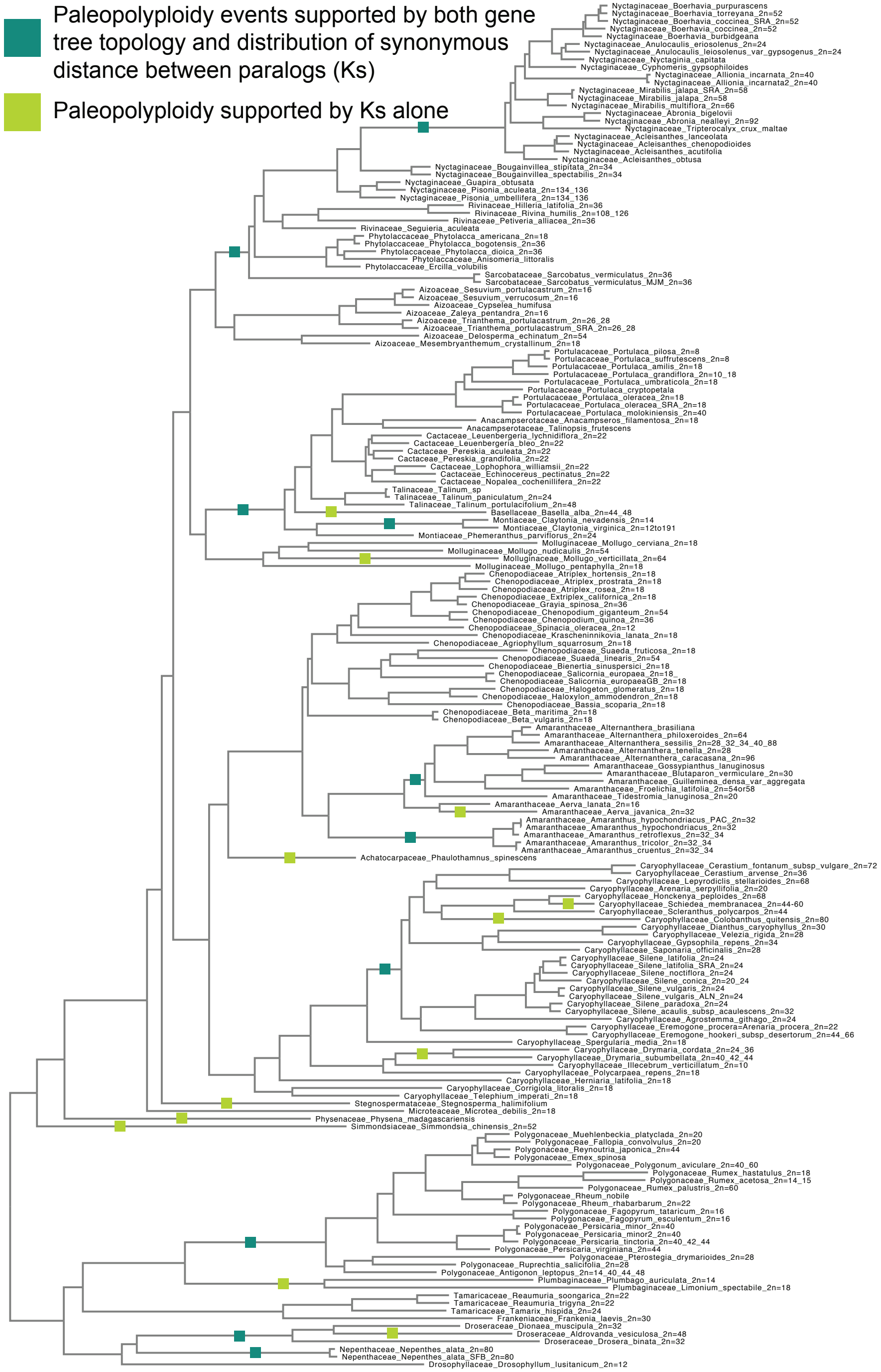






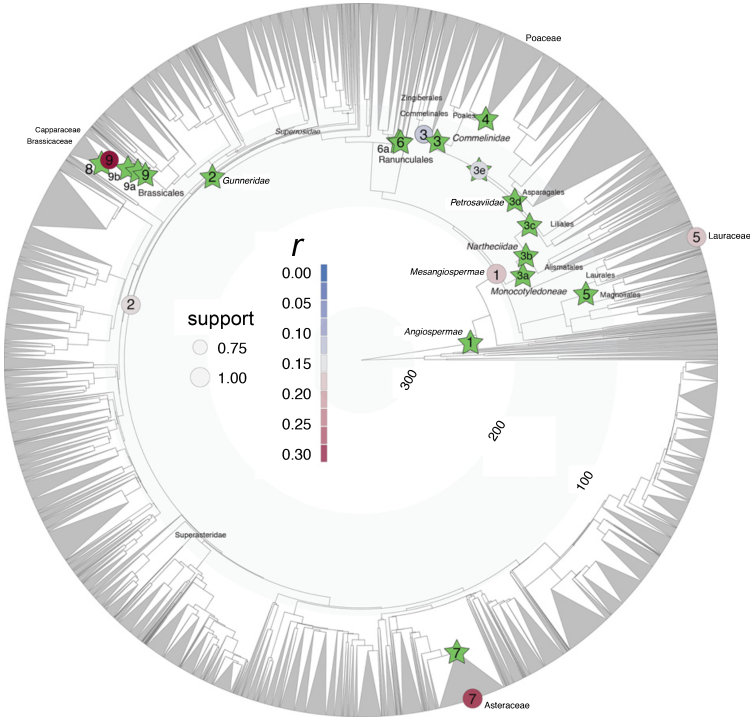
Genome duplications in Caryophyllales

- expanded analysis with 50 more taxa added
- 22 duplications (so far – analyses are on-going)
- Likely to continue to find more
- Yang et al. *in review*



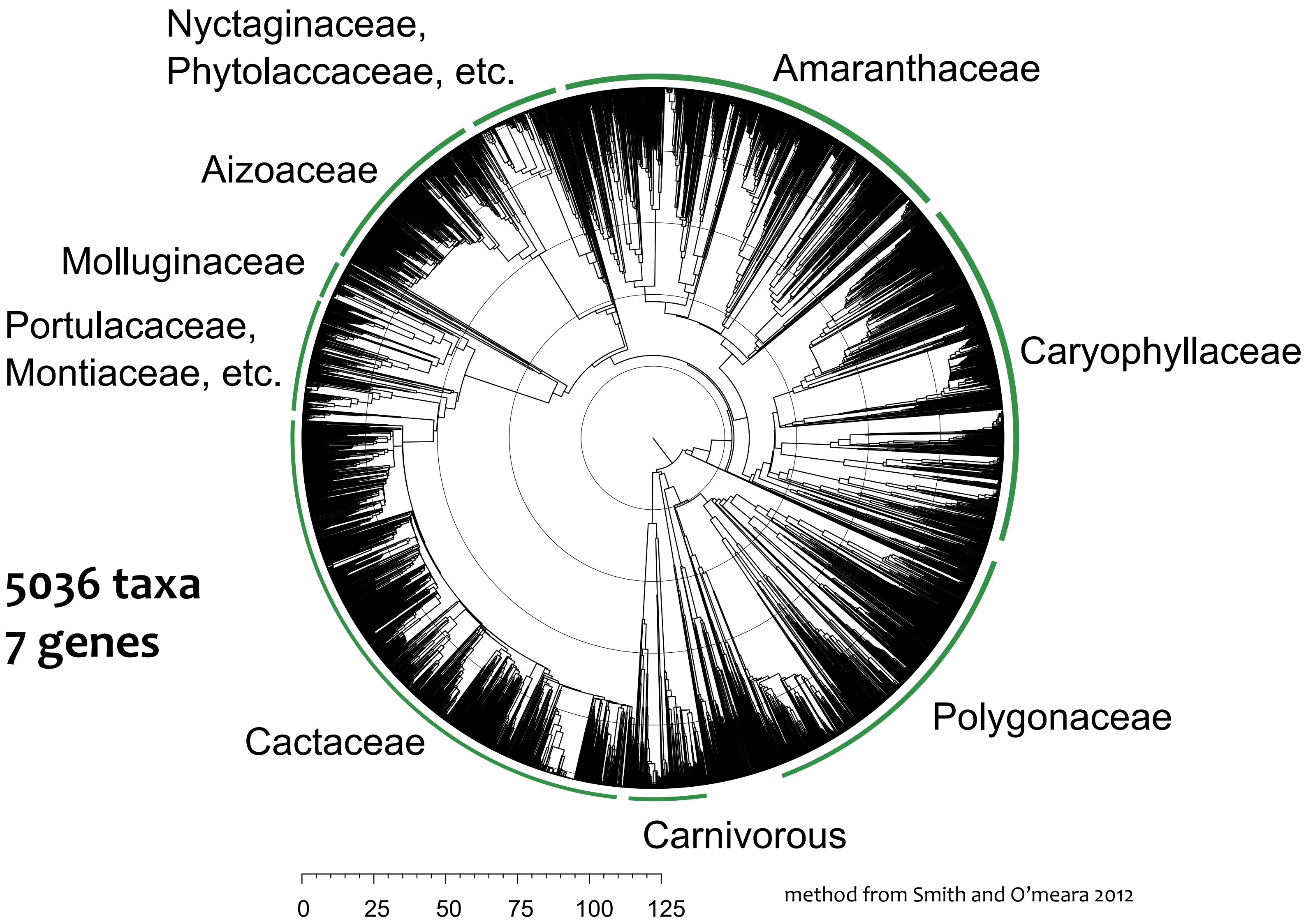
Duplications and diversification

- It has been suggested that **diversification** is associated with **genome duplication**
- This is highly debated in the literature



- From Tank et al. 2015
- Suggests that duplications (in green) general precede diversification shifts (in red/blue)
- Using 9 duplications

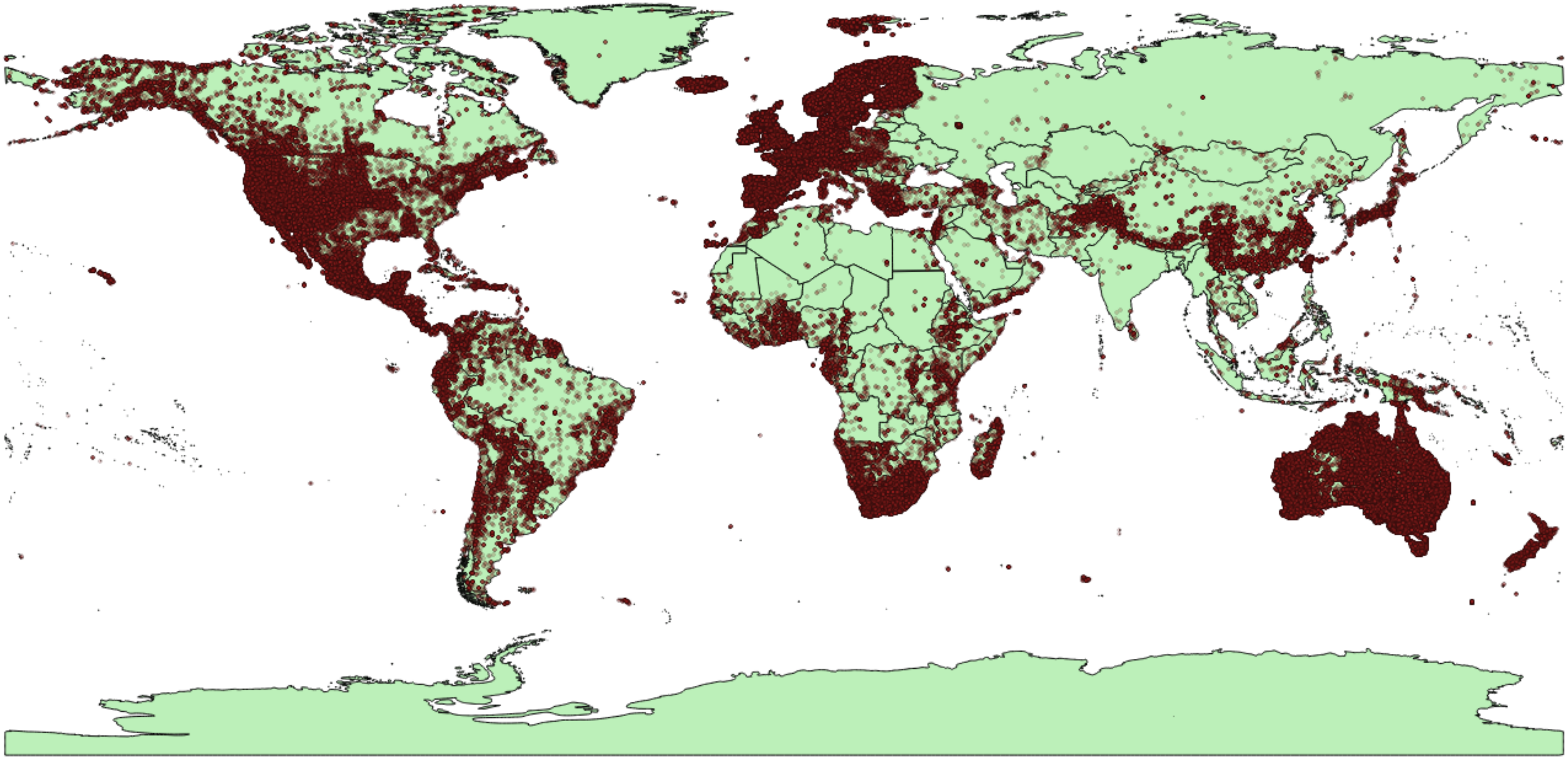
Dated phylogeny of Caryophyllales



- **5036 taxa**
- **7 genes**

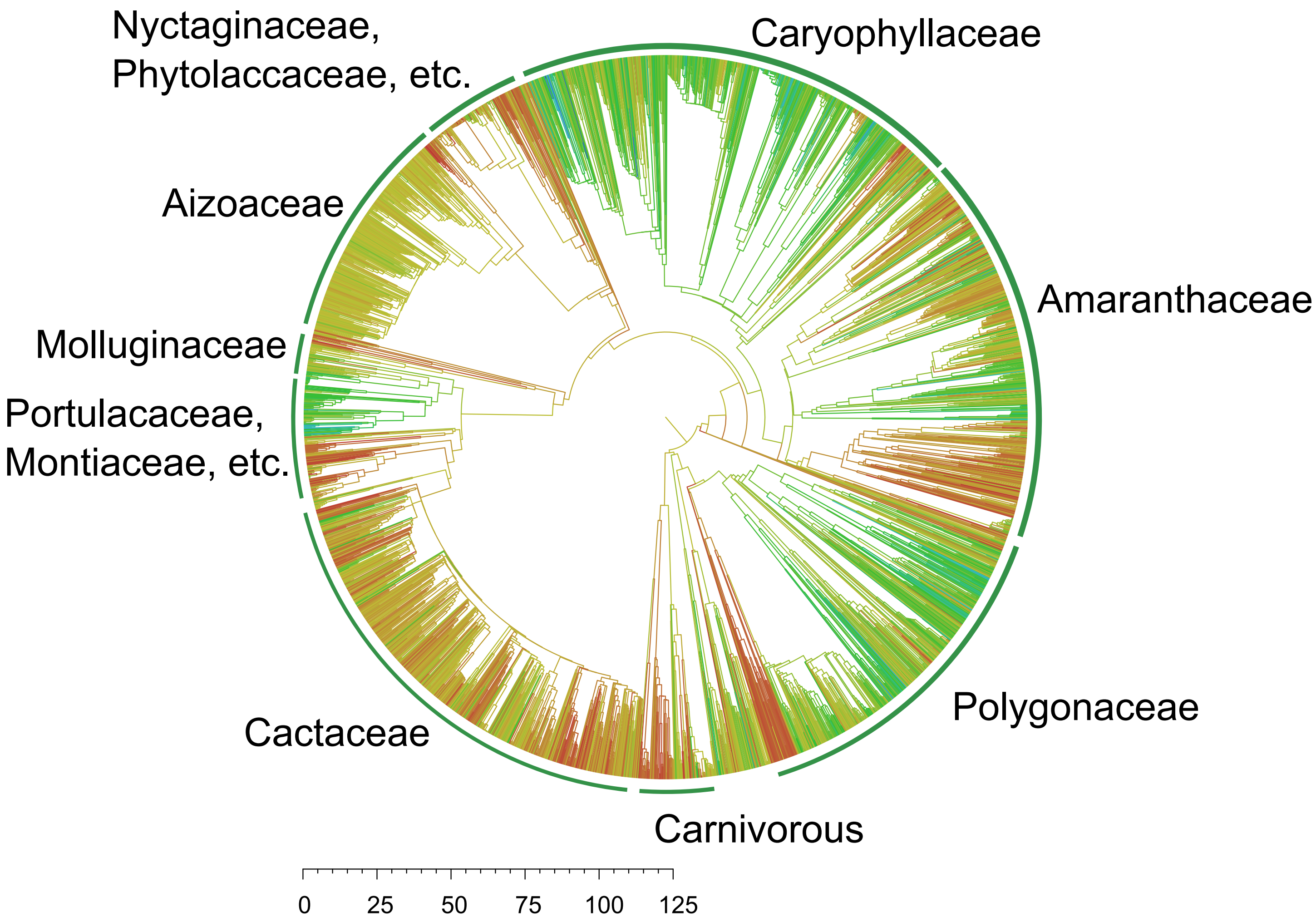
method from Smith and O'meara 2012

Caryophyllales geography

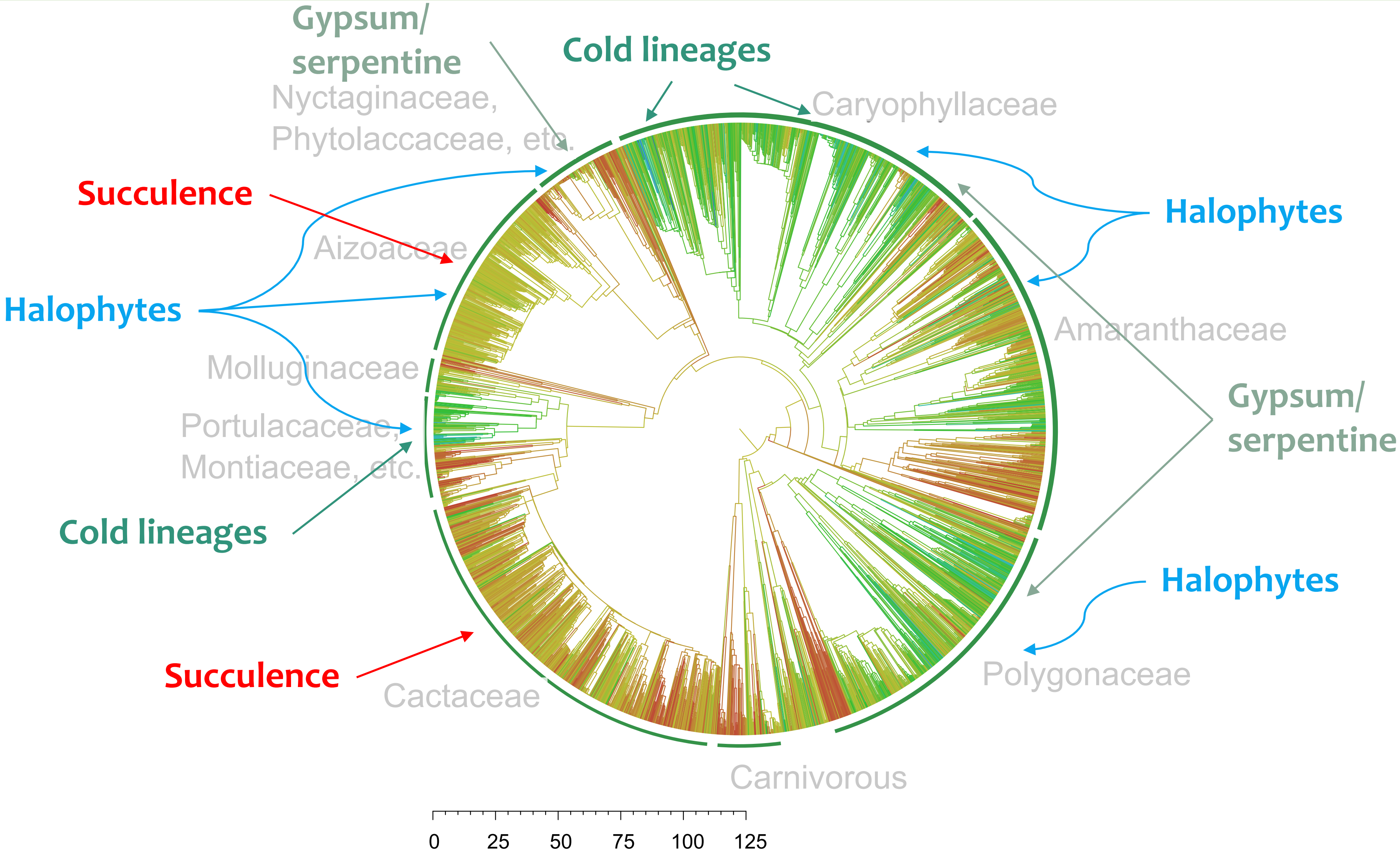


>6,000,000 filtered coordinates

Caryophyllales: annual mean temperature



Adaptations



Gypsum/ serpentine

Cold lineages

Succulence

Halophytes

Halophytes

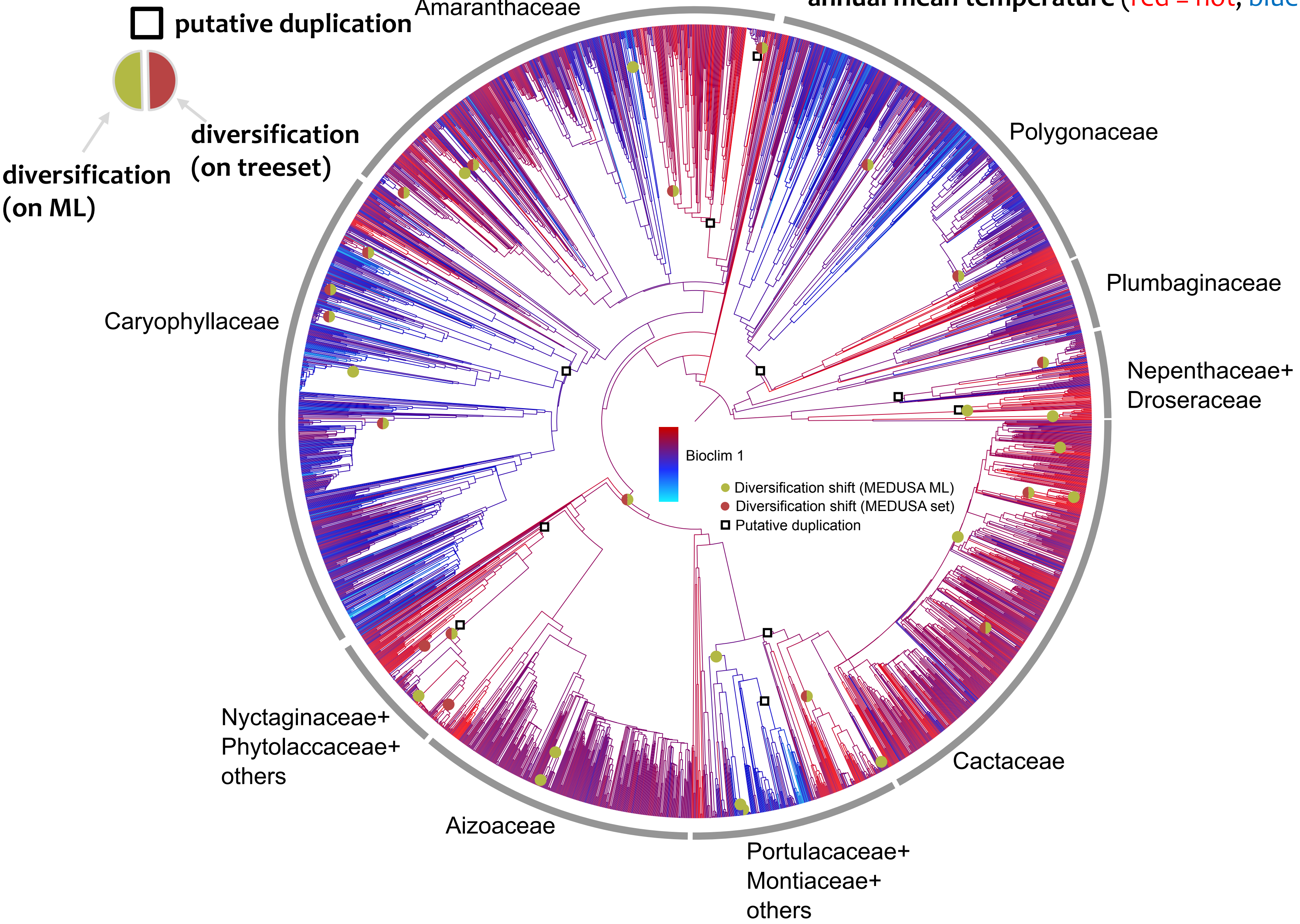
Gypsum/ serpentine

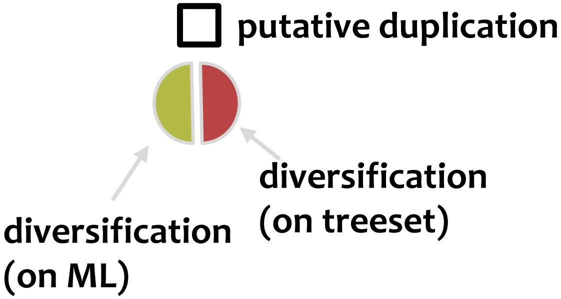
Cold lineages

Halophytes

Succulence

annual mean temperature (red = hot, blue = cold)



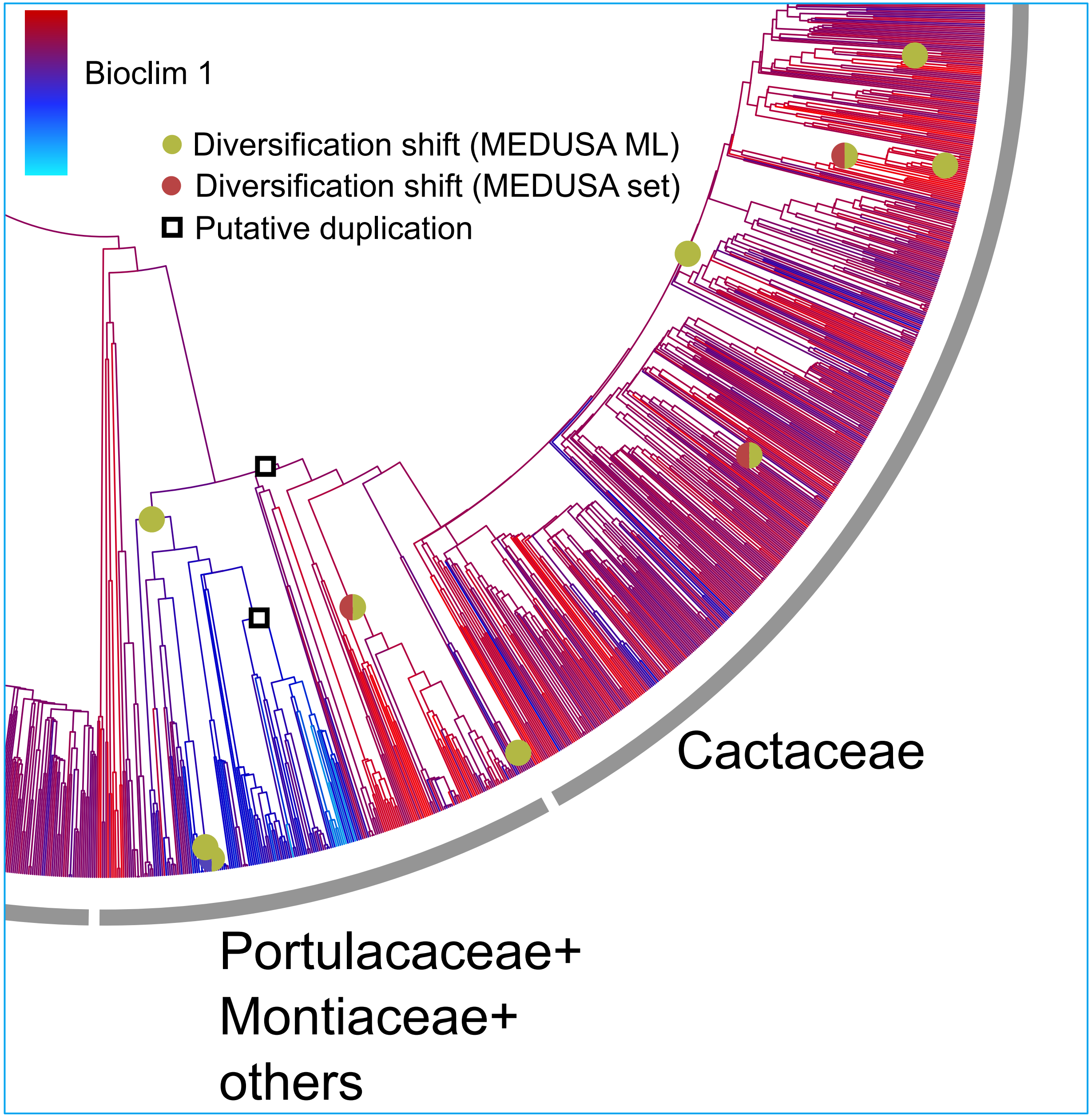


☐ **putative duplication**

diversification (on treeset)

diversification (on ML)

annual mean temperature (red = hot, blue = cold)

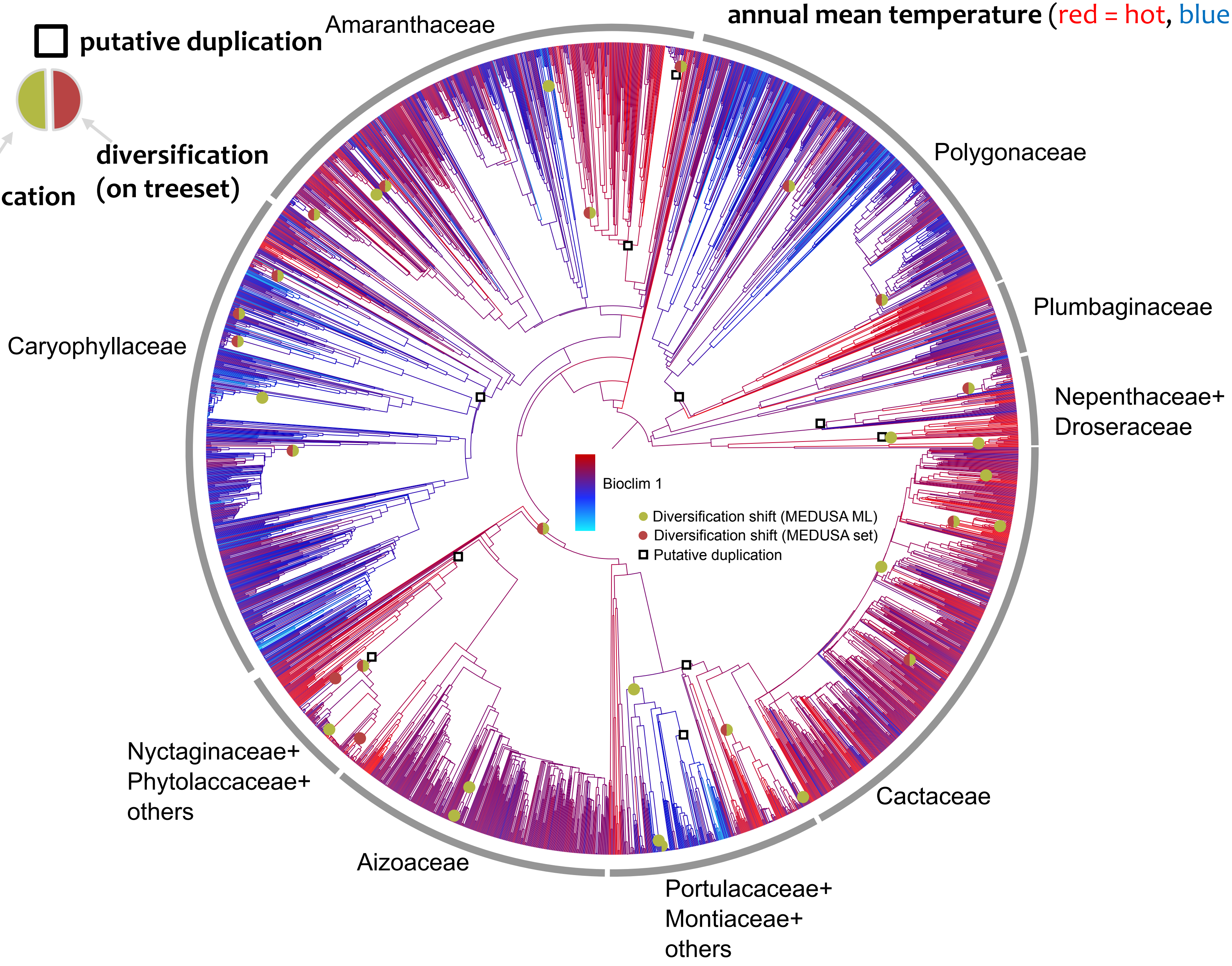


annual mean temperature (red = hot, blue = cold)

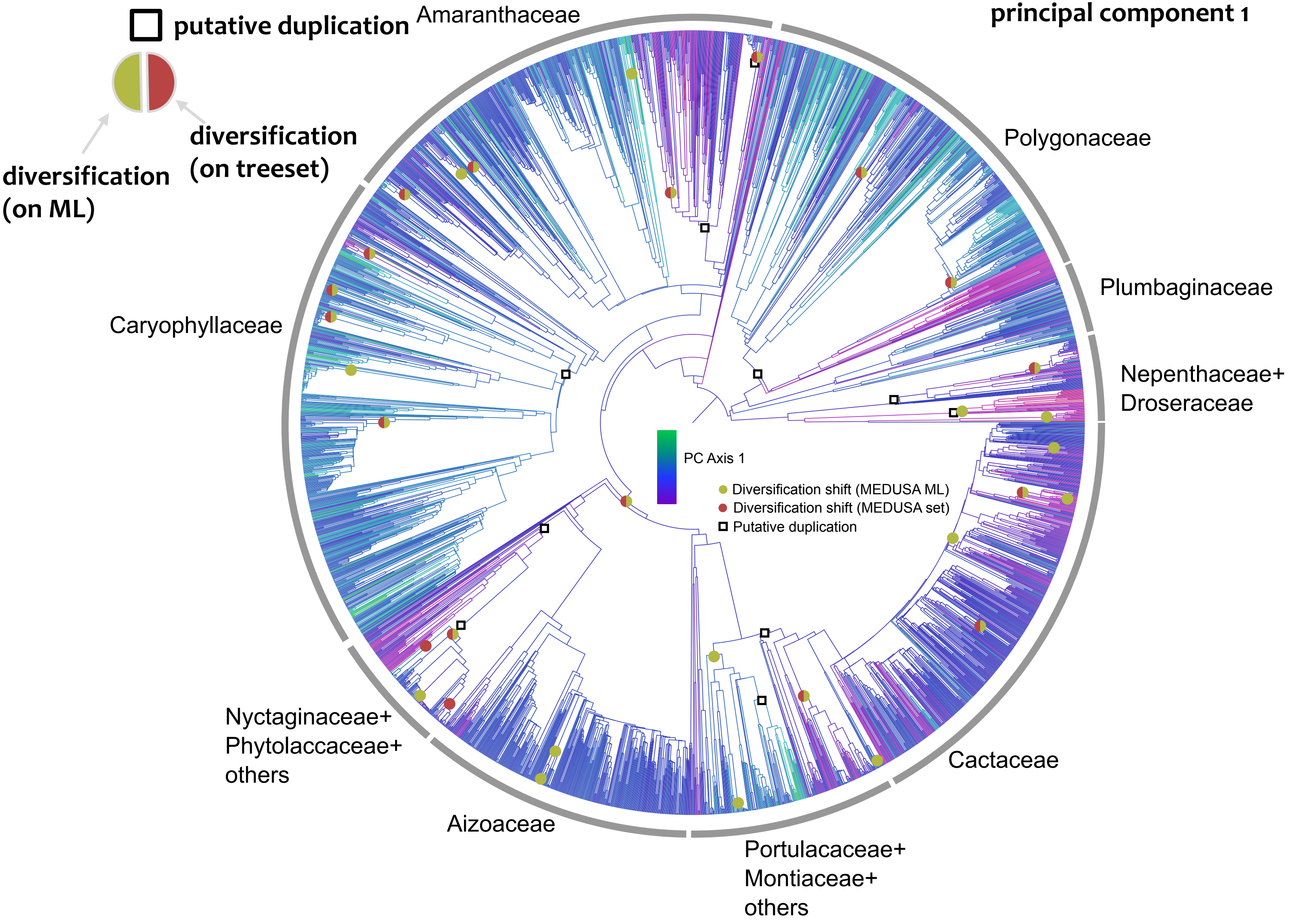
☐ **putative duplication**

diversification (on ML)

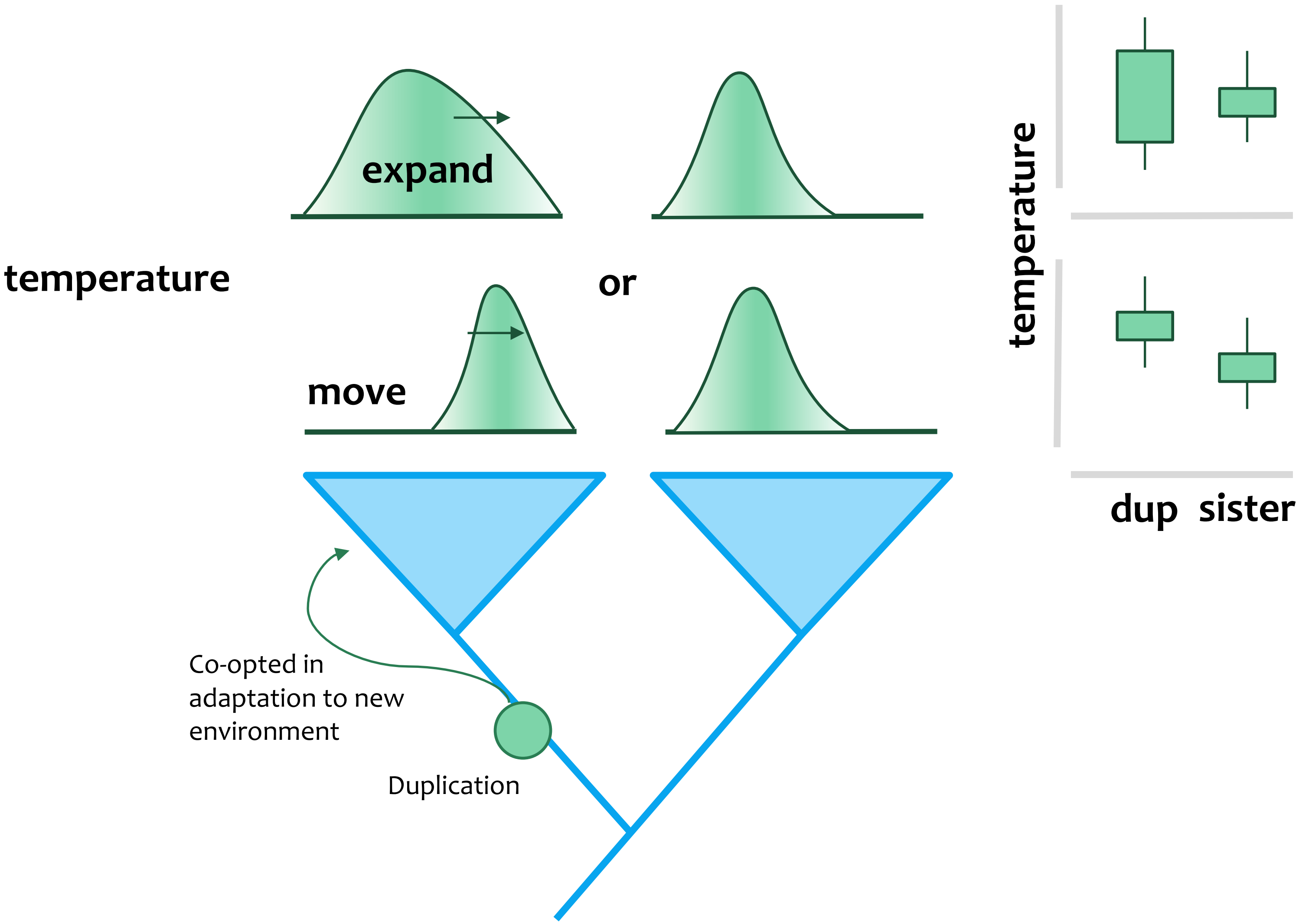
diversification (on treeset)



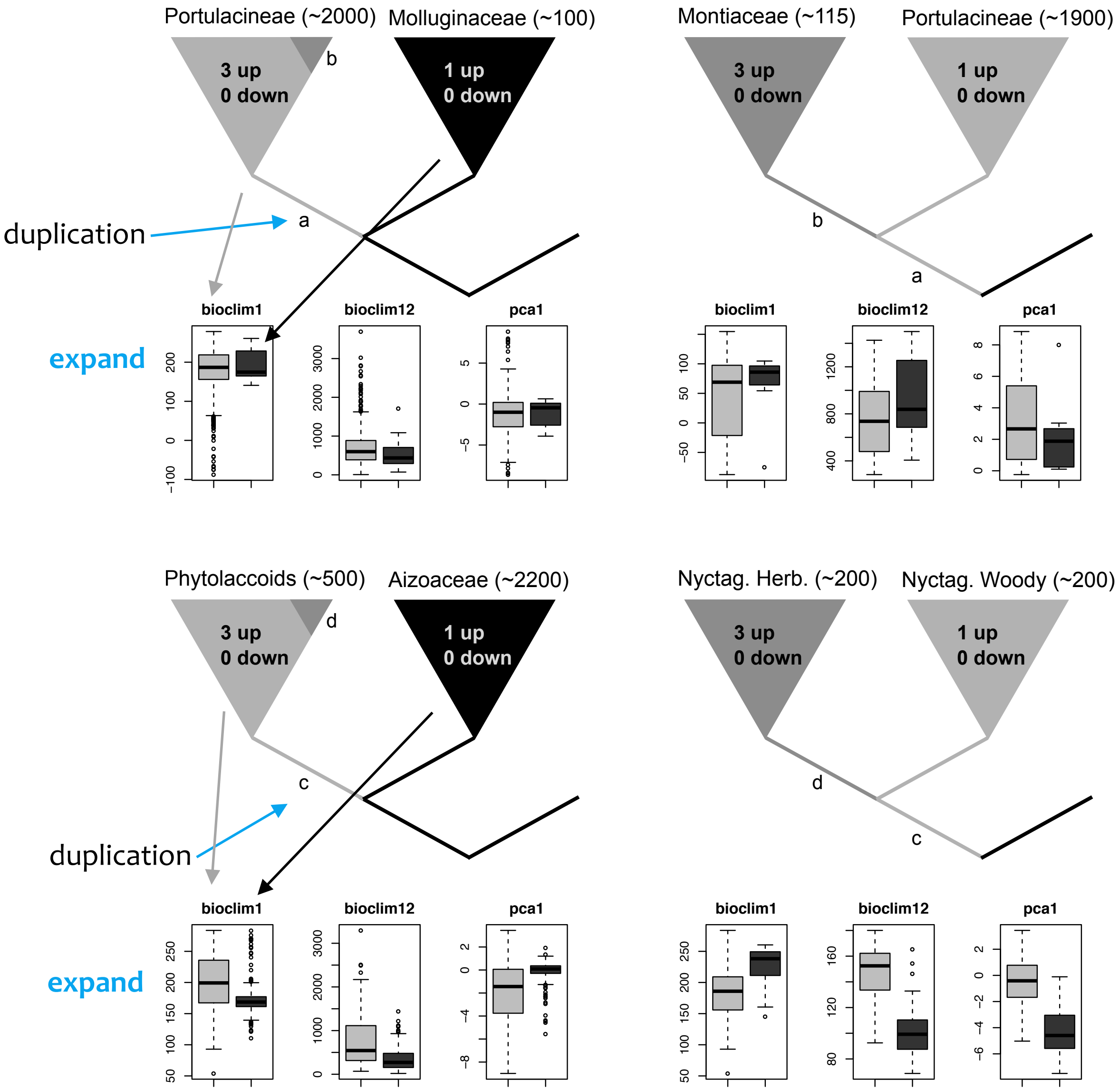
principal component 1



Climate shifts and genome duplications

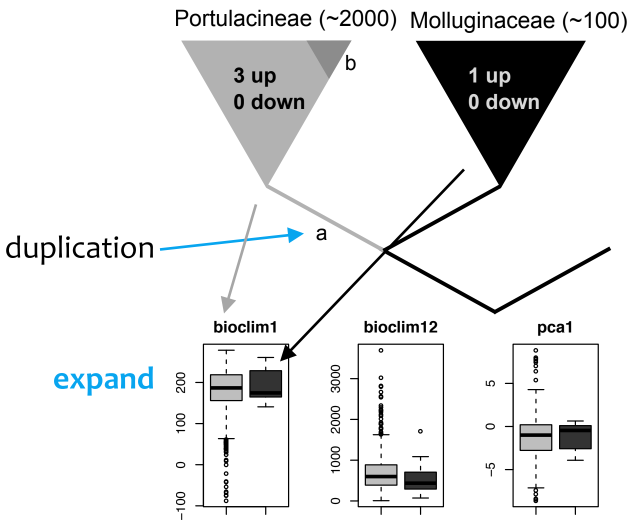


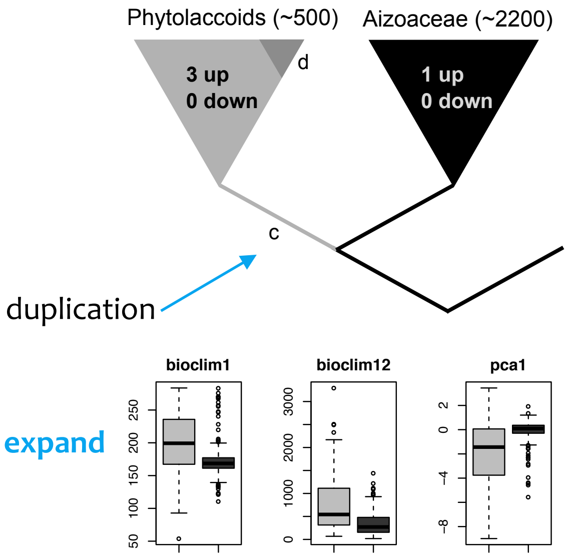
Summary of results



Summary of results

- Duplications **are not** associated with diversifications
- Duplications occur **before** diversification
- We need a better description of the biological model that would associate diversification and dups
- Many duplications associated with expanded climate ranges
- Combining these paradigms will likely transform evolutionary biology
- Smith et al. *in review*





What would we like to do?

We would like a phylogeny that is

- Accurate and precise
- Has branch lengths relative to time
- Comprehensive

Why is constructing a tree for all life hard?

Challenges

Computational challenges

- Tree space is large
- Datasets can be large
- Basic limits on computational time and memory

Biological challenges

- Some nodes are difficult to resolve because of biological reasons
  - Incomplete lineage sorting
  - Horizontal gene transfer
  - Duplications and loss

Data challenges

- We haven’t collected all of the data (yet)
- Some of the data is difficult to include in datasets

Tree space is large

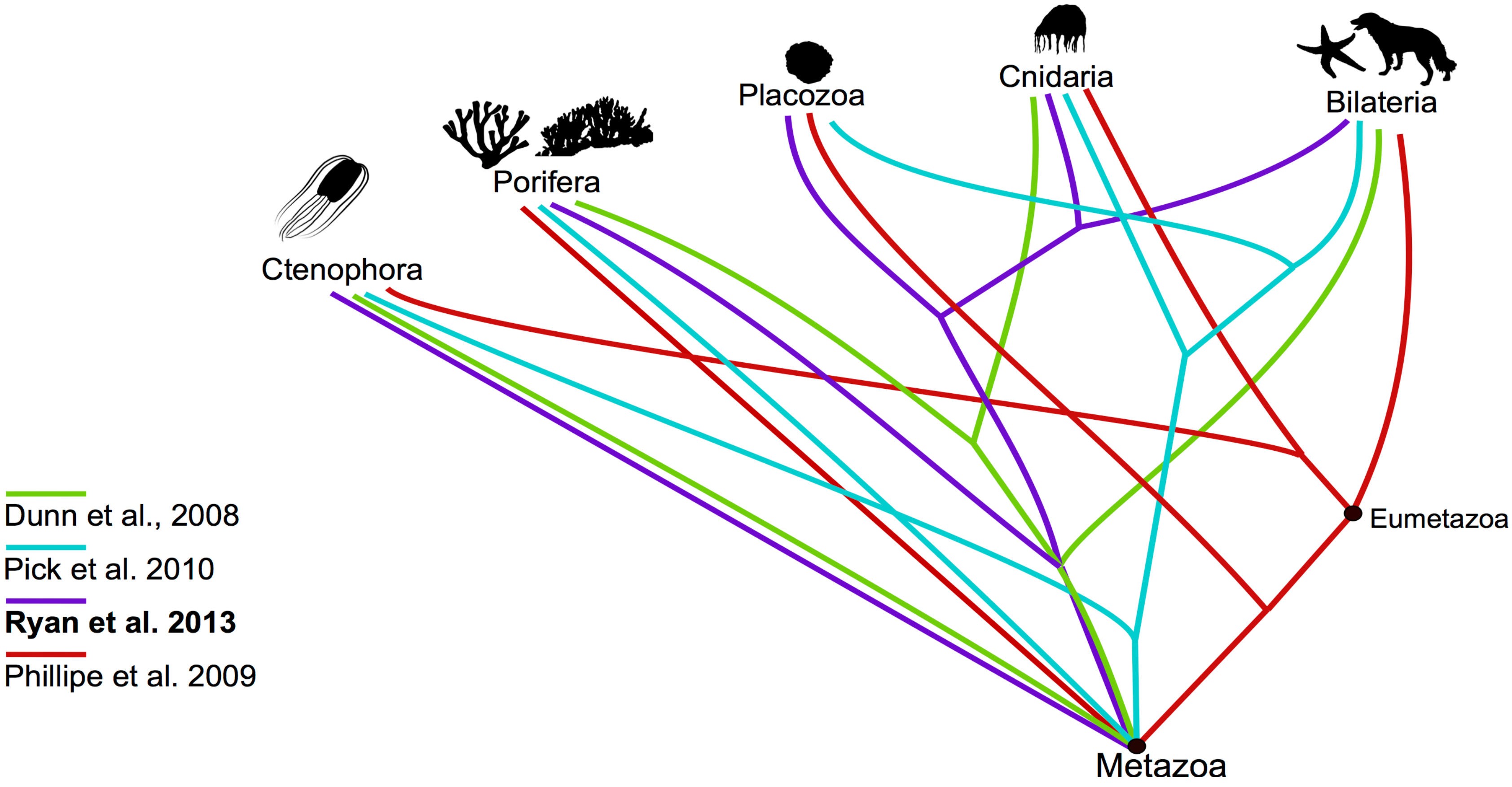
Number of rooted trees = $\frac{(2n-5)!}{2^{n-3}(n-3)!}$

- 5 species = 105 trees
- 6 species = 945 trees
- 7 species = 10,395 trees
- 9 species = 2,027,025 trees
- 20 species = 8,200,794,532,637,892,000,000 trees
- 30 species = 495,179,769,008,019,800,000,000,000,000,000,000,000 trees
- 40 species =57 zeros
- 60 species = 96 zeros
- 80 species = 139 zeros

Number of stars in the universe is between 20 and 30 species

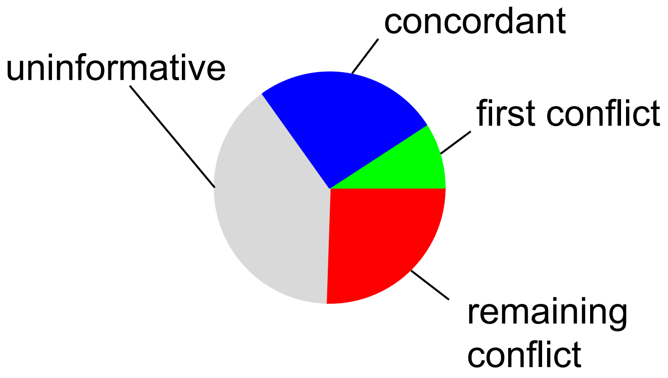
Finding the best tree is NP-hard

Edges can be difficult to resolve

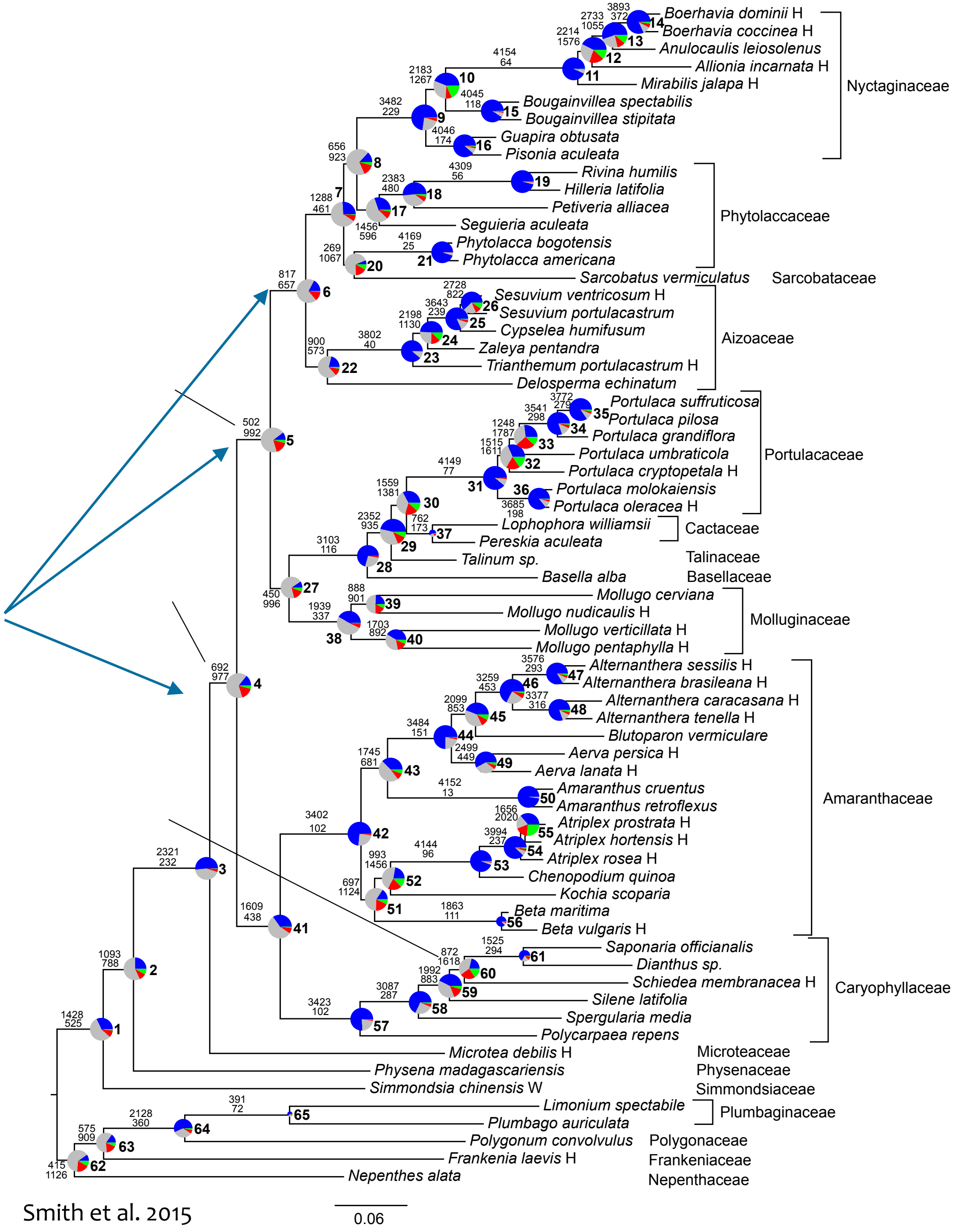


Hinchliff et al. 2015

Proportion of Homologs



few informative genes



Smith et al. 2015

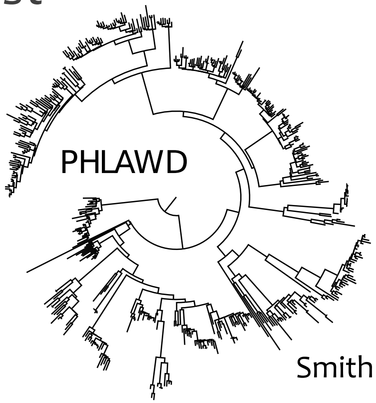
Nature of molecular data

- major sequence database (GenBank) stores species taxonomy but not gene taxonomy
- hard to get gene identification for all of GenBank
- all genes for species x
- NOT all the species for gene y

PHLAWD

Phylogenetic dataset assembly with databases

- multithreaded C++ program for assembly large phylogenetic datasets
- automatically updating (as of Izquierdo-Carrasco et al. 2014)
- baited Smith-Waterman
- you provide sequences and it does the rest



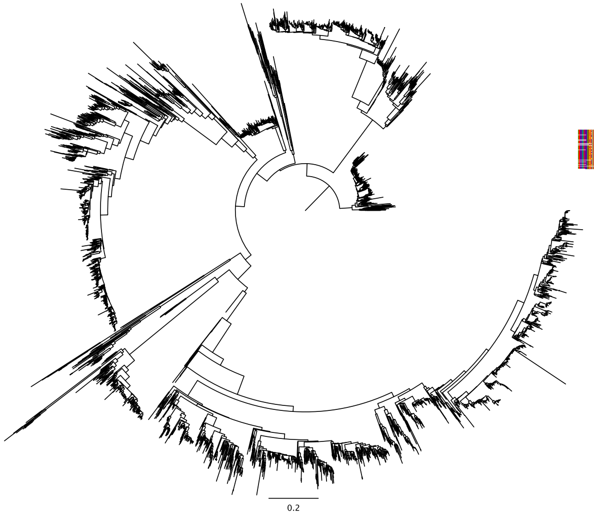
Algorithm

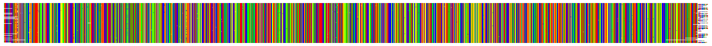
1. pick a group of organisms
2. pick a gene region -- identify sequences (x) that represent these genes
3. compare these seqs with those in the database (n) -- so $\boldsymbol{xn}$ comparisons not $\boldsymbol{n}^2$
4. test for saturation
5. divide and conquer if saturated

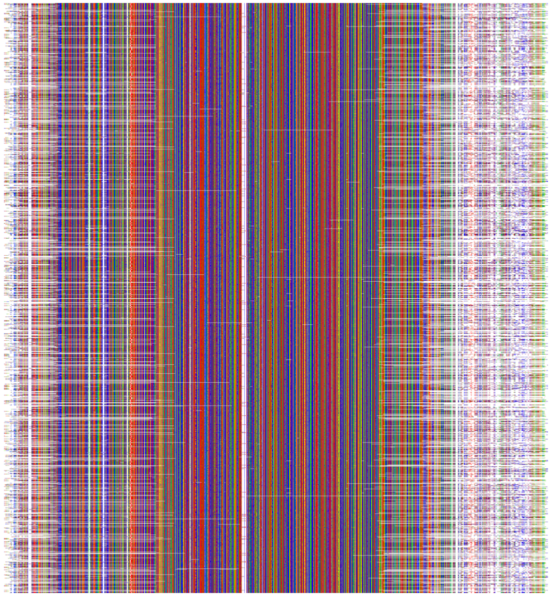
- repeat 2-4 for additional genes

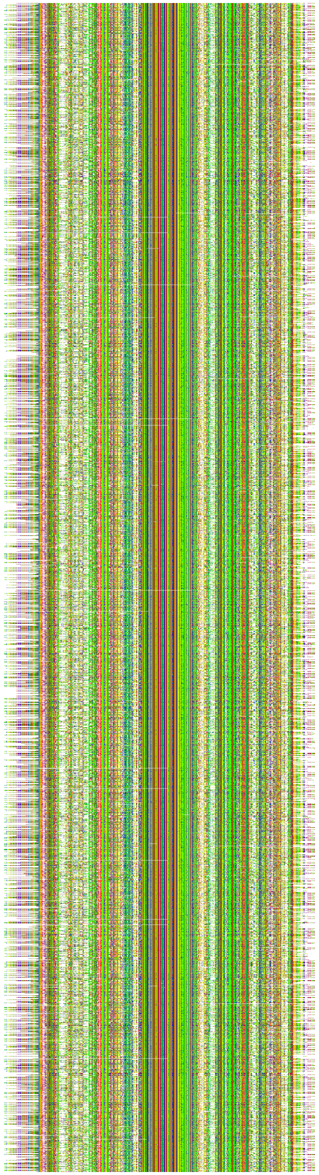
Smith et al. 2009, Izquierdo-Carrasco et al. 2014

Concatenate and construct a tree









Are data available to construct a tree for all life?

Data for building trees are growing in both directions

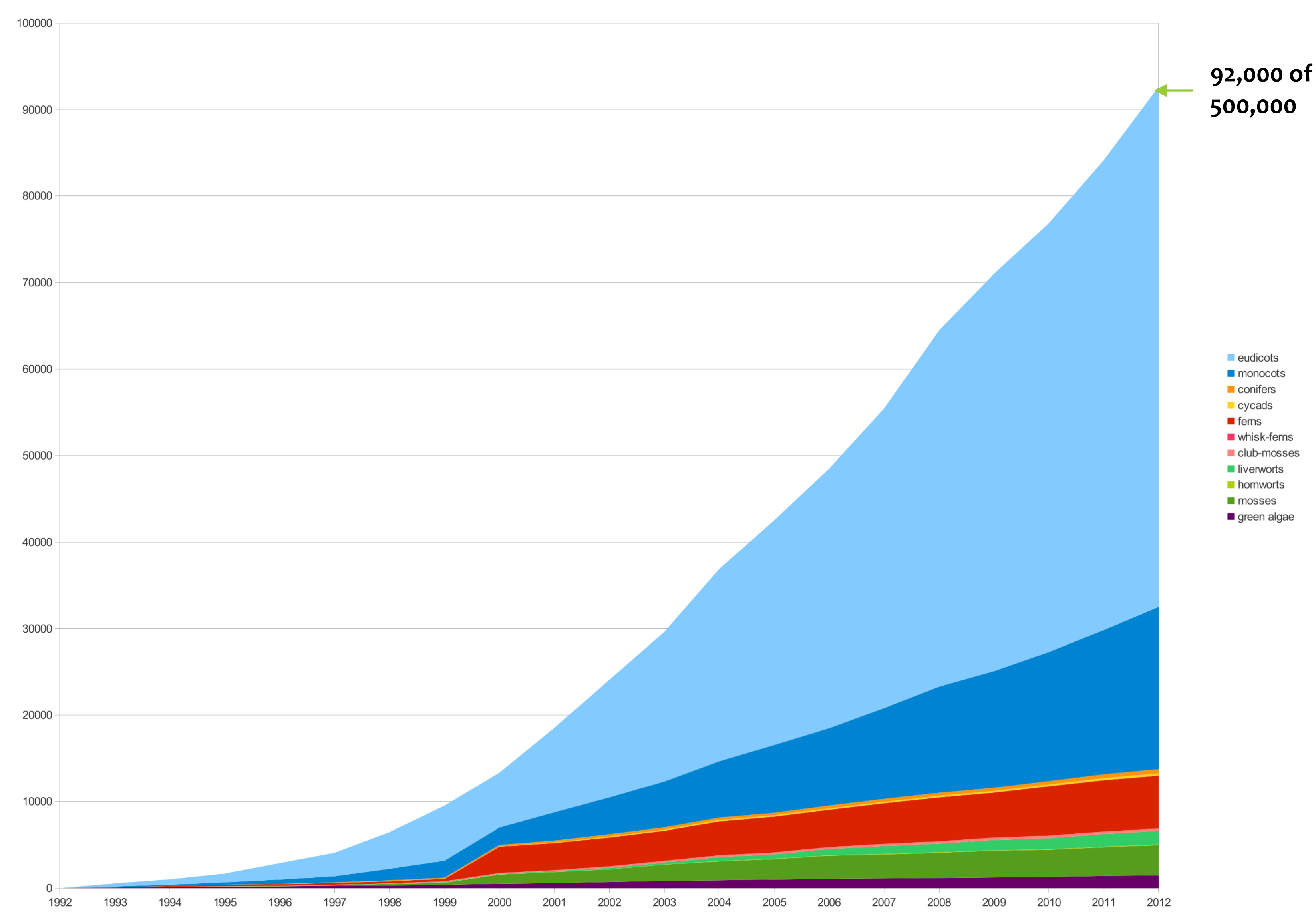
- >30,000,000 sequences in GenBank
- >500,000 species in Genbank
- >2.5 million described species
- >10 million species

Are data available to construct a tree for life?

- NO
  - > 500,000 species in GenBank

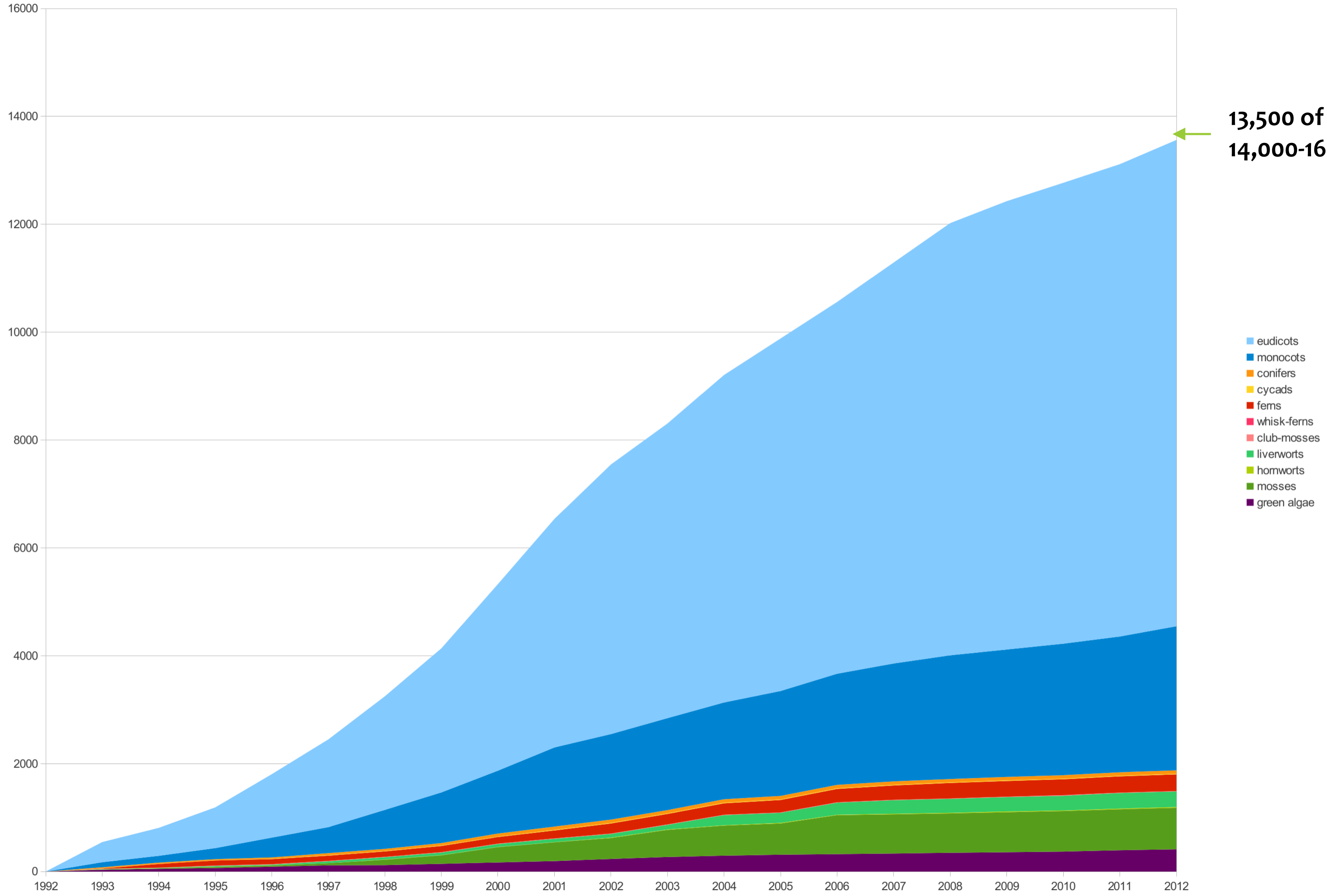
Are data available to construct a tree for plants?

Species of plants with molecular data



92,000 of 250,000-500,000

Genera of plants with molecular data

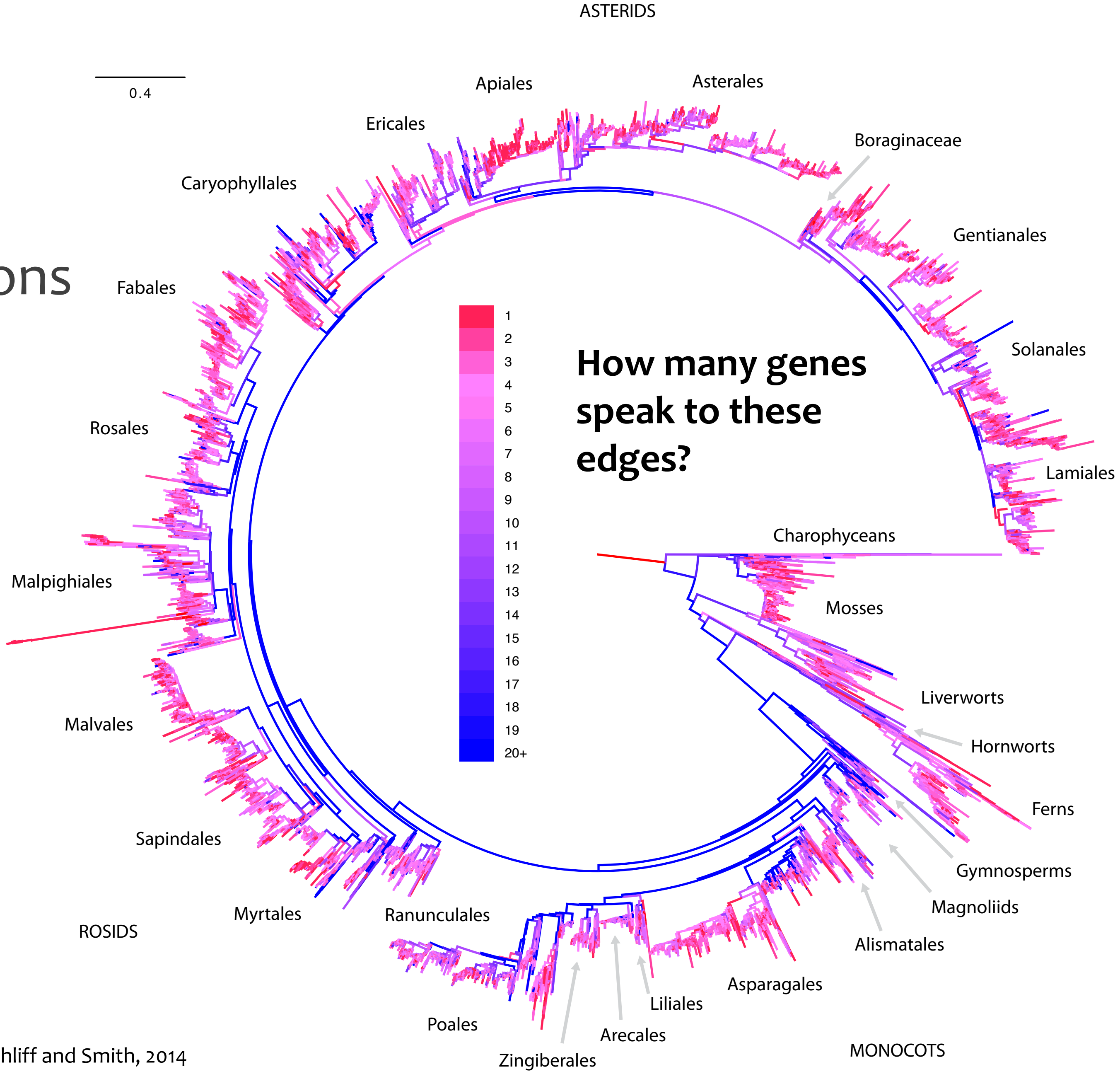


13,500 of 14,000-16,000

13,083 genera

128 gene regions

148,150 sites



How many genes speak to these edges?

Hinchliff and Smith, 2014

Are data available to construct a tree for life?

- NO

Are data available to construct a tree for plants?

- NO

What is the Open Tree of Life project?

OPEN
Tree of Life

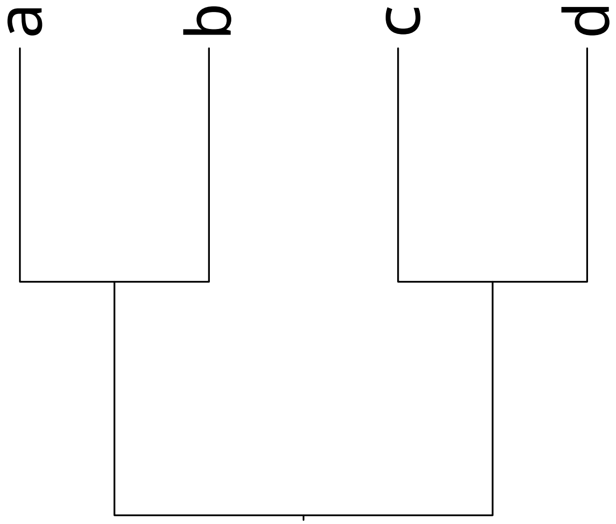
Open Tree of Life

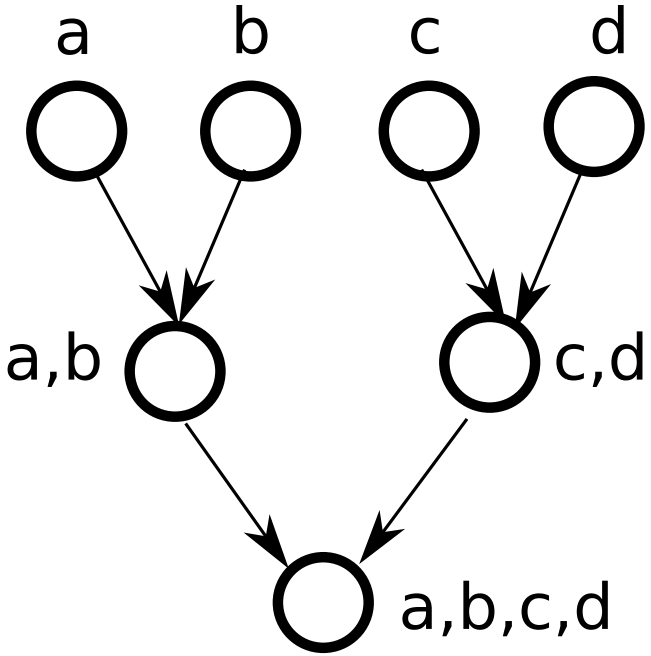
- Open Tree of Life (NSF AVATOL)
  - 11 PIs working collaboratively
- > 2.3 million species, extinct and extant
- Synthesize data from existing sources (trees and datasets)
- Allow for annotations, comments, and comparison

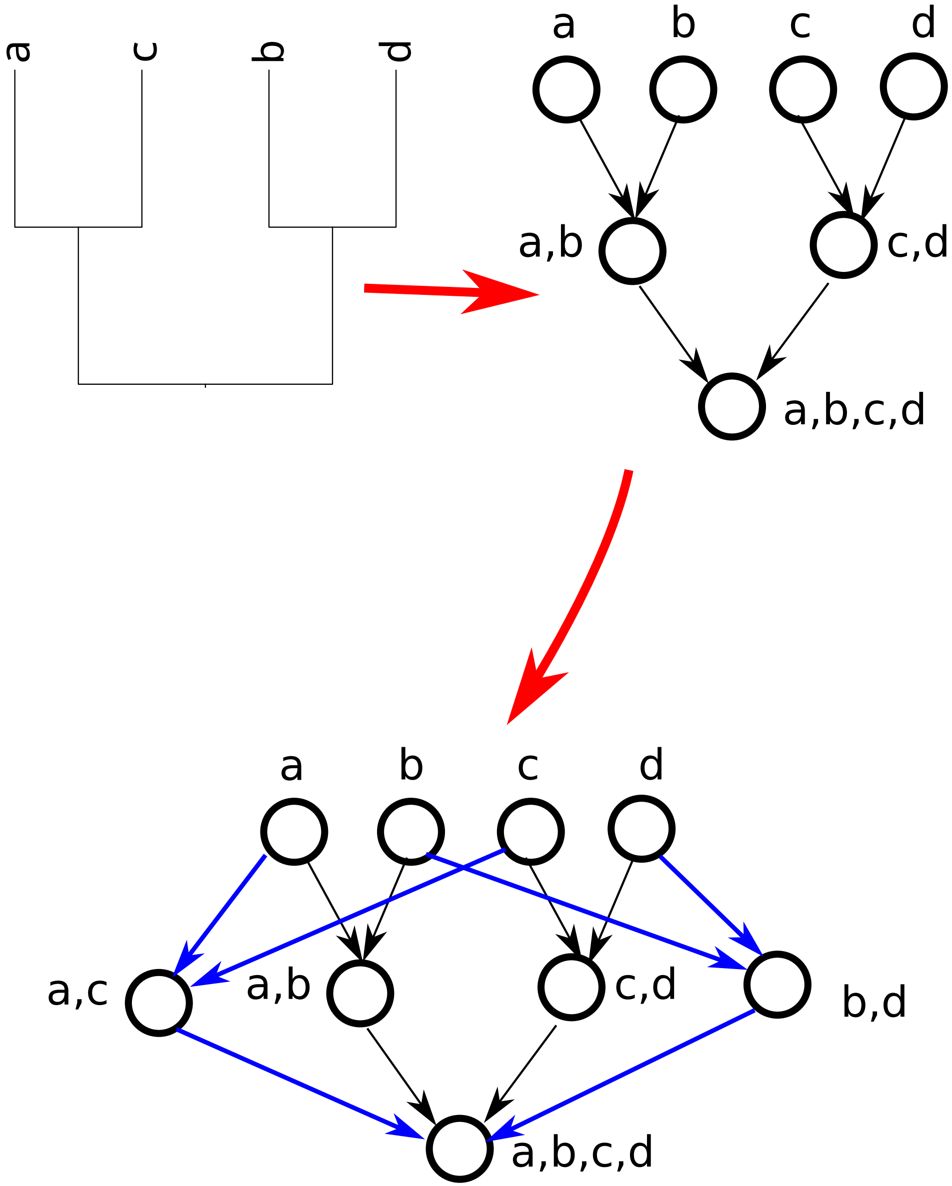
Open Tree of Life

- **We don't have genetic data for all species?**
  - We can use taxonomy and other sources of trees
- **We have to scale to the millions**
  - Few existing methods can do this
- **Solution:** We need a novel way to combine trees and to develop new methods for synthesis

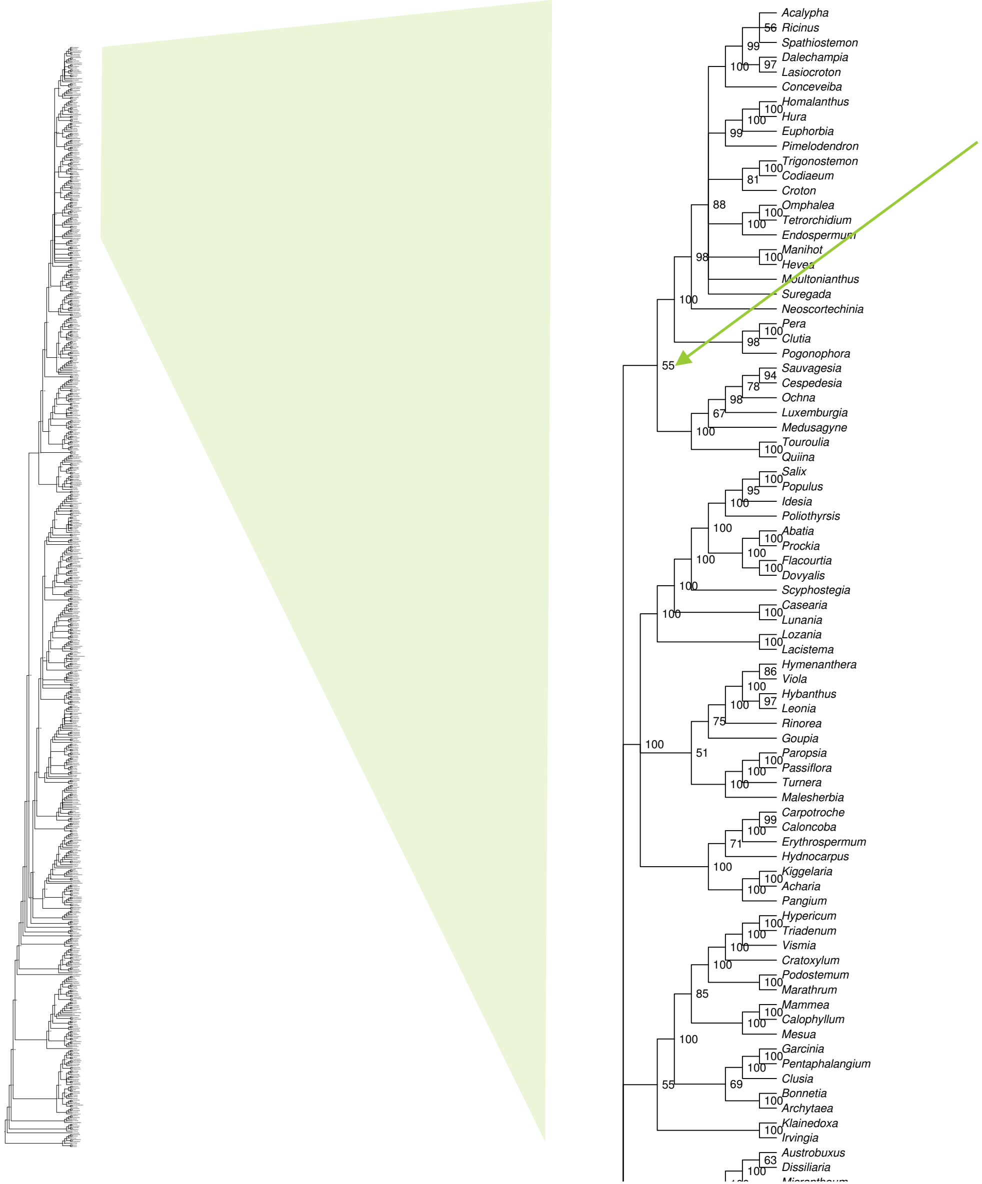
OPEN
Tree of Life


a
b
c
d

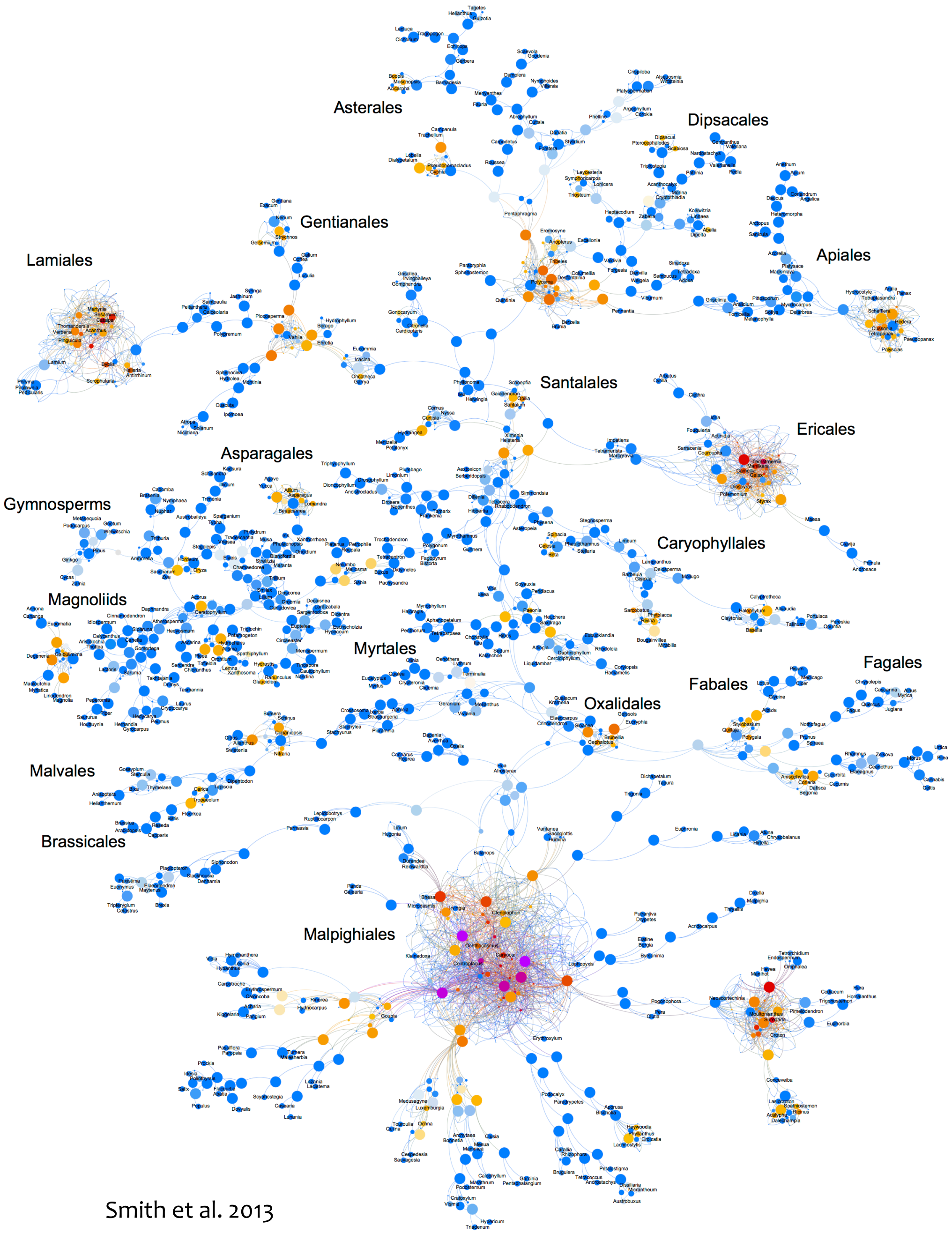

a
b
c
d
a,b
c,d
a,b,c,d


a
c
b
d
a
b
c
d
a,b
c,d
a,b,c,d
a
b
c
d
a,c
a,b
c,d
b,d
a,b,c,d

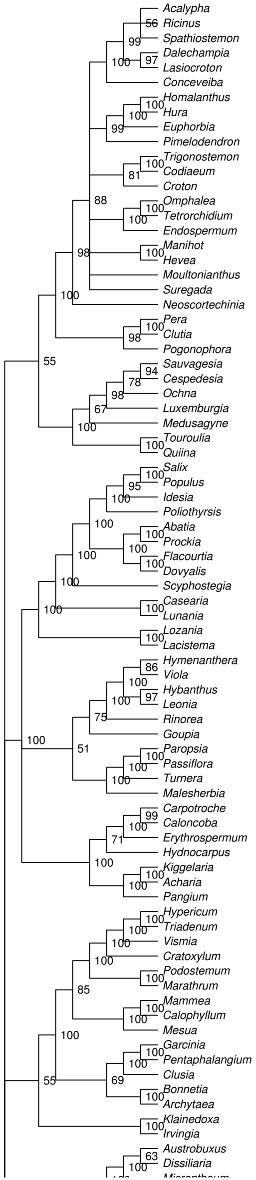
Angiosperm Tree of Life (Soltis et al., 2011)



uncertainty

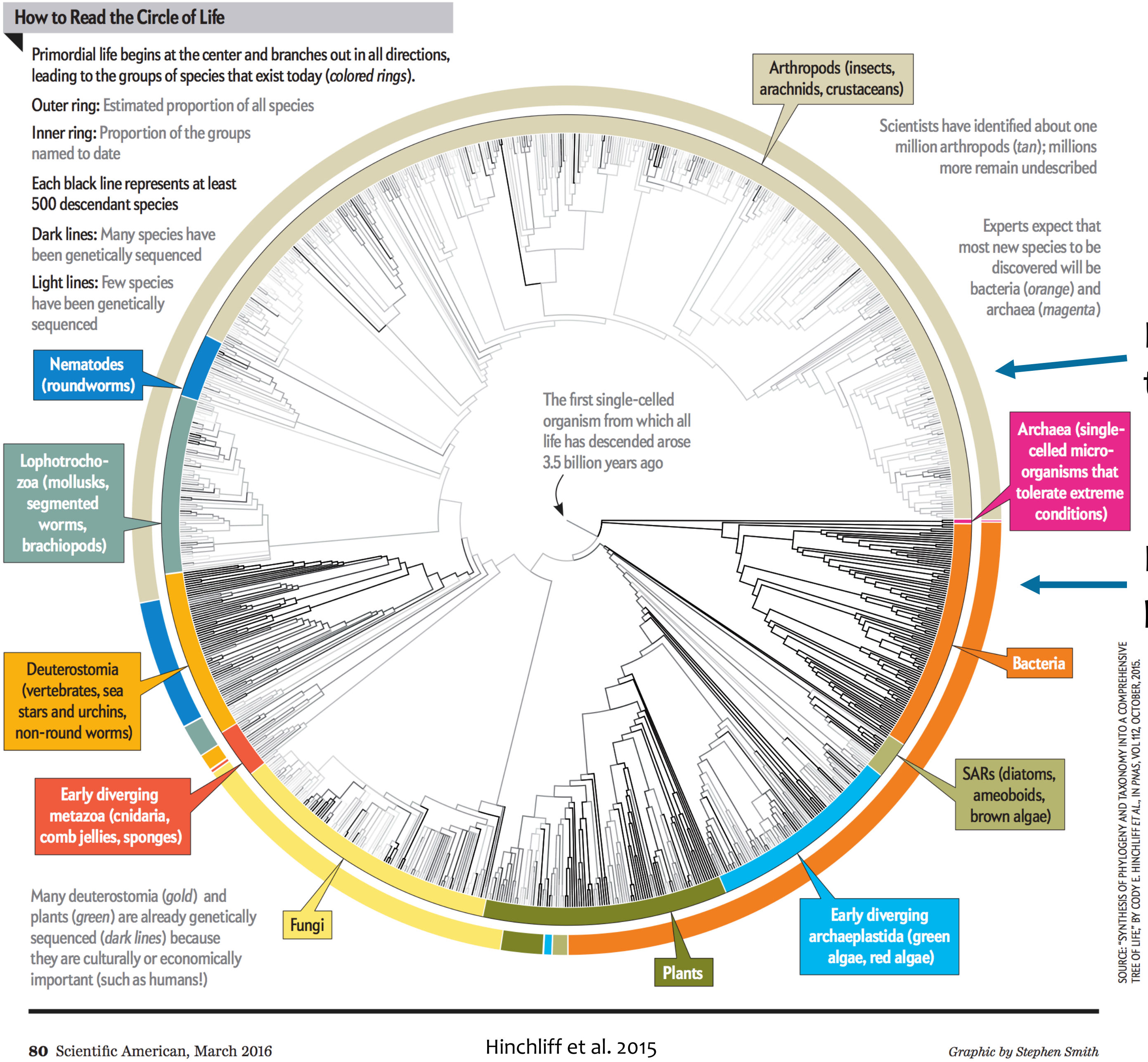


Smith et al. 2013



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



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

The first single-celled organism from which all life has descended arose 3.5 billion years ago

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*)

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

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

80 Scientific American, March 2016

Graphic by Stephen Smith

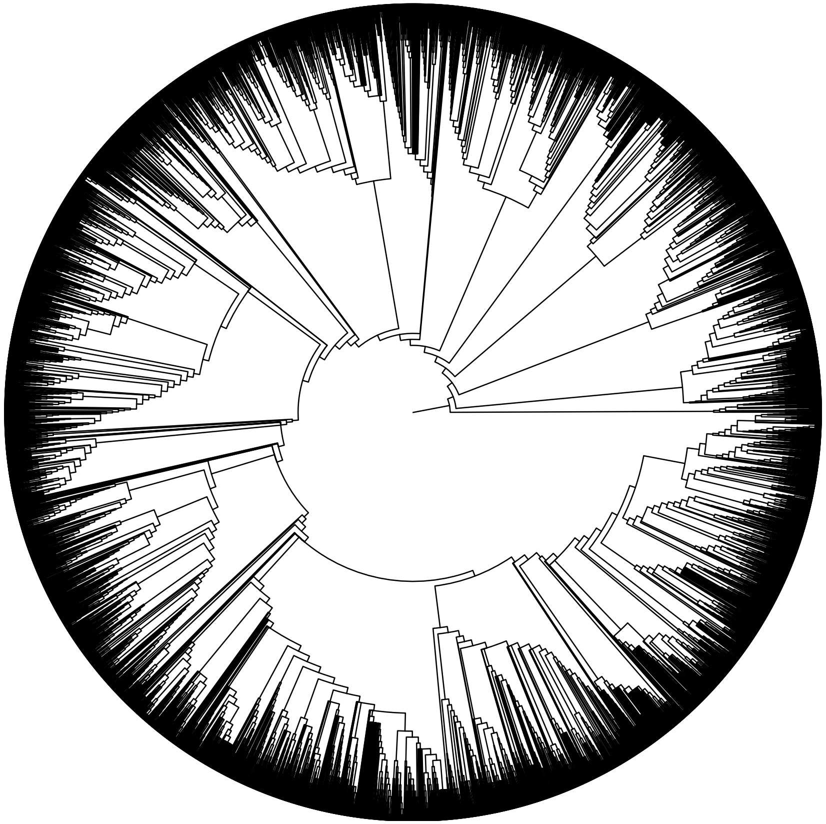
Informed by taxonomy

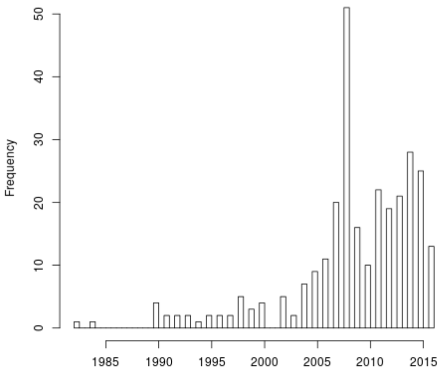
Informed by phylogenies

Hinchliff et al. 2015

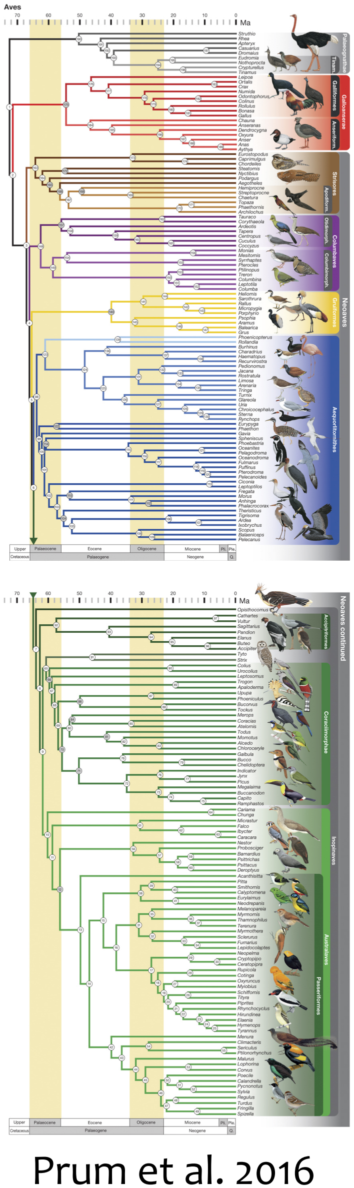
Birds and conflict

- combine 233 trees into a new tree for birds
- includes 13,579 taxa
- what if we compare to other trees



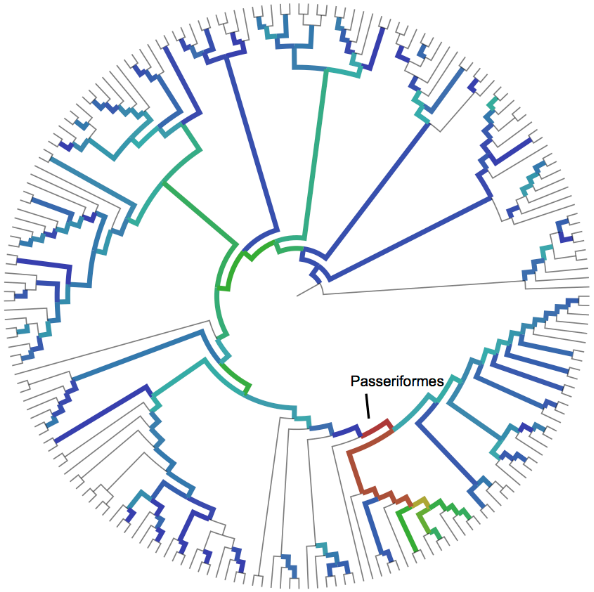


Brown et al. *in revision*

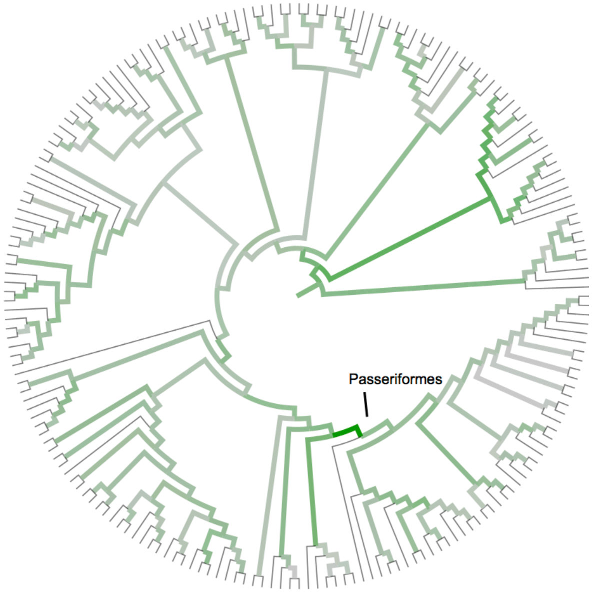


Prum et al. 2016

Conflict

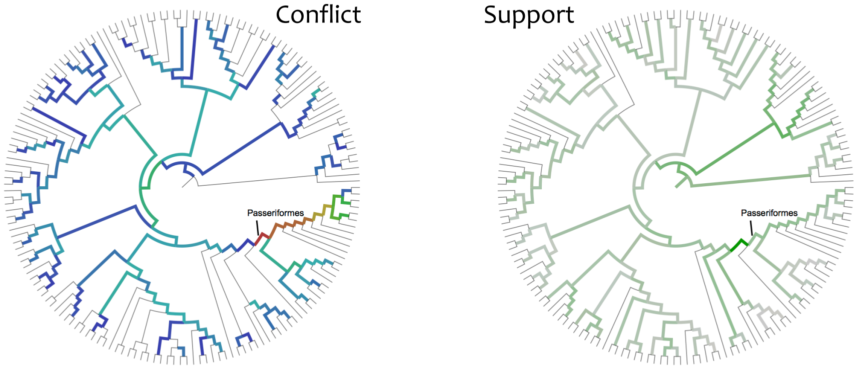


Support

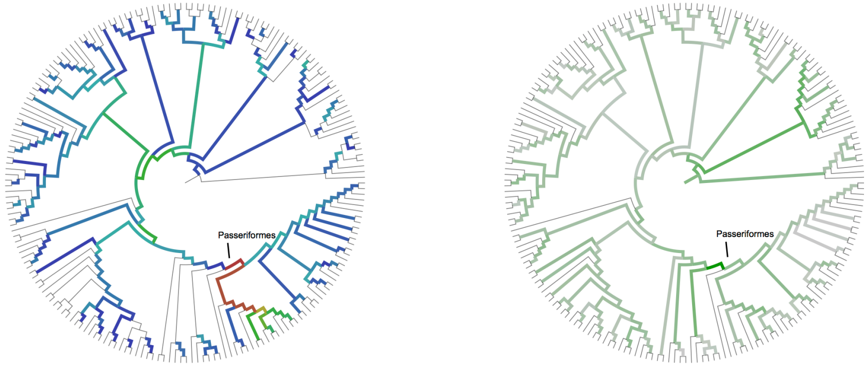


- Support for a clade of Passeriformes
- Large conflict for the early diverging branches

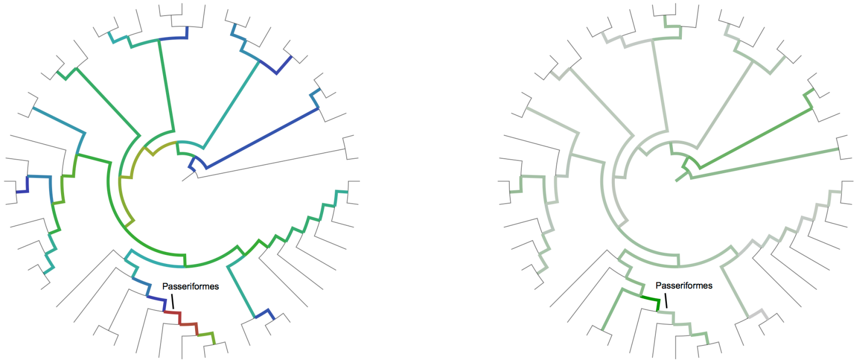
Hackett



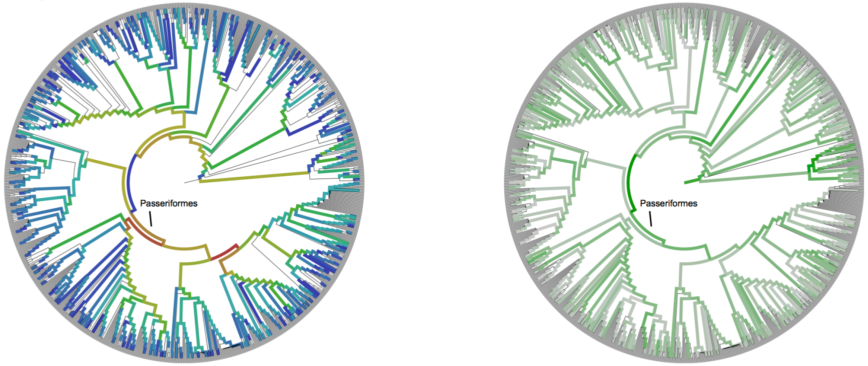
Prum



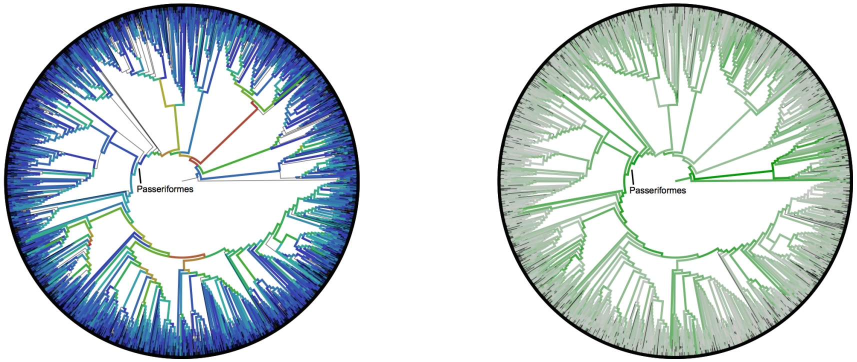
Jarvis



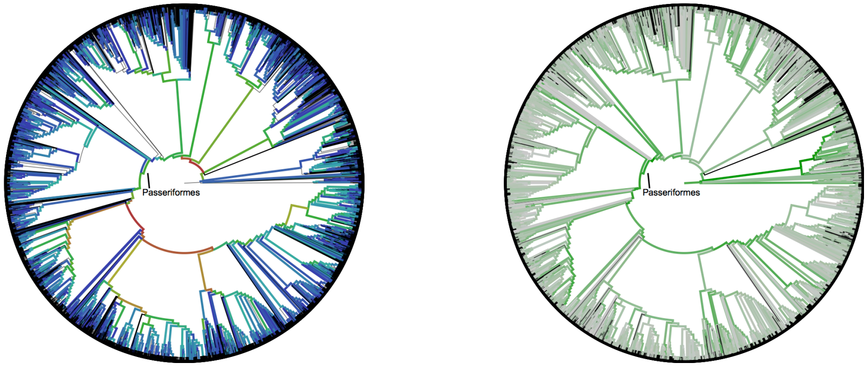
Sibley



Jetz

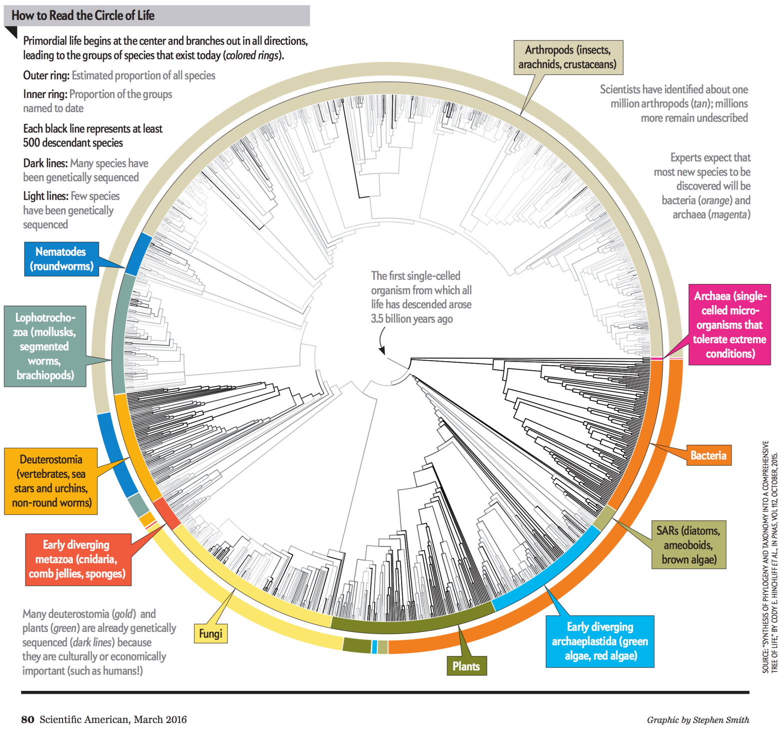


Opentree



Brown et al. *in revision*

Michigan phylogenies



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

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*)

The first single-celled organism from which all life has descended arose 3.5 billion years ago

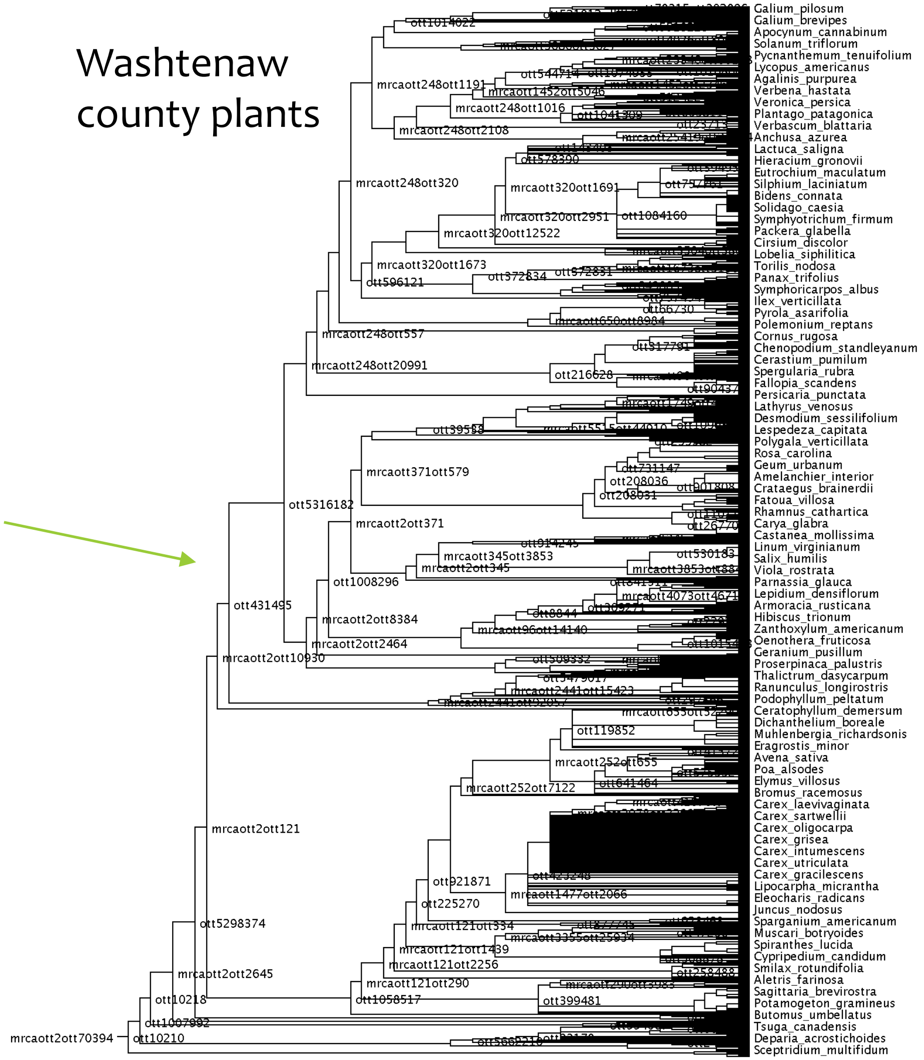
Many deuterostomia (*gold*) and plants (*green*) are already genetically sequenced (*dark lines*) because they are culturally or economically important (such as humans!)

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

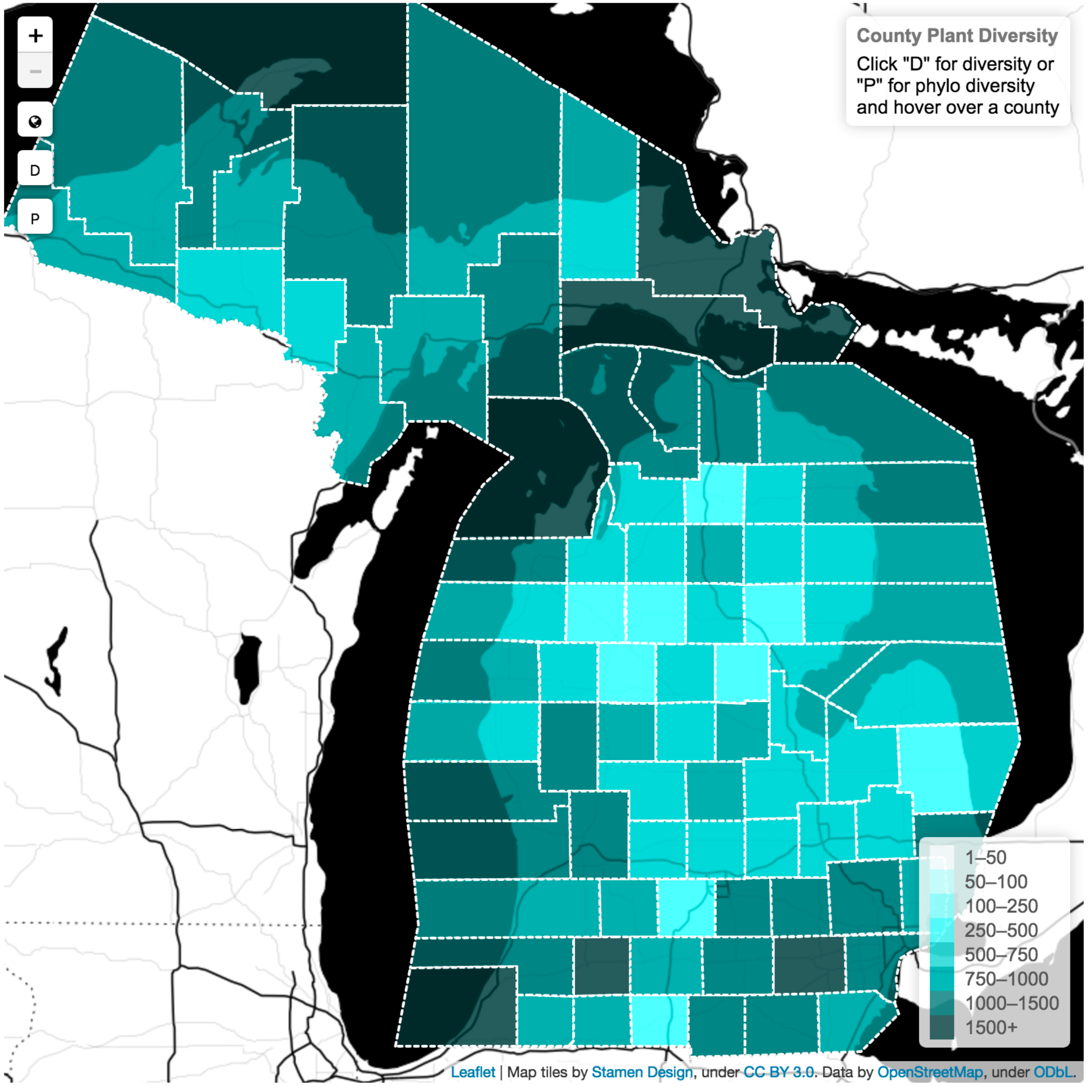
80 Scientific American, March 2016

Graphic by Stephen Smith

Washtenaw county plants



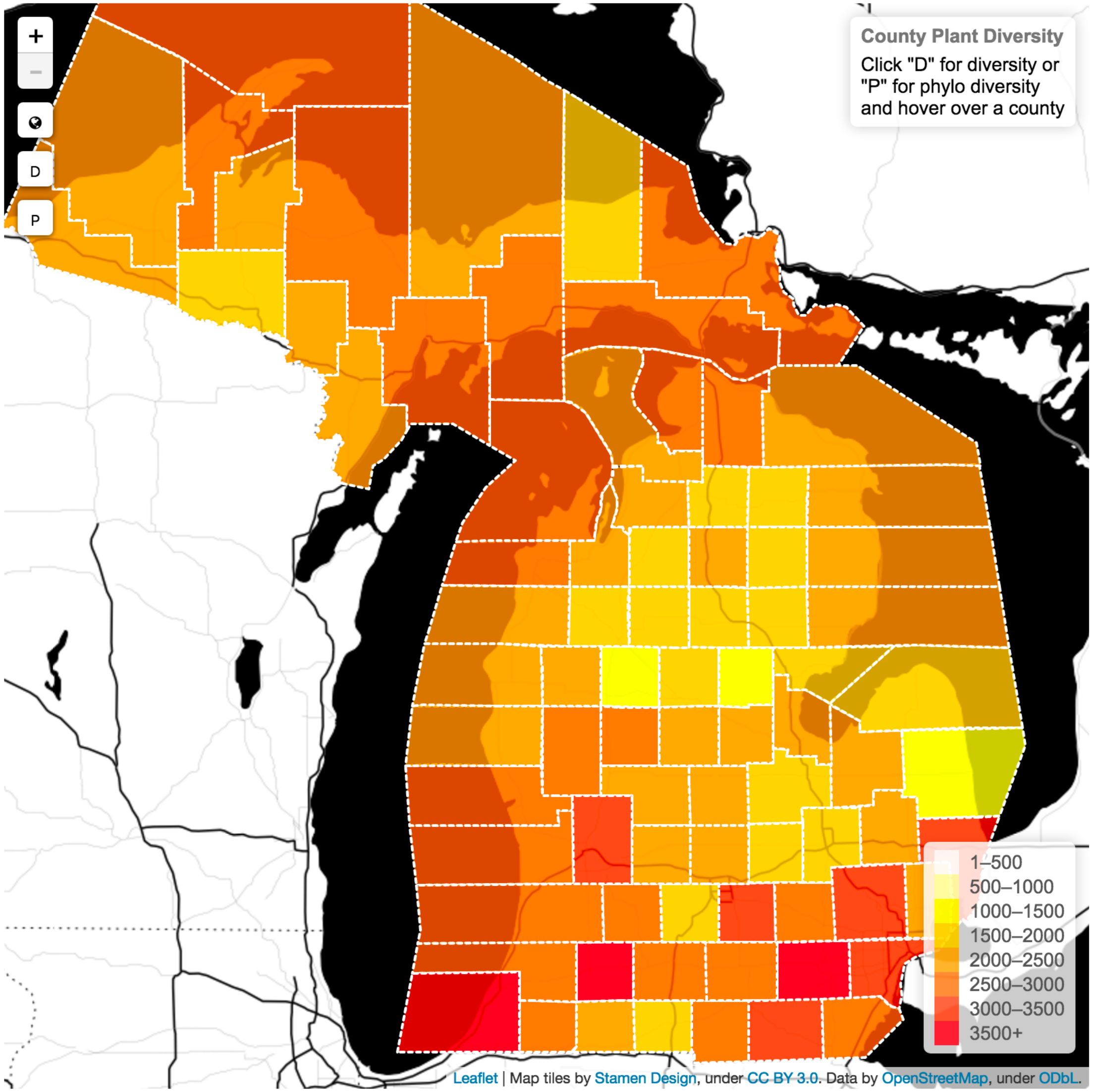
Michigan diversity



County Plant Diversity
Click "D" for diversity or "P" for phylo diversity and hover over a county

Leaflet | Map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under ODbL.

Michigan phylogenetic diversity



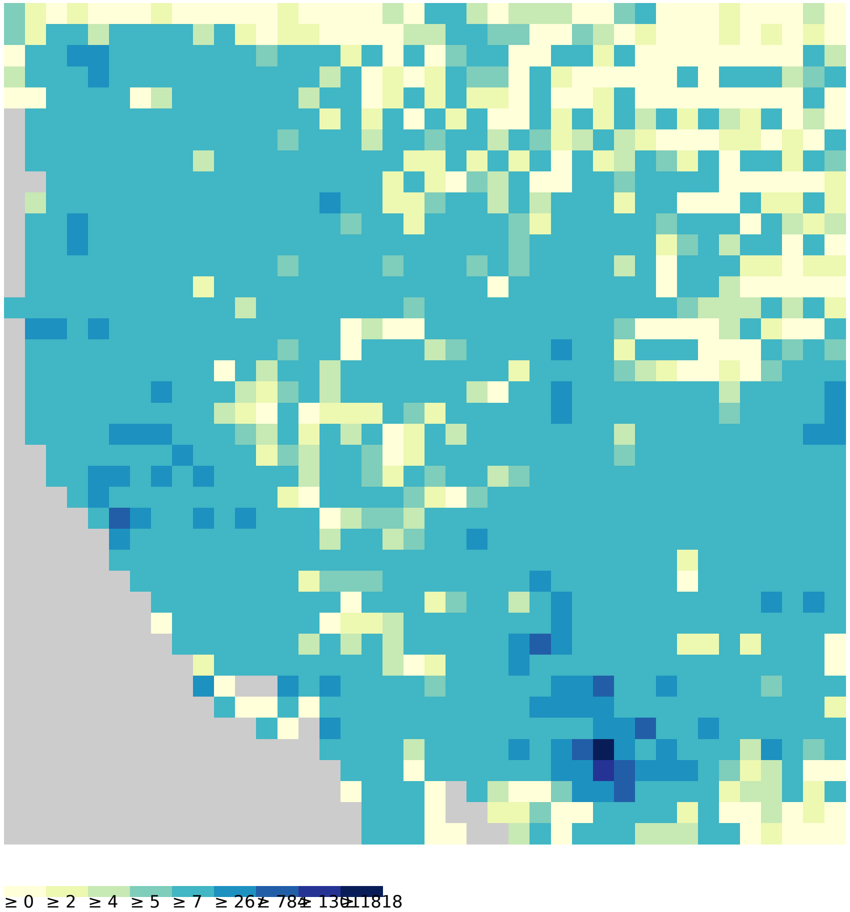
County Plant Diversity

Click "D" for diversity or "P" for phylo diversity and hover over a county

Leaflet | Map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under ODbL.

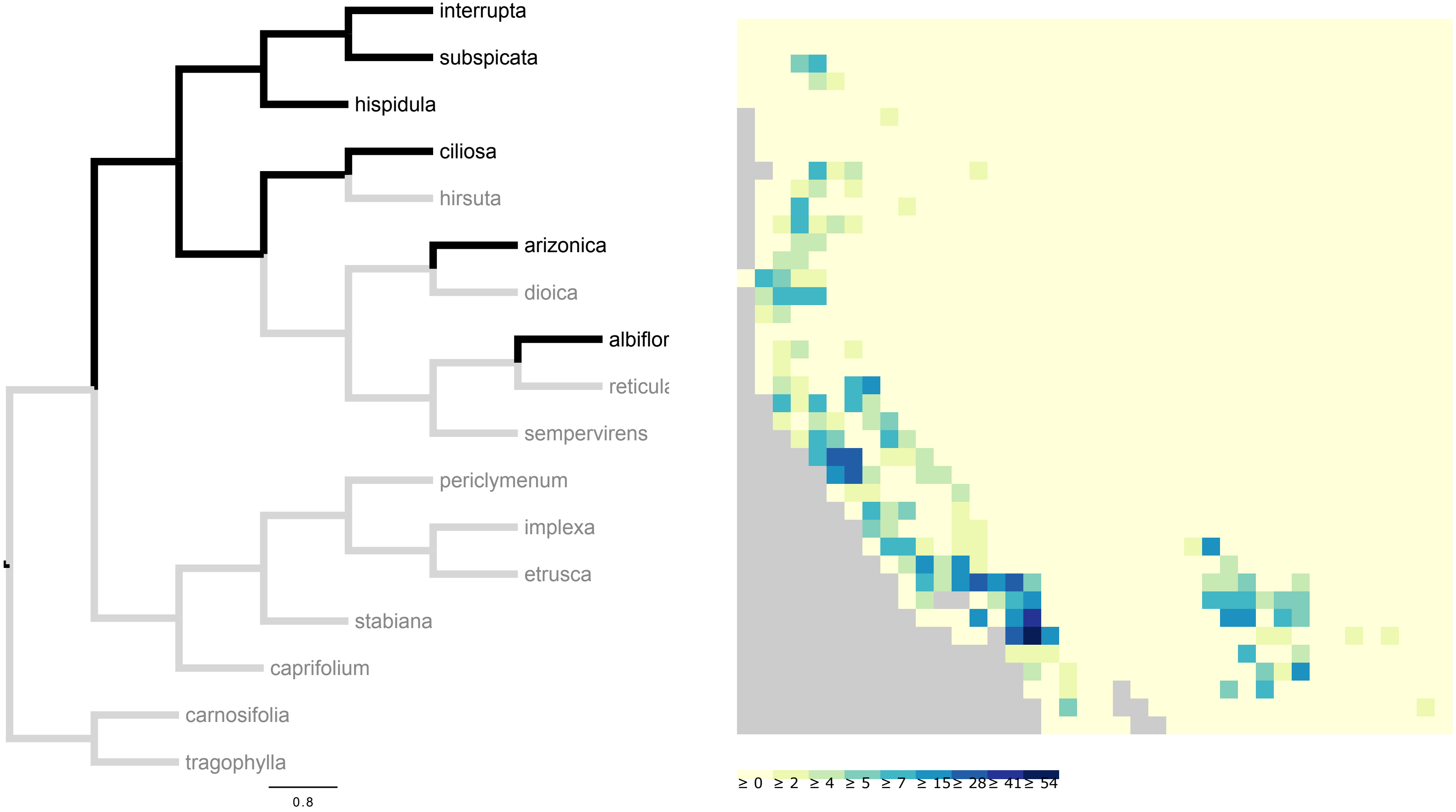
What about biodiversity?

How does diversification relate to rates of biogeographic movement?

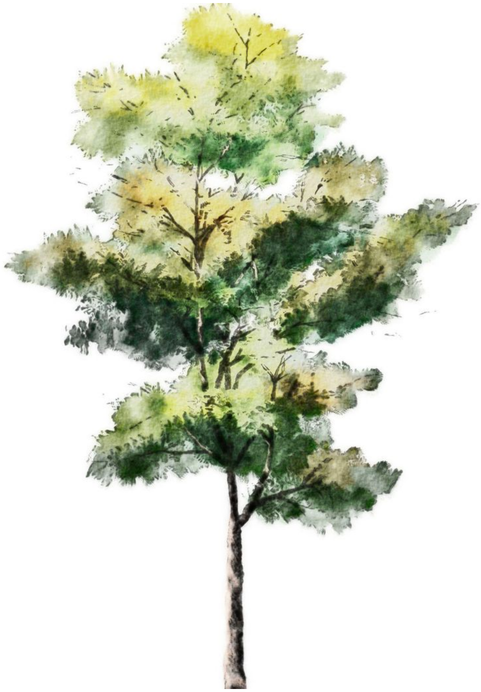


What about biodiversity?

How does diversification relate to rates of biogeographic movement?



Comprehensive and dated tree for seed plants



LET’S SEE A BIG PLANT TREE

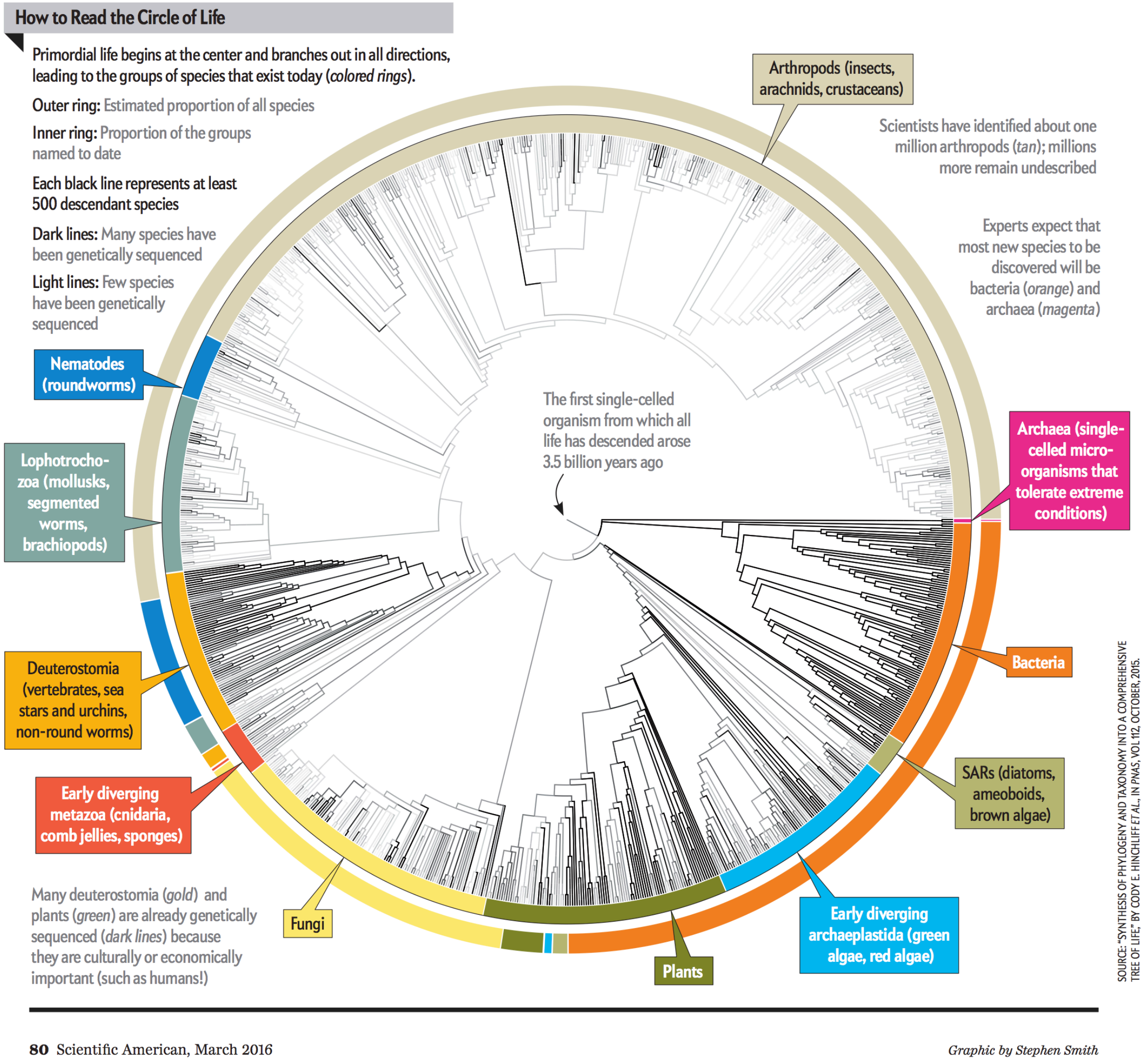
Tree from Open Tree of Life

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



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

The first single-celled organism from which all life has descended arose 3.5 billion years ago

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*)

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

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

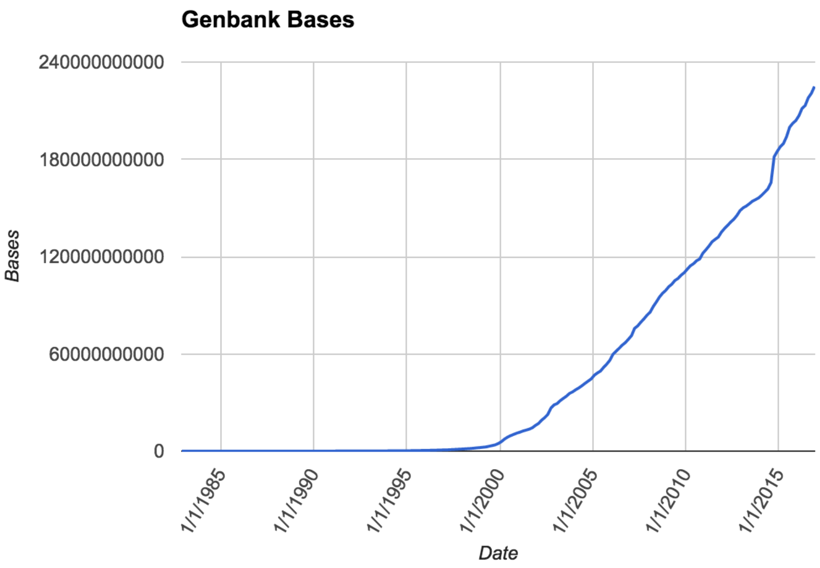
80 Scientific American, March 2016

Graphic by Stephen Smith

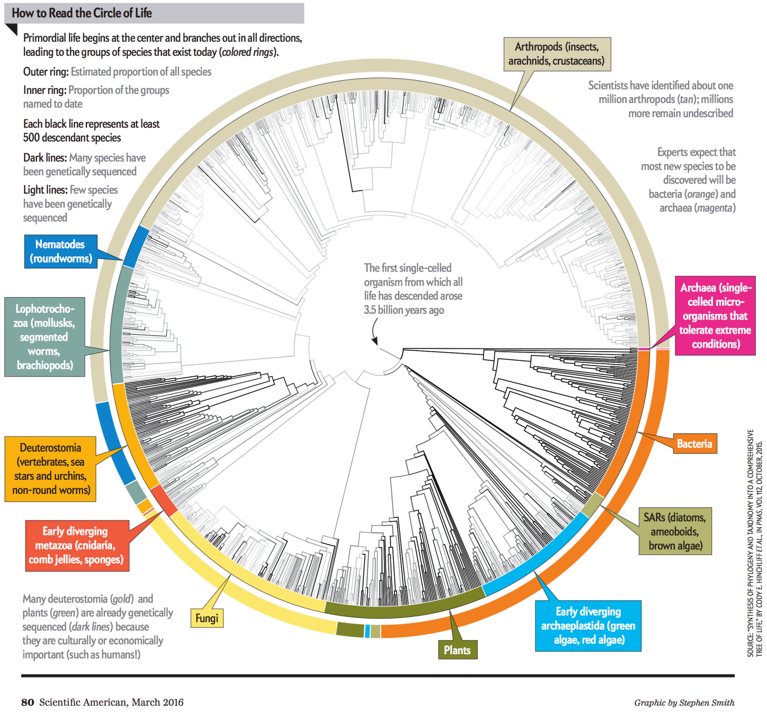
Combine GenBank and Open Tree

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

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



+



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

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*)

The first single-celled organism from which all life has descended arose 3.5 billion years ago

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

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

80 Scientific American, March 2016

Graphic by Stephen Smith

Procedure

Get a list of clades

- We construct trees for monophyletic orders (roughly)

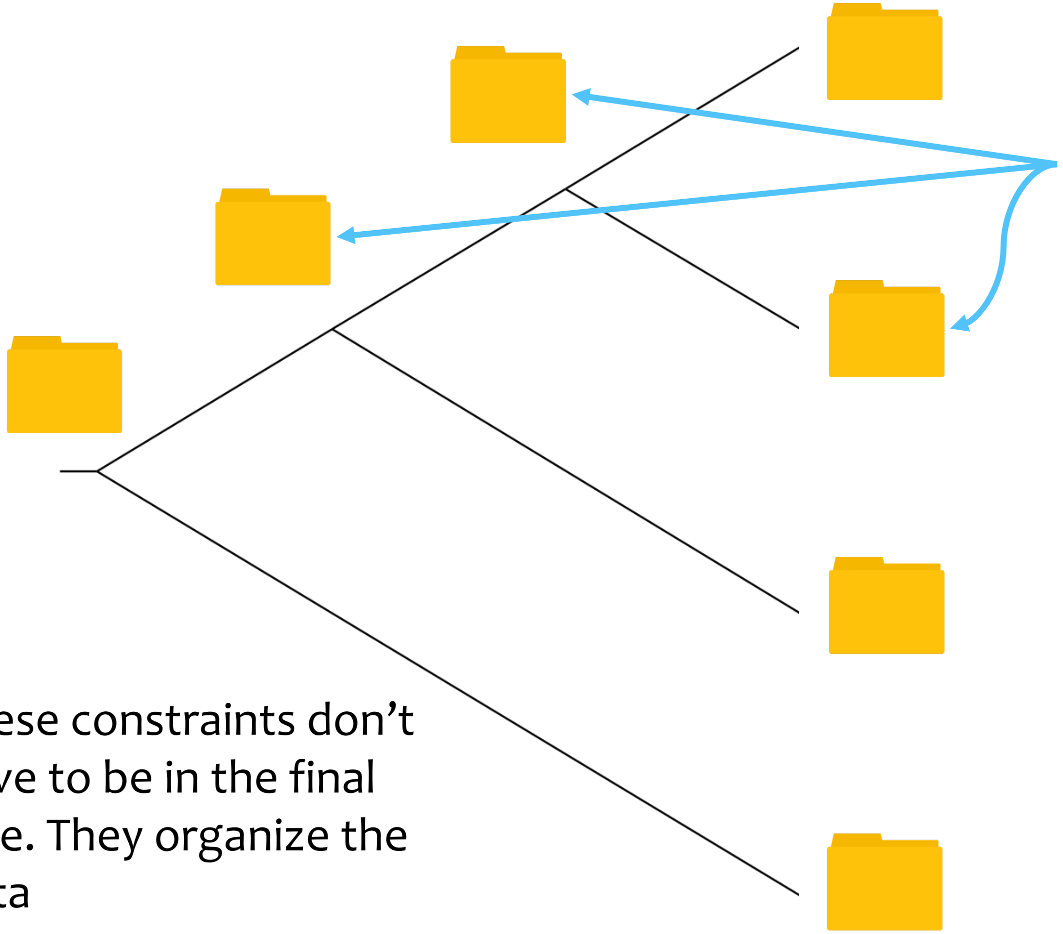
For each clade

- 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 - grad student in my lab)

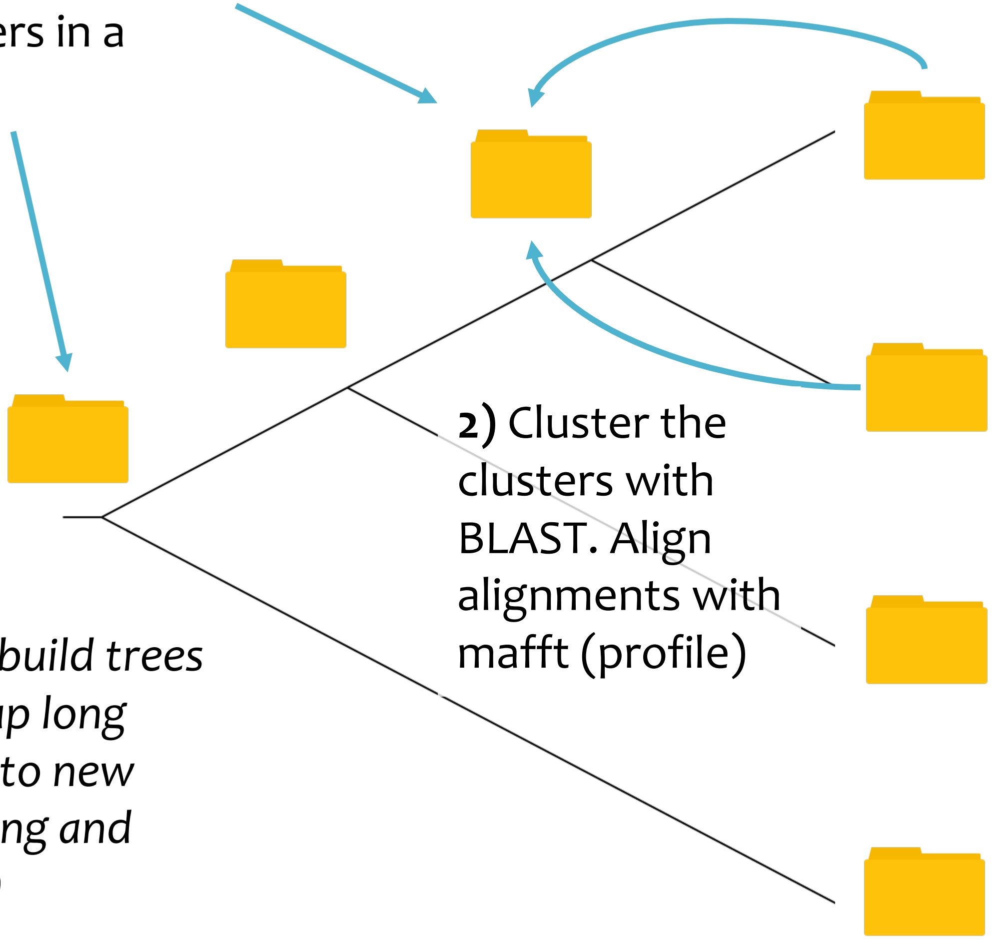
Create folders that replicate the constraint tree



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



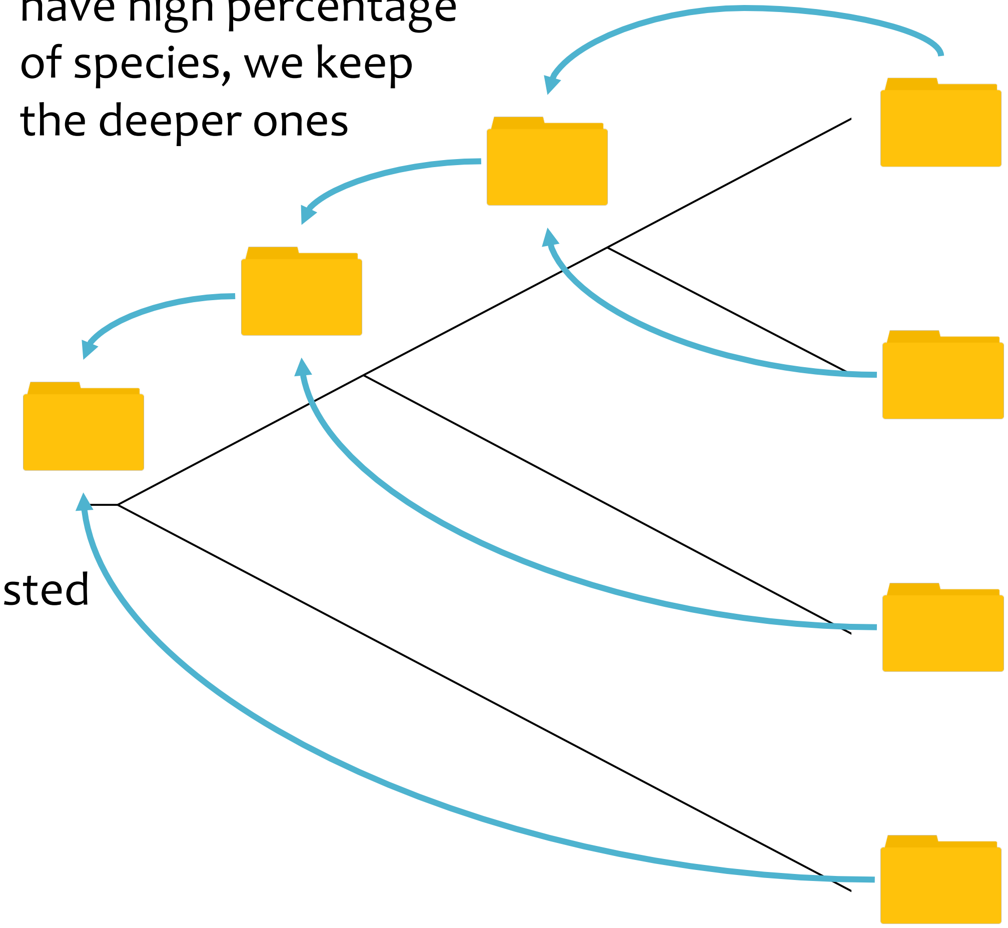
3) At each node in the tree, we record the clusters in a folder

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

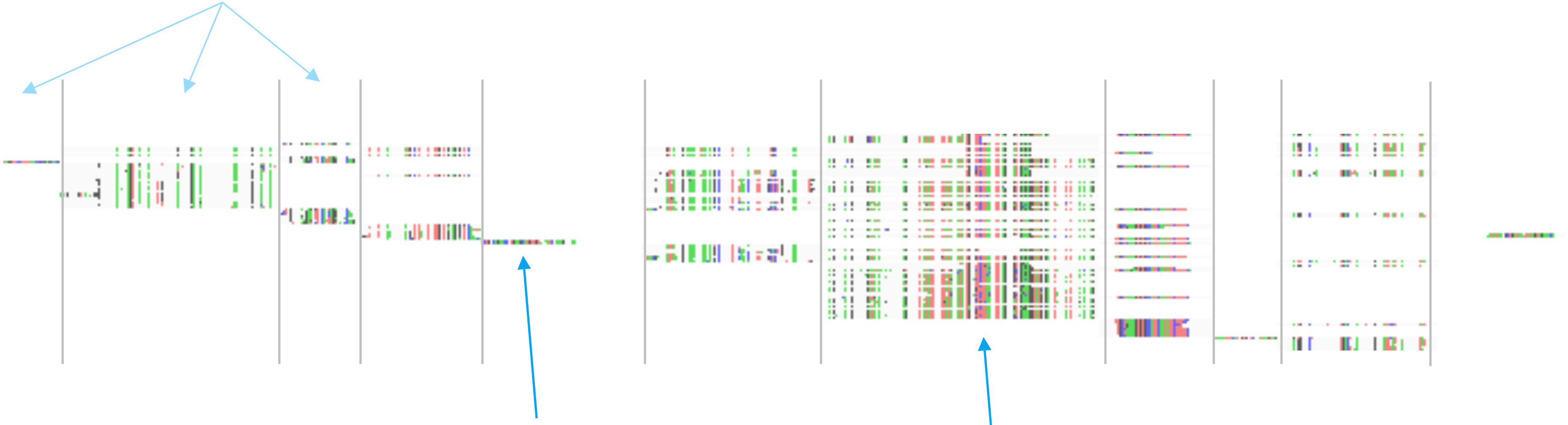


1) Start at the tip folders and get the number of species and those clusters with high percentage of species are kept

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

3) Keep good nested and good deep clusters

An alignment example



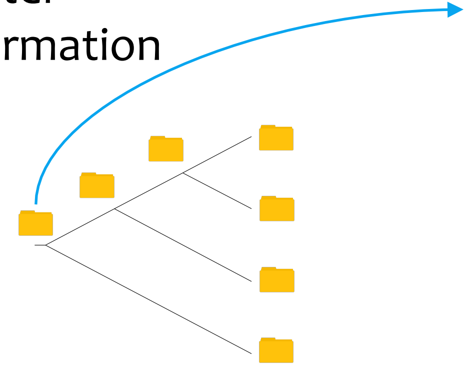
Gene regions

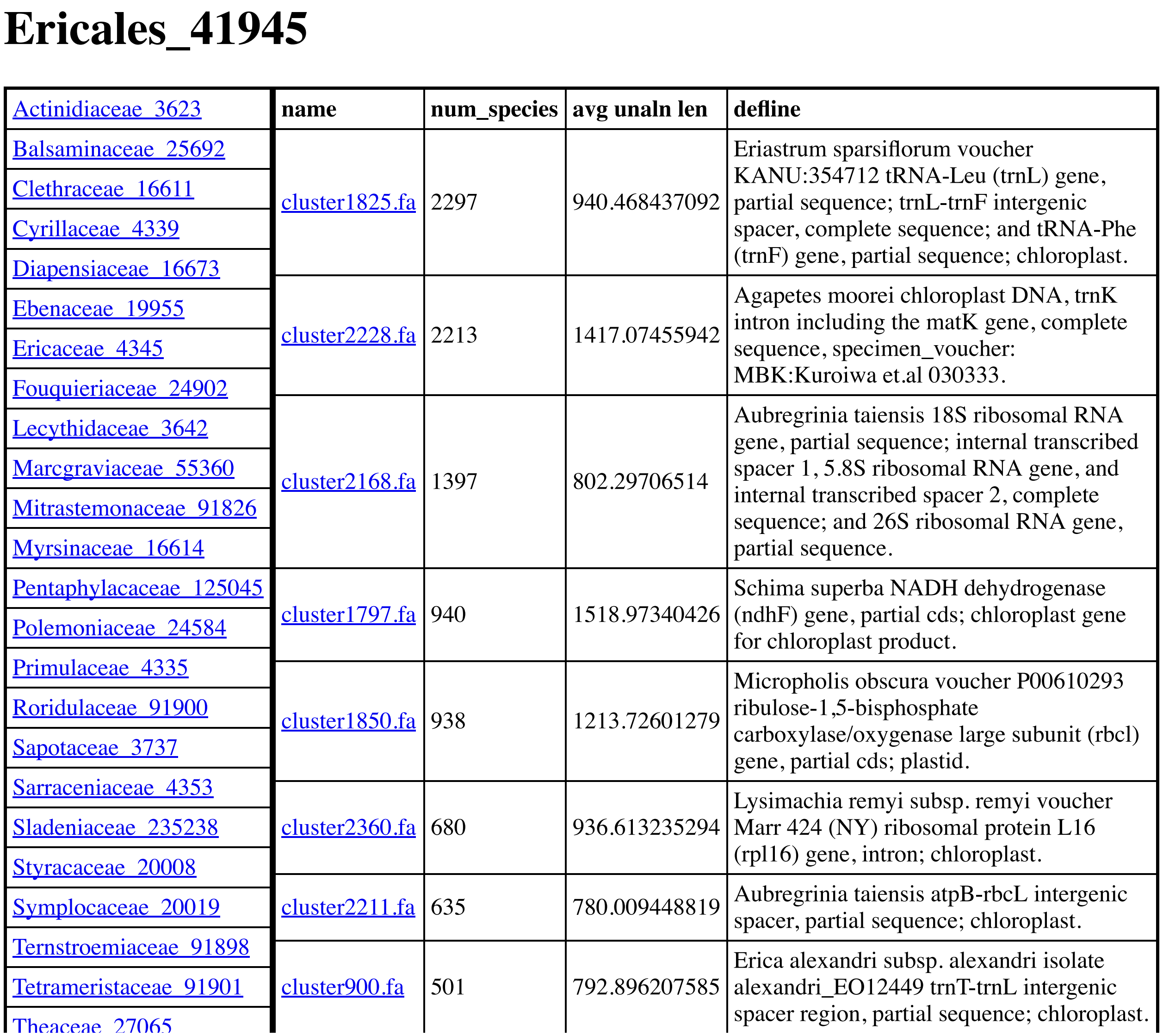
Sampling in one clade

Broader sampling across clades

Website for each node/folder

Each clade has a website with cluster information





Ericales_41945

Actinidiaceae_3623
Balsaminaceae_25692
Clethraceae_16611
Cyrillaceae_4339
Diapensiaceae_16673
Ebenaceae_19955
Ericaceae_4345
Fouquieriaceae_24902
Lecythidaceae_3642
Marcgraviaceae_55360
Mitrastemonaceae_91826
Myrsinaceae_16614
Pentaphylacaceae_125045
Polemoniaceae_24584
Primulaceae_4335
Roridulaceae_91900
Sapotaceae_3737
Sarraceniaceae_4353
Sladeniaceae_235238
Styracaceae_20008
Symplocaceae_20019
Ternstroemiaceae_91898
Tetrameristaceae_91901
Theaceae_27065

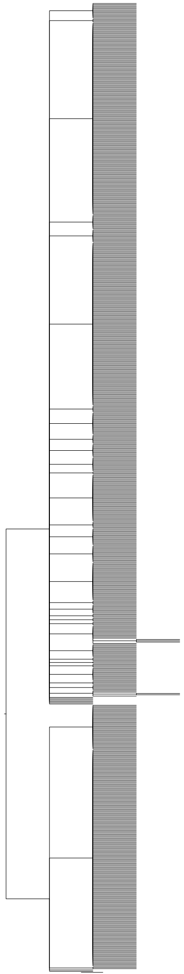
name	num_species	avg unaln len	defline
cluster1825.fa	2297	940.468437092	Eriastrum sparsiflorum voucher KANU:354712 tRNA-Leu (trnL) gene, partial sequence; trnL-trnF intergenic spacer, complete sequence; and tRNA-Phe (trnF) gene, partial sequence; chloroplast.
cluster2228.fa	2213	1417.07455942	Agapetes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

Assume monophyly until proven otherwise

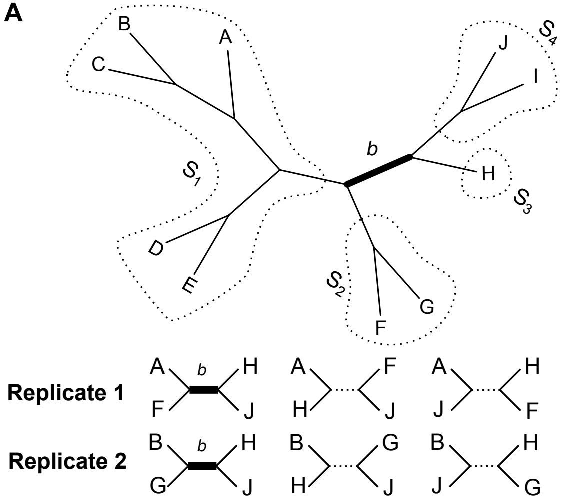
- Some edges are *impossible* to reconstruct without genomic data or otherwise difficult to assemble datasets
- Can we assume that things are monophyletic?
  - Certain clades?
  - Major Linnaean groups? Orders, Families, Genera?
- Do we have to confirm this every run?

Constraint procedure

1) Calculate phylogeny with a constraint tree based on taxonomy

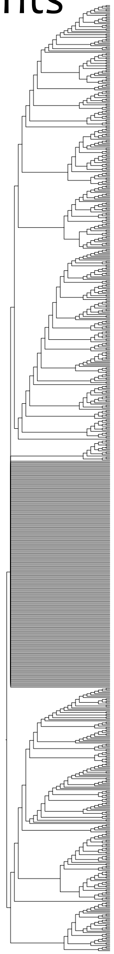


2) Test constraints using a quartet procedure (when there is strong conflict, constraints are removed

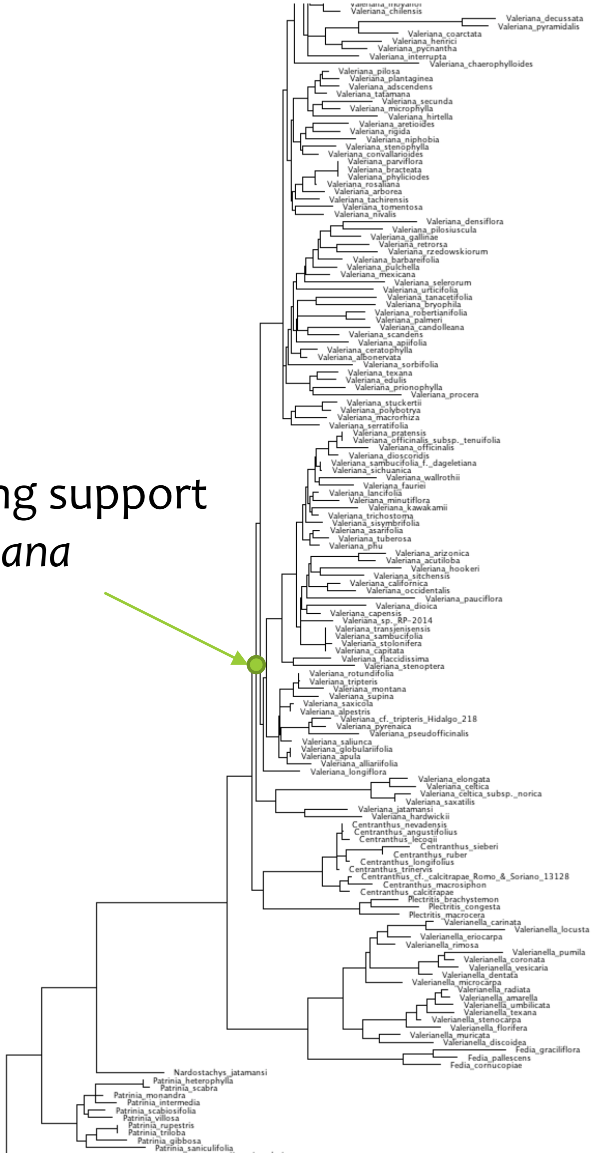


Pease et al. *submitted*

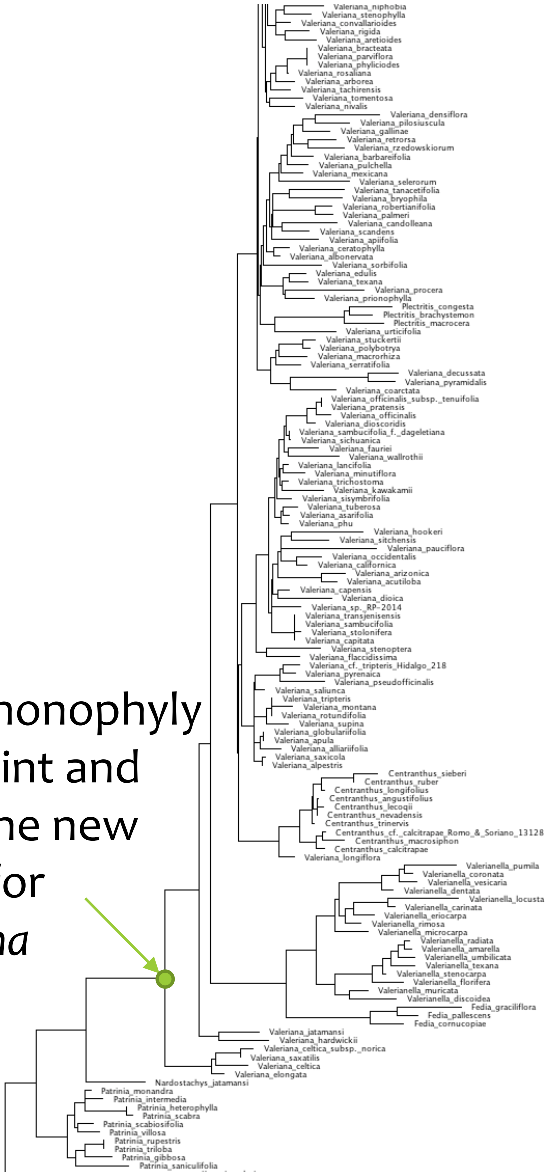
2) Recalculate the tree with the constraint tree as the ML tree with constraints removed



Example with Dipsacales

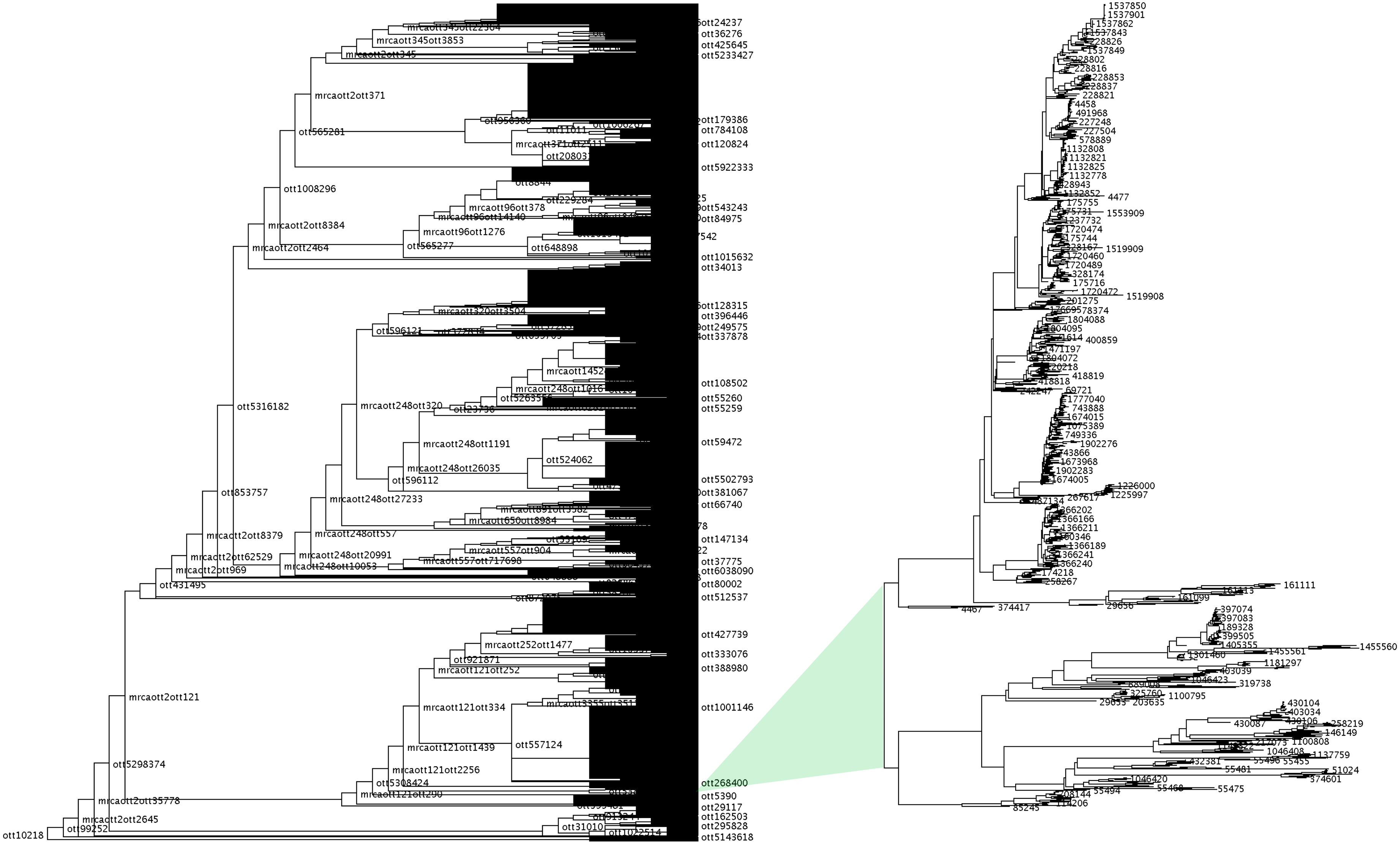


Conflicting support for *Valeriana*



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

Place clades in the backbone/synth tree



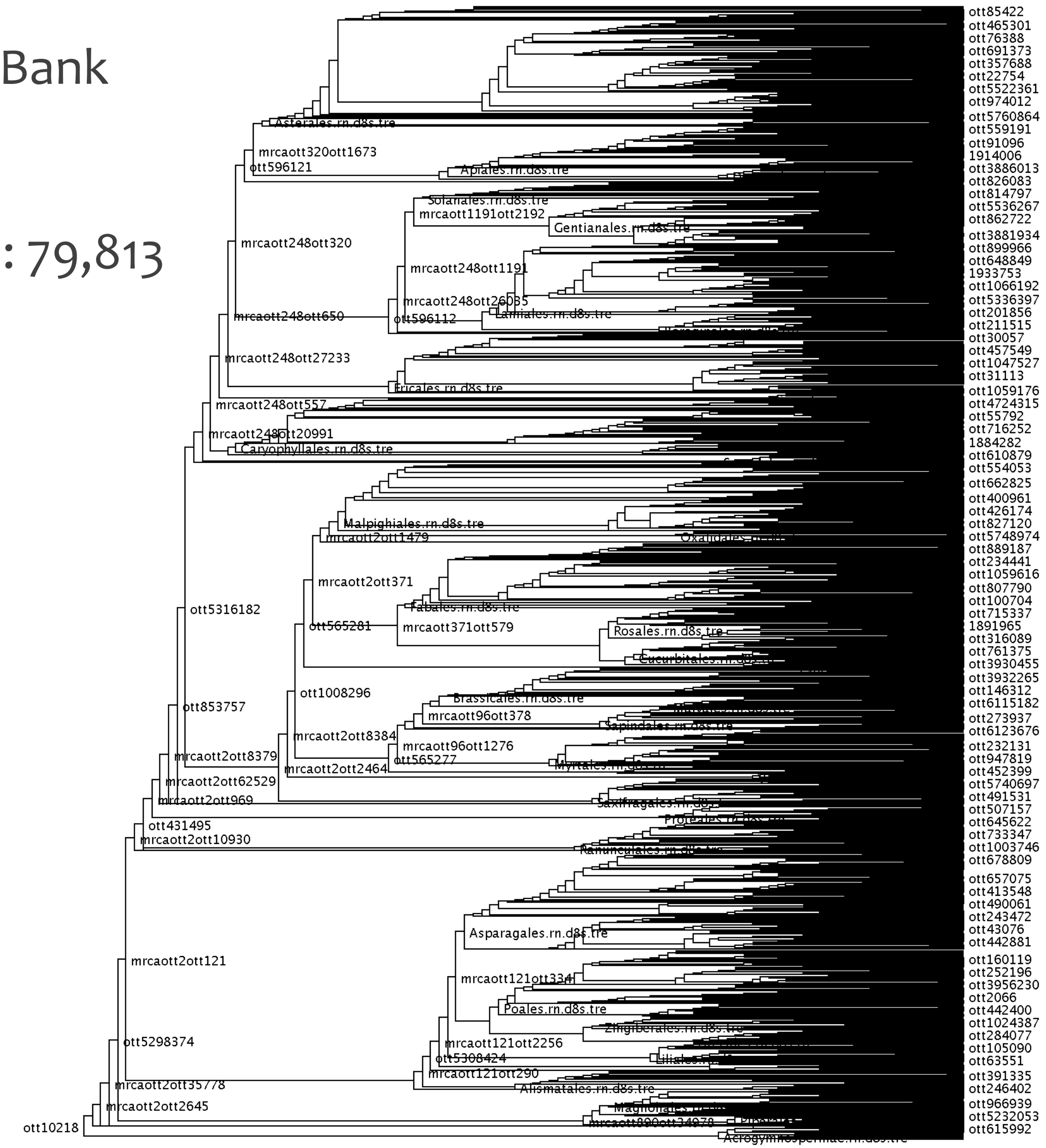
What does the data look like?

79,759 sampled from GenBank

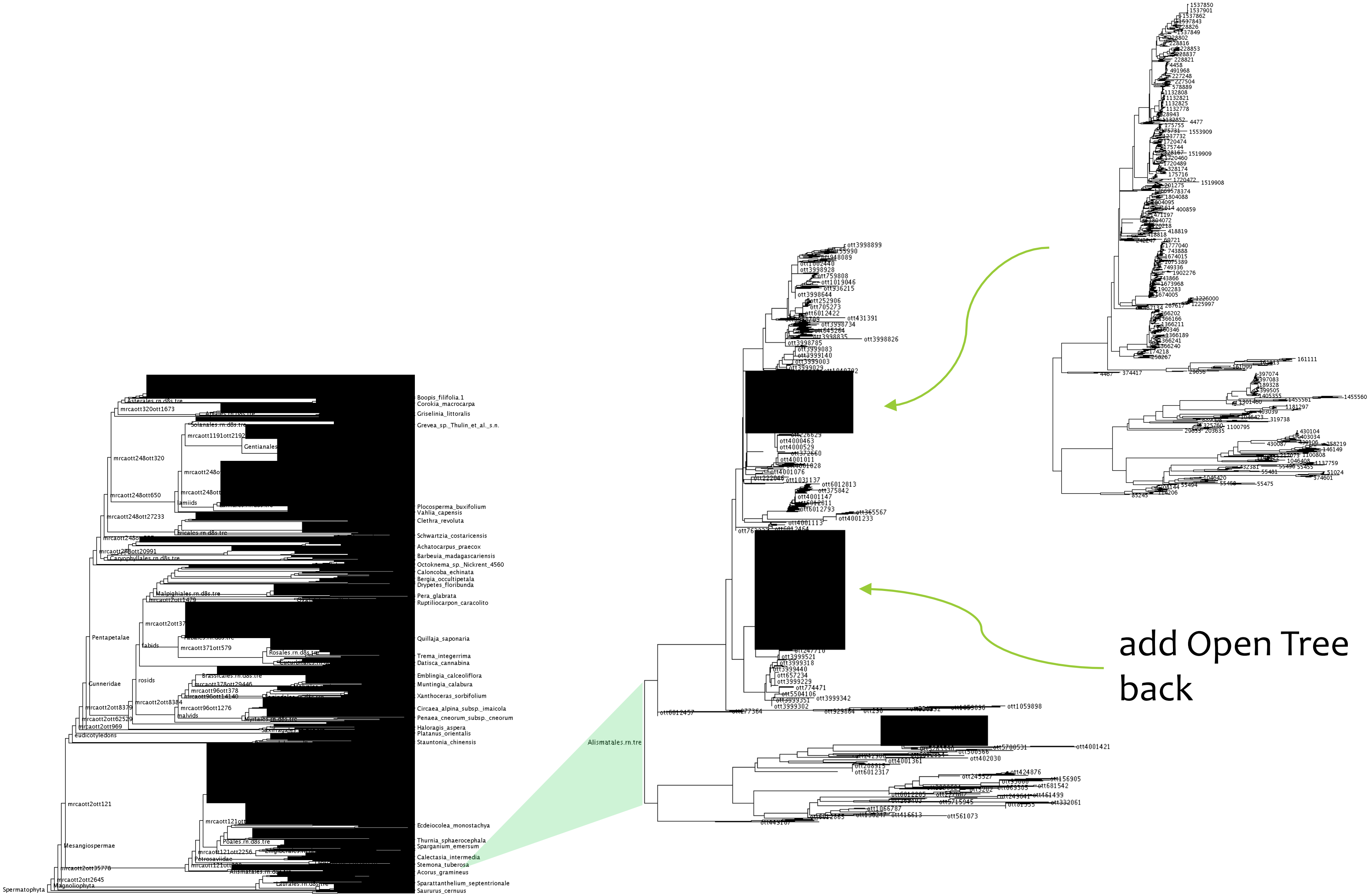
Number of tips: 79,875

Number of internal nodes: 79,813

61 unresolved nodes



Place clades in the backbone/synth tree



add Open Tree back

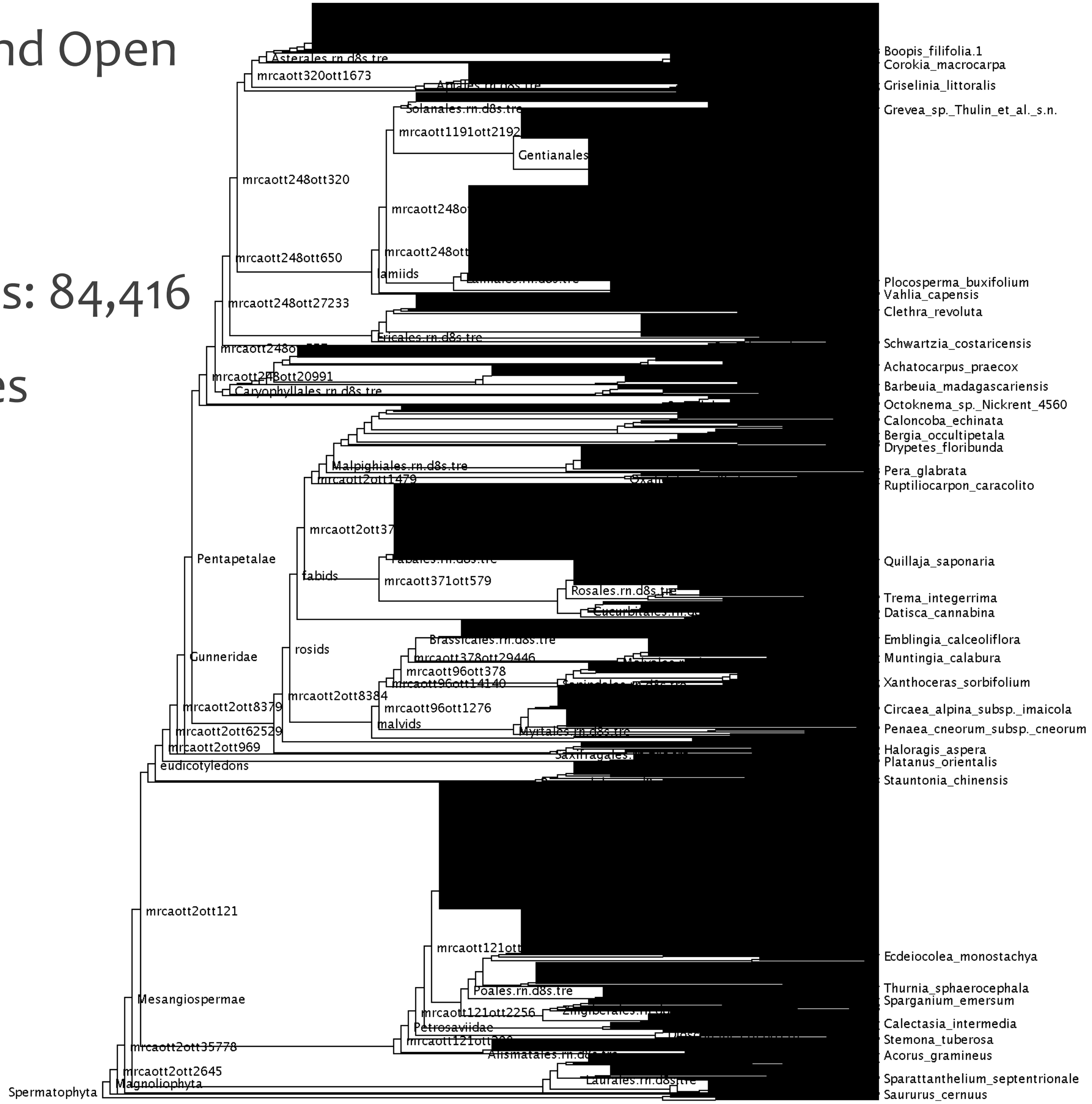
What does the data look like?

356,807 with GenBank and Open Tree taxonomy

Number of tips: 356,807

Number of internal nodes: 84,416

272,390 unresolved nodes



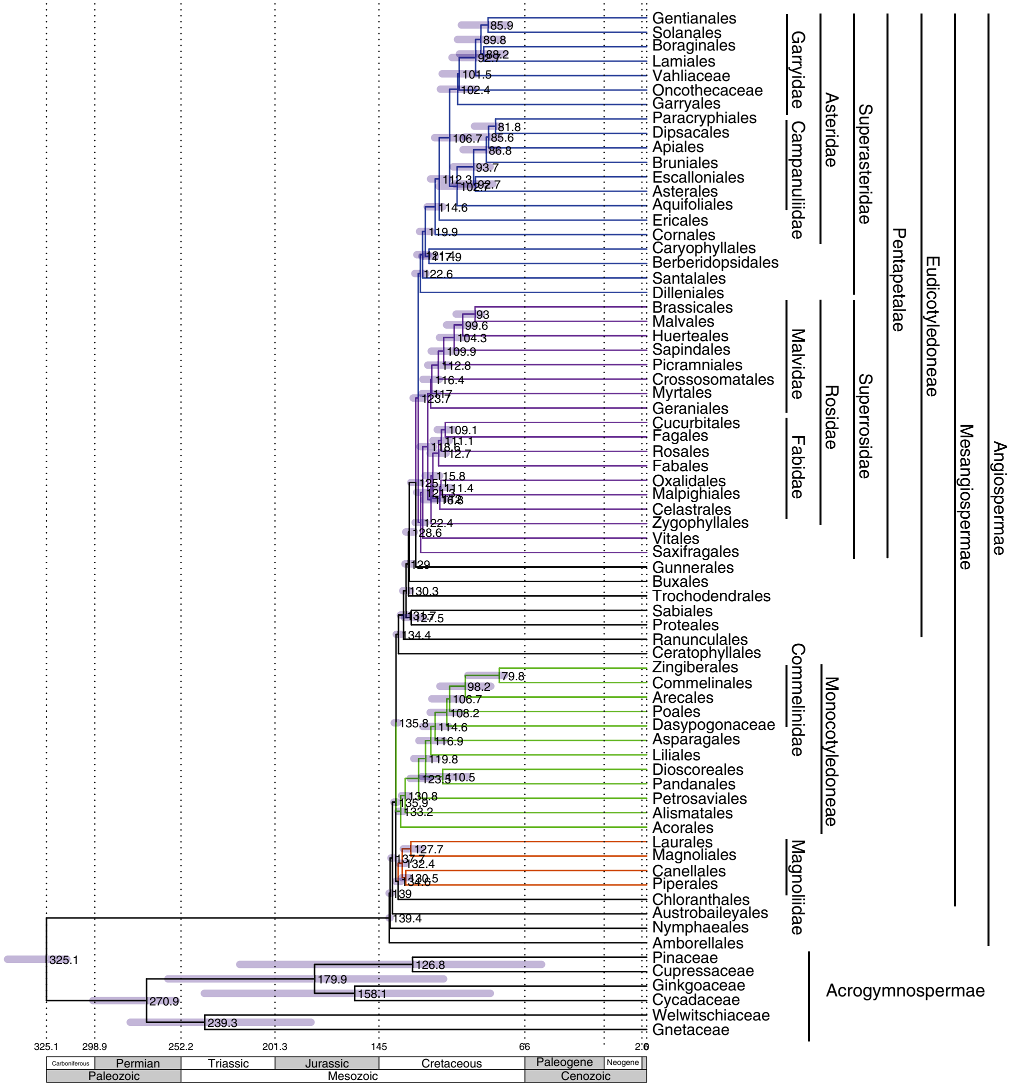
Divergence time estimation

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

Details:

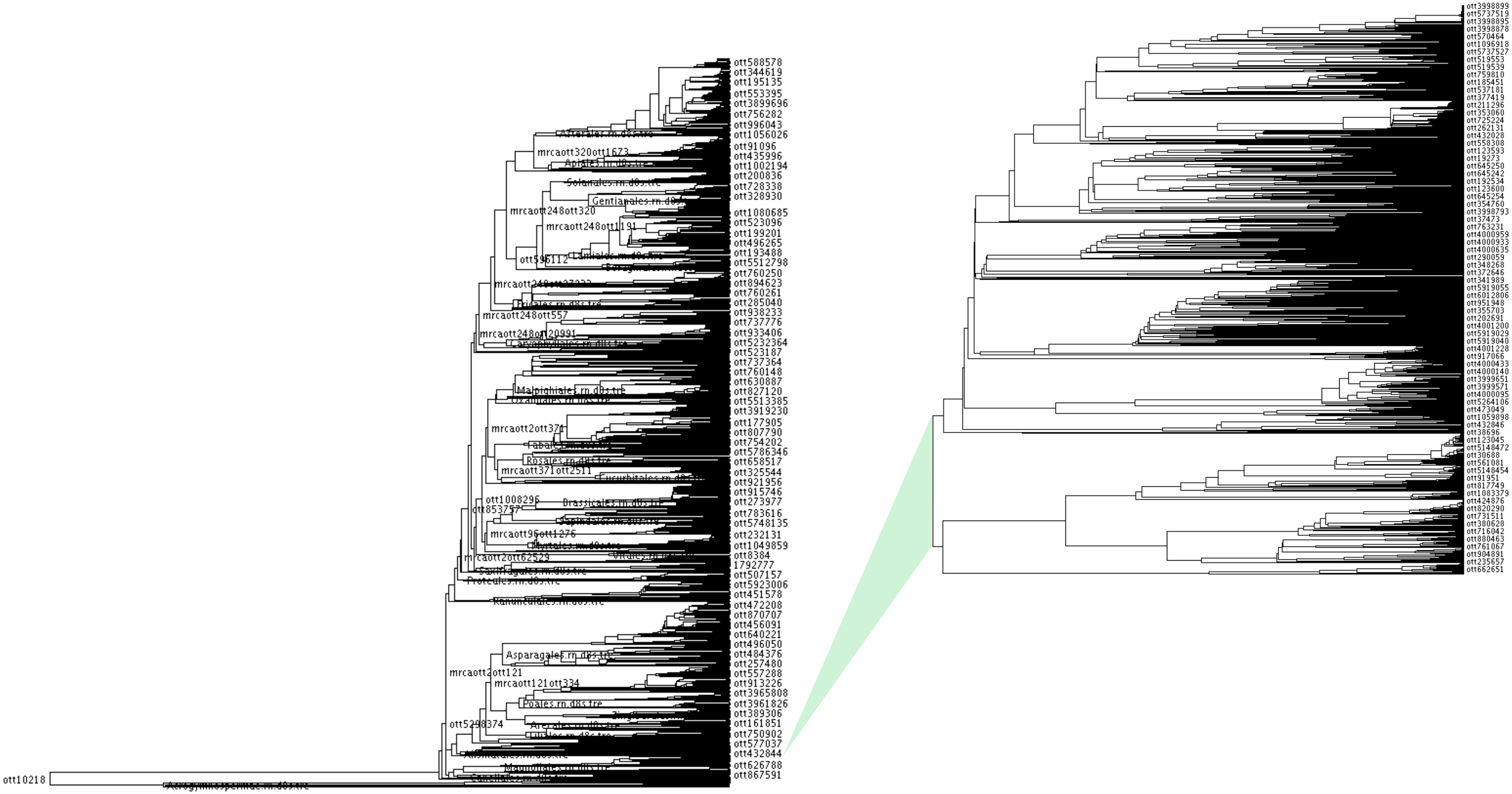
We use hard constraints for clades that overlap with Magallon

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

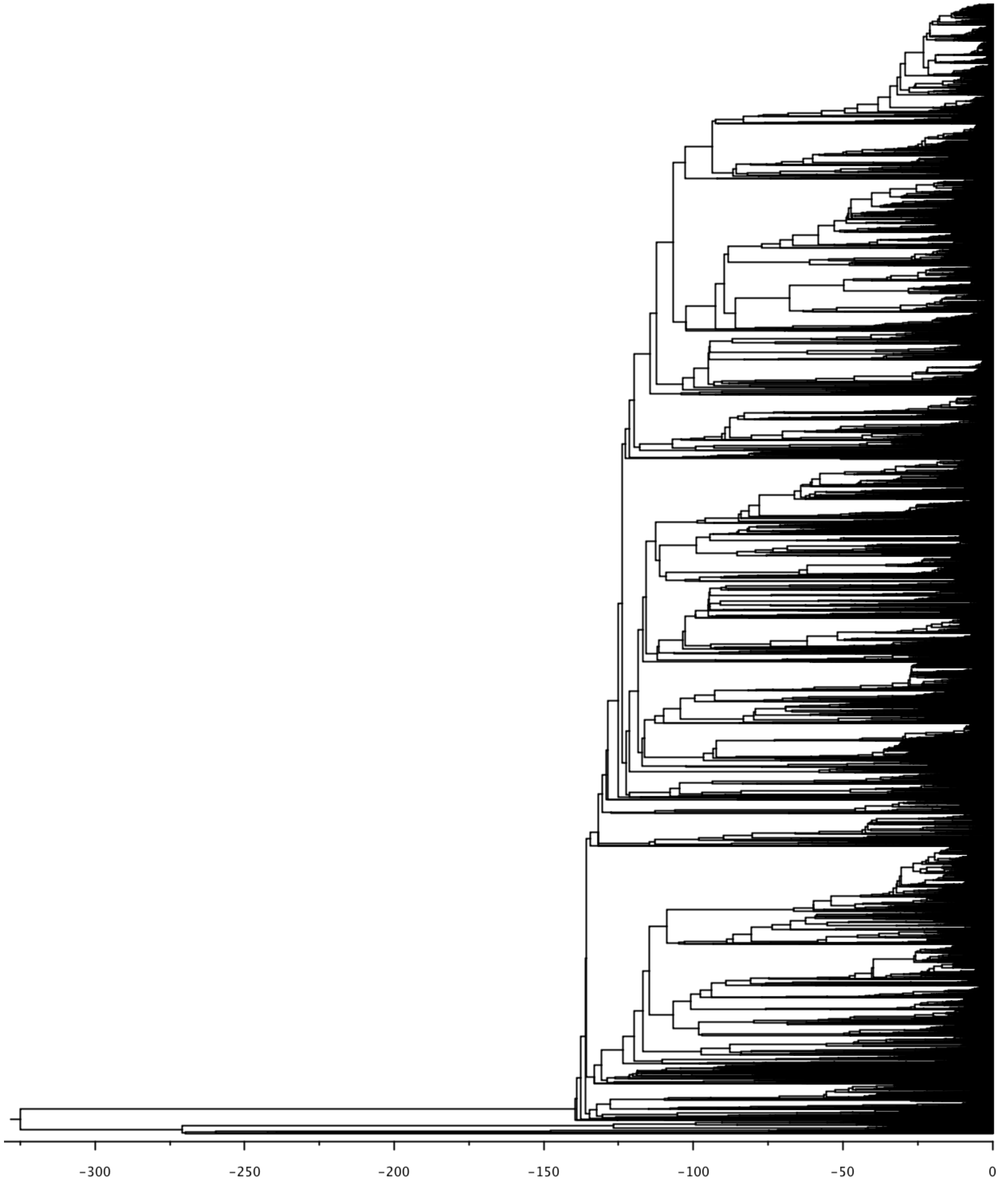


Magallon et al. 2015

Dated Alismatales



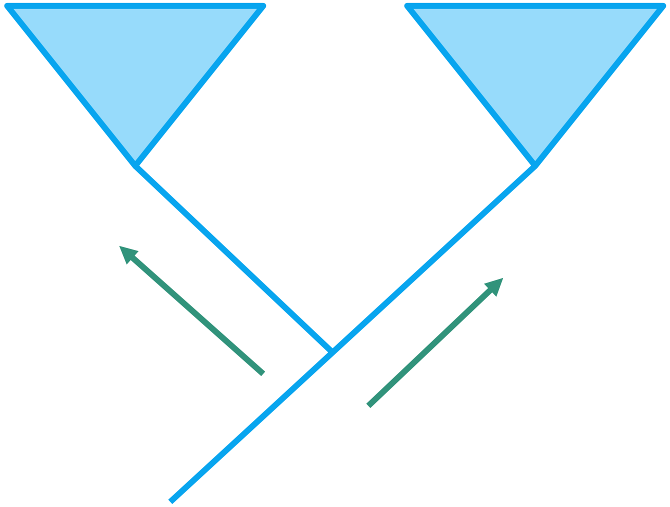
With divergence times



With divergence times and Open Tree taxa

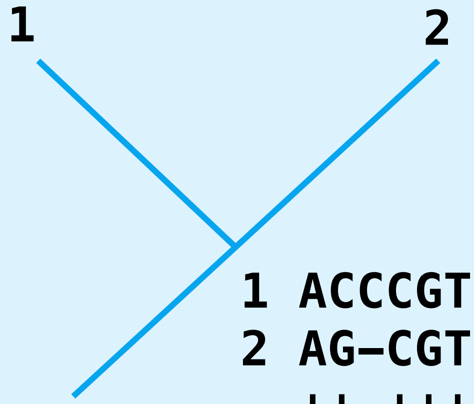
How about data overlap

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



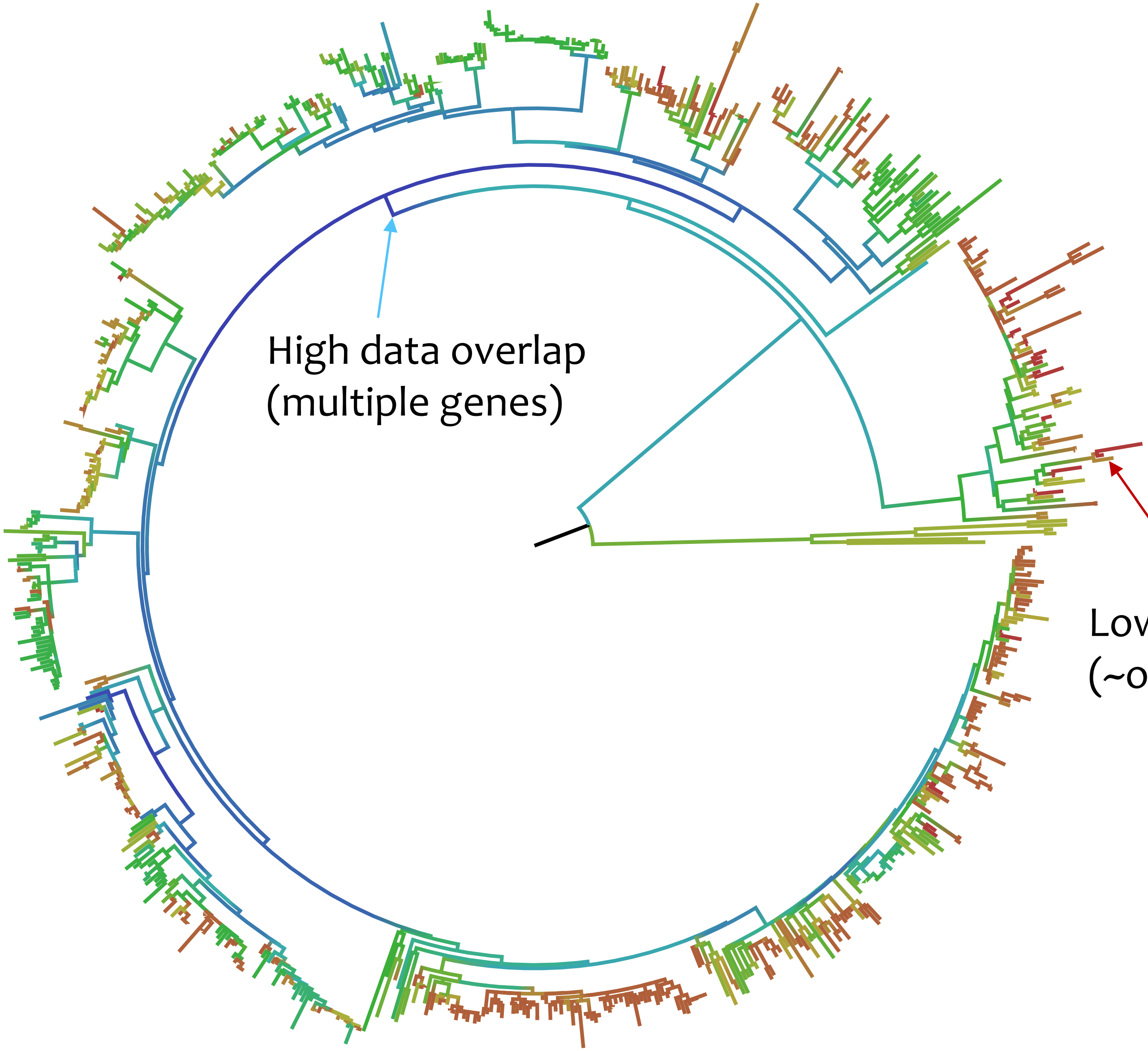
How many sites overlap between any taxa on the left and any taxa on the right?

Example



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

Arecales data overlap

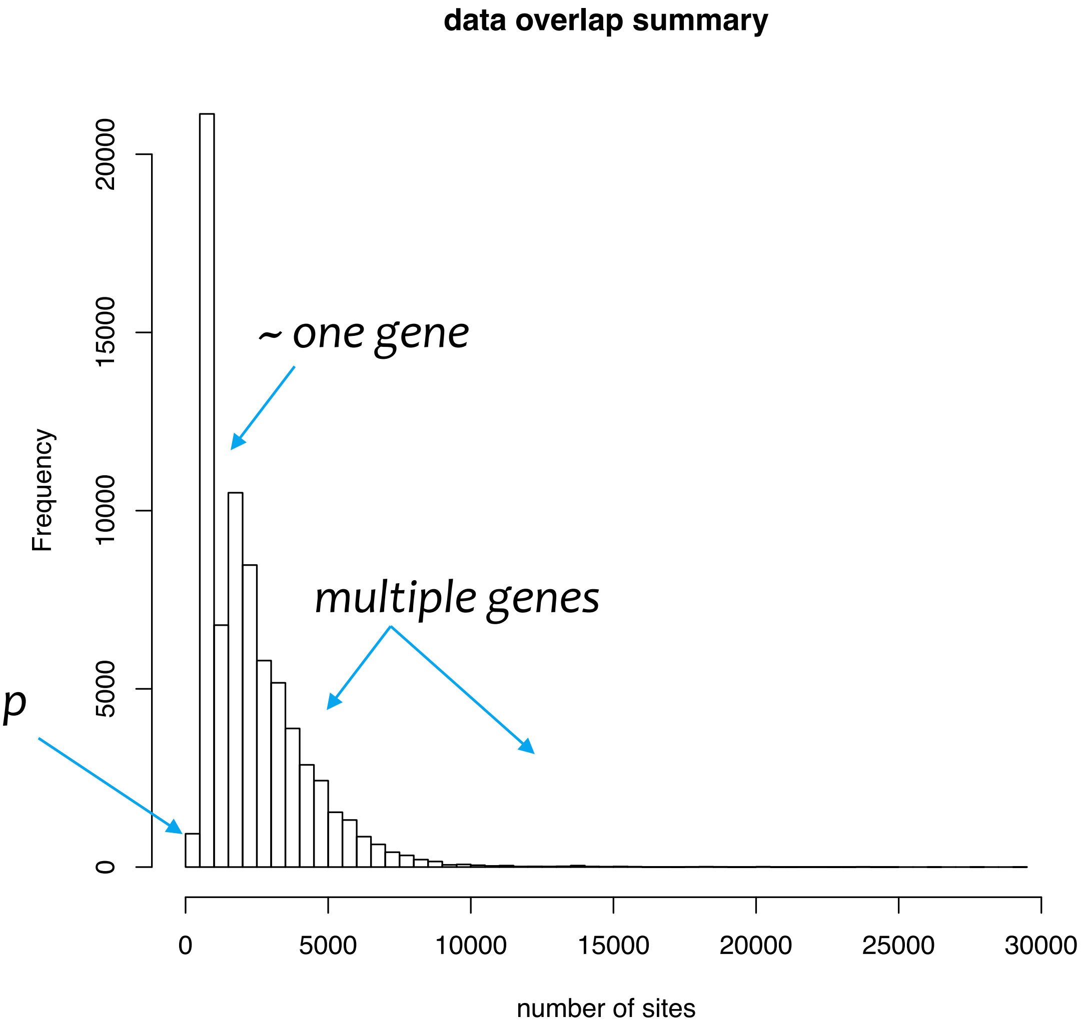


High data overlap
(multiple genes)

Low data overlap
(~one gene)

Summary

Minimum:	0
Median:	1792
Mean:	2339.7
Maximum:	29229



~ one gene

multiple genes

no overlap

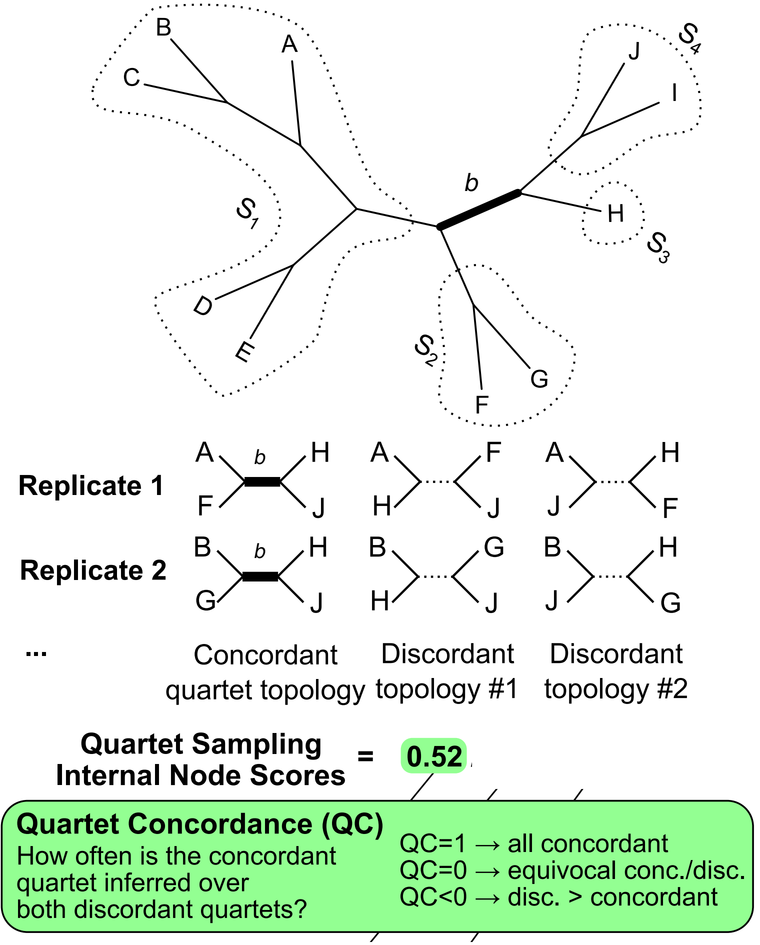
How about support?

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. (submitted)

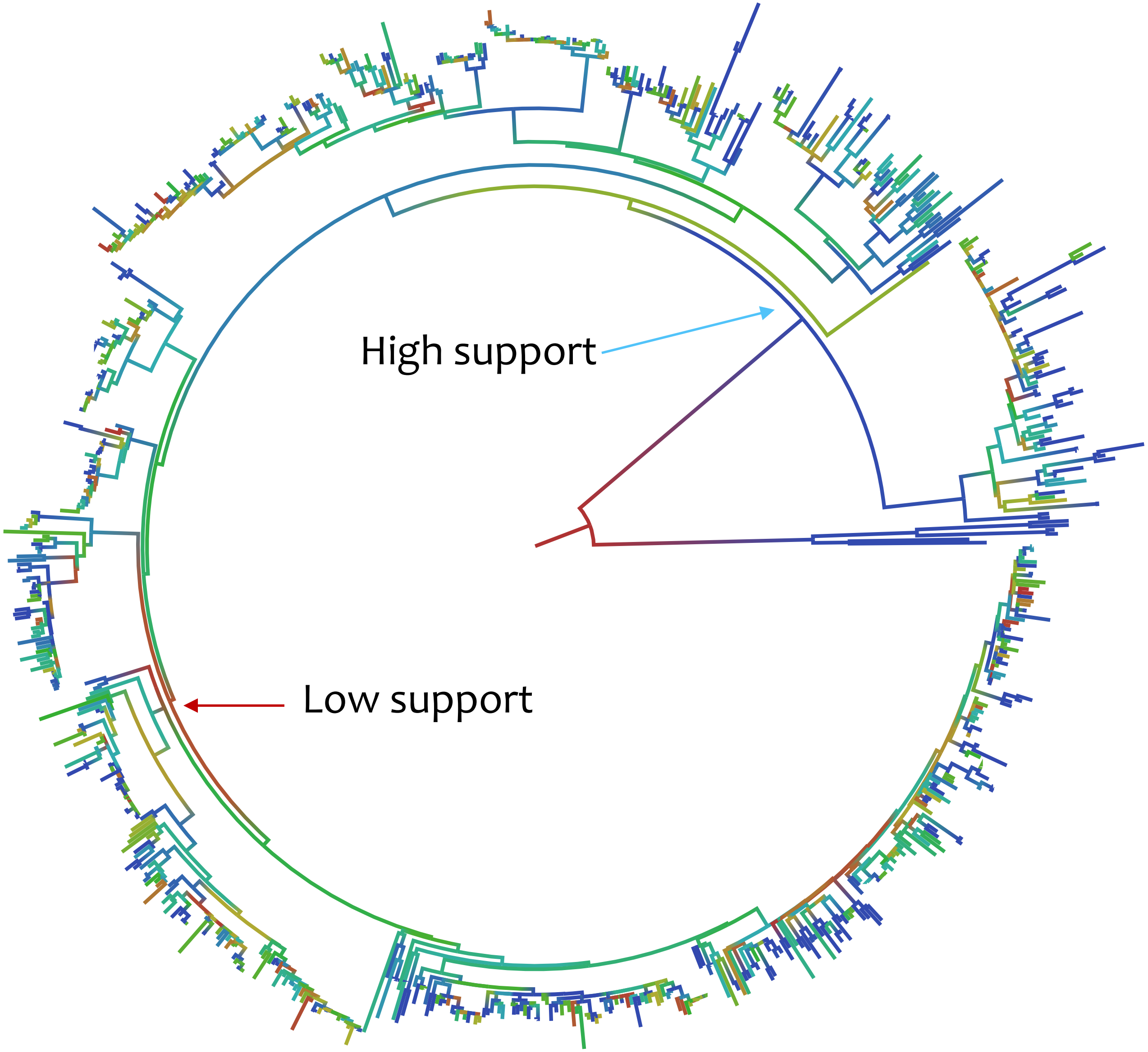


Quartet Sampling Internal Node Scores = **0.52**

Quartet Concordance (QC)
How often is the concordant quartet inferred over both discordant quartets?

QC=1 → all concordant
QC=0 → equivocal conc./disc.
QC<0 → disc. > concordant

Arecales support

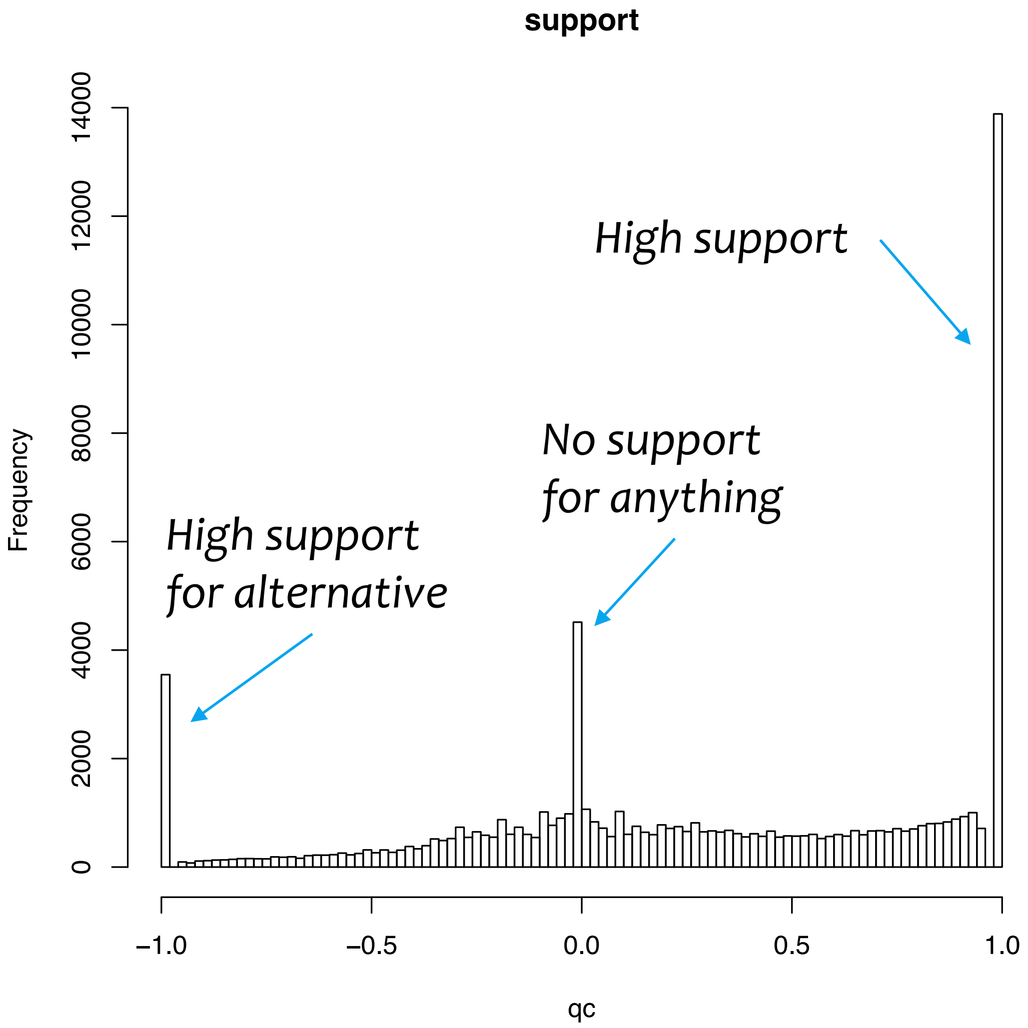


High support

Low support

QC support distribution

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



High support

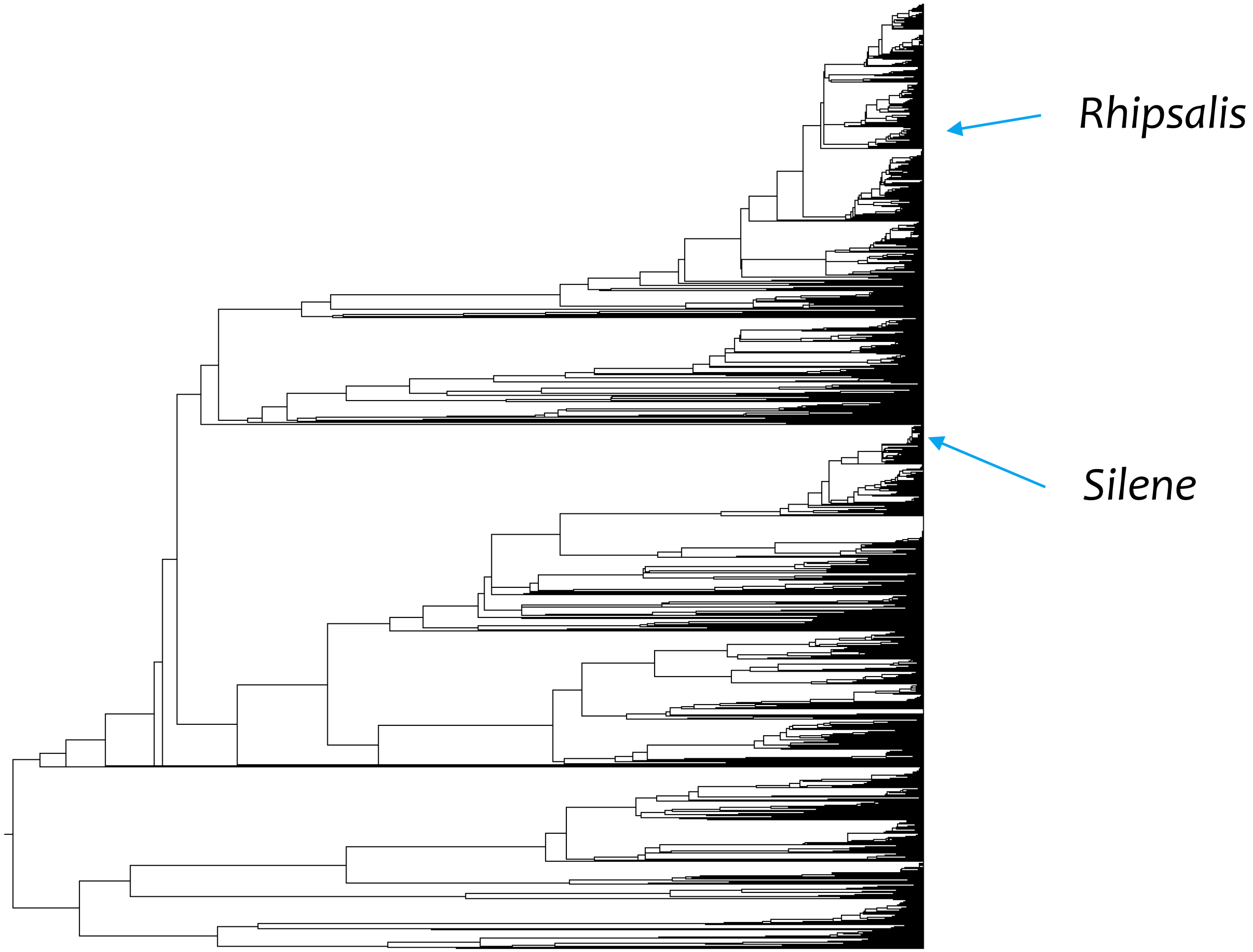
No support for anything

High support for alternative

Do we need to recalculate the tree each time?

Updating trees

If we want to add a sequence to *Rhipsalis*, do we need to recalculate the resolution of *Silene*



Why not more taxa?

The challenge of barcoding for phylogenetics

Barcoding efforts

- These are popular (and more prevelant on GenBank)
- Promise of many species that can then be placed phylogenetically or used for identification

Problems

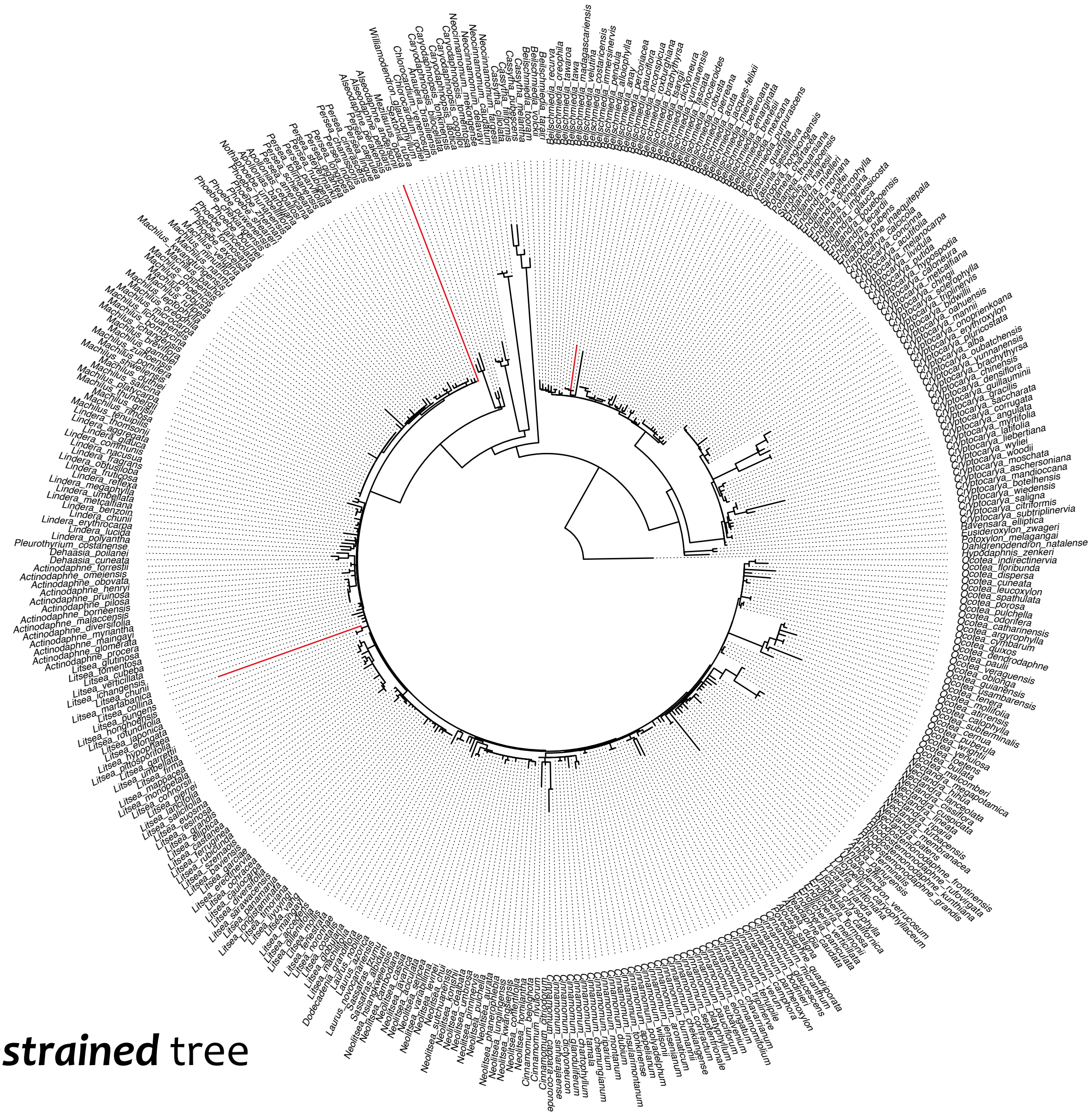
- Misidentification
  - Submitted sequences may not be the species they say they are
  - It is difficult to get NCBI to correct these and there is little information to distinguish from other sequences
- No phylogenetic information
  - Some of the sequences can be very short and not carry much if any phylogenetic information
  - This renders them useless for resolution and potentially misleading for maximum likelihood methods without uncertainty calculations

The challenge of barcoding for phylogenetics

Outlying species

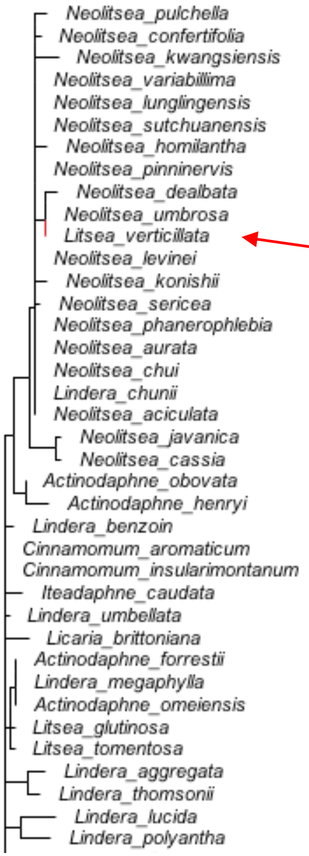
- *Litsea collina*
- *Alseodaphne andersonii*
- *Beilschmiedia pendula*

Clear pattern from constrained tree

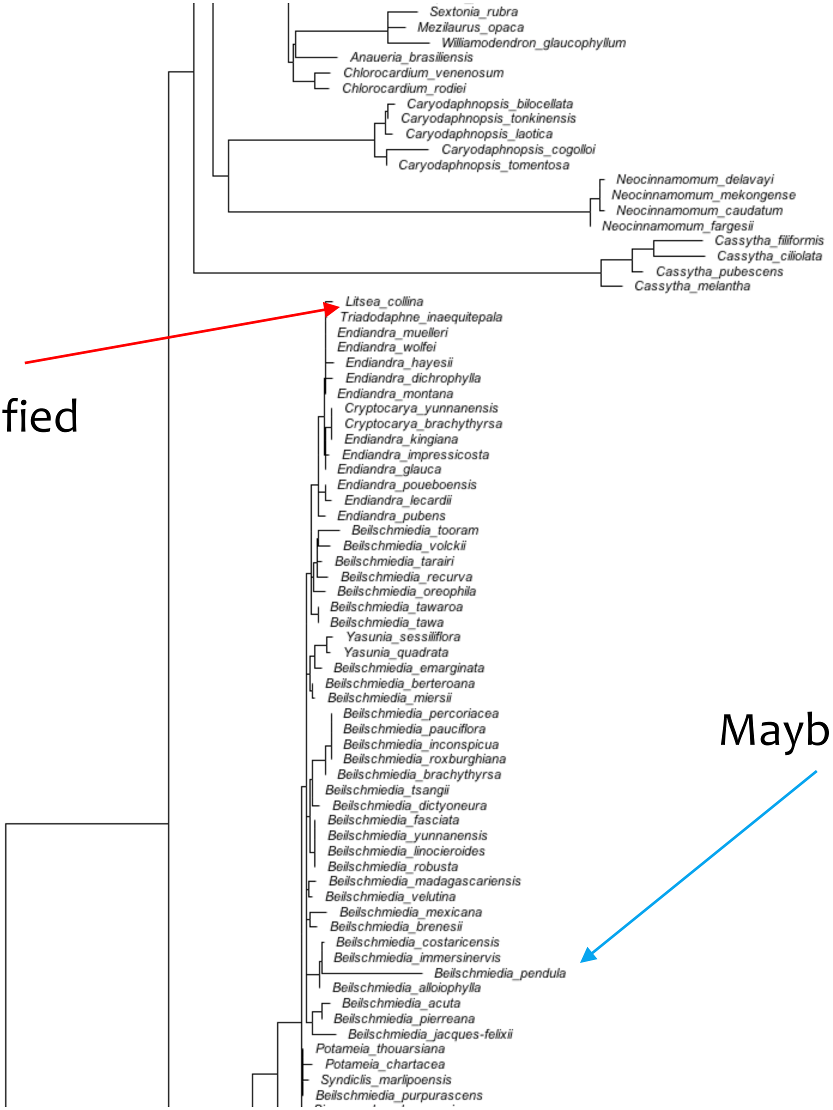


Taxonomically ***constrained*** tree

The challenge of barcoding for phylogenetics



Misidentified

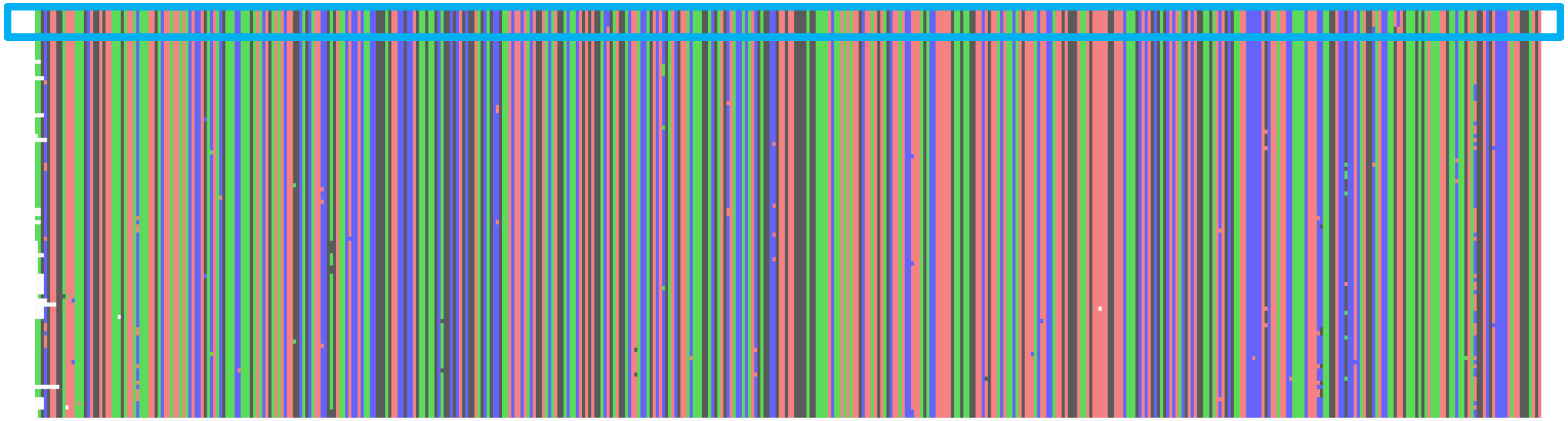


Maybe ok

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
  - We may be able to *attempt* to automate the discovery of these with the constrained searches

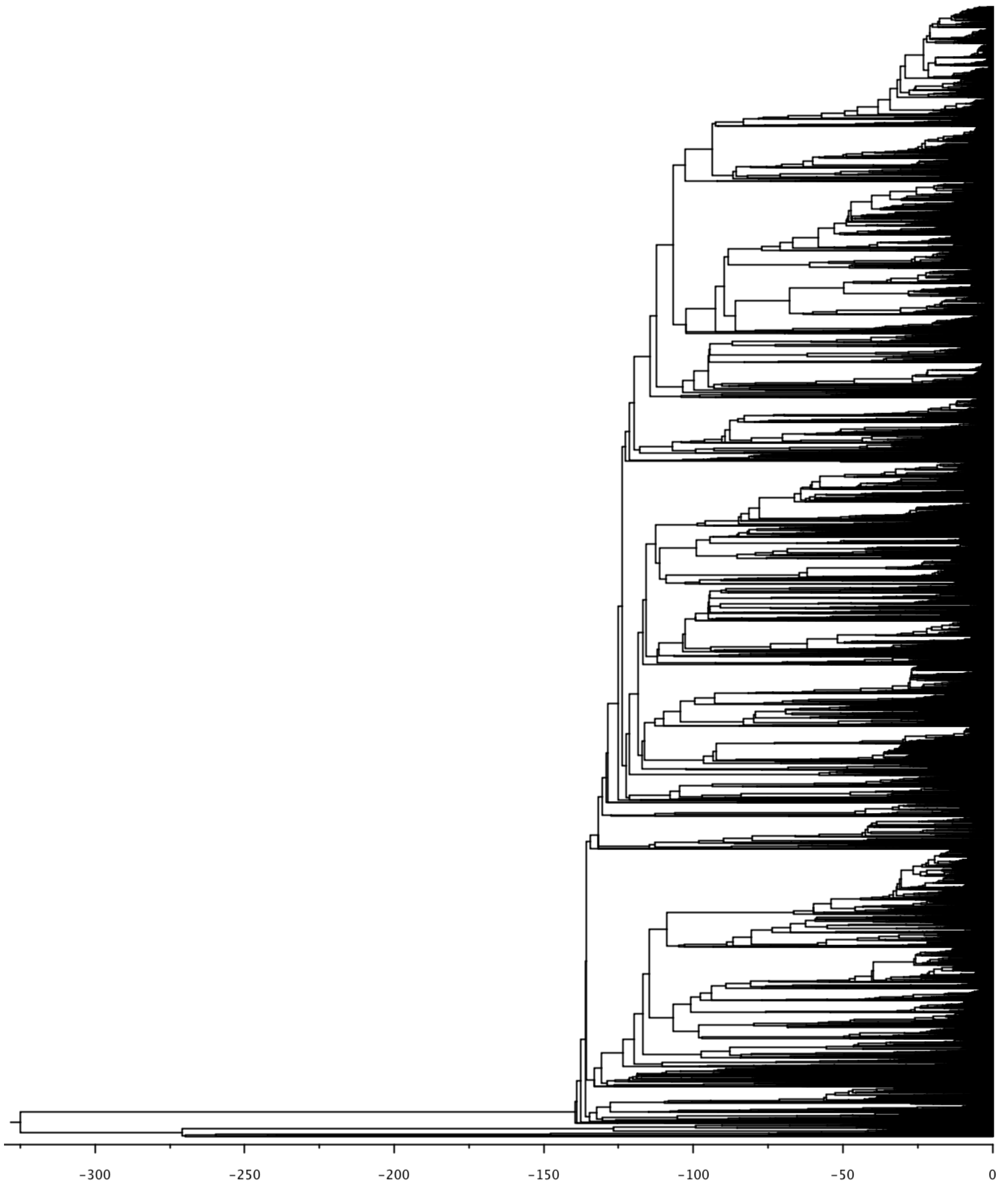
No phylogenetic information

- Maximum likelihood may be positively misled if uncertainty is not incorporated
- You may be able to use for ids to some general place in a phylogeny but not more specifically
- 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?

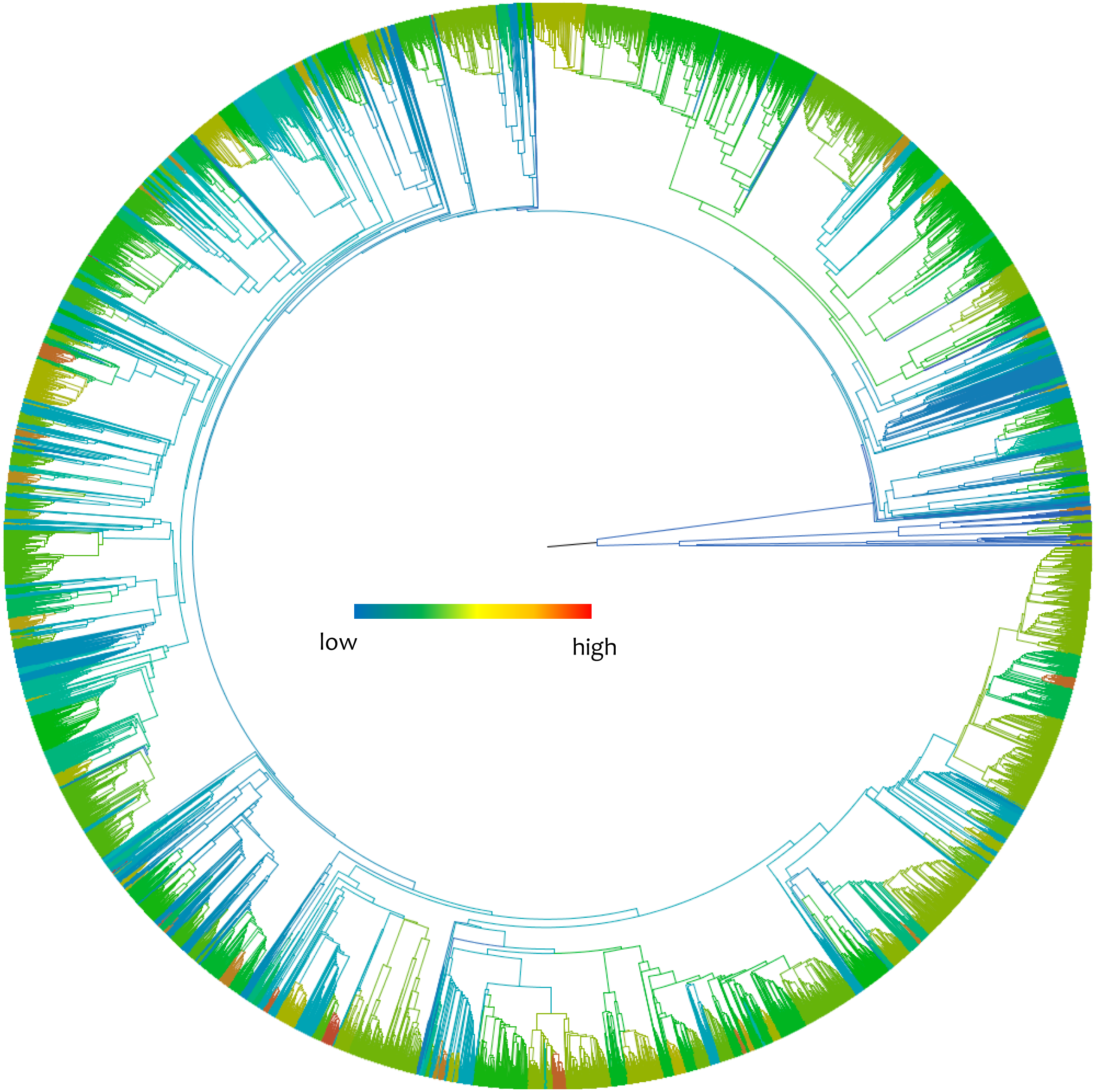
Don’t despair



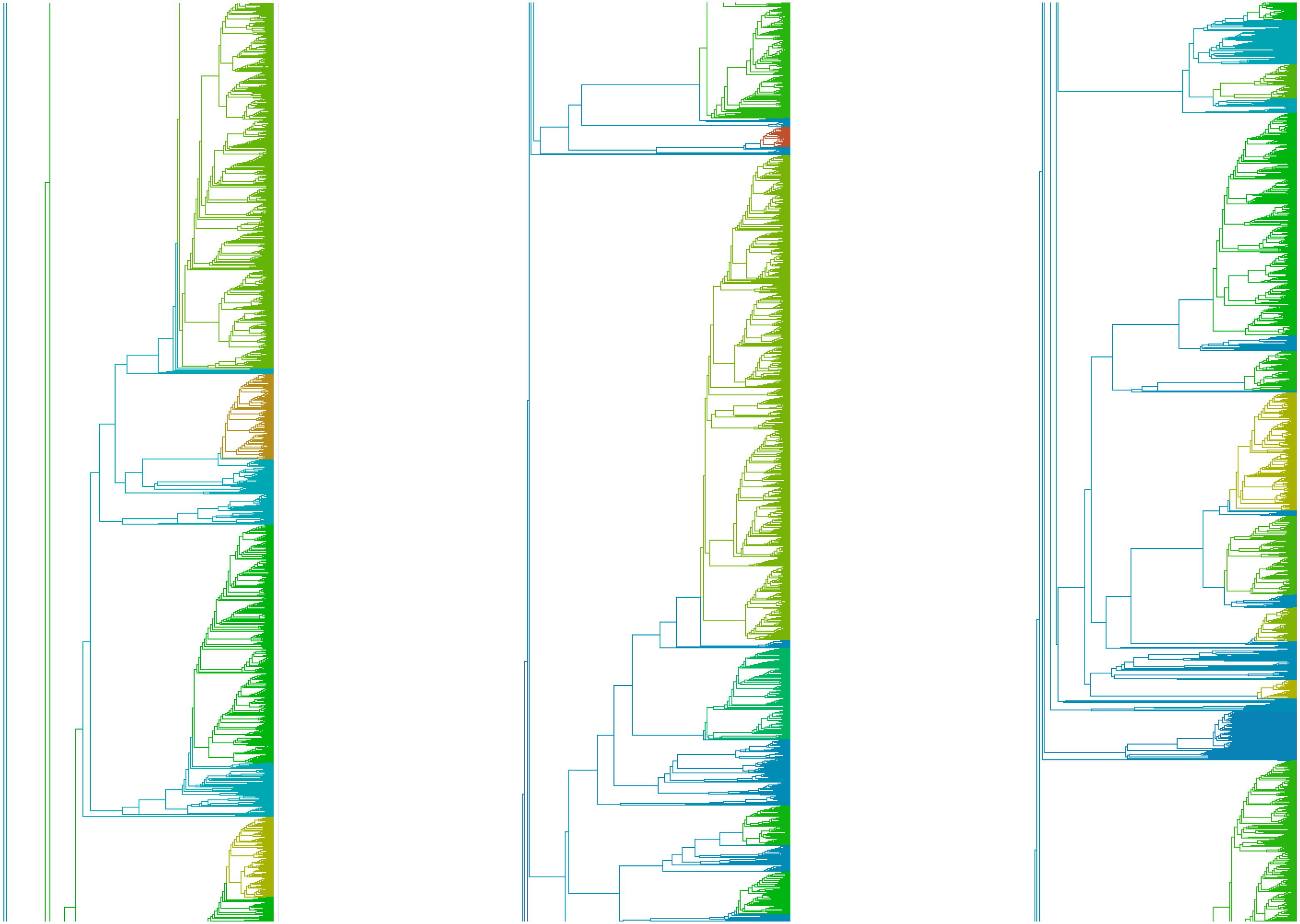
Rates of diversification

MEDUSA analyses of rates of evolution using birth-death models

Mean rate of evolution plotted

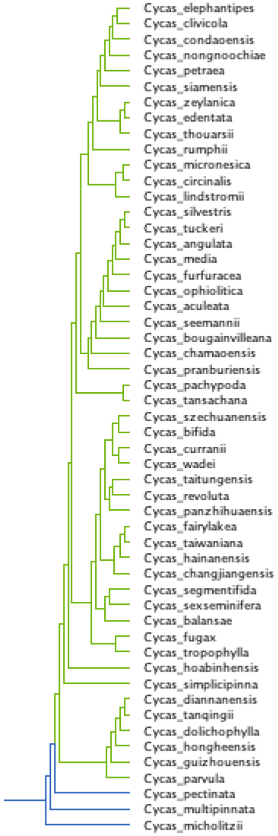


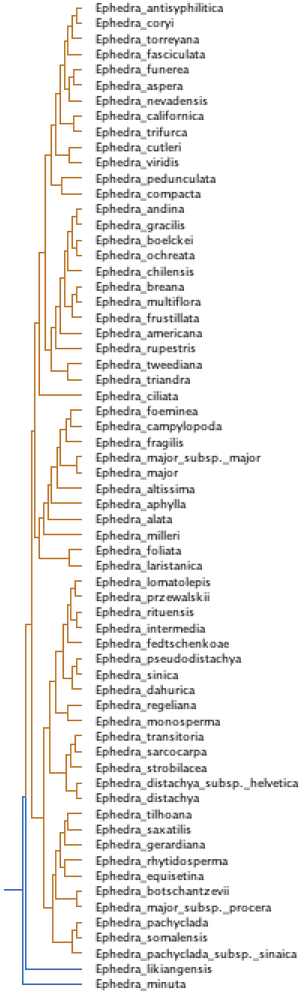
Repeated nested patterns

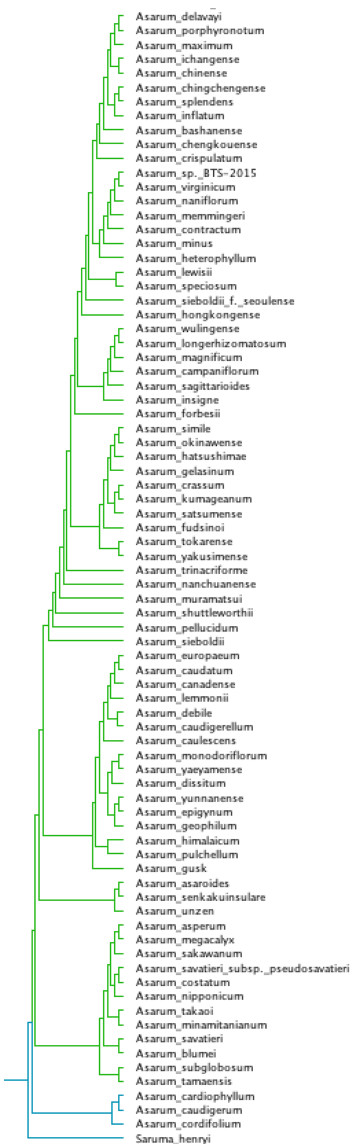


Many repeated patterns

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

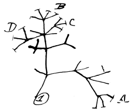






General conclusions

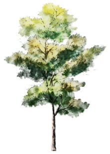
Methodological



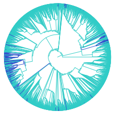
With some methodological advances, ***we can construct comprehensive trees*** with branch lengths

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