Phylogenomics and Population Genomics:
Inference and Applications

Testing the robustness of your phylogeny: sensitivy analyses

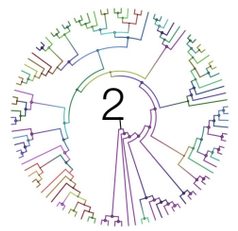
Gemma I. Martínez-Redondo & Rosa Fernández

Biodiversity Program
Institute of Evolutionary Biology (CSIC-UPF)

rosa.fernandez@ibe.upf-csic.es

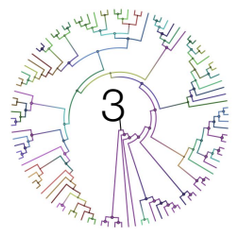
www.metazomics.com

Can I trust my results (or the results of others)?



Can I trust my results (or the results of others)?

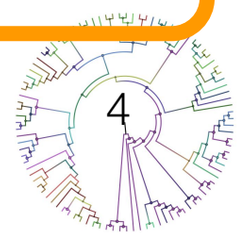
High support in an analysis *<u>does not</u>* mean that you can trust your tree!!



Can I trust my results (or the results of others)?

High support in an analysis *<u>does not</u>* mean that you can trust your tree!!

Wait, what?? And WHY is that?

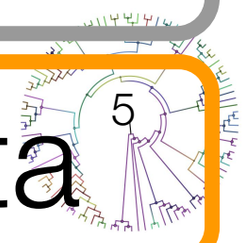


Can I trust my results (or the results of others)?

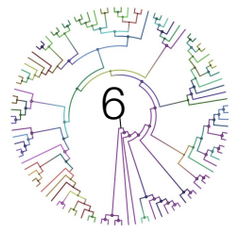
High support in an analysis *<u>does not</u>* mean that you can trust your tree!!

Wait, what?? And WHY is that?

because of the ***properties*** of the data

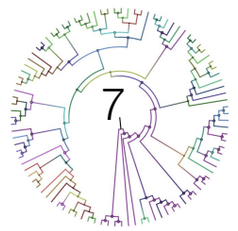


Which properties of the data are we talking about?



Which properties of the data are we talking about?

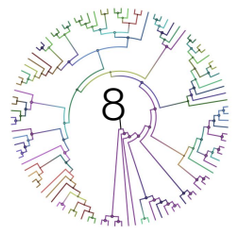
(1) *Intrinsic* properties



Which properties of the data are we talking about?

(1) *Intrinsic* properties

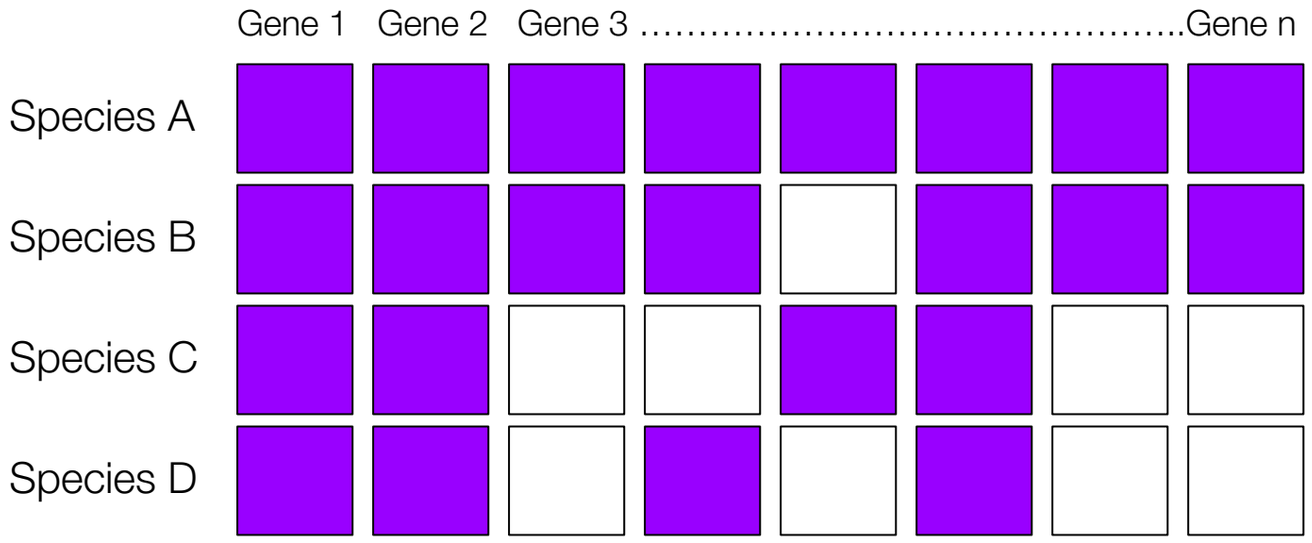
Missing data

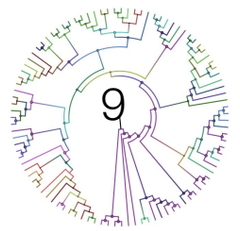


Which properties of the data are we talking about?

(1) *Intrinsic* properties

Missing data

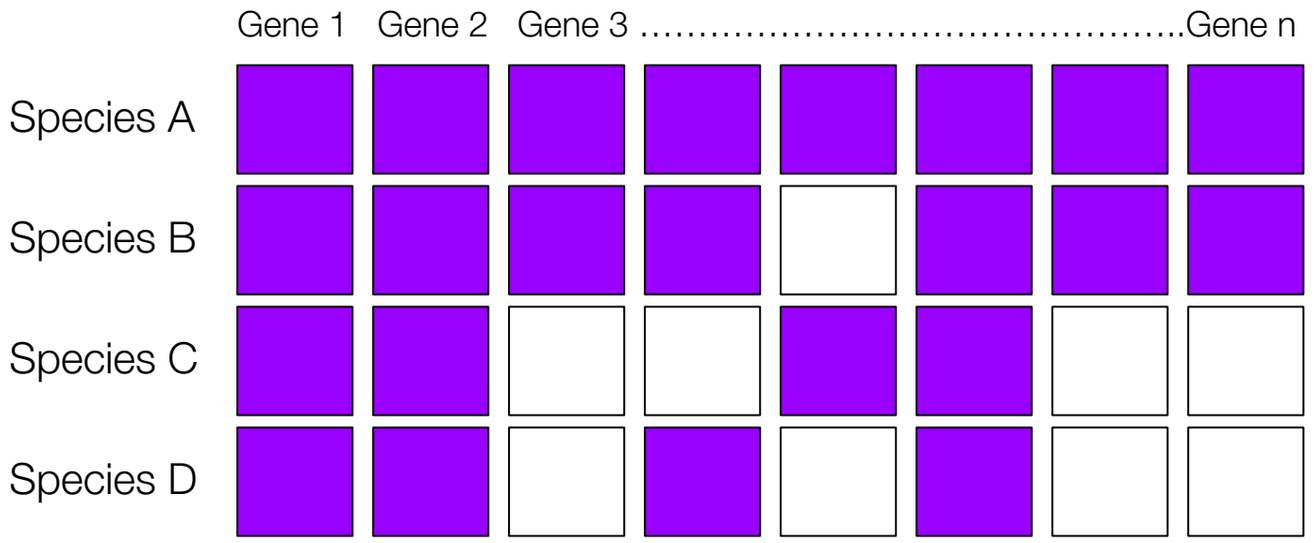




Which properties of the data are we talking about?

(1) *Intrinsic* properties

Missing data



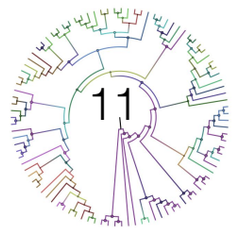
“Gruyère effect”

Which properties of the data are we talking about?

(1) *Intrinsic* properties

Missing data

Saturation

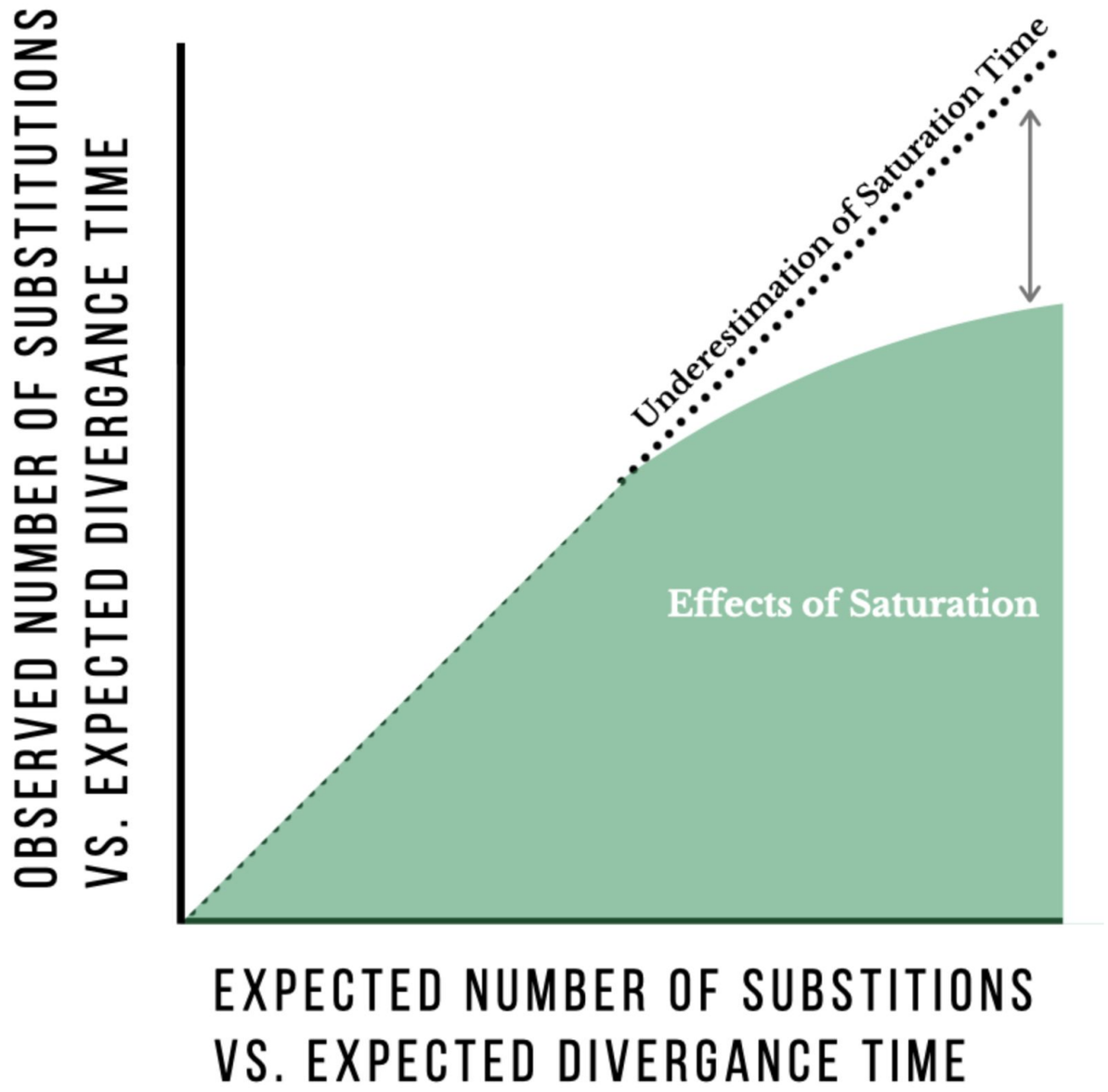


Which properties of the data are we talking about?

(1) *Intrinsic* properties

Missing data

Saturation



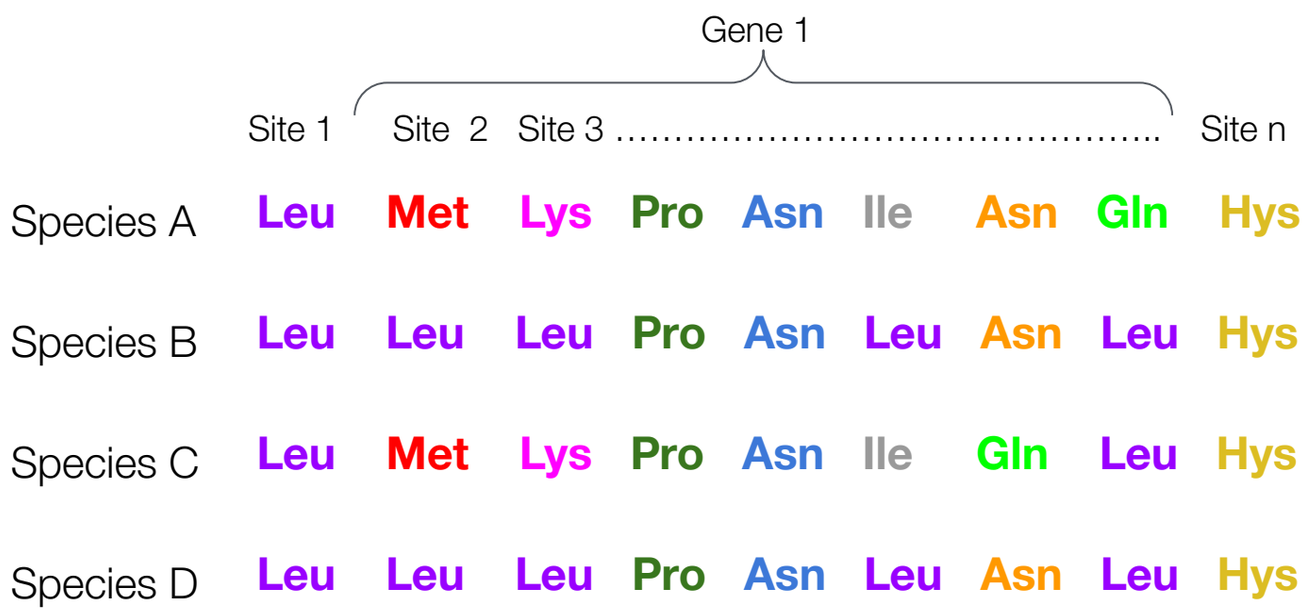
Which properties of the data are we talking about?

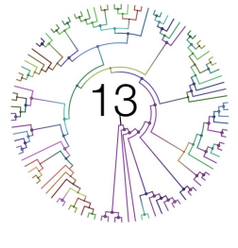
(1) *Intrinsic* properties

Missing data

Saturation

Compositional Heterogeneity





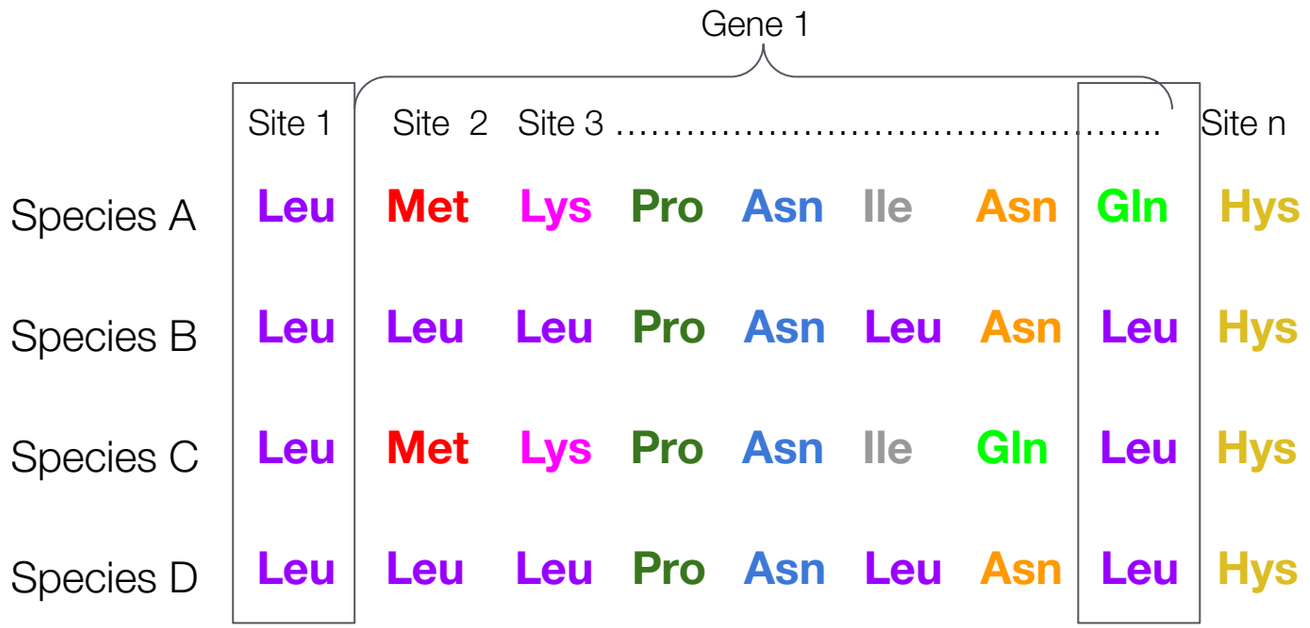
Which properties of the data are we talking about?

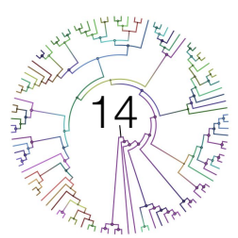
(1) *Intrinsic* properties

Missing data

Saturation

Compositional Heterogeneity





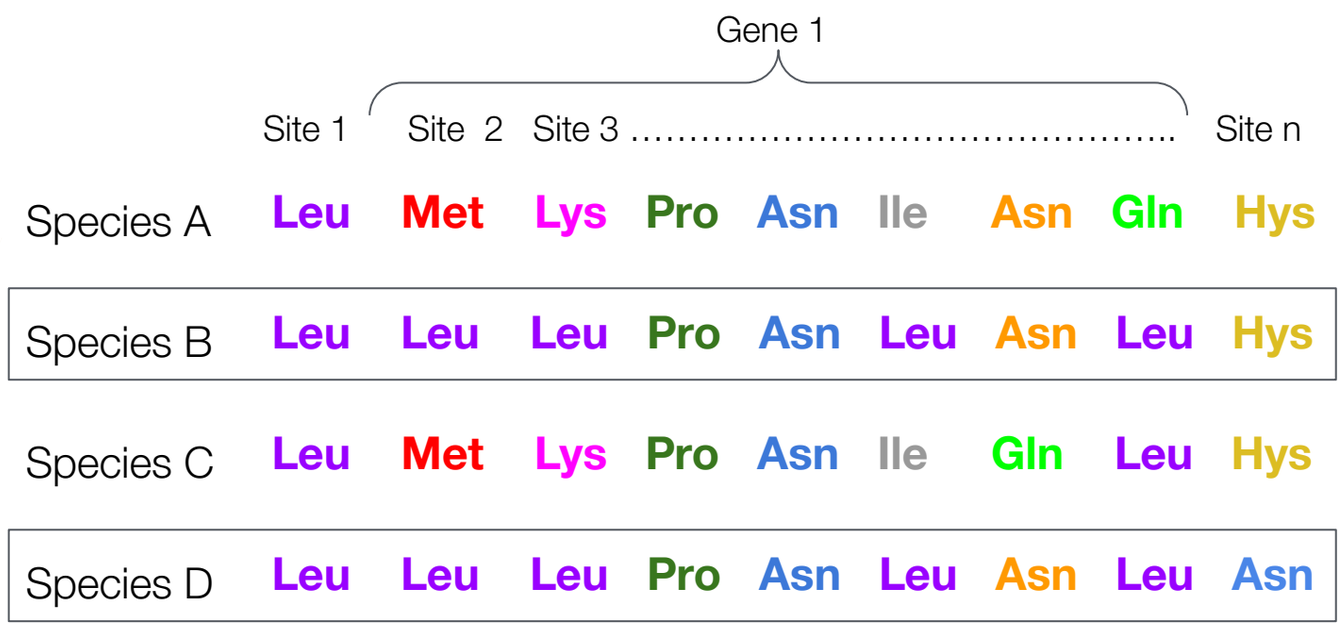
Which properties of the data are we talking about?

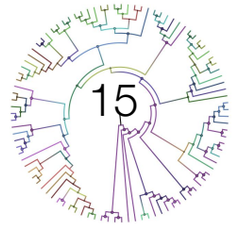
(1) *Intrinsic* properties

Missing data

Saturation

Compositional Heterogeneity





Which properties of the data are we talking about?

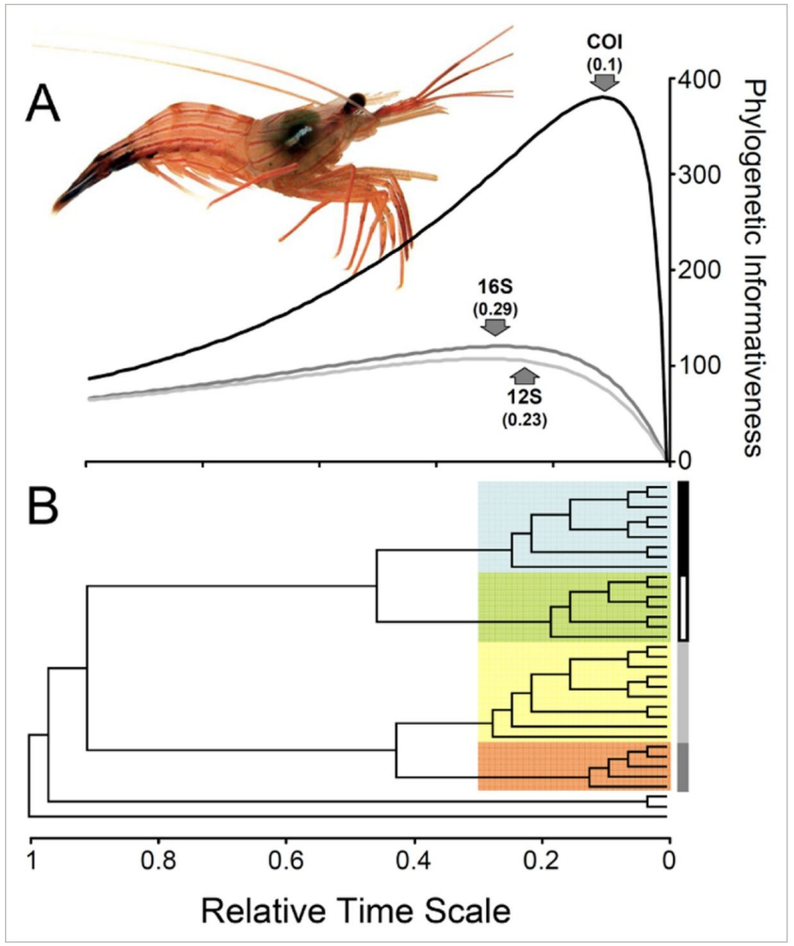
(1) *Intrinsic* properties

Missing data

Saturation

Compositional Heterogeneity

Lack of phylogenetic signal for your node of interest



Baeza & Fuentes 2013

Which properties of the data are we talking about?

(1) *Intrinsic* properties

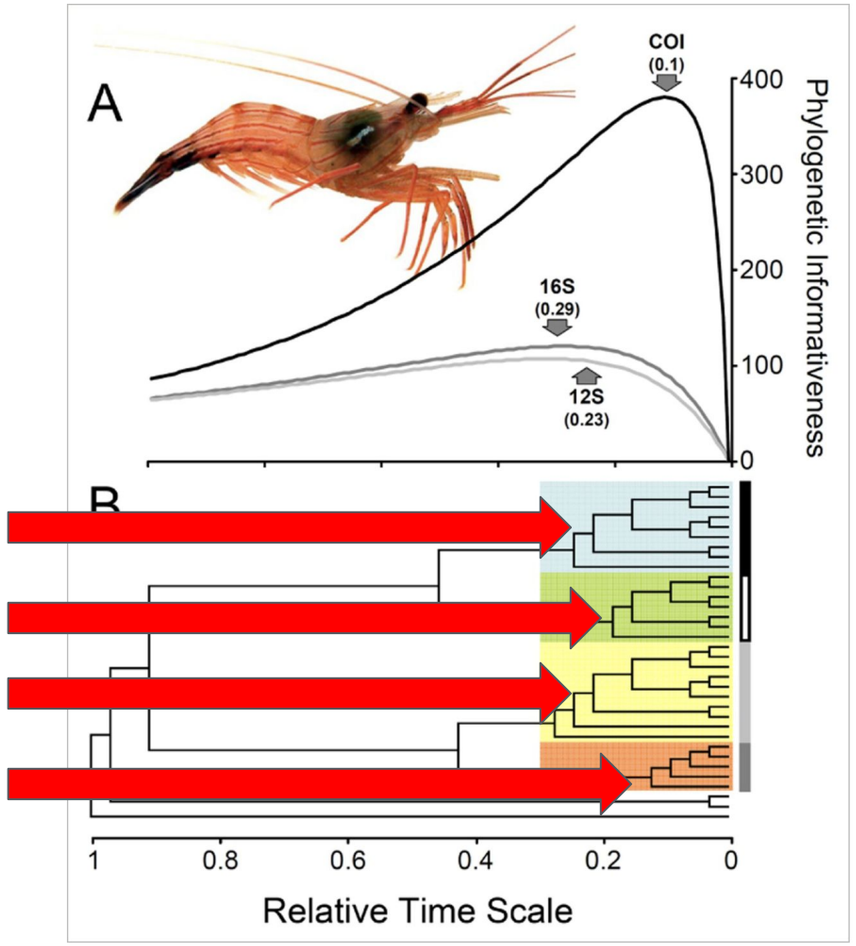
Missing data

Saturation

Compositional Heterogeneity

Lack of phylogenetic signal for your node of interest

Good information to resolve these nodes



Baeza & Fuentes 2013

Which properties of the data are we talking about?

(1) *Intrinsic* properties

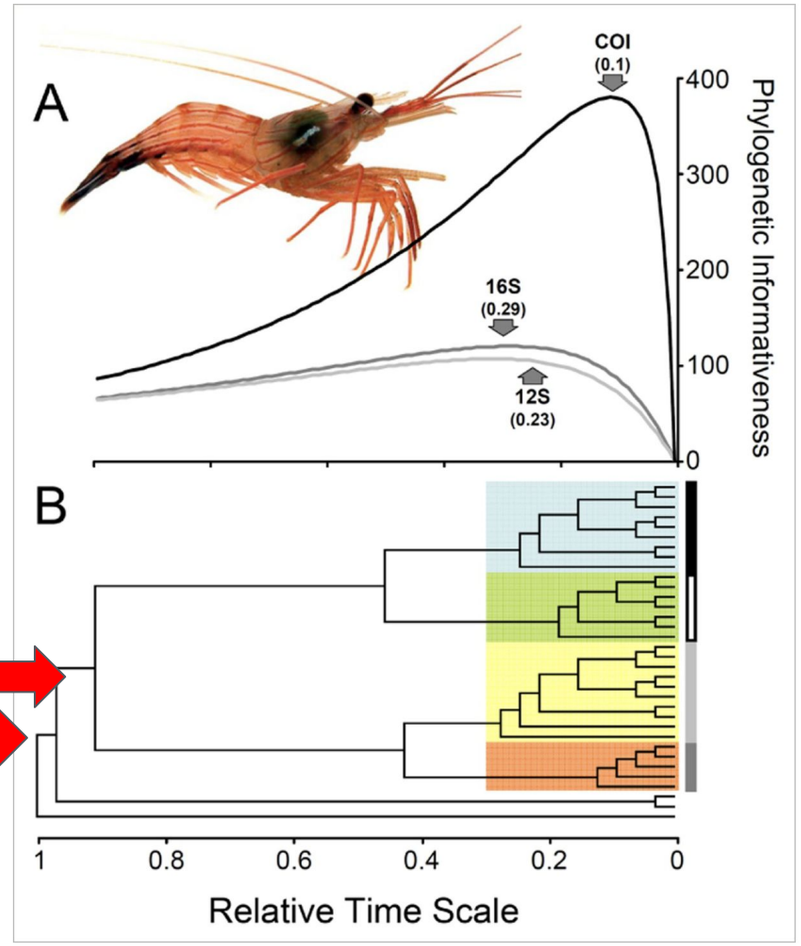
Missing data

Saturation

Compositional Heterogeneity

Lack of phylogenetic signal for your node of interest

<u>Not</u> enough information to resolve these nodes



Baeza & Fuentes 2013

Which properties of the data are we talking about?

(1) *Intrinsic* properties

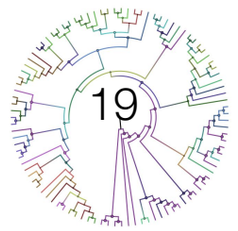
Missing data

Saturation

Compositional Heterogeneity

Lack of phylogenetic signal for your node of interest

etc.



Which properties of the data are we talking about?

(1) *Intrinsic* properties

Missing data

Saturation

Compositional Heterogeneity

Lack of phylogenetic signal for your node of interest

etc.

(2) Conflict between individual gene trees and the ‘real’ species tree

Which properties of the data are we talking about?

(1) *Intrinsic* properties

Missing data

Saturation

Compositional Heterogeneity

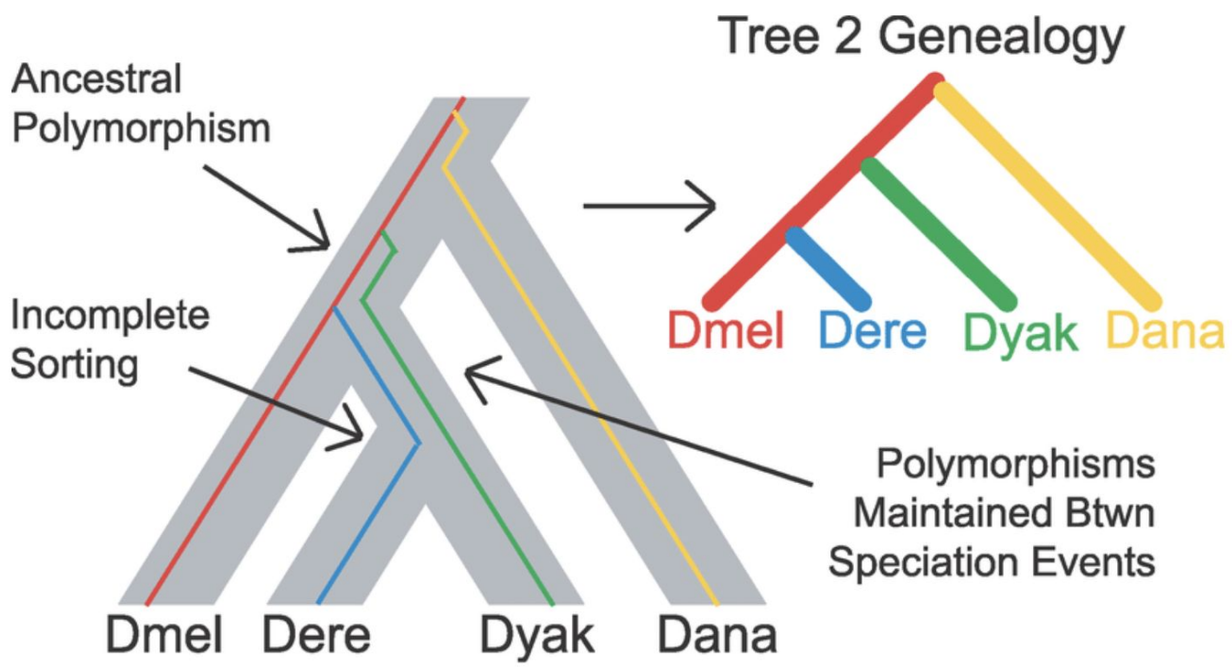
Lack of phylogenetic signal for your node of interest

etc.

(2) Conflict between individual gene trees and the ‘real’ species tree

Incomplete lineage sorting

Which properties of the data are we talking about?



Incomplete Lineage Sorting The history of a gene (colored lines) is drawn in the context of a species tree (gray bars). New lineages arising from new polymorphisms in the gene are drawn in different colors. In this case, the two alleles in the population prior to the split of Dmel are maintained through to the split of Dere and Dyak, leading to incomplete lineage sorting and an incongruent genealogy (tree 2). The greater the diversity in the ancestral population and the shorter the time between speciation events, the more likely nonspecies genealogies are.

Pollard et al. 2006

(2) Conflict between individual gene trees and the ‘real’ species tree

Incomplete lineage sorting

Which properties of the data are we talking about?

(1) *Intrinsic* properties

Missing data

Saturation

Compositional Heterogeneity

Lack of phylogenetic signal for your node of interest

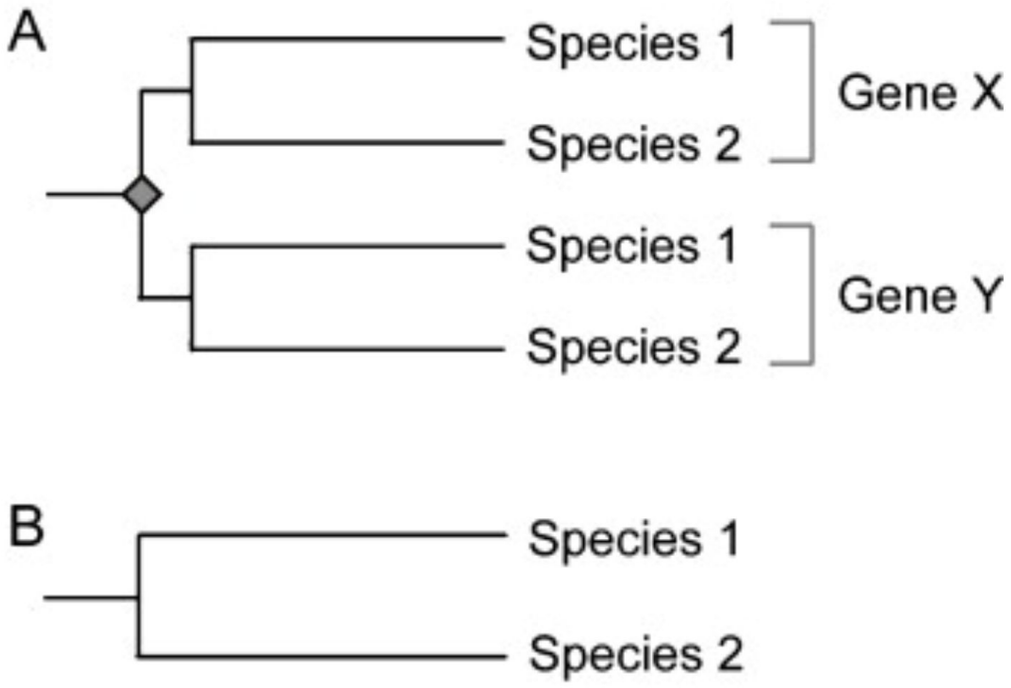
etc.

(2) Conflict between individual gene trees and the ‘real’ species tree

Incomplete lineage sorting

Gene loss (eg, hidden paralogy)

Which properties of the data are we talking about?



Schematic illustration of hidden paralogy. (A) Hypothetical situation in which two species (Species 1 and 2) have the same set of genes (Gene X and Y) that were duplicated before the speciation between the two species. (B) Phylogenetic tree without any obvious gene duplication. If only one gene is sampled from each species without exhaustive sampling, they might not be orthologous to each other. (C) Possible explanation of the tree topology in B. Misidentification or loss of Gene Y of Species 1 and Gene X of Species 2 occurred, and thus the situation B represents paralogy between Gene X of Species 1 and Gene Y of Species 2.

Kuraku 2013

(2) Conflict between individual gene trees and the ‘real’ species tree

Incomplete lineage sorting

Gene loss (eg, hidden paralogy)

Which properties of the data are we talking about?

(1) *Intrinsic* properties

Missing data

Saturation

Compositional Heterogeneity

Lack of phylogenetic signal for your node of interest

etc.

(2) Conflict between individual gene trees and the ‘real’ species tree

Incomplete lineage sorting

Gene loss (eg, hidden paralogy)

Hybridization

Which properties of the data are we talking about?

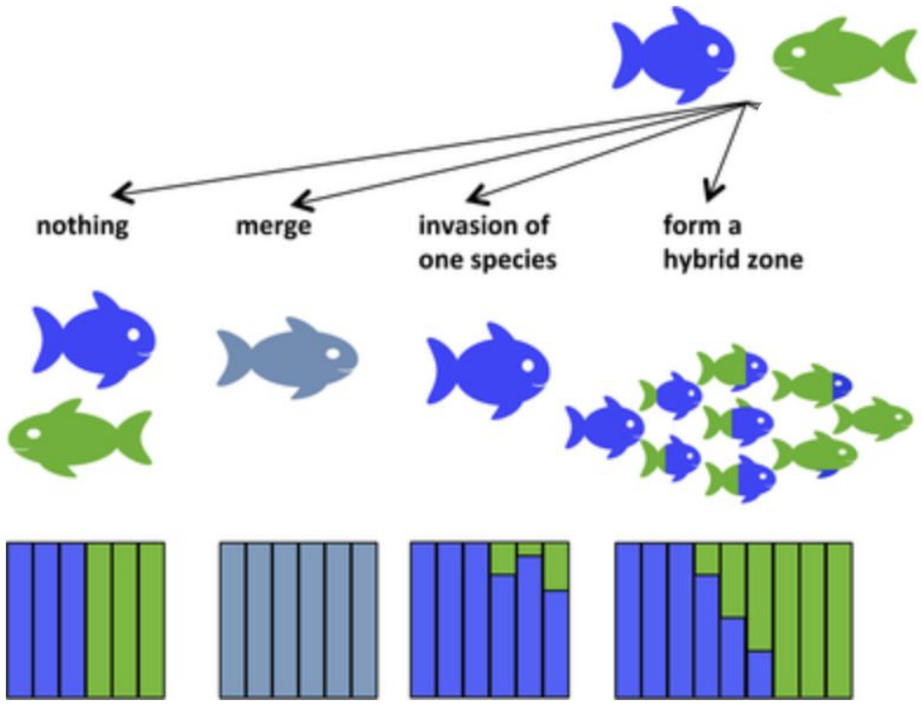


Fig 2.

Schematic representation of homoploid and allopolyploid hybrid speciation.

Runemark et al. 2019

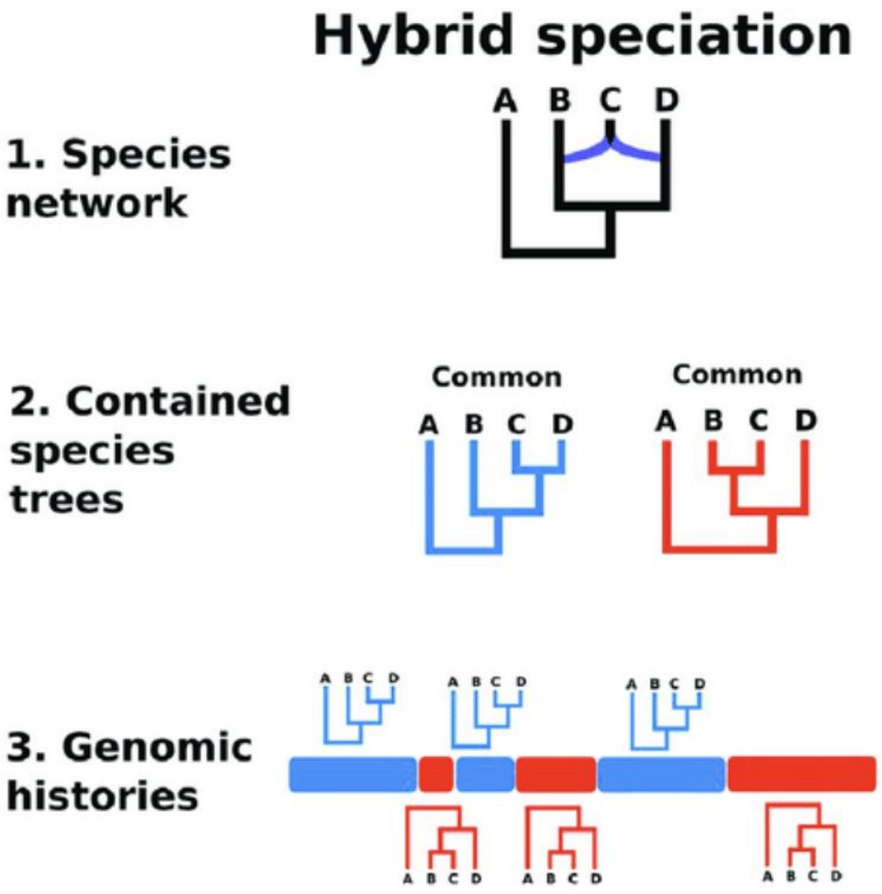
(2) Conflict between individual gene trees and the ‘real’ species tree

Incomplete lineage sorting

Gene loss (eg, hidden paralogy)

Hybridization

Which properties of the data are we talking about?



Folk et al. 2018

(2) Conflict between individual gene trees and the ‘real’ species tree

Incomplete lineage sorting

Gene loss (eg, hidden paralogy)

Hybridization

Which properties of the data are we talking about?

(1) *Intrinsic* properties

Missing data

Saturation

Compositional Heterogeneity

Lack of phylogenetic signal for your node of interest

etc.

(2) Conflict between individual gene trees and the ‘real’ species tree

Incomplete lineage sorting

Gene loss (eg, hidden paralogy)

Hybridization

Introgression

Which properties of the data are we talking about?

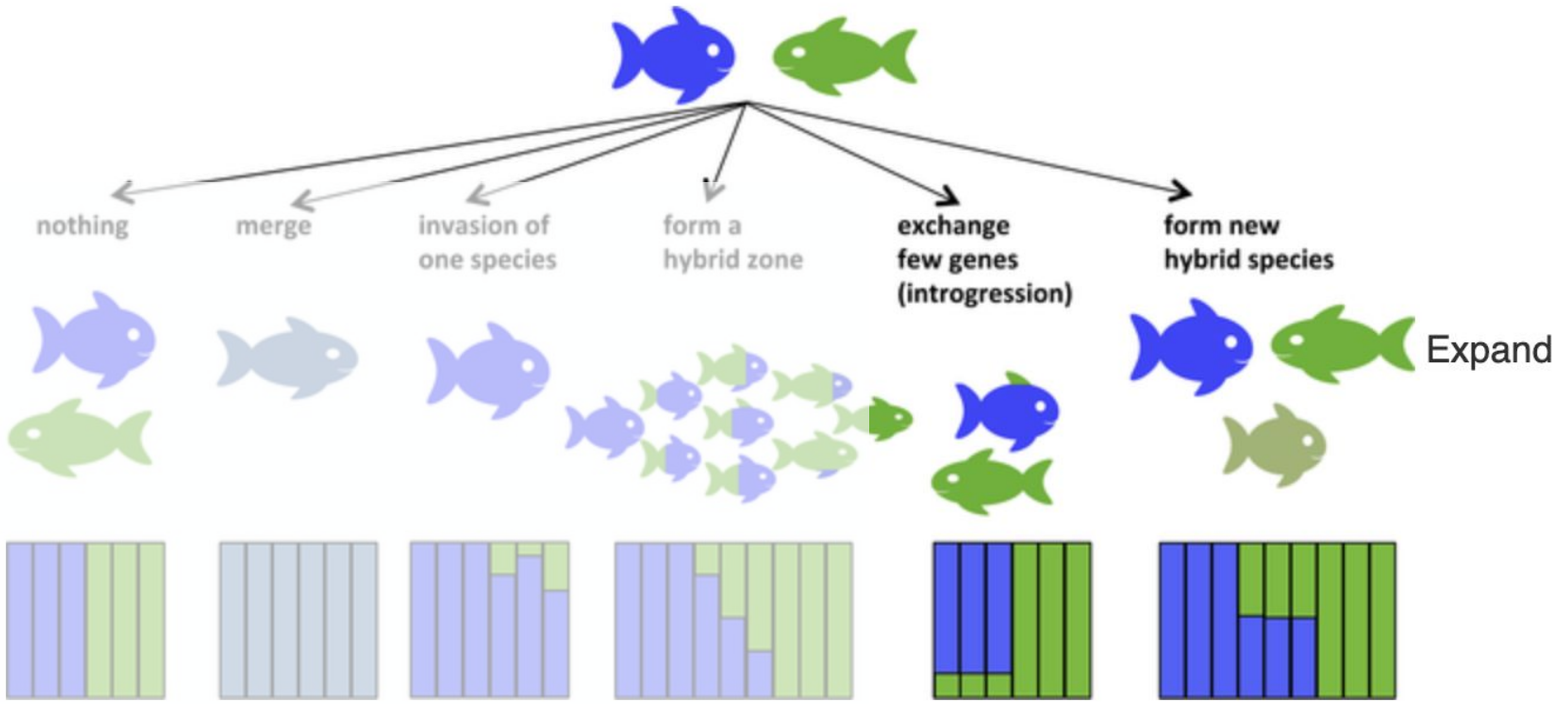


Fig 2.

Schematic representation of homoploid and allopolyploid hybrid speciation.

Runemark et al. 2019

(2) Conflict between individual gene trees and the ‘real’ species tree

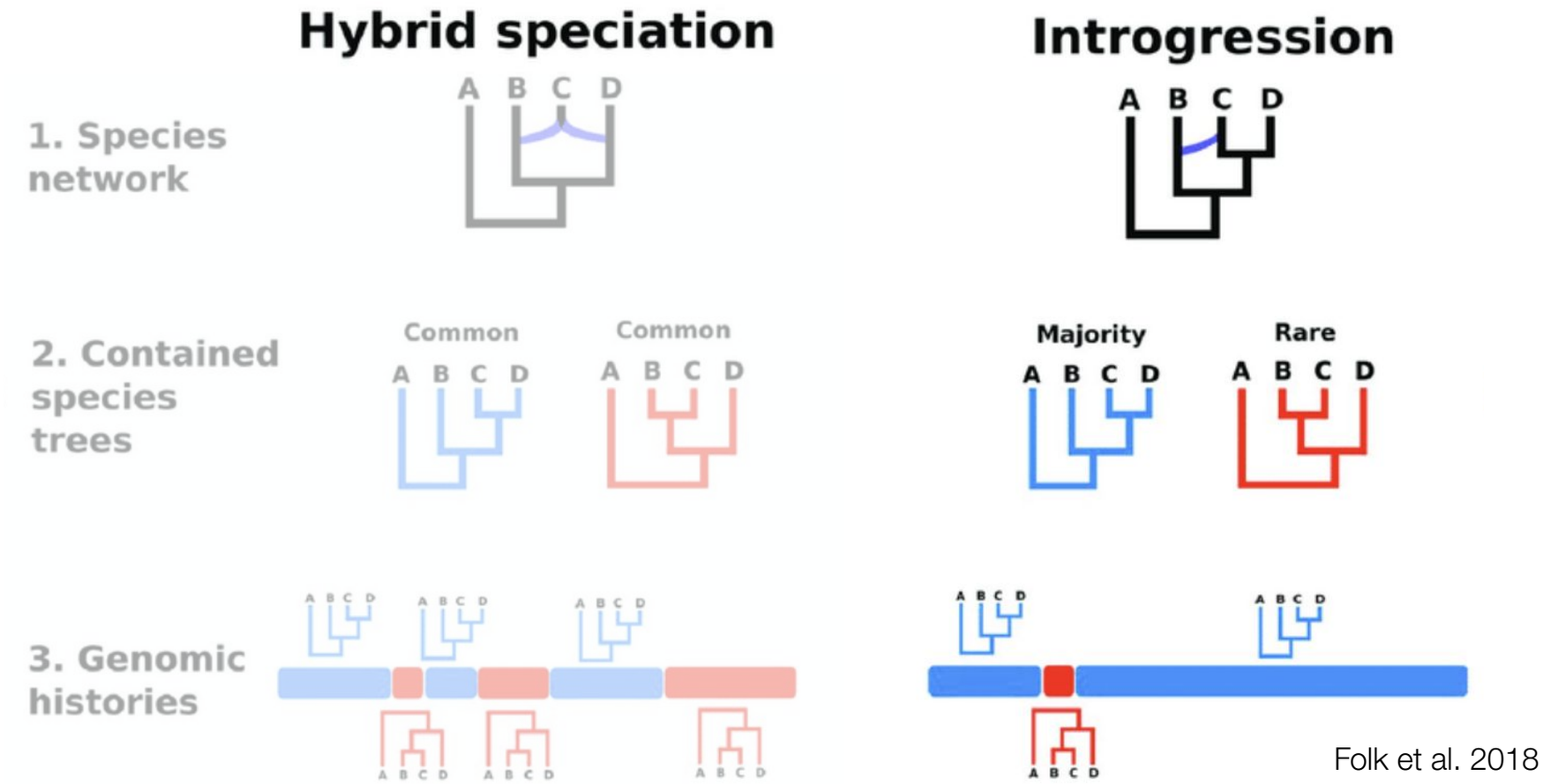
Incomplete lineage sorting

Gene loss (eg, hidden paralogy)

Hybridization

Introgression

Which properties of the data are we talking about?



Folk et al. 2018

(2) Conflict between individual gene trees and the ‘real’ species tree

Incomplete lineage sorting

Gene loss (eg, hidden paralogy)

Hybridization

Introgression

Which properties of the data are we talking about?

(1) *Intrinsic* properties

Missing data

Saturation

Compositional Heterogeneity

Lack of phylogenetic signal for your node of interest

etc.

(2) Conflict between individual gene trees and the ‘real’ species tree

Incomplete lineage sorting

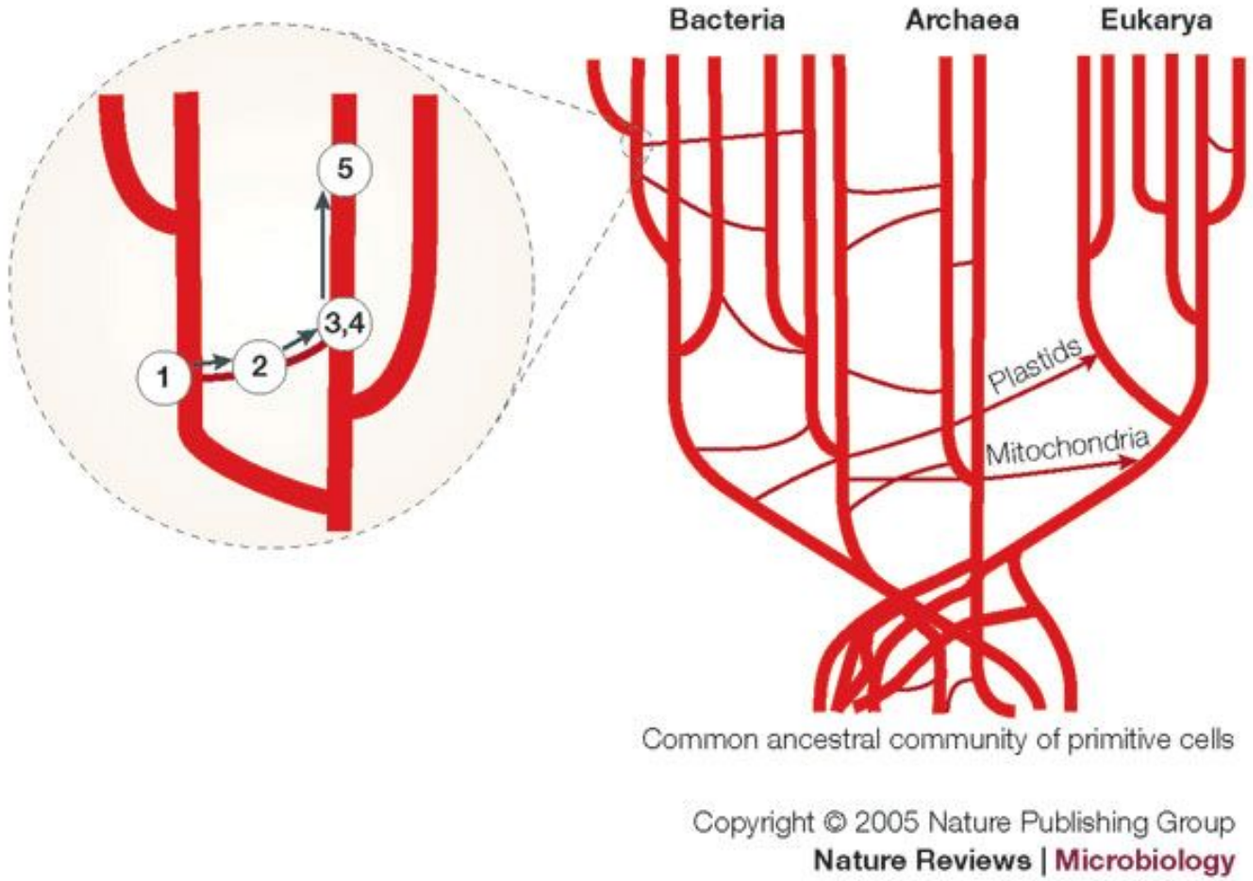
Gene loss (eg, hidden paralogy)

Hybridization

Introgression

Horizontal gene transfer

Which properties of the data are we talking about?



Smets and Barkay 2005

(2) Conflict between individual gene trees and the ‘real’ species tree

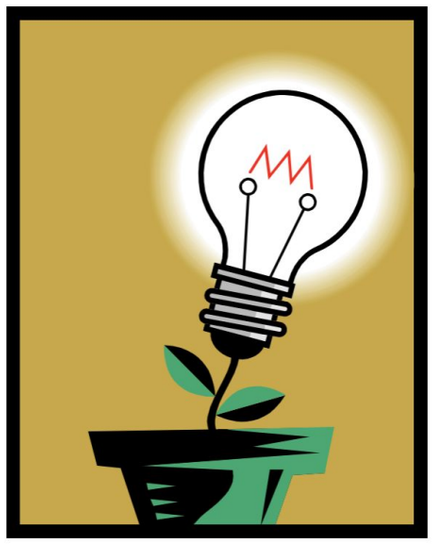
Incomplete lineage sorting

Gene loss (eg, hidden paralogy)

Hybridization

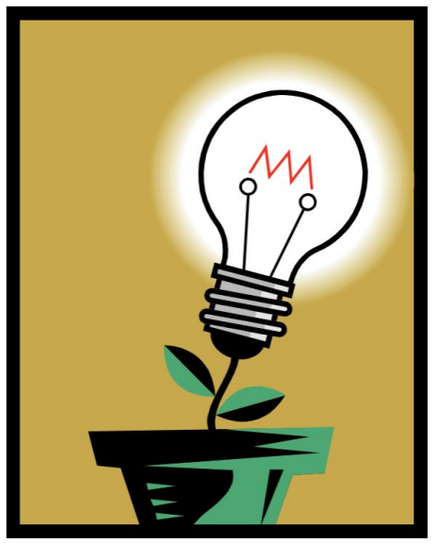
Introgression

Horizontal gene transfer



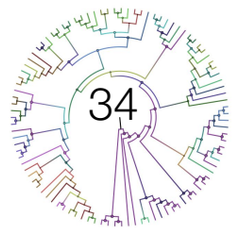
Welcome!

Critical Thinking Corner

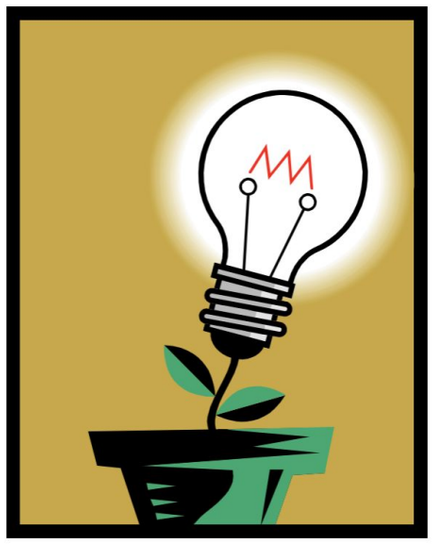


Welcome!

Which of these properties have a *biological origin* and which ones a *methodological* one?



Critical Thinking Corner



Welcome!

Which of these properties have a *biological origin* and which ones a *methodological* one?

Missing data

Incomplete lineage sorting

Saturation

Gene loss (eg, hidden paralogy)

Compositional Heterogeneity

Hybridization

Introgression

Lack of phylogenetic signal for your node of interest

Horizontal gene transfer

Why may these properties result in a highly supported ‘wrong’ tree?

Why may these properties result in a highly supported ‘wrong’ tree?

Because of:

Why may these properties result in a highly supported ‘wrong’ tree?

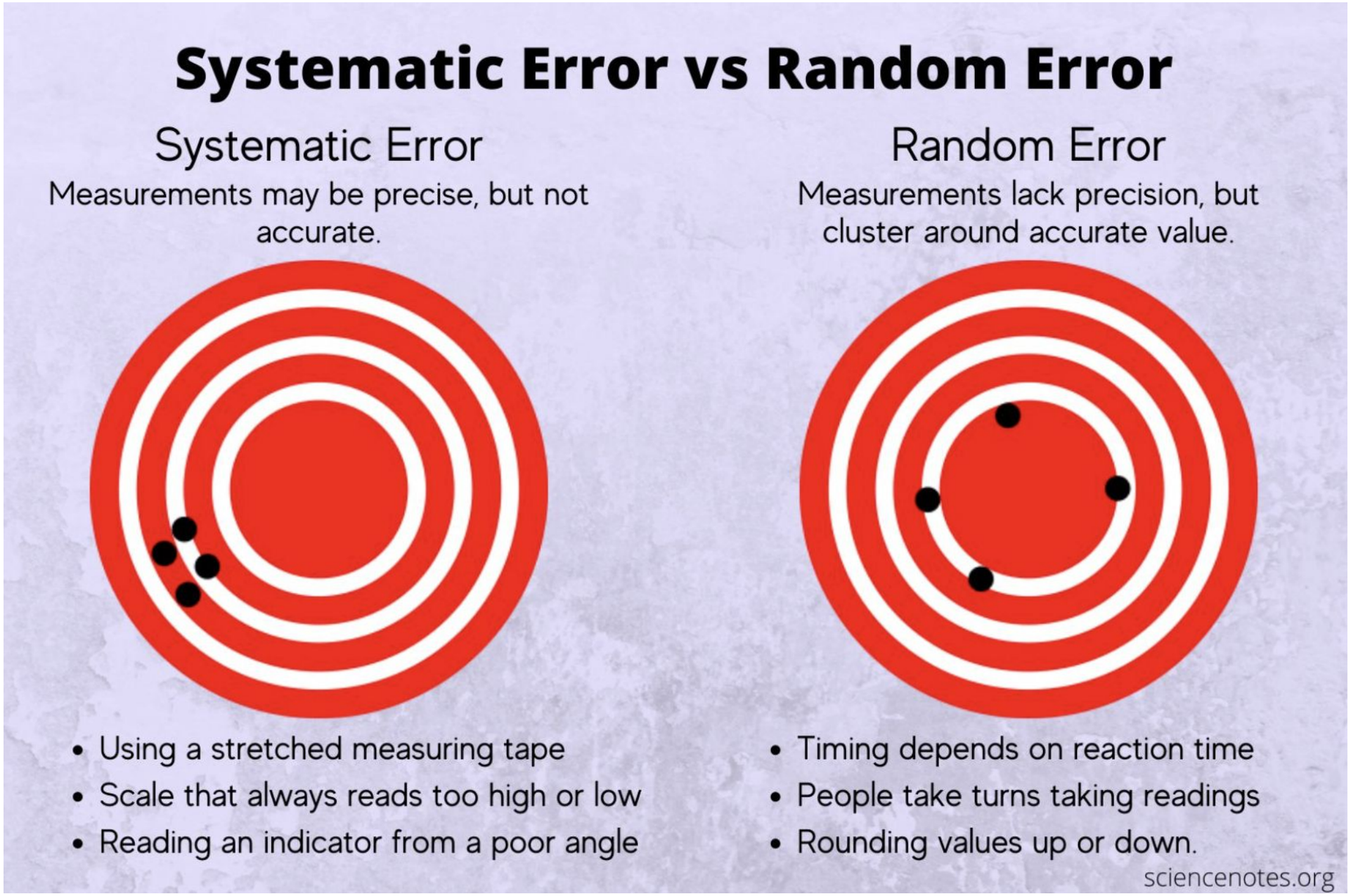
Because of:

1) Systematic error

Why may these properties result in a highly supported ‘wrong’ tree?

Because of:

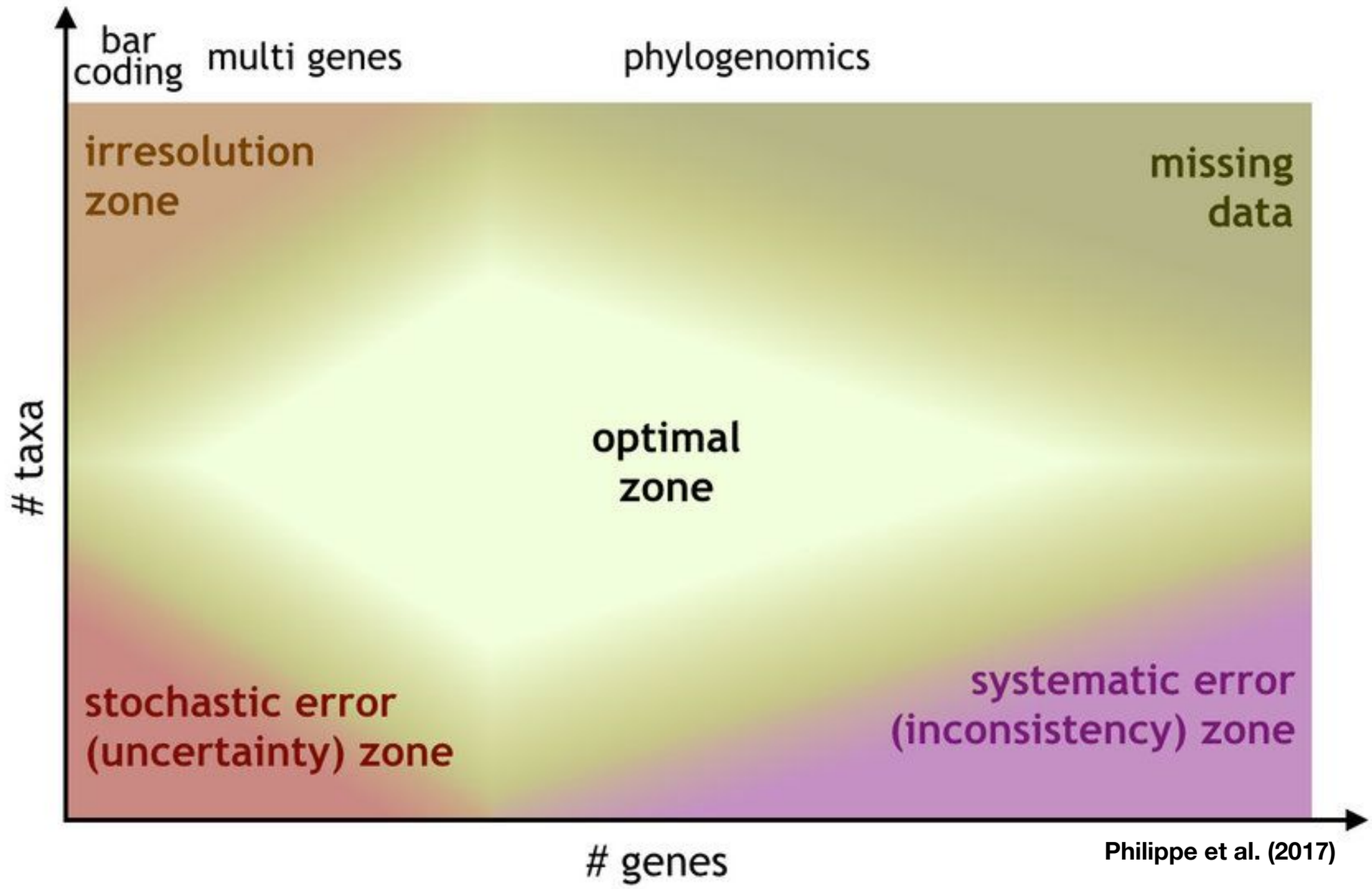
1) Systematic error



Why may these properties result in a highly supported ‘wrong’ tree?

Because of:

1) Systematic error

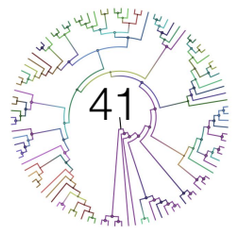


Philippe et al. (2017)

Why may these properties result in a highly supported ‘wrong’ tree?

Because of:

1) Systematic error
2) **Model violation**

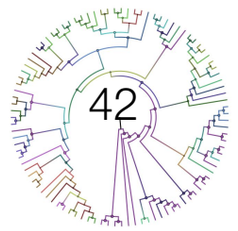


Why may these properties result in a highly supported ‘wrong’ tree?

Because of:

1) Systematic error
2) **Model violation**

Eg 1, compositional homogeneity in the gene sequence to correctly infer/apply a substitution model



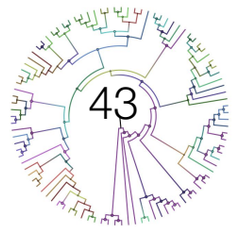
Why may these properties result in a highly supported ‘wrong’ tree?

Because of:

1) Systematic error
2) **Model violation**

Eg 1, compositional homogeneity in the gene sequence to correctly infer/apply a substitution model

Eg 2, no recombination



Why may these properties result in a highly supported ‘wrong’ tree?

Because of:

1) Systematic error
2) **Model violation**

Eg 1, compositional homogeneity in the gene sequence to correctly infer/apply a substitution model

Eg 2, no recombination

Eg 3, genes evolved through speciation and not through duplication

Why may these properties result in a highly supported ‘wrong’ tree?

Because of:

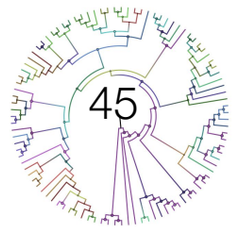
1) Systematic error
2) **Model violation**

Eg 1, compositional homogeneity in the gene sequence to correctly infer/apply a substitution model

Eg 2, no recombination

Eg 3, genes evolved through speciation and not through duplication

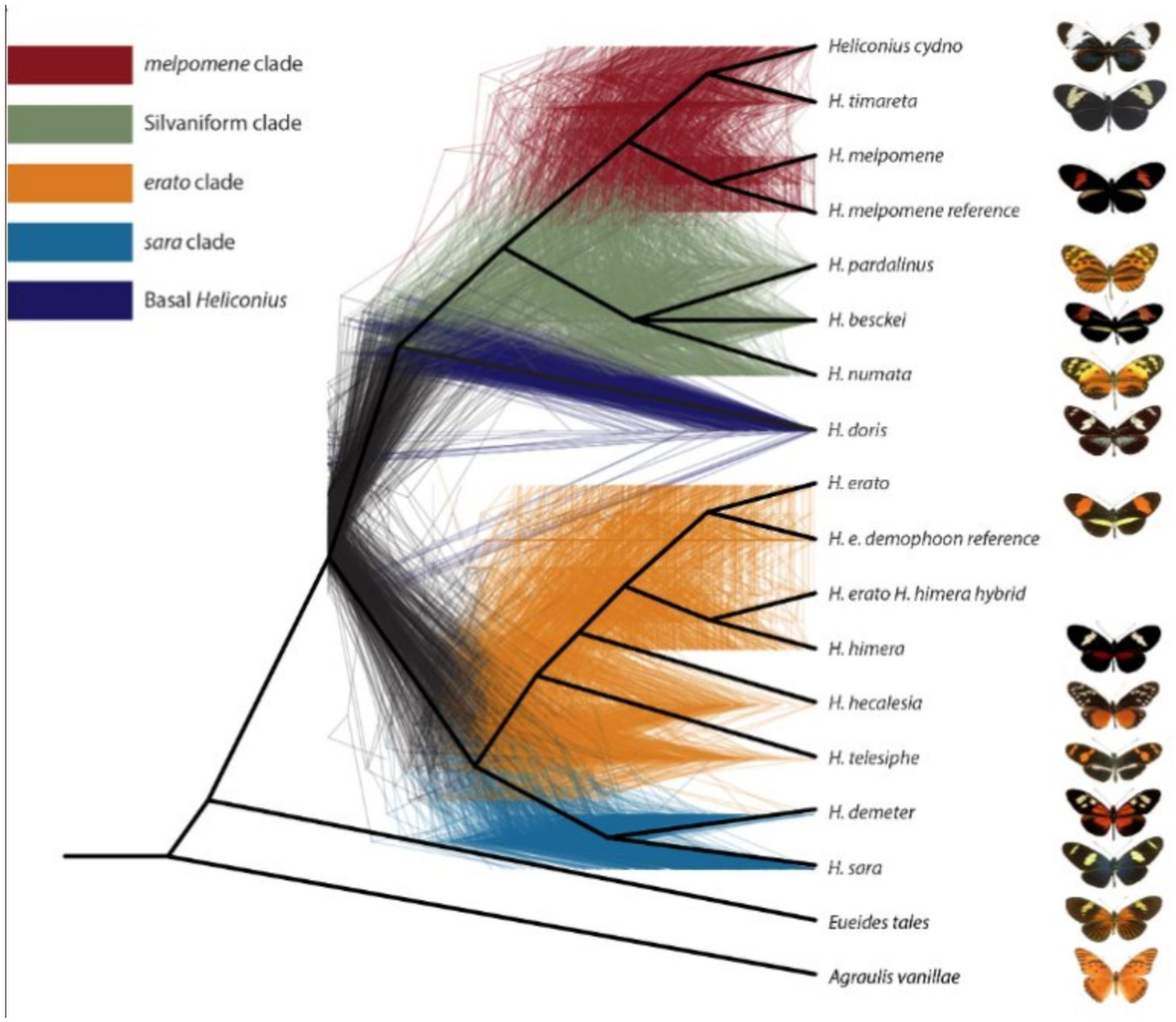
etc.



Why may these properties result in a highly supported ‘wrong’ tree?

Because of:

1) Systematic error
2) Model violation
3) **Gene tree/species tree discordance**



Edelman et al. 2019

So… what do we do to test the robustness of our tree?

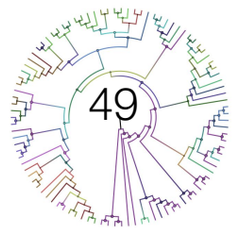
So… what do we do to test the robustness of our tree?

1) **Build different subsets of your data through a subsampling strategy selecting genes with different properties**

So… what do we do to test the robustness of our tree?

1) Build different subsets of your data through a subsampling strategy selecting genes with different properties

2) **Run different analyses that rely on different assumptions or apply different models**

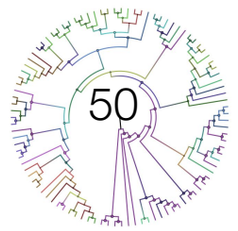


So… what do we do to test the robustness of our tree?

1) Build different subsets of your data through a subsampling strategy selecting genes with different properties

2) Run different analyses that rely on different assumptions or apply different models

3) Do 1) and 2) both at the level of *supermatrix* and *subset of individual gene trees*

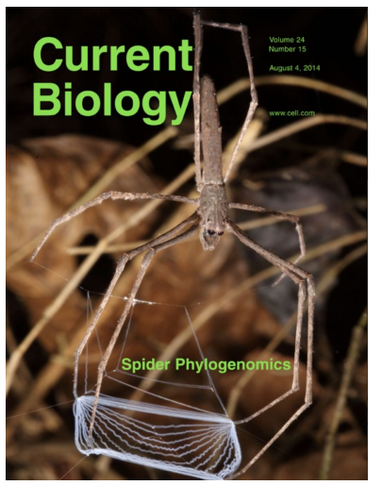


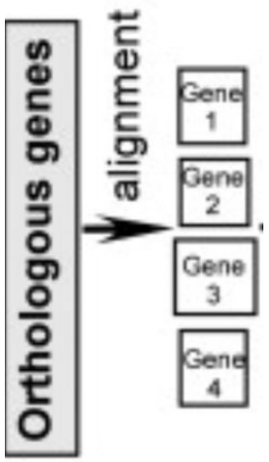
So… what do we do to test the robustness of our tree?

1) Build different subsets of your data through a subsampling strategy selecting genes with different properties

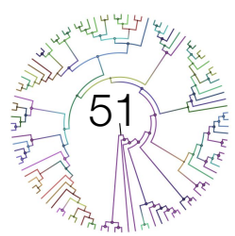
2) Run different analyses that rely on different assumptions or apply different models

3) Do 1) and 2) both at the level of *supermatrix* and *subset of individual gene trees*





Fernández, Hormiga and Giribet 2014

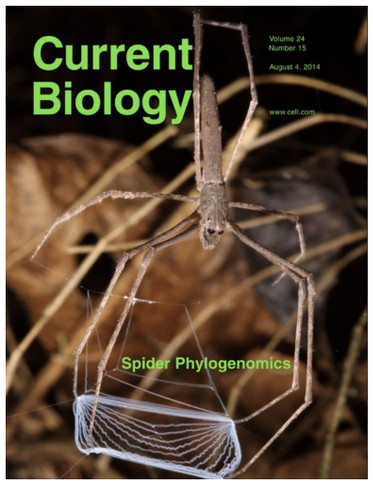


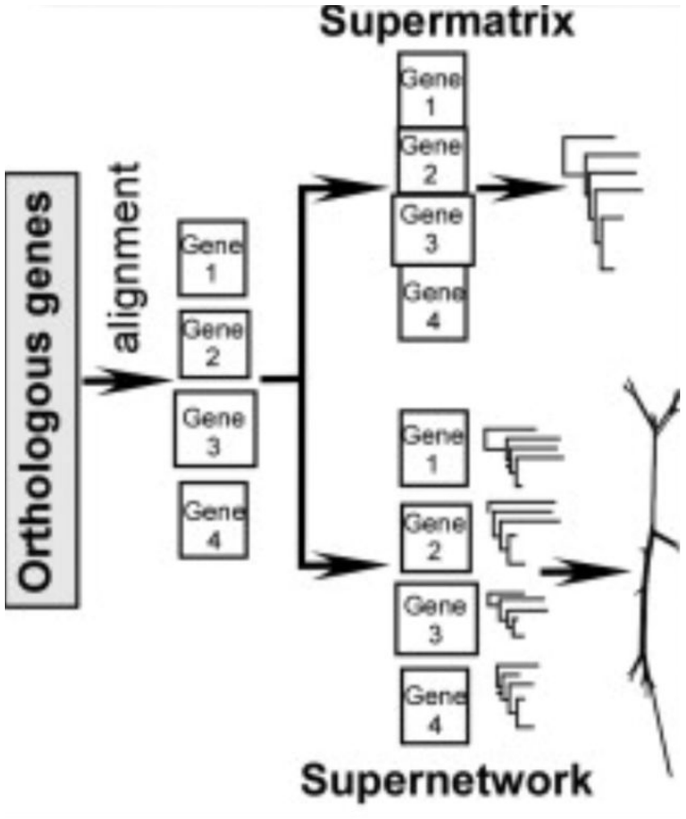
So… what do we do to test the robustness of our tree?

1) Build different subsets of your data through a subsampling strategy selecting genes with different properties

2) Run different analyses that rely on different assumptions or apply different models

3) Do 1) and 2) both at the level of *supermatrix* and *subset of individual gene trees*





(subset of indiv. gene trees)

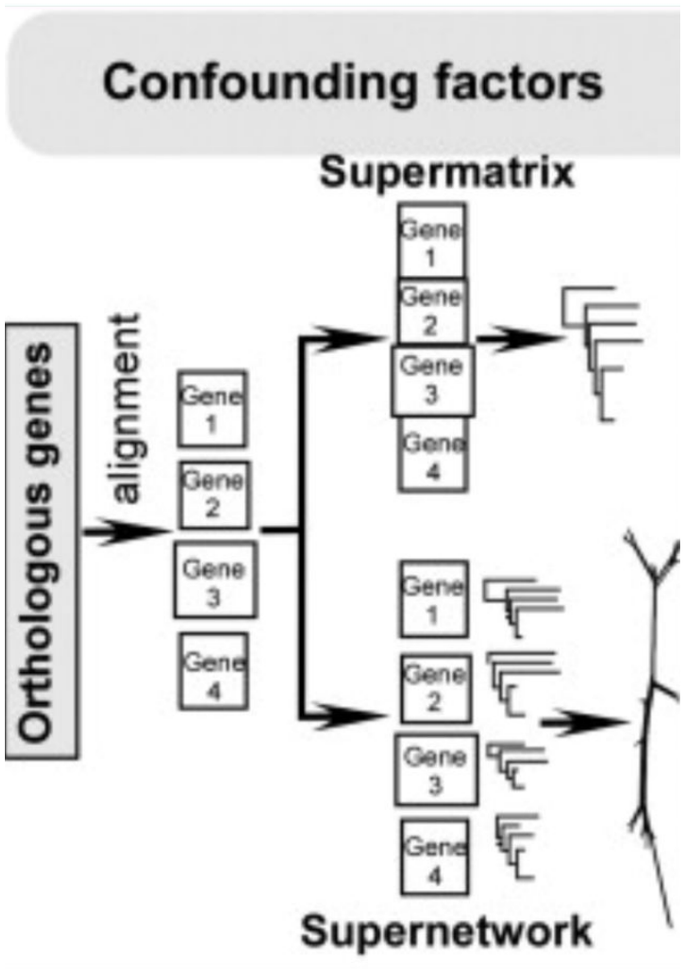
Fernández, Hormiga and Giribet 2014

So… what do we do to test the robustness of our tree?

1) Build different subsets of your data through a subsampling strategy selecting genes with different properties

2) Run different analyses that rely on different assumptions or apply different models

3) Do 1) and 2) both at the level of *supermatrix* and *subset of individual gene trees*



(subset of indiv. gene trees)

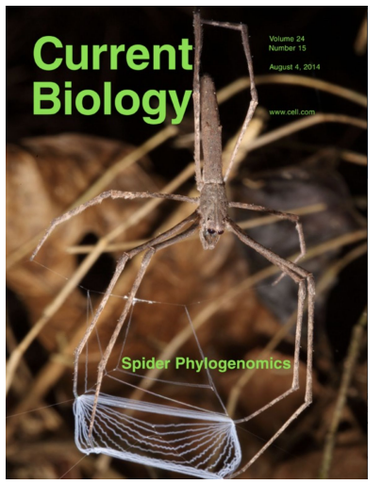
Fernández, Hormiga and Giribet 2014

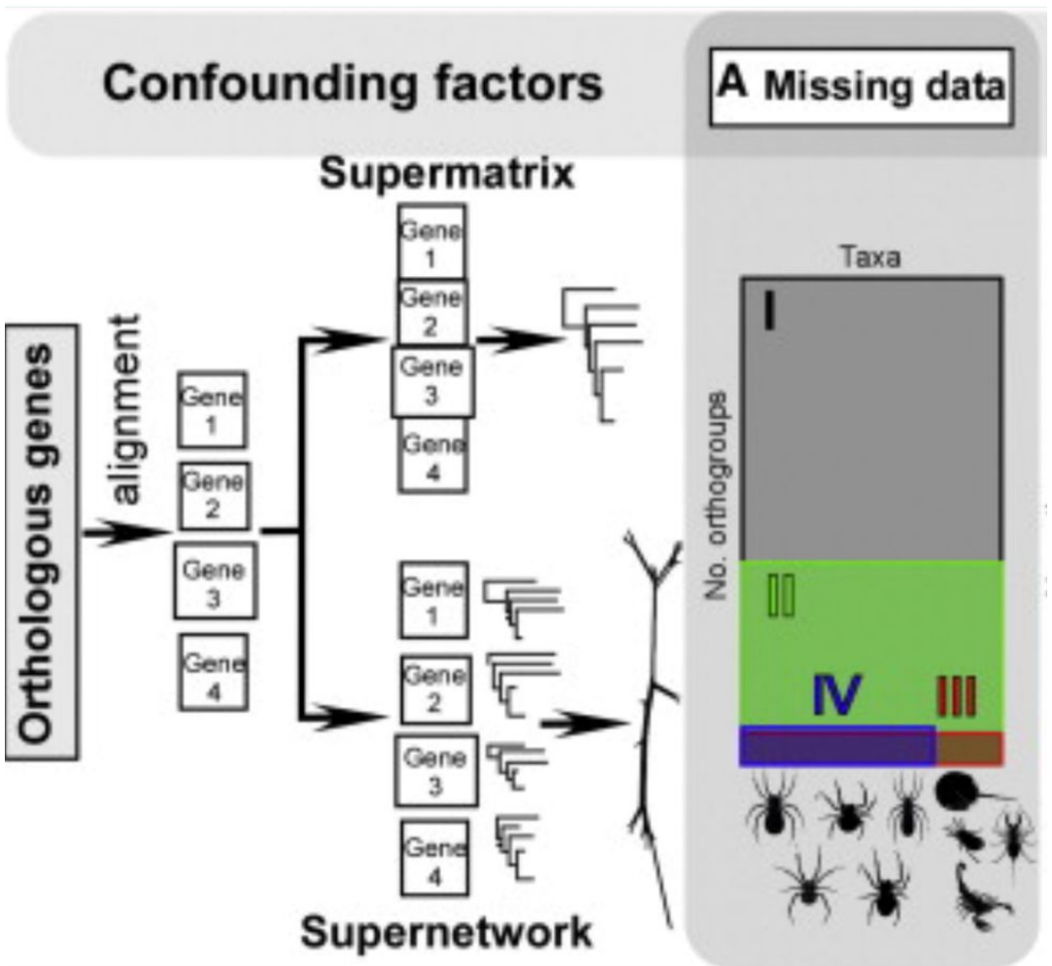
So… what do we do to test the robustness of our tree?

1) Build different subsets of your data through a subsampling strategy selecting genes with different properties

2) Run different analyses that rely on different assumptions or apply different models

3) Do 1) and 2) both at the level of *supermatrix* and *subset of individual gene trees*





(subset of indiv. gene trees)

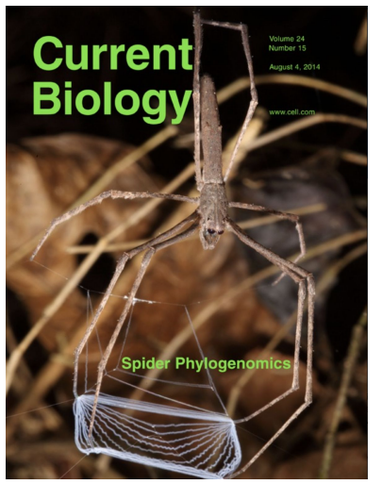
Fernández, Hormiga and Giribet 2014

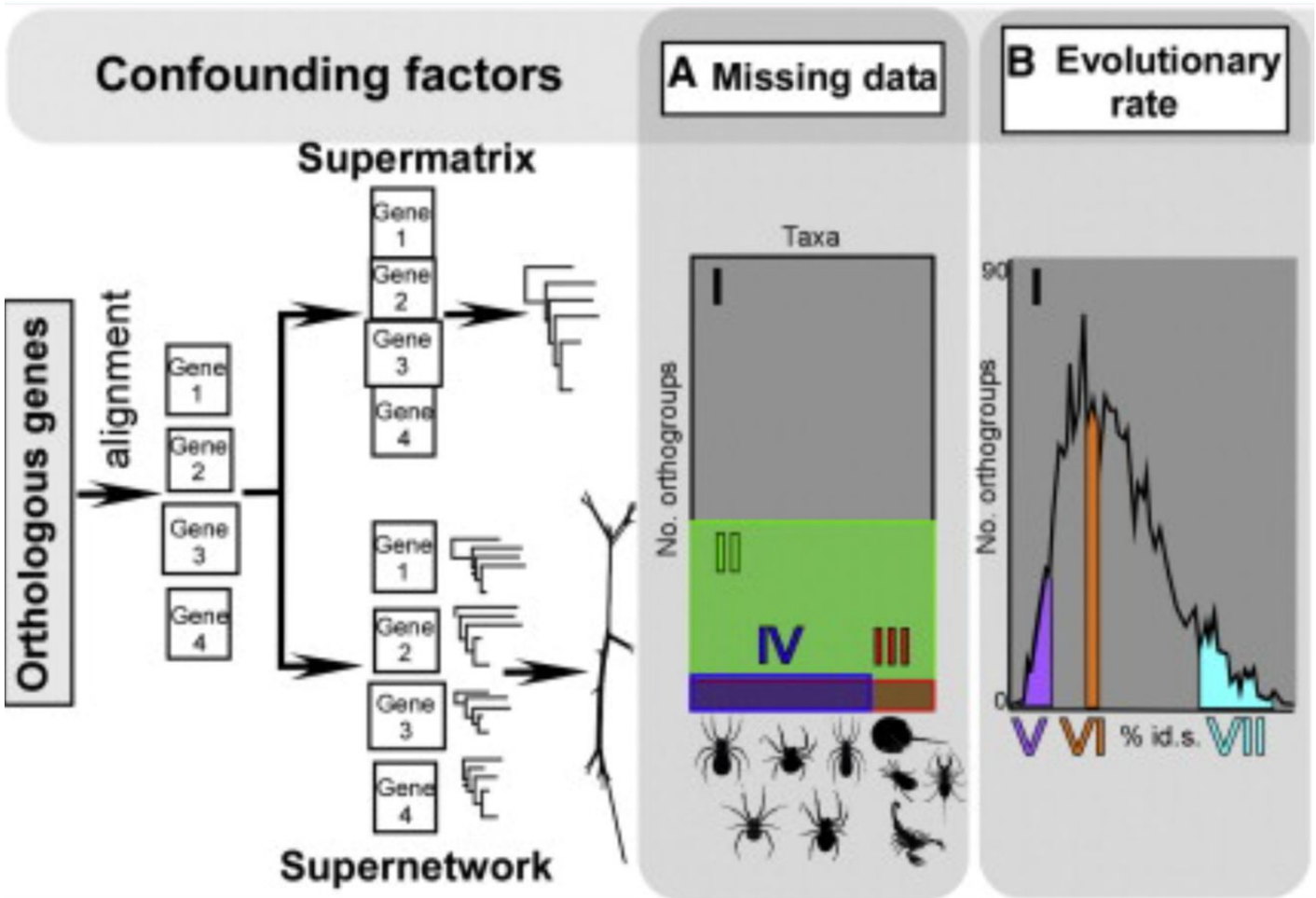
So… what do we do to test the robustness of our tree?

1) Build different subsets of your data through a subsampling strategy selecting genes with different properties

2) Run different analyses that rely on different assumptions or apply different models

3) Do 1) and 2) both at the level of *supermatrix* and *subset of individual gene trees*





(subset of indiv. gene trees)

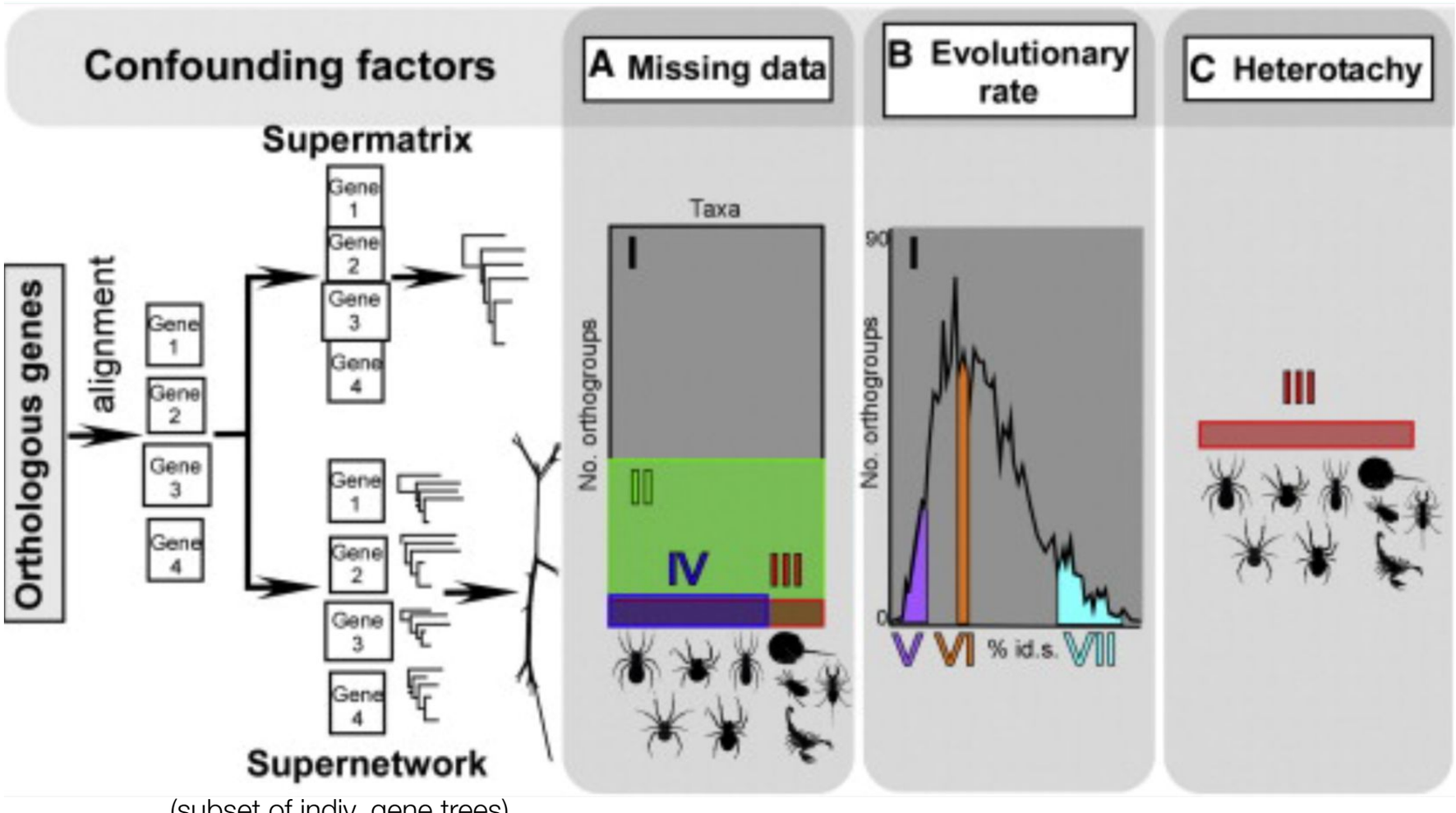
Fernández, Hormiga and Giribet 2014

So… what do we do to test the robustness of our tree?

1) Build different subsets of your data through a subsampling strategy selecting genes with different properties

2) Run different analyses that rely on different assumptions or apply different models

3) Do 1) and 2) both at the level of *supermatrix* and *subset of individual gene trees*



(subset of indiv. gene trees)

Fernández, Hormiga and Giribet 2014

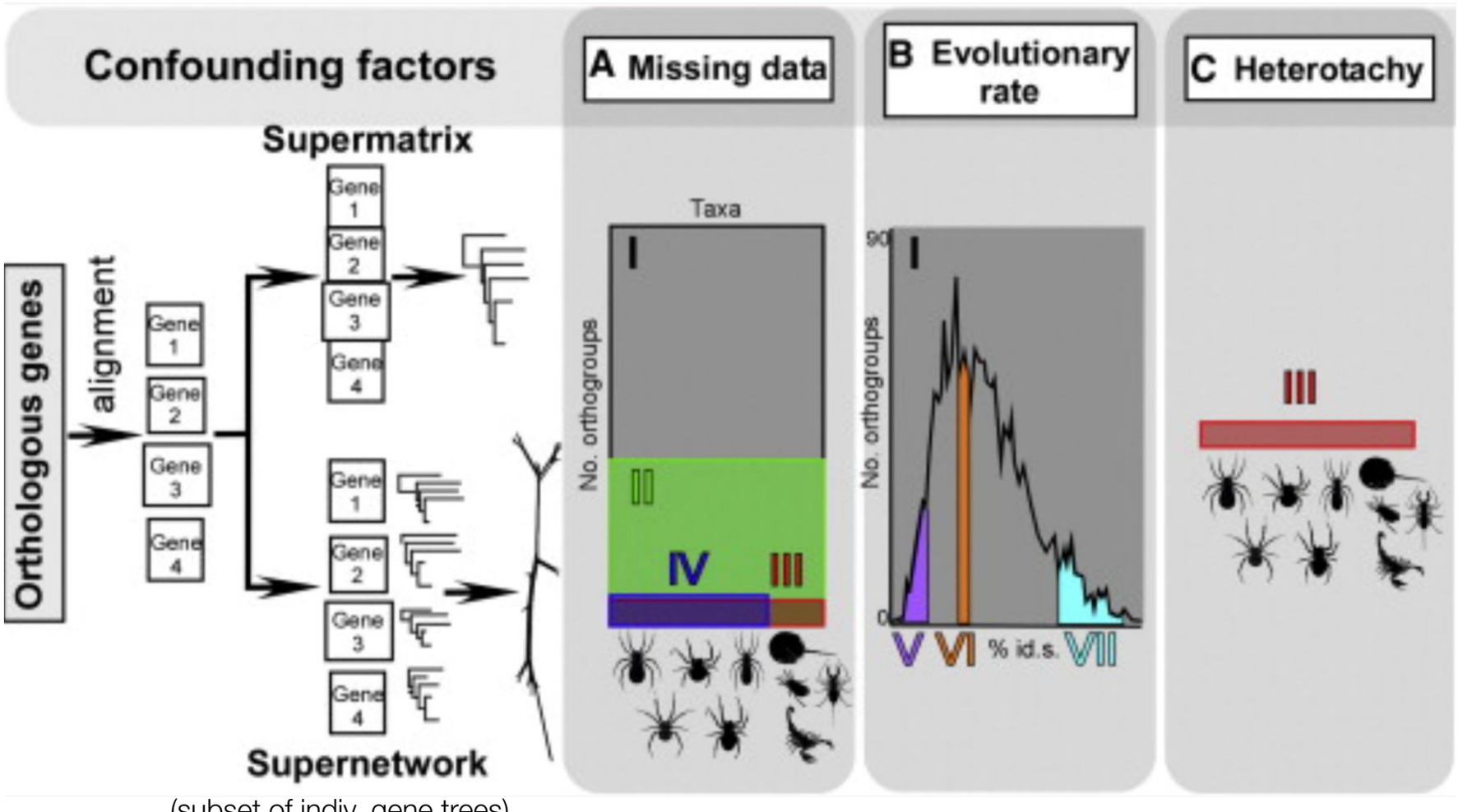
So… what do we do to test the robustness of our tree?

1) Build different subsets of your data through a subs
properties

2) Run different analyses that rely on different assump

3) Do 1) and 2) both at the level of *supermatrix* a

Heterotachy refers to the phenomenon of **a site in a gene-sequence changing its rate of evolution throughout the tree** (ie, sometimes evolving fast, some others evolving slow)



(subset of indiv. gene trees)

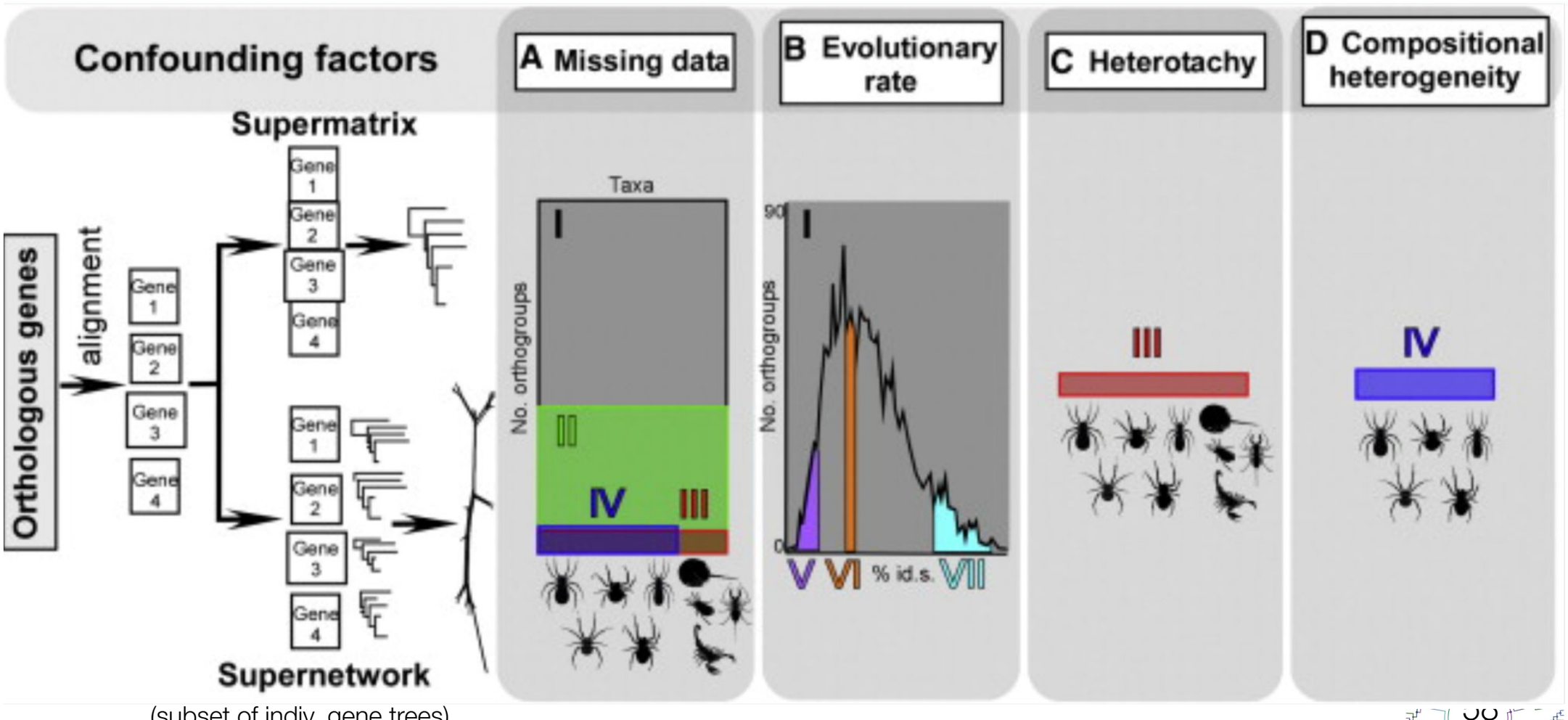
Fernández, Hormiga and Giribet 2014

So… what do we do to test the robustness of our tree?

1) Build different subsets of your data through a subsampling strategy selecting genes with different properties

2) Run different analyses that rely on different assumptions or apply different models

3) Do 1) and 2) both at the level of *supermatrix* and *subset of individual gene trees*



(subset of indiv. gene trees)

Fernández, Hormiga and Giribet 2014

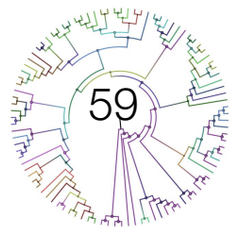
So… what do we do to test the robustness of our tree?

1) **Build different subsets of your data through a subsampling strategy selecting genes with different properties**

2) **Run different analyses that rely on different assumptions or apply different models**

3) **Do 1) and 2) both at the level of *supermatrix* and *subset of individual gene trees***

Fernández, Hormiga and Giribet 2014



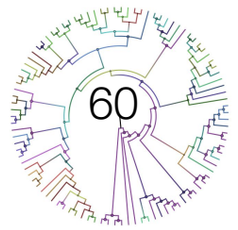
So… what do we do to test the robustness of our tree?

1) **Build different subsets of your data through a subsampling strategy selecting genes with different properties**

2) **Run different analyses that rely on different assumptions or apply different models**

3) **Do 1) and 2) both at the level of *supermatrix* and *subset of individual gene trees***

THIS IS THE MOST IMPORTANT TAKE-HOME MESSAGE FROM TODAY'S LESSON!!



So… what do we do to test the robustness of our tree?

1) **Build different subsets of your data through a subsampling strategy selecting genes with different properties**

2) **Run different analyses that rely on different assumptions or apply different models**

3) **Do 1) and 2) both at the level of *supermatrix* and *subset of individual gene trees***

THIS IS THE MOST IMPORTANT TAKE-HOME MESSAGE FROM TODAY'S LESSON!!

Let's take a minute to interiorize it and learn it by heart.

This is what you need to know to think critically when *designing* or *evaluating* a phylogenomic study

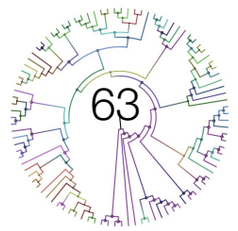
So… what do we do to test the robustness of our tree?

1) **Build different subsets of your data through a subsampling strategy selecting genes with different properties**

2) **Run different analyses that rely on different assumptions or apply different models**

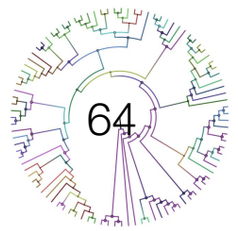
3) **Do 1) and 2) both at the level of *supermatrix* and *subset of individual gene trees***

And… how do I *chose* a subset of genes to run these analyses?



And… how do I *chose* a subset of genes to run these analyses?

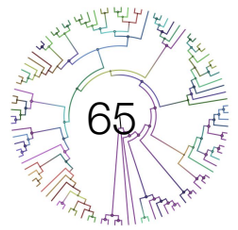
a) Random subsampling (eg, select randomly 30% of your initial data)



And… how do I *chose* a subset of genes to run these analyses?

a) Random subsampling (eg, select randomly 30% of your initial data)

b) Check the properties of the genes and chose the ones that behave ‘well’ (eg, discard the outliers).

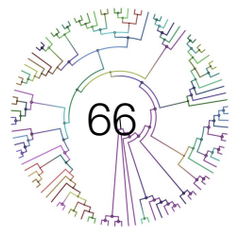


And… how do I *chose* a subset of genes to run these analyses?

a) Random subsampling (eg, select randomly 30% of your initial data)

b) Check the properties of the genes and chose the ones that behave ‘well’ (eg, discard the outliers).

-> Custom scripts (eg, select genes with less 50% of missing data)



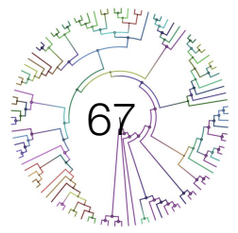
And… how do I *chose* a subset of genes to run these analyses?

a) Random subsampling (eg, select randomly 30% of your initial data)

b) Check the properties of the genes and chose the ones that behave ‘well’ (eg, discard the outliers).

-> Custom scripts (eg, select genes with less 50% of missing data)

-> Software to measure some of these properties (eg, compositional heterogeneity, saturation, etc.)



And… how do I *chose* a subset of genes to run these analyses?

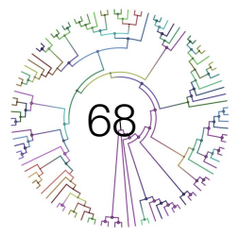
a) Random subsampling (eg, select randomly 30% of your initial data)

b) Check the properties of the genes and chose the ones that behave ‘well’ (eg, discard the outliers).

-> Custom scripts (eg, select genes with less 50% of missing data)

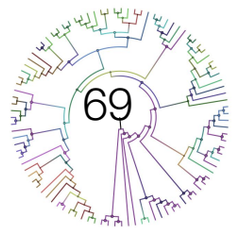
-> Software to measure some of these properties (eg, compositional heterogeneity, saturation, etc.)

We will be doing this today in the first half of our hands-on session



Generating phylogenomic data matrices

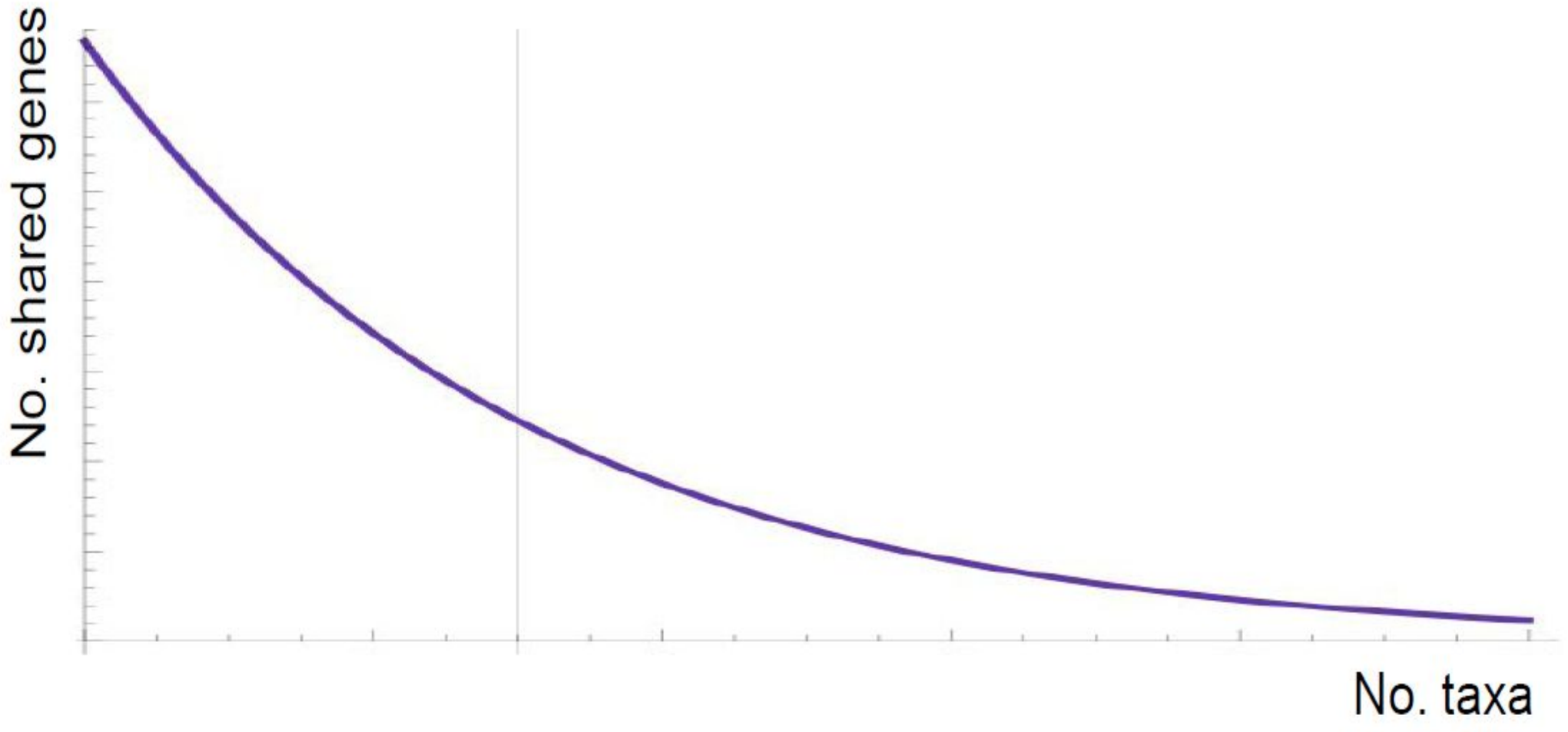
- **MISSING DATA:** depends on the type of data



Generating phylogenomic data matrices

- **MISSING DATA:** depends on the type of data

Empirical evidence from different datasets in transcriptomics

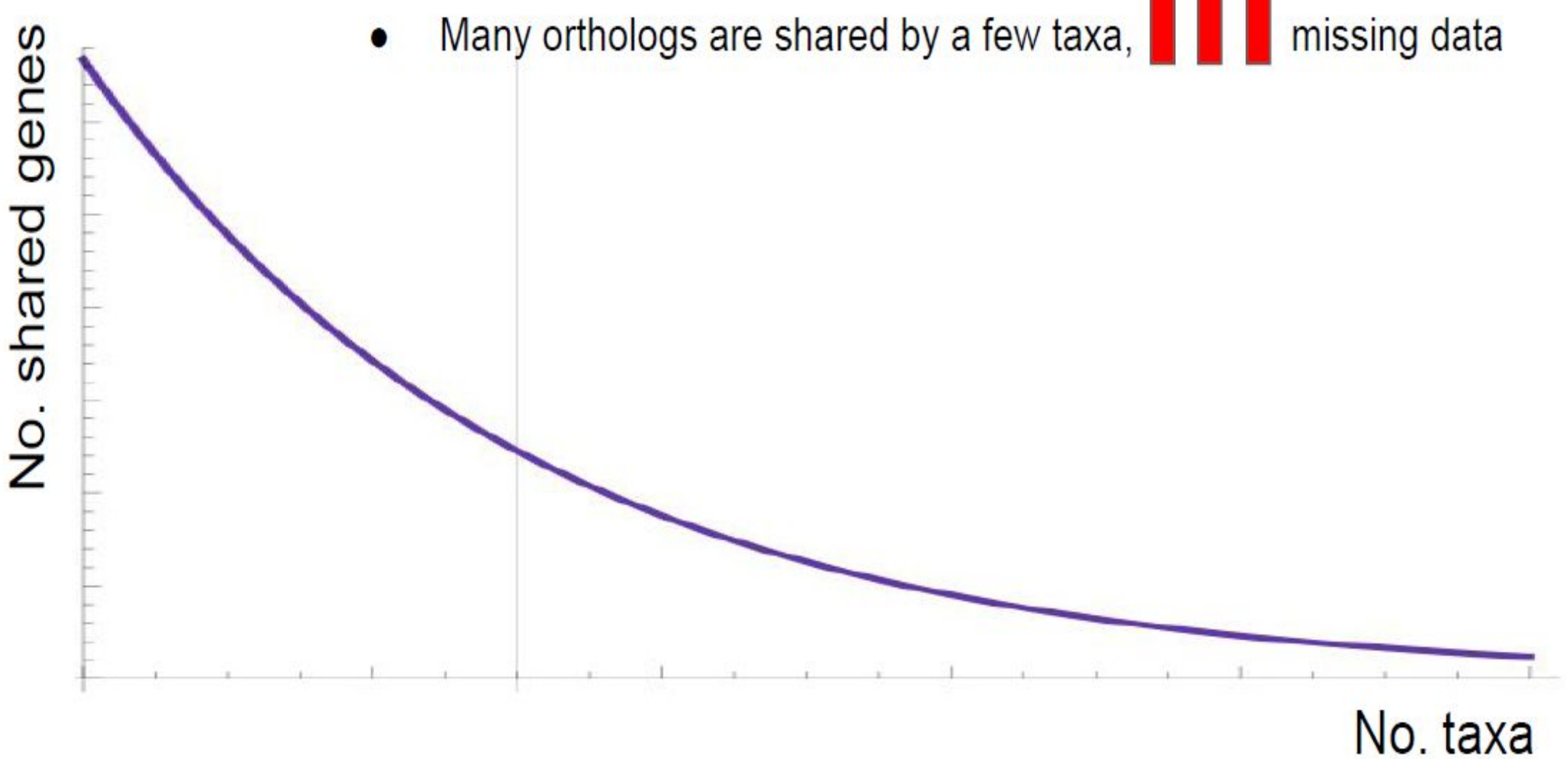


Generating phylogenomic data matrices

- **MISSING DATA:** depends on the type of data

Empirical evidence from different datasets in transcriptomics

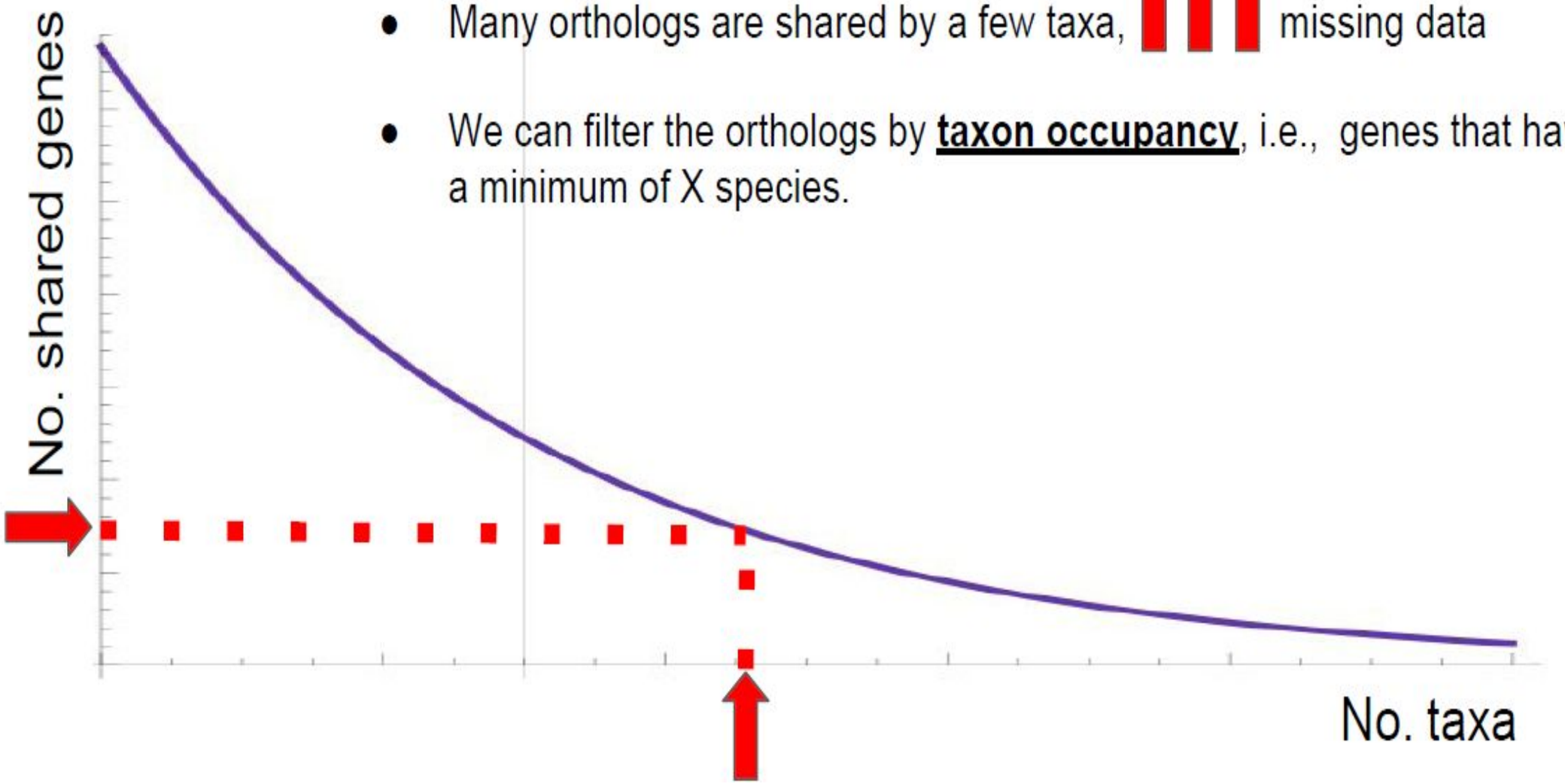
- Many orthologs are shared by a few taxa, ⇑⇑⇑ missing data



Generating phylogenomic data matrices

- **MISSING DATA:** depends on the type of data

Empirical evidence from different datasets in transcriptomics

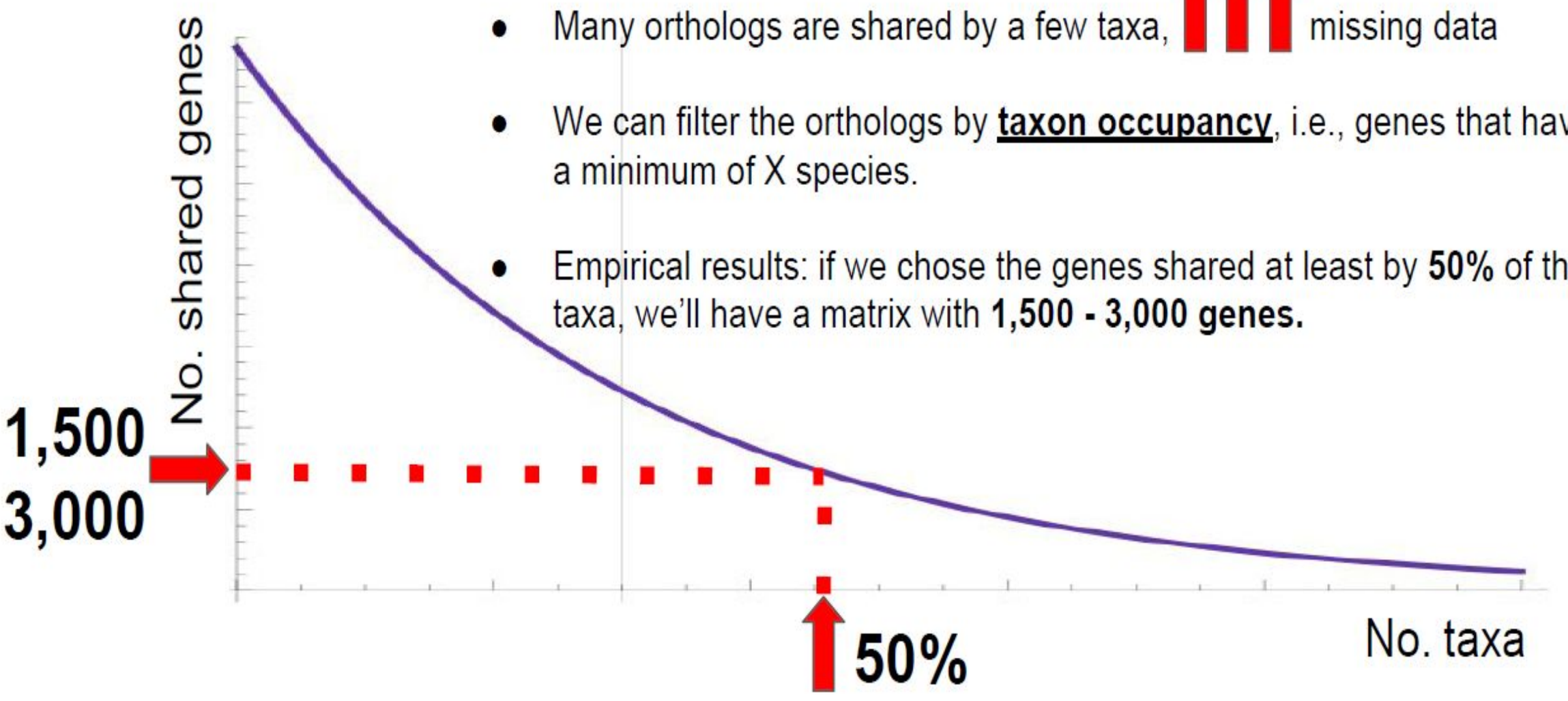


- Many orthologs are shared by a few taxa, missing data
- We can filter the orthologs by **taxon occupancy**, i.e., genes that have a minimum of X species.

Generating phylogenomic data matrices

- **MISSING DATA:** depends on the type of data

Empirical evidence from different datasets in transcriptomics

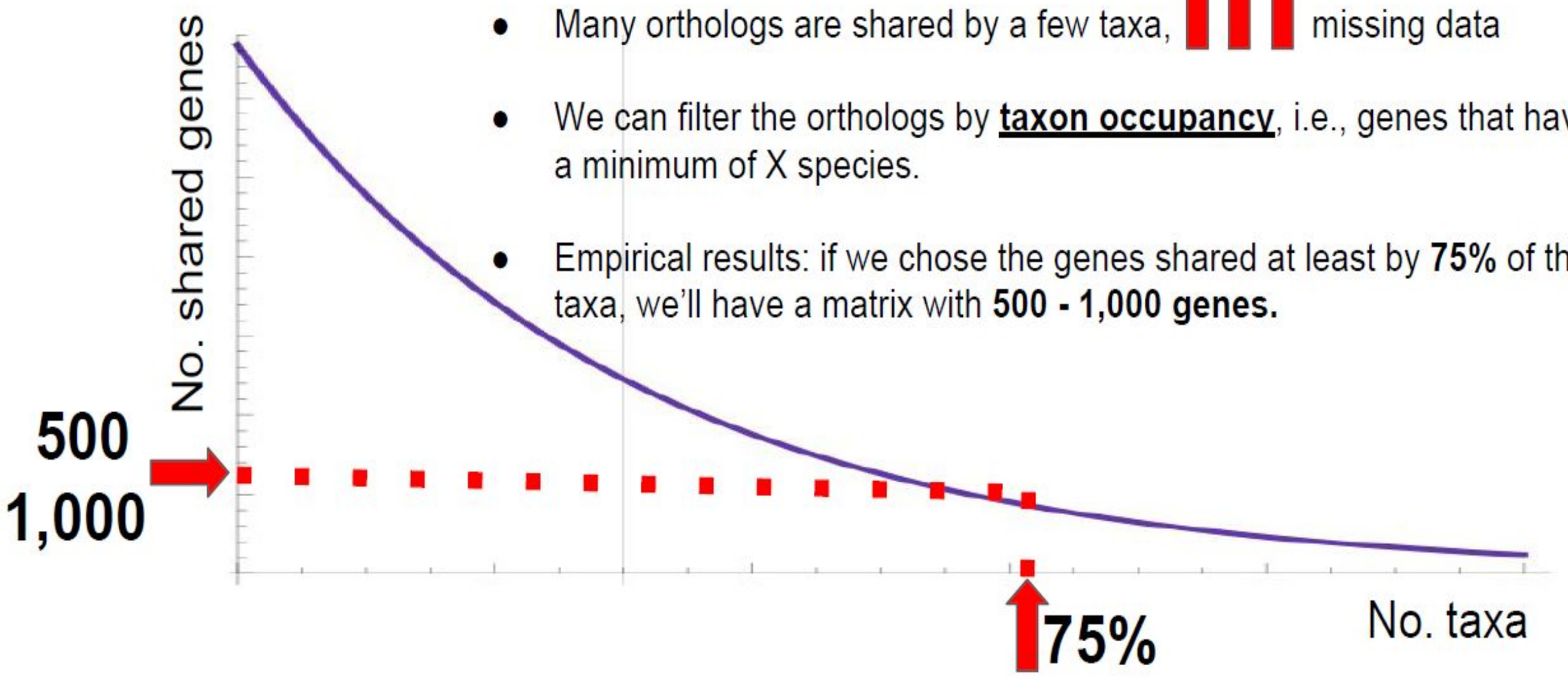


- Many orthologs are shared by a few taxa, ⇑⇑⇑ missing data
- We can filter the orthologs by **taxon occupancy**, i.e., genes that have a minimum of X species.
- Empirical results: if we chose the genes shared at least by **50%** of the taxa, we'll have a matrix with **1,500 - 3,000 genes.**

Generating phylogenomic data matrices

- **MISSING DATA:** depends on the type of data

Empirical evidence from different datasets in transcriptomics

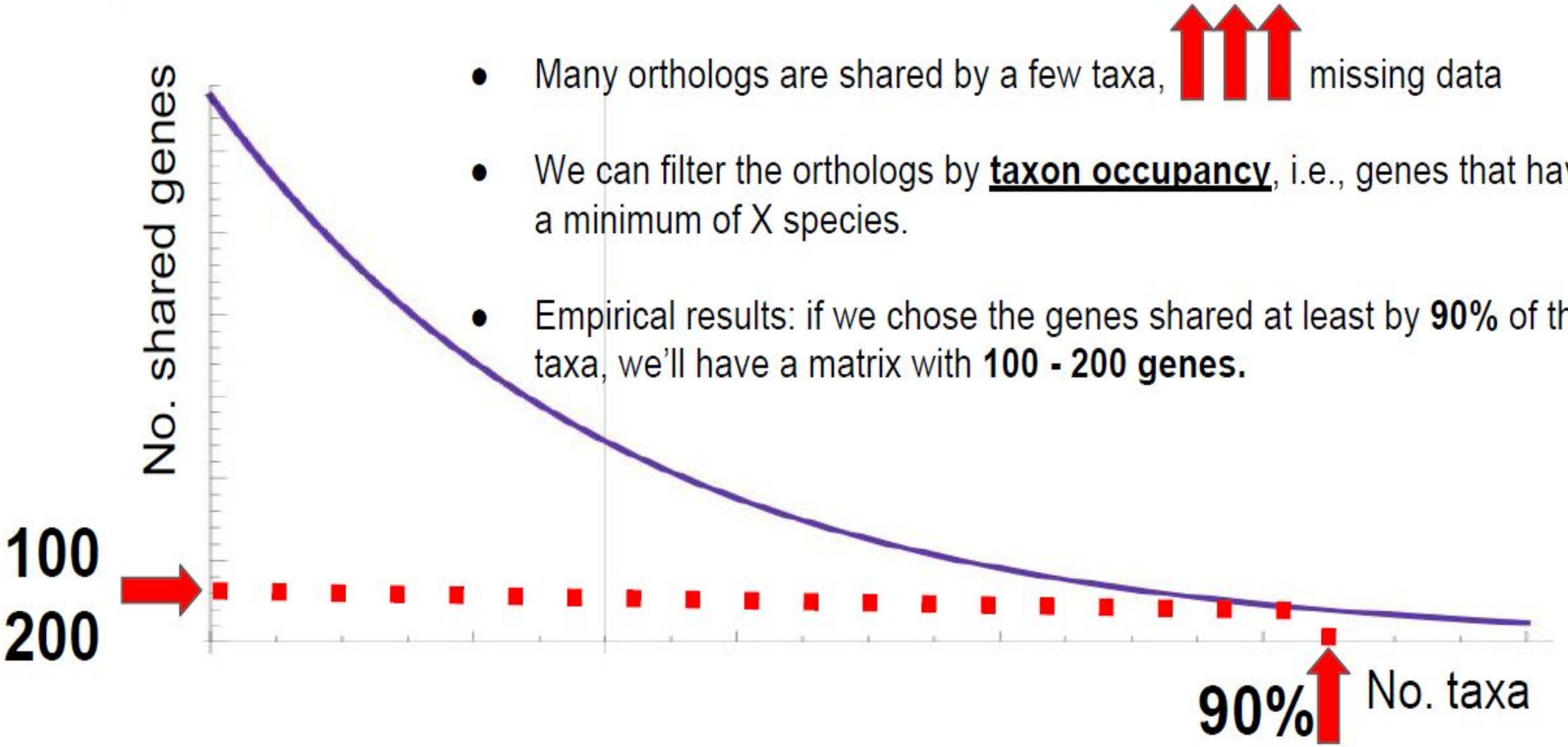


- Many orthologs are shared by a few taxa, ⇑⇑⇑ missing data
- We can filter the orthologs by **taxon occupancy**, i.e., genes that have a minimum of X species.
- Empirical results: if we chose the genes shared at least by **75%** of the taxa, we'll have a matrix with **500 - 1,000 genes.**

Generating phylogenomic data matrices

- **MISSING DATA:** depends on the type of data

Empirical evidence from different datasets in transcriptomics

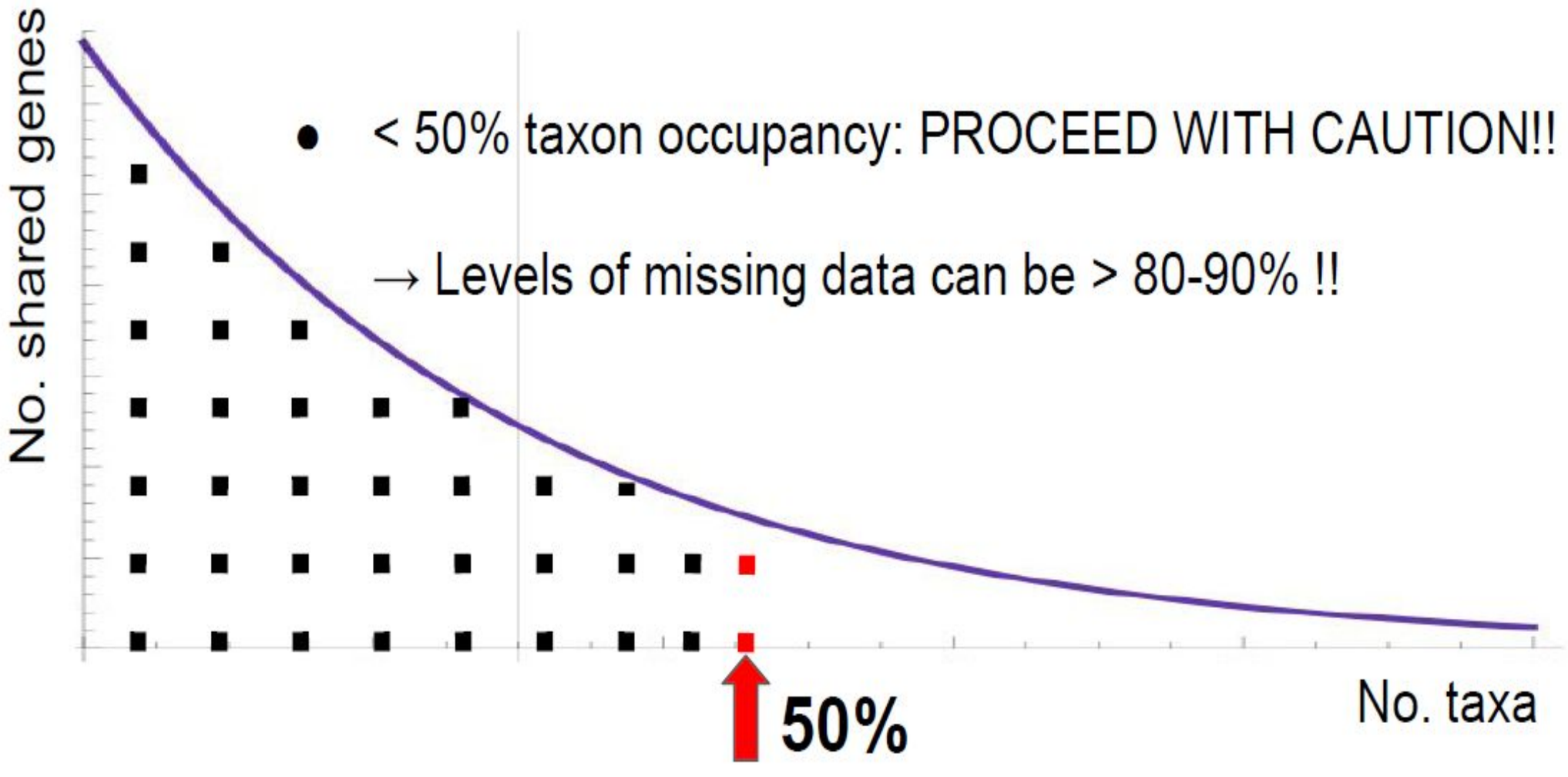


- Many orthologs are shared by a few taxa, missing data
- We can filter the orthologs by **taxon occupancy**, i.e., genes that have a minimum of X species.
- Empirical results: if we chose the genes shared at least by **90%** of the taxa, we'll have a matrix with **100 - 200 genes.**

Generating phylogenomic data matrices

- **MISSING DATA:** depends on the type of data

Empirical evidence from different datasets in transcriptomics



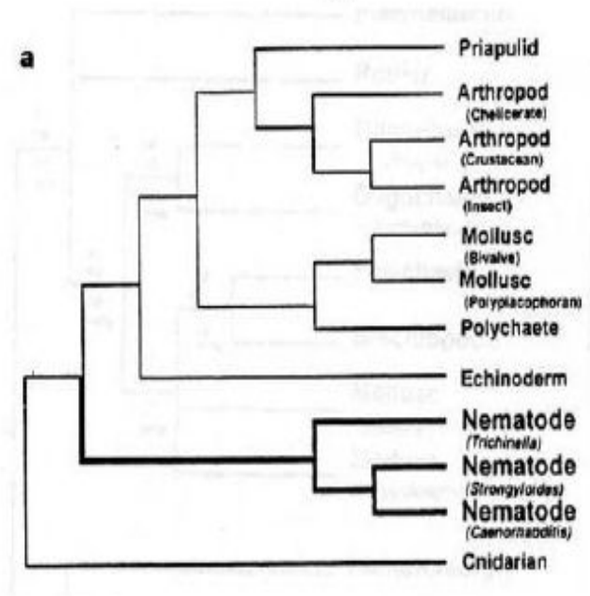
- < 50% taxon occupancy: PROCEED WITH CAUTION!!

→ Levels of missing data can be > 80-90% !!

Generating phylogenomic data matrices

- **LONG BRANCH ATTRACTION (LBA)**

Long-Branch Attraction Problem



Some taxa have **fast-evolving DNA**

Often drop out at **base** of tree, clustered with:

(a) basal lineages / outgroups

(b) other fast-evolvers, whom they may not be related to

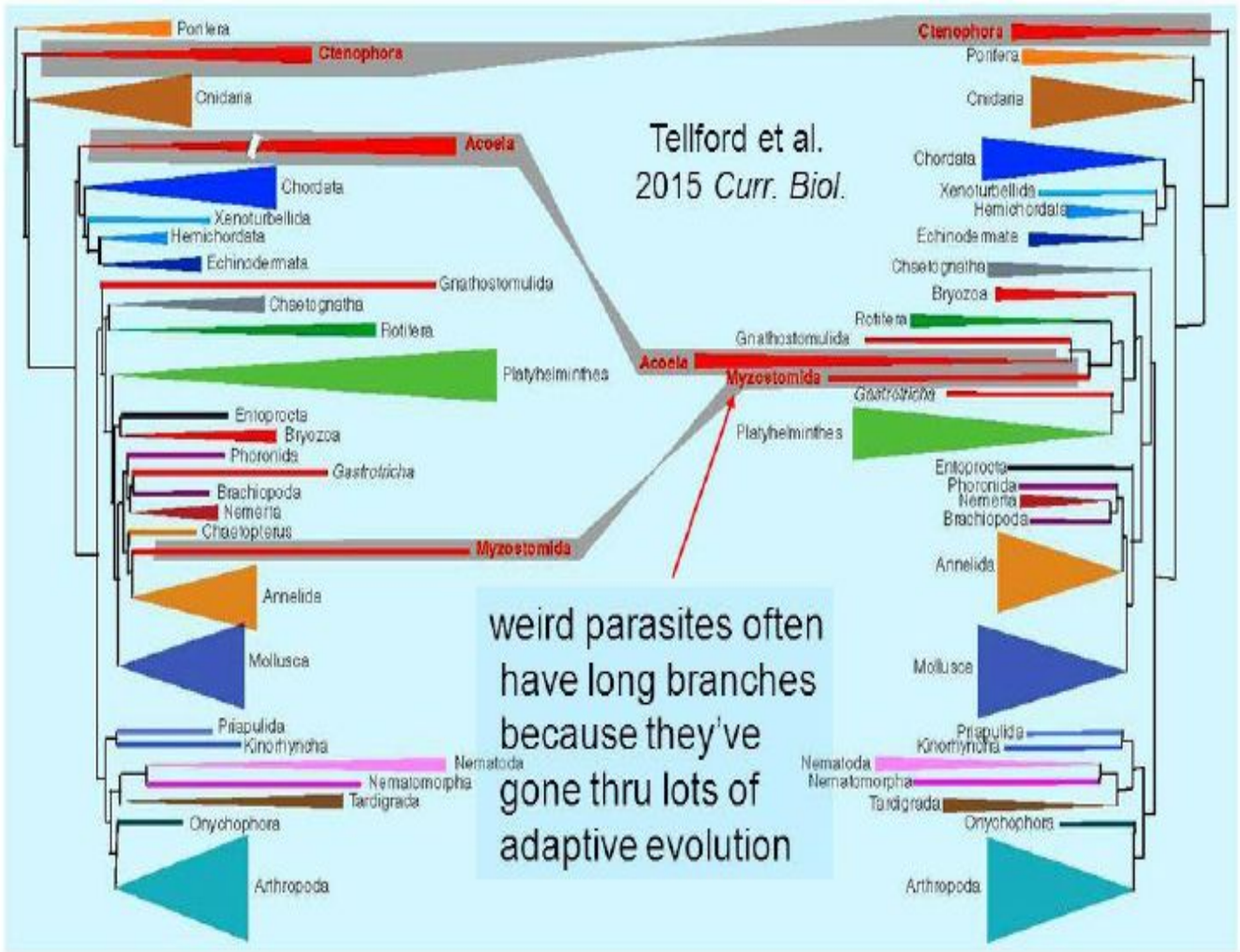
This is an artifact (a false result) of how computer programs analyze DNA sequences, called **long-branch attraction**

- sequences that are fast-evolving give very long branches on trees, which tend to "attract" other long branches

→ sequences that are very different (fast mutating) get lumped together with other fast-evolving sequences

Generating phylogenomic data matrices

- **LONG BRANCH ATTRACTION (LBA)**



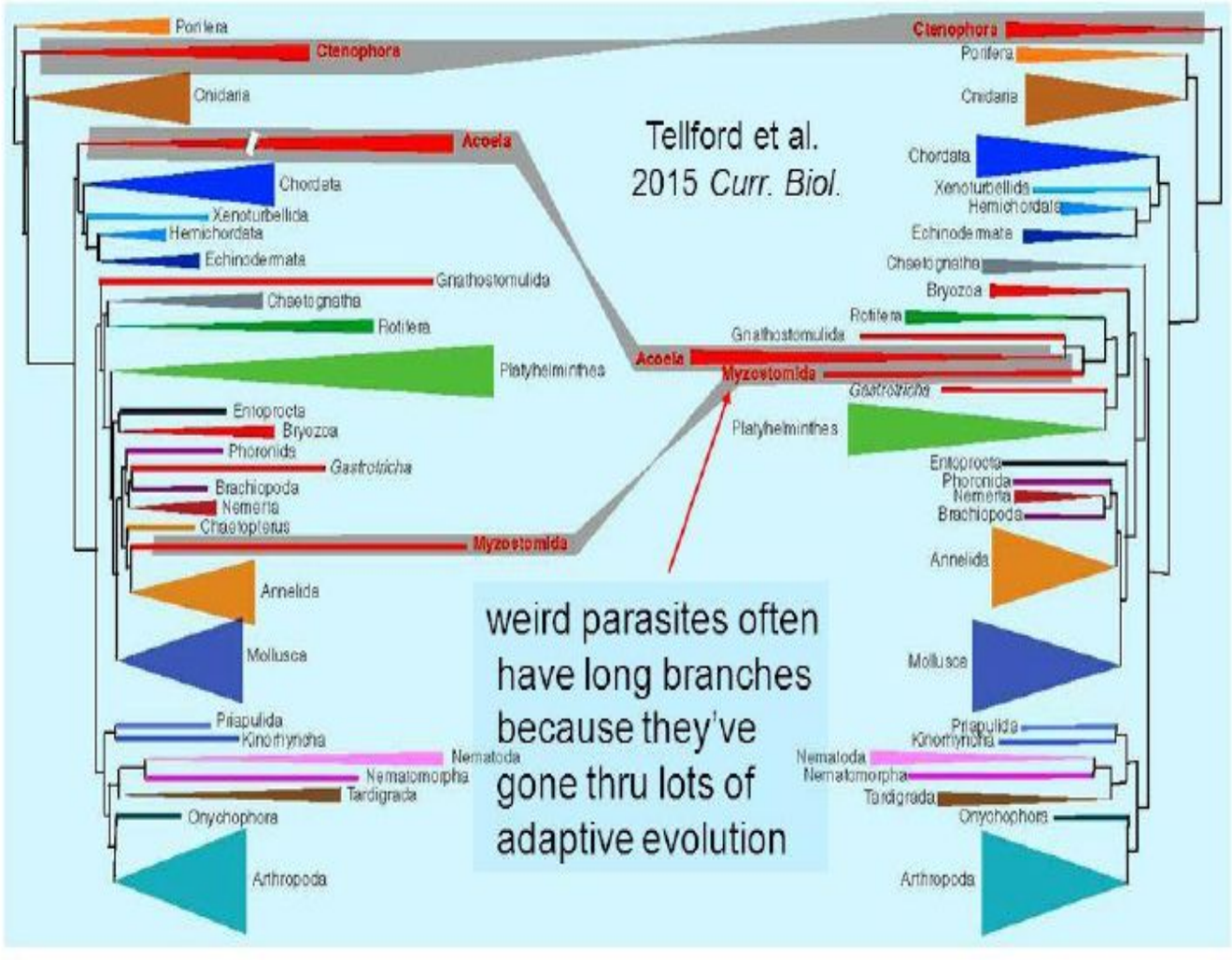
correcting for long-branch attraction

not correcting for long-branch attraction

- **Correcting for LBA:**
  - Compositional heterogeneity

Generating phylogenomic data matrices

- **LONG BRANCH ATTRACTION (LBA)**



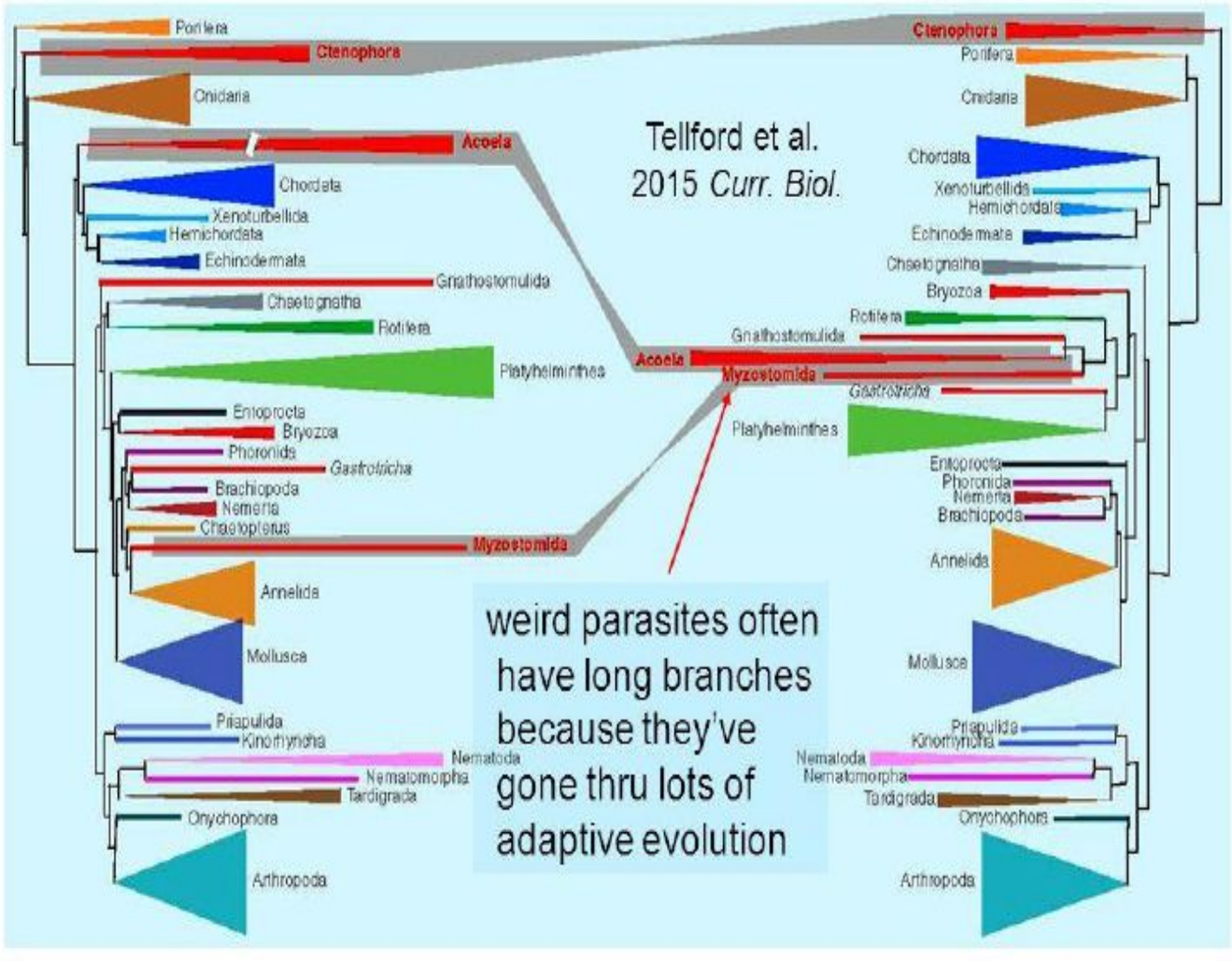
correcting for long-branch attraction

not correcting for long-branch attraction

- **Correcting for LBA:**
  - Compositional heterogeneity
  - Evolutionary rate (also related to resolving deep nodes vs shallow ones)

Generating phylogenomic data matrices

- **LONG BRANCH ATTRACTION (LBA)**



correcting for long-branch attraction

not correcting for long-branch attraction

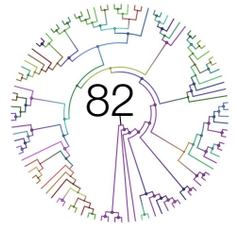
- **Correcting for LBA:**
  - Compositional heterogeneity
  - Evolutionary rate (also related to resolving deep nodes vs shallow ones)

So… how many matrices/subsets/analyses should I analyze?

So… how many matrices/subsets/analyses should I analyze?

Many.

As many as you can!

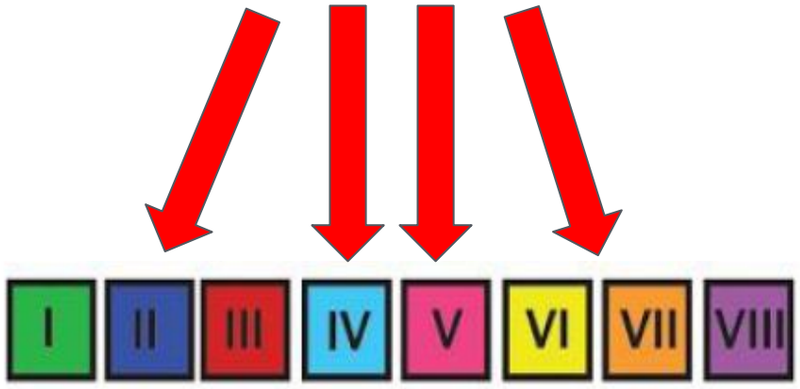


So… how many matrices/subsets/analyses should I analyze?

Fernández, Edgecombe & Giribet (2016) Syst Biol

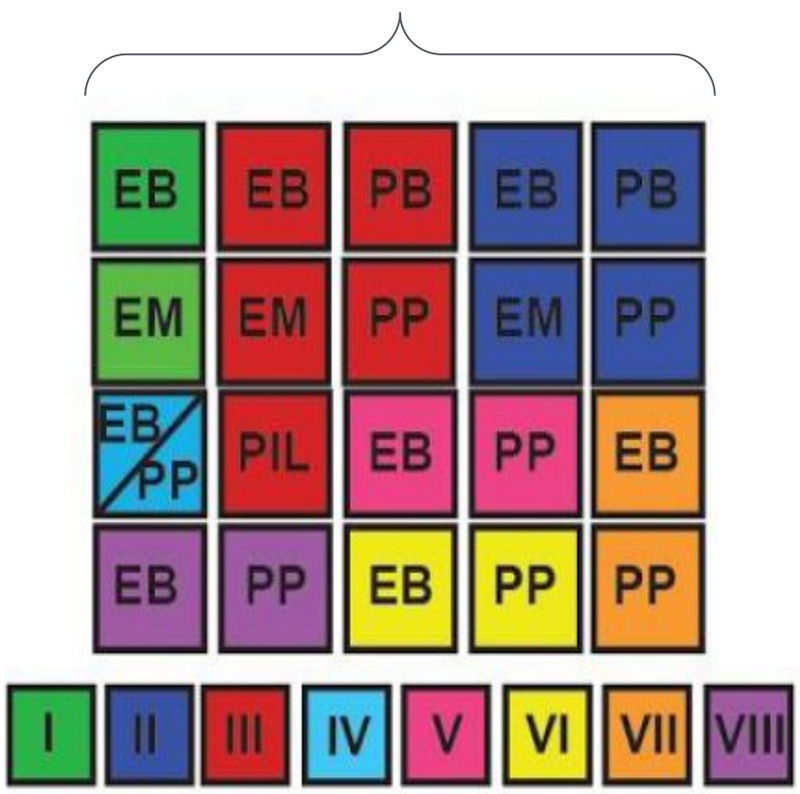
So… how many matrices/subsets/analyses should I analyze?

These are **matrices/subsets** of individual gene trees



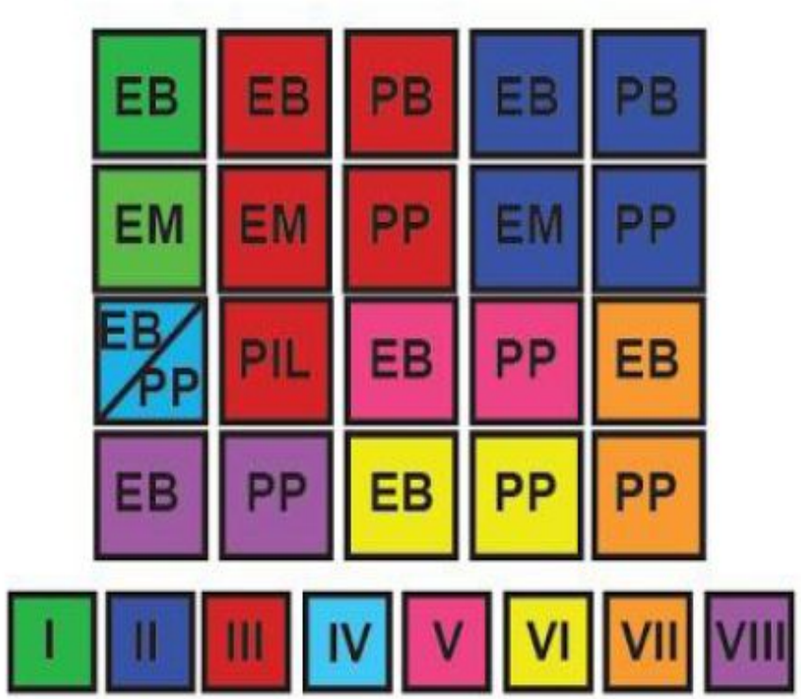
Fernández, Edgecombe & Giribet (2016) Syst Biol

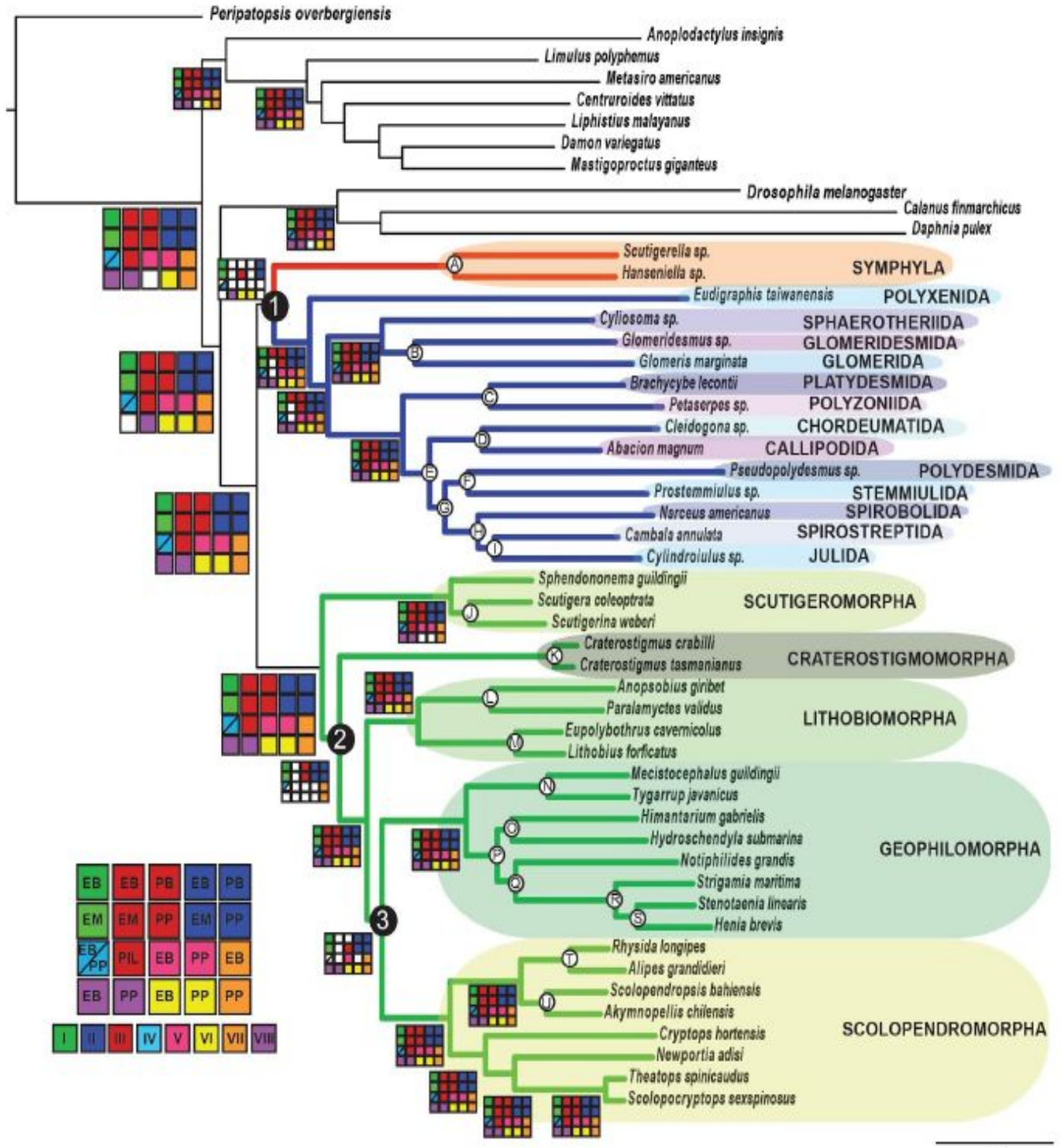
So… how many matrices/subsets/analyses should I analyze?



Fernández, Edgecombe & Giribet (2016) Syst Biol

So… how many matrices/subsets/analyses should I analyze?





Fernández, Edgecombe & Giribet (2016) Syst Biol

Generating phylogenomic data matrices: hands-on session

Case study: phylogenomic interrogation of the genus *Ursus*

The *goals* of the project are as follow:

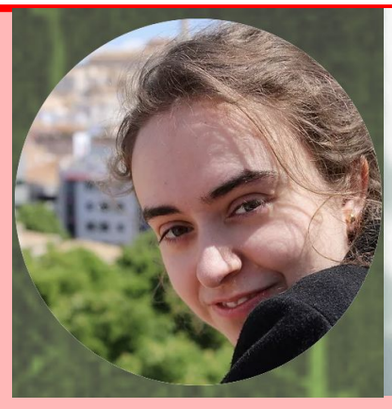
1) **Understand the phylogenetic relationships within the genus *Ursus*. Is the polar bear (*U. maritimus)* more closely related to the American black bear or the brown bear?**

1) **Understand the phylogenetic relationships within the genus *Ursus*. Is the polar bear (*U. maritimus)* more closely related to the American black bear or the brown bear?** *Type of data: transcriptomes and genomes*

- 5th July, *Cinta Pegueroles* & *Ana Rojas*:
  - **Raw data preprocessing**
    - adapter trimming, quality checking
  - ***De novo* transcriptome assembly**
  - **Multiple Sequence Alignments (I)**

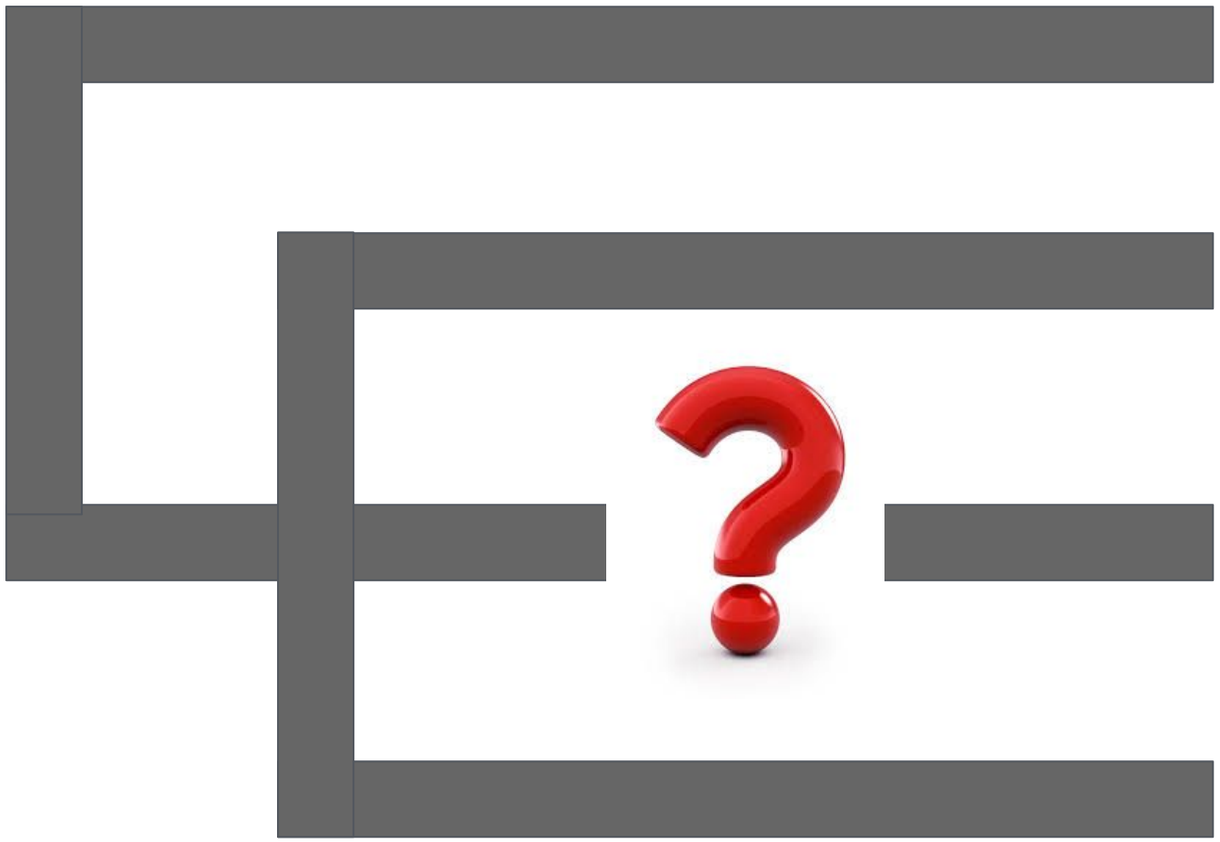

- 6th July, *Toni Gabaldón* & *Marina-Marcet Houben*:
  - From nucleotides to amino acids: **inferring protein-coding sequences**
  - **Orthology inference**
    - Tree-based approaches
    - Graph-based approaches
  - **Multiple Sequence Alignment (II)**
  - **Genome-wide phylogenetic inference**


- 8h July, *Gemma I. Martínez* & *Rosa Fernández*:
  - Testing the robustness of your results: **sensitivity analyses in phylogenomics**
  - **Supermatrix** vs **Multispecies coalescent**




Generating phylogenomic data matrices: hands-on session

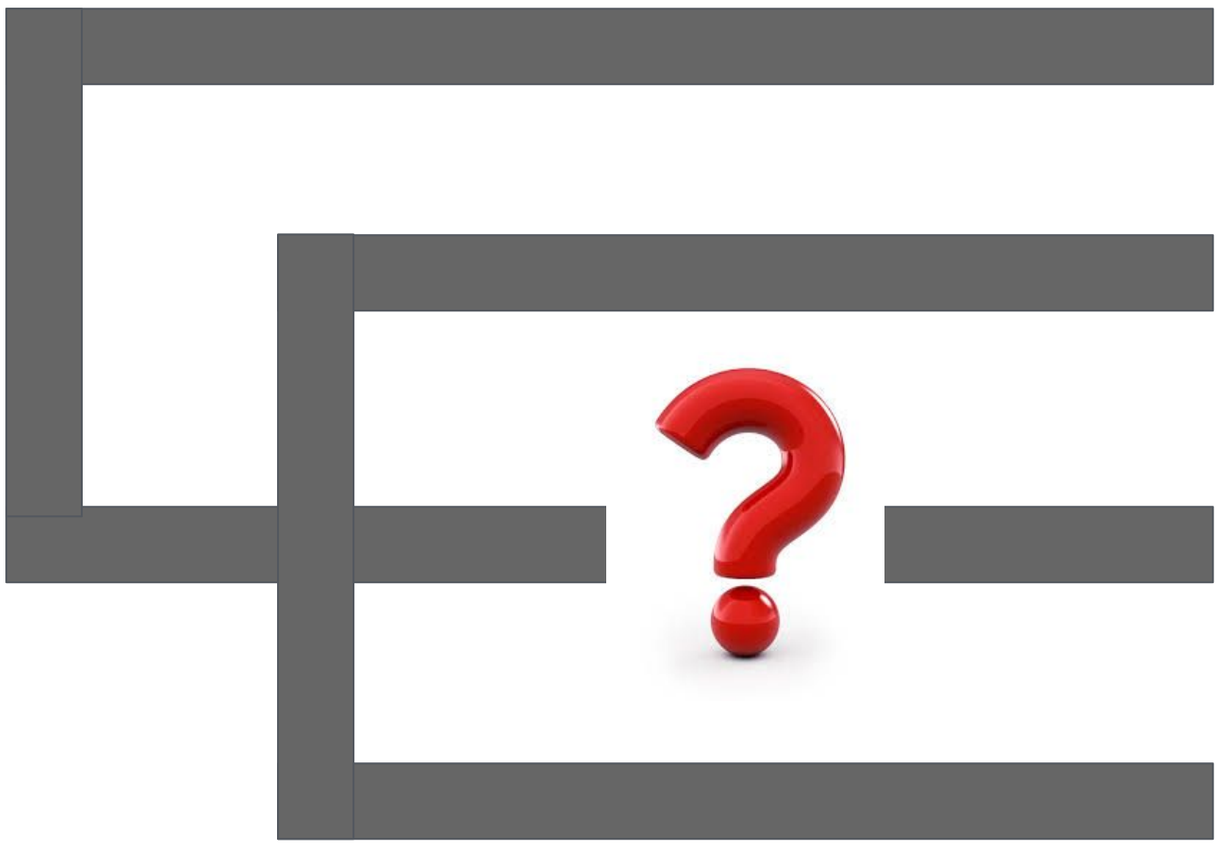
Is the polar bear the sister group to the American black bear or the brown bear?



Generating phylogenomic data matrices: hands-on session

Is the polar bear the sister group to the American black bear or the brown bear?

TOTAL: 16 samples

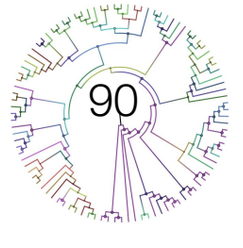


Siro
Luisa
Pepe
Juan

Noah
Oskar
Summer
Montana

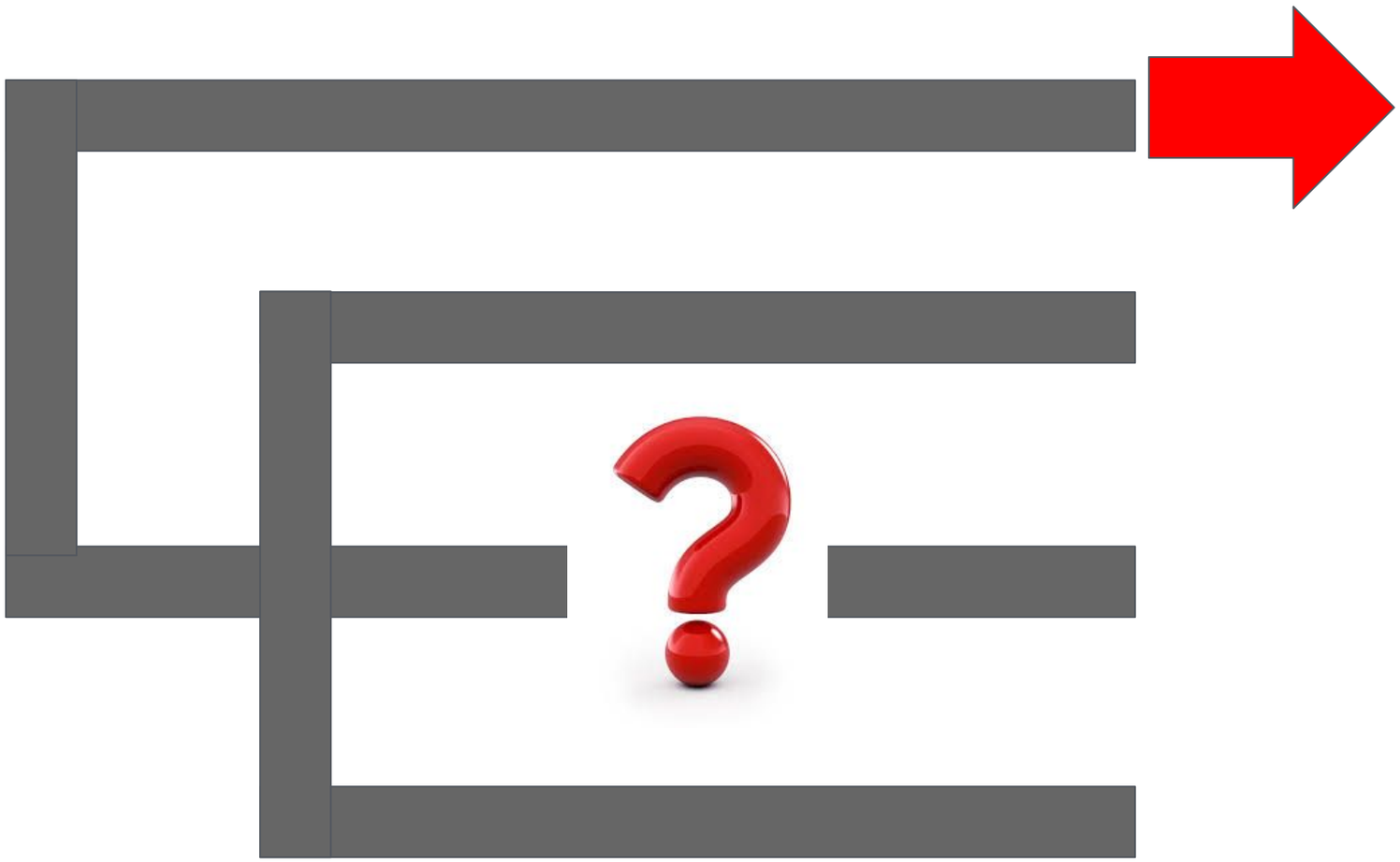
Joseph
Margaret
Maripepa
Maria

Amparo
Paco
Adelaide
Margo



Generating phylogenomic data matrices: hands-on session

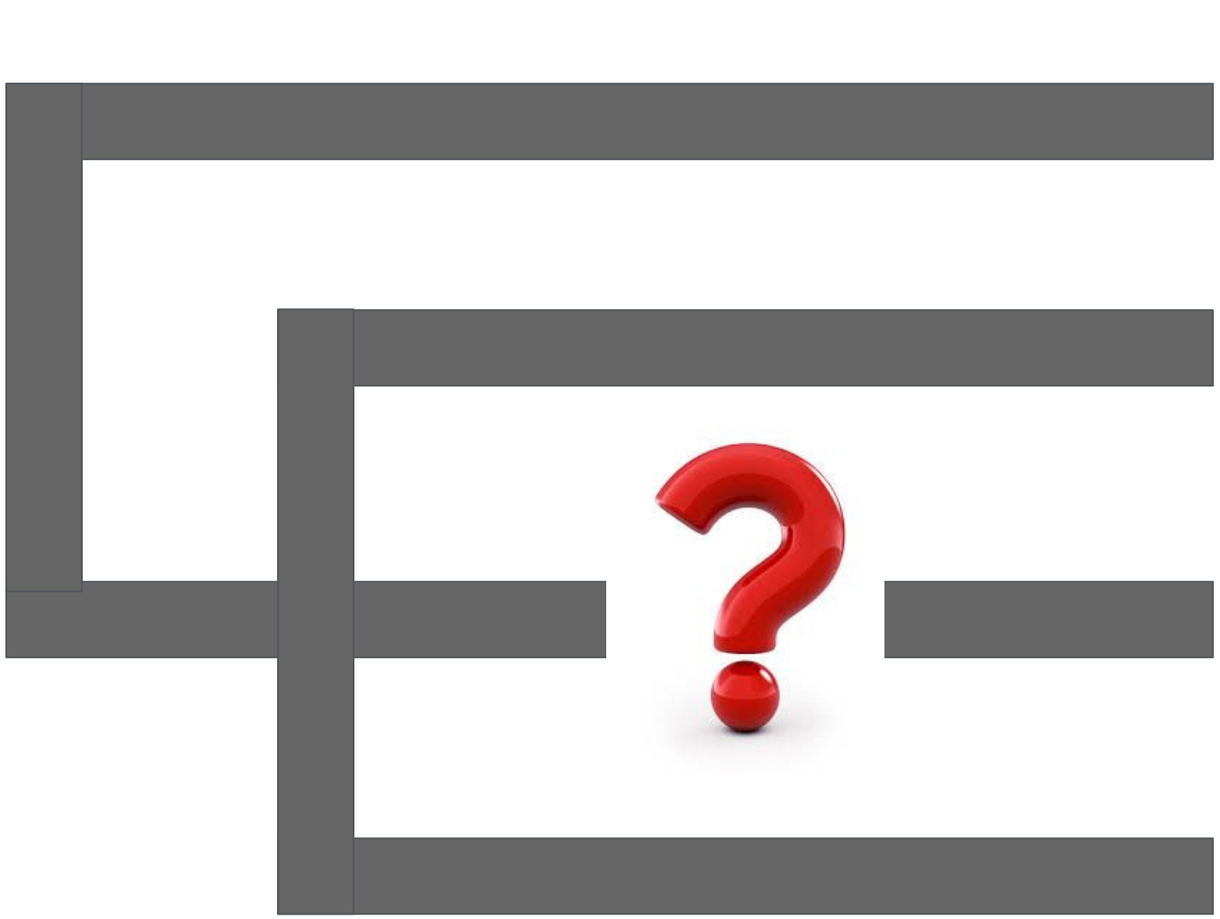
Is the polar bear the sister group to the American black bear or the brown bear?



SAMPLE COLLECTION

Generating phylogenomic data matrices: hands-on session

Is the polar bear the sister group to the American black bear or the brown bear?

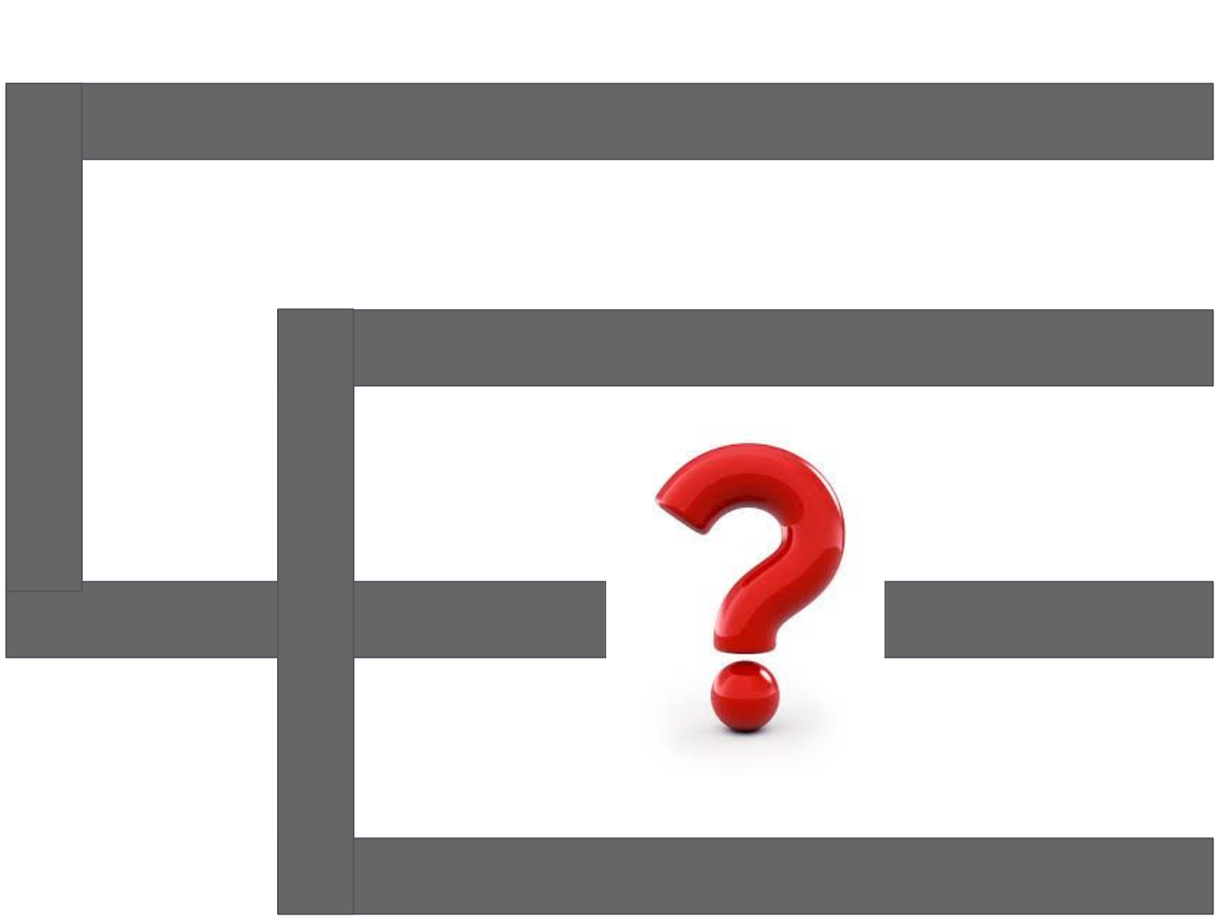


SAMPLE COLLECTION

SEQUENCING

Generating phylogenomic data matrices: hands-on session

Is the polar bear the sister group to the American black bear or the brown bear?



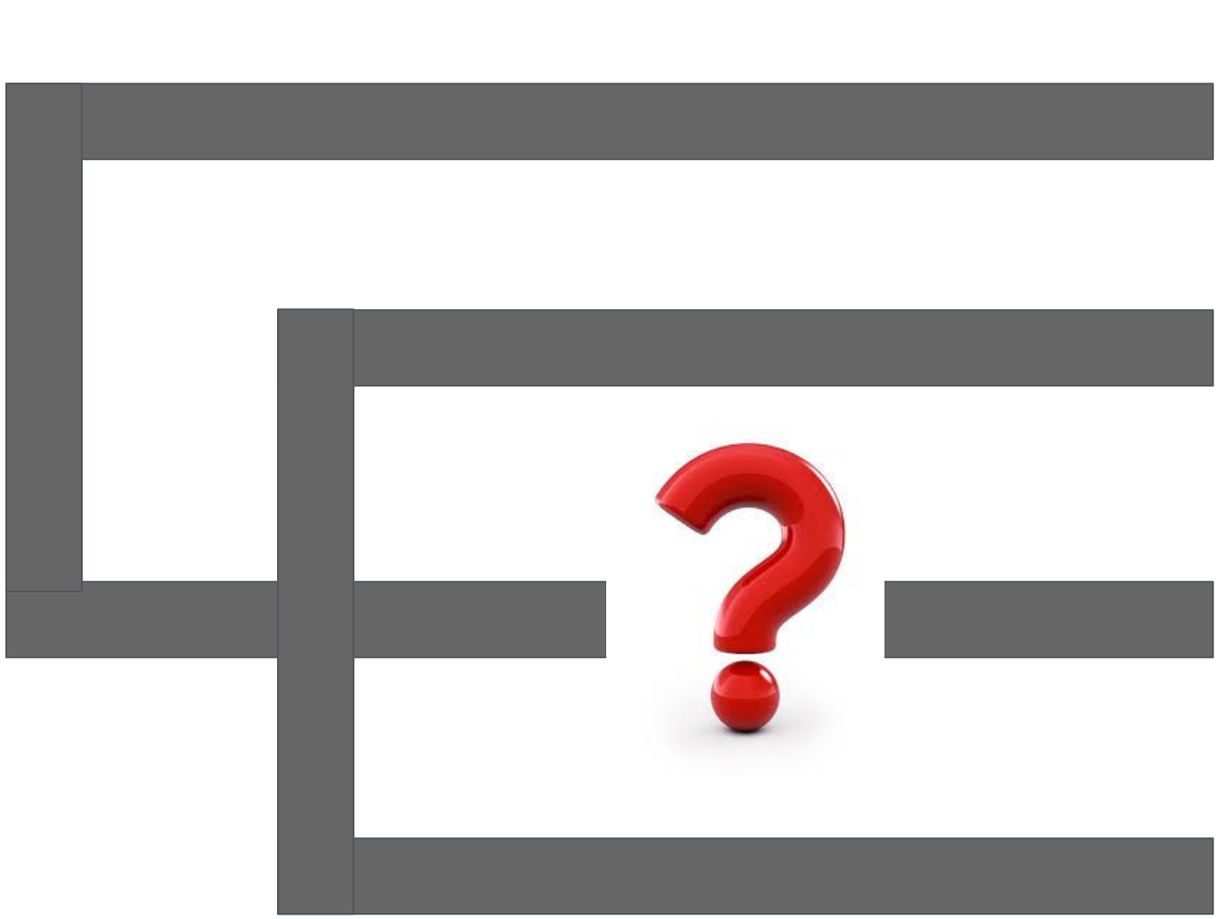
SAMPLE COLLECTION

SEQUENCING

ORTHOLOGY INFERENCE

Generating phylogenomic data matrices: hands-on session

Is the polar bear the sister group to the American black bear or the brown bear?



SAMPLE COLLECTION

SEQUENCING

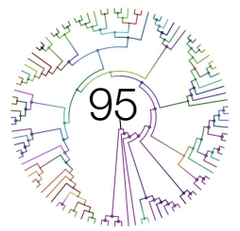
ORTHOLOGY INFERENCE

READY TO CREATE A MATRIX… BUT… WHICH GENES SHOULD WE INCLUDE??

Generating phylogenomic data matrices: hands-on session

Is the polar bear the sister group to the American black bear or the brown bear?

- **MISSING DATA**

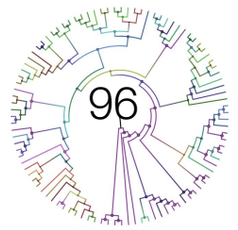


Generating phylogenomic data matrices: hands-on session

Is the polar bear the sister group to the American black bear or the brown bear?

- **MISSING DATA**

Let's create **different matrices with different taxon occupancy** to account for the effect of missing data.



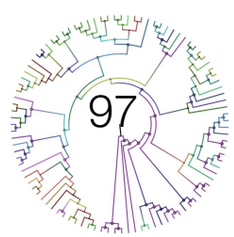
Generating phylogenomic data matrices: hands-on session

Is the polar bear the sister group to the American black bear or the brown bear?

- **MISSING DATA**

Let's create **different matrices with different taxon occupancy** to account for the effect of missing data.

1) The data is located in the folder **Bear_dataset_SensitivityAnalyses/MISSING_DATA**



Generating phylogenomic data matrices: hands-on session

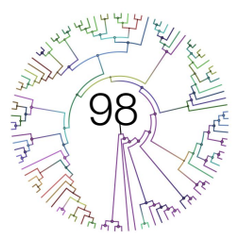
Is the polar bear the sister group to the American black bear or the brown bear?

- **MISSING DATA**

Let's create **different matrices with different taxon occupancy** to account for the effect of missing data.

1) The data is located in the folder **Bear_dataset_SensitivityAnalyses/MISSING_DATA**

2) If you check the list of files in the folder (ls), you'll see that there are 50 orthologous genes ('number.fa'). They're already aligned.



Generating phylogenomic data matrices: hands-on session

Is the polar bear the sister group to the American black bear or the brown bear?

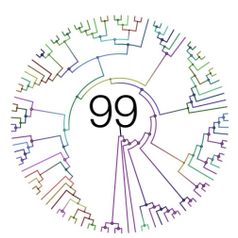
- **MISSING DATA**

Let's create **different matrices with different taxon occupancy** to account for the effect of missing data.

1) The data is located in the folder **Bear_dataset_SensitivityAnalyses/MISSING_DATA**

2) If you check the list of files in the folder (ls), you'll see that there are 50 orthologous genes ('number.fa'). They're already aligned.

3) There are also 3 python scripts. For them to run, we'll need the python libraries **numpy** and **cogent** (already installed in our cluster).



Generating phylogenomic data matrices: hands-on session

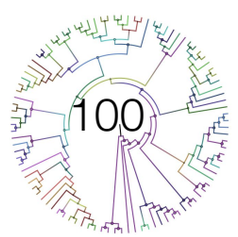
Is the polar bear the sister group to the American black bear or the brown bear?

- **MISSING DATA**

4) Let's explore the amount of missing data that we have in each taxon. Let's run the script:

python count_genesPerSpecies.py

Explore the amount of missing data in each taxon. Which individuals are poorly represented in each species?



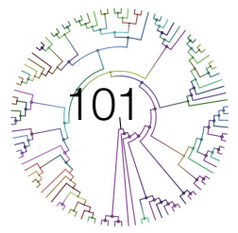
Generating phylogenomic data matrices: hands-on session

Is the polar bear the sister group to the American black bear or the brown bear?

- **MISSING DATA**

5) Now let's select the genes that have a taxon occupancy above a certain threshold (i.e., we want to create a matrix only with the genes that have a minimum of, let's say, 3 species). Let's run the script:

python select_taxon_occupancy.py



Generating phylogenomic data matrices: hands-on session

Is the polar bear the sister group to the American black bear or the brown bear?

- **MISSING DATA**

5) Now let's select the genes that have a taxon occupancy above a certain threshold (i.e., we want to create a matrix only with the genes that have a minimum of, let's say, 3 species). Let's run the script:

python select_taxon_occupancy.py

It will ask you to select the minimum taxon occupancy. Let's start by 3. It will create a folder called '**orthologs_min_[number]_taxa**'. Open it and check how many genes were selected with this threshold.

Generating phylogenomic data matrices: hands-on session

Is the polar bear the sister group to the American black bear or the brown bear?

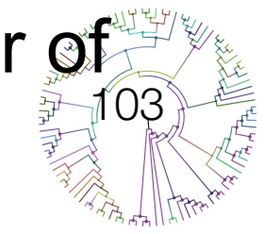
- **MISSING DATA**

5) Now let's select the genes that have a taxon occupancy above a certain threshold (i.e., we want to create a matrix only with the genes that have a minimum of, let's say, 3 species). Let's run the script:

python select_taxon_occupancy.py

It will ask you to select the minimum taxon occupancy. Let's start by 3. It will create a folder called '**orthologs_min_[number]_taxa**'. Open it and check how many genes were selected with this threshold.

Run the script with different thresholds and check how the number of selected genes varies.



Generating phylogenomic data matrices: hands-on session

Is the polar bear the sister group to the American black bear or the brown bear?

- **MISSING DATA**

6) Let's now think again on our goal: to resolve the interrelationships between *Ursus* species. If we select genes just based on taxon occupancy, we may select some that do not include representatives of one or more of the species, and we'll have a strongly biased dataset.

Let's then select genes that have an homogeneous representation of all the four species.

Generating phylogenomic data matrices: hands-on session

Is the polar bear the sister group to the American black bear or the brown bear?

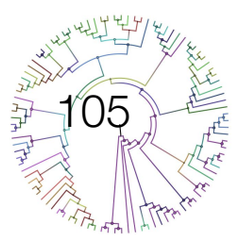
- **MISSING DATA**

7) Let's open the **decisive_genes.py** script and inspect it together.

Notice that at the end of the script we're defining our four species and choosing a minimum number of individuals representing each species in the genes that will be selected (3 in this case).

Run the script:

python decisive_genes.py



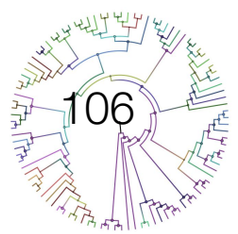
Generating phylogenomic data matrices: hands-on session

Is the polar bear the sister group to the American black bear or the brown bear?

- **MISSING DATA**

8) We now have 2 folders called **‘Decisive_genes3’** and **‘NonDecisive_genes3’**. Check how many genes you have in the ‘Decisive_genes3’ one. Change the threshold in the script, rerun it and check how the selected (=decisive) genes change.

How do you think this may affect the inferred phylogenetic relationships?



Generating phylogenomic data matrices: hands-on session

Is the polar bear the sister group to the American black bear or the brown bear?

- **MISSING DATA**

8) We now have 2 folders called **'Decisive_genes3'** and **'NonDecisive_genes3'**. Check how many genes you have in the 'Decisive_genes3' one. Change the threshold in the script, rerun it and check how the selected (=decisive) genes change.

How do you think this may affect the inferred phylogenetic relationships?

9) Now (or at the end of the class) you can play with these scripts to create different matrices (see guidelines for instructions), run some trees and see how the topology and the support for each node/lineage changes. But first let's learn how to create a matrix with these genes.

Generating phylogenomic data matrices: hands-on session

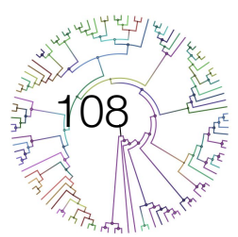
Is the polar bear the sister group to the American black bear or the brown bear?

- **MISSING DATA**

To create your matrix, first you will need to concatenate the genes selected for each matrix. We will use the software **catsequences**

To download it, go back to the folder **Bear_dataset_SensitivityAnalyses** and execute the following command:

git clone https://github.com/ChrisCreevey/catsequences.git
cd consequences
cc catsequences.c -o catsequences -lm



Generating phylogenomic data matrices: hands-on session

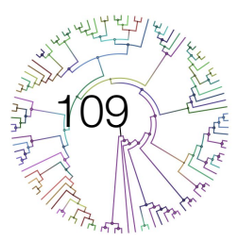
Is the polar bear the sister group to the American black bear or the brown bear?

- **MISSING DATA**

We have to create a file with the list of genes that we'd like to concatenate. Let's create a file with the list of decisive genes. For that, go the folder containing your decisive genes (orthologs_3_fasta) (**cd MISSING_DATA/orthologs_min_3_fasta**) and execute the following command: **ls *.fa > list_decisive_genes.txt**

To concatenate the genes, run the following command: **../../catsequences/catsequences list_decisive_genes.txt**

It will create two files: one with the information of the partitions (**allseqs.partitions.txt**) and the other one with a concatenated fasta with all genes (**allseqs.fas**). This is your matrix!!



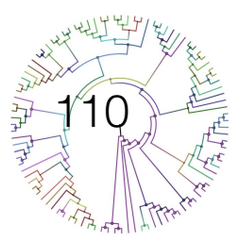
Generating phylogenomic data matrices: hands-on session

Is the polar bear the sister group to the American black bear or the brown bear?

- **MISSING DATA**

Once the genes are concatenated, use one of the phylogenetic inference programs that you learned in the previous days to run a quick tree and test how the phylogeny varies depending on the genes that you include.

Do you see many differences? What is the factor that is affecting the most in the tests you run?

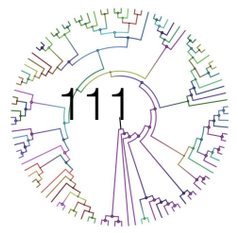


Generating phylogenomic data matrices: hands-on session

Is the polar bear the sister group to the American black bear or the brown bear?

- **OTHER PROPERTIES: genesortR**

10) To further test the robustness of your phylogenomic hypothesis you should also generate matrices accounting for other confounding factors, such as evolutionary rate, compositional heterogeneity, heterotachy, etc.



Generating phylogenomic data matrices: hands-on session

Is the polar bear the sister group to the American black bear or the brown bear?

- **OTHER PROPERTIES: genesortR**

10) To further test the robustness of your phylogenomic hypothesis you should also generate matrices accounting for other confounding factors, such as evolutionary rate, compositional heterogeneity, heterotachy, etc.

There are many softwares to do so that you can explore: **BMGE** (compositional heterogeneity at the level of site), **BaCoCa** (compositional heterogeneity at the level of gene), **TIGER2** (oder genes by evolutionary rate), etc.

Generating phylogenomic data matrices: hands-on session

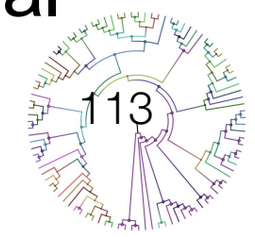
Is the polar bear the sister group to the American black bear or the brown bear?

- **OTHER PROPERTIES: genesortR**

10) To further test the robustness of your phylogenomic hypothesis you should also generate matrices accounting for other confounding factors, such as evolutionary rate, compositional heterogeneity, heterotachy, etc.

There are many softwares to do so that you can explore: **BMGE** (compositional heterogeneity at the level of site), **BaCoCa** (compositional heterogeneity at the level of gene), **TIGER2** (oder genes by evolutionary rate), etc.

We are going to try **genesortR**, an R package that explores several of these properties at the same time.

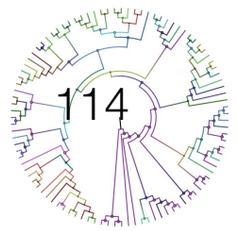


Generating phylogenomic data matrices: hands-on session

Is the polar bear the sister group to the American black bear or the brown bear?

- **OTHER PROPERTIES: genesortR**

11) Let's take our 50 orthogroups and analyze them with **genesortR** to see which ones are the most adequate to analyze. We will use species tree 1 for this analysis.



Generating phylogenomic data matrices: hands-on session

Is the polar bear the sister group to the American black bear or the brown bear?

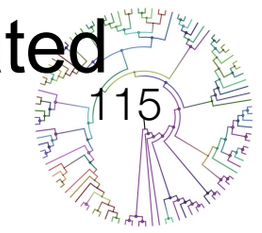
- **OTHER PROPERTIES: genesortR**

11) Let's take our 50 orthogroups and analyze them with **genesortR** to see which ones are the most adequate to analyze. We will use species tree 1 for this analysis.

Data and scripts are located in:

Bear_dataset_SensitivityAnalyses/GENESORTR. Go to that folder.

You will see 3 R scripts, the species tree, the 50 gene alignments concatenated (50_genes.fa), its correspondent partitions file (50_genes.partitions.txt), and the newick gene trees concatenated (50_genes.nwk).



Generating phylogenomic data matrices: hands-on session

Is the polar bear the sister group to the American black bear or the brown bear?

- **OTHER PROPERTIES: genesortR**

12) We will execute genesortR with default parameters on our 50 genes with this command:

module purge
module load r-adephylo r-phytools r-phangorn r-cowplot
Rscript genesortR.R

We'll obtain a copy of our concatenated alignment, partition file and gene trees sorted by their phylogenetic usefulness, from most to least useful.

Take a look at the sorted_figure_50_genes.pdf file obtained. How do gene properties vary according to their phylogenetic usefulness?

Generating phylogenomic data matrices: hands-on session

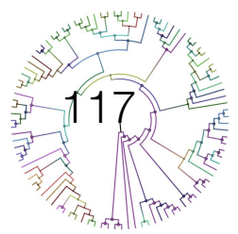
Is the polar bear the sister group to the American black bear or the brown bear?

- **OTHER PROPERTIES: genesortR**

13) We will now test how selecting the *most* and the *least* phylogenetically useful genes affects the tree inferred.

To obtain the 10 best genes run: **Rscript select_10_best_genes.R**

To obtain the 10 worst genes run: **Rscript select_10_worst_genes.R**



Generating phylogenomic data matrices: hands-on session

Is the polar bear the sister group to the American black bear or the brown bear?

- **OTHER PROPERTIES: genesortR**

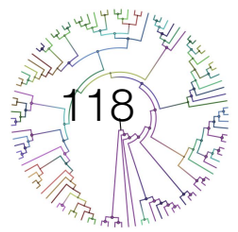
13) We will now test how selecting the *most* and the *least* phylogenetically useful genes affects the tree inferred.

To obtain the 10 best genes run: **Rscript select_10_best_genes.R**

To obtain the 10 worst genes run: **Rscript select_10_worst_genes.R**

Once we have obtained the smaller alignments, use one of the phylogenetic inference programs that you have used before to run a quick tree and test how the phylogeny varies when using genes with different phylogenetic usefulness.

Do you see any differences?



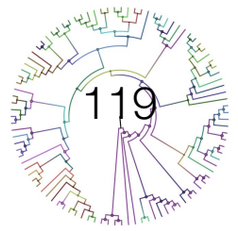
So… what do we do to test the robustness of our tree?

✓ **Build different subsets of your data through a subsampling strategy selecting genes with different properties**

2) **Run different analyses that rely on different assumptions or apply different models**

3) **Do 1) and 2) both at the level of *supermatrix* and *subset of individual gene trees***

THIS IS THE MOST IMPORTANT TAKE-HOME MESSAGE FROM TODAY'S LESSON!!



So… what do we do to test the robustness of our tree?

✓ **Build different subsets of your data through a subsampling strategy selecting genes with different properties**

2) Run different analyses that rely on different assumptions or apply different models

3) Do 1) and 2) both at the level of *supermatrix* and *subset of individual gene trees*

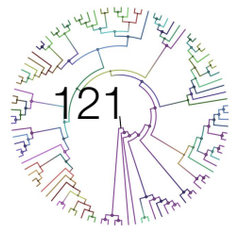
So… what do we do to test the robustness of our tree?

✓ **Build different subsets of your data through a subsampling strategy selecting genes with different properties**

2) Run different analyses that rely on different assumptions or apply different models

3) Do 1) and 2) both at the level of *supermatrix* and *subset of individual gene trees*

We're not going to focus on this today, but it would involve analyses with different assumptions or models that you can explore in the following software:



So… what do we do to test the robustness of our tree?

✓ **Build different subsets of your data through a subsampling strategy selecting genes with different properties**

2) Run different analyses that rely on different assumptions or apply different models

3) Do 1) and 2) both at the level of *supermatrix* and *subset of individual gene trees*

We're not going to focus on this today, but it would involve analyses with different assumptions or models that you can explore in the following software:

➢ Comparing Maximum Likelihood vs Bayesian Inference vs Parsimony (with some of the software you've already learnt)

So… what do we do to test the robustness of our tree?

✓ **Build different subsets of your data through a subsampling strategy selecting genes with different properties**

2) Run different analyses that rely on different assumptions or apply different models

3) Do 1) and 2) both at the level of *supermatrix* and *subset of individual gene trees*

We're not going to focus on this today, but it would involve analyses with different assumptions or models that you can explore in the following software:

➢ Comparing Maximum Likelihood vs Bayesian Inference vs Parsimony (with some of the software you've already learnt)

➢ Comparing more simple models (eg, LG) vs more complex ones (eg, CAT)

So… what do we do to test the robustness of our tree?

✓ **Build different subsets of your data through a subsampling strategy selecting genes with different properties**

2) Run different analyses that rely on different assumptions or apply different models

3) Do 1) and 2) both at the level of *supermatrix* and *subset of individual gene trees*

We're not going to focus on this today, but it would involve analyses with different assumptions or models that you can explore in the following software:

- ➢ Comparing Maximum Likelihood vs Bayesian Inference vs Parsimony (with some of the software you've already learnt)
- ➢ Comparing more simple models (eg, LG) vs more complex ones (eg, CAT)
- ➢ IQ-TREE2: GHOST model -> heterotachy

GHOST: Recovering Historical Signal from Heterotachously Evolved Sequence Alignments FREE

Stephen M Crotty ✉, Bui Quang Minh, Nigel G Bean, Barbara R Holland, Jonathan Tuke, Lars S Jermiin, Arndt Von Haeseler Author Notes

Systematic Biology, Volume 69, Issue 2, March 2020, Pages 249–264,
https://doi.org/10.1093/sysbio/syz051
Published: 31 July 2019 **Article history** ▾

So… what do we do to test the robustness of our tree?

✓ **Build different subsets of your data through a subsampling strategy selecting genes with different properties**

✓ **Run different analyses that rely on different assumptions or apply different models**

3) Do 1) and 2) both at the level of *supermatrix* and *subset of individual gene trees*

Analyzing gene tree/species tree conflict: multispecies coalescent

- The use of **transcriptomic and genomic datasets** for phylogenetic reconstruction has become increasingly common as researchers attempt to resolve recalcitrant nodes with **increasing amounts of data**.

Analyzing gene tree/species tree conflict: multispecies coalescent

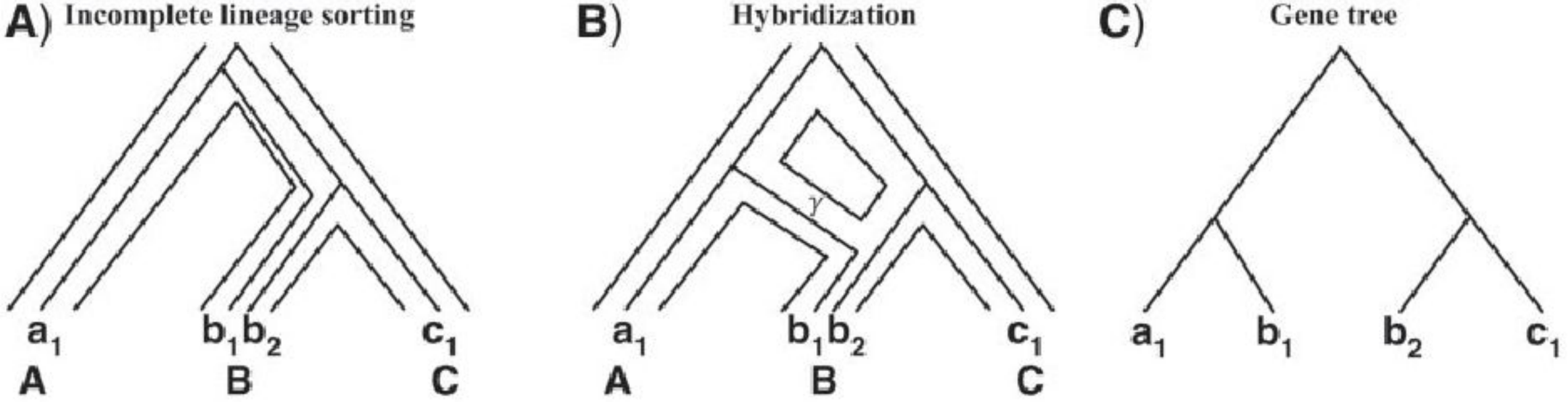
- The use of transcriptomic and genomic datasets for phylogenetic reconstruction has become increasingly common as researchers attempt to resolve recalcitrant nodes with increasing amounts of data.
- The **large size and complexity** of these datasets introduce **significant phylogenetic noise and conflict** into subsequent analyses.

Analyzing gene tree/species tree conflict: multispecies coalescent

- The use of transcriptomic and genomic datasets for phylogenetic reconstruction has become increasingly common as researchers attempt to resolve recalcitrant nodes with increasing amounts of data.
- The large size and complexity of these datasets introduce significant phylogenetic noise and conflict into subsequent analyses.
- The **sources of conflict** may include **hybridization, incomplete lineage sorting, or horizontal gene transfer**, and may vary across the phylogeny.

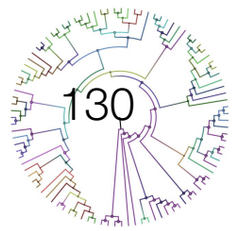
Analyzing gene tree/species tree conflict: multispecies coalescent

- The use of transcriptomic and genomic datasets for phylogenetic reconstruction has become increasingly common as researchers attempt to resolve recalcitrant nodes with increasing amounts of data.
- The large size and complexity of these datasets introduce significant phylogenetic noise and conflict into subsequent analyses.
- The **sources of conflict** may include **hybridization, incomplete lineage sorting, or horizontal gene transfer**, and may vary across the phylogeny.



Analyzing gene tree/species tree conflict: multispecies coalescent

- The use of transcriptomic and genomic datasets for phylogenetic reconstruction has become increasingly common as researchers attempt to resolve recalcitrant nodes with increasing amounts of data.
- The large size and complexity of these datasets introduce significant phylogenetic noise and conflict into subsequent analyses.
- The sources of conflict may include hybridization, incomplete lineage sorting, or horizontal gene transfer, and may vary across the phylogeny.
- For **phylogenetic analysis**, this noise and conflict has been accommodated in one of several ways:
  - by **binning gene regions into subsets** to isolate consistent phylogenetic signal
  - by using **gene-tree methods for reconstruction**, where conflict is presumed to be explained by incomplete lineage sorting (ILS)
  - through **concatenation**, where noise is presumed to be the dominant source of conflict.

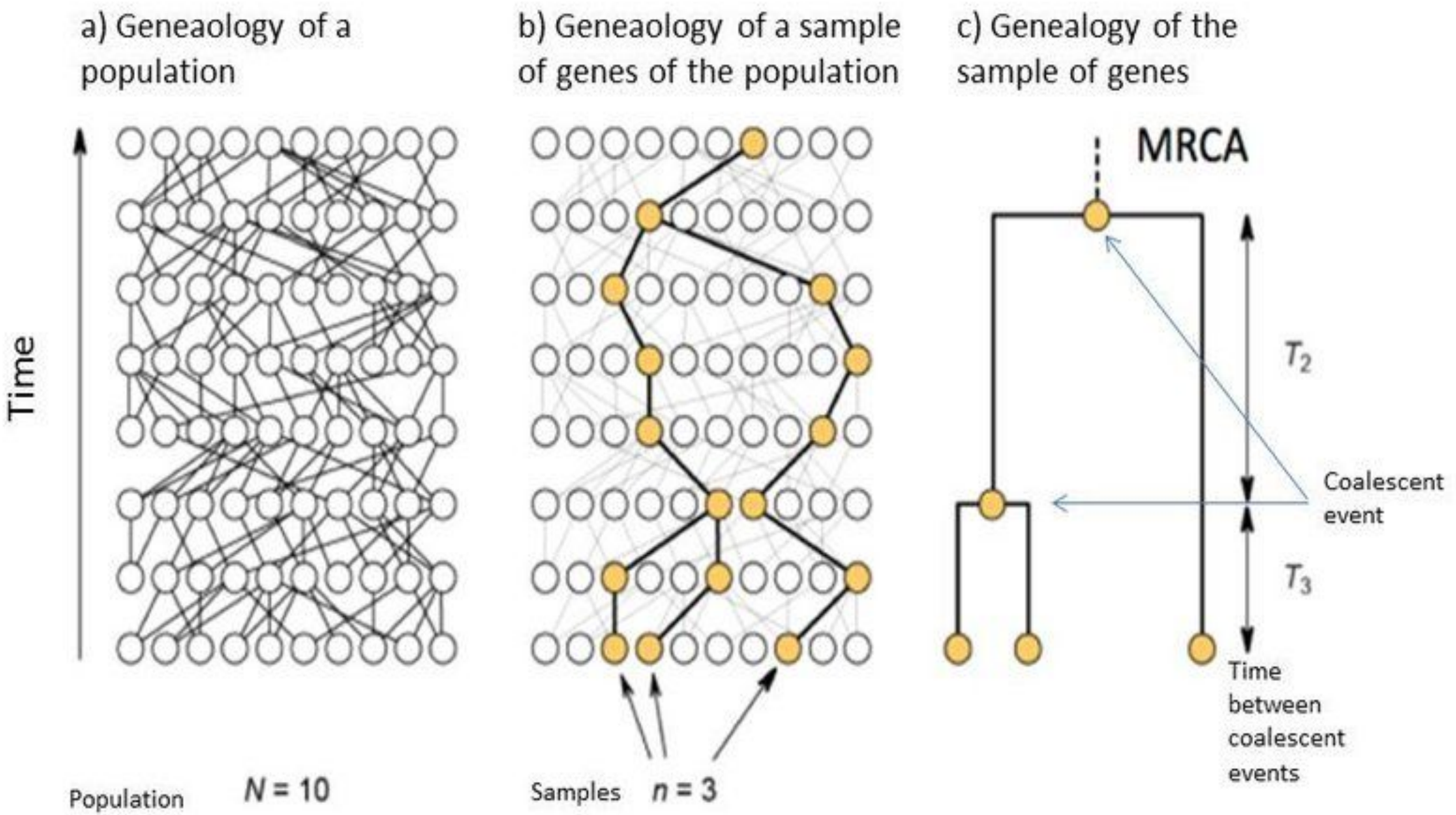


Analyzing gene tree/species tree conflict: multispecies coalescent

- The use of transcriptomic and genomic datasets for phylogenetic reconstruction has become increasingly common as researchers attempt to resolve recalcitrant nodes with increasing amounts of data.
- The large size and complexity of these datasets introduce significant phylogenetic noise and conflict into subsequent analyses.
- The sources of conflict may include hybridization, incomplete lineage sorting, or horizontal gene transfer, and may vary across the phylogeny.
- For phylogenetic analysis, this noise and conflict has been accommodated in one of several ways:
  - by binning gene regions into subsets to isolate consistent phylogenetic signal
  - by using gene-tree methods for reconstruction, where conflict is presumed to be explained by incomplete lineage sorting (ILS)
  - through concatenation, where noise is presumed to be the dominant source of conflict.
- The analysis of individual homologous gene regions can greatly improve our understanding of the underlying conflict within these datasets.

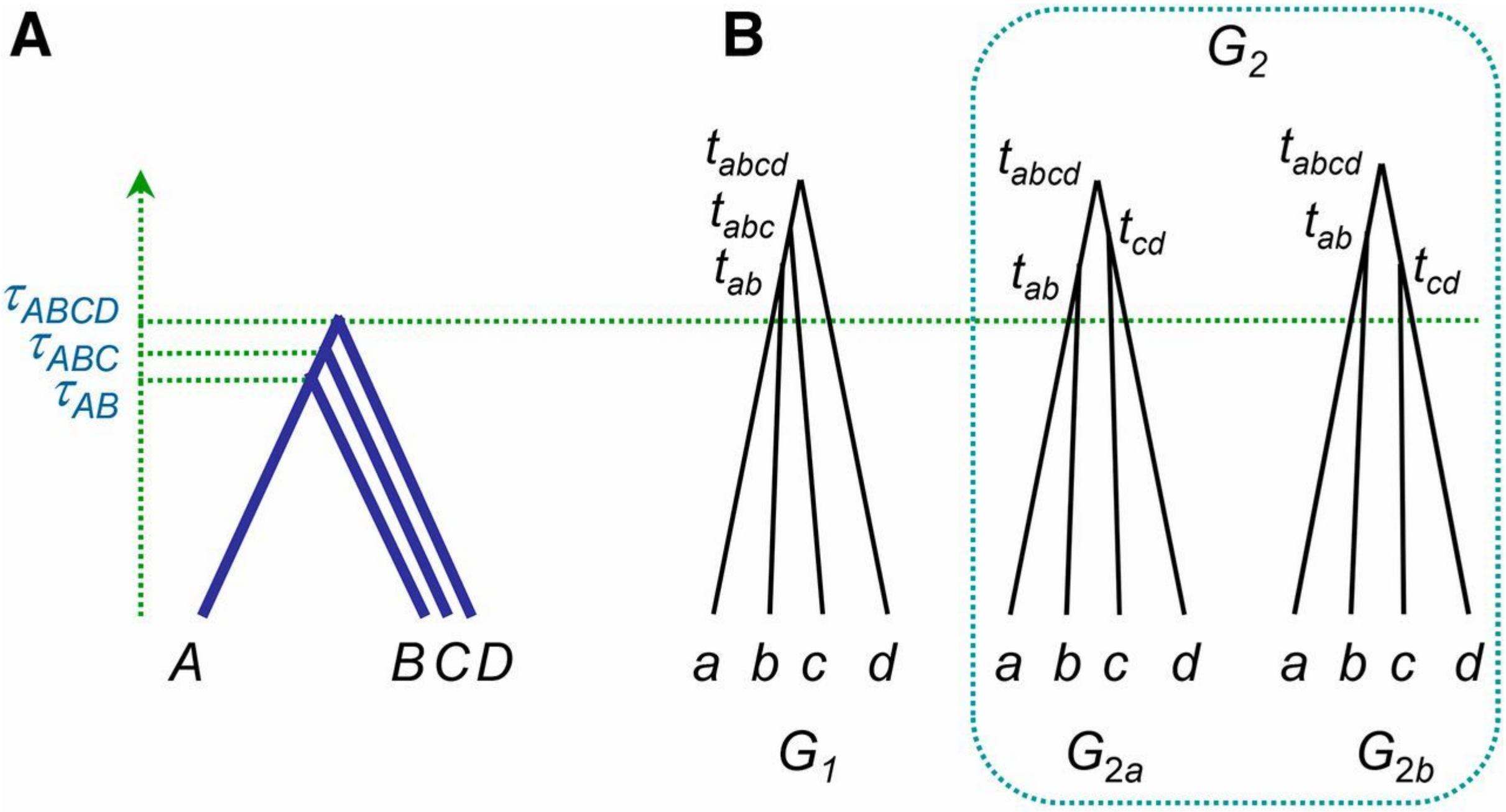
Analyzing gene tree/species tree conflict: multispecies coalescent

Brief introduction to coalescent theory



(more on the coalescent theory next week with Julio Rozas)

Analyzing gene tree/species tree conflict: multispecies coalescent

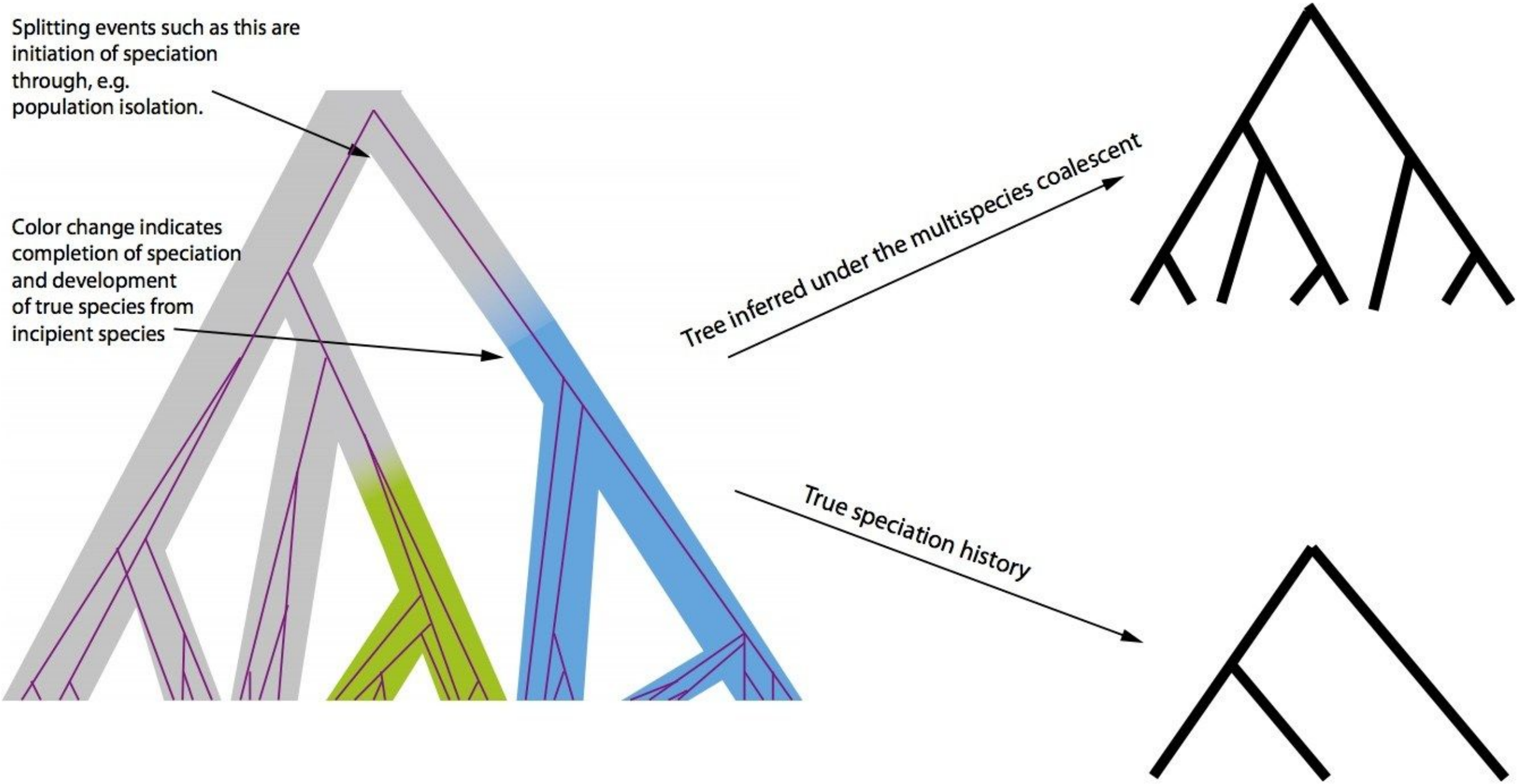


Species tree *S*

Three labelled histories

Analyzing gene tree/species tree conflict: multispecies coalescent

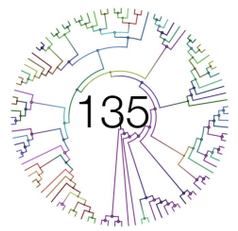
The **Multispecies Coalescent** model provides a framework for inferring species phylogenies while accounting for ancestral polymorphism and gene tree-species tree conflict.



Analyzing gene tree/species tree conflict: multispecies coalescent

Many different softwares:

- ASTRAL-III
- PoMo (allele-frequency based, does not use the multispecies coalescent model)
- STAR
- BPP
- etc.



Analyzing gene tree/species tree conflict: hands-on session

Many different softwares:

- ASTRAL-III
- PoMo (allele-frequency based, does not use the multispecies coalescent model)
- STAR
- BPP
- etc.

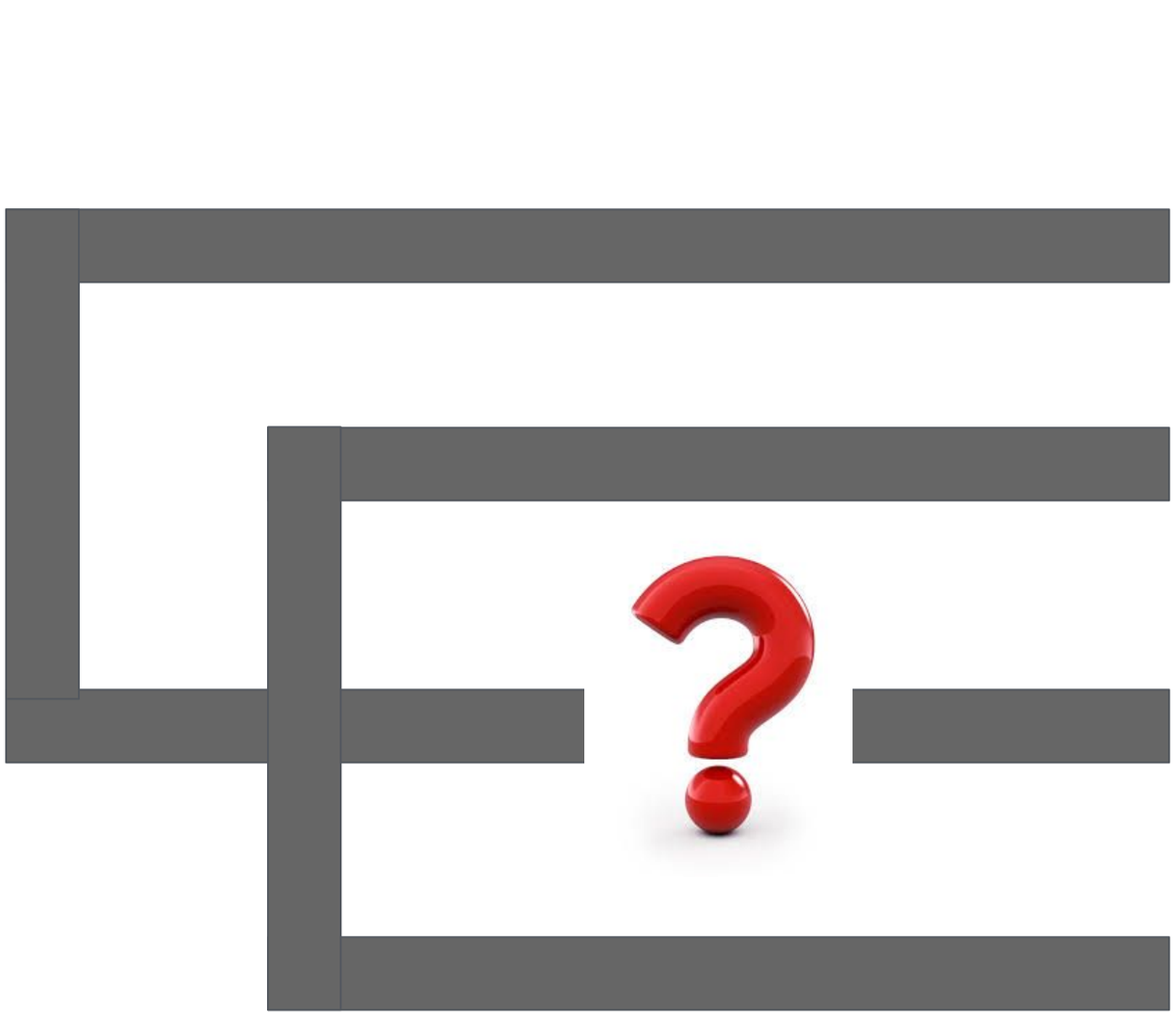
Analyzing gene tree/species tree conflict: hands-on session

Is the polar bear the sister group to the American black bear or the brown bear?

ASTRAL is a tool for estimating an unrooted species tree given a set of unrooted gene trees.
ASTRAL is statistically consistent under the multi-species coalescent model (and thus is useful for handling incomplete lineage sorting, i.e., ILS).
ASTRAL finds the species tree that has the maximum number of shared induced quartet trees with the set of gene trees, subject to the constraint that the set of bipartitions in the species tree comes from a predefined set of bipartitions.

Analyzing gene tree/species tree conflict: hands-on session

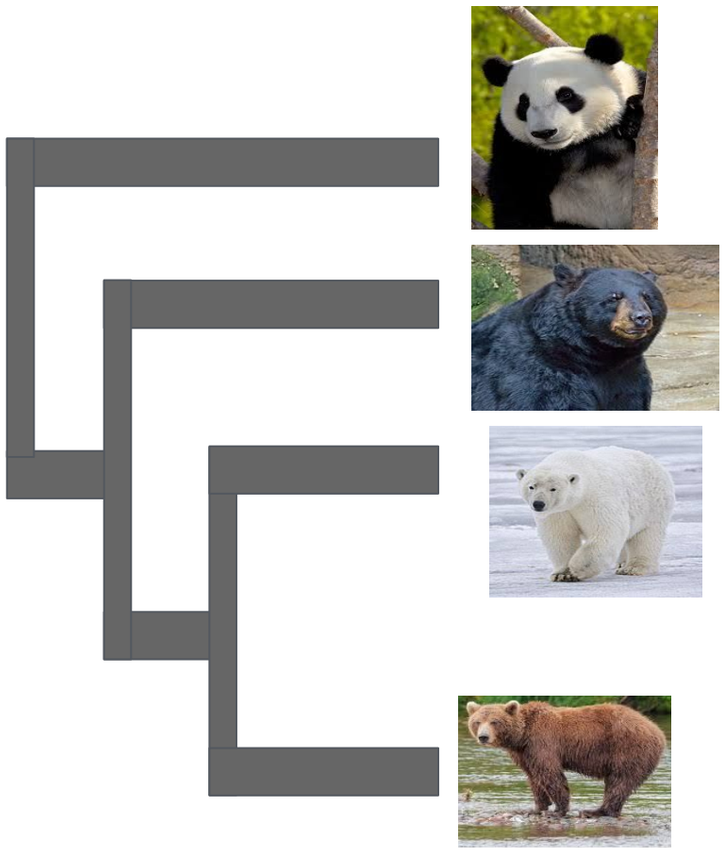
Is the polar bear the sister group to the American black bear or the brown bear?



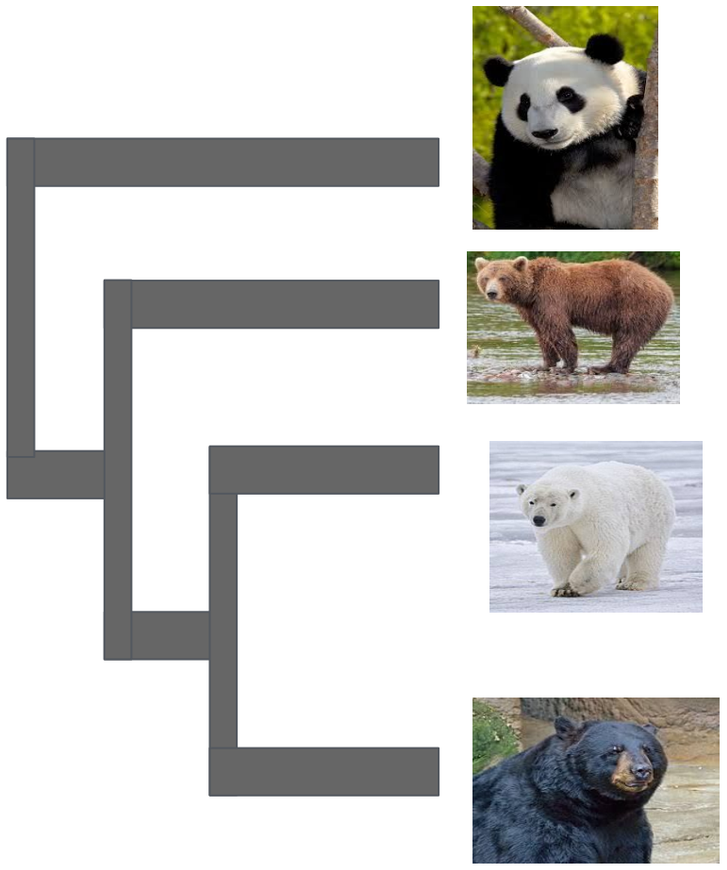
Analyzing gene tree/species tree conflict: hands-on session

Is the polar bear the sister group to the American black bear or the brown bear?

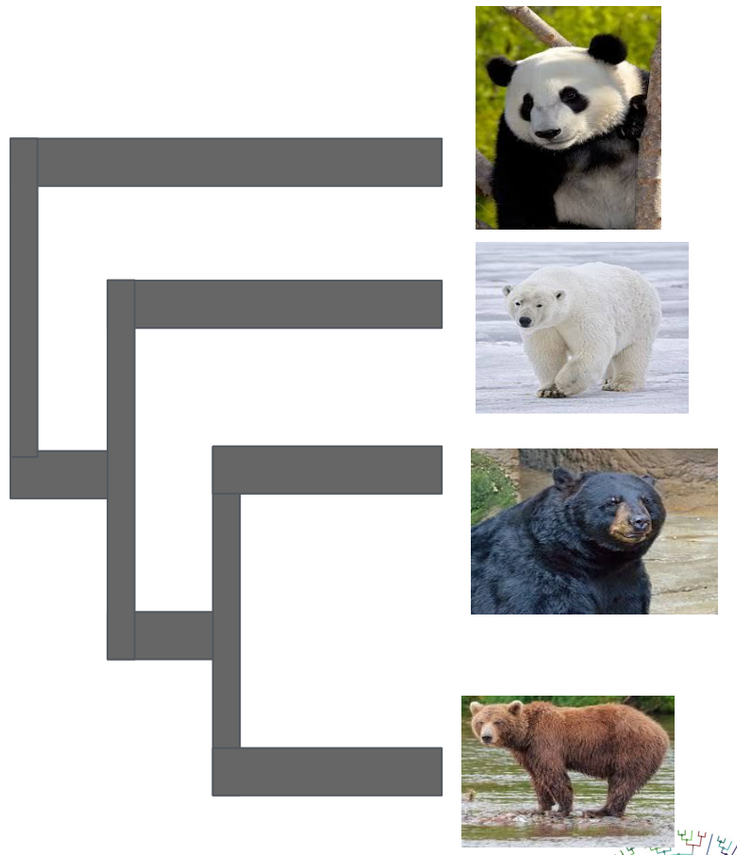
1) Let's analyze conflict between individual gene trees to see which phylogenetic hypothesis is the most robustly supported:



Species Tree 1



Species Tree 2



Species Tree 3

Analyzing gene tree/species tree conflict: hands-on session

Is the polar bear the sister group to the American black bear or the brown bear?

2) We have selected 50 orthologous genes and have run individual gene trees with IQTree. Let's have a look at them here:

Bear_dataset_SensitivityAnalyses/PHYLIP (.tree files)

Analyzing gene tree/species tree conflict: hands-on session

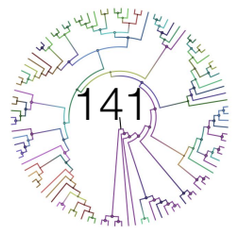
Is the polar bear the sister group to the American black bear or the brown bear?

2) We have selected 50 orthologous genes and have run individual gene trees with IQTree. Let's have a look at them here:

Bear_dataset_SensitivityAnalyses/PHYLIP (.tree files)

3) ASTRAL-III needs all gene trees in the same file. For that, let's concatenate them:

cat *trees > bears_allTrees.tre



Analyzing gene tree/species tree conflict: hands-on session

Is the polar bear the sister group to the American black bear or the brown bear?

2) We have selected 50 orthologous genes and have run individual gene trees with IQTree. Let's have a look at them here:

Bear_dataset_SensitivityAnalyses/PHYLIP (.tree files)

3) ASTRAL-III needs all gene trees in the same file. For that, let's concatenate them:

cat *trees > bears_allTrees.tre

And create a folder called ASTRAL and move the files there:

mkdir ASTRAL

mv bears_allTrees.tre ASTRAL

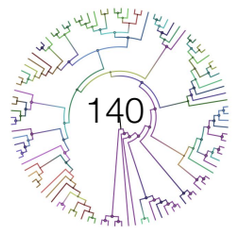
And go to that folder: **cd ASTRAL**

Analyzing gene tree/species tree conflict: hands-on session

Is the polar bear the sister group to the American black bear or the brown bear?

4) Let's now run an analysis on the 50 individual gene trees:

astral -i bears_allTrees.tre



Analyzing gene tree/species tree conflict: hands-on session

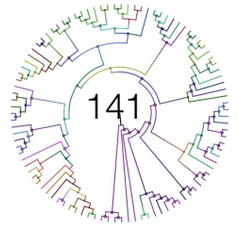
Is the polar bear the sister group to the American black bear or the brown bear?

4) Let's now run an analysis on the 50 individual gene trees:

astral -i bears_allTrees.tre

You'll see the output in the screen. To save it in an output file run:

astral -i bears_allTrees.tre 2> output_ASTRAL.txt



Analyzing gene tree/species tree conflict: hands-on session

Is the polar bear the sister group to the American black bear or the brown bear?

4) Let's now run an analysis on the 50 individual gene trees:

astral -i bears_allTrees.tre

You'll see the output in the screen. To save it in an output file run:

astral -i bears_allTrees.tre 2> output_ASTRAL.txt

Examine the output. What is the optimal tree inferred by ASTRAL? What is the final normalized quarted score?

->The normalized quartet score is the proportion of input gene tree quartet trees satisfied by the species tree. This is a number between zero and one; the higher this number, the *less* discordant your gene trees are.

Analyzing gene tree/species tree conflict: hands-on session

Is the polar bear the sister group to the American black bear or the brown bear?

4) Let's now run an analysis on the 50 individual gene trees:

astral -i bears_allTrees.tre

You'll see the output in the screen. To save it in an output file run:

astral -i bears_allTrees.tre 2> output_ASTRAL.txt

Examine the output. What is the optimal tree inferred by ASTRAL? What is the final normalized quarted score?

->The normalized quartet score is the proportion of input gene tree quartet trees satisfied by the species tree. This is a number between zero and one; the higher this number, the *less* discordant your gene trees are.

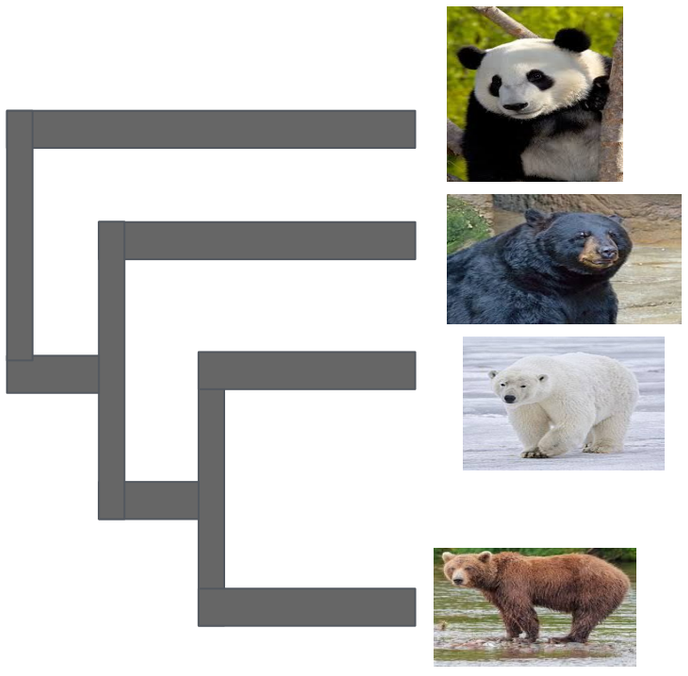
Visualize it in your preferred software or online in **phylo.io** (just copy-paste the newick tree and click 'Render').

Analyzing gene tree/species tree conflict: hands-on session

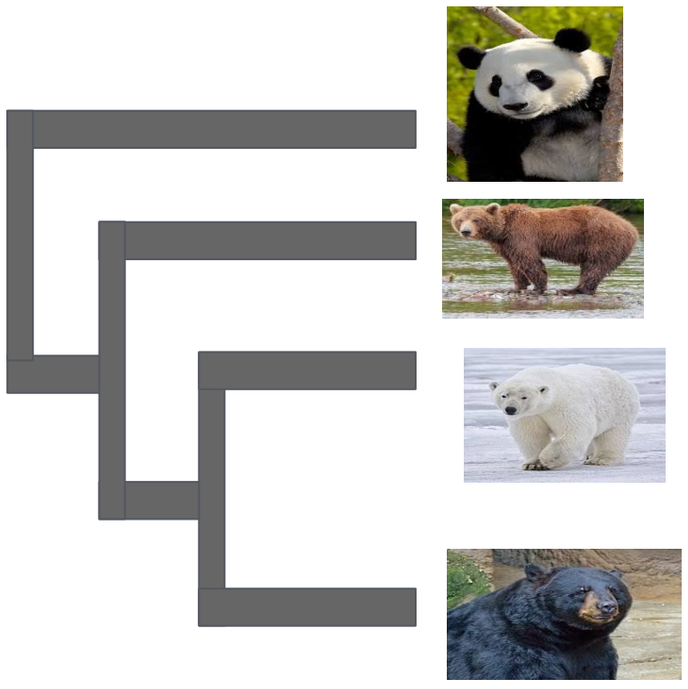
Is the polar bear the sister group to the American black bear or the brown bear?

5) So far ASTRAL showed us the preferred togology. Let's now check how our individual gene trees support the alternatives topologies-

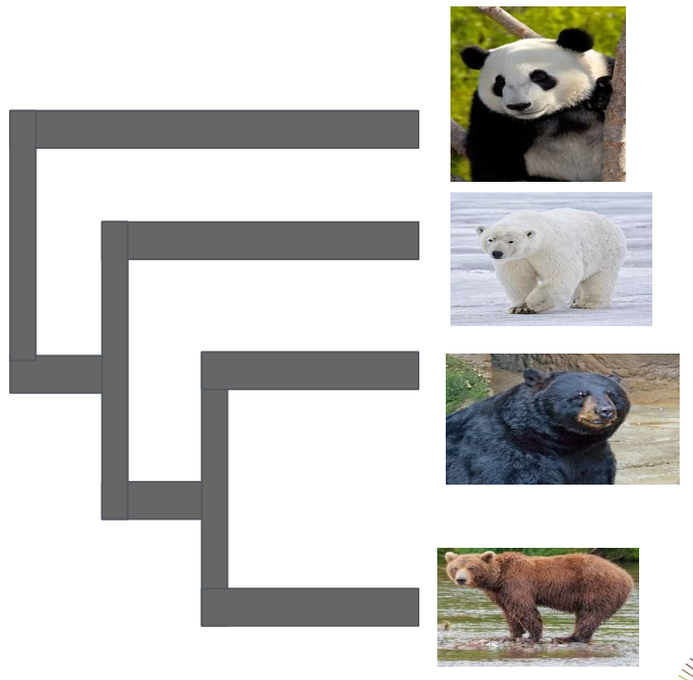
For that, let's score each species tree topology and compare the normalize quartet score for each one.



Species Tree 1



Species Tree 2



Species Tree 3

Analyzing gene tree/species tree conflict: hands-on session

Is the polar bear the sister group to the American black bear or the brown bear?

5) So far ASTRAL showed us the preferred togology. Let's now check how our individual gene trees support the alternatives topologies-

For that, let's score each species tree topology and compare the normalize quartet score for each one.

Go to the folder
Bear_dataset_SensitivityAnalyses/SPECIES_TREES

and check the three provided species trees. Visualize them and identify the differences.

Analyzing gene tree/species tree conflict: hands-on session

Is the polar bear the sister group to the American black bear or the brown bear?

5) So far ASTRAL showed us the preferred togology. Let's now check how our individual gene trees support the alternatives topologies-

For that, let's score each species tree topology and compare the normalize quartet score for each one.

Go to the folder
Bear_dataset_SensitivityAnalyses/SPECIES_TREES

and check the three provided species trees. Visualize them and identify the differences.

Let's now score them with ASTRAL.

Analyzing gene tree/species tree conflict: hands-on session

Is the polar bear the sister group to the American black bear or the brown bear?

6) Let's score the first species tree. From the ASTRAL folder, run:

astral -i bears_allTrees.tre -q ../SPECIES_TREES/bear_species_tree1.tre 2> score_speciesTree1.txt

Analyzing gene tree/species tree conflict: hands-on session

Is the polar bear the sister group to the American black bear or the brown bear?

6) Let's score the first species tree. From the ASTRAL folder, run:

astral -i bears_allTrees.tre -q ../SPECIES_TREES/bear_species_tree1.tre 2> score_speciesTree1.txt

Do the same with the species trees 2 and 3.

Analyzing gene tree/species tree conflict: hands-on session

Is the polar bear the sister group to the American black bear or the brown bear?

6) Let's score the first species tree. From the ASTRAL folder, run:

astral -i bears_allTrees.tre -q ../SPECIES_TREES/bear_species_tree1.tre 2> score_speciesTree1.txt

Do the same with the species trees 2 and 3.

Compare the results. Which phylogenetic hypothesis is the most robustly supported?

Which branches are not supported by many genes in each analyses? Does this affect the overall preferred phylogeny of *Ursus*?

Phylogenomics and Population Genomics:
Inference and Applications

Testing the robustness of your phylogeny: sensitivy analyses

Gemma I. Martínez-Redondo & Rosa Fernández

Biodiversity Program
Institute of Evolutionary Biology (CSIC-UPF)

rosa.fernandez@ibe.upf-csic.es

www.metazomics.com