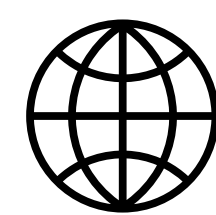


Internode certainty and related measures



@JLSteenwyk



<https://jlsteenwyk.com/>

Outline

- PhyKIT, what is it and why?
- A refresher on incongruence
- Technical comments for practical
- Quiz at 4:40pm

PhyKIT



**a toolkit for examining multiple
sequence alignments and trees**

Motivation

Motivation

Motivation

- “Code available upon request....”

Motivation

- “Code available upon request....”
- “Can you send me your script?”

Motivation

- “Code available upon request....”
- “Can you send me your script?”
- I can't read their code

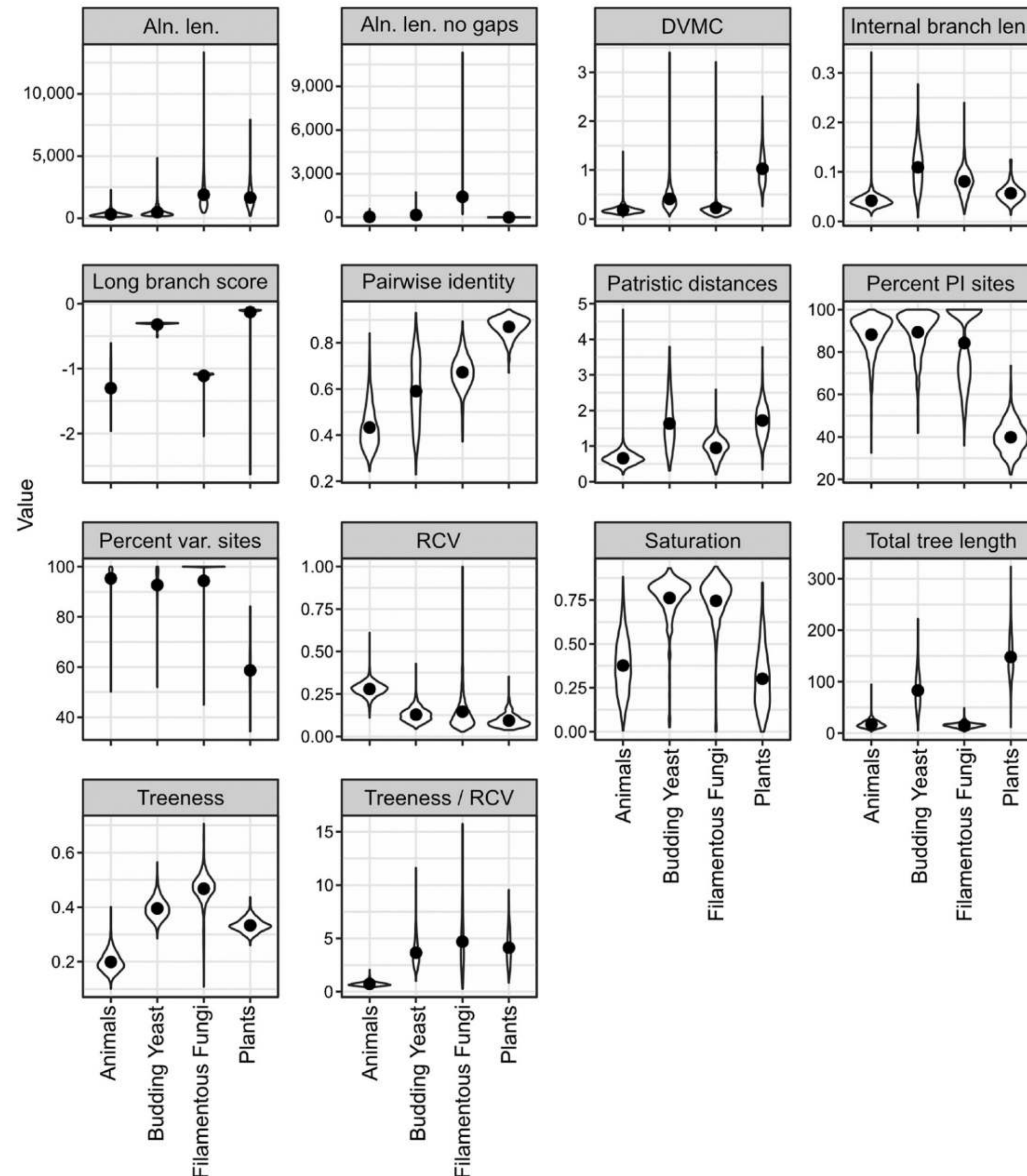
Motivation

- “Code available upon request....”
- “Can you send me your script?”
- I can’t read their code
- Documentation is horrendous or nonexistent

PhyKIT, a Swiss-army knife toolkit

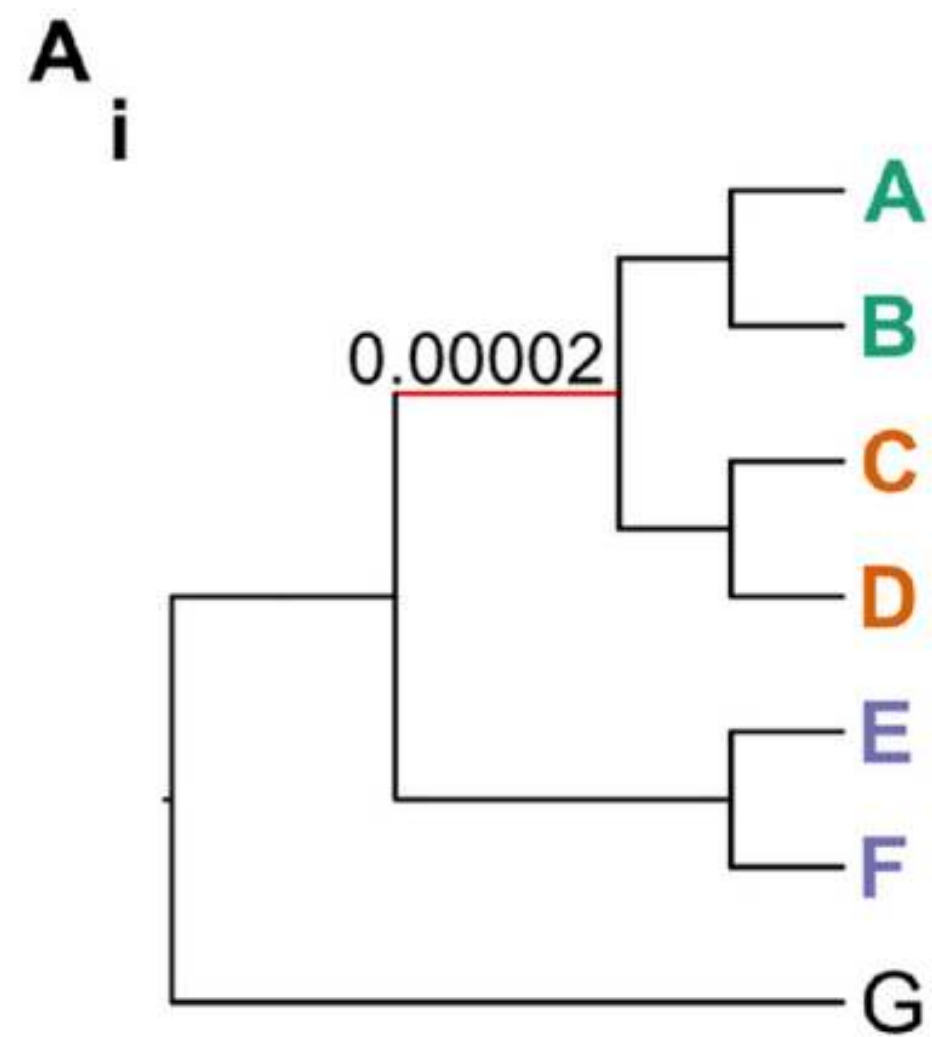
- Helps with processing and analyzing MSAs and trees
- Three exemplary use cases
 - Summarize information content
 - Identify radiations / polytomies
 - Quantify gene-gene coevolution

Information theory-based summary of data

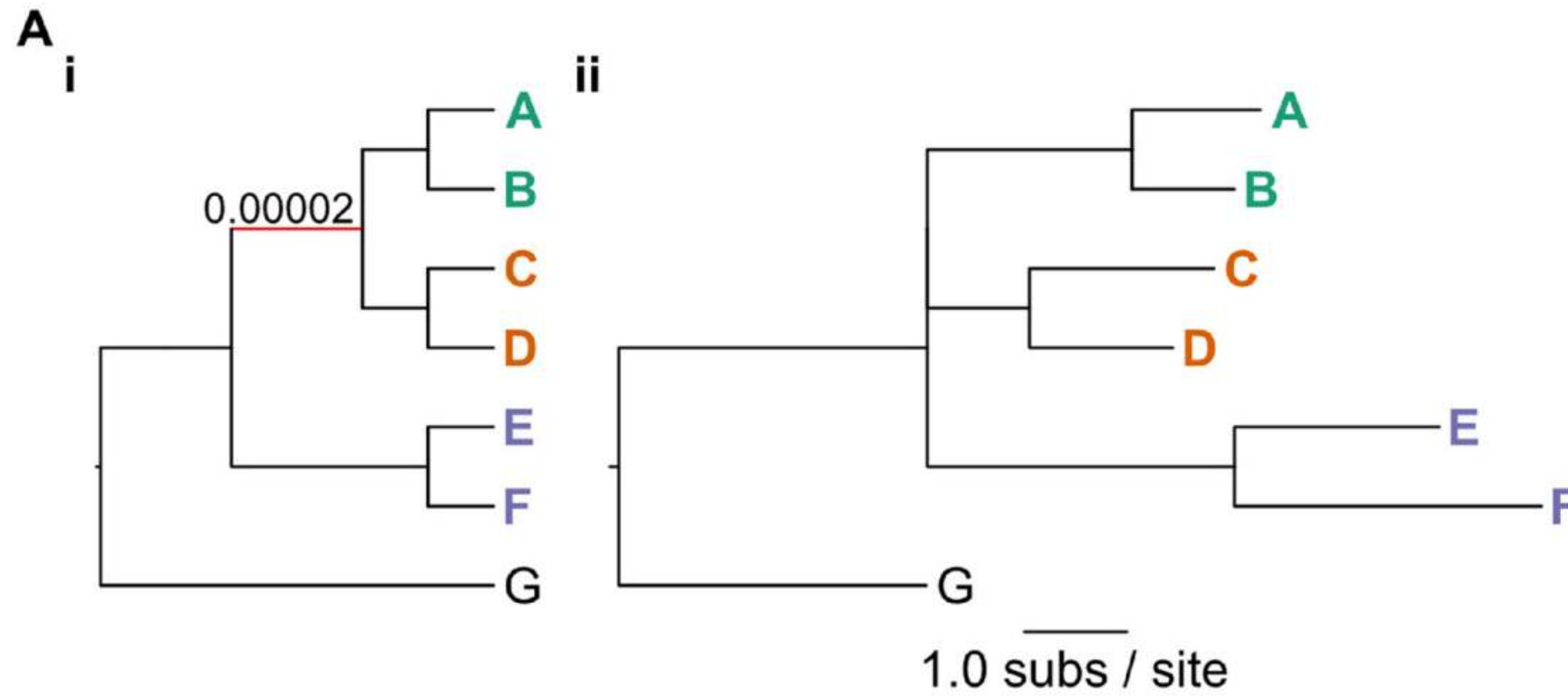


- Phylogenomic subsampling
- Identifying potential sources of error
- Quantifying evolutionary rate, etc.

The polytomy test

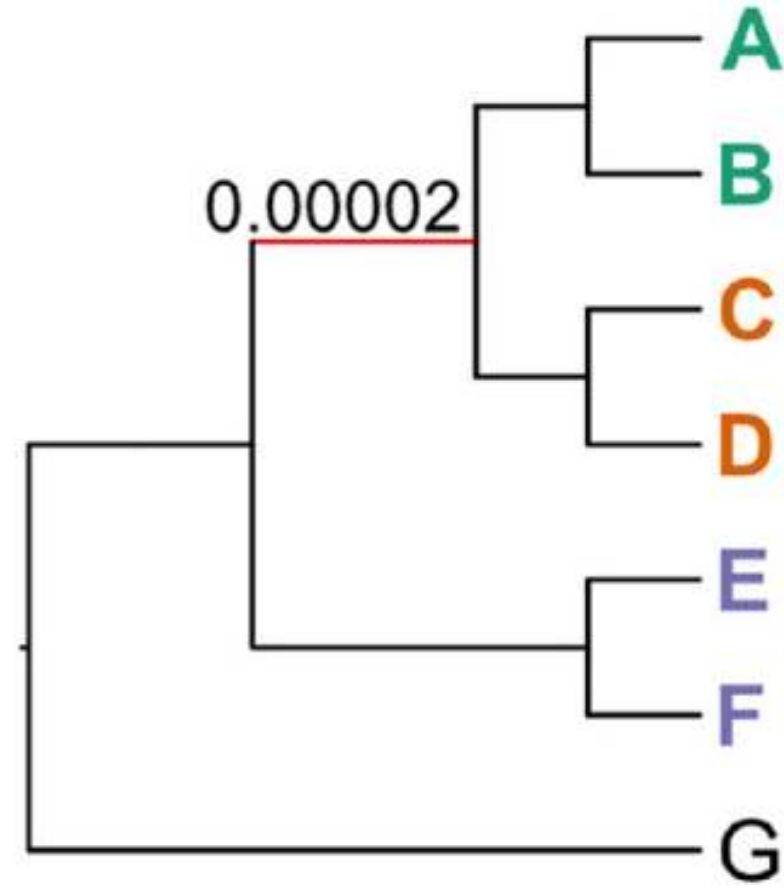


The polytomy test

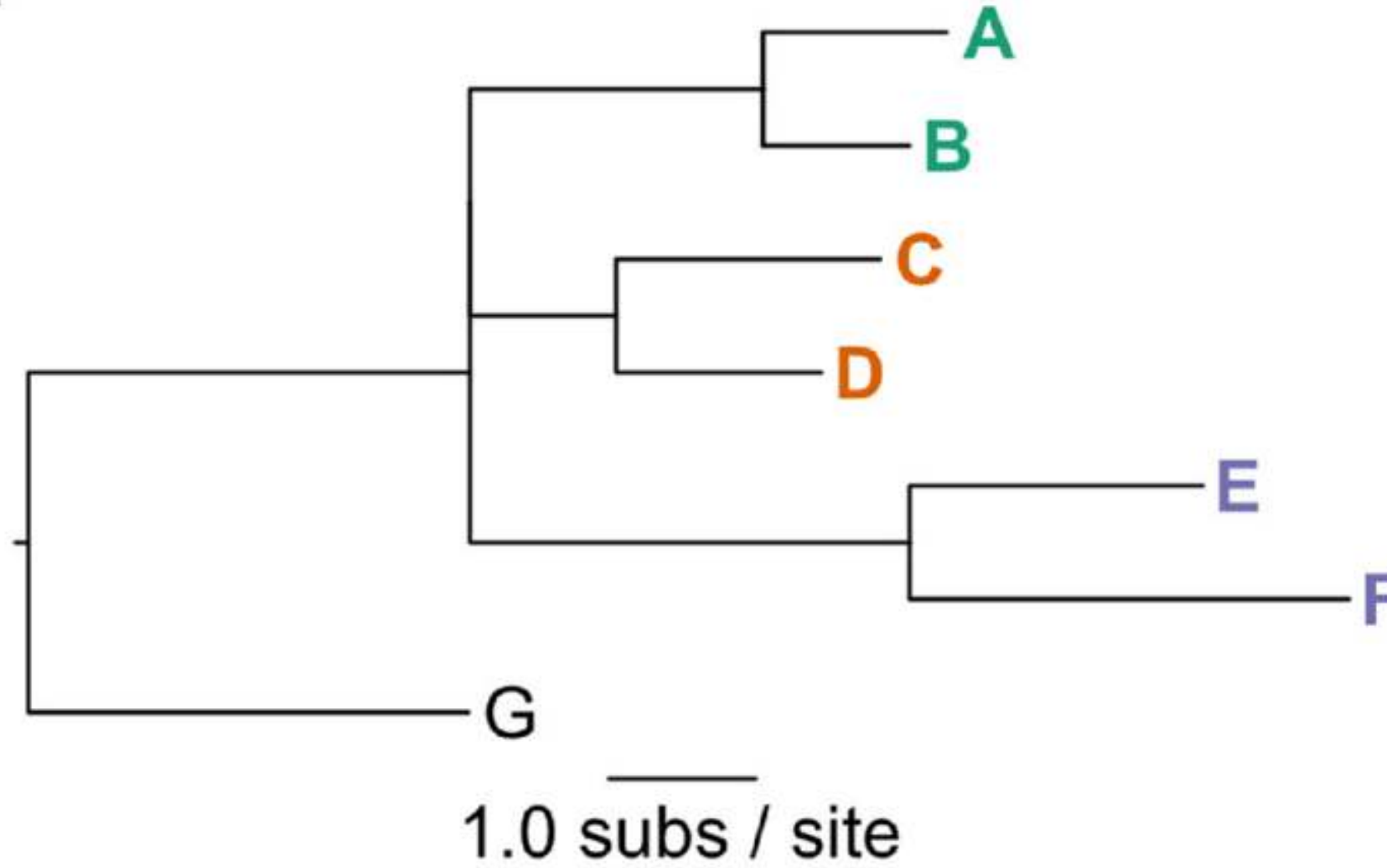


The polytomy test

A
i



ii

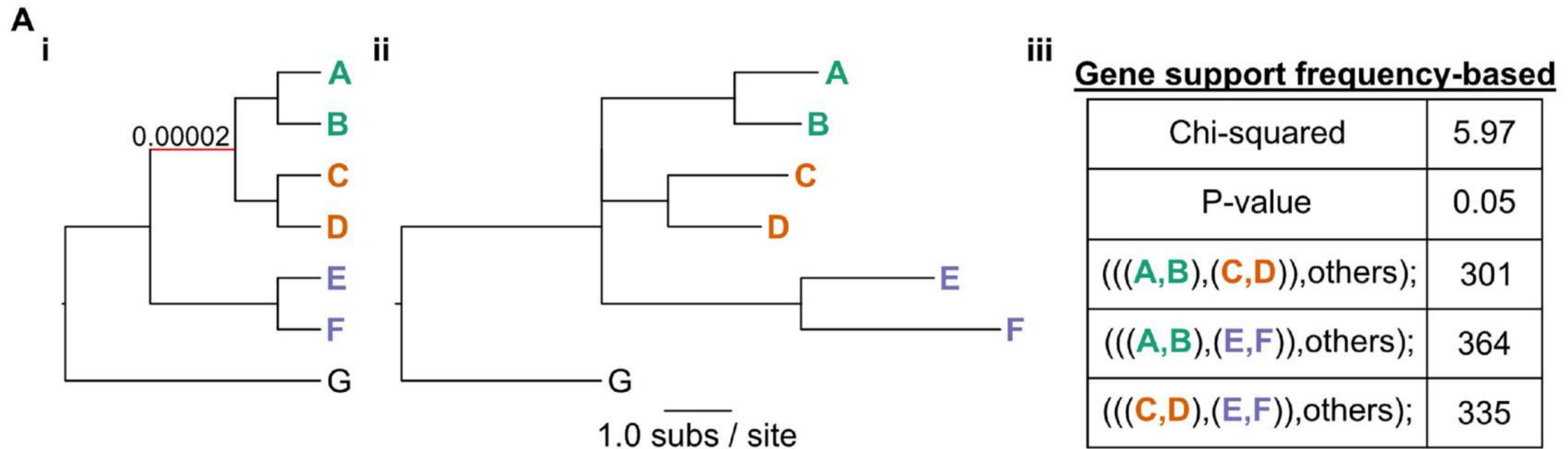


iii

Gene support frequency-based

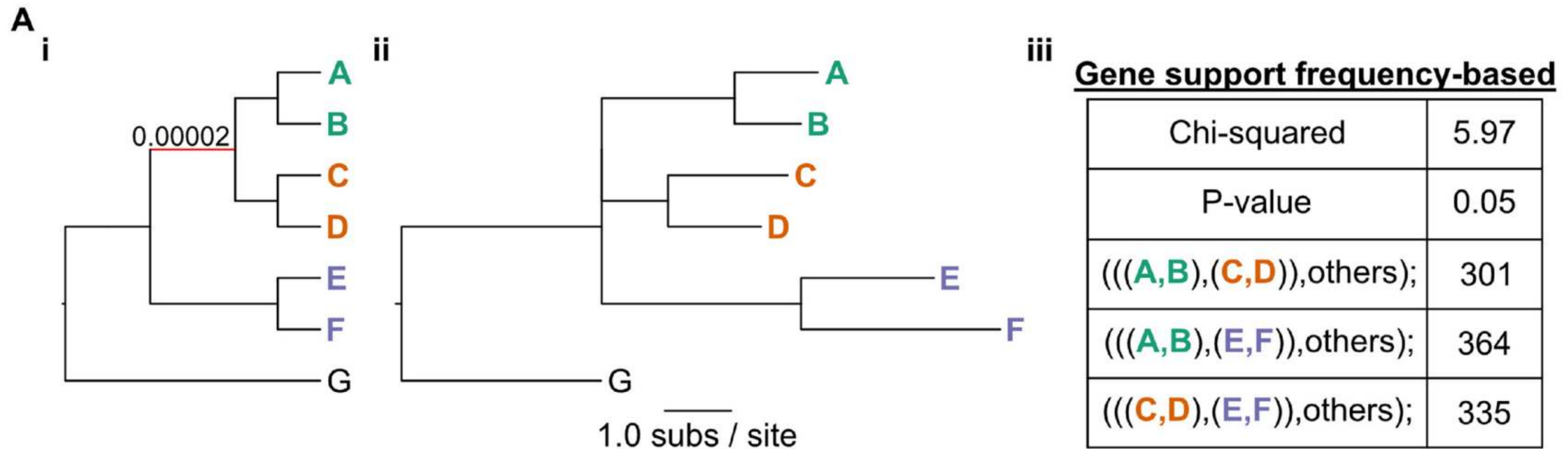
Chi-squared	5.97
P-value	0.05
(((A,B),(C,D)),others);	301
(((A,B),(E,F)),others);	364
(((C,D),(E,F)),others);	335

The polytomy test



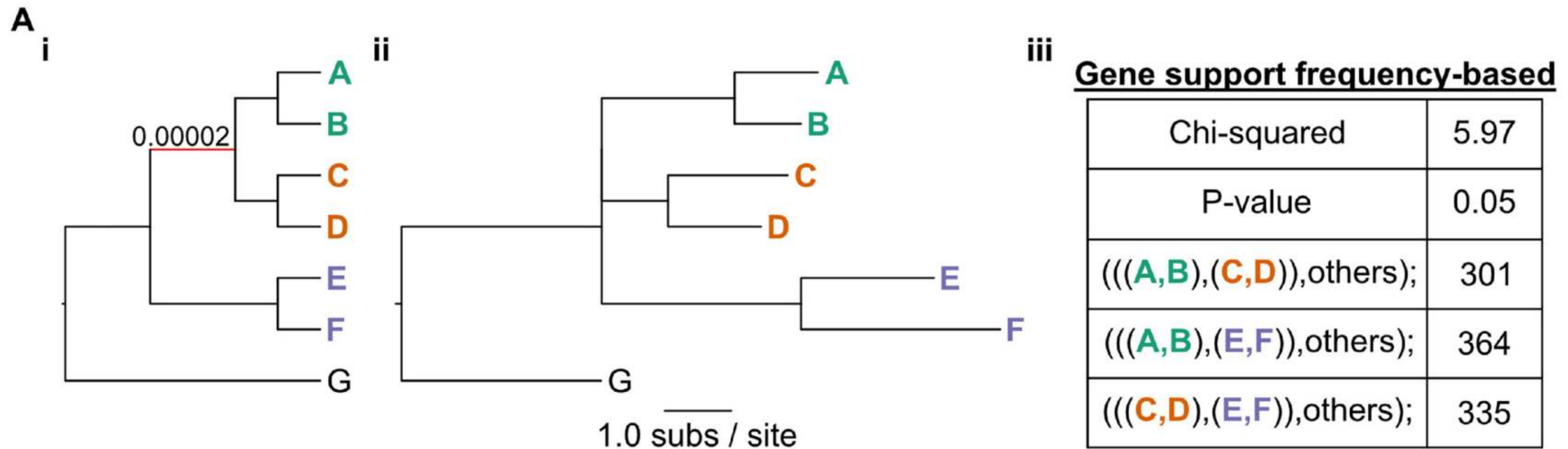
- H0: There is **no difference** in the GSF for the three possible topologies

The polytomy test



- H0: There is **no difference** in the GSF for the three possible topologies
- HA: There is a **difference** in the GSF for three possible topologies

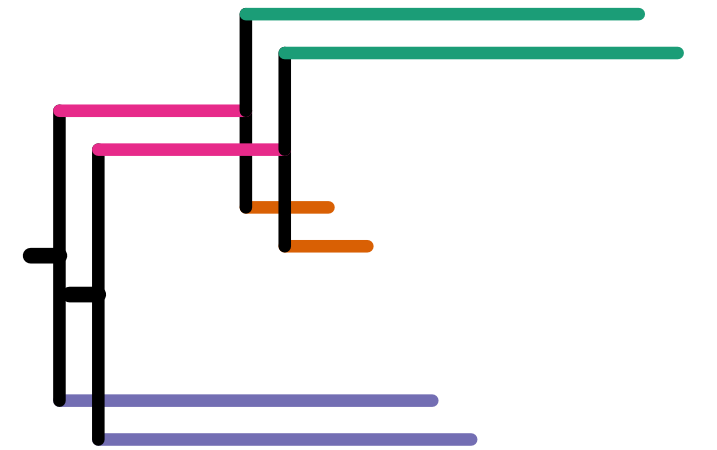
The polytomy test



- H0: There is **no difference** in the GSF for the three possible topologies
- HA: There is a **difference** in the GSF for three possible topologies

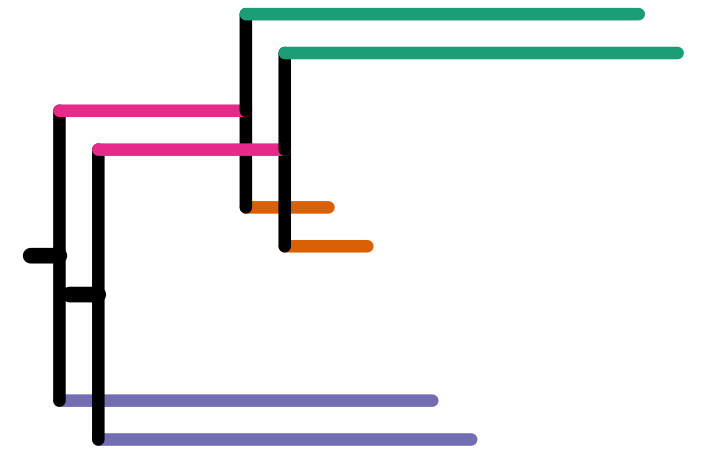
Gene-gene coevolution predicts shared function

- gene coevolution refers to:
 - two genes that covary in parallel across speciation events
 - often observed among genes that share function, are coexpressed, or are part of the same multi-meric complexes



Gene-gene coevolution predicts shared function

- gene coevolution refers to:
 - two genes that covary in parallel across speciation events
 - often observed among genes that share function, are coexpressed, or are part of the same multi-meric complexes



PhyKIT

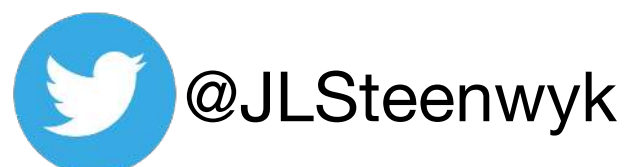
a toolkit for examining multiple
sequence alignments and trees

PhyKIT: a broadly applicable UNIX shell
toolkit for processing and analyzing
phylogenomic data

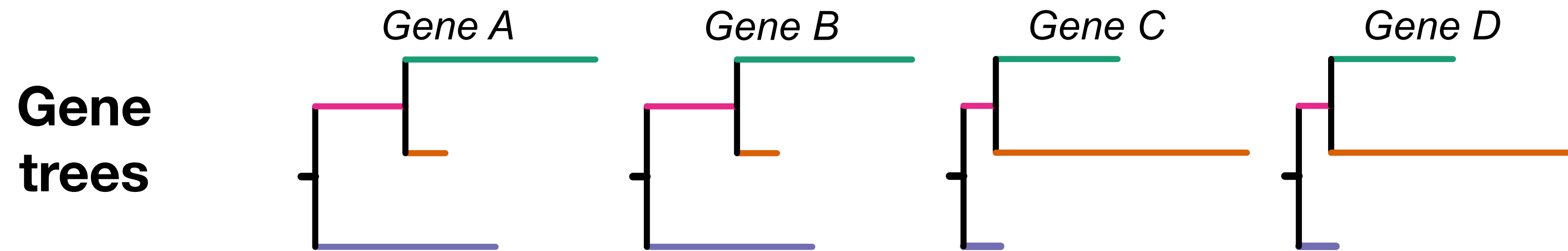
Jacob L Steenwyk ✉, Thomas J Buida, III, Abigail L Labella, Yuanning Li,
Xing-Xing Shen, Antonis Rokas ✉

Bioinformatics, btab096, <https://doi.org/10.1093/bioinformatics/btab096>

Published: 09 February 2021 Article history ▼

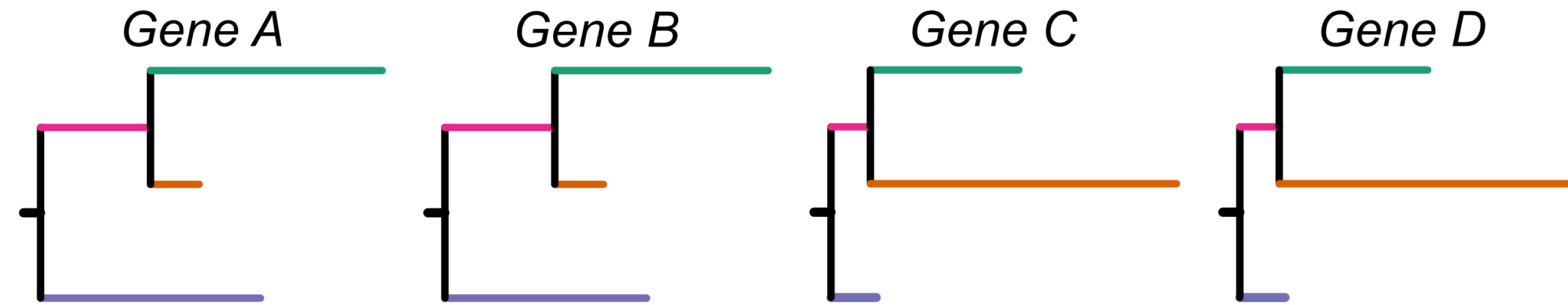


The mirror principle to detect gene coevolution

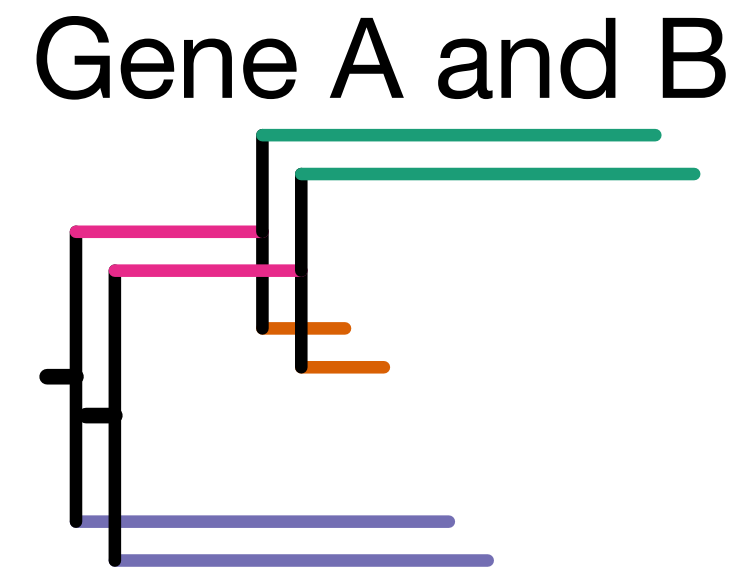


The mirror principle to detect gene coevolution

**Gene
trees**

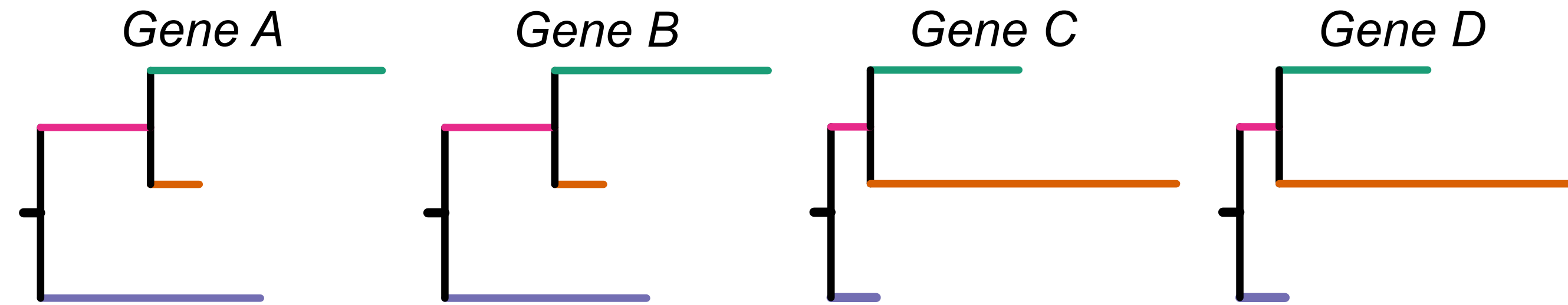


**The
'mirror'
principle**

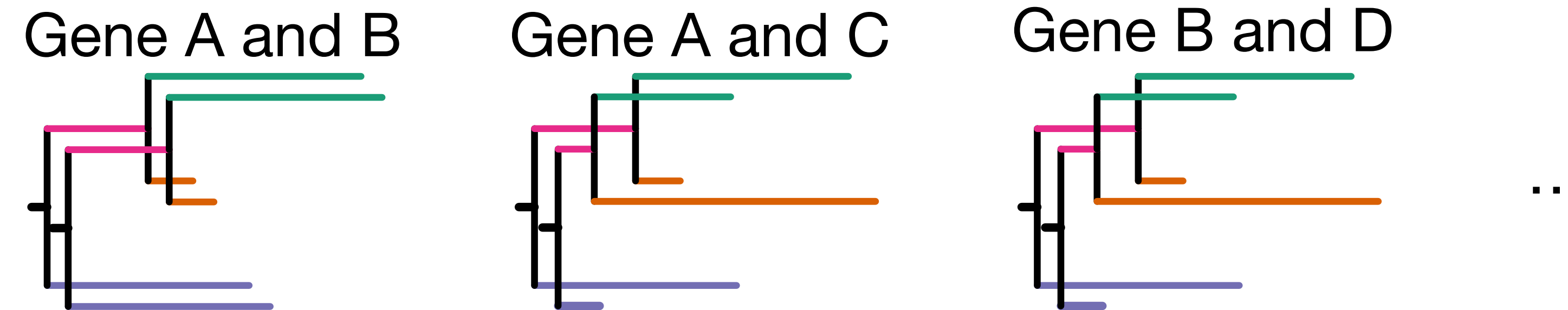


The mirror principle to detect gene coevolution

**Gene
trees**

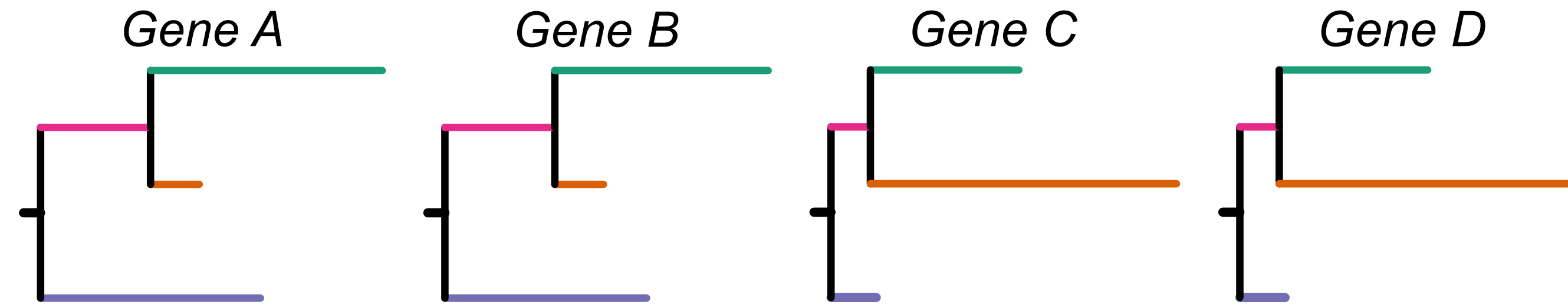


**The
'mirror'
principle**

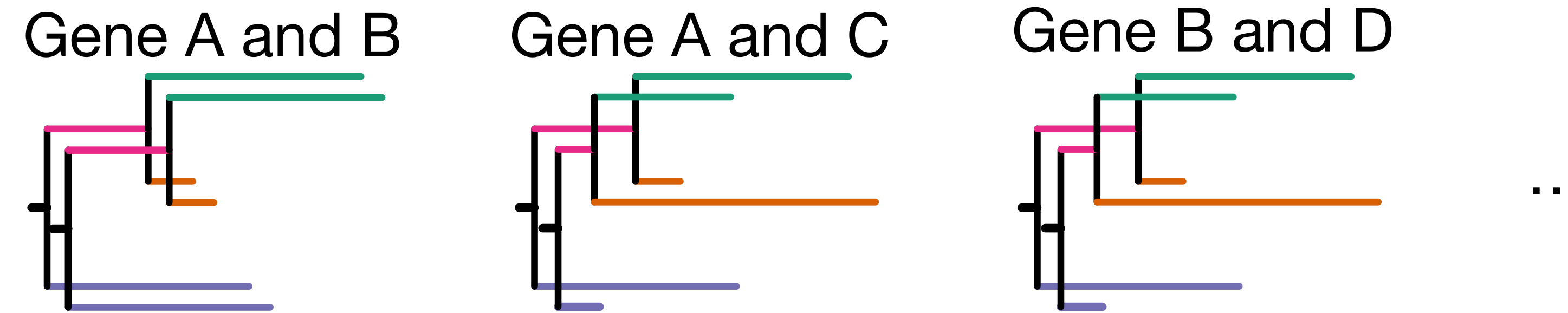


The mirror principle to detect gene coevolution

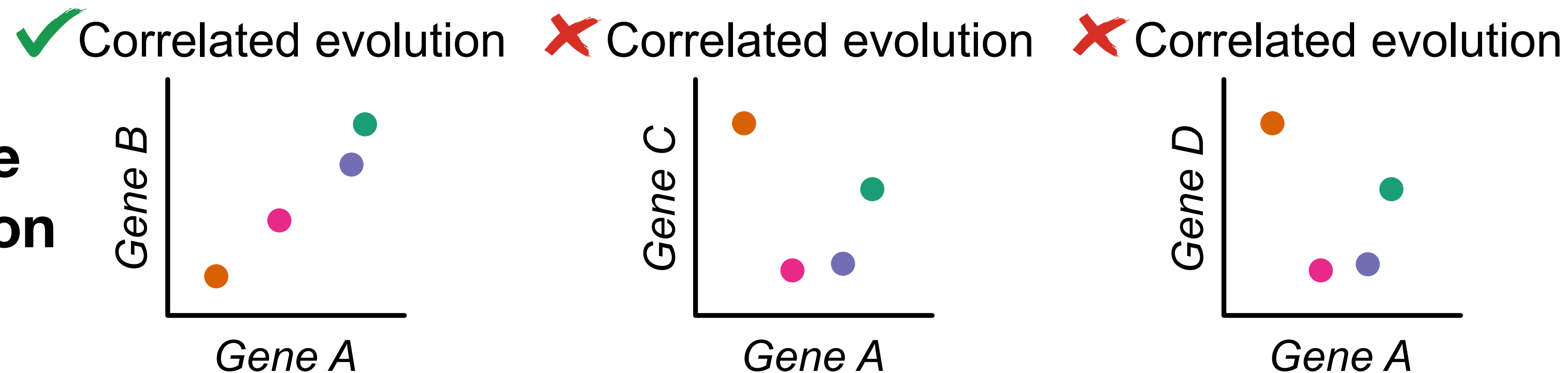
Gene trees



The 'mirror' principle

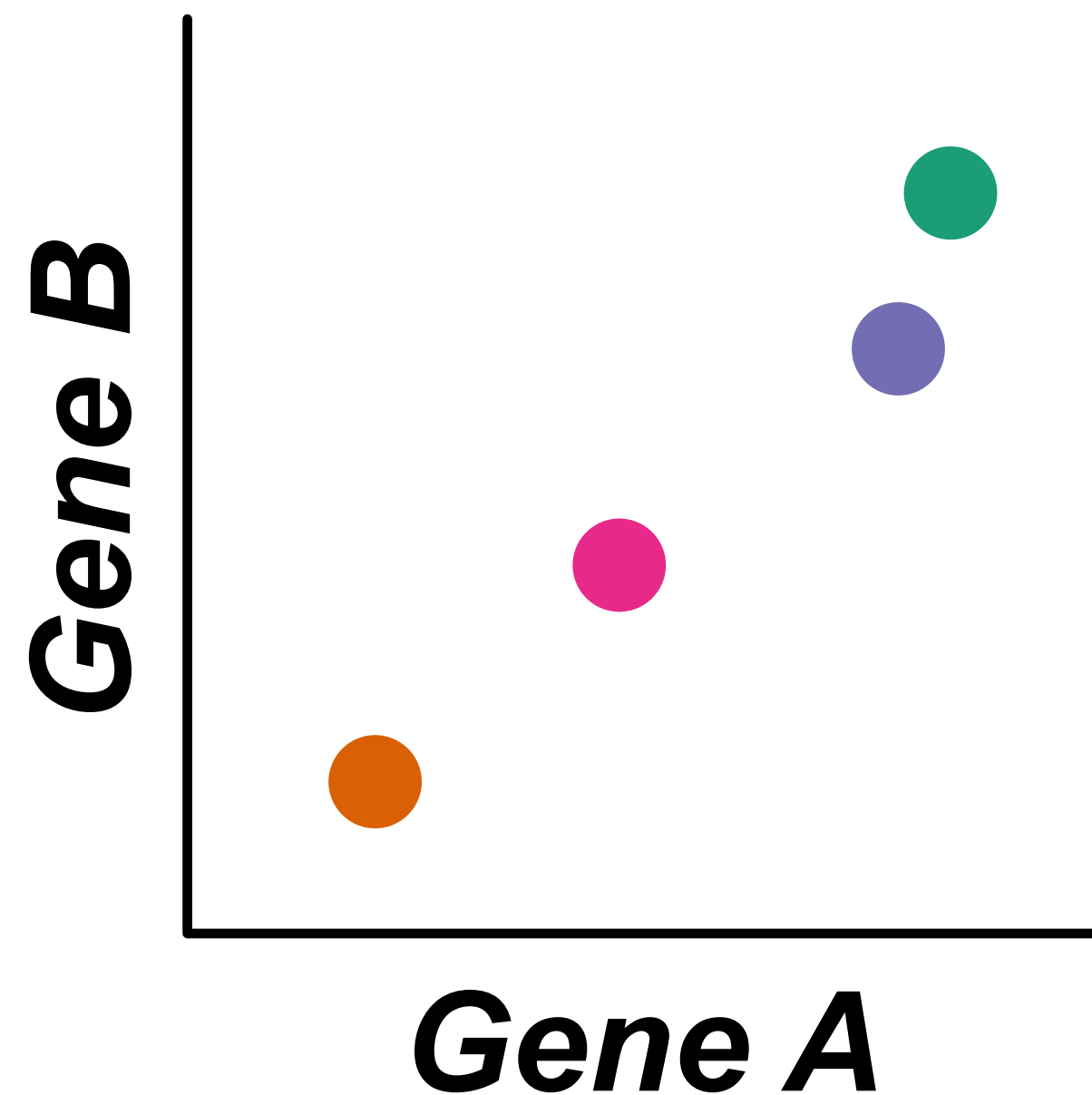


Examine covariation



Genes of a feather evolve together

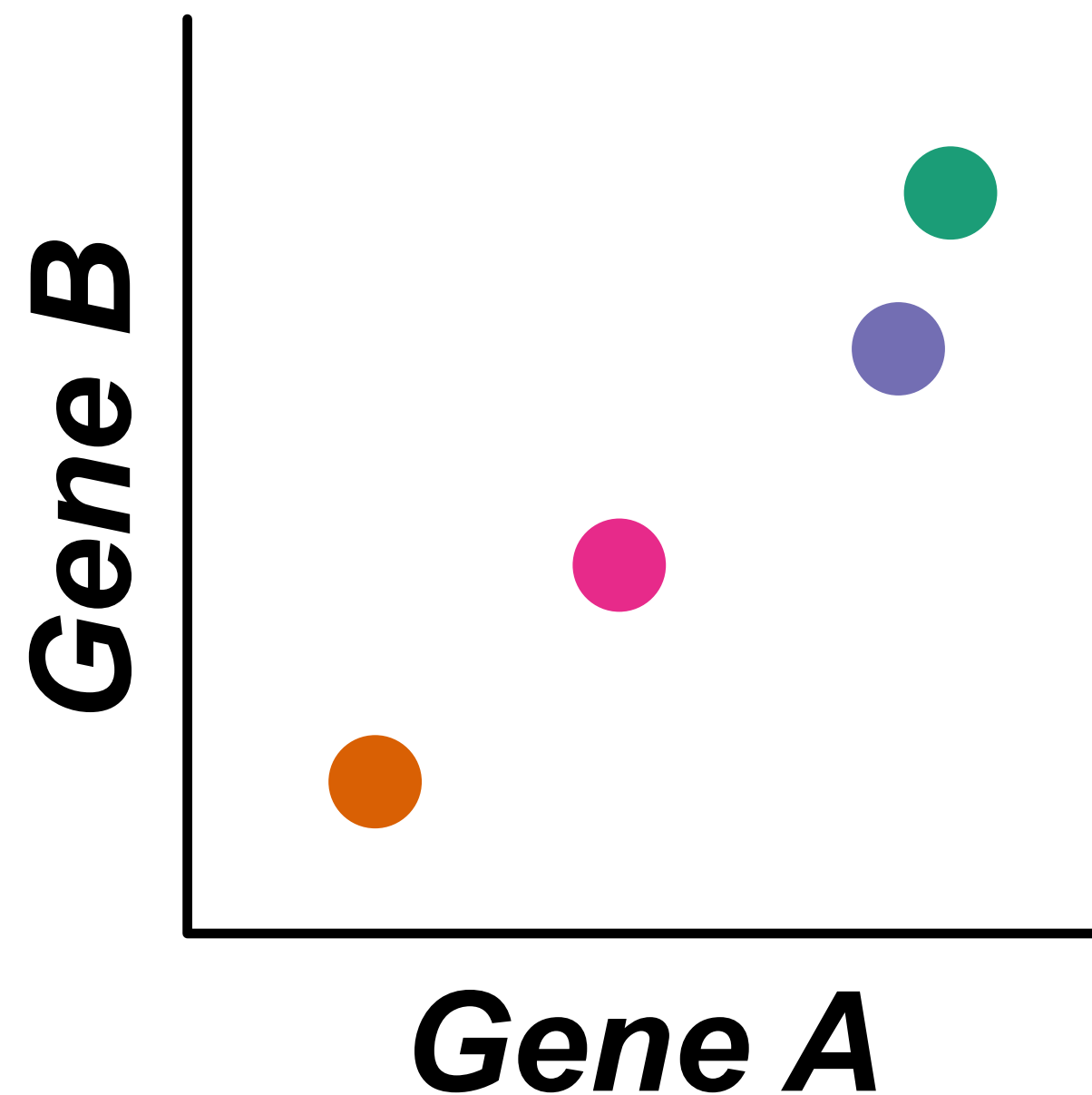
✓ Correlated evolution



- Coevolving genes tend to share function, be coexpressed, or are part of the same multimeric complexes

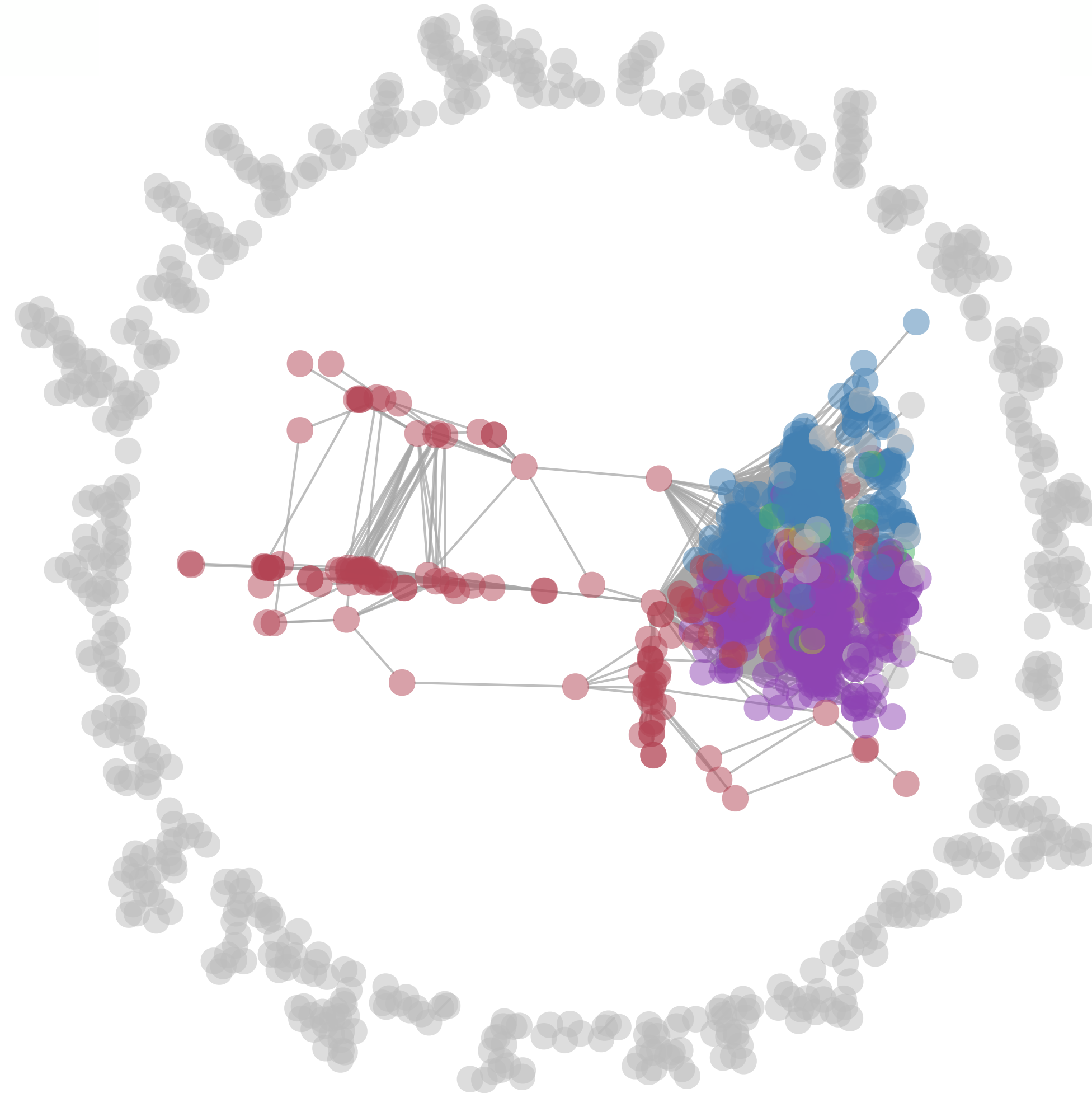
Genes of a feather evolve together

✓ Correlated evolution



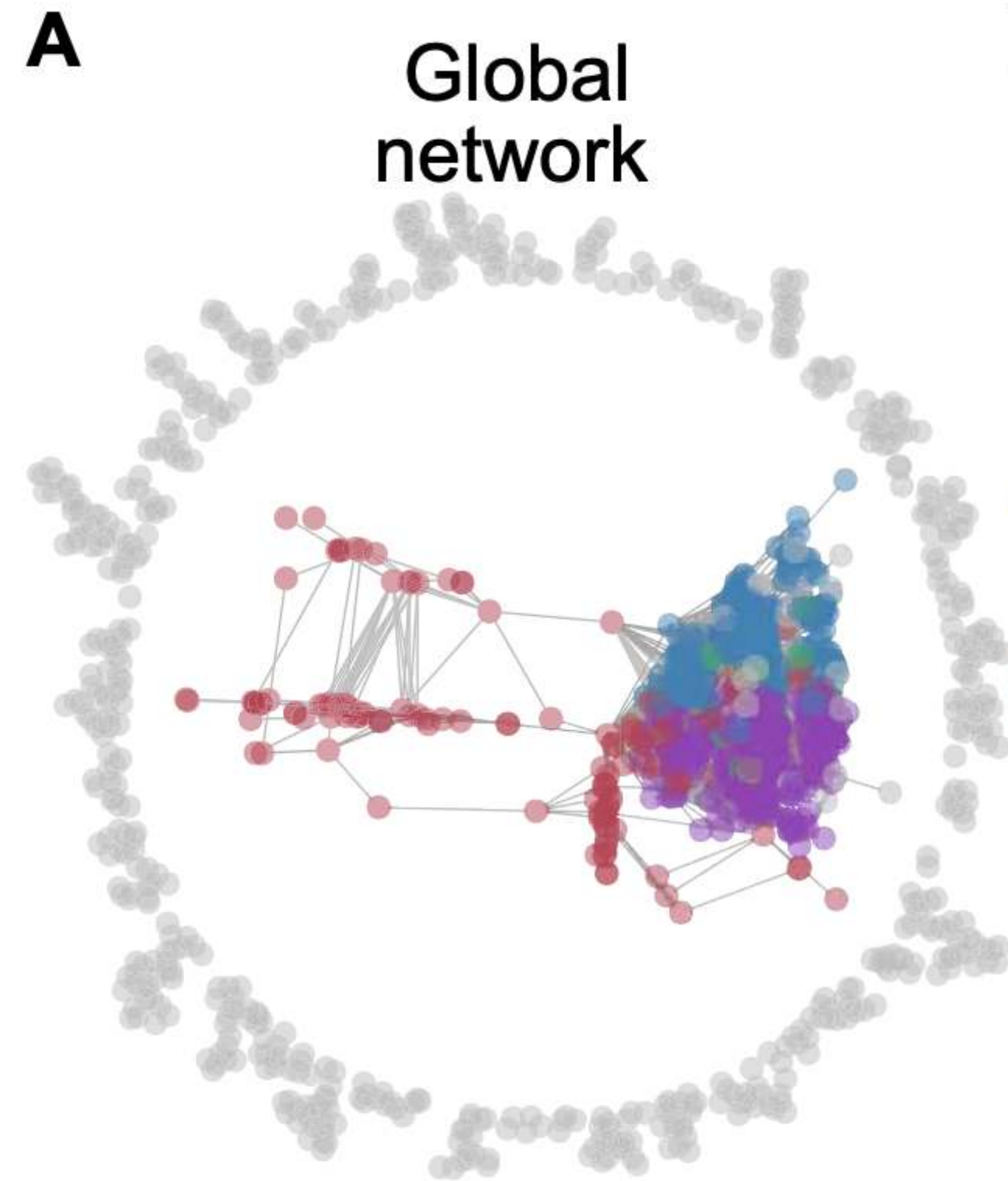
- Coevolving genes tend to share function, be coexpressed, or are part of the same multimeric complexes
- **But can we build a genetic network?**

A gene coevolutionary network for budding yeast

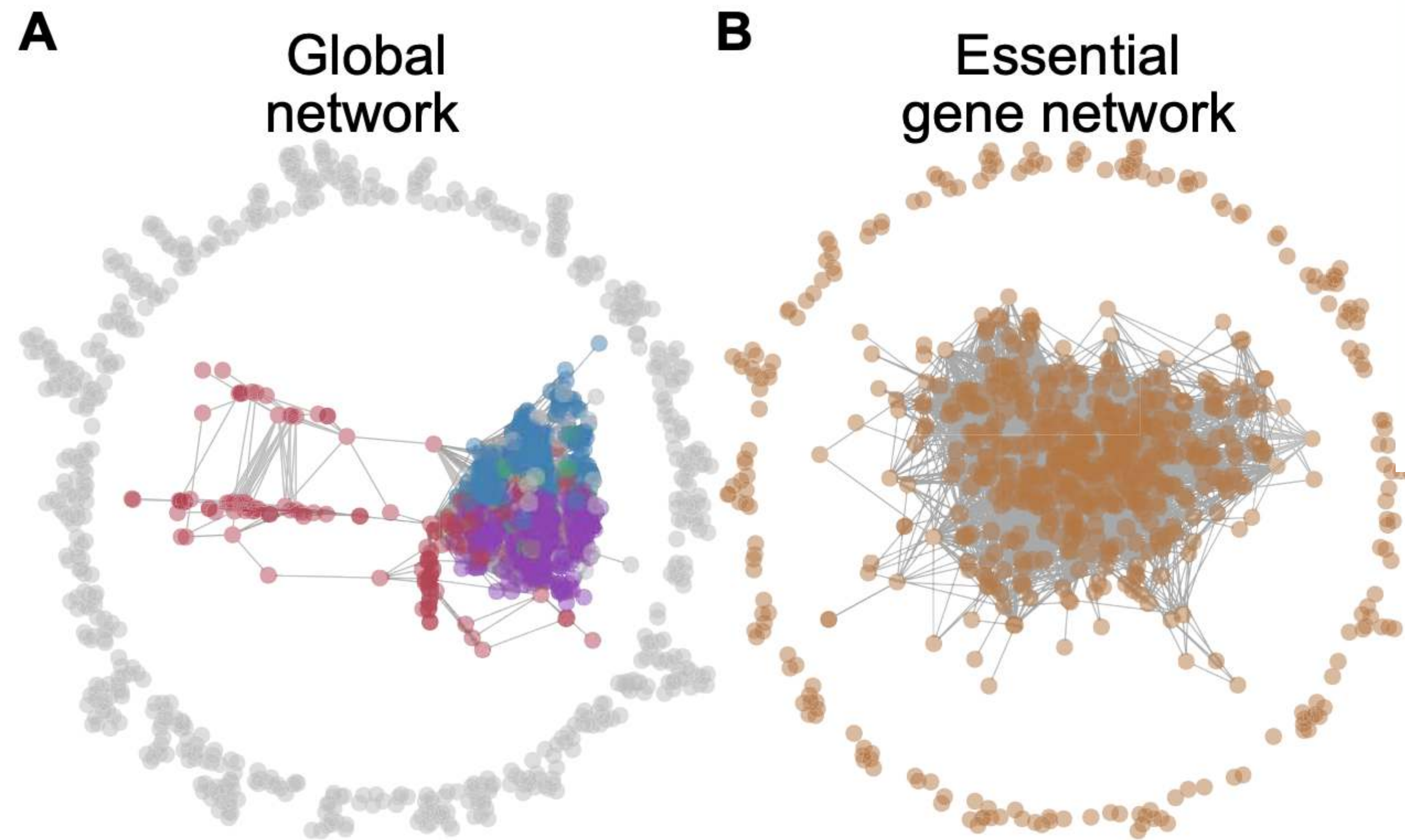


- Nodes are genes
- Genes that are predicted to be cofunctional are connected

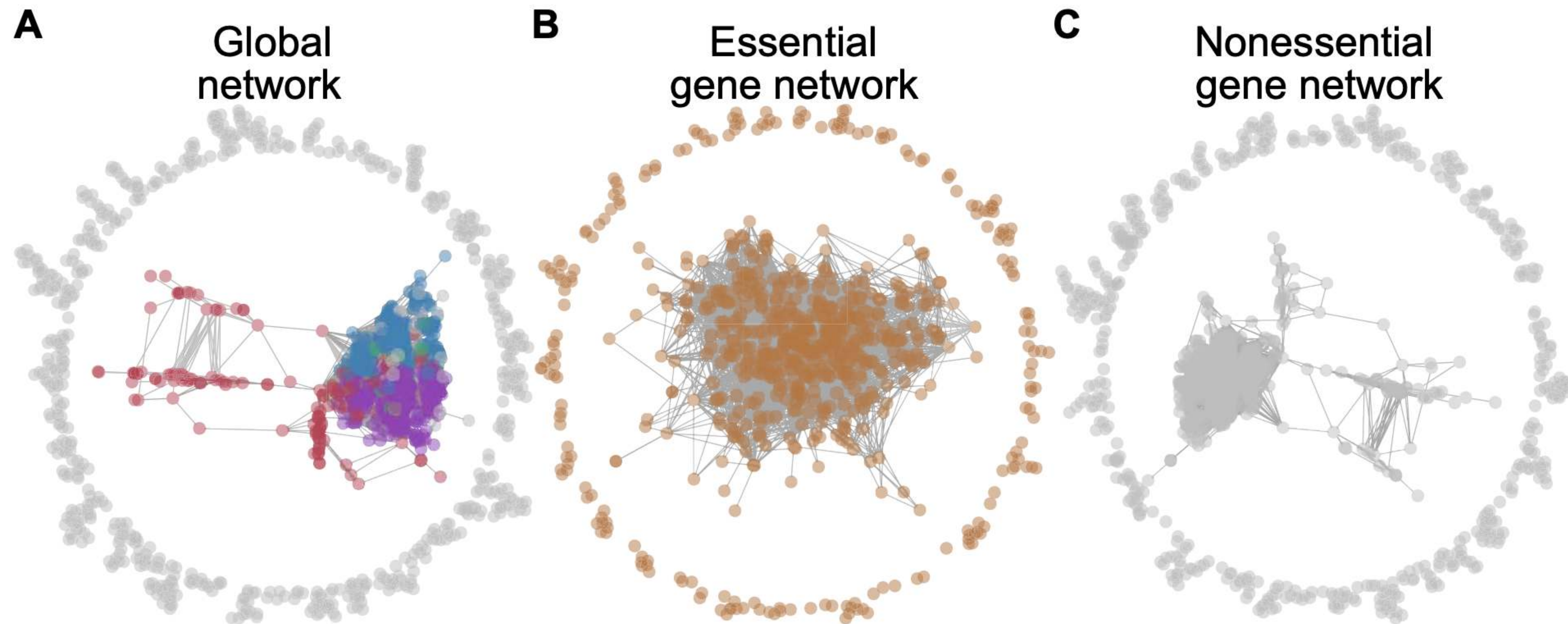
Gene coevolution networks capture bio-information



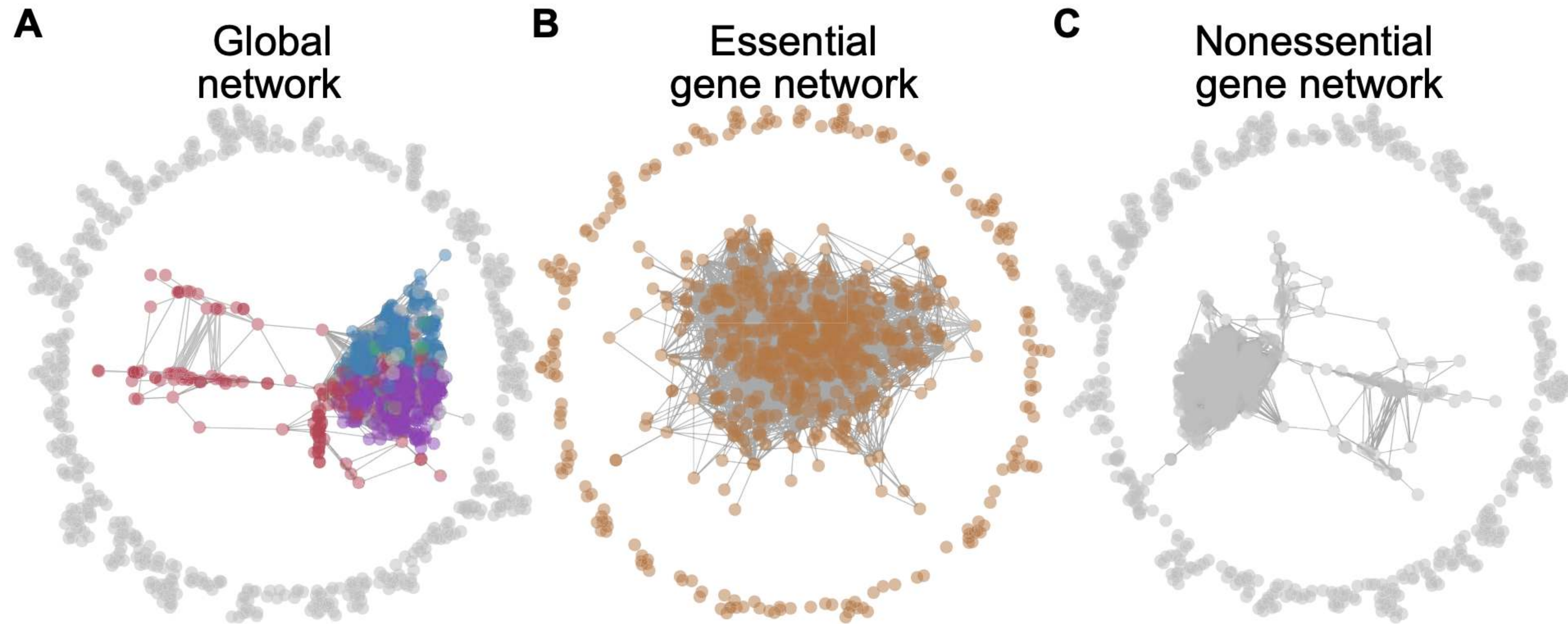
Gene coevolution networks capture bio-information



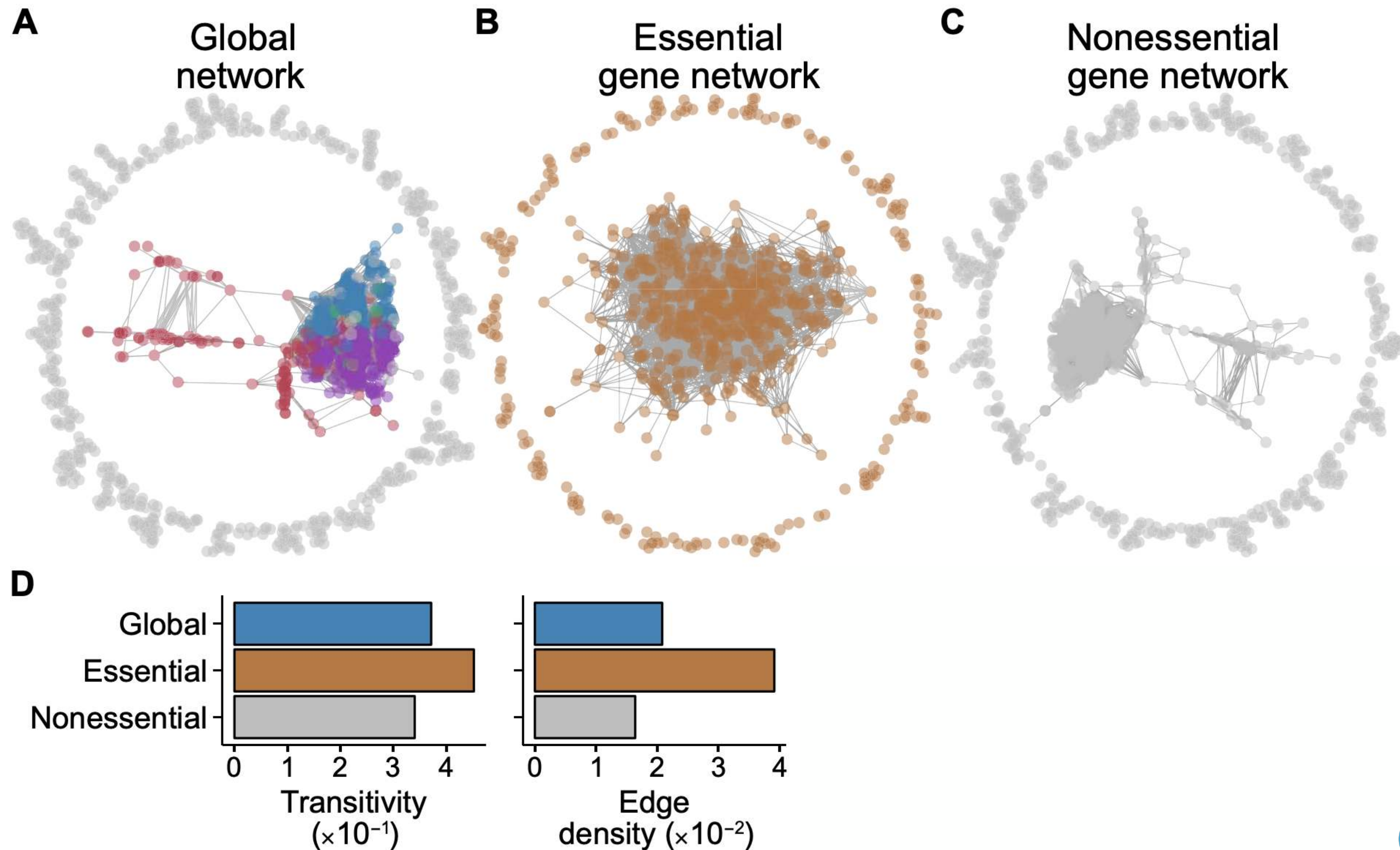
Gene coevolution networks capture bio-information



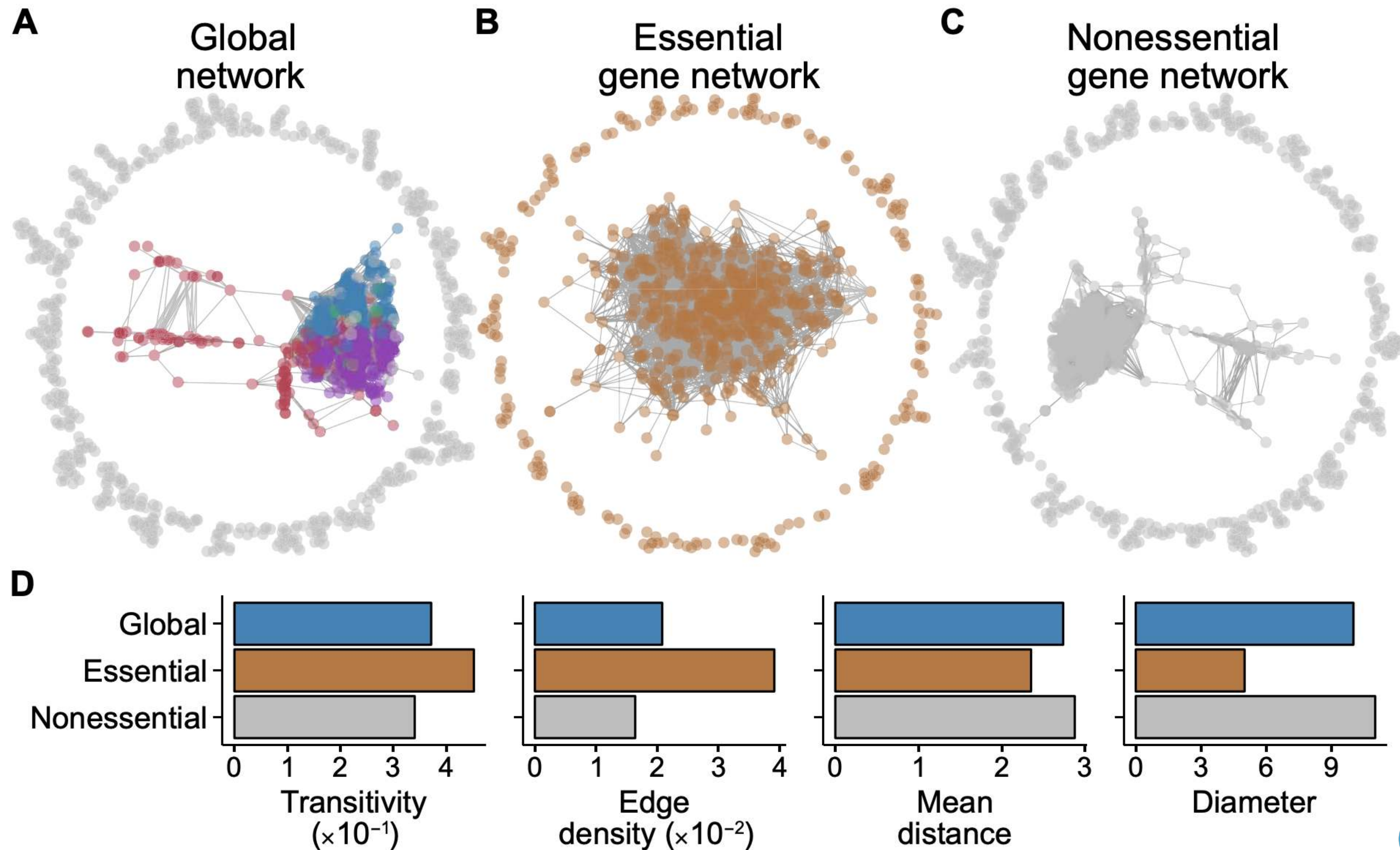
Gene coevolution networks capture bio-information



Gene coevolution networks capture bio-information



Gene coevolution networks capture bio-information



PhyKIT, a Swiss-army knife toolkit

- Helps with processing and analyzing MSAs and trees
- Three exemplary use cases
 - Summarize information content
 - Identify radiations / polytomies
 - Quantify gene-gene coevolution

PhyKIT is not your only option

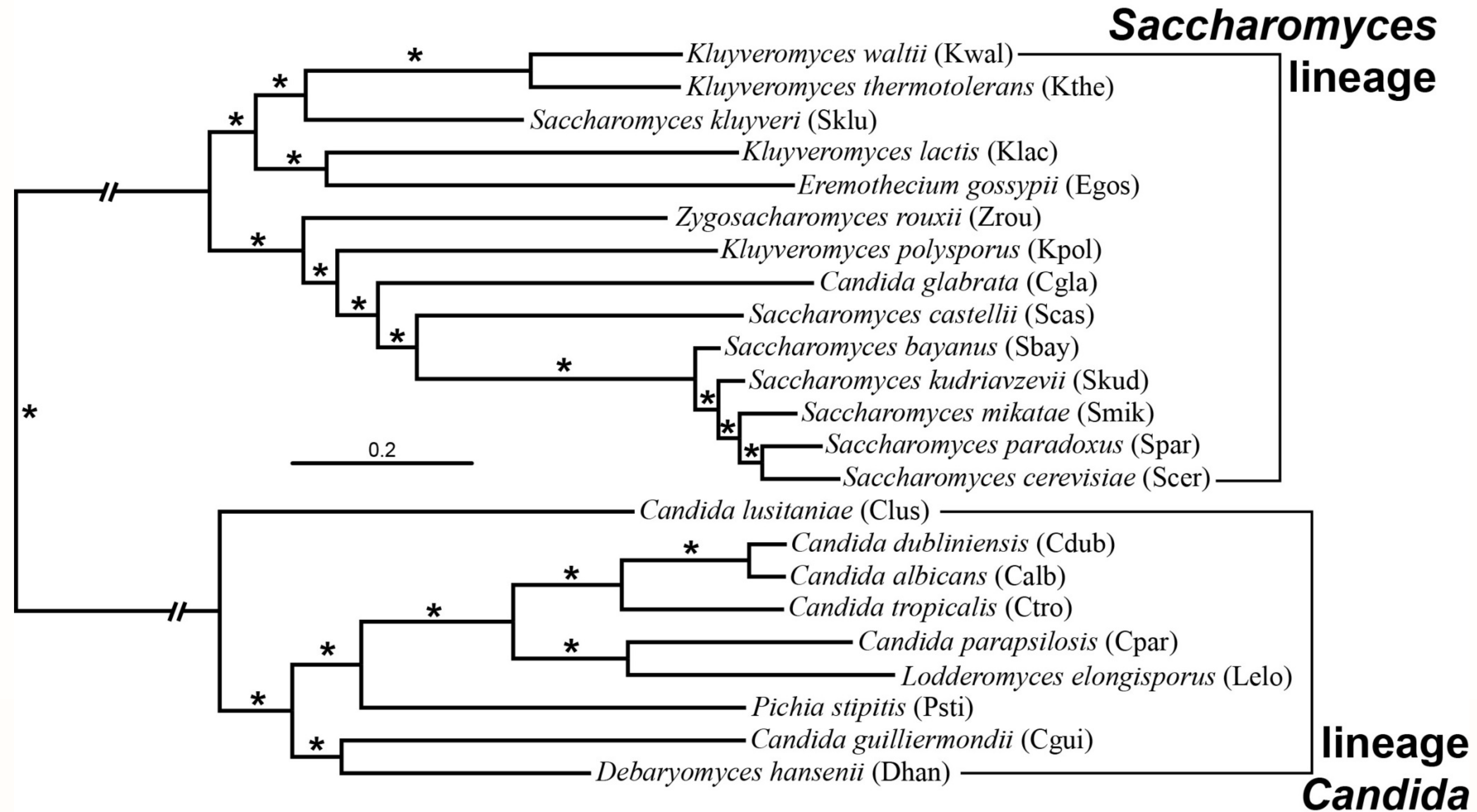
- Phyx
- Newick utilities
- Phylommand
- Gotree

A refresher...

A refresher...

**The next few slides
are from Antonis Rokas**

Concatenation Yields an Absolutely Supported Phylogeny



Bootstrap Support is Misleading When Used in Large Datasets

53% 47%

taxonA AAAAAAAAAATTTTTTTTTT

taxonB AAAAAAAAAACCCCCCCCCC

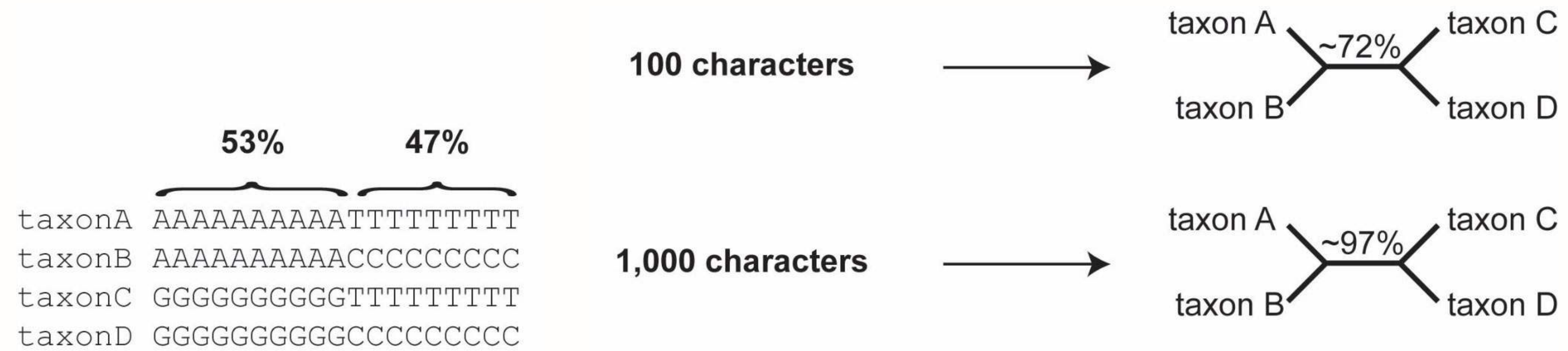
taxonC GGGGGGGGGGTTTTTTTTTT

taxonD GGGGGGGGGGCCCCCCCCCC

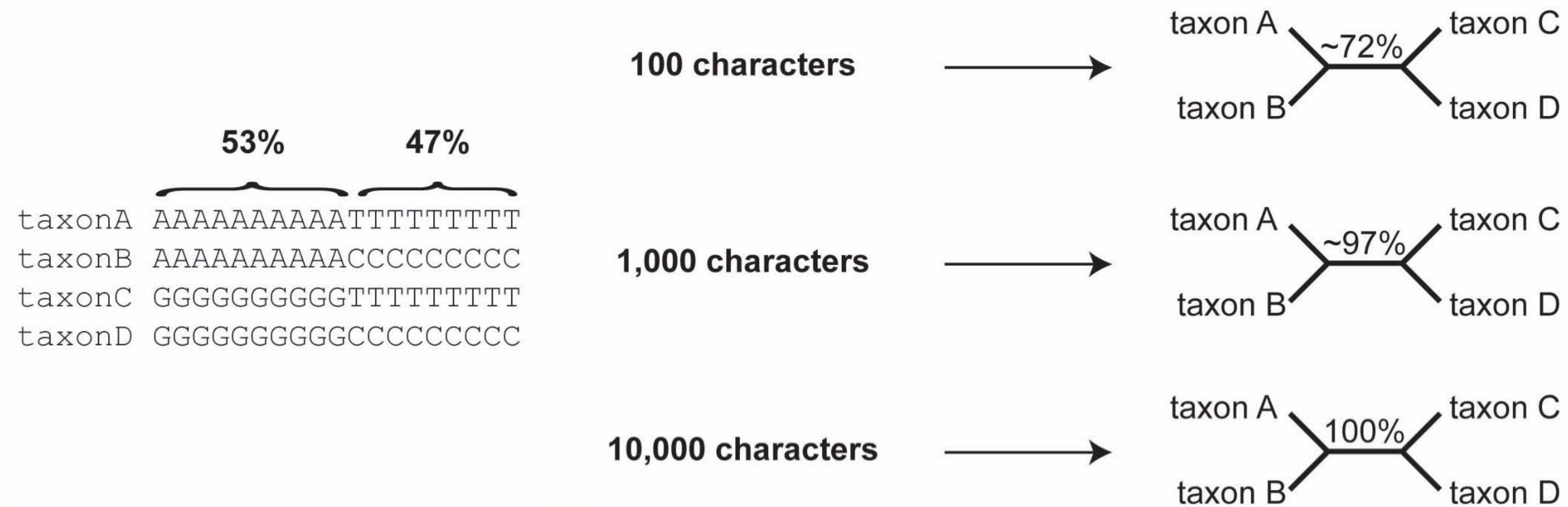
Bootstrap Support is Misleading When Used in Large Datasets



Bootstrap Support is Misleading When Used in Large Datasets



Bootstrap Support is Misleading When Used in Large Datasets

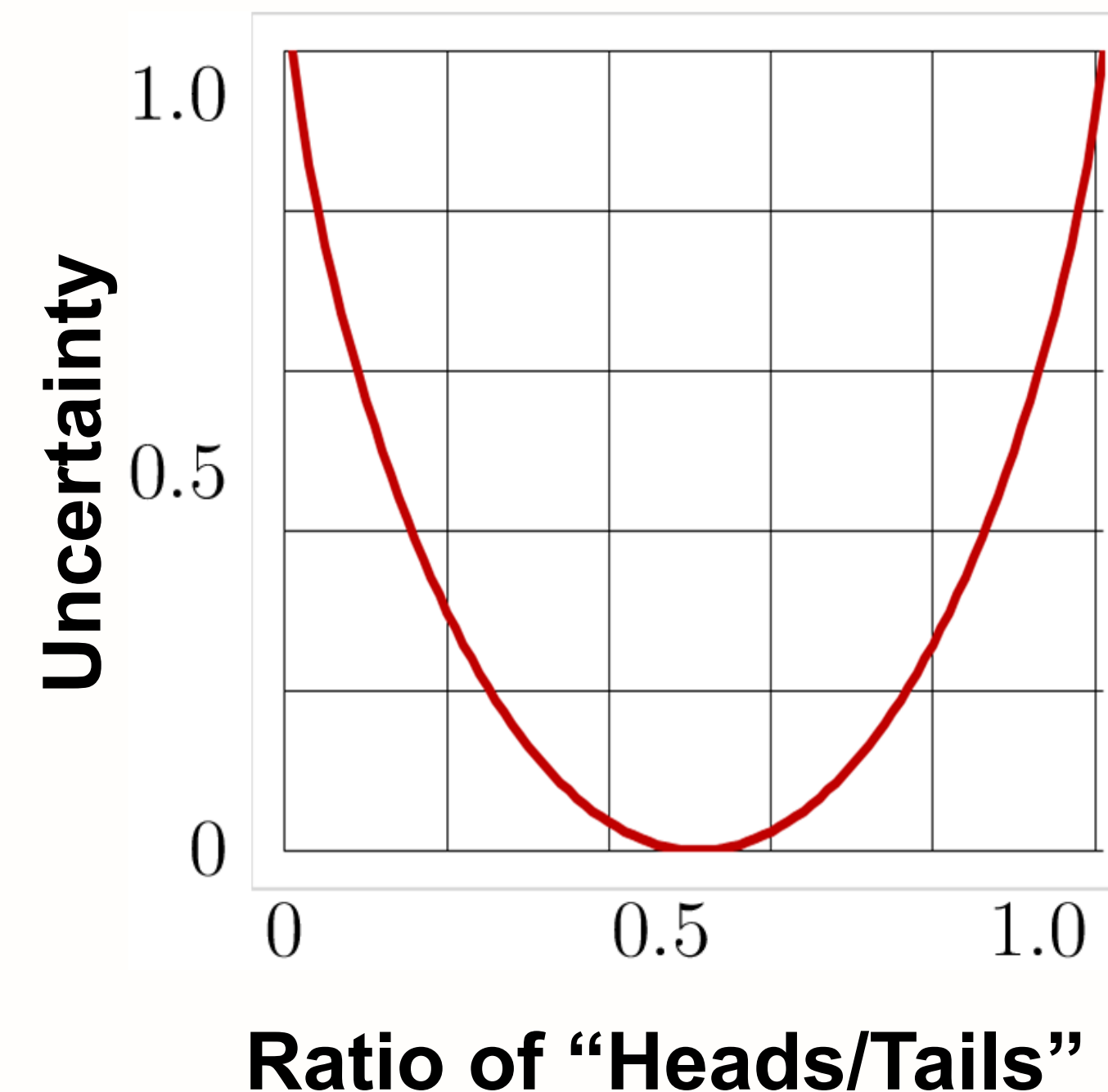


Quantifying Incongruence

Internode Certainty (IC): a measure of the support for a given internode by considering its frequency in a given set of trees jointly with that of the most prevalent conflicting internode in the same set of trees

Tree Certainty (TC): the sum of IC across all internodes

IC and TC are implemented in the latest versions of RAxML

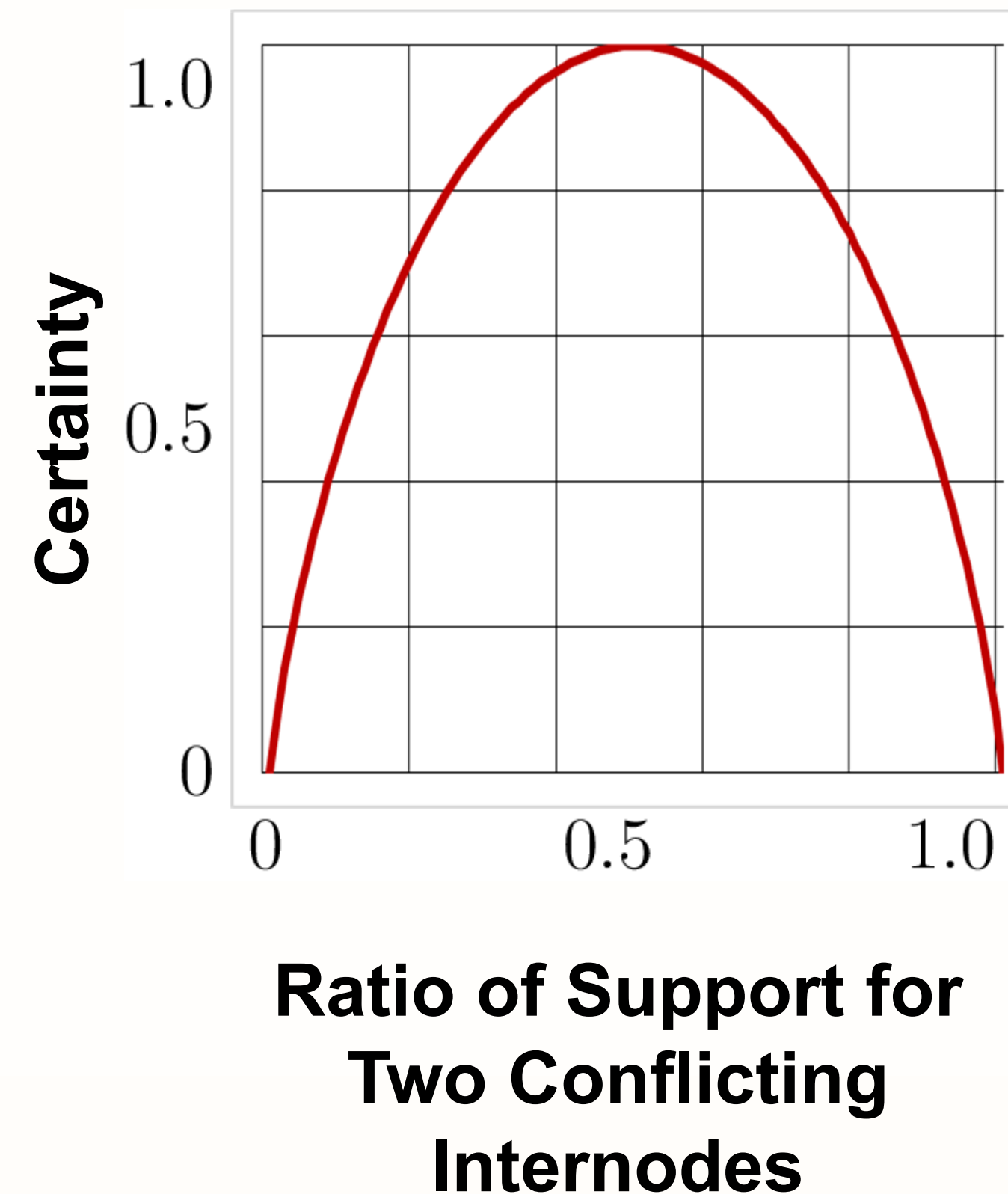


Quantifying Incongruence

Internode Certainty (IC): a measure of the support for a given internode by considering its frequency in a given set of trees jointly with that of the most prevalent conflicting internode in the same set of trees

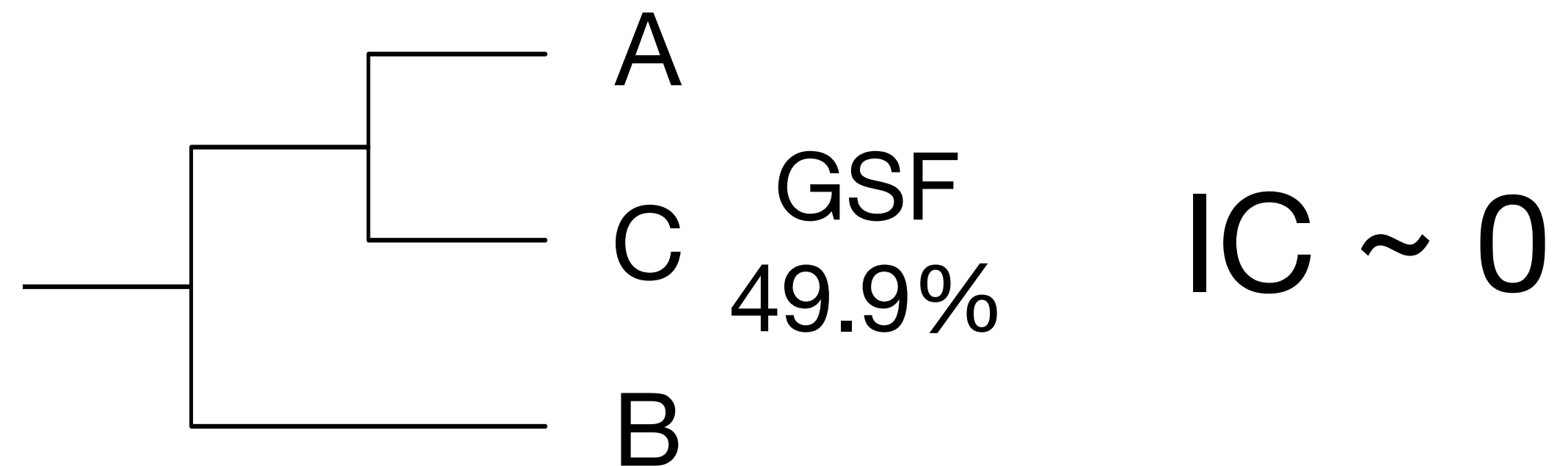
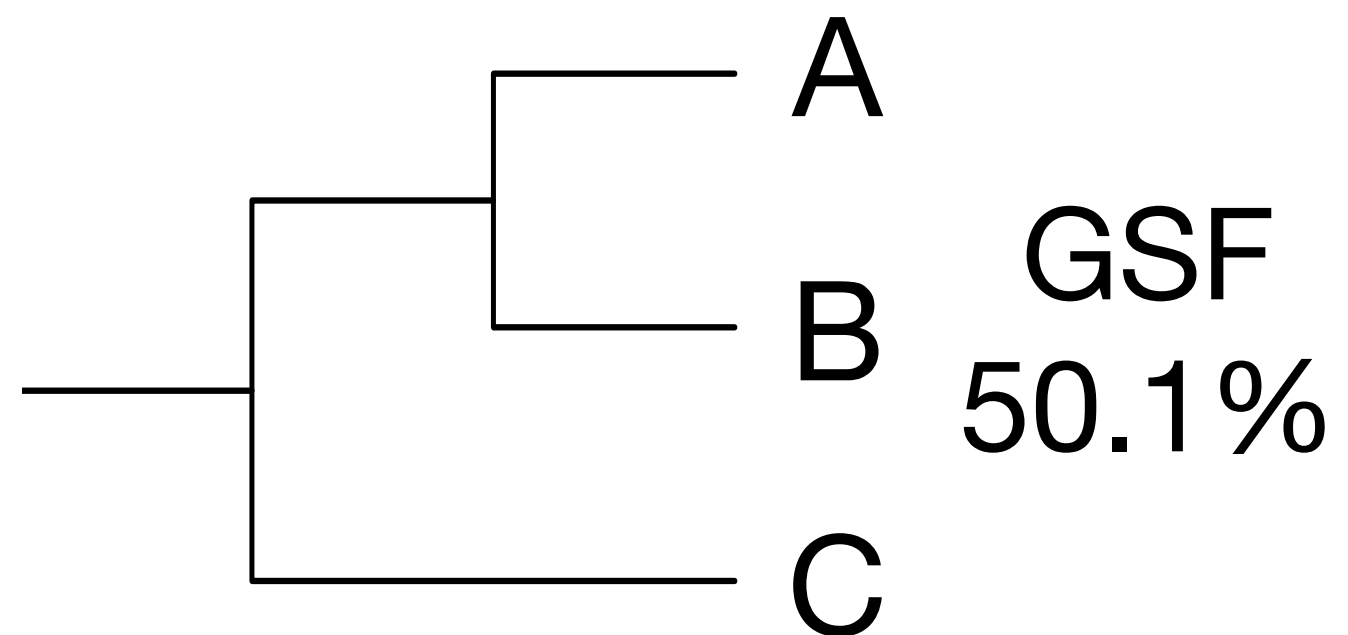
Tree Certainty (TC): the sum of IC across all internodes

IC and TC are implemented in the latest versions of RAxML



Internode certainty (in other words...)

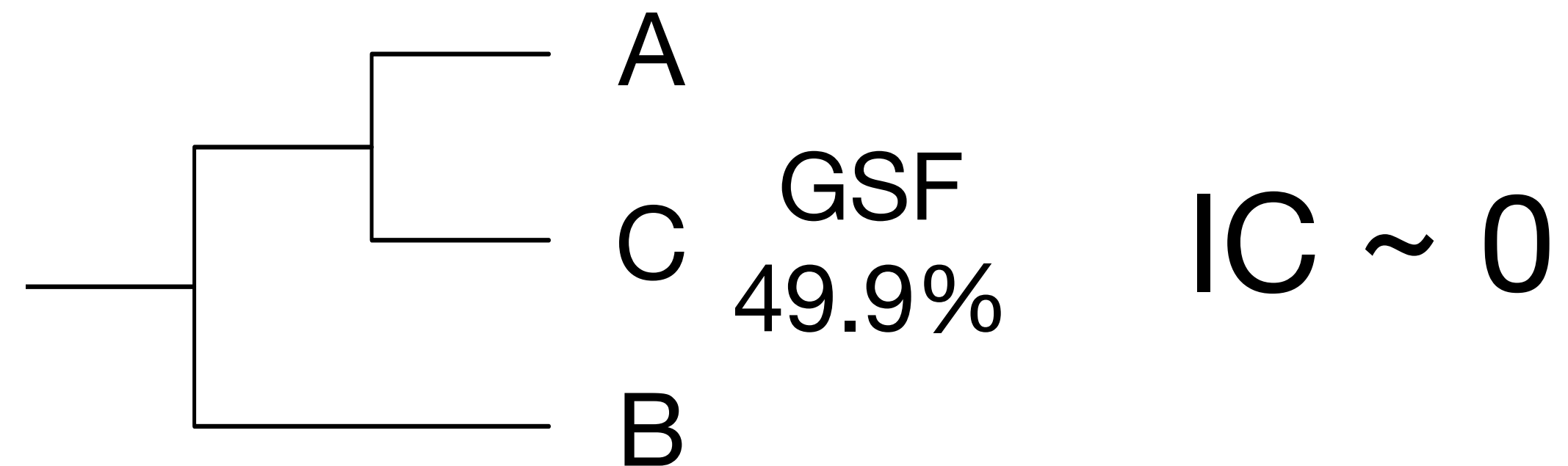
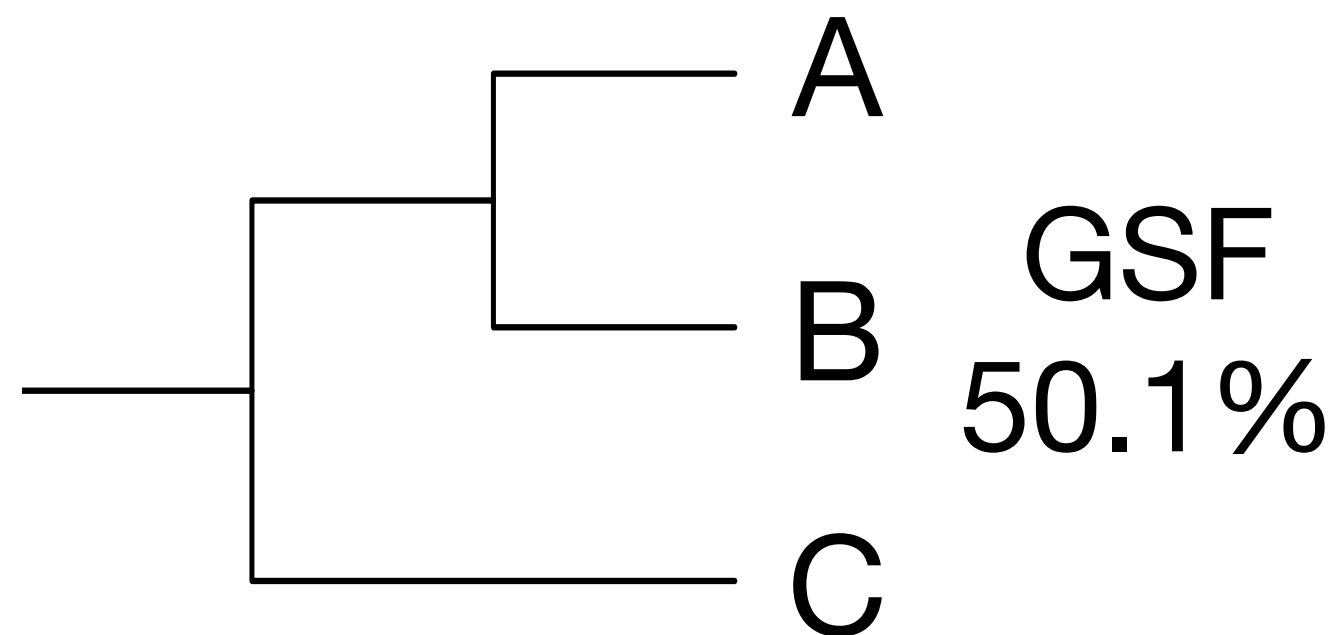
Case 1: High conflict



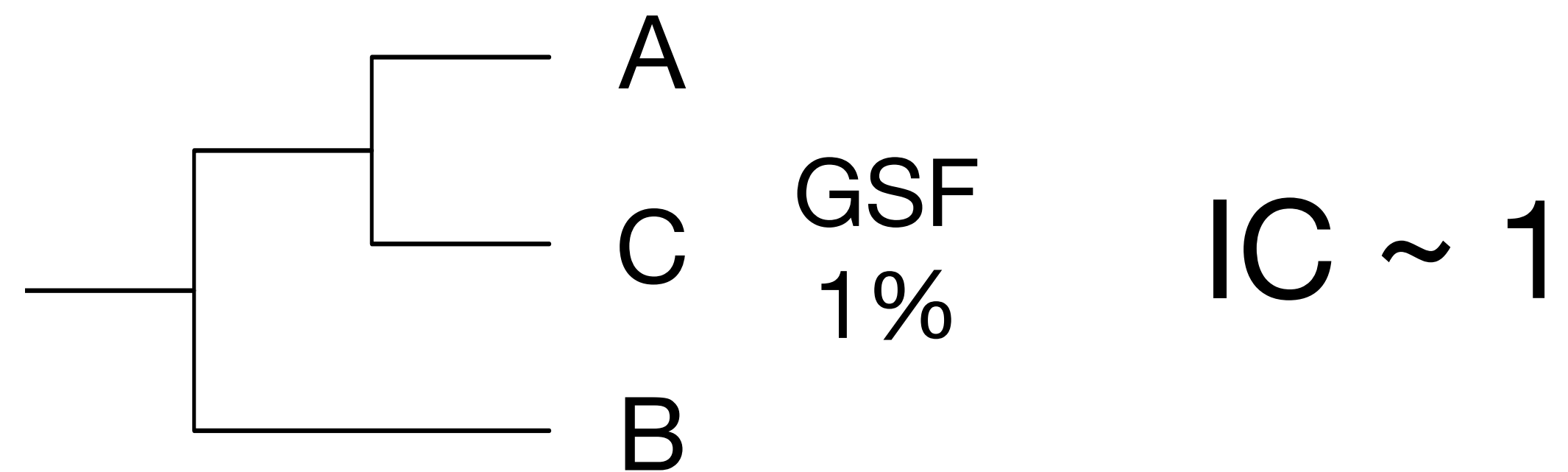
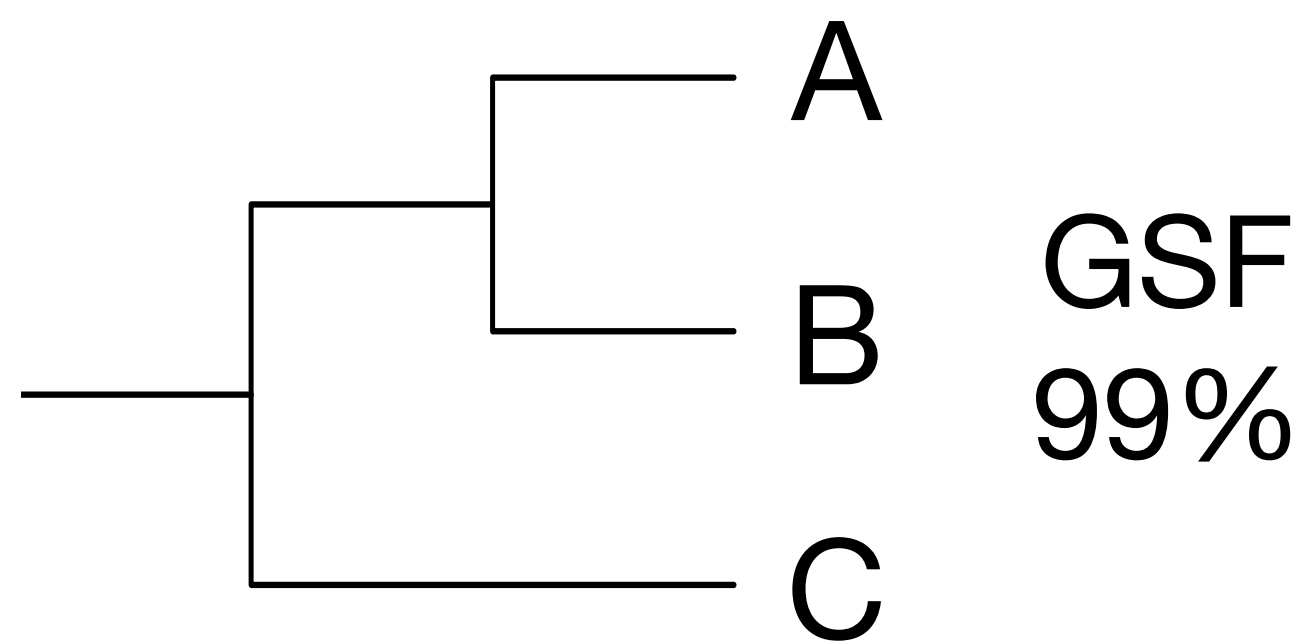
IC ~ 0

Internode certainty (in other words...)

Case 1: High conflict



Case 2: Low conflict



Developments of internode certainty

- The original implementation was originally developed for phylogenies with complete taxon representation
 - **Salichos and Rokas (2013) Nature**
 - **Salichos *et al.* (2014) Mol. Biol. Evol.**

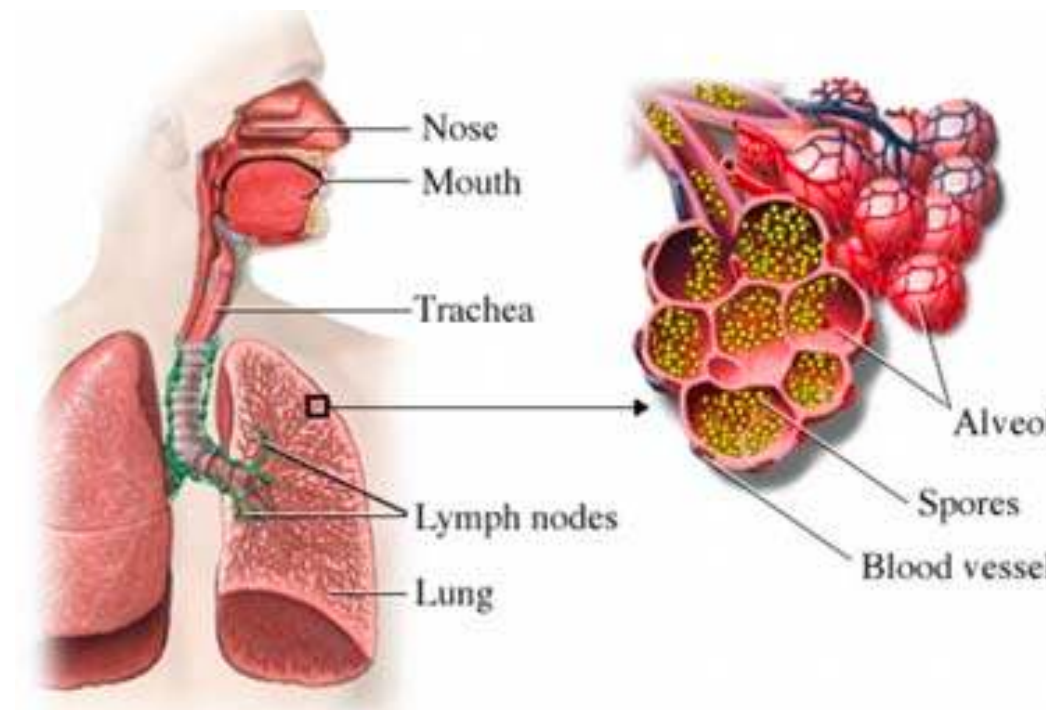
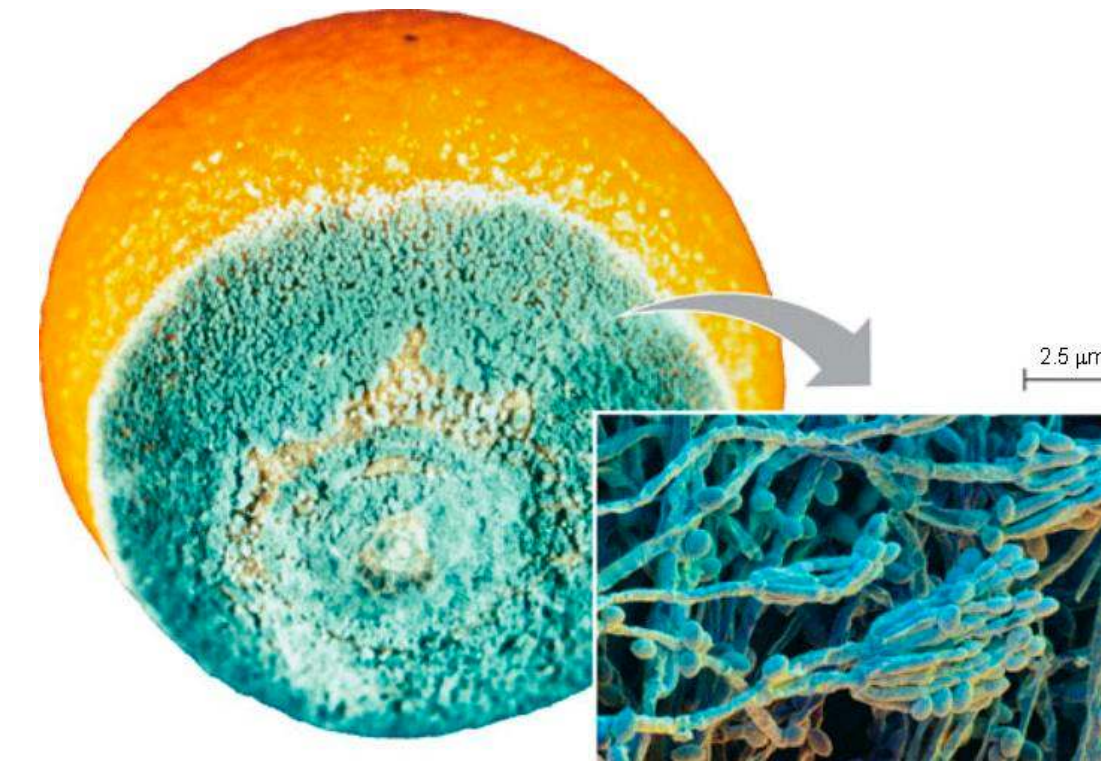
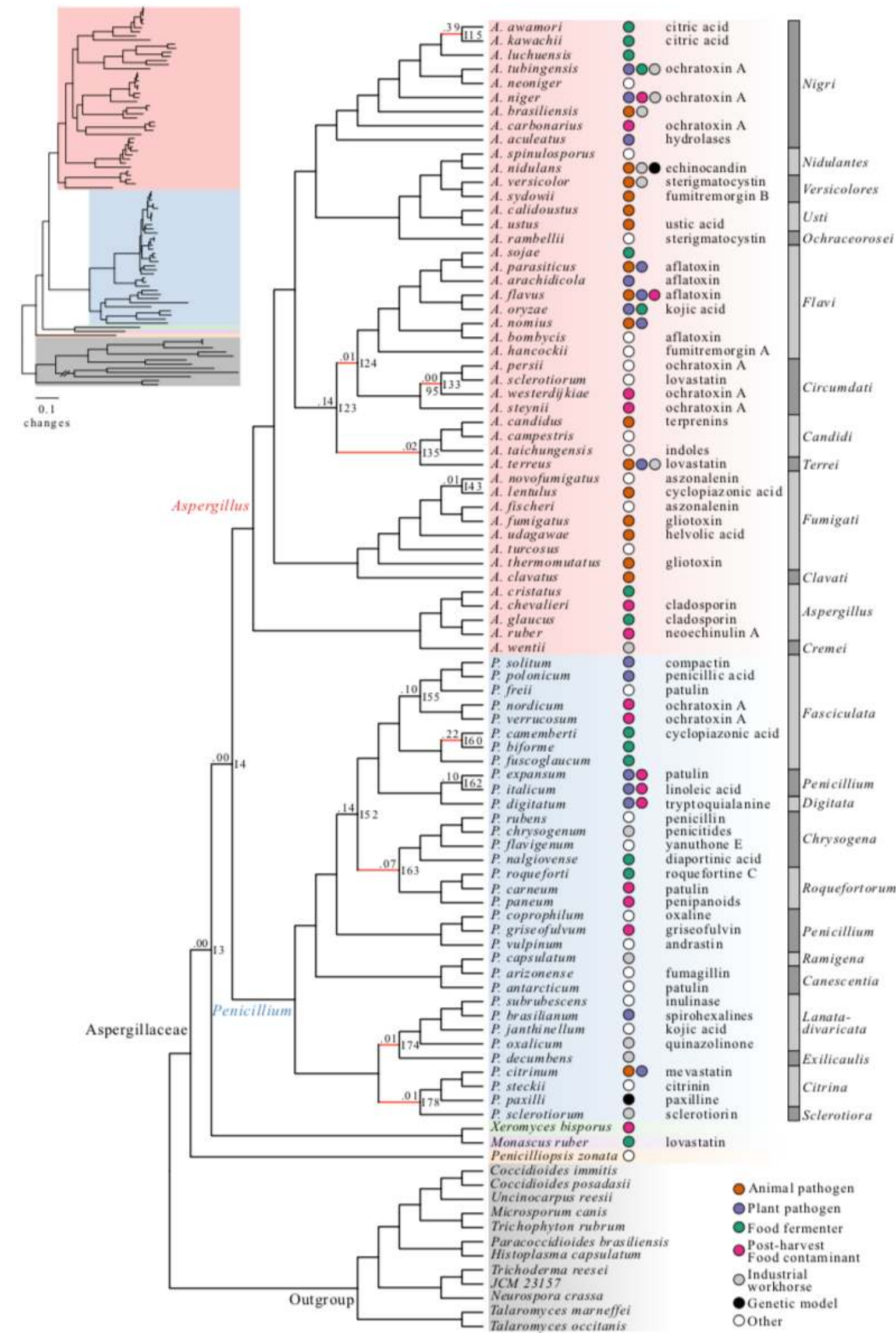
Developments of internode certainty

- The original implementation was originally developed for phylogenies with complete taxon representation
 - **Salichos and Rokas (2013) Nature**
 - **Salichos *et al.* (2014) Mol. Biol. Evol.**
- Corrections for partial taxon representation has been implemented in current versions of RAxML.
 - **Kobert *et al.* (2016) Mol. Biol. Evol.**

Developments of internode certainty

- The original implementation was originally developed for phylogenies with complete taxon representation
 - **Salichos and Rokas (2013) Nature**
 - **Salichos *et al.* (2014) Mol. Biol. Evol.**
- Corrections for partial taxon representation has been implemented in current versions of RAxML.
 - **Kobert *et al.* (2016) Mol. Biol. Evol.**
- Quartet based IC measures, QuartetScores, are more accurate with partial gene trees
 - **Zhou *et al.* (2020) Systematic Biology**

81 genomes from mainly Aspergillus and Penicillium



Notes on implementation

- Verbose usage of RAxML's calculations of IC provides detailed information about conflicting bipartitions

RAxML_verboseSplits.suffix

Notes on implementation

- Verbose usage of RAxML's calculations of IC provides detailed information about conflicting bipartitions

RAxML_verboseSplits.suffix

```
1. Uncinocarpus_reesii
2. Coccidioides_posadasii
3. Penicillium_zeae
4. Xeromyces_bisporus
5. Monascus_ruber
6. Penicillium_camemberti
7. Penicillium_digitatum
8. Penicillium_roqueforti
9. Aspergillus_fumigatus
10. Aspergillus_niger
11. Aspergillus_oryzae

partition:
---** ----- - 1189/92.385392/0.850774
--*-* ***** * 26/2.020202/0.850774
```



Notes on implementation

- Verbose usage of RAxML's calculations of IC provides detailed information about conflicting bipartitions

RAxML_verboseSplits.suffix

```
1. Uncinocarpus_reesii
2. Coccidioides_posadasii
3. Penicillliopsis_zonata
4. Xeromyces_bisporus
5. Monascus_ruber
6. Penicillium_camemberti
7. Penicillium_digitatum
8. Penicillium_roqueforti
9. Aspergillus_fumigatus
10. Aspergillus_niger
11. Aspergillus_oryzae
```

```
partition:
```

```
---**  ----- -   1189/92.385392/0.850774
--*-*  ***** *   26/2.020202/0.850774
```



Taxa names

Notes on implementation

- Verbose usage of RAxML's calculations of IC provides detailed information about conflicting bipartitions

RAxML_verboseSplits.suffix

```
1. Uncinocarpus_reesii
2. Coccidioides_posadasii
3. Penicillium_zeae
4. Xeromyces_bisporus
5. Monascus_ruber
6. Penicillium_camemberti
7. Penicillium_digitatum
8. Penicillium_roqueforti
9. Aspergillus_fumigatus
10. Aspergillus_niger
11. Aspergillus_oryzae
```

partition:

```
---**  -----  -   1189/92.385392/0.850774
--*-*  *****  *   26/2.020202/0.850774
```

Taxa names

Partition information

xx/yy/zz

xx = Trees supporting ref.

yy = gene support freq.

zz = Internode certainty

Notes on implementation

- Verbose usage of RAxML's calculations of IC provides detailed information about conflicting bipartitions

RAxML_verboseSplits.suffix

```
1. Uncinocarpus_reesii
2. Coccidioides_posadasii
3. Penicillium_zeae
4. Xeromyces_bisporus
5. Monascus_ruber
6. Penicillium_camemberti
7. Penicillium_digitatum
8. Penicillium_roqueforti
9. Aspergillus_fumigatus
10. Aspergillus_niger
11. Aspergillus_oryzae

partition:
---** ----- -      1189/92.385392/0.850774
--*-* ***** *      26/2.020202/0.850774
```

*Verbose can only be used
with trees that have full
taxon representation*

Taxa names

Partition information

xx/yy/zz

xx = Trees supporting ref.

yy = gene support freq.

zz = Internode certainty

Notes on implementation

- Exact bipartition topology for a given bipartition can be examined among files with the following syntax
RAxML_verbose/C.suffix.0 ... RAxML_verbose/C.suffix.N-1

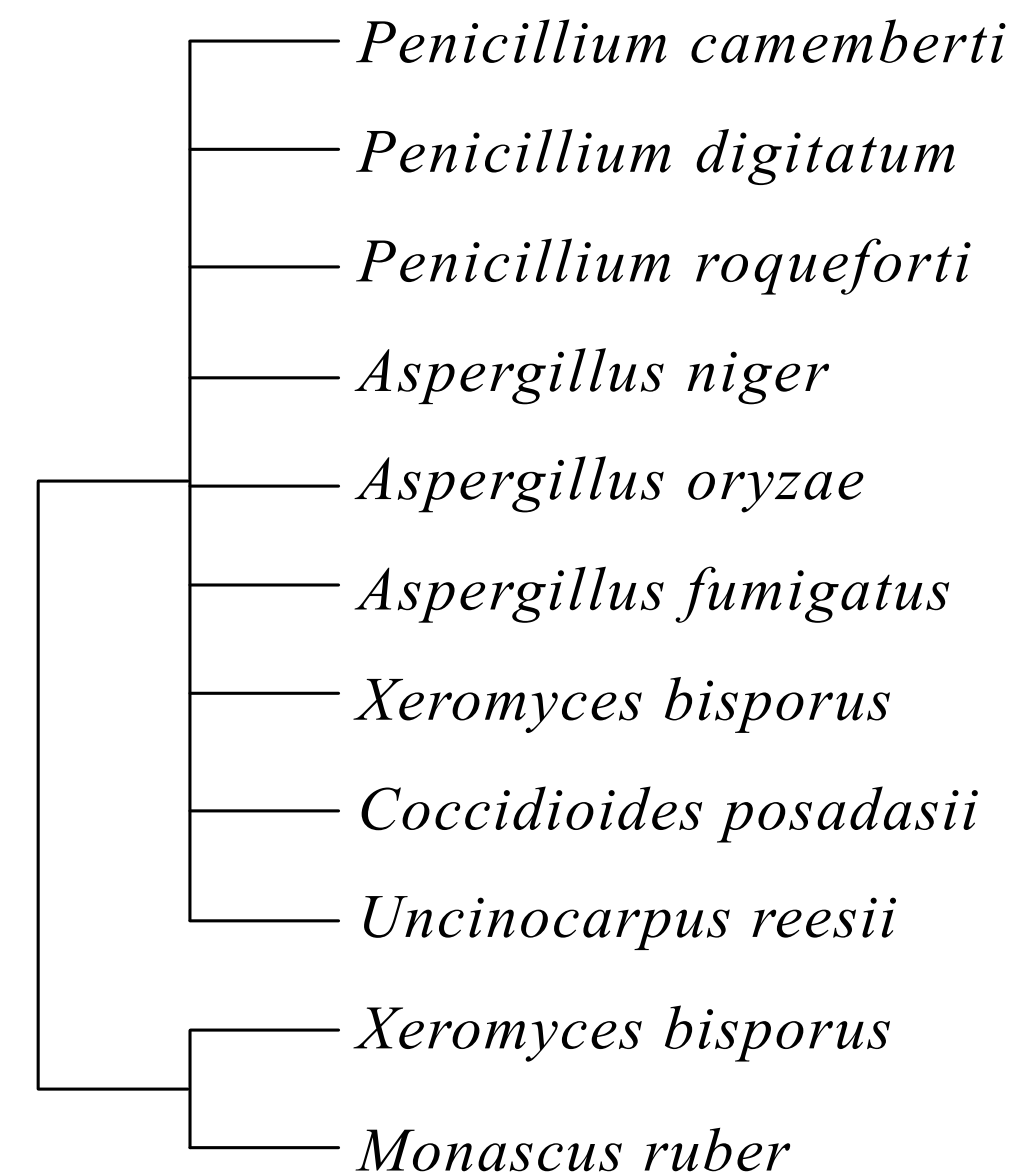
```
1. Uncinocarpus_reesii
2. Coccidioides_posadasii
3. Penicillium_izonata
4. Xeromyces_bisporus
5. Monascus_ruber
6. Penicillium_camemberti
7. Penicillium_digitatum
8. Penicillium_roqueforti
9. Aspergillus_fumigatus
10. Aspergillus_niger
11. Aspergillus_oryzae
```

Notes on implementation

- Exact bipartition topology for a given bipartition can be examined among files with the following syntax
RAxML_verbose/C.suffix.0 ... RAxML_verbose/C.suffix.N-1

```
1. Uncinocarpus_reesii
2. Coccidioides_posadasii
3. Penicillium_camemberti
4. Penicillium_digitatum
5. Penicillium_roqueforti
6. Aspergillus_fumigatus
7. Aspergillus_niger
8. Aspergillus_oryzae
9. Xeromyces_bisporus
10. Monascus_ruber
```

Topology 1



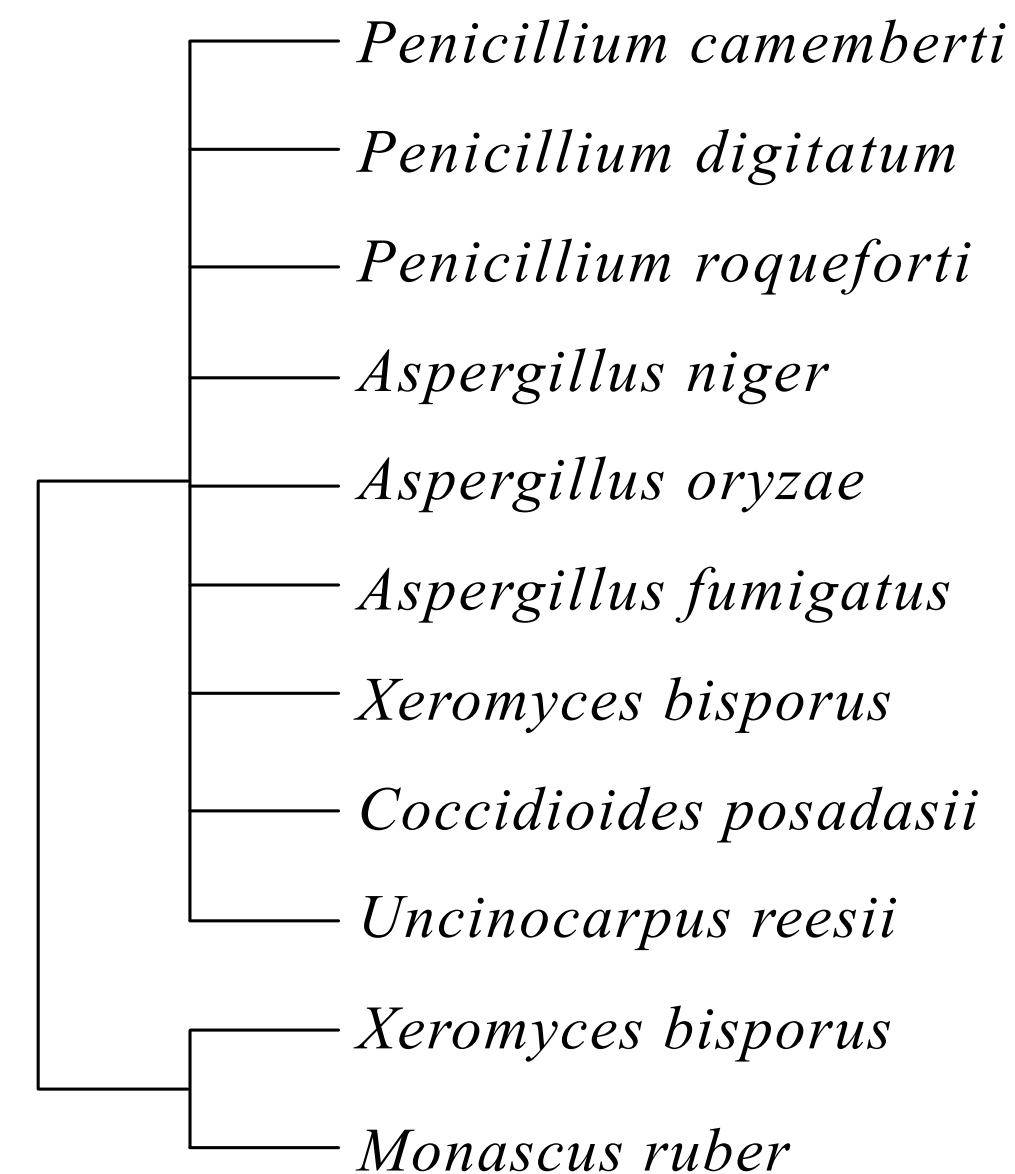
----**-----
1189/92.385392/0.850774

Notes on implementation

- Exact bipartition topology for a given bipartition can be examined among files with the following syntax
RAxML_verbose/C.suffix.0 ... RAxML_verbose/C.suffix.N-1

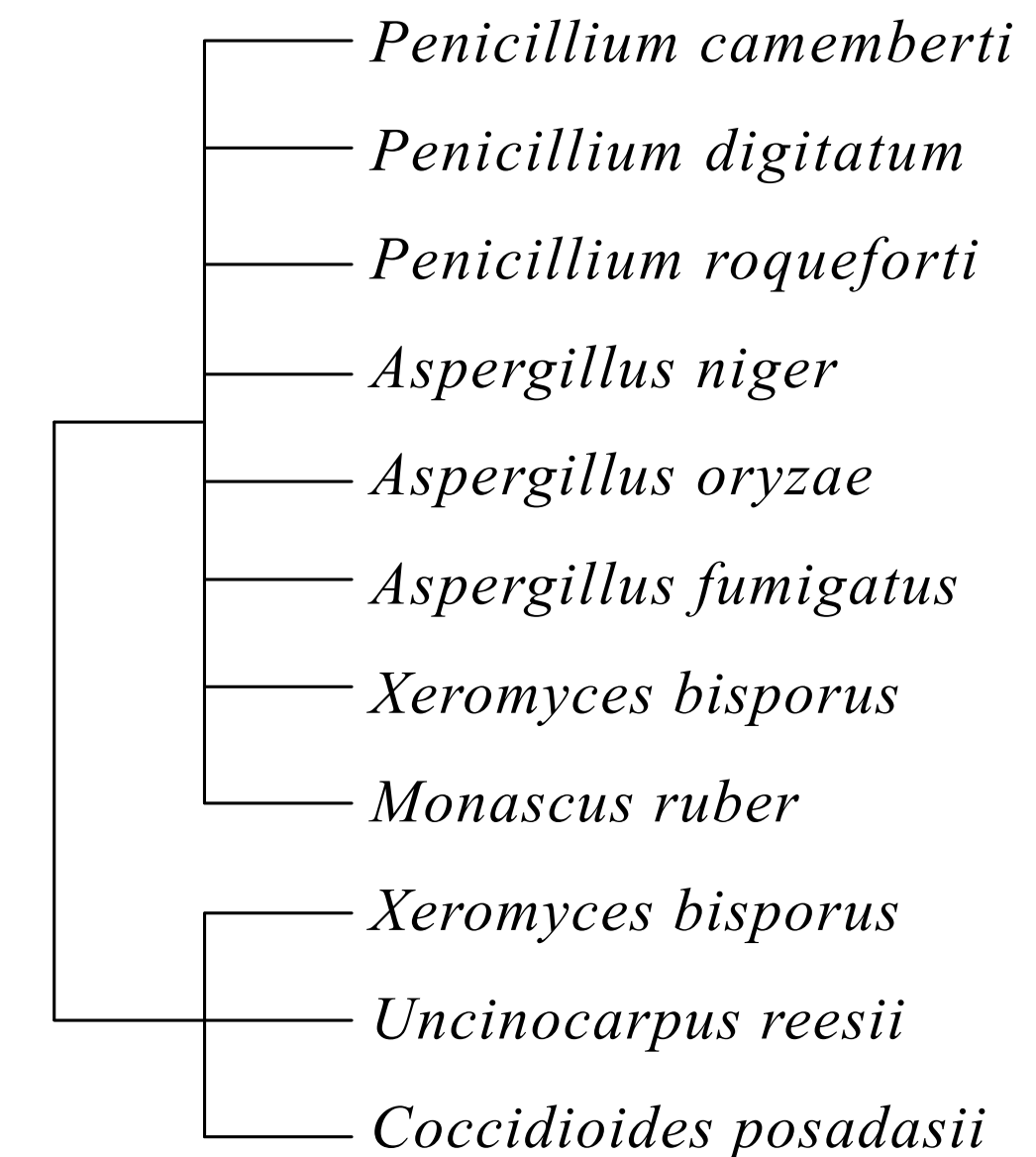
```
1. Uncinocarpus_reesii
2. Coccidioides_posadasii
3. Penicilliopsis_zonata
4. Xeromyces_bisporus
5. Monascus_ruber
6. Penicillium_camemberti
7. Penicillium_digitatum
8. Penicillium_roqueforti
9. Aspergillus_fumigatus
10. Aspergillus_niger
11. Aspergillus_oryzae
```

Topology 1



----** -----
1189/92.385392/0.850774

Topology 2

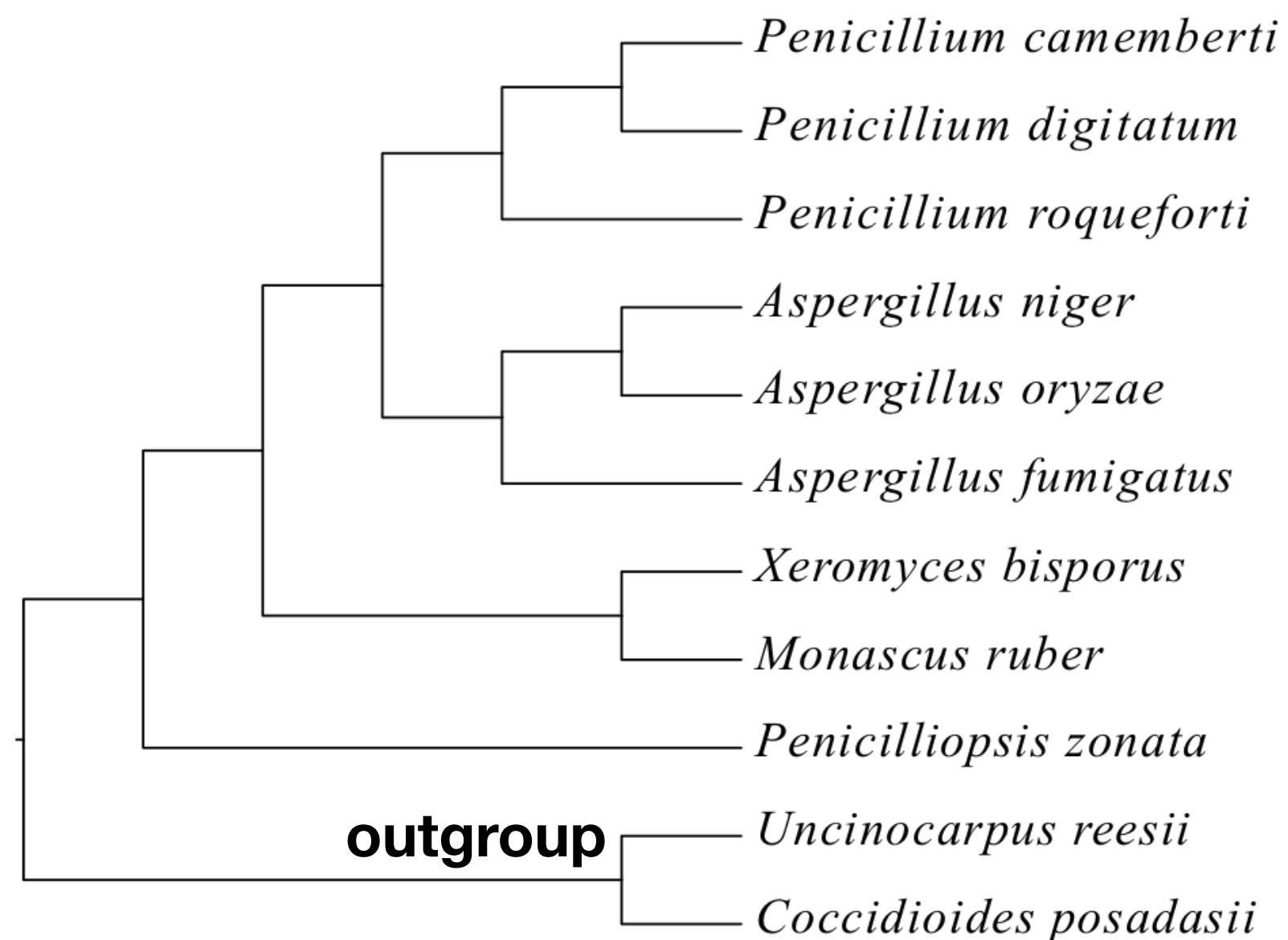


---*- * * * * *
26/2.020202/0.850774

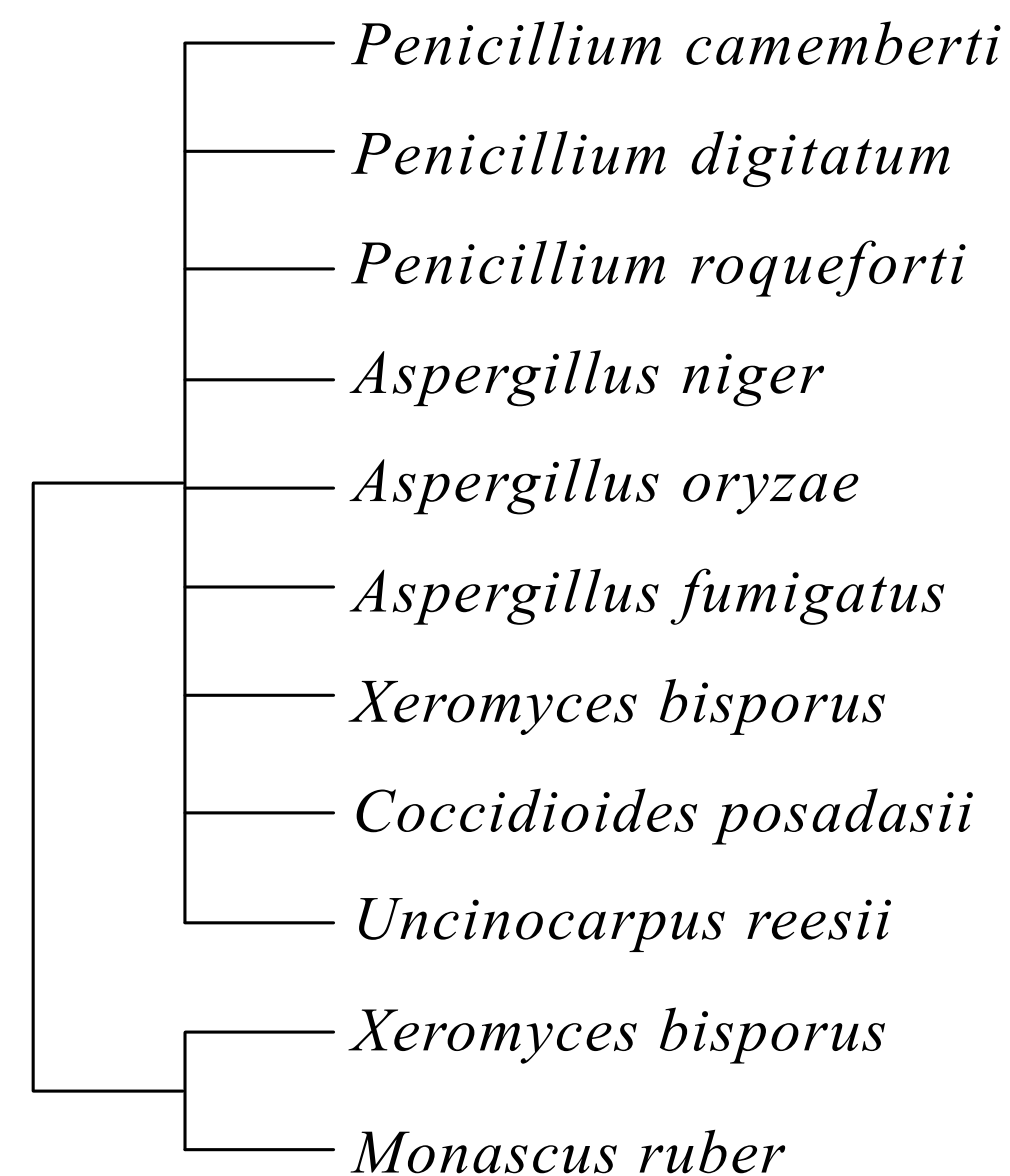
Notes on implementation

- Exact bipartition topology for a given bipartition can be examined among files with the following syntax
RAxML_verbose/C.suffix.0 ... RAxML_verbose/C.suffix.N-1

Concat topology

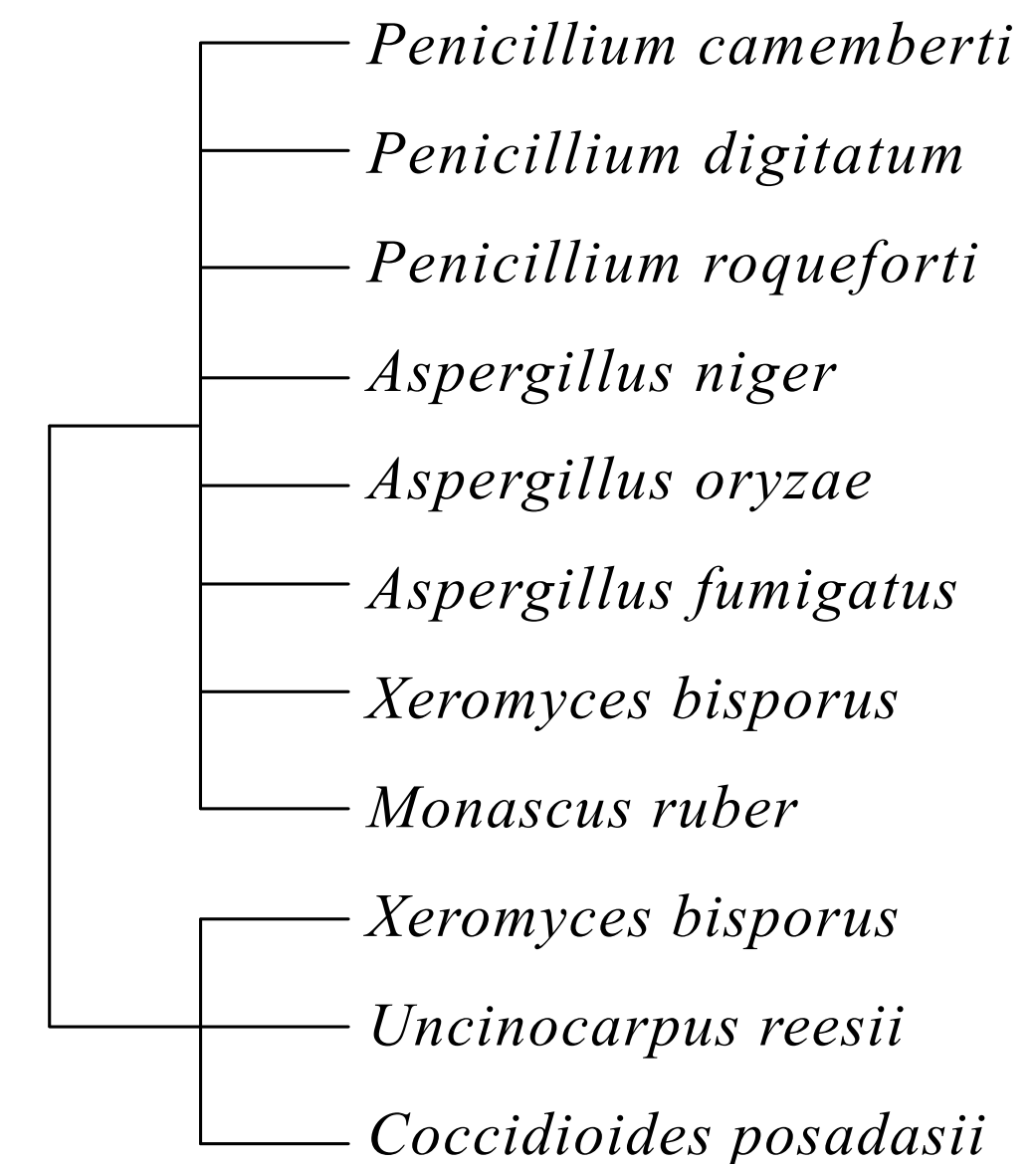


Topology 1



---** --- -
1189/92.385392/0.850774

Topology 2



---** ***** *
26/2.020202/0.850774

Measures related to internode certainty

IC-All

- computed by taking into account all conflicting bipartitions with that have $\geq 5\%$ support and not only the most supported conflict

TC-All

- The sum of IC-All values

Relative tree certainty

- A value from 0 (no certainty) to 1 (high certainty)

Concordance factors

Concordance factors

- proportion of the genome for which a given clade is true

Baum (2007) Taxon

Gene or site concordance factors

- proportion of genes or sites for which a given clade is true
- more precisely, percentage of decisive gene trees (or sites) with a given branch

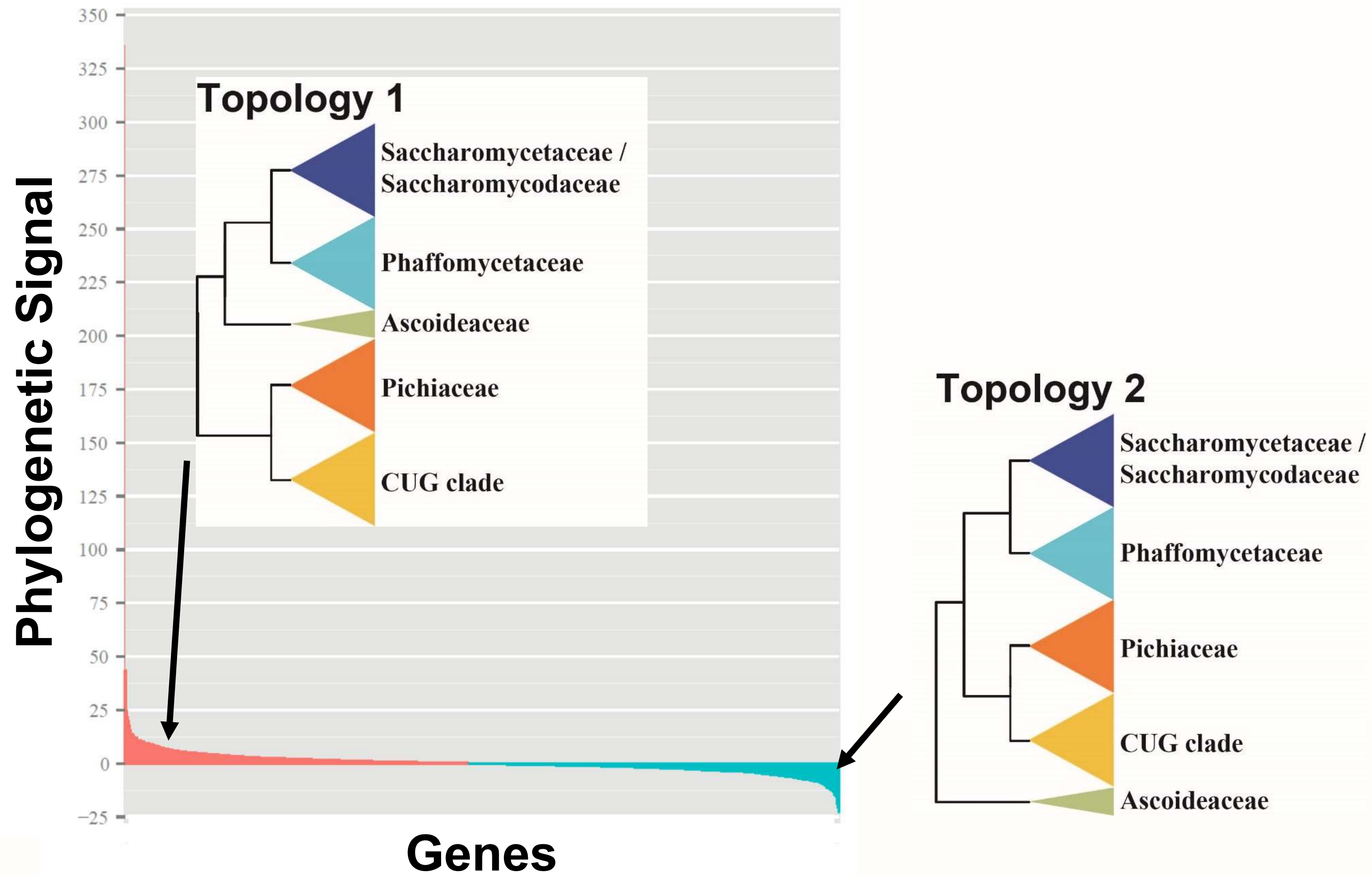
Minh (2020) Molecular Biology and Evolution

**Great additional analysis when
bootstrapping becomes unreliable!!**

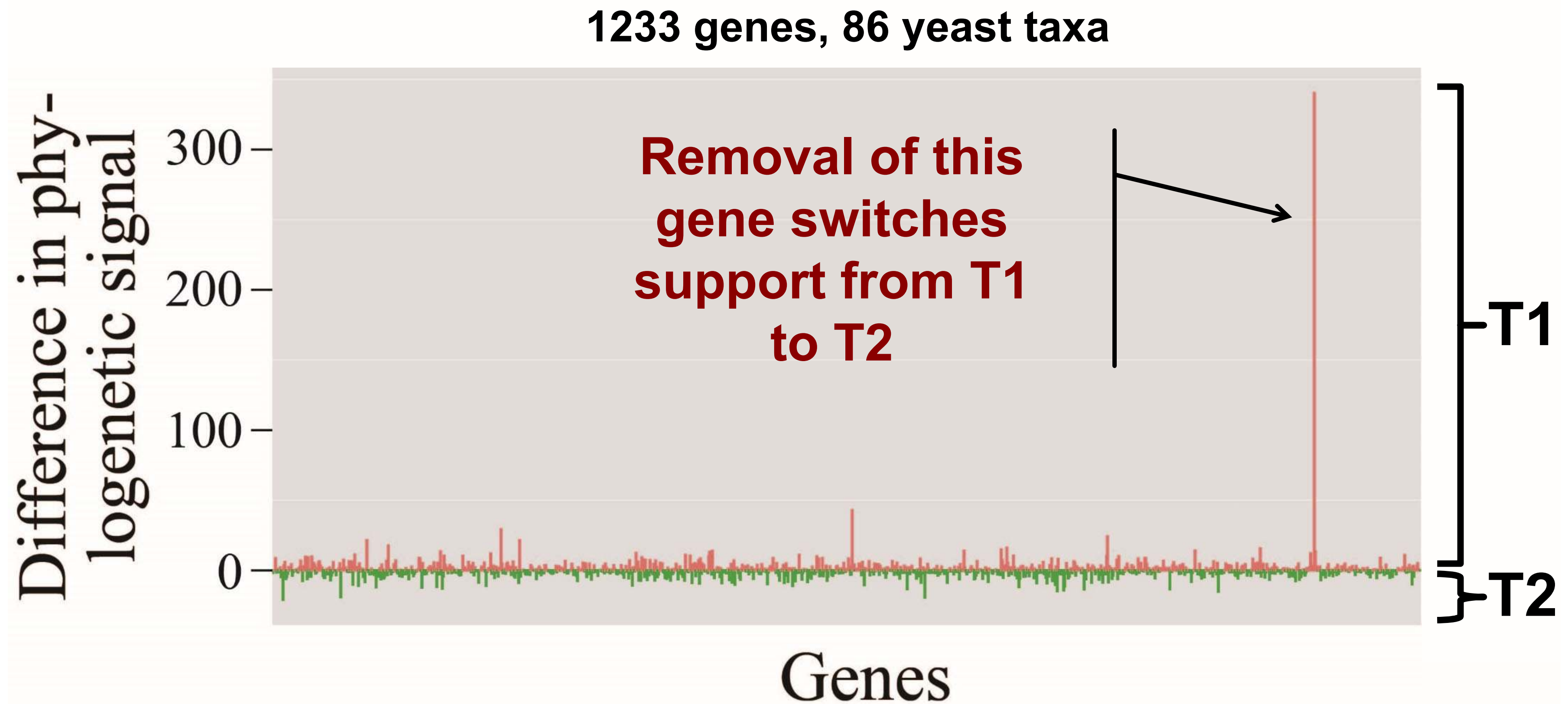
A refresher...

**The next few slides
are from Antonis Rokas**

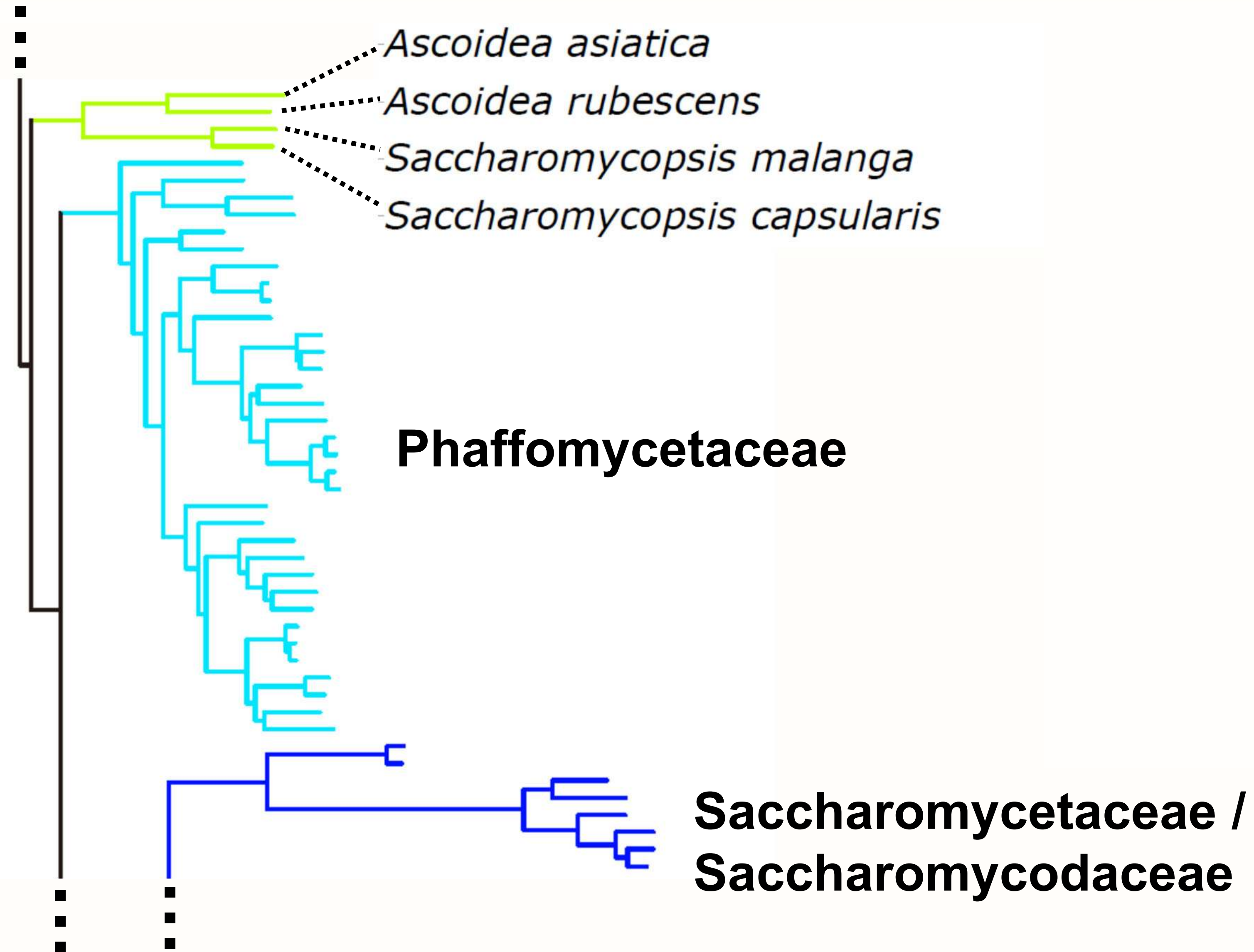
...But in Others It Stems from One or Two Genes



Phylogenetic Signal per Gene for the Two Hypotheses

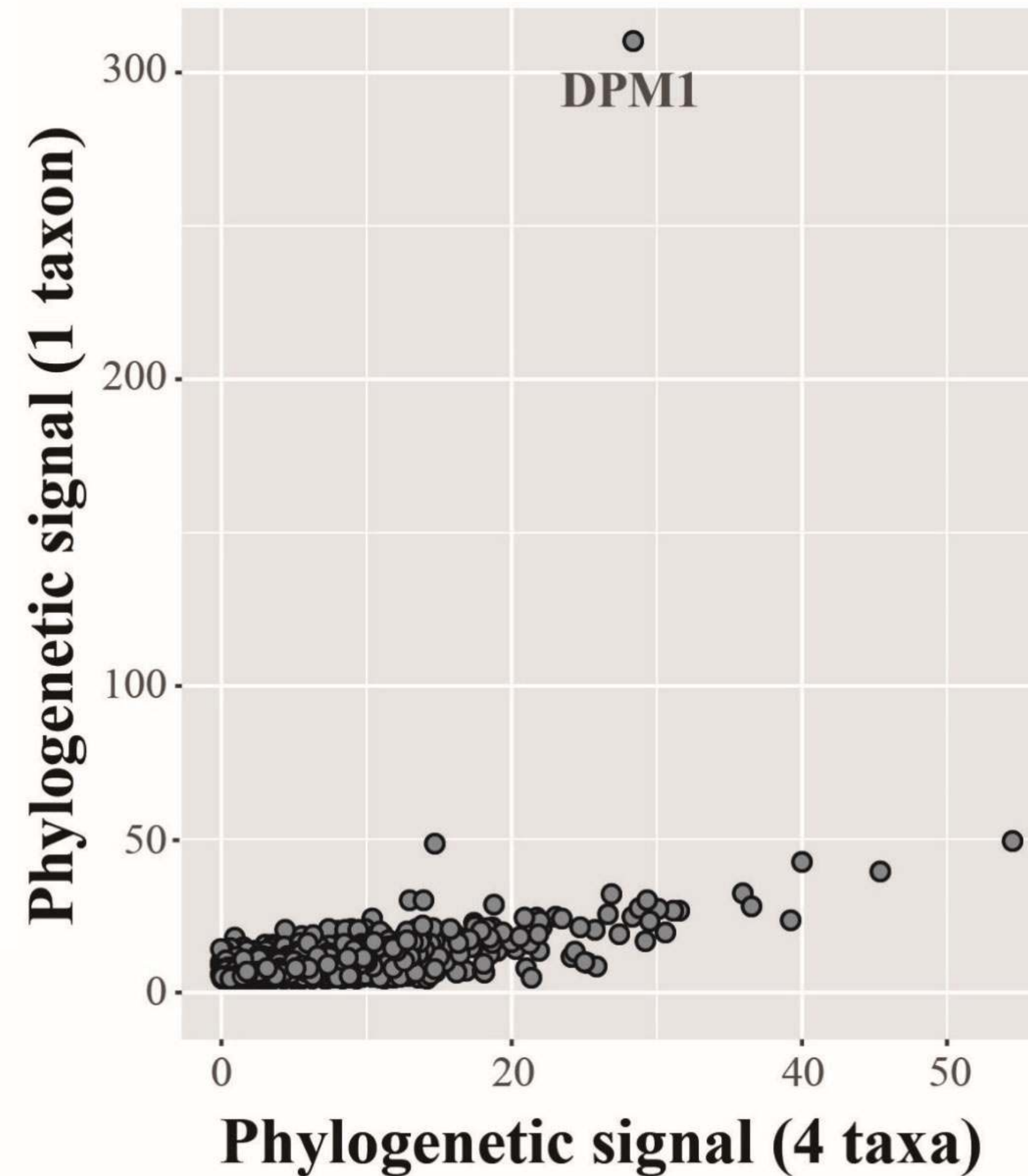


Sampling of 3 Additional Taxa “Breaks” the Long Branch



Sampling of 3 Additional Taxa Decreases Gene's Signal

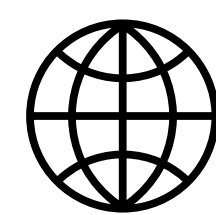
2,408 genes, 329 – 332 yeast taxa



Internode certainty and related measures



@JLSteenwyk



<https://jlsteenwyk.com/>

Quiz time :D



Internode certainty and related measures

- When viewing the trees, root on the clade with *U. reesii* and *C. posadasii*

4iii)

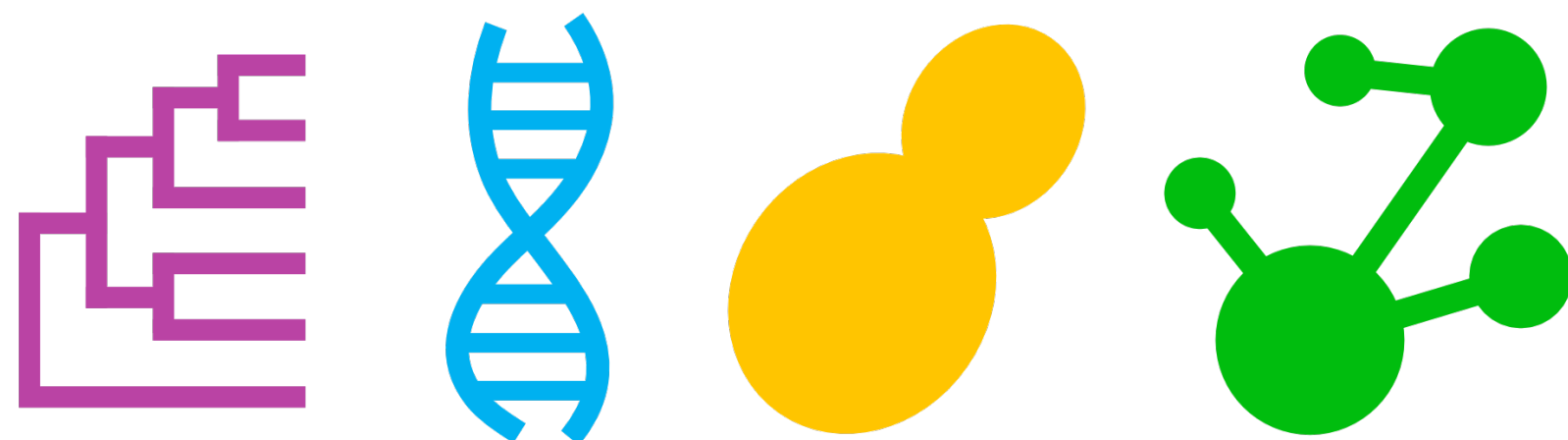
- Don't forget to put the path to IQ-Tree

5i)

- “Do not execute the following commands” only pertains to 5i

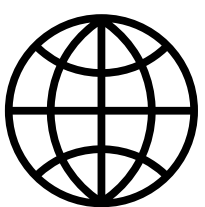
5xi)

- This command takes some time...try subsetting the data before reading it into *R*



Jacob L. Steenwyk

 @JLSteenwyk

 <https://jlsteenwyk.com/>

- Tree certainty:

4.190492

- Relative tree certainty:

0.523811

- Tree certainty all including all conflicting

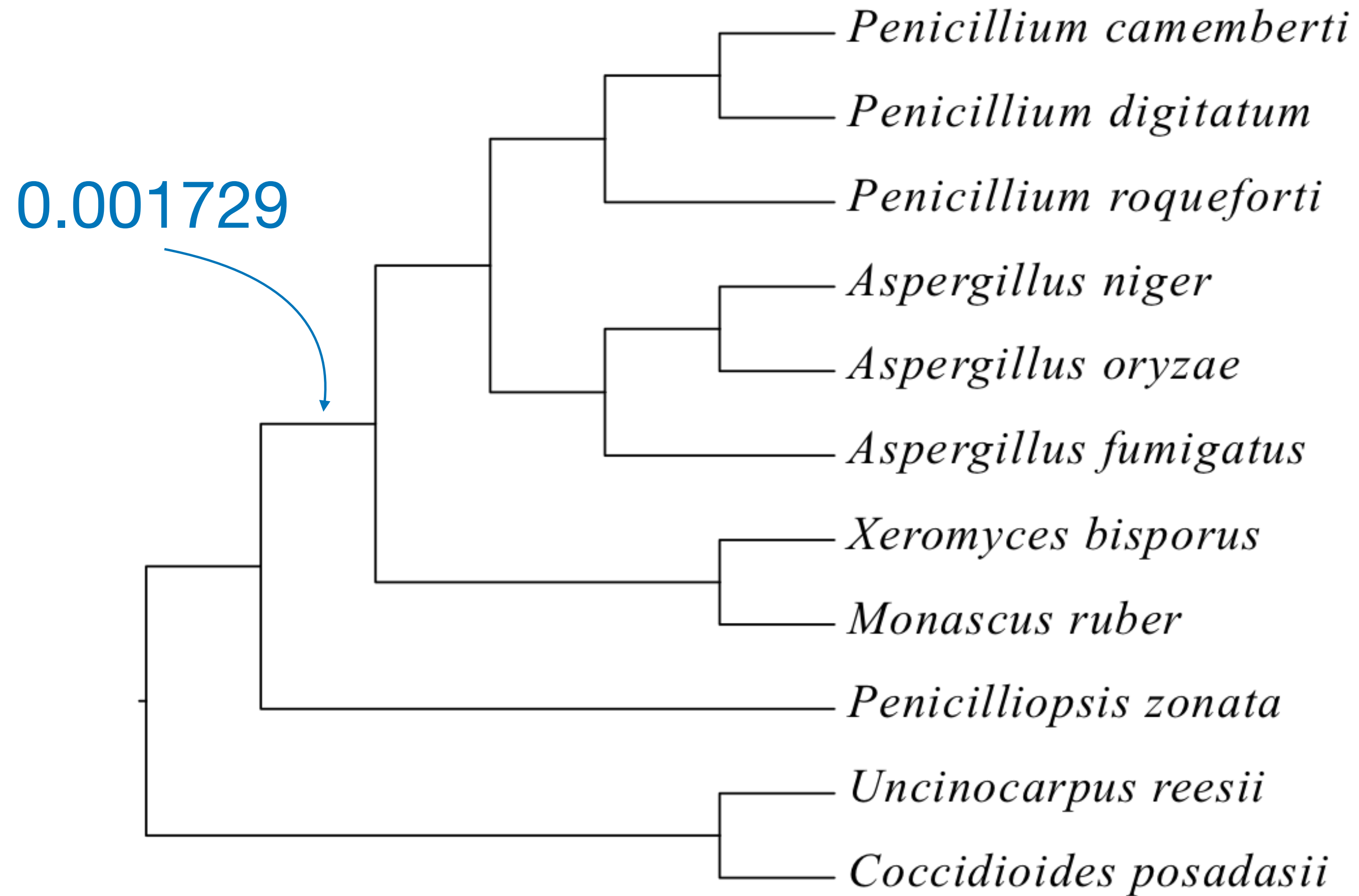
bipartitions:

4.128041

- Relative tree certainty all including all conflicting

bipartitions:

0.516005



What is the number of genes and gene support frequency that support the two topologies?

First topology:

375/29.137529

Second topology:

340/26.418026

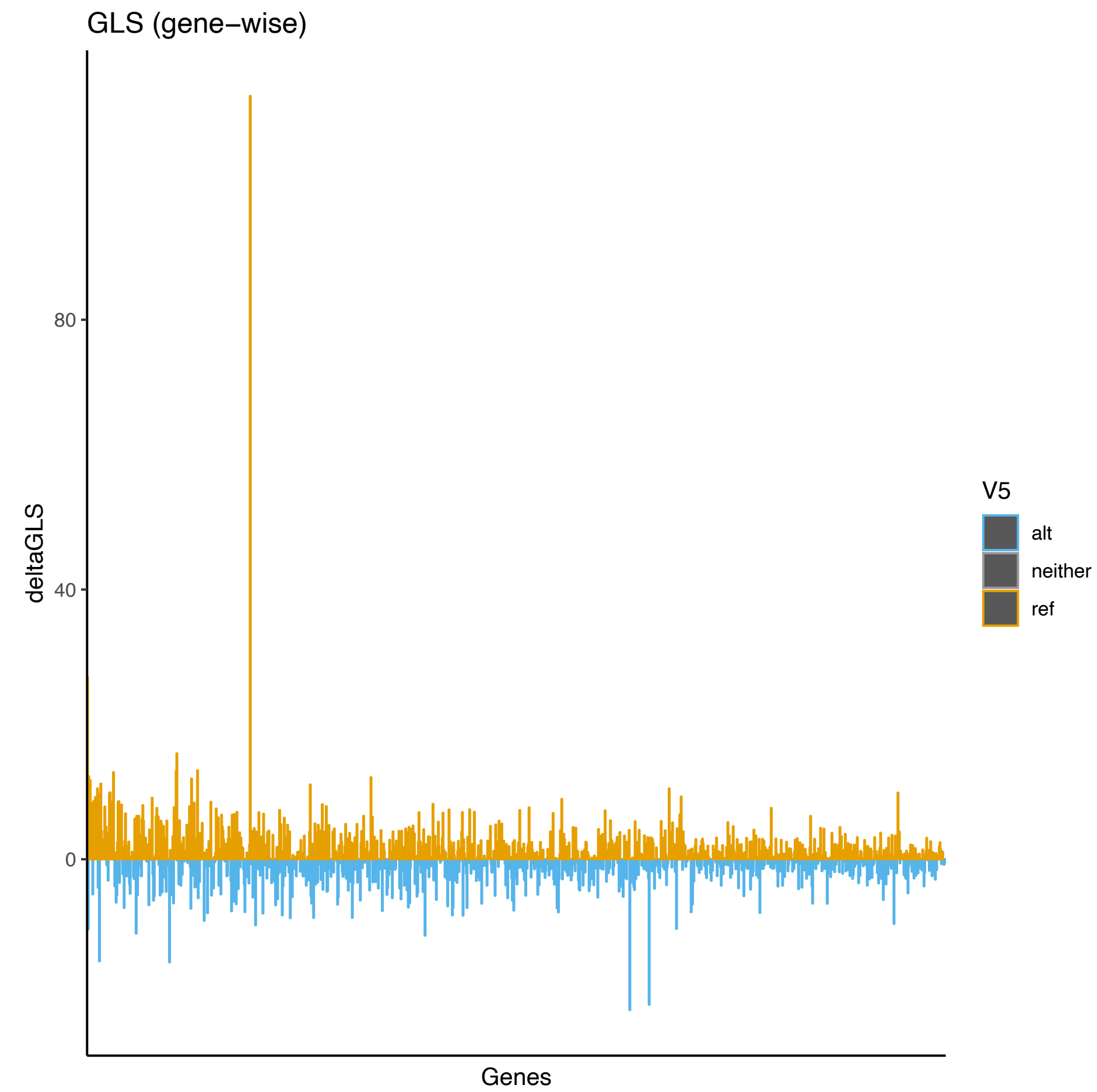
What is gene and site concordance factor values for the reference topology and alternative topology?

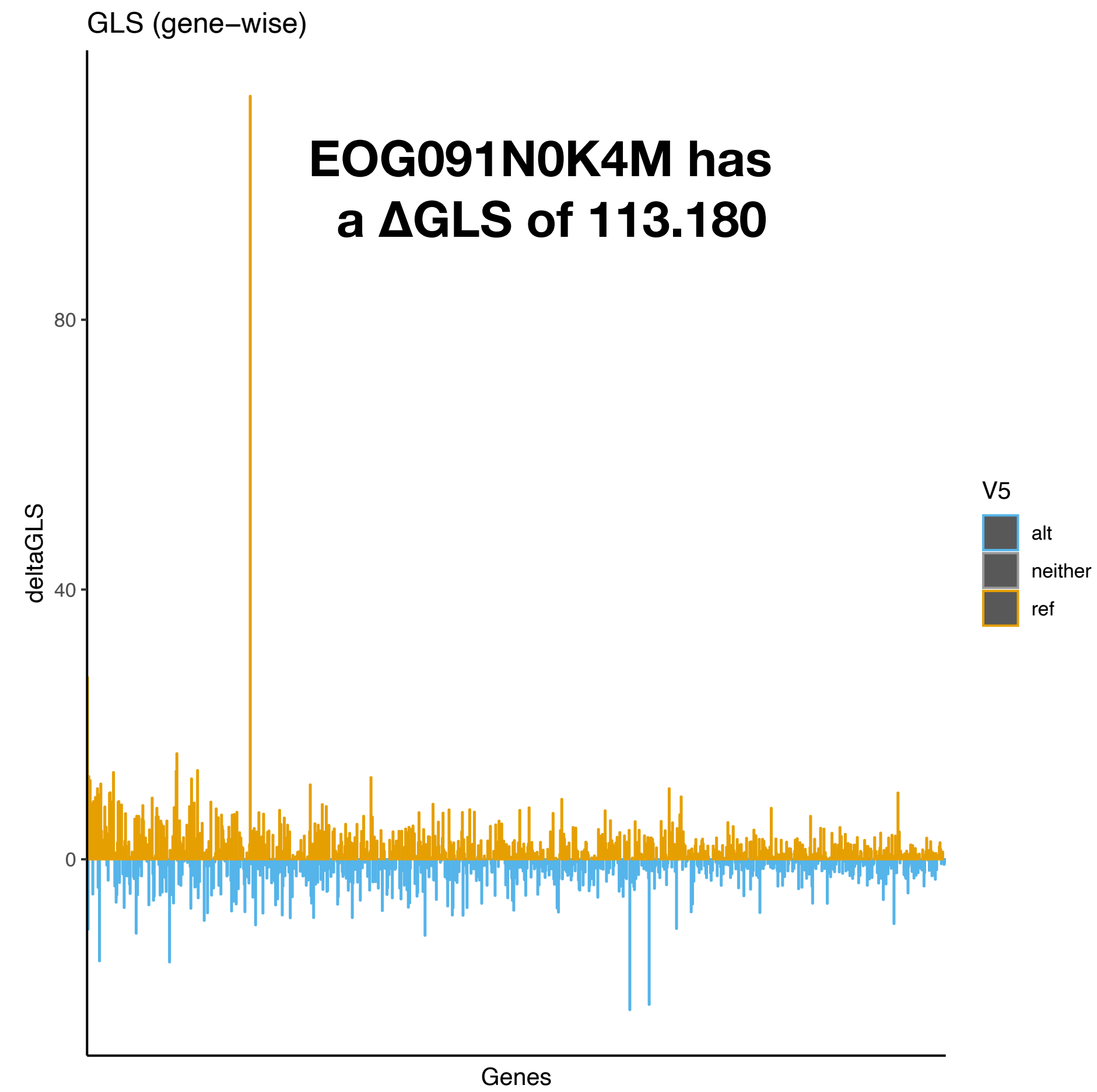
Reference topology:

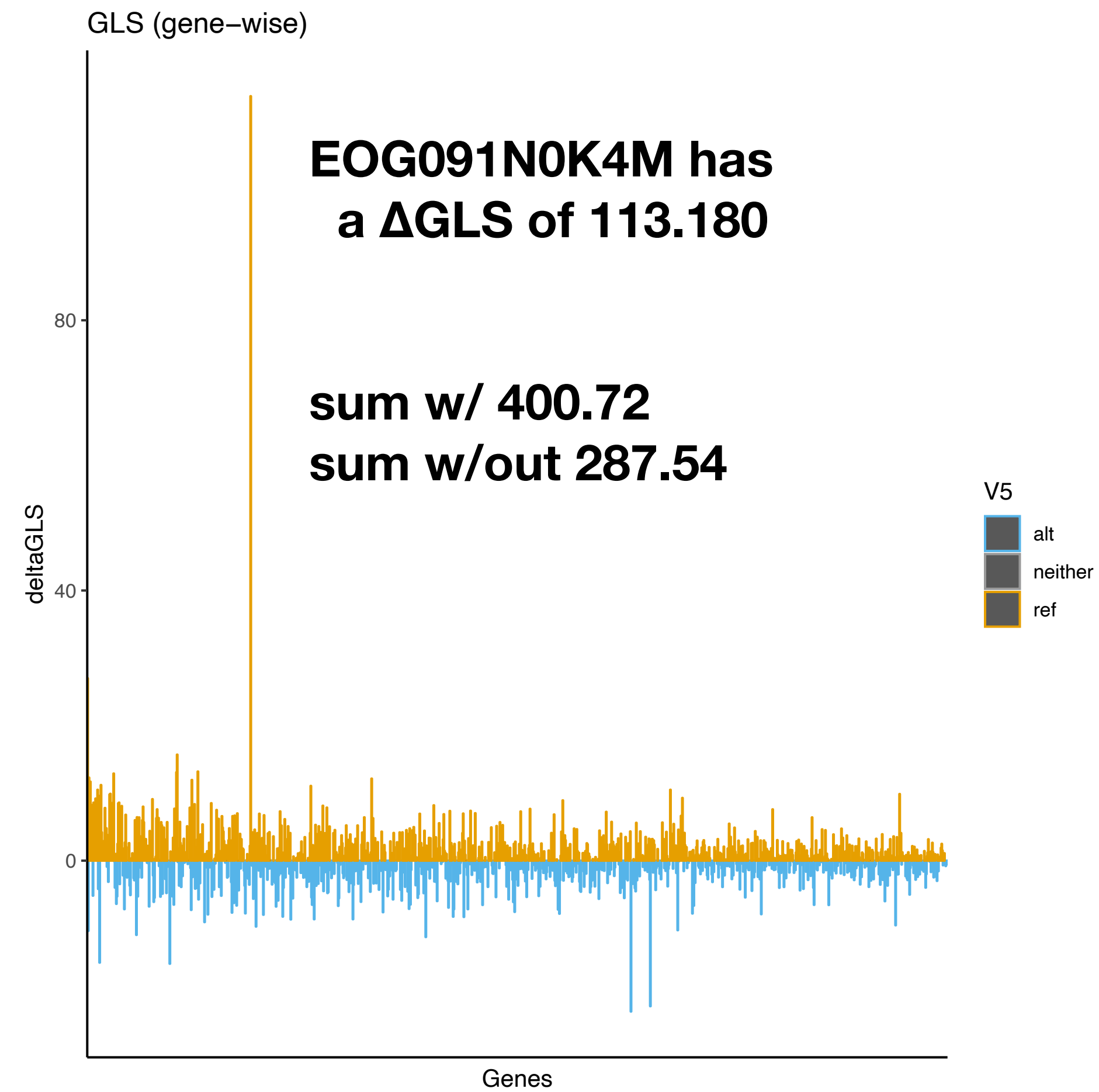
0.155/32.1

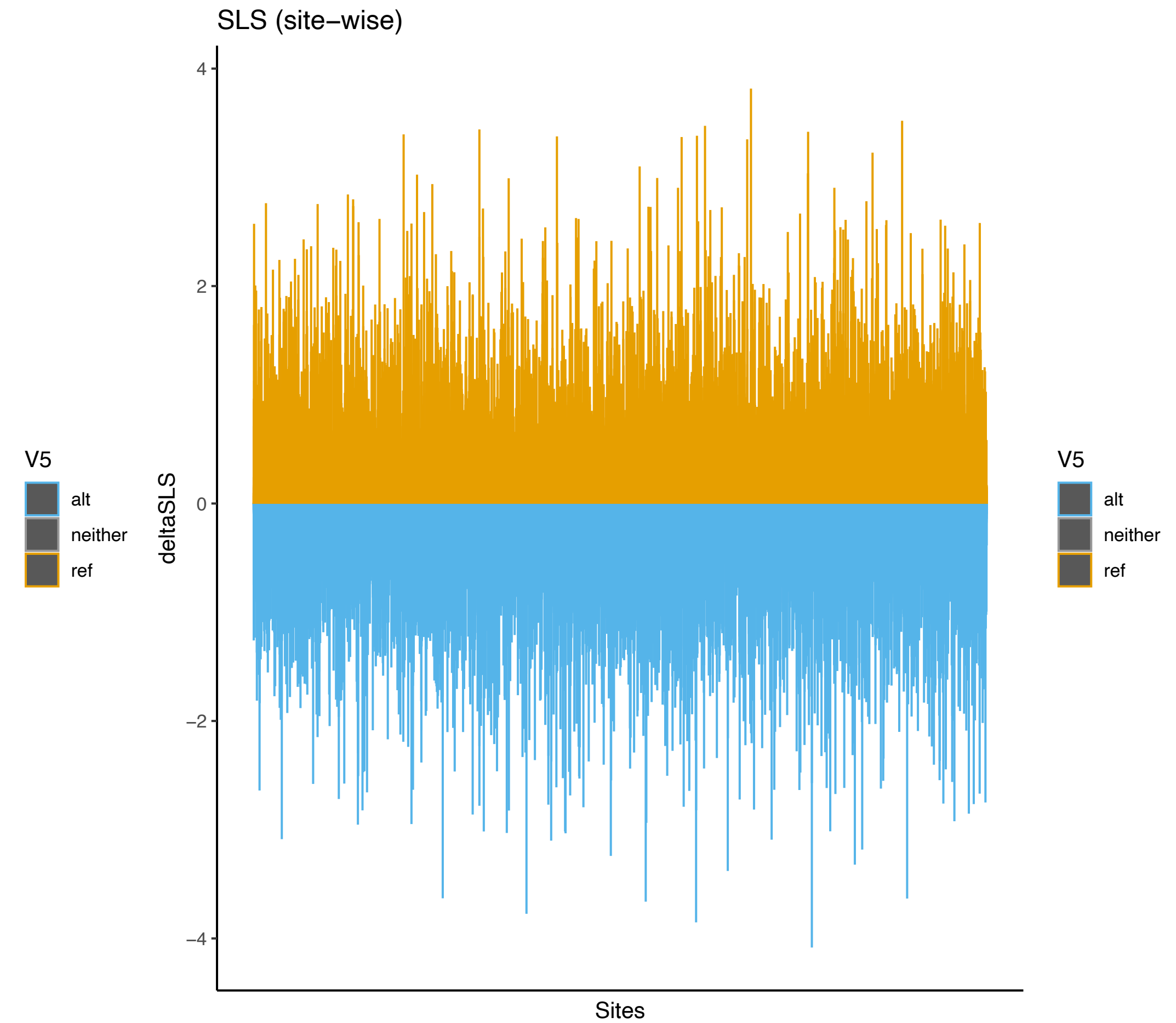
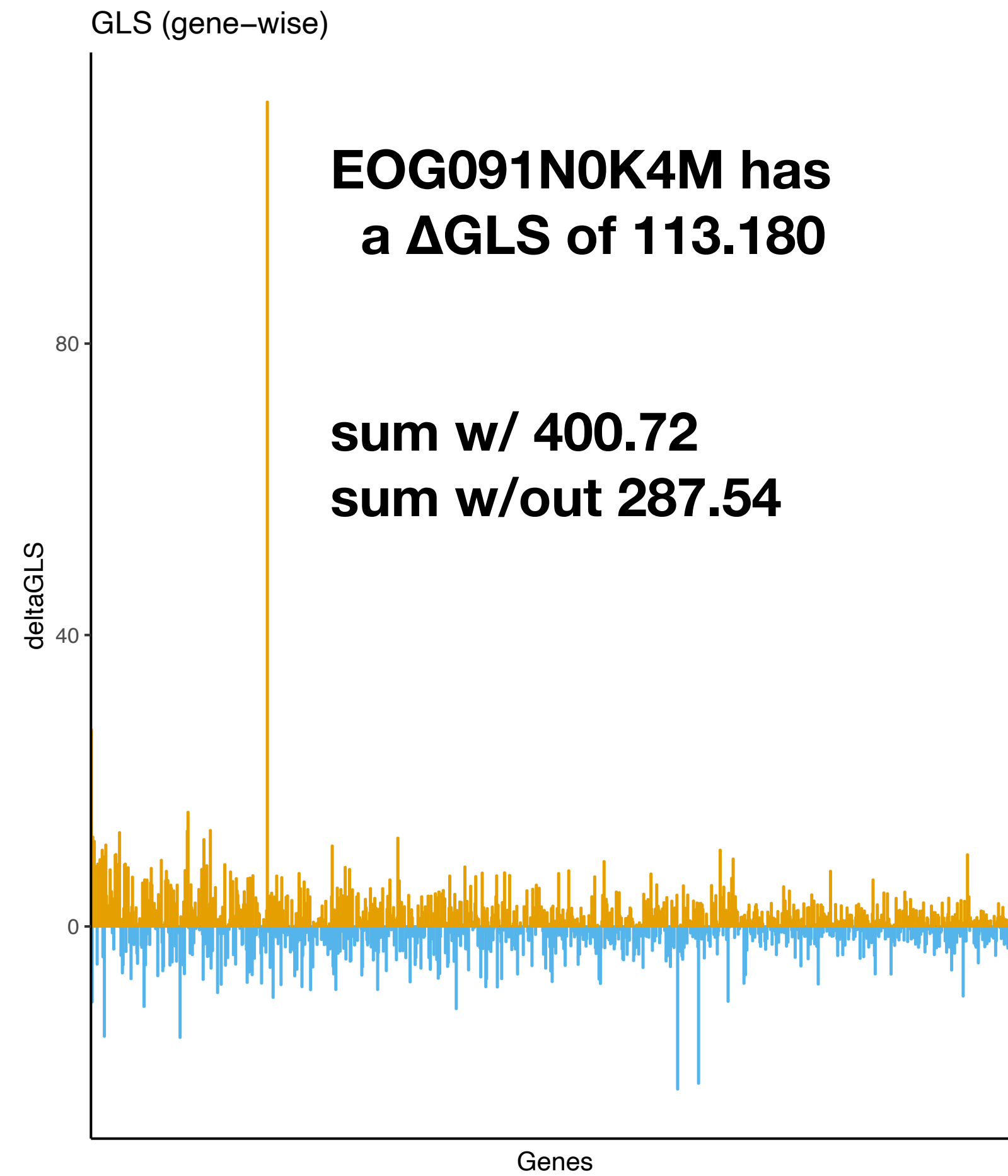
Second topology:

0/36.2





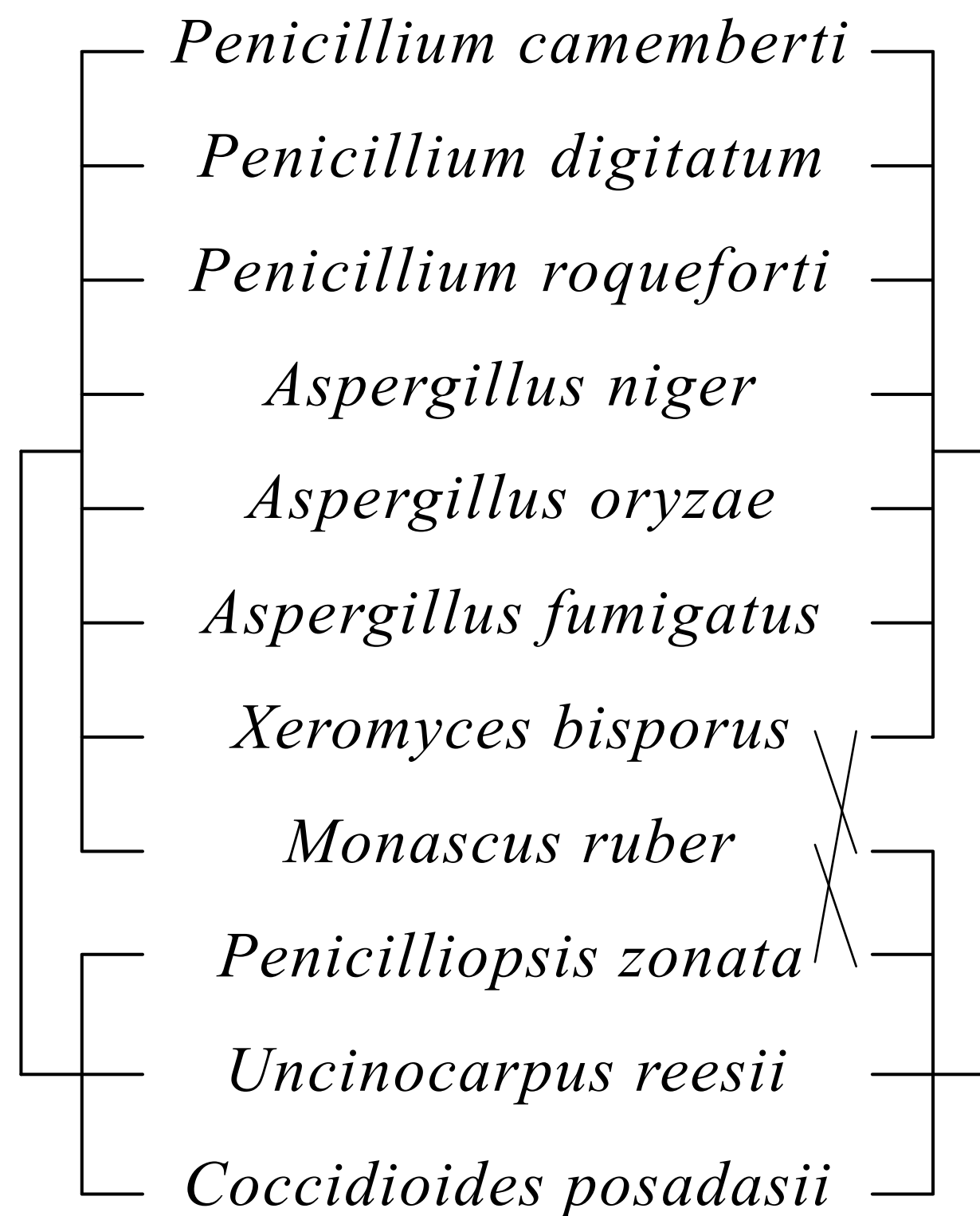




Why do gCF and GSF differ?

Consider the topologies being examined

RAxML



IQ-Tree

